Animal Minds in Time

Simon Alexander Burns Brown

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Animal Minds in Time
Simon Brown

Humans, octopuses, bees, scrub jays, and dogs have all been shown to have rich cognitive capacities, but their minds work very differently. What are the most important factors contributing to the kind of mind an animal has? I show that the ways in which an animal relates to time—its ways of representing the temporal and the nature of its memory—are crucial to the rest of its capacities. I develop empirically-informed philosophical accounts of the natures of episodic memory, and of the representation of temporal properties, temporal frameworks, and narratives. And I use these accounts to interpret the relevant empirical literature, and to show how these capacities can transform the extent to which animals can understand the complexity of the world.

Table of Contents

Acknowledgments

This dissertation would not have been possible without the support, advice, and insight of a huge number of people. I would like to extend special thanks to Christopher Peacocke and John Morrison for advising me and helping me in many ways over the years. I would also like to thank everyone who has read and discussed earlier versions of different parts of this dissertation with me, including especially Peter Godfrey-Smith, Jenann Ismael, and Katie Tabb. Many people deserve mention for contributing friendship with huge contributions to my intellectual development, but perhaps above all Marjorie Xie, Matt Heeney, Kate Pendoley, Nemira Gasiunas, Jorge Morales, and Andrew Richmond. I have also benefitted greatly from everyone involved in PoPRocks, the Columbia Philosophy of Neuroscience group, and the Proposal Preparation and Dissertation Seminars at Columbia. Various parts of this dissertation have also benefited from discussions with audience members and commentators at the Joint Session in Cardiff in 2016, the Memory, Mental Time Travel, and Self-Control Workshop in Rome in 2017, The SSPP in Cincinnati in 2019, the TIF in Valencia in 2019, Evolution Evolving at Cambridge in 2019, the International Symposium on Thinking In and About Time in Milan, 2019, Issues in Philosophy of Memory 2 in Grenoble, 2019, and the APA Eastern Division in Philadelphia, 2020. Finally, for their support and patience I would like to thank Kirsty Brown, Matthew Brown, my parents Colin Brown and Ann Davison (who were the first to introduce me to philosophy), and Cromarty Brown (who was the first to introduce me to the importance and interest of the minds of different species).

Chapter 1: Questions of Cognitive Significance

1.1 The Cognitive Significance Question

What is the difference between humans and other animals? Many have offered answers to this question. Aristotle thinks the answer is that we are rational.[1] Descartes is more extreme, claiming that while substance dualism is true of humans, non-human animals are simply machines, incapable of thoughts.[2] Hobbes attributes more sophistication to animals than this, but emphasizes that they lack language.[3] Locke emphasizes instead that they lack the power of abstraction.[4] It is not just philosophers who make such claims. Consider the following two passages from Steinbeck's *East of Eden*:

> "At such a time it seems natural and good to me to ask myself these questions. What do I believe in? What must I fight for and what must I fight against?
> Our species is the only creative species, and it has only one creative instrument, the individual mind and spirit of a man."[5]

> "think of the glory of the choice! That makes a man a man. A cat has no choice, a bee must make honey. There's no godliness there."[6]

[1] Aristotle, *Metaphysics* I.1 980a25-981a12; *Nicomachean Ethics* I.7 1098a1-1098a17, I.13 1102b12-1103b4. Many philosophers followed him in this, albeit with various elaborations. For two particularly influential cases, see Avicenna, whose views on this issue are summarized by Alwishah 2016, and Aquinas, *Summa Contra Gentiles* 3.2.112.

[2] Descartes' most explicit, worked-out discussions of this issue are in letters (e.g. Descartes 1646), although he also summarizes his position in the *Discourse on Method* (Descartes 1911: 116-118).

[3] Hobbes 1991.

[4] Locke (1689: I.11.10-11).

[5] Steinbeck (2001: 446).

[6] Steinbeck (2001: 630).

And UK Prime Minister Edward Heath once defined his political philosophy as: "What distinguishes man from animals is his desire and his ability to control and to shape his environment."[7]

Contemporary philosophers (amongst others) still make such claims, ranging from claims offered without any empirical support, such as Frankfurt (1971: 6f.)'s suggestion that only humans have second-order desires, to suggestions made on the basis of detailed empirical arguments, like Stotz (2010: 488)'s suggestion that "what is most distinctive about humans is the reaction of extremely developmentally plastic brains to a total immersion and involvement into a well-engineered, cumulatively constructed cognitive–developmental niche".

However, the question as I formulated it at the outset is a bad question. It pushes us in an unhelpful direction, because it makes (or at least strongly suggests) several closely related presuppositions.

Firstly, it presupposes that there is One Big Difference, *the* difference between humans and other animals. It *might* turn out that there is truth in this: it might turn out that one feature of humans has consequences which ramify in many directions, and explains all the other differences between humans and other animals. For example, it might turn out that having Reason provides for the ability to talk, think, be creative, and all the rest. But we should not presuppose that the answer will turn out this way. It should be a live alternative that there is a constellation of smaller differences, mutually reinforcing and interacting with one another, without any one being more important than all the rest.

Secondly, thinking about things as 'humans and other animals' pushes us towards thinking that 'other animals' picks out a unified class. But there is a huge variety in non-human

[7] Langdon 2015.

animals, including in their minds. The differences between octopuses, pigs, bees, and owls will be just as great as the differences between humans and other great apes. Neglecting this variety is likely to distort our picture of animals and of humans' place in the animal kingdom.

Thirdly, thinking of non-human animals as forming a unified class tends to be embedded in a hierarchical understanding of kinds of mind. It suggests the idea that we can think of human minds as animal minds plus some other feature, such as rationality. And if one thinks *that*, then it is natural to try to recognize diversity within animal minds by introducing not a complicated distribution of different mental features across the tree of life, but a scale of nature which evolution marched up in a linear fashion, from creatures with no minds, through creatures with increasingly complex minds (vertebrates, mammals, primates…), until it reached humans. But there are many different kinds of mind, specialized in different ways for different niches. None of these minds is the best adapted in all respects. As Hull (1969: 248) memorably put it, "Man is very efficient, flexible in his adaptiveness, and so on. So are cockroaches. Man is unique. So are cockroaches."

The question 'what is the difference between humans and other animals?' is a bad question, but it picks up on important issues. We should not treat all non-human minds as a mass. But neither should we treat them as a mess. We should not do away with cross-species comparisons and simply focus on human psychology, octopus psychology, owl psychology etc. independently of one another.[8] It is worth studying the mind in general, and this means studying the different forms it can take and their relationships to one another. There is huge variety in kinds of minds, but there are important, systematic patterns in that variety. And while we should not assume there is One Big Difference, we should expect that some differences will be more

[8] Cf. Buckner (2013: 856).

important than others. But what question should replace our opening question, if we are to capture all this?

One way of reframing our question is to look to evolution. Maynard Smith and Szathmáry (1995) discuss what they claim are 'major transitions in evolution', including the emergence of RNA, multicellular organisms, and language. Perhaps we should look for those mental capacities whose emergence constituted major evolutionary transitions in the mind. What are major evolutionary transitions? Maynard Smith and Szathmáry describe such transitions as involving changes in "the way in which genetic information is transmitted between generations."[9] In a closely related tradition, Dennett (1995) and Godfrey-Smith (2018) discuss the emergence of increasingly complex forms of learning during evolution, on the grounds that these can be seen as "ways of realizing the Darwinian pattern on different scales and with different degrees of sophistication".[10] So perhaps we could replace our initial question with something like "What are the features of the mind which mark a major transition in evolution, and which animals have which of these features?"

This is an important question. However, it may be too narrow to capture the full underlying interest behind "what is the difference between humans and other animals?" Many traits will make a big difference to how the mind works and what it can do, without fundamentally changing the way *evolution* operates. Take Locke's power of abstraction. It would not seem to detract from the interest of Locke's suggestion if it turned out that the power of

[9] Maynard Smith & Szathmáry (1995: 3). Within biology, and philosophy of biology, the emphasis is often on slightly different properties of the transitions Maynard Smith & Szathmáry discuss. Michod 1999 and Godfrey-Smith (2009: 7-8, 121-128) emphasize such as changes in the scale of the individuals who are being operated on by natural selection (e.g. individual cells, operating in concert with one another, vs. a genuinely multicellular organism, whose individual cells are unable to reproduce on their own). Calcott & Sterelny 2011 emphasize changes in "the core components of the evolutionary process" more generally.

[10] Godfrey-Smith (2018: 226).

abstraction did not change the way information is transmitted, the way natural selection operates, or the units of selection. Furthermore, restricting our attention to evolutionary issues would seem to limit the kinds of mind we can focus on. We should not rule out the possibility that there could be genuine minds in artificial systems, which have interestingly different forms of minds to humans. And if they do so, it will be because they either lack capacities that humans have or have capacities that humans do not have, even if they are not subject to Darwinian natural selection at all.

We can reframe our original question in a different way: what are the most cognitively significant traits, and which species have them?[11]

This will be the guiding question of this dissertation, so it is worth unpacking a little. The key notion is 'cognitive significance'. I will use this in the following way: a trait is cognitively significant relative to a kind of mind M to the degree that adding it to M transforms M's capabilities. To explain what this means, I will go through examples of cognitively significant traits, then explain the relationship between cognitive significance and other kinds of significance. This will lay the groundwork for explaining why we have an important question here—why we should care about cognitive significance, and what the underlying interest of the initial human uniqueness question really was.[12]

---

[11] I will be using 'mental trait' as a shorthand for something like 'mental state, system, capacity, feature, or process'. Often philosophers do the work I want to do with this term by using 'mental state' to cover all of these. In many contexts this is unproblematic, but when thinking about the dynamics of the mental, as we will be doing in this dissertation, it can be. So I prefer the more neutral 'trait'. This term has baggage of its own, which needs to be disowned. One should not read my use of 'trait' as applying only to character traits or only to lifelong, fixed features of individuals' minds. Furthermore, one should not assume that traits in my sense will always be heritable (§3.5.6 discusses this issue further).

[12] This is not to claim that the only motivation people have had for asking about human uniqueness is curiosity about how minds are different from one another: sometimes claims about human uniqueness are used for other purposes, such as exhorting humans to behave in certain ways because these ways are allegedly distinctive of humans (a tradition which goes back to Aristotle, *Nicomachean Ethics* I.7), or for a kind of self-flattery, or just to sound grandiose. But such motivations have typically been more effective at prompting pronouncements about the

One way of being cognitively significant—of transforming a mind's capabilities—would be dramatically expanding the range of things the mind can represent. For example, we might think (though Chapter 2 will consider reasons for thinking otherwise) that adding the ability to represent duration would have a big impact on the kinds of things a mind could do. It might allow for precisely timing actions and (if the creature is also able to perform division) one way of estimating rates of change. Gaining the ability to perform division on represented magnitudes would be an example of another way of transforming a mind's capabilities—expanding the range of cognitive *operations* that a creature can perform. And we can transform the capabilities of a mind by expanding both the operations which it can in principle perform, and its efficiency in doing so. If a creature can learn a wide range of actions, but only after thousands of trials, then radically cutting this number would be highly significant.

By contrast, gaining the ability to represent the concentration of nitrogen levels in the air would (assuming this ability did not bring with it a general understanding of chemistry) be rather insignificant for most animals, as there would be little they could do with this information. They would not gain new kinds of understanding or perform new mental operations: they would just represent one more property.

Cognitive significance can be brought into sharper focus by considering its relationship to other kinds of significance. A trait may be *morally* significant to the extent that a creature's possessing that trait implies we morally ought to treat that creature differently. A trait may also be *personally* significant, to the extent that the individual who possesses it values doing so. Or it

---

difference between humans and others, than at motivating pursuing the question seriously. Aristotle may be an exception to this: he did carefully study animal behaviour. But it is debatable how far this study was motivated solely by his concern for understanding humans' distinctive function for ethical purposes, and how far by questions more like the cognitive significance question. And in any case, the idea that humans have a distinctive function which is ethically relevant looks less attractive to us than it did to him.

might be *evolutionarily* significant, to the extent that it increases fitness (this is a different notion of evolutionary significance to being a 'major transition'). I will discuss the relation of each of these to cognitive significance, starting with evolutionary significance.

A cognitive trait can enhance fitness a great deal without being cognitively significant. Adding a system for detecting nearby snakes might radically improve fitness. But it would have limited effects on how a mind operates and the sorts of things it can do. Conversely, a trait can be cognitively significant, transforming the capabilities of a mind, even in creatures who exercise these capabilities in a fitness-neutral, or even evolutionarily maladaptive, fashion. To take an extreme example, if we add certain philosophical abilities to some creature, it may gain profoundly new ways of thinking about mereology, but it might take so enthusiastically to metaphysics that it entirely neglects breeding and collecting food. So cognitive significance is distinct from evolutionary function.

However, cognitive significance can *ground* evolutionary significance, given the right background conditions. If a trait radically changes the capabilities of a mind, this will affect the creature's ability to achieve its goals more reliably and successfully. Where fulfilling these goals promotes successful reproduction or survival, this can contribute to evolutionary success. Evolutionary functions of cognitive systems will therefore often advert to the reasons for their cognitive significance—to the ways in which they transform the mind.

Similar points can be made about personal and moral significance.

There are personally important traits which are cognitively insignificant. Some people strongly value their keen sense of artistic beauty, but do not use this sense to fundamentally change the way they think about other things. Conversely, any case where a subject has some trait which makes a big difference to how their mind works, but which they are not sufficiently

well-versed in psychology to even think about, let alone value (like using a particular learning algorithm) would trivially count as cognitively but not personally significant. However, given the right background conditions, cognitive significance can ground personal significance: if someone appreciates how important their ability to reason is to how their mind operates and to their achieving their goals, this will provide a strong justification for valuing their ability to reason.

There are examples of traits whose moral significance is much greater than their cognitive significance. The possession of moral concepts arguably falls into this category (note that the claim is not that this is cognitively insignificant—just that it is less cognitively significant than it is morally significant). Purely intellectual capacities like arithmetical abilities, meanwhile, may be traits which are much more significant cognitively than morally. But given background conditions, cognitive significance can ground moral significance. We might have a moral theory on which cognitively significant traits are thereby morally significant: some people think that intelligence, understanding and rationality should be valued in themselves. But even if we do not, expanding the range of things a creature's mind can do will typically expand its morally relevant interests. While this dissertation will not for the most part address moral issues directly, we will discuss one case of moral significance in §4.6: the claim that grasping a narrative of one's own life is required for one's life as a whole (as opposed to one's momentary pleasures, pains, desires and aversions) to matter morally.

One final point that is important to unpacking the definition of cognitive significance above. Cognitive significance is *relative to a kind of mind*. This is because a mind's existing capacities shape both the kinds of use it can make of an additional trait, and whether those uses are genuinely new. A trait could be significant but only given a certain background. For

example, adding a new way of estimating numerosities may allow a creature to develop further mathematical capabilities, such as accurate addition and multiplication. But this may depend on also having the ability to simultaneously hold multiple sets and their numerosities in working memory without confusing them. A creature who could only hold one numerosity in mind at once would not be able to represent both of the numerosities to be added together, the answer, and the numerosity it needs to be compared to. And a trait could be rendered insignificant because some other trait does the thing it was supposed to be special for doing. For example, there may be two ways of estimating the number of items in a set. In such a case, each would be significant for a mind which did not also have the alternative available, but would not be significant for a mind which had both.

There are several advantages to reframing the question about the difference between humans and animals as a question about which traits are cognitively significant. This question carries none of the unhelpful connotations and presuppositions of our original question. It allows for (though does not entail) the view that the evolution of the mind involves many important innovations, not a single important leap. It does not collapse the diversity of animal minds into one category, or single out humans as the only creature with cognitively significant traits. Neither does it imply an objectionable form of hierarchy. It does not assume that cognitively significant traits will be more adaptive, let alone that evolution produced a number of animals who can be straightforwardly ranked. But it does capture a great deal of the interest of the human uniqueness question, as cognitive significance is both important, and fruitful to study.

The fact that cognitive significance can often ground other kinds of significance provides one reason why it is important. If we are interested in understanding the evolutionary, moral,

and personal significance of episodic memory, understanding its cognitive significance will probably be a large part of this project.

We can go further: it would be beneficial to replace the current focus on potential evolutionarily adaptive benefits in the literatures on many cognitive traits, such as episodic memory, with a focus on their cognitive significance. One reason for this develops the last point: cognitive significance is plausibly explanatorily prior not only to evolutionary significance but also to moral and personal significance. By contrast, evolutionary significance is unlikely to play a role in explaining moral and personal significance. It is not plausible that the ethics of how we treat a creature should depend on the survival value of its various cognitive traits in anything like the way that it does plausibly depend on the range of cognitive abilities the creature has.

Another reason to move away from the focus on evolutionary significance and towards a focus on cognitive significance is that we can often gather more appropriate evidence to decide the question of cognitive significance. As we will see in this dissertation (especially Chapters 4 and 5), there are often promising research questions opened up for testing proposals about cognitive significance, especially using computational simulations of possible minds in combination with the study of actual brains and behaviour. By contrast, the literature on the evolutionary function of traits like episodic memory is largely speculative or thinly supported, and (unlike those claims about cognitive significance in this dissertation which are thinly supported) not amenable to gaining thicker support through future research. [13] For example, the main arguments for postulated evolutionary functions for episodic memory revolve around roles

[13] Suddendorf & Corballis 1997, 2007; Klein et al. 2002; Buckner & Carroll 2007; Schacter et al 2007, 2011; Boyer 2008, 2009; Rasmussen & Berntsen 2009; Allen & Fortin 2013; Templer & Hampton 2013; De Brigard 2014; Michaelian 2016; Mahr & Csibra 2018; Mar & Spreng 2018; and Rau & Botterill 2018.

it seems to play in contemporary humans. But we should be very cautious about assuming that the role played by a trait now tracks its historical roles, especially without fossil evidence to indicate the context in which episodic memory actually arose, and should be wary even of assuming of any trait (especially a complex one like episodic memory) that it was selected for and has (or had) adaptive benefits at all.[14]

Another reason to focus on the question of cognitive significance is that, as I will be arguing in Chapter 5, it can be used to individuate hard-to-define traits. It can be controversial what it takes for an animal to count as having a trait like episodic memory, language, or reasoning—what its core features are. I show in Chapter 5 that cognitive significance provides a non-arbitrary, illuminating way of answering questions of this sort (roughly, looking for combinations of features which are intuitively important to the trait but also combine to account for the cognitive significance of the trait).

1.2 Time and Cognitive Significance

We have our question, then: what are the most cognitively significant traits, and which species have them? What is the answer?

As stated above, we should not expect a single answer. There may be a number of traits which are highly cognitively significant. Complicating the story further, there may be traits whose significance is explained in terms of even more significant traits, and traits which are highly significant but only in the context of the presence (or absence) of certain others. But with these qualifications in mind, most of the even vaguely plausible answers to the original question

[14] Gould & Lewontin 1979, Lloyd 2005. See Chapter 3 for more issues with the epistemology of the evolution of the mind.

about human uniqueness will be plausible answers to the cognitive significance question (indeed, traits which were proposed to be unique to humans, but then were found to exist in some form in some animals will remain good candidates for cognitive significance, which does not rest on any claims about being *unique* to any given species). Language, reasoning, and theory of mind are all plausible candidates, for example.

This dissertation focuses on the cognitive significance version of a different cluster of suggested answers, made in different forms in a wide range of authors, to the human uniqueness question. Many have suggested that a major difference between humans and animals has something to do with humans' having a richer relationship to time. Animals are (it is often asserted without much argument or evidence) in some sense stuck in time, unable to reflect on other times; and this makes a big difference to their minds generally.

Avicenna thought that animals are incapable of recalling memories at will,[15] and cannot anticipate the future. They fear only events that directly relate to the present, and act in preparation only by instinctually experiencing a future event as if it were happening now.[16] Similar ideas are echoed in Burns' address *To a Mouse*, centuries later and in a rather different cultural context:

> "The present only touches thee:
> But och! I backward cast my e'e
> On prospects drear!
> An' forward, tho' I canna see,
> I guess an' fear!"[17]

[15] See also Aristotle, *On Memory*.

[16] Alwishah (2016: 79).

[17] Burns (1786: 140).

John Wesley, meanwhile, emphasizes that 'thoughtless brutes' do not think about time (specifically 'That time shall shortly end'), thanks to being distracted by "the wild whirl of time's pursuits".[18] Kant emphasizes the *cross-temporal* unity of a self-conscious, in passages still taken seriously by Kantian ethicists today, who respond to Kant with points like:

> "People like to say that animals live in the moment, and in one sense that is probably right: unlike human beings, they do not seem to spend a lot of time planning for the future or fretting about problems that may or may not arise. But in another sense, I do not think it is true. Or perhaps what I should say is that at least for many animals, the moment itself does not live merely in the moment, but reverberates with the character of the other moments in the animal's life."[19]

Nietzsche makes the following claims about cows (although he immediately slips into talking about 'the animals' generally):

> "Consider the cattle, grazing as they pass you by: they do not know what is meant by yesterday or today, they leap about, eat, rest, digest, leap about again, and so from morn till night and from day to day, fettered to the moment and its pleasure or displeasure, and thus neither melancholy nor bored."[20]

Wittgenstein seems to tie at least some forms of explicitly temporal content to possessing language:

> "§649. "So if someone has not learned a language, is he unable to have certain memories?" Of course — he cannot have linguistic memories, linguistic wishes or fears, and so on. And memories and suchlike in language are not mere threadbare representations of the *real* experiences; for is what is linguistic not an experience?
>
> 650. We say a dog is afraid his master will beat him; but not: he is afraid his master will beat him tomorrow. Why not?"[21]

[18] Wesley (1868: 3371).

[19] Korsgaard (2018: 33). Korsgaard Ch. 2 catalogues the relevant passages in Kant.

[20] Nietzsche (1997: 60f.).

[21] Wittgenstein (2009: 174). He repeats similar ideas in Wittgenstein (2009: 183) and Wittgenstein (1989: 282f.).

These are not the only thinkers to make this sort of claim: others include Köhler 1925: 276; Bergson (1991: 82ff.); Velleman (1991: 68f.); and still more enumerated in Roberts 2002 and Hoerl 2008.

These are diverse thinkers, making these claims for diverse reasons. They do not all offer these issues about the relationship animals have to time as one of the *most* important differences between humans and other animals. Nonetheless, it is striking that they all seem to hit on ideas which are at least resonant with one another, and thought this alleged limitation in thinking about, caring about, or understanding anything but the present to be worth remarking on.

What does the contemporary science of animal minds have to say about all this? To many scientists, it will look bizarre to say that animals are stuck in time. There is a great deal of evidence of sophisticated behavior, that is often taken to show that many animals represent temporal properties including time of day, durations, and rates, and that many animals can predict and plan for the future and remember events and facts from the past.[22] However, there are two overlapping strands of the contemporary literature where the idea that animals are stuck in time has legs. Firstly, there are puzzles about what it takes to genuinely represent other times, and suggestions that animals do not do that.[23] Secondly, there are many who claim that episodic memory and 'mental time travel' into the future are either unique to humans, or untestable in non-linguistic species.[24] Furthermore, even for those who do not take the claim that these capacities are unique to humans seriously, the claim that our relationship to time is cognitively significant should be a very live option. Instead of asking whether any animals represent time,

[22] See e.g. Gallistel 1990.

[23] Roberts 2002; Hoerl 2008; Hoerl & McCormack 2019.

[24] See especially Tulving 1983, 2005; Suddendorf and Corballis 1997, 2007.

our question should be: how cognitively significant are episodic memory and the representation of time, and how widespread are they? And Hoerl and McCormack and Tulving should be seen as offering a particularly extreme pair of answers to these questions ('very significant' and 'unique to humans'). But it might turn out that temporal representation and episodic memory are extremely significant and possessed by a larger range of species.

I will be arguing that both episodic memory and certain kinds of temporal representation are highly cognitively significant. And I will be arguing that we do not have strong evidence about which animals have these traits, but that it is a live option that several species do, including scrub jays, monkeys, and rodents, and I will be suggesting ways of gaining stronger evidence on these issues. I will also be emphasize the variety of forms that temporal representation temporal competence generally can take in different species. One consequence of this is that some forms of temporal representation are likely to be much more widespread, but also less cognitively significant, than others. Another is that, *pace* many in the literature, who assume that episodic memory requires temporal representation, neither episodic memory nor temporal representation of any kind require one another. They are constitutively independent of one another, and significant for different reasons.

1.3 Overview of the Dissertation

One repeated theme in this dissertation will be the need to ask the right questions about animal minds, and to understand the distinctions between different possible questions. We have already seen reasons to replace a question about human uniqueness with a question about cognitive significance. But we should distinguish a further set of questions, which one can ask about any mental trait X:

(**Cognitive Significance Question**) How significant is X, and why? What important kinds of capacity can this trait underpin (in combination with other features possessed by a relevant kind of mind)?

(**Constitutive Question**) What is it to possess X? What are the core features one has to have to count as unequivocally possessing X?

(**Epistemic Question**) What would be evidence that a species has X, and how strong would different potential forms of evidence be?

(**Distribution Question**) Which species have X?

All of these questions are related to one another, and our answers to any of them should be responsive to thinking about the others. But they should not be confused with one another. And it is common to confuse at least some of these questions. For example, it is common to confuse epistemic issues of whether we can get enough evidence that some species has X with constitutive issues about whether it would be possible for that species to possess X, and with distribution issues about whether that species in fact has X. We will see authors confusing questions in this kind of way at several key points in this dissertation.

I will be arguing that the constitutive question often needs to be answered in tandem with the cognitive significance question. Which features are core to a trait depends on which features contribute to that trait's distinctive cognitive significance, but cognitive significance will depend on what the features of the trait are. The answer to the constitutive question should also heavily inform one's views on the epistemic question, which will in turn determine how one interacts with the evidence to answer the distribution question.

This dissertation will (more or less) follow this order. For several traits, I will be starting with questions about cognitive significance and the nature of the trait in question, then drawing

on this account to answer the epistemic and distribution questions. I will be discussing these questions with respect to temporal representation; various sophisticated kinds of representation (in particular, narrative representation, representation of temporal frameworks, and representation of essentially dynamical entities); and episodic memory.

I will begin with the nature and significance of temporal representation generally, in Chapter 2. Here, much of the discussion will focus on the nature of temporal representation, which is not at all clear: As we will see, there is a puzzle in that many of the capacities we might think would be distinctive of temporal representation could also be possessed by systems which we would not want to call representational. I will show that temporal representation is distinctive in that it allows for a certain kind of flexibility in decoupling from very specific temporal contexts. But a consequence of this discussion will be that temporal representation itself is not all that significant. Chapter 3 draws on this account of the nature of temporal representation to discuss evidence for temporal representation. But its focus will largely be on introducing various ideas about the interpretation of evidence concerning animal minds, which will be drawn on throughout the rest of the dissertation: for example, it develops an account of the conditions which determine the strength of evidence provided by different sorts of evolutionary considerations when trying to determine which animals have which mental traits.

Chapter 4 develops the ideas developed by Chapters 2-3 about temporal representation to discuss the potential significance, constitutive nature of, and evidence for specific sophisticated *kinds* of temporal representation: representation of temporal frameworks, narratives, and essentially dynamical entities. We will begin to see more clearly how considerations of cognitive significance can be used to answer constitutive questions, by shaping which distinctions between different temporal representations we should care about. I will show that

narrative representations and representations of essentially dynamical entities could both be cognitively significant, because both could be used as ways of coping with otherwise intractable complexity in the dynamics of the world around us. However, as Chapter 4 will introduce a number of new (or underappreciated) distinctions and ideas about these sorts of representations, there will be only limited empirical evidence to speak to which animals have these traits.

At this point, we will have in place both a general sense of what different kinds of temporal representation can buy a mind, and what a mind can do without them. This will lay the groundwork for understanding episodic memory, in Chapters 5-7. Chapter 5 argues that episodic memory should be defined by three core features: (i) imagistic representation, of (ii) particular past events, based on (iii) a memory trace. It should be so defined because the combination of these features could be extremely cognitively significant, leading to a form of learning that is in an important sense unlimited. Chapter 6 defends and elaborates this account in light of the discussion of temporal representation in Chapters 1-4: we will be in a position to argue that temporal representation is not core to episodic memory, although it may be typical of many specific kinds of episodic memory. I will also be arguing that episodic memory can come in many varieties in different species and within one species, including unconscious varieties. Finally, Chapter 7 will answer the epistemic and distribution questions for episodic memory in light of the discussion of evidence in Chapter 3, Chapters 5-6's account of the nature of episodic memory, and the most prominent empirical paradigms. Again, given the novelty of my constitutive account of episodic memory, many of the most direct ways of testing for episodic memory have not been carried out yet. However, there is suggestive evidence that several animals do have episodic memory.

There will be several important themes to this dissertation. Some have already been touched on: the importance of cognitive significance; the need to keep the different questions one can ask about animal minds distinct, whilst also recognizing how addressing some of these can help answer the others; and the huge, systematic variety that is possible in kinds of mind. But three further themes will emerge.

One is the importance of understanding all minds in terms of imperfect ways of coping with complexity. The world is an incredibly complex place, and even humans cannot hope to fully understand or predict everything that happens. Instead, we have ways of simplifying the world into something more manageable, and systems which can help us gradually capture more of that complexity with special kinds of learning. Different species are more or less limited in their ability to grasp (or at least cope with) various forms of complexity, and the most cognitively significant traits will often be those that improve those grasping abilities.

A related point is that learning will often be crucial to the most significant traits. The capacity to learn better models can create far more flexibility, and leads to more impressive further capacities, than the kinds of features philosophers often focus on, such as having a complex model of the world that includes the understanding that a particular domain or kind of entity has certain properties, and immediate reasoning and problem-solving abilities. One reason time is so significant is that although it is in a sense a specific domain with its own quirks, temporal sophistication can lead to much more effective general learning capacities.

Finally, we will repeatedly run up against the main issue of chapter two: it is easy to over-intellectualize what is required for various capacities. Often it is tempting to claim that some ability requires a certain sort of representation or computation, where in fact it turns out that with a little ingenuity, we can find ways in which a fixed architecture could perform all the

functions required of the representation or computation. This is crucial for understanding animals: it often somewhat undercuts apparent evidence for certain sorts of representation or computation. But it makes sophisticated-sounding capacities available even to animals with very limited representational and computational capacities. To understand this issue more fully, we will need to launch into the examples from Chapter 2.

Chapter 2: Temporal Competence and Temporal Representation

2.1 Introduction

Some dogs get excited a few minutes before their owners are due to get home, well before they are close enough to be smelled, heard or seen. How? One possibility is that these dogs keep track of how long it has been since their owners left, have a sense of how long their owners generally go out for, and thereby anticipate that their owners will get home soon. However, something rather different might be going on. When you leave your house, you leave behind a cloud of odour that will change in predictable ways over time, gradually interacting with other molecules in the air and slowly dissipating. These changes — and quite possibly the initial odour itself — are typically too subtle for humans to notice. But dogs have a much keener sense of smell. Thus, it is possible that a dog could associate the composition and intensity that is always reached, say, seven hours after their owner left, with their owner's coming home. This association could be enough to explain the anticipatory excitement, even if the dog has no sense of time at all.[25]

This is an example of a more general phenomenon. Animals often behave in ways which are appropriate to the dynamics of their environment. This temporal competence *might* be explained by positing that the animals represent some temporal feature like duration, time of day, the temporal order of a sequence of events etc. But alternative explanations of the competence, which do not posit such representations, are also live options.

Consider another example. Nearly all organisms — including plants and bacteria — go through regularly repeating cycles of activity; indeed, individual cells within multicellular

[25] This explanation is suggested by Horowitz 2016: 22f. and seems to be borne out by an informal test on one dog for the BBC's *Inside the Animal Mind* Ep. 1 'You Are What You Sense', viewable at https://www.youtube.com/watch?v=Ftr9yY-YuYU, where surreptitiously introducing a fresh sample of the owner's smell in the middle afternoon meant that the dog no longer anticipated their owner's return.

organisms go through circadian (24 hour) cycles.[26] This is often adaptive: it might be beneficial if the movement of your leaves and flowers throughout the day maximizes exposure to the sun, or if you hunt at night when you cannot be easily seen by your prey. But again, very different mechanisms could produce temporally appropriate behaviour. One could keep track of the time of day and use this to intentionally perform activities at their optimal time — using an alarm clock to leave home at 08:15 to ensure one gets to work by 09:00 would be a clear case of this, as would a less specific, less technologically dependent case like going for a walk in late evening in the hope of seeing badgers. But many regular cycles in nature are not cognitively mediated in any way. The rotation of the Earth is kept regular by the laws of motion, not through the careful consideration of a successfully-appeased Sun-God. And many cycles in single cells and in our bodies are regular because they consist of sequences of chemical reactions which each last a reliable amount of time and then directly cause the next reaction in the chain, or because they are governed by external circadian changes in light and heat, not because there is cognition of time involved. This is so even if evolution has selected the features of the mechanisms involved partly so that it will synchronize with the environment, producing different behaviour for different times of day as appropriate.

What is the difference between a system with temporal representation and a system which is temporally competent — which reliably produces temporally appropriate behaviour — without temporal representation? What distinctive capacities can temporal representation underpin, or what can we explain by positing temporal representations?

Getting clear on this issue is crucial to understanding the mind generally. Dynamics are fundamental to how the mind works. All cognition, perception, action, and emotion involves

[26] Gallistel (1990: 221f.); Robins & Craver 2009; Montemayor (2013: 35).

processes which unfold over time, and these need to be coordinated with each other and with the world for any kind of successful behaviour. This makes it tempting to many to claim that *all* representations must have temporal content to ensure this sort of coordination,[27] or that all action rationalizations must involve at least one state with temporal content.[28] Others have claimed that the whole computational theory of mind is a gross oversimplification precisely because (on their view) it cannot capture the complex dynamics involved in cognition.[29] If we can make sense of how coordination and temporally appropriate behaviour can often be achieved in minds understood as computational, representational systems without temporal representation, this tells us something important about how those minds work.

Clarity about the representation of time will also shed light on the nature of representation itself. Giving an account of the nature of representation in general is a fundamental issue in the philosophy of mind, and I will not attempt to fully solve it here.[30] However, the issues in this chapter cause problems for many of the standard accounts of representation. Standard treatments often start by ruling out simple accounts of representation, like representation amounting to reliable covariation with the representata. One of the standard counterexamples to that view—where we have reliable covariation without representation—is a temporal case, the covariation of number of tree rings with the tree's age. But it is not usually noticed that this is an instance of a more general phenomenon: representation-like phenomena

[27] I will address versions of this view in §2.5 and §4.7 below.

[28] Morgan 2019.

[29] Perspectives on this view can be found in Van Gelder 1995; Clark 1998; McClelland et al 2010; Chirimuuta (2020).

[30] See Dretske 1981; Millikan 1984; Fodor 1987; Burge 2010; Neander 2017; and Shea 2018 for important treatments of the general issue.

often arise in the temporal domain, because so much of being a successful organism is about synchronizing to the dynamics of the environment, to internal processes, to fluctuations of resources and to conspecifics' behaviour. Compared to tree-rings, there are far more complex, less accidental-looking and representation-like, yet still non-representational, cases of synchronization; and they make trouble for several leading theories of representation. This is at least in part because those theories tend to be built around accounting for other cases, especially perception or language, where our interactions with representata work very differently to our interaction with temporal properties.[31]

This paper has a negative part and a positive part. On the negative side, I will show that many kinds of temporal competence, and even a robust form of diachronic rationality, are possible without temporal representation. This raises a puzzle: if so much can be achieved without temporal representation, what does temporal representation actually add to the mind? The positive side to the paper makes progress in answering this question, proceeding in two stages. First, I will open up space for the positing of temporal representations adding explanatory value, even where it appears redundant due to the presence of other explanations. Second, I will develop a positive account of distinctive roles for temporal representation, by separating out two kinds of alternatives to temporal representation and dealing with each in turn. First, temporal competences can be explained in terms of systems with states whose functional roles include having certain kinds of dynamics, even if those states do not have temporal contents. To these kinds of system, temporal representation adds certain kinds of flexibility: states need not be tied to such a specific functional role and so can be used in a wider range of

[31] Neander 2017 explicitly only tries to give an account of representation in perception, with the hope that this can be generalized or built on to ultimately have an account of representation in general. But if this is the approach, we should have multiple starting points, including temporal representation, in the hope of convergence.

operations on the basis of their more explicit relationship to time. Second, temporal competences can be explained in terms of states which represent some property which is not temporal, but which does have a reliable temporal profile, as in the case of dogs representing odours and thereby acting at certain times after their owners left. Adding temporal representation to such systems adds a different kind of flexibility: here the main difference is that temporal properties are more abstract than their rivals, so allow for combinations with a wider range of other properties and hence more systematic coordination.

The key to the proposal, then, will be flexibility: to count as representing time, a system needs to behave appropriately with respect to time even though what that appropriateness is is not fixed, but is determined in combination with many other contextual features. A representation needs to be able to take on different roles, in different combinations with other representations, varying in content with the represented feature but also varying in use according to the relevance of that represented feature: it needs to be usable in different domains where the feature appears, in contexts where the target feature has different relationships to other features, and so on. This view has some features in common with consumer-based views: it implies that there is a great deal of untapped potentially representational correlations and synchronizations between mind and world, which do not count as representational partly because they are never used in computations. But it does not require clear consumer and consumed systems: rather, the key to the flexibility which representations give is multiple systems interacting together to decouple from fixed relationships to the environment in the right kinds of way.

I will proceed as follows. §2 will get some more key examples of temporal competence without temporal representation on the table and begin to pull out some general lessons from these examples. §§3-4 will show how even a robust form of diachronic rationality is possible

without the representation of time. §5 considers the view that many or all of the cases discussed in earlier sections are in fact cases of temporal representation after all. I argue that we should reject these views, but draw a lesson from them in §6: we need to distinguish between the implementation of temporal representation and rivals to it. §§7-8 offer a positive account of when we should count these explanations as implementing temporal representation, by dividing this question into two: when we should count these explanations as involving representation at all, and when we should count them as involving temporal representation specifically.

2.2 Temporal Competence Without Temporal Representation

2.2.1 Coordination and Updating

Any mind will only be successful if its different processes coordinate appropriately with each other and with the environment. This might make us suspect that temporal representation is extremely widespread. We might think that successfully using perception, memory and anticipation requires each of these systems to at the very least mark out their contents as past, present, or future, and that coordinating systems which operate with different lags — like the lag between seeing a match being lit and smelling it — or with different time frames — like long term and short term memory — requires more temporal detail to be built into the contents of those systems.

Not so. Coordination can be achieved in other ways. This point has been developed at length by Hoerl and McCormack, who suggest many explanations that do not positing temporal representation, for a wide range of temporally appropriate behaviour.[32]

[32] Hoerl 2008, Hoerl & McCormack 2011, 2018, 2019

Hoerl and McCormack's most basic case like this involves the coordination of a kind of memory with current perception and action, in what they call a 'Temporal Updating System'. This system takes information from *previous* experiences, and uses it appropriately to produce behaviour suitable for the *current* environment. One could do this by marking out memories of the past as past and perceptions of the present as present, and inferring using bridge principles like 'features of the world in the past will hold true of the world today unless I have evidence that they changed'. A temporal updating system uses no such mechanisms. It achieves co-ordination by not maintaining representations of past environments *as past* at all. It keeps a model of the world, which retains information from past encounters with the world beyond what is immediately experienced. But when it gets new information contradicting its current model about some point (e.g. while the model has it that there is a tree in a certain location, the system is told that there is not), it never concludes that the world must have changed (e.g. that there used to be a tree but it must have fallen down): it changes its model and discards all information about how it used to think the world was. It does not distinguish between its memories of the past being false and the world having changed from how it remembers it.

The basic move here is to build the temporal competence into the functional roles of the states involved, obviating the need for the states' *contents* to mark out the times they relate to. Another simple case of this kind would be the coordination of beliefs about the present and desires for the future. One might think that in order to make any practical decisions, a creature would need to use some sort of belief about how the world is *now* and a desire for the world to be different in a certain way *in the future*. However, we need to be careful not to overintellectualize the requirements for action here. A creature could have two kinds of states: both represent a state of affairs without building into their *content* when (or if) the state of affairs

occurs: they will both have tenseless representations like *apple on the table* rather than *apple is now on the table*. Coordination could be achieved by the different kinds of states being formed and used in systematically differing set ways. The belief-like state could be formed and modified only given certain kinds of perception or inference, while the desire-like state could be formed from preferences. The two could interact according to fixed patterns like: the combination of a belief-like state with content of the form *if action A, then P* and a desire-like state with the content *P* gives rise to action *A*.

Human beliefs and desires are no doubt more complicated than this, precisely because we do have temporal contents. We can have explicitly tensed beliefs including beliefs about the way the world will be but currently is not; and we can have desire-like wishes about how the world could have turned out in the past but did not. But the simpler system described above would work without building tense into contents.

Furthermore, even in the case of humans, we should be wary about trying to build too much of the functional role of different states into their contents. For one thing, it is not clear that we *can* capture every feature of a state's functional role in this way. Consider desire: we might think that we could capture its special role in terms of representing a state of affairs as *a good future state of affairs for me*. But it seems possible for certain kinds of depression to leave a person unmotivated with respect to P, and failing to actually desire that P, even if they wholeheartedly accept (at least in principle) that P would be a good future state of affairs for them. Perhaps a different content could be found which will guarantee that any state with that content will have the functional role of a desire, but there is no reason to expect this. Fundamentally, positing a representational content where we could just describe the functional role is explanatorily redundant.

The doctrine that all of the important features of a state's functional role will be captured in its content, as this gives rise to implausibly strong intellectual requirements for most states. For example, if we are going to require that all creatures with belief-like states have tensed contents, we might as well also require that their contents also include 'It is *actually true in my current context* that…'. This would be misleading about what a creature capable of the belief-like state can achieve. It forces us into a dilemma when it comes to most animals and many humans. Either we are deflationary about what it takes to represent truth, actuality etc., and blur the distinction between creatures which have simple belief-like states representing the location of food, shelter, predators and potential mates around them but not much more, and a creature who has a robust concept of truth, an understanding of modality etc. Or we are demanding about what it takes to have a belief-like state and claim that most animals and many humans do not have any belief-like states at all, leaving us unable to capture the sophistication of many of the minds in the world around us. [33]

One might still insist that there is a discrepancy between time and other properties here, and that tensed contents — especially 'now' — are required for all creatures where 'actually' is not. But it is unclear what the argument for this would be. Morgan 2019 argues for such a discrepancy, but his argument is effectively that coordination between memory and perception requires temporal content, and this would beg the question in the current context. There is some temptation to think there is something special about tense here because it is so difficult to state what the contents would be without using tense: 'an apple on the table' is not a complete sentence in English and there is no way to complete it without adding a verb with a tense, while 'There is an apple on the table' is a complete sentence, despite not obviously and explicitly

[33] Davidson 1975 effectively endorses the second horn of this dilemma. See Camp 2009 for criticism.

marking that it is about the actual world and relying on context to fill out which table is being referred to. However, we should not rely heavily on natural language as a guide to mental content, especially the content of states in non-linguistic animals.[34] Furthermore, the use of present tense in English does not always straightforwardly imply a commitment to a temporally bound state of affairs in any case: those who think that mathematical truths are eternal will still use expressions like 'two plus two is four' to express them.

Suppose we accept that coordination can be achieved without temporal representation, between states with functional roles that guarantee they will always be about the past, states with functional roles that guarantee they will always be about the present, or states whose functional roles guarantee they will always be about the future. More complex versions of this general move — offloading temporal coordination from representational content onto carefully designed but fixed functional roles — can be used to explain more complicated kinds of coordination. Take the case of sensory systems with different lags. If the lag between, say, vision and audition, is either fixed (e.g. vision always projects information about external events 200ms before audition) or depends in a straightforward way on a few variables (e.g. if the lag is a linear function of distance from source), then integrating them successfully need not involve representing the lag. Instead, coordination could be achieved through, for example, the downstream systems automatically delaying the processing of visual information by the appropriate duration before comparing it to auditory information.

In general, there are many ways of achieving coordination in complicated systems without temporal representation: mechanical engines and clockwork provide numerous examples of such coordination. A much-discussed example is the Watt Governor in steam engines, which

[34] Beck 2013.

keeps a flywheel's speed constant in the face of varying steam pressures. The Watt Governor does not achieve this by calculating speed via representations of numbers of rotations and time. Instead, arms are connected to the flywheel which open the throttle to a degree depending on how high the arms are thrown up, which depends in turn on how fast the flywheel is spinning, so that when the flywheel spins faster than the desired speed the amount of steam being used is automatically modified downwards and the wheel is automatically slowed.[35] Systems that traffic in representations rather than steam could also be set up such that their dynamics automatically play out in a certain way.

2.2.2 Duration Sensitivity

One might think that while internal coordination is possible without temporal representation, it is quite another thing to learn to produce behaviour of an appropriate duration or in response to a stimulus of an appropriate duration. And this is a capacity which is extremely widespread: many classic findings in conditioning rats and pigeons involved learning to respond to stimuli like tones and lights of particular durations, to produce stimuli for certain durations, and variations on this theme.[36] This capacity too, however, can be achieved without temporal representation.

Again, we can appeal to states with dynamics built into their functional roles. Suppose you want to maximize the amount of nectar you get from a flower, and that the flower refills at a fixed rate after being emptied, until it is full and stops increasing, so that it is optimal to come back to the flower a specific duration after last emptying it.[37] One way to do this is to represent

[35] Van Gelder 1994; Clark & Toribio 1994; Clark 1998.

[36] Gallistel (1990: 294, 301ff.)

[37] Gill 1988 discusses the relevance of such flowers for the hermit hummingbird.

the duration since your last visit and use this to estimate the level of the nectar now. But instead, you might have a state whose representational content is the level of nectar, and whose functional role includes its having the relevant dynamics: it automatically increases its representation of nectar levels at the right rate, without any computation at all. This does not seem any stronger than the dynamics built into states for the purposes of coordination.

This is not yet enough to explain *learning* to respond to durations. If different flowers replenish at different rates, one would have to learn to come back at the appropriate time. But this need not occur through representation of time either. Instead, the system just described could be modified so that instead of having completely fixed dynamics, the rate at which its represented nectar levels grows can be modified by feedback. Suppose that when you visit the flower and find it already completely full, the rate gets sped up slightly; and when you visit the flower and it is still far from full, the rate is slowed down slightly. This does not seem to require representation of duration any more than the Watt governor did in modifying how much steam drives the engine. But it will produce learning of the relevant kind: the system will gradually converge on the dynamics appropriate to that flower.

Building dynamics into states' functional role is not the only way to produce duration sensitivity without temporal representation. An alternative strategy is to represent some other property one will represent anyway, but which covaries with the duration. A recipe can instruct us to sauté onions 'until translucent' instead of 'for eight minutes', and someone who had no conception of minutes would be able to follow it. This is also what is going on in dogs, if Horowitz's suggestion discussed above was correct: rather than representing duration, the dogs represent smell qualities which will bottom out in concentrations of different chemicals. The represented property need not be a perceivable one like translucence or odour: if one had the

system described above for estimating nectar levels even in their absence, for example, one could learn a different duration-appropriate behaviour, by using this system. One could, for example, learn to visit a location where friendly scientists periodically refill a feeder, not every 3 hours, but every half-filling-of-the-flower-with-nectar.

2.2.3 Rate Sensitivity

Producing appropriately timed behaviour and being sensitive to the timing of external events is all well and good, but what about sensitivity to variables which seem to require calculations with respect to time to appreciate? Surely sensitivity to rates of change requires either differentiation of a function of time, or dividing the difference between a start and end value of a variable over some interval by the duration of that interval? And surely this kind of thing requires computation over states that explicitly represent time?

Whether this is so is important, as sensitivity to rates is extremely important and widespread. As Burge (2010: 445) emphasizes, a huge part of perception is detecting change and motion. We can think of many advantages to being able to represent speed and other rates: being able to anticipate the precise moment at which an object in motion will arrive at a certain location, for example. Keeping track of rates, especially rates of reward or success of various kinds, will also be important for fundamental issues of cognitive control, like making speed/accuracy tradeoffs, and choosing how long to persist at a particular activity before moving on to another potentially more rewarding one. Indeed, Gallistel argues that *all* classical conditioning involves representation of time and rates of reinforcement. The basic idea here is that rather than just learning that two stimuli tend to occur together, one learns to increase one's expected rate of occurrence of one stimulus after observing the other stimulus, or one learns that

the other stimulus is likely within a certain window of time. This makes sense of various pieces of data that other frameworks for modelling classical conditioning fail to explain.[38]

However, like in the case of duration sensitivity, rate sensitivity can be achieved without temporal representation, either through functional roles that incorporate the right dynamics, or through representing other variables.

One way of designing functional roles to give sensitivity to rates would be as follows. Suppose that the variable whose rate of change we want to measure is temperature. Have a component that generates readings of temperature whenever it is told to, and hook this component up to a regular oscillator, say some homeostatic process in the brain that automatically happens every 1s just as a byproduct of normal functioning. So the system takes readings every 1s automatically, without representing time. Each reading generates a representation *m*. Suppose also that each period, just before it is replaced by a new reading, *m* is copied and a new state is formed with the most recent reading, *n*. We can compare *m* and *n*: if *m* - *n* is positive, temperature is increasing. Indeed, *(m* - *n*)/(period of the oscillator) gives an estimate of the average rate of change during the period. Because the oscillator has a fixed period of 1s, there is no need to use the period in the calculation: *m* - *n* gives an estimate of change in temperature per second.

If all we want is a one-off estimate of average rate of change, we do not need an oscillator: instead, we can just use a process that takes a regular duration to govern the taking of readings. This is in effect the mechanism used during Chemotaxis by *E. Coli*, where a system responds to whether the chemical composition of the surrounding water is getting better or worse given current behaviour, by either maintaining its current course or changing behaviour. It can

[38] Gallistel & King (2009: 226ff.).

respond to changes in encountered chemical composition because the chemicals encountered a few moments ago are still slowly interacting with certain components of the cell, so can be 'compared' to incoming chemicals. The interpretation of this case is controversial. Lyon 2015, van Dujin et al 2006 and others have argued that this constitutes 'cognition', with 'memory' being compared to 'perception' to make 'decisions' about 'action'. I do not take a stand here on whether these descriptions are accurate; but I do claim that if there is any cognition going on here, it does not involve computation using *temporal* representations: there would be no need for this, given that the length of time between a chemical entering the cell and reaching the 'memory' stage is effectively fixed by chemistry, and not tracked within the cell by a further system.

Instead of using either duration-sensitivity- or oscillator-based dynamics in the functional roles of relevant states, rate sensitivity can be achieved by representing alternative variables which correlate with either duration or with the rate itself. For example, to achieve sensitivity to how fast I have moved, I do not need to represent anything about distance and time, if I instead use efference copies of my motor commands or my expenditure of energy, provided these reliably correlate with speed. Or I could use a duration-dependent representation of nectar levels like that described in the previous section, and track changes in temperature per half-flower-of-nectar rather than per-hour.

Even more complicated computations can be built out of these kinds of components. For example, take Dead Reckoning, a means of navigation. This is normally described as follows. I leave my location and travel for some duration, and make a note of my average speed and direction of movement during that period. This vector has the information to calculate a direct route back to the nest. If I then carry on moving around, I can take note of the duration I have

travelled and the average speed and direction and add this vector to the original vector. I can keep doing this and I will end up with an estimate of my location, expressible as a vector giving direction, and distance I need to travel to get back to my starting point. But this sort of navigation does not strictly require representation of the duration of the periods involved: if readings are governed by a regular oscillator, so that they are taken at regular intervals, and if average speed is calculated using one of the methods described above in this section, one can avoid ever representing duration and still succeed in the navigation.

2.2.4 Building Blocks for Temporal Competence Without Temporal Representation

So far, we have seen two broad ways of doing without temporal representation: operations over states whose functional roles are partly delineated in terms of having specific dynamics; and operations over representations of properties which in fact have a reliable temporal profile even though nothing within the system recognises this. The former kind of states included states whose functional roles include their specific relationship to the past, present, or future. It includes states which are outputted by a process governed by a regular oscillator. And it includes states which are outputted by a process which lasts the same duration (or more generally, shows the same dynamic path) every time it occurs. The latter could include states representing nectar levels, light, one's own hunger and so on: any regular process in the environment.

Our task will be to find what genuine temporal representation actually adds to the mind over and above these ways of achieving temporally appropriate behaviour. This can be seen as special cases of more general problems in the theory of representation. The problem of what explicit representation adds to carefully constructed functional roles is closely related to issues of 'implicit' or 'tacit' vs. 'explicit' representations, which arise especially in the context of certain

kinds of neural network models but also in issues like language-learning and knowledge how. The problem of what representation of *time* adds to representation of other properties is closely related to the so-called 'disjunction problem' which arises for many theories of naturalized semantics, where theories relying on properties like causal covariation and natural selection seem to fail to assign determinate contents to states, as it is unclear what difference it makes to the organism's behaviour and success in the environment which of a bundle of correlated properties it represents.

The temporal case has special features here, however. It is particularly plausible that states' functional roles should include their dynamics, so explanations that appeal to functional roles instead of represented content are particularly plausible in the temporal case. For the version of the disjunction problem we have here, meanwhile, there are often reasons to favour alternatives to temporal content. For example, it is often thought to be more mysterious how *durations* could cause events in my brain than it is how *changes in nectar levels* could: the latter are less abstract.

2.3 Diachronic Rationality

Hoerl and McCormack develop many explanations of temporal competences without temporal representation along the lines of those above.[39] How do they face the challenge of saying what temporal representation adds to such systems? While they have made different suggestions in different places, Hoerl & McCormack 2019 emphasizes two properties of a

[39] One major difference is that they do so as part of an argument that only humans have genuine temporal representation; on my view, this is too fast, as we will see below.

system with temporal representation: it allows sensitivity to a temporal order which comes apart from the order in which items are presented to a subject; and it is necessary for diachronic rationality.[40] Surprisingly, however, both of these can be achieved with a reinforcement learning (RL) algorithm which does not traffic in temporal representations.

RL algorithms are an extremely important tool in AI and in models of animal behavior. Treating animals as performing reinforcement learning algorithms can explain many classic and novel features of animal behaviour in conditioning experiments, and details of activity in dopamine systems.[41] The general idea of RL algorithms is to learn a policy—a way of acting for each state of the environment the system finds itself in—which maximizes total rewards. They do this by incrementally updating estimates of the values (expected contribution to total rewards) of different actions or different states, on the basis of the outcomes they encounter.

The total rewards these policies learn to maximize, V(.), are the sums of immediate rewards, R, received after each of a sequence of actions, with rewards further in the future discounted by γ per period. So, when an action leads to immediate rewards, RL algorithms will learn that this action is valuable; but an action might also be valuable due to its longer-run effects, despite low initial returns, and RL algorithms will take this into account. An action might be valuable due to its longer-run effects because it puts the system in a position to take even more valuable actions down the road, leading to greater rewards overall. So these algorithms can learn to take actions with payoffs some way down the road in preference to actions that have a more immediate but ultimately smaller payoff.

---

[40] Hoerl & McCormack 2019. I am reading their claims about 'temporal reasoning' and 'thinking about time' here as relating to temporal representation. It is plausible that temporal reasoning is more demanding than temporal representation, but given that I will be arguing that the most important features of the capacities they claim are distinctive of temporal reasoning can be achieved without even temporal representation, making this distinction will not help their suggestion. We will discuss some of their other accounts of temporal representation in Chapter 4.

[41] Sutton & Barto 2018 chs. 14-15 and Ludvig et al 2011 summarize these results.

There are many subtly different RL algorithms, but we can consider pseudo-code for just one well-known algorithm to get a better sense for how these algorithms work — a Temporal Difference Learning (TD) algorithm: [42]

1. *Input: the policy π to be evaluated*
2. *Algorithm parameter: step size $\alpha \in (0,1]$*
3. *Initialize V(s), for all $s \in \mathcal{S}^+$, arbitrarily except that V(terminal) = 0*

4. *Loop for each episode:*
    5. *Initialize S*
    6. *Loop for each episode:*
        7. *A ← action given by π for S*
        8. *Take action A, observe R, S'*
        9. *$V(S) \leftarrow V(S) + \alpha [R + \gamma V(S') - V(S)]$*
        10. *$S \leftarrow S'$*
    11. *until S is terminal.*

Many of the details of this algorithm do not matter for our purposes. The important parts are as follows. π is the policy: it specifies an action for each state the system finds itself in. V(S) is the value of moving to a particular state S, on the assumption that the subject will subsequently take the actions dictated by π. V(S) is the sum of the immediate reward R from moving into S, and all the rewards the subject will get until the end of the episode, discounted by rate γ per time step. S' is the state system will move into after S, given that it follows π.

The point of this algorithm is that it arrives at an estimate for the value of V(.) for each state, through repeatedly taking actions from different states and updating its estimate for V(.) on the basis of the results. Although the true value which V(.) is trying to track is the discounted sum of rewards from a sequence of actions, the algorithm does *not* work by representing an entire sequence and simply summing over the discounted rewards at each stage. If it did, this would involve representing the temporal order of the stages of the sequence, in order to discount

[42] This specific TD algorithm is copied from Sutton & Barto (2018: 120), who call it 'Tabular TD(0) for estimating v_π'.

each stage appropriately, as later stages need to be more heavily discounted. Instead, at any given time, all that is represented are S and S`, V(.), R, and the policy π. Nonetheless, the algorithm estimates the discounted sum of a whole sequence of rewards.

It does this by implicitly relying on two facts in its key update of V(S) at line 9. First, the value of S just is the expected reward from S, plus the expected discounted future stream of rewards from the next period onwards. Second, the expected discounted future stream of rewards from the next period onwards is just V(S`), discounted for being one period into the future. V(S') already (initially inaccurately) incorporates information about the future stream of rewards, so this stream need not be represented explicitly to be responsive to that future stream. Line 9 therefore uses V(S') as a key part of an error signal (the expression inside the square brackets), which allows it to adjust its estimate of V(S) towards its true value each time it visits S. As more states are visited on more occasions and this operation is performed repeatedly, this algorithm will end up converging on the correct values for V(.) for all states it visits.

Given correct estimates of V(.) for every state, the optimal policy—the one which maximizes discounted total rewards—will be the one which always chooses to move into the state, of those available, which has the highest value of V(.).[43]

None of this requires representing the temporal structure of the sequences of actions and states involved. All that need be represented in each period are S, S', R, V(.), and π. S and S' are just states: it is important that the algorithm treats these differently, and does so in a way

[43] This is a simplification: this algorithm only estimates the values of different states *given policy π*. To turn this into an algorithm for learning what the policy should be, various complications need to be introduced which are irrelevant for our purposes (the basic idea will be to keep iterating a process of estimating the values of states given a policy, then changing the current policy to one which moves into states that would maximize the estimated values-given-the-old-policy, then estimating the values of states according to this new policy, and so on). Yet more complications are introduced by a probabilistic environment where an action's results are not determined. But none of these complications introduce an advantage in representing time, so I leave them out here for ease of exposition.

which reflects the fact that S' comes after S. But it can do this by having slots for each of them with fixed functional roles, much as in many of the other cases of temporal coordination above. V(.) and π, meanwhile, are not explicitly temporal: V is just a function from states to real numbers; and π is just a function from states to actions (or from states to states). If one buys that the systems for calculating rate etc. discussed above did not require temporal representation, one should equally buy the parallel claim about TD-learning.

So, given enough attempts at the task, a system can use this algorithm to converge on a policy which always take the actions which maximize its total discounted stream of rewards over time, without representing that stream (or the temporal relations between its component events). Surprisingly, we seem to have here a form of diachronic rationality—at least, something extensionally equivalent to diachronic rationality in the actions it performs in the long run—without temporal representation.

2.4 Immediate Computation for Diachronic Rationality

The achievement of this kind of diachronic rationality without temporal representation is surprising and interesting, but there is a severe limitation here: to learn V(.) for every state, the system needs to have multiple attempts at the task. During these attempts, it will decidedly not act in a diachronically rational manner. It often acts in what would look to us like very irrational ways for the sake of exploring the value of different states which we would immediately know to avoid through common sense, and for the sake of allowing changes to the estimated values of different states to gradually converge to their correct values. In a large state-space, especially, this will mean that the system only approaches the production of diachronically rational sequences of actions in the long run.

However, we can supplement the algorithm, without including explicit representations of time, so that it can overcome these limitations. If a system makes use of model-based simulation in the right way, it can determine its optimal policy, given its model of the world, before it has taken any actions at all.

A model is a subsystem which takes as inputs a represented state S and action A, and outputs representations of another state S` and reward R.[44] Introducing a model does not introduce temporal representation. The model does not represent long sequences: like the policy-representation, it need only represent two states (and an action and reward) at once, and it can do so with slots that have distinctive functional roles rather than representing the relationship between the two states as temporal. These functional roles will correspond to the fact that we want the model to output states and rewards according to those that will temporally follow a given action and state, but this correspondence need only involve the algorithms that update and use the model implicitly respecting these constraints. For any simulations it produces to be useful in guiding behaviour, the model will need to correspond to the actual transitions which will happen in reality, but this can be brought about through learning that itself uses only representations of pairs states (or quadruples of two states, a reward and an action) in functionally defined slots.

While the model only represents pairs of states, it can be used to generate longer sequences of states by stringing these pairs together. A creature can use the model to generate a prediction about what will happen if it takes action A in S, then generate another prediction about taking action A` in the state it is predicted to be in, S`, and so on. And it can update its

[44] In the stochastic case, the model will produce represented states and rewards according to a probability distribution p(S`,R|S,A).

value representations each simulated period much like the algorithm for dealing with the real world above. If the model can generate states and rewards for any action and any starting state, and it can do this sufficiently frequently, the system will be able to learn the values of different states under any policy without taking any actions in the real world. If the model is accurate, these value representations will be accurate. So if it has time to all this computation, the system will be able to converge on the diachronically rational policy before it has taken any actions, and immediately act in the diachronically rational manner.[45]

One might at this point double down and claim that genuine diachronic rationality requires temporal representation. One could appeal to internalist intuitions that rational choices need to be produced by the right kind of reasoning, and that merely extensionally conforming to a certain calculus is not good enough. But it is not clear that this line answers our puzzle of what temporal representation adds to alternative explanations of temporal competences. Instead, it leaves us in a situation where the results of even rational operations over temporal representations can all be systematically mimicked by systems without temporal representation. Perhaps one of these systems should be counted as rationally superior to the other. But our question is whether there is any non-normative fact about such systems that such representations actually explain.

We have seen a number of ways in which systems which do not seem to use temporal representation can nonetheless achieve temporal competence. This leaves us without an obvious way of finding any distinctive capacities that temporal representation can underpin. We have

[45] This presentation neglects many of the options for using model-based simulation in reinforcement learning; various tricks can be used to cut down on the number of simulations that need to be run, for example. The details quickly get complicated, and are being rapidly developed by Artificial Intelligence Researchers—some of these ideas are among the key ingredients of algorithms that have recently achieved news-worthy feats, like AlphaZero.

seen that the main issues here relate to two kinds of alternative to temporal representation: states with certain dynamics built into their functional roles, and states which represent other properties which correlate with the temporal properties of interest. Reinforcement Learning combined these two kinds of alternative. It relied on rather specific functional roles, especially for generating and using simulations of sequences of states. And in representing V(.), it effectively collapses information about a temporal stream of rewards into one atemporal representation.

2.5 Deflationary Accounts

One response to all this would be to claim that in some (if not all) of the cases discussed so far, we were tacitly talking about temporal representation after all. There are quite different flavours of this response, which I will classify jointly as 'deflationary accounts', as they all suffer from the same basic problem.

One kind of deflationist simply appeals to differing intuitions about the cases. Perhaps we should be inclined to call some of the above 'representation of durations', 'representation of temporally ordered sequences' etc., and that is enough. Such views will face the challenge of meeting three desiderata simultaneously: first, saying how positing representations in such cases adds any explanatory value to simply giving the apparently non-representational alternative explanation; second, making sense of the apparent discrepancy between temporal representation and the other representations cited in some of these cases, which seemed to be doing real work (e.g. the representations of states and values in RL, or the representation of nectar levels); and third, avoiding answering these issues in a way that over-generates representation elsewhere, for example implying that desires always include the content *It would actually be good for me in the future that…* or that we perceive not just *tables* and *chairs* but *collections of atoms*. Meeting

these challenges will require moving beyond intuitions and giving some sort of account of what representation adds in the favoured cases.

One could appeal to a general theory of semantics at this point, such as a version of teleosemantics.[46] This account might say that where we have a state whose dynamics are shaped by a process of selection so that they mirror the dynamics of a process in the environment, then this state's dynamics represent the external dynamics. The extra explanation from positing representations here, the account would go, would relate to an explanation of why the state robustly has dynamics matching the environment: and this explanation would appeal to the process of selection. And as teleosemantic accounts of representation tend to face disjunction problems in other areas too, different teleosemantic theories have solutions that one might hope can be wheeled out in this case.

One might respond here with more general, well-rehearsed objections to teleosemantics. But the approach also faces a problem specific to this case. It is unclear that such theories can really point to deep similarities between temporal representations so understood and other representations. And it is unclear that they can point to a deep difference between cases of 'temporal representation' and nonrepresentational systems whose dynamics have been shaped by selection. Pretty much every component of a living system is at least partly selected on the basis of performing its function at the right times, at the right speed, and so on. Teleosemantics typically requires more than just *selection* for representation, of course: advocates of teleosemantics will often build in some sort of use requirement. Suppose they require selection for carrying information in a computationally accessible format. But temporal information is *not* carried in a computationally accessible format in our cases: instead, it is baked

[46] E.g. Montemayor 2013 argues that temporal representation is widespread on the basis of an undemanding version of teleosemantics.

into the functional role, unlike information about e.g. which state one is in and the values of different states one could move into.

Burge 2010 gives a rather different deflationary account of temporal representation. Burge is adamantly opposed to deflationary (and teleosemantic) accounts of the representation of properties like colour and shape. For Burge, there is a robust difference between mere sensitivity to such properties, as found very early in the visual system, and genuine representation of such properties, as found in later perception. According to Burge, representation plays a very distinctive explanatory role when it comes to constancies and the solution of underdetermination problems. For example, he thinks we can only explain the visual system's distinguishing between the colour of a surface and the effects of the colour of the light, and the consequent ability to respond to the surface the same way even when it is under different colours of light, in terms of representation. However, Burge thinks that we do not have constancy-like phenomena in the case of time. Instead, he argues that we count as representing time whenever a 'sensitivity' to time (i.e. a system with dynamics that mean it produces temporally appropriate outputs) is 'harnessed' to a system that does use constancies to help it run properly. Burge's main examples of this 'harnessing' involve coordination of memory-like and perception-like states. A strong indication of just how deflationary Burge is willing to be about temporal representation is that he argues that *all* perception requires enough temporal coordination to imply temporal representation.

There is a great deal to say about the details of Burge's discussion of time, much of which has been said by Gross (2016). The basic problem for Burge here is quite simple, however: he does not point to any explanatory payoff to describing these harnessed sensitivities as 'representation'. He does not offer considerations suggesting that the visual system forms

states corresponding to temporal variables and computes over them in a way that presupposes they have veridicality conditions, like he does for visual perception of colour. Rather, he emphasizes that the very same systems and states which are used for coordinating non-representational activity are used in the same way when coordinating representational activity (when they are 'harnessed'). If these considerations support the idea that we have temporal representation in such cases, then they also suggest that other non-representational processes harnessed by the visual system, such as the flow of potassium ions involved in action potentials, should count as representational.

2.6 Non-Representational and Anti-Representational Accounts

While it is a mistake to call *all* of the above examples representational, and also a mistake to appeal to intuitions to support ascribing temporal representations without articulating any explanatory value to doing so, there is an important insight in the deflationary attitude. This is that it does not follow from our finding a mechanism that produces all the behaviour we are interested in, despite being describable without talking about temporal representations, that there are no temporal representations here. We could be describing a system that has computations over representations of time, but describing the *implementation* of that computation over representations.

One of the attractive features of the computational approach to the mind is that it shows how different levels of explanation fit together, because we can understand how higher level computational processes are implemented by lower level computational processes, how these are implemented by lower levels still, until at the lowest level we have operations implemented by physical systems. On this picture, we should *expect* that for any computation over explicit representations there is in the mind, we will be able to find a lower-level implementation of that

computation which does not involve explicit representations of the same things the higher level representation is implementing.

For example, a Turing machine which just represents strings of ones and zeros and has three states can implement a computation for deciding whether an input number is even or not, despite having no explicit symbol for division.[47] It can represent numbers by strings of '1's, and can move along any string of '1's , flipping between two of its states at each step. The state it is left in at the end of this sequence will correspond to even or odd, and it can output a '0' or '1' according to which state it is in at the end of its sequence (before moving into its third state, its halt state) to have a readable symbol of whether it determined its input number to be even or odd.

If another system *used* such a Turing machine to determine whether numbers are even or odd, feeding it numbers and reading the results, then we should think of that Turing machine as implementing the relevant function, even though it does not ever represent instructions like 'divide by two', and even though it just uses the '1' and '0' symbols, which have to be read by the right kind of system, in the right kind of context, to count as representations of 'odd' and 'even'. After all, *any* function a (digital) computer can implement will be implemented at some level by simple operations on '1's and '0's.

We should distinguish between explanations of behavior which are *non*-representational from explanations of behavior which are *anti*-representational. In the former case, we have an explanation, but it is not a rival to the representational explanation: it could be describing an implementation of that explanation. In the latter case, we have a genuine rival.

So why should we not think of our reinforcement learning algorithms as non- rather than anti-temporal-representational, implementing the computation of determining the diachronically

[47] This example is from Gallistel & King (2009: 111f.)

rational sequence of actions given a certain set of states, available actions and rewards? Why not think of the slots with distinctive functional roles as implementing representations of *state I am now in* and *state I might go to next*? Or the sequence of representations of states in the simulation as a representation of a sequence of states?

The answer is that in some cases, these *are* implementations of the temporal contents and computations. We should not infer from the fact that one of these algorithms is used to there being no representation of time here. However, this need not *always* the case. Sometimes we could have a system like this which is set up to operate in this way and produce the temporally appropriate behaviour, but there is no reading and writing of representations with temporal contents, no use of the system as representing, no high-level explanation invoking the representations which adds anything to the low-level explanation which does not.

The key question remains then, albeit in a reshaped form: what does it add to an explanation of a temporal capacity in terms that do not invoke temporal representation, to claim that we have described the implementation of a computation over temporal representations?

To answer this question, we need to again return to our division of the problem into two, depending on the relevant alternatives to temporal representation: representation as opposed to fixed functional roles which automatically have the correct dynamics, and *temporal* representation as opposed to representation of properties which happen to correlate with temporal properties.

2.7 Temporal *Representation*.

Let us take the case of the distinction between temporal representation and appropriate dynamics being built into functional roles. Fundamentally, a representation offers flexibility in how it is used, where a functional role is fixed. The same representation can be used in multiple

contexts. It can be combined with multiple different representations, it can be formed on the basis of different routines, and it can be used in different ways depending on the task at hand. Where a system is used flexibly in these ways because of its dynamics, there is reason to treat it as representational: calling it a temporal representation points to an explanation of why it is being used in the particular context, and its role in the mind.

Temporal representations can be combined with many different variables. For example, consider findings suggesting scrub jays represent durations since caching bits of food. There is an extensive literature on this capacity, which started with Clayton & Dickinson 1998, which aims to establish that scrub-jays' recovery of cached food is responsive to combining information about what kind of food was cached in a particular location (where), and how long ago this was (when), and we will be coming back to these findings at multiple points in this dissertation, so it is worth describing the experimental set-up in some detail. This study turns on scrub-jays revealing in their behaviour in advance of the experiment that they prefer eating fresh worms to peanuts, and peanuts to degraded worms.

Scrub-jays were first allowed to cache one kind of food (peanuts or worms), then, 120 hours later and in a different location, the other. 4 hours after that, they were given access to both locations, without any other food. Each subject belonged to one of three groups, defined by what they found when searching in locations where worms had been cached for 124 hours (which experimenters manipulated by interfering with the caches whilst the birds were out of sight):

> *Degrade*: non-degraded worms after 4 hours but degraded worms after 124 hours
> *Replenish*: fresh worms, even after 124 hours.
> *Pillage*: fresh worms after 4 hours, nothing after 124 hours

The subjects were given the opportunity to learn how quickly worms degraded (for their group) in pretraining trials. Of interest was which location they searched first and how often they searched in each location during the test phase.

Degrade jays choose the worm-location when they cache worms second (i.e. just 4 hours before the test phase), but the nut-location when they cache worms first (i.e. 124 hours earlier). *Pilfer* jays behave similarly (though with a weaker effect, presumably because finding a rotten worm is worse than finding nothing). *Replenish* choose worms in all conditions.

Subsequent work built on these results, finding surprisingly flexible use of this temporal sensitivity. There is evidence that scrub-jays use their sensitivity to (or representation of) time differently when its significance changes—even if its significance changes between caching and recovery. Significance can change in two ways: (i) rate of degradation might change, so that a given cache will/won't now degrade before time of recovery; and (ii) the individual's preference-ordering might change. Clayton et al 2003 show scrub jays respond to (i), by using a more complex set-up involving distinct opportunities to cache and recover from several of trays (the details are not important to what follows). Clayton & Dickinson (1999) and Correia et al 2007 (Experiment 1) provide evidence for (ii), by manipulating scrub jays' preference orderings mid-experiment through satiation (if a scrub jay is fed only, say, peanuts, for a few hours, then for a period it will like peanuts less compared to, say, pine seeds). After satiation, scrub jays search in the location of the food now preferred at recovery, not the food preferred at caching. It has also been shown that information concerning how long ago food-caching took place is integrated with 'who' information: whether and when jays try to recover or move cached food (and even where they move it to) depends on whether they are observed by other jays during caching that food, and on which scrub jay the observer was, especially their status (whether they are a mate of,

subordinate to or dominant over the subject).[48] This is because that scrub-jays are prone to pilfer one another's caches when given the opportunity.

All this suggests that scrub jays represent durations. This suggestiveness is largely because the duration information seems to be flexibly combined with many other kinds of information: type of cache, location, observers, rate of decay, and different preferences. McCormack (2002) does offer a deflationary explanation of the scrub jay results, in which rather than representing time since caching, they simply have representations of the cache being governed by a process with dynamics matching the decay process, so that when this process reaches its end the tasty-cache-representation is deleted or changed to a rotting-worm representation. However, the responsiveness to changing rates being moderated by kind of food, and the integration of information about possible pilferers mean that even if some such process is governing the cache representation, we should think of this as an implementation of rather than a rival to the temporal representation. For the particular length of time the process takes will vary according to context in predictable ways, as will the selection of which process to use; and pointing to its functioning as a representation points to answers to why this is.

Flexibility in combination with other variables is not the only kind of flexibility distinctive of representations. We also have reason to posit representations where they are formed via numerous patterns of input, integrating different streams of information under one heading because of shared environmental relevance. The case of constancies can be thought of in these general terms: for example, many combinations of retinal inputs giving rise to the same state because of their shared connection to a particular colour, under different lighting conditions. In the case of time, a system that integrates multiple sources about the timing of

[48] Emery & Clayton 2001, Clayton et al 2006, Dally et al 2006

some event or duration, because there will be computations drawing on the system that rely on its getting that timing right, will be implementing a temporal representation.

Representation also allows for flexibility in use. In some cases, even a TD-learning version of RL should count as representing *discounted present value*, implementing a calculation over a stream of rewards, because it is chosen for that purpose by another system. In more complex RL algorithms, there are often different parameters of the algorithm that can be manipulated, including parameters over e.g. how many time-steps ahead a simulation considers. In such cases, we can describe an implementation in terms of the components discussed above: states with fixed functional roles representing only one state ahead and so on. But describing the system in terms of temporal representation helps make sense of the system and its operation as a whole, by capturing and explaining patterns like how many of these states with fixed functional roles there are on particular occasions.

This point is clearer when considering data on monkey sequence learning in the Simultaneous Chain Paradigm. This requires subjects to learn to produce a *chain* (sequence) of actions in the correct temporal order, where the actions taken at each stage will be selecting a certain stimulus (pressing a certain image on a touch screen). The chain is 'simultaneous' because the stimuli are presented simultaneously—the subject will be presented with a number of images in different locations on one touch screen— and what the subject sees on the screen does not change after each action. Subjects are never presented with the stimuli in order during training. Instead they are always presented with all the stimuli at once, and on each trial, if they choose all of the stimuli in the correct order, they are rewarded, whilst as soon as they make a mistake, the trial ends immediately without a reward. Experimenters can choose what the items in the sequence are—they could be specific images, or image kinds. For example, the sequence

could be shape-based triangle-square-circle-pentagon, with triangles, squares, circles and pentagons being presented on each trial but in different colours and sizes. The spatial locations and other features of the different stimuli will be varied each trial to ensure that subjects are learning a sequence of stimuli of the right kind rather than a sequence of motor movements.[49]

This kind of sequence is exactly the kind of thing that can be learned easily by an RL algorithm (although the state space would have to be carefully defined so that the monkey counted itself as in a new state each time it selected an image). But a number of results from the SimChain paradigm and extensions of it suggest monkeys us a representation of temporal order in performing this task.

First, when monkeys are trained on numerous lists, with different kinds of stimuli and different numbers of items, they gradually get better—reach higher accuracy more quickly— at learning new lists.[50]

Next, there are positive reasons to think monkeys represent the lists they learn in an ordinally structured way. One such reason is that list-learning in monkeys in the SimChain paradigm shows *ordinal position effects*.[51] Suppose that a monkey learns 4 lists, A-D, each consisting of 4 previously unseen items, which we can label 1-4, so that list A consists of A1-A2-A3-A4, list B consists of B1-B2-B3-B4, and so on. If the monkey is then trained on new lists, made up of these same stimuli in new combinations, they will find some of these lists easier to learn than others. Specifically, they will find it easier to learn chains where each stimulus

[49] In Altschul et al 2017, monkeys learned to order stimuli according to categories like <*bird, cat, flower, human*> and even by painting styles <*Van Gogh, Dali, Gerome, Monet*>, with different exemplars of these categories on different trials.

[50] Terrace et al 2003.

[51] Chen 1997; Orlov et al 2000 and D'Amato & Colombo 1988 find related effects.

appears in the same ordinal position as it did in the list they were originally trained on. So, for example, they would find it easier to learn the list <B1, C2, A3, D4> than the list <C2, A3, D4, B1>.

If monkeys represent these lists in their entirety, as ordered structures, this ordinal position effect is relatively easy to explain: it is natural to think that the monkey recognizes a stimulus as having appeared *second* in a particular list in the past, and this either raises their prior that it will appear second in a new list, or leads to some kind of association with the position *second*, depending on exactly how we should think of the computations involved. McCormack (in conversation) suggested to me that a non-representational system could produce this effect if it learns to produce sequences by associating each stimulus with the phase of an endogenous oscillator.[52] If this were the case, learning a new list would be quicker if it requires associating stimuli to the same phases of the oscillator they were already associated with. However, there are a few problems with this account. One is that it will struggle to explain distance effects, to be described below. And in any case, the reason this oscillator was used would relate to its implementing efficient learning of sequences.

Another reason to suspect that monkeys represent these chains in full is that they are able to spot patterns in how the chain is constructed and use these to extend the chain. In particular, they are able to spot when the chains they learn are based on increasing numerosity.[53] While many SimChain experiments use random stimuli or patterns of shapes, kinds of object etc., some teach subjects to pick e.g. the stimulus with just 1 item, then the stimulus with 2, then the

[52] Brown et al 2000 propose a model of human serial order memory works in something like this fashion.

[53] Brannon & Terrace 1998, 2000, Ohshiba 1997, Cantlon & Brannon 2006, Drucker & Brannon 2014 for monkeys, Matsuzawa & Kawai 2000 and Smith et al 2003 for great apes.

stimulus with 3 etc. This can be the only basis for picking, with other features (shape, colour etc.) varying trial-by-trial.

Monkeys find lists with a discernible pattern like this easier to learn than arbitrary lists.[54] When presented with new numerosities (e.g. when trained to order stimuli with numerosities 1-4, and presented with stimuli with numerosities 5-9), monkeys generalized the rule of ordering by ascending numerosity in their first trial.[55]

Monkeys' responses to these lists show distance and magnitude effects which are hard to explain unless they represent the entire list. When subjects are first trained on a serial chain of, say, 5 items A1-A5, they can then be given a further task: given just 2 stimuli from that chain, e.g. A2 and A4 select them in order. Monkeys (just like humans) find this task *easier* (i.e. they can do it more quickly, with fewer errors) for items which are further apart in the list (e.g. for A5-A1- 4 spaces apart rather than A5-A4 - 1 space apart). And, given a distance between items (e.g. 2 spaces apart), they find it easier to compare items earlier in the list (i.e. A3-A1 rather than A5-A3).[56] This effect is not just found with lists where there is a clear order (e.g. based on numerosity of the stimuli), but for arbitrary lists too.

If we allow that monkeys represent ordered sequences as ordered sequences, there is a natural explanation of magnitude and distance effects. These representations of the sequences involve mapping items onto some sort of analogue magnitude representation, according to those

[54] Brannon and Terrace 1998, 2000.

[55] However, they may be better at doing this for lists based on ascending rather than descending numerosities - Brannon and Terrace 2000.

[56] See Terrace 2005 Fig. 3 for graphs superimposing results of this kind for numerous studies in humans and animals.

items' ordinal positions. Given that analogue magnitude representations display magnitude and distance effects, the representations of order would inherit these.

2.8 *Temporal* Representation.

So much for what representation of time adds to systems with dynamics built into their functional roles. But what of representation of the *temporal* as opposed to representation of related properties? To return to the Scrub Jays case, we could think that instead of representing duration since caching, they are representing *freshness* in a way governed by a timer. What difference does *temporal* representation make?

Temporal properties tend to be rather abstract relative to their rivals. They are certainly less situation-specific, and are relevant to nearly all systems. This means that they can combine with virtually any variable, and relate virtually any two variables as a kind of common currency. We saw some of this in the monkey sequence learning where we had transfer of temporal properties between somewhat dissimilar stimuli, with faster learning of new sequences composed of existing items depending on their ordinal positions.

However, it might be that many forms of temporal representation are not pure and unambiguously about time. There may be degrees of ambiguity, with representations which could be interpreted as being about duration or as being about some other, closely correlated, property. We should not shy away from the possibility of somewhat confused representations, especially in animals. Carey 2009 points to multiple cases where young children use representations which systematically confuse e.g. weight and mass. Likewise, there may well be animals who systematically confuse duration and freshness, or at any rate use representations which do not discriminate between them. This is a controversial position: some would deny that

genuine representation is possible unless all relevant alternative contents are ruled out,[57] while some versions of the claim would be vulnerable to arguments that genuine vagueness is impossible.[58] But it is not an uncommon response to versions of the disjunction problem.

If this is right, then we need, in any given case where we have a representation involved in explaining a temporal competence, to decide between the options of the representation being about some other property entirely, usually a quite specific one, being about time, or being a confused representation which has features of both of these options. Which category it falls into will depend on how the representation behaves. If it is only about freshness, say, then it will only be usable in the context of food. If it is unambiguously about time then it will be usable in a much broader range of contexts. It is possible to use freshness as a means of achieving competence with respect to other durations, just like we could count intervals in terms of how full of nectar a flower would need to be; but such representations would need to be tied to a representation of a caching event, and we would expect this use to be specialized and uncommon. As the representational practice of using 'freshness' representations to in effect measure other intervals of time becomes more abstracted from particular instances of food: as imaginary caching events with only skeletal details become enough, and as the creature uses the technique in a wider range of circumstances to achieve temporal competence, it is more and more useful to think of the 'freshness' representation not as a rival to temporal representation, but as ambiguous between a specifically freshness-related representation and a temporal representation.

[57] Burge (2010: 466).

[58] Williamson 1994.

2.9 Conclusion

In human experience, temporal representation seems to us primitive and ubiquitous: while we have to learn to use clocks and calendars and think about long time periods, our ordinary perceptual experience is shot through with representation of the order, duration, and rhythm of experienced events. Does this suggest that animals probably have such experience too? Or is animals' relationship to time fundamentally different to our own?

There are a few options here. One is that temporal representation *is* widespread. Nothing we have said in this paper should be taken to deny this possibility. Hoerl and McCormack tend to argue from the possibility of explanations of temporal competences that do not invoke temporal representation to the claim that animals with those competences do not have temporal representation. But I have not offered any reason why our credence should be higher in the alternative explanations offered here, whether in terms of functional role or representation of alternative properties, than in the explanations in terms of temporal representation. In many cases, these rivals are not even more parsimonious than the temporal representation explanations, positing representations of other properties or additional functionally defined kinds of state. Rather, the point of the alternative explanations has been to highlight the issue of what distinctive explanatory value positing temporal representation gives.

A second option is that our temporal experience is closely based on a more widespread experience whose representational content is ambiguous between being about temporal properties and more domain-specific properties like freshness or number of actions per event.

A third option is that many animals have quite a different experience to ours. What's widespread is temporal competence; in humans temporal experience is phenomenologically important but only because our culturally learned concepts/perceptual categorizations affect our

experience, or because of our way of metacognizing our experience means that we routinely recognize the temporal upshots of what in animals are processes that do not involve temporal representation.

We have seen that all of these are live options, but we have also seen how to decide between such options. To determine whether an animal is using temporal representation, and the degree of ambiguity about its being temporal, we need to consider two question. First, how far do they have a representation rather than their temporal competence being underpinned by a dynamic functional role? This question will be answered by the kinds of flexibility in using the representation and how far these would imply variation in the functional role in question. Second, should we think of that representation as temporal, as having some other kind of content, or as ambiguous between the two? Here, the answer will depend on the degree to which the representation is usable in contexts foreign to the alternative kind of content, and the extent to which the mechanisms involved eschew the alternative kind of content in those contexts, using just more general mechanisms suitable for the more abstract temporal properties.

Chapter 3: Do Animals Represent Time? Evidence for Temporal Representation and for Other Traits in Animals

3.1 Introduction

Chapter 2 showed that a great deal of temporally appropriate behaviour is possible without explicit representation of time, and that only certain kinds of flexibility require it. One might conclude from this that temporal representation is not widespread in the animal kingdom. If most or all of the behaviour we observe could be explained without positing temporal representation in animals, why do so? At the extreme, one might even deny that the evidence discussed concerning scrub jays and primates establishes that they have temporal representation, given that even there our evidence for the relevant kind of flexibility has its limits, such as the scrub jay evidence nearly all relating to food caching rather than being domain general. Pressing this kind of line, one might claim that temporal representation is unique to humans. Hoerl and McCormack 2019 do argue along these lines, arguing that animals and young children only use a Temporal Updating System.

This line of thinking is mistaken. It reflects a confusion about methodology which infects many other discussions of animal and infant psychology.

It does not follow from the existence of anti-representational (or representational but anti-temporal-representational) explanations of experimentally demonstrated animal behaviour that these animals do not have temporal representation. It does not even follow that the behaviour in question is not evidence for temporal representation in these animals. This is so even if we establish that these alternative explanations really are anti-representational, rival explanations rather than descriptions of representational phenomena at a lower, implementational level.

To see why, we should return to Chapter 1's distinction between questions about the constitutive core of a trait, on the one hand, and questions about what constitutes good evidence

for that trait, on the other. We can allow that something can be evidence that an animal has some mental trait, without establishing this via being evidence for the animal's having all the core features which are unique to that trait. Instead, such evidence can be more circumstantial. It can involve behaviour which could be explained by positing the trait in question but could also be explained under other hypotheses. It can operate via behavioural, neural or any other features thought to correlate with the trait in question even if they are not constitutively implied by it. It can involve considerations about which mechanism for explaining some behaviour is not just in principle possible but likely, given what we know about how such mechanisms might be implemented or might have evolved. In principle, evidence can come from a wide range of sources.

The upshot of this way of thinking will be that we should have some confidence that at least simple forms of temporal representation are quite widespread in the animal kingdom, even if that confidence should vary by species and often dip considerably below 100%.

But this way of thinking needs to be thought through. It runs counter to the thought (which does have something importantly right about it) that in science we should not believe in a mechanism on the basis that it *could* explain some phenomenon if there are alternative mechanisms available. And it runs counter to a principle which has been hugely influential in animal psychology and continues to exert a pull in many quarters—Morgan's Canon:

> "*In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale*".[59]

[59] Morgan (1904: 53). Morgan (1904: 59) qualifies this principle, claiming it "by no means excludes the interpretation of a particular activity in terms of the higher processes, if we already have independent evidence of the occurrence of these higher processes in the animal under observation." But he does not say how strong this independent evidence would need to be or what forms it might take.

It runs counter to a tendency in animal and developmental psychology to look for crucial experiments—experimental paradigms such that success or failure on this one task is taken to demonstrate possession or lack of some specific trait.[60] Furthermore, it is a complicated question exactly which considerations should enter into deciding which species have which mental traits, and how much different considerations should count for: In particular, the question of how (if at all) to use evolutionary considerations is a fraught one.

Because of controversy over these sorts of methodological issues, there is considerable controversy about how ready we should be to attribute different kinds of mental states to animals, even within the community of animal psychologists/cognitive ethologists/comparative evolutionary psychologists (there is also related dispute about the appropriate term for the field). Some people are very willing to attribute various impressive-sounding mental states to animals even when the behavioural and neural evidence we have is consistent with a range of alternative hypotheses that attribute only less-impressive-sounding mental states. The more sophisticated defenders of this position often appeal to considerations like an animal's being relatively closely related to humans in evolutionary terms—the same kinds of considerations that might drive us to suspect that our mystery ape had chimp- or even human-like mental states. Others are much more cautious, demanding a higher amount of evidence before they are willing to endorse 'rich' over 'lean' interpretations of animal behaviour, often appealing to

[60] The mirror test is often treated this way as a test for a kind of self-consciousness, and the Sally-Anne false belief task as a test for theory of mind. In the present context, we can see this sort of thinking at least implicitly exerting an influence in, the following: "From a developmental point of view, a key question we need to ask is: what are circumstances that would provide clear evidence as to whether or not children have a form of understanding of sequences that cannot be explained by the possession of a script?" Hoerl & McCormack 2011: 451. This can be read innocuously, if it allows for the 'circumstances' in question to be quite broad; but Hoerl and McCormack go on to look for a particular experiment which will do this work.

principles like Morgan's Canon. It has been shown that both approaches have problems, however.

I will show that reframing this debate in terms of the degrees of credence different kinds of evidence can sustain helps us capture the best of both worlds—taking into account the limitations of evolutionary considerations and anthropomorphism which the 'conservative' position espouses, whilst allowing that such considerations have some force. Furthermore, thinking about these issues in these terms allows us to generalize the issue of cross-species comparison and use of evolutionary considerations to issues about how we should use model organisms, and to incorporate different kinds of (sometimes conflicting) evolutionary considerations relating to phylogeny on the one hand and selection pressures on the other.

This is not to say that the methodological position proposed here is entirely novel. As I will show as I go along, many scientists do think in these terms already, at least some of the time. Furthermore, I will be drawing on critiques of Morgan's Canon developed by other philosophers. However, the alternatives to Morgan's Canon proposed by these philosophers (and by scientists who discuss their methodological assumptions explicitly) are too underspecified to guide the use of evolutionary and other sorts of consideration in science. Partly as a result, scientists and philosophers have not been entirely consistent in their application of these ways of thinking.

The general issue of how to form beliefs about which animals have which traits, on the basis of direct behavioural and neural evidence, mere correlations, and evolutionary considerations, is important for a few reasons. Sometimes the kind of careful observation of behaviour we would need to do to rule out *all* alternative hypotheses about how a given animal's mind is working is extremely costly, both in financial terms and in terms of inconvenience and

worse to animal subjects. Sometimes it is arguably *impossible* to decide these questions one way or another using behavioural considerations, although nothing I say here depends on this claim.[61] The same goes for neural evidence—it is typically highly costly to obtain, and can only be obtained in small samples from a small number of organisms from a limited number of species. Given this costliness, if there are cheaper forms of evidence that would allow us to rationally shift our credences about which creatures have which capacities more (relative to the same amount of (or less) behavioural and neural testing), we should be availing ourselves of that evidence, even if it is not decisive. Such considerations should shape our research programmes. If we can have considerations in advance of running new experiments that already provide *strong* evidence that a certain creature has a certain capacity, it may not be worth running those experiments. Conversely, if we know that finding the status of a particular mental trait in a particular animal would tell us a lot about the distribution of that trait across many other species, this should be one factor in deciding whether to test for that trait in that animal: model organisms should not be selected solely on the basis of how easy they are to work with, but also their potential for inferences to more general or more important conclusions.

Finally, the case of the epistemology of animal minds can be treated as a case study in more general issues in epistemology and the philosophy of science, such as consilience (how

[61] There are a number of potential candidates like this, though all are controversial. It has sometimes been claimed that we are in this situation when it comes to distinguishing between 'mind-reading'—genuinely representing others as having mental states, and 'behaviour-reading'—merely being able to predict others' complex behaviour, without realizing that behaviour springs from the others' mind. There are also classic philosophical examples where it is controversial how (if at all) we should assign determinate contents to representational states in animals - such as the question of whether we should think that when a frog catches a fly, it visually represents *fly*, *edible prey*, *moving dot* or some other content. Another such case would be distinguishing between a creature's having conscious states and being a philosophical zombie, i.e. a creature with all the behaviour and physical states of a conscious creature but without the same phenomenology. All these sorts of issues, of course, are not restricted to *animal* minds - for example, see Haith 1998 and Carey 2009: 14, 186, 202, 211 for discussion of related problems in infant psychology.

considerations from different areas can inform one another and be unified in drawing a single conclusions), and in the use of parsimony considerations or Ockham's Razor.[62]

This chapter will proceed by first discussing Morgan's Canon and the most important criticisms of that Canon in more detail (§2), before laying out the alternative, more ecumenical Bayesian-based approach I have already suggested in §3. §4 briefly discusses how these issues generalize to illuminate recent discussions about Model Organisms. §5 discusses how evolutionary considerations can be used in inferences about animal minds. One advantage of the Bayesian framework is that it allows us to move away from simply asking whether evolutionary considerations should be used at all, to asking under what conditions different kinds of evolutionary considerations should have more or less weight. §6 explicitly draws out the upshots of the general methodological views for what we should think about which animals have temporal representation. The considerations from Chapter 2 do show that the view that temporal representation is extremely widespread has not been demonstrated to the degree that many of its proponents assume; but the experiments that they appeal to, in combination with evolutionary considerations, do nonetheless provide good reasons to have high credence in temporal representation in a number of animals and medium credence for many more—even in species where these experiments have not been run.

3.2 Morgan's Canon and its Problems

A standard way of telling a simplified history of animal psychology in the second half of the nineteenth century and early twentieth century goes like this: Darwin emphasized the continuity of animal minds with human minds as a consequence of evolution by natural

[62] E.g. Lewis (1973: 87) uses the question of whether positing spirits for animals or just for people as a key motivation for distinguishing qualitative and quantitative parsimony, and Sober 2015 discusses issues concerning animal minds at length as one of the main case studies in his treatment of Ockham's Razor.

selection. This prompted Darwin and his successor George Romanes to try to study animal minds systematically, but they relied heavily on compiling anecdotes of suggestive behaviour and ended up attributing all kinds of things — religious belief, complicated forms of grief etc. — to animals on rather scant evidence.[63] In the late nineteenth and early twentieth century, people started studying animal behaviour with systematic experiments in a controlled laboratory environment, the most famous such work being Pavlov's and Thorndike's research on conditioning. At the same time and in a similar spirit, Conwy Lloyd Morgan published his influential textbook *An Introduction to Comparative Psychology*, which advocated behavioural experiments over anecdotes, and most famously argued for principle quoted above.

This principle became known as 'Morgan's Canon'. In the context of the new paradigms for studying conditioning, it helped along the rise of behaviourism, which (at least officially) eschewed positing mental states at all—especially to non-human animals—treating their behaviour as caused solely by 'lower', simpler sorts of mechanisms. Now, it is not entirely clear how far Morgan's *discussion* of his Canon actually influenced the behaviourists. Morgan is explicit that animals have minds, and seems to advocate not non-mentalistic explanations of animal behaviour, but 'lower' sorts of *mental* processes. But the Canon was certainly quoted approvingly by behaviourists for decades, becoming "the probably most quoted recommendation in all of psychology" in the process.[64] And both philosophers and scientists have continued to appeal to similar principles since the demise of behaviourism in the second half of the twentieth century, claiming that it is a good idea to be conservative in the mental abilities one attributes to

[63] Darwin 1889, especially Chapters 3-4; Romanes 1883, 1892.

[64] De Waal (2016: 42). For catalogues of numerous examples of behavioural scientists quoting and giving rather misleading glosses on the Canon, see Allen-Hermanson (2005: 608) and Wozniak 1997.

organisms to explain their behaviors,[65] that one "should attribute to an organism as little intelligence or consciousness or rationality or mind as will suffice to account for its behavior",[66] or that behaviour "should be explained by reference to less sophisticated representational structures...other things equal".[67]

Morgan's Canon has faced considerable criticism in the last few decades. Its two most important problems are its excessive and unmotivated conservatism, and its use of an unelaborated notion of some faculties or processes being 'higher' on a 'psychological scale' than others.[68] I will explain each in turn.

Morgan's Canon is conservative in the sense of demanding a high degree of evidence before we are permitted to believe in a given animals' having a given mental trait. If an animal's behaviour *can* be interpreted as resulting from an unsupplemented lower faculty, Morgan's Canon instructs us not to attribute it to a higher faculty. At least if we read 'can' here in the way Morgan's Canon is typically interpreted by both its advocates and detractors, this means that if the probability of the behaviour on the hypothesis that the animal just has the lower and not the higher faculty is anything other than miniscule, we should refuse to believe it resulted from the higher faculty. This can be thought of as a policy of *always* allowing worries about type I errors - erroneously attributing the higher faculty when it is not in fact present - to trump worries about type II errors - failing to attribute the higher faculty when it is in fact present. This seems to be a

[65] Cheney and Seyfarth 1990, cf. Povinelli and Vonk 2004.

[66] Dennett (1976: 182).

[67] Burge (2010: 61).

[68] Both of these problems have been developed in detail in the literature on Morgan's Canon of the last 20 years, especially in Sober 2005.

mistake. Both kinds of errors matter, and we should trade off the risks here in a way which gives at least *some* weight to each sort of error.[69]

One might have the thought here that there are two possible forms of conservatism, and one is more objectionable than the other. Consider the question of whether one should believe in God: if one decides that there is not strong evidence to believe in God, one could become an atheist, and actively believe that there is no God. But one could instead become an agnostic, and refrain from either believing that there is a God or believing that there is not. Similarly, one could respond to multiple possible explanations of some animal's behaviour by refraining from believing—or disbelieving—any of them; or one could respond by actively believing the explanation positing only the 'lowest' faculties.[70] These attitudes would have very different practical consequences: confidence that an animal is not feeling pain (or not representing time, or not reasoning) will result in very different behaviour from uncertainty about whether it is. However, even an 'agnostic' attitude here is too conservative. In sometimes failing to believe in true propositions about animals having certain mental states despite having some (inconclusive) evidence for those propositions, one will miss out on the truth. The opportunity cost of systematically always missing out on such true beliefs could be very large, especially in scenarios like deciding whether to engage in some trivial activity which might produce pain in an animal, where the costs of having a false belief that the animal feels pain (missing out on the trivial activity) are low, while the costs of failing to truly believe that the animal has pain are

[69] De Waal 1999, Sober 2005. The thought here is close to the point in James 1896, that we should aim both to minimize error and to maximize truth.

[70] Cf. Sober (2005: 91); Buckner (2013: 854). Sober (2015: 12, 50, 55, 243) also contrasts different versions of Ockham's Razor as a general principle of reasoning: A 'razor of denial' will advocate for believing hypotheses which minimize the number of entities or kinds of entities posited and disbelieving others; a 'razor of silence' will just favour refusing to believe hypotheses where alternative explanations positing fewer entities are available, without forcing disbelief.

high. These sorts of costs will stack up given that it is typically very difficult, and (as was briefly discussed above) perhaps impossible, to rule out all alternative explanations of a given behaviour. Minds and brains are very complex and flexible, and different hypotheses about them can usually be made compatible with an extremely wide range of data through auxiliary hypotheses, especially auxiliary hypotheses concerning pretence, or (as we saw in Chapter 2) extremely intricate architectures. Furthermore, as we will discuss below, the options should not be construed as simply believing, disbelieving, and withholding both belief and disbelief thinking in terms of credences allows us a much broader range of options, which can do much better justice to the different degrees of evidence in different cases.

Not only is it excessively conservative, but Morgan's Canon seems to make ineliminable use of the notion of a 'psychological scale' on which we can rank different faculties.[71] The challenge is to spell out exactly what this means, without committing to some sort of *scala naturae* view at odds with a serious understanding of evolution by natural selection, whilst also making it plausible both that there should be some relatively a priori bias to believing in 'lower' over 'higher' faculties, and that the Canon is widely applicable. We can ask two questions here. First, we can ask the historical question of what exactly *Morgan* meant by 'higher' and 'lower' and whether his Canon looks justified from within his framework. Second, we can ask whether, independently of what Morgan actually said, we can reconstruct a version of the Canon using a different notion of 'psychological scale' which is a reasonable principle to use for our purposes.

[71] One might also worry that it uses a notion of 'faculty', which has largely fallen out of favour in psychology, but as Morgan himself points out (1904: 59) we could replace his talk of 'faculties' here with 'processes'—or, if we liked, 'systems', 'mechanisms', 'traits' etc.—without much changing the spirit of the principle or its appeal.

I will not dwell on the historical question here.[72] Suffice to say that though Morgan does often seem to slip into *scala naturae*-like talk about evolution, including in the passages where he offers a rather difficult to interpret argument for the Canon, and in passages where he sets out his general world view,[73] it is not clear that this way of thinking is essential to the Canon itself or Morgan's justification for it. Rather, Morgan's use of the notion of a 'psychological scale' may depend heavily on the relationship between the particular faculties he believes in: 'instinct', 'intelligence', and 'reason'. As Morgan understands these three faculties, possessing a 'higher' faculty *entails* that one possesses those lower than it, and this may be all that he means by his talk of 'higher' and 'lower' in this context.[74]

Insofar as Morgan's own argument for and understanding of his Canon turns on a particular conception of which faculties/processes there are, and insofar as his conception limits us to just three faculties/processes, it will not result in a useful methodological principle for our purposes. But it is very natural to talk of some mental traits as 'higher' or 'lower', and there are many things which such talk might mean, and these can give rise to different versions of Morgan's Canon. Most people who cite the Canon have one of these other meanings in mind. I will not go through every possible such meaning.[75] Instead, I will note that different options tend

[72] There is particularly helpful discussion of the historical question in Allen-Hermanson 2005 and Clatterbuck 2016, who offer slightly different interpretations in light of what Morgan says elsewhere.

[73] Morgan (1904: 54ff.) and Morgan (1904: 9f.) respectively.

[74] There is controversy about this reading of Morgan in the secondary literature. Allen-Hermanson 2005 makes the case that Morgan understands some kind of temporal or modal 'supervenience' relation to be at work here. The resulting principle is still rather limited, however, as it is not at all clear that, for example, high-level reasoning supervenes on operant conditioning; but the latter is certainly taken to be 'lower' than the latter by most people who wield versions of Morgan's Canon.

[75] See Sober (2005: 91f.) and (2015: 96) for discussion of several options not discussed here, such as 'higher' traits simply being those which in fact appeared later in the history of evolution.

to face one of two problems. Either they do not apply widely enough to help provide for general principles of the kind we might hope for, or they do not seem to be relevant to how ready we should be to attribute mental traits to animals.

The former problem infects, for example, any version of Morgan's Canon that trades simply on an animal's possessing one trait logically entailing its possessing another. If we do not buy into Morgan's particular picture of reasoning and intelligence, we might well think that learning by optimally updating one's probabilities across a hypothesis space according to Bayes' rule does not *entail* also being capable of operant conditioning. But we might well want a principle that tells us how to assess the possibilities that a series of behaviours in response to stimuli was due to optimal Bayesian updating or operant conditioning. And what's more, we might think that such a principle should take into account the point that the Bayesian process is in some sense other than entailment 'more sophisticated' than the conditioning process.

The latter problem, of a notion of 'psychological scale' being irrelevant for our purposes, infects different options slightly differently. One might think that we should rank mental traits by how modular they are. If we buy into something like Fodor 1983's architecture of the mind, it includes a number of highly specialized, informationally encapsulated and inflexible modules in addition to more general-purpose, flexible 'central cognitive' processes, and it is common to think of the more 'central' systems as 'higher' than the more modular ones. One might be able to mount some argument for expecting that animals are more likely to have modules than central cognition. But it is difficult to see what such an argument would look like without appealing to some antecedent principles about expecting animals to rely on 'lower' traits in some further, deeper, less well-defined sense. And without such an argument, any principle in the style of

Morgan's Canon, biased towards assigning only 'lower' traits, would seem to be simply question-begging on important issues about the structure of animal minds.

Other ways of cashing out the 'psychological scale' for the purposes of developing such a principle are not exactly question-begging, but do seem to lead to very implausible principles if plugged into Morgan's Canon. For example, assuming we can make sense of the idea that systems in the mind have particular functions, we could rank systems by how reliably they fulfill their functions. But it seems very implausible to think that we should be biased towards thinking animal minds are more likely to be 'lower' than 'higher' in this sense: if anything, we should assume that most systems shaped by evolution are pretty reliable at fulfilling their functions.

There may well be many notions of 'psychological scale' which are not infected by such problems. Indeed, I will be arguing below that the notion of evolutionary dependence often allows us to recover some sense of a 'psychological scale' in particular contexts, and does support a principle that is somewhat like Morgan's Canon in the sense of motivating a bias towards attributing traits to non-human animals that are slightly 'lower' than human mental traits. But before we can develop such a principle, we need to have a better alternative to Morgan's conservatism in hand.

3.3 Bayesian Alternatives to Morgan's Canon

We should not be excessively conservative and insist that whenever we find evidence that an animal has some mental trait (e.g. temporal representation), we should disbelieve that the animal has that mental trait—or at least completely withhold belief—as soon as we articulate an alternative explanation of that evidence. Instead we should take a more graded approach, allowing different forms of evidence to shift our credences according to their strength. We should not expect to to get overwhelming evidence for one position, and should not wait until we have done so before gaining any degree of belief in any position. We can keep multiple

positions live, according them different weight based on the evidence interpreted in light of our priors.

For example, when considering the question 'do alpacas have temporal representation?', we should afford some credence in each of:

1. Alpacas do not have any temporal competence. Any appearance of such a competence is a fluke.
2. Alpacas have a temporal competence which is underwritten entirely by non-representational coordination.
3. Alpacas have a temporal competence underwritten entirely by non-representational coordination in combination with representation of other, time-correlated factors.
4. Alpacas have representations which are indeterminate between temporal and other contents.
5. Alpacas have determinately temporal representations (and within this broad position, we could have many varieties, concerning exactly which kinds of temporal representation they have and how they use them).

This does not answer all our questions: there are still substantive issues about what should shift the credences in each of these up or down and by how much. But it is significantly better than the belief/disbelieve/suspend judgement entirely way of thinking.

One advantage of this approach is that it allows for, but does not demand, conversion of these degrees of belief into more precise characterizations of probability distributions, with explicit use of Bayesian updating in response to evidence. This requires not only being prepared to specify quite specific priors (although we can get round this sort of issue to some extent by performing the analysis multiple times by considering different priors), but also modelling some of the assumptions one is using explicitly, for example modelling processes of evolution when one draws on evolutionary considerations, and committing to precise formulations of any psychological hypotheses. Such modelling can be illuminating given enough evidence, and Sober 2015 derives useful conclusions about evolutionary reasoning from taking this approach, as we will discuss below. But it is not obligatory to actually carry out any of these procedures in order to derive the main benefits of the view on offer here: these just result from accepting that

we can allow evidence of different strengths to make a difference to our beliefs about animal minds, provided that these beliefs are graded rather than absolute.

One advantage of accepting such a view, even in broad terms, is that it allows us to recognize a much wider range of phenomena as evidence that some animal has temporal representation. It allows us to be responsive to forms of evidence which are genuine evidence but by no means conclusive. *Being evidence* need not be a demanding notion: on a standard Bayesian formalization of the notion of evidence, E is evidence for H just in case $P(H)<P(H|E)$. This picture allows that ruling out—or even rendering less probable—*some* (but not all) alternative explanations will often count as evidence. For example, finding impressive temporal coordination in alpacas might raise our credence in *all* of hypotheses 2-5 above (i.e. in *both* hypotheses that attribute temporal representation and in (some) hypotheses that do not). This is because it allows us to discard hypothesis 1, that alpacas cannot coordinate with the temporal feature in question at all, (as well as any hypotheses on which they cannot do so as extensively, accurately or impressively as our new evidence suggests).[76] But besides direct evidence of temporal competence, we can also recognize weaker forms of evidence, like evolutionary considerations combined with temporal coordination findings in *other* species, as well as mere correlates of temporal representation, like other forms of intelligence and representation. These will provide some evidence in the above sense, even if not as strong evidence as direct evidence for the core features of temporal representation.

One might worry at this point that the notion of evidence in question is too weak to have any bite. Having evidence in this sense does not, after all, imply raising one's credence by a

[76] Strictly speaking, rather than entirely discarding hypothesis 1, we should cut its probability to close to 0, given that the observed behaviour which looks like temporal coordination could be a fluke, at least in principle.

large amount. Surely, we need to worry not about evidence, but about *strong* evidence: evidence such that P(H|E) is much higher than P(H). And we should worry about whether we have good reason to have a *high* credence that some animal has temporal representation: 0.2 vs. 0.1 is not all that interesting. After all, we can discover intricate kinds of temporal coordination in bacteria and plants, and this counts as evidence that they have temporal representation on this view of evidence. But it is not strong evidence: at least if we are not deflationist about temporal representation and think of it as allowing for much more than just temporal coordination, as was suggested in Chapter 2. In this case, at least, our credence in the temporal representation hypothesis should go up, but should still be dwarfed by the credence in non-representational but sophisticated temporal coordination.

But the shift to thinking in terms of evidence allows another move here. Evidence can accrue from multiple considerations, many of which may be somewhat weak evidence on their own but can, in combination, add up. And a key advantage of moving to the degrees of belief framework is that it allows us to see how different forms of evidence can be *combined*, such as different forms of evolutionary consideration, as we will see below.

One might have two kinds of worry about this approach. Firstly, one might offer direct reasons to be suspicious of degrees of belief, or at least to demand independent principles for determining outright, binary belief. Secondly, one might agree with what has been said so far but wonder whether any more can be said in general besides very general principles like 'update your credences in response to incoming evidence according to Bayes' law', with the application of these principles requiring careful modelling of very specific details of any given case.

There may be some contexts, like in framing legislation or in judicial contexts, where a non-graded approach may be more desirable than a graded approach. For example, Birch 2017

advocates using precautionary principles for thinking about animal sentience in the context of animal welfare legislation. This would involve demarcating exactly which species should be protected and offering them all the same and this sort of protection. The reasoning used to decide on which species should be protected would go more like this:

> *There is good evidence that species A can feel pain. So all the species in A's order might feel pain too, so all the species in A's order should be protected.*

than like this:

> *There is good evidence that species A can feel pain. So species in A's genus probably do too and should be protected; species in A's family might well do so probably should be protected/should be afforded some protections; it is plausible that all/many species in A's order feel pain so they should be afforded some lesser protections/maybe should be afforded protections; and it is possible that many species in A's class feel pain too*.

The reasons to avoid the latter sort of reasoning in legislation relate to many considerations which do not replicate at the individual level, however: the fact that more complex reasoning is less transparent to non-experts and therefore easier to manipulate to suit special interests, the need for decisive policies and the unavailability of 'in-between' options, difficulties with pooling complicated preferences and beliefs in a democratic way, risks of unintended consequences of actions, and so forth. Deciding public policy neither can nor should proceed on the same basis as deciding what we should believe; and I do not know of strong reasons to drop the more sophisticated framework of credences for the individual-level project.[77]

[77] Buchak 2014 argues that in general, outright belief is neither necessary nor sufficient for high credences, and that the main use of outright belief has to do with normative and especially legal contexts, like deciding guilt and innocence. As with Birch, it is not clear that her considerations apply to the individual level or just for collective decisions like legal cases.

The other worry with what I have said so far is that I will not be able to say any more, at least outside of very specific cases of assigning credences to attributions of particular kinds of temporal representation to particular animals. Both Sober 2005 and Fitzpatrick 2008 advocate for such views, on which Morgan's Canon is redundant not just because it is excessively conservative and relies on a potentially suspect notion of 'higher' and 'lower', but because any function it could allegedly serve is better served by much more general epistemological principles, having nothing in particular to do with animal minds. For example, perhaps we could get by with *don't believe any hypothesis without sufficient evidence,*[78] or just using Bayes' Rule?

If we *were* ideal Bayesian agents, this *would* be enough. We would start with a probability distribution over all possible hypotheses, which would already take into account all the evidence available to us up to this point. To update our beliefs, we would simply apply Bayes' rule in the standard way, using our priors, the observed behaviour, and the probability of the observed behaviour under the different hypotheses. However, scientists do often appeal to general methodological slogans, including Morgan's Canon, for a reason.

It is difficult for ordinary humans to figure out which considerations count as relevant and how strongly to weight them. It is often not obvious exactly what *is* evidence pertaining to these particular questions, so in the context of any given single experiment it will not be the case that we actually start with priors, yet alone priors that incorporate all available evidence rationally: evolutionary considerations especially are hard to assess. In this context, it is useful

[78] Fitzpatrick 2008. Note that this principle as formulated unhelpfully presumes a believe/disbelieve dichotomy. Cf. also "Perhaps these default principles deserve to be swept from the field and replaced by a much simpler idea - that we should not indulge in anthropomorphism *or* in anthropodenial until we can point to observations that discriminate between these two hypotheses...the best way to minimize the risk of *both* types of error is not to embrace an a priori prejudice. The only prophylactic we need is empiricism" (Sober 2005: 97)—which again is too conservative, as argued above.

discussing some of the pitfalls and advantages of such kinds of evidence in general, especially as the same kinds of potential evidence are liable to arise in many of these cases, with many of the same potential pitfalls and advantages, which are harder to spot when reinventing the wheel due to considering the cases in isolation from one another.

Furthermore, thinking about these issues in general, principled terms is particularly important in the context of animal minds, because there are often strong biases and ideologies which play out differently in particular cases, relating to motivated reasoning to avoid the potential ethical implications of animals having complex minds,[79] to the cuteness or otherwise of particular animals,[80] the human tendency to anthropomorphize and attribute minds like ours even to inanimate objects,[81] and to the kinds of ideologies about what it means for science to be 'hard-nosed'.[82] Given all this potential for sloppy thinking and the likelihood that it will play out inconsistently across different cases and in different literatures, it is worth stepping back and discussing these issues at a somewhat more general level than the individual case.

When we step back in this way, we can see that the relevant issue is not just whether, say, alpacas have temporal representation, or even extrapolation from humans to animals, as much of the Morgan's Canon literature presupposes. Rather, the framework here will unify a number of different kinds of case where the same issues arise. When we should attribute mental traits that we have studied in humans to other animals is one of these cases. But so is the question of when

[79] Midgley 1973, Andrews 2009, Singer (2009: 9ff.),

[80] Wilder (1996:35) claims in an oft-quoted passage that Morgan's Canon is 'a welcome prophylactic against sentimental and unwarranted attributions of finer sensibilities to our favorite animals'.

[81] Heider & Simmel 1944, Horowitz & Bekoff 2007, Horowitz 2009, Carey 2009 Ch. 5.

[82] Keeley 2004, De Waal 1999, 2016, Andrews 2009. As Sober 2005: 86 points out, James 1896's discussion of 'soft' and 'tender' dispositions is highly relevant here.

we should attribute mental traits we have studied in non-human animals to humans — an issue discussed a great deal recently in philosophy of neuroscience and biology in the context of model organisms — and when we should attribute mental traits we have studied in one species to another species more generally (not just for cases where one of the species in question is humans). In all of these cases, it is natural to appeal to evolutionary considerations.

Issues of needing to appeal to evolutionary and similar considerations due to insufficient direct evidence are particularly systematic in the case of the mind. The mind and brain are so complex and flexible that hypotheses about the mind are nearly always somewhat underdetermined by data: we nearly always *can* come up with alternative hypotheses, by positing different background preferences or beliefs, an intention to mislead about one's mind,[83] or (as we saw in chapter 2) anti-representational alternatives turning on the architecture of the brain being intricately set up to produce the behaviour in question.

However, related kinds of evolutionary reasoning also arise for questions about non-psychological but non-directly-observable properties of animals. For example, do okapi calves defecate in the woods? In captivity, okapi often do not defecate in the first 4-10 weeks of their life. Some have speculated that this also holds true in the wild, suggesting that it has some adaptive benefit: for example, they tend to spend much of their first few weeks hiding in a bush, and creating strong smells could alert predators to their presence.[84] But it is also possible that such behaviour is due to captivity, caused by stress, boredom, confinement, or other features affecting either calf or mother (mothers do often behave abnormally in captivity, notably tending

[83] Dennett (1996: 13f.).

[84] Bennett & Lyndaker Lindsey (1992: 439).

to excessively lick their young's rear, to the point of causing serious health problems).[85] It is extremely difficult to observe okapi in the wild: they are extremely good at hiding. So deciding between the hypotheses here will depend significantly on considerations about how we should test adaptationist claims, and how far we can extrapolate between the behaviour of different populations of one species.

Another case of extrapolation between species, and between groups within one species subjected to different environments, is the case of model organisms. Model organisms are used both for studying psychological (and neural) traits, and for purely physiological research in medicine, again bringing out the generality of these issues.

3.4 Model Organisms

Neuroscience, like other areas of biology, tends to proceed by studying model organisms: a small range of species with systems studied in great detail, and treated as models for other versions of those systems in other species (sometimes called 'target species'). Often this will involve studying some non-human animal as a model for human systems. This practice extends back a long way in neuroscience: even the discovery of nerves and their having a role in motor responses was first established by Galen by doing experiments on pigs.[86] But the target systems are not always human, and often the model systems are studied in place of target systems from a wider range of species.

Many considerations inform decisions about which species to study, such as ease of dealing with particular species in captivity, ease with which those species can be acquired (perhaps they can be captured in the vicinity of the lab, or bought at a local pet shop; perhaps

[85] Warhol & Benirschke (1986: 44f.).

[86] Gross 1998.

they are extremely expensive), how much is known already about that organism, and of sheer institutional inertia. But particularly important will be some combination of simplicity of the system as it appears in the model, and the extent to which findings can be projected to the target system.

For example, here is a philosopher giving a fairly standard account of discussing animals to support her general view of representation:

> "The frog and toad are also good subjects for our purposes. Biologists believe that much of what they've learned about the anuran nervous system applies to a wide range of vertebrate species. Yet the relevant information processing is, while complex, simpler than in many other species, relatively speaking...It is much easier to see in such a case how content ascriptions can fit or fail to fit the information-processing explanations of the relevant cognitive capacities that scientists provide."[87]

There is considerable controversy about how extrapolation from model to target species works, and how far it is legitimate, especially in neuroscience, and especially in cases where model systems are much simpler, or importantly different, to targets.[88] For example, the extent to which rodents are used relative to other organisms is often criticized, on the grounds that rodents may well not be representative of other organisms of interest—indeed, lab rodents are often so inbred that they may not be representative of wild rodents.

Sometimes similarities and differences between the model and the target can be determined through direct observation, as (in Kaplan 2018's example) when *Streptomyces lividans* was confirmed as a good model for potassium channels by MacKinnon et al 1998 showing its conservation. But in other cases, other considerations need to be appealed to.

[87] Neander (2017: 99).

[88] Preuss 2000; Hedges 2002; Manger et al 2008; Schnabel 2008; Steel 2008; Bechtel & Richardson 2010; Kaplan 2018.

Current discussion (e.g. Kaplan 2018, Weber 2004) of using evolutionary considerations in guiding model organism selection emphasizes considering whether the system under study is 'phylogenetically conserved'; the considerations below will shed light on how such considerations work and when they provide good evidence, but also on other sorts of relevant evolutionary consideration which could be appealed to, relating to selection pressures.

One reason it is useful to expand our concerns to include model organisms here is that in this case, the conservative approach found in Morgan's Canon looks particularly implausible. It would amount to insisting on studying each species one by one, because whenever one extrapolated there would be some alternative explanation of the findings positing a difference between the model and the target. Given that model organisms are often used successfully in apparently hard-nosed science as models for other systems, the conservative would owe us a good explanation of what is going on here.

Another point which the model organisms case brings out is a further advantage of the approach in this paper (and in the thesis more generally): everything I say is compatible with a revisionist attitude towards folk psychology and introspection. We need not assume, as Morgan 1904 did, that the best way to study the mind will always be to discover the workings of psychology in our own case and then project, as far as possible, to similar cases. We can allow that we might learn more about our minds from studying a range of species, developing good models of their minds, and extrapolating those models to humans then making appropriate modifications, than we can learn about animals' minds from studying our own. This view is required by any use of model organisms for studying the mind. It also has independent plausibility.[89] And it has a long history: something like this point can be found in some of

[89] See Wynne 2004; Shettleworth 2010; Penn 2012; Buckner 2013; Andrews (2015: 15ff.); and even Morgan (1930: 250) for defences of similar views.

Hume's attacks on explanations of human behaviour that appeal to 'Reason'.[90] But it is not always appreciated in discussions of animal minds, where attention is often focused on whether animals *really* have versions of certain traits humans are assumed to have. We will return to this issue in Chapter 5.

3.5 Evolutionary Considerations and Animal Minds

When inferring from one species to another, we can draw on two rather different kinds of evolutionary consideration. Most treatments of model organisms or alternatives to Morgan's Canon focus on one or the other sort of evolutionary consideration, but it is important to understand both, and how they interact. We might expect animals to share similar traits if the two animals are closely related—like wolves and dogs—or if they face similar recent selection pressures—like wolves and thylacines. I will discuss both sorts of consideration in turn, seeking to understand the conditions determining the strength of evidence they provide in different cases; I will then discuss some limitations which will often apply to evolutionary reasoning in these sorts of contexts.

3.5.1 Phylogenetic Relationships

If both scrub jays and monkeys are found to succeed on the very same temporal task, such as a what-where-when task or a serial order task, should we treat such findings differently in these two species? Should we be more inclined to think that to pass this task, the monkeys use similar cognitive processes to those humans would use than we are with scrub jays, and more willing to countenance the scrub jays using very different processes? As just discussed,

[90] See especially Hume 1998, §IX.; for commentary on this point, see Kail 2007 and Clatterbuck 2016.

considerations of how related organisms are often play an important role in extrapolations from model organisms. How does such reasoning work, and when is it most secure?

Morgan himself thought that it is in some sense *simpler* to attribute human-like mental traits to animals which are closely related to us, although he did not think that this sort of simplicity was epistemically important. Following Sober (2005, 2012, 2015), we can formalize the kind of simplicity involved and the idea that it is epistemically important in terms of *cladistic parsimony*. The basic idea of cladistic parsimony is that in advance of firmer evidence to the contrary, we should assume the smallest number of changes in traits in any given branch of the tree of life we can.[91] That is, in effect, we should choose the hypothesis about the distribution of a given trait over the tree of life as our starting point that minimizes the number of times a new trait springs up or gets lost within different branches. So if we know that A is the ancestor of B and B is the ancestor of C, and we know that both A and C have trait T, we should choose that hypothesis B had T too over the hypothesis that A had T, it was then lost in B, and then it re-appeared in C. And if we know that M and N both have trait T, and we do not know of any descendants of M and N's last common ancestor L lack trait T, then we should choose the hypothesis that L had trait T over the hypothesis that L did not (which would require that T arose twice, once in the lineage leading from L to M and once in the lineage leading from L to N).

[91] This is rather different, and often cuts in an opposite direction, to a rather different version of 'parsimony', trying to minimize the complexity of the mind you attribute to any given species. This idea of parsimony is sometimes appealed to, for example when Camp (2009b: 124) describes the view that baboons represent dominance hierarchies with a domain-specific tree-like structure as 'more parsimonious' than the view that they represent them with a more elaborate, domain-general language-like structure. The basic problem with this form of parsimony is that it is unmotivated, particularly if it favours refusing to attribute mental kinds to some species on parsimony grounds when we have already allowed them into our ontology for other species. Furthermore, if it is applied on a species-by-species basis without taking into account evolutionary considerations, it tends to result in positing very large jumps in capacities in short evolutionary time-spans, simply because we have evidence for some capacity in one species and do not (yet) have direct evidence for anything like it in any nearby species.

A common (though not entirely uncontroversial) use of cladistic parsimony in biology is in using knowledge about which creatures have which traits to reconstruct the genealogical relationships between those creatures.[92] If we know the traits of various historical species, we can choose our hypotheses about who descended from whom so as to minimize the number of changes in traits between ancestors and descendants. Another fairly standard use is to utilize knowledge about genealogical relationships and knowledge of current creatures' traits to infer the traits of long-dead species where those traits aren't known from the fossil record.[93] An instance of this sort of reasoning is implicit here: "Technically, fossil feathers of neither *Troodon* nor young *T. rex* are known, but in both cases, fossils have been found of very closely related species bearing feathers".[94]

The parallel inference between the minds of contemporary animals would look similar to the inference in the dinosaur case, with two main differences. First, the reason we don't know about the psychological traits of a given species aren't because it has died out leaving few traces, but because exactly what mental processes are the cause of observable behaviour is difficult to observe.[95] Second, and more importantly for the strength of the inference, there may well be fewer closely related species whose psychology is well-understood to compare to the species of

[92] For details, see Sober (2015: 153ff.).

[93] See Sober (2015: 161ff.).

[94] Losos (2017: 321).

[95] As Sober (2015: 192) points out, there is nothing special about mental traits which makes this sort of inference work better or look more appealing than other sorts of traits. However, it is more likely to be useful in the case of mental traits because it is more difficult to observe them - there would be no need to use such convoluted reasoning to make a guess at whether some creature has feathers if it is alive today and we are in a perfectly good position to simply look and see.

interest than there are closely related species who are known to be feathered or unfeathered in the dinosaur case.

Roughly, cladistic parsimony implies that more closely related creatures are more likely to share traits, and it can seem to license attributing human-like mental traits to apes, primates, mammals and beyond. There is cladistic-parsimony-like reasoning at work in some statements to this effect from various comparative psychologists, such as the following from De Waal:

> "the most parsimonious Darwinian assumption is that if closely related species - whether they be squid and octopus or humans and apes - show similar solutions to similar problems, they probably involve similar cognitive mechanisms"[96]

> "Anthropomorphism [i.e. attributing human-like mental traits to animals] and anthropodenial [i.e. denying that animals have human-like mental traits] have an inverse relationship: the closer another species is to us, the more anthropomorphism will assist our understanding of this species and the greater will be the danger of anthropodenial. Conversely, the more distant a species is from us, the greater the risk that anthropomorphism will propose questionable similarities that have come about independently". [97]

It is not just De Waal who appeals to this kind of reasoning. For example, the following is fairly typical:

> "Prospection-like abilities exist to varying degrees in great apes. Given their close evolutionary lineage to humans, abilities in primates are more likely [than abilities in corvids] to be proto-forms of human abilities, so they might reflect common ancestral capacities rather than convergent evolution such as that seen in scrub jays."[98]

More precisely, cladistic parsimony supports two relevant kinds of inference. First, if two species X and Y are known to have trait M, it supports attributing M to those creatures

[96] De Waal & Ferrari (2010: 202).

[97] De Waal (2016: 25).

[98] Buckner and Carroll (2007: 54). Cf. also "Given that animal minds are organized along massively modular lines, then normal biological reasoning should lead us to expect that massively modular architectures will be preserved in the minds of members of Homo sapiens, too." - Carruthers 2006: 149, critiqued in Wilson 2008 and Brown (2019).

between X and Y phylogenetically. Second, where there is a behaviour in a relative Y of species X that is equally well explained by attributing some mental trait M_X which X has, and attributing some other mental trait M_N which cannot be found anywhere else in the animal kingdom, it supports choosing the first of these options. This is because it would only require positing one emergence—of M_X in the last common ancestor of X and Y—whereas the alternative would require positing at least two emergences— of M_X and of M_N.

Sober shows in detail that we can use detailed models of evolutionary processes to support versions of cladistic parsimony, in particular showing that under reasonable assumptions about evolutionary processes giving rise to a trait M, we can derive that Pr(Chimps have M|Humans have M) > Pr(Chimps have M), via cladistic parsimony-like reasoning.[99] The basic intuition for why evolution supports cladistic parsimony is that natural selection requires traits to be inherited relatively stably, with variations occurring only occasionally: otherwise there would not be any sustained process of selection on traits, but simply random variation generation to generation.

However, Sober also shows that such considerations typically will not provide *strong* evidence for inferences between chimpanzees and humans: Pr(Chimps have M|Humans have M)/Pr(Chimps have M) may be close to 1. This is so particularly where the only species known to have M is humans. In such a case, to get *strong* evidence for chimps having M from humans having M, we need to assume that P(humans have M| most recent common ancestor of humans and chimps has M)/P(humans have M| most recent common ancestor of humans and chimps does not have M) to be high.[100] But this will be high only if we can rule out the possibility that M

[99] Sober (2015: 193ff.).

[100] Sober (2015: 195).

developed rapidly. Intuitively, the issue here is that if M might have developed rapidly, and the only species we know to have it is humans, it might just as well have developed since we split with chimps as before, whereas if it developed slowly, its process of evolution is likely to stretch back past our last common ancestor with chimps, who are in turn likely to have conserved a version of it in their branch of the lineage from that LCA.

We need not always worry about rapid evolution. In some cases, such worries can be assuaged by independent direct evidence for the relevant brain areas being homologous and conserving key features, whether that evidence takes the form of structural similarities in the brain areas, ontogenetic evidence that they grow from the same embryological structures, or direct genetic evidence.[101] However, this strategy depends on knowing quite a lot both about the brains of all the species involved and how the cognitive trait of interest is realized (as well as the genetics and/or ontogeny of those brain structures).

Often, the cognitive scientists who appeal to the idea that closely related creatures are likely to have similar psychological profiles do not just have cladistic parsimony in mind, but are also making a more or less explicit assumption about rapidity of evolution. This is explicit, for example, in De Waal:

> "Morgan's Canon was seen as a variation on Occam's razor, according to which science should seek explanations with the smallest number of assumptions. This is a noble goal indeed, but what if a minimalist cognitive explanation asks us to believe in miracles? Evolutionarily speaking, it would be a true miracle if we had the fancy cognition that we believe we have while our fellow animals had none of it. The pursuit of cognitive parsimony often conflicts with evolutionary parsimony. No biologist is willing to go this far: we believe in gradual modification. We don't like to propose gaps between related species without at least coming up with an explanation. How did our species become rational and conscious if the rest of the natural world lacks any stepping-stones? Rigorously applied to animals—and to animals alone!—Morgan's Canon promotes a saltationist view that leaves the human mind dangling in empty evolutionary space."

[101] See Bechtel & Richardson (2010: xlii) for discussion, and the discussion of homology in §3.5.6 below.

Indeed, Darwin himself seems to have this sort of point in mind, when he says things like "there is no fundamental difference between man and the higher mammals in their mental faculties" on the basis that there must be "numberless gradations" in mental faculties to avoid large jumps in evolution,[102] or:

> "On the theory of natural selection, we can clearly understand the full meaning of that old canon in natural history, 'Natura non facit saltum' ['nature makes no jumps']. This canon, if we look only to the present inhabitants of the world, is not strictly correct, but if we include all those of past times, it must by my theory be strictly true".[103]

However, this last qualification about "the present inhabitants of the world" is crucial in our context. Even if it is right that there had to be a series of small developments rather than a big jump leading to cognition, this does not imply that there are no jumps at all when we compare species *that still survive today*. Just as we do not see species today which are rather like giraffes but with necks somewhere in between normal length and giraffe-length (the okapi seems to be the closest we have to that), there may be a gap between humans and chimpanzees in our cognitive traits which was filled by species which have now died out.

Furthermore, rapid evolution *is* possible, at least where rapid means the timescales since the last common ancestors of species like humans and our closest contemporary relatives. There are many important physiological differences between humans and chimpanzees in our teeth, posture and so forth, and lactose tolerance may well have evolved multiple times in just the last 10 000 years. The famous, striking differences between Galapagos finches arose in just the last

[102] Darwin (1889: 65f.)

[103] Darwin (1859: 226).

2.3m years. By comparison, our Last Common Ancestor with chimpanzees is thought to be 6-7 million years ago.[104]

Rapid evolution may be particularly likely in cognition. Here we have phenotypes which depend in rather complex ways on genotypes along with environmental influences, and small changes in the genotype may thereby have large, unpredictable effects on such phenotypes, just as we can get big changes in capacities due to relatively small changes in learning algorithms (as I emphasize in Chapters 4 and 5). Furthermore, many hypotheses have been defended on which cognition can advance rapidly due to some combination of setting up arms races, allowing for more sophisticated cooperation and sharing of knowledge, and rapidly changing the technological and environmental situation each generation is born into.[105]

However, this problem need not lead us to throw out this sort of reasoning altogether. And to avoid throwing away such reasoning, we need not just assume away evolutionary jumps. We can instead consider when rapidity is likely to be a particular problem, and use this to moderate the strength we treat this sort of evidence as having. The likelihood of divergence since the last common ancestor of two species will depend on two factors: the length of time since that last common ancestor, and speed of evolution. The first factor can generally be known reasonably well through the normal methods of phylogenetics and natural history. The second, speed, is more difficult to estimate as it pertains specifically to the trait in question. It will depend on the strength of the selection pressures on that trait and the frequency of the relevant mutations. Getting precise, reliable estimates of these values will be extremely difficult

[104] Sato et al (2001) (this example is from Brown 2019).

[105] E.g. Byrne & Whiten 1988; Dunbar & Shultz 2007; Stotz 2010; Csibra & Gergely 2011; Sterelny 2012; Heyes 2018.

without studying the evolutionary process directly, which would require already knowing the sorts of things we are trying to figure out (which species actually have the trait in question). But we can at least consider whether the factors likely to give rise to a high speed of evolution are likely to arise for our particular trait: is it an area where there are likely to be arms races, feedback effects etc.?

To use this kind of reasoning at all successfully we will need at least working possibilities of how our trait might have evolved: the sorts of intermediate stages it might have evolved from, and the selection pressures it will likely have been subject to. This may well involve a certain amount of speculation and should not be heavily relied on, at least at first. As we gain more direct evidence about which species have which mental traits, however, the projects of reconstructing how these evolved and who has which traits may be mutually reinforcing.

Furthermore, reconstructing possible routes of evolution of our trait—chains of evolutionary dependence consisting of traits and the intermediate stages which need to be in place before they can evolve (or multiple such possible precursors), the intermediate stages which need to be in place before those intermediate stages can evolve, and so on—provides additional opportunities to say things about the distribution of mental traits across species with more confidence than we would have just considering each species' mind based only on what is known about that species' individual brain and behaviour.

We can use such considerations to weaken the hypotheses that we try to support with cladistic parsimony.[106] Instead of aiming to show that things that are closely related to us are

[106] Sober 2015: 183 does consider something very much like chains of evolutionary dependence in the context of trying to infer an ancestor's state. He considers a model with an 'ordered n-state character' i.e. one where the character state will take a value that has to be within 1 of the value of my immediate ancestor: given this model, he shows that maximum likelihood will say that where we should expect the ancestor to be depends on how strong

more likely to have the *same* traits as us, we can aim to show that things more closely related to us are likely to have *similar* traits to us, where similarity is understood as *close to us along chains of evolutionary dependence*. So while we may get only very weak support for the hypothesis that monkeys represent order in the same way as great apes, we will get stronger support for the hypothesis that either the two groups represent order in the same way, *or* the two groups represent order using related systems, which both use essentially the same mechanism although the great apes' (say) is slightly more complicated in certain respects and this gives them the capacity to represent slightly longer sequences and to do slightly more complex computations over those sequences. These weaker hypotheses will apply with at least somewhat higher confidence than the stronger hypotheses, over a wider range of target species from our model (or known about) species, so we can have higher confidence for more similarity in very closely related species, but much lower confidence for species further away.

If we impose constraints of evolutionary dependence on the class of hypotheses cladistic parsimony is trying to decide between, we will get something like this result. If there is a chain of five mutations that we have to posit to get to a trait that we know humans have, the hypotheses which posit these changes all happening, but only once, will be more cladistically parsimonious. And given either another creature besides humans known to have reached a certain point in this chain, or given behaviour that could only be explained by positing one of the traits from a certain rung of our chain or above, or by positing some other trait that requires its own series of mutations, we will be able to use cladistic parsimony to motivate believing that some creatures are at a certain point in the same chain of evolutionary dependence as humans.

---

selection is (whereas cladistic parsimony doesn't really capture this as it predicts no change). But he does not develop the model further by considering how we might use such notions to weaken our possible principles into something more probable but still useful.

The notion of evolutionary dependence or necessary precursors is not completely novel and unused in evolutionary biology. Biologists often talk of 'pre-adaptation', 'evolutionary building blocks', etc. One particularly interesting example comes from the Long Term Evolution Experiment, an experiment where several colonies of *E. coli* taken from the same initial strain have been evolving in parallel, with every generation being studied genetically and stored, for tens of thousands of generations. Years into this experiment, one of these 12 populations underwent a particular mutation that allowed it to grow much more quickly: it developed the ability to metabolize both glucose and citrate - a rare ability (at least in the presence of oxygen) for *E. coli.*[107] None of the other populations developed this trait. Careful genetic work on the ancestors of this particular population revealed a series of mutations that had to be in place before a critical mutation gave the right ability. The lab in question called the mutations in this series *potentiating mutations*, but this is just synonymous with saying that the final trait *evolutionarily depended* on the mutations in this series.

Nonetheless, one might worry about using evolutionary dependence in this way. In the Long Term Evolution Experiment, the path of evolution of *E. coli* was directly observed, with each generation being genetically sequenced. In the case of reconstructing the evolution of the representation of time, we have much scantier evidence. Again, we should expect the situation here to be a back and forth between estimates of who has which traits today and how those traits might have arisen historically, with both projects informing and driving the other forward. We should not expect either to proceed without the other. But how might the project of reconstructing the history (let alone the pressures which give rise to this particular history but would result in similar histories counterfactually) proceed at all?

[107] See Losos 2017, Ch. 10.

A starting point will be an account of the trait of interest from which we can extrapolate possible simpler versions, looking at the trait's components and how they might be developed individually or in concert. Sometimes the key features might include abstract computational structures, in which case we need to remain alive to the possibility of those computational structures developing in other functional contexts and being exapted. Evolution can proceed by fine-tuning existing mechanisms, building through combining them, replicating multiple versions of the same mechanism then allowing these to diverge through specialisation, but also repurposing them. We will also need to take into account the possibility of domain specific mechanisms becoming more flexible and domain-general in function, as was suggested in Chapter 2 for temporal representation. Evidence for the paths evolution might have taken or must take will include observed convergent evolution, but also similar paths taken in multiple simulations of evolution, similar paths taken in development, and evidence for similarities in the underlying components. Most of these sources of evidence will be much stronger when the status of the trait is known for multiple animals, preferably including several animals per lineage for multiple lineages.

One interesting point to note about the strategy of thinking about chains of evolutionary dependence to assign greater confidence to weaker hypotheses, is the way the justification for this strategy works, a form of conservatism is preserved. We should not put the most probability on animals having the *same* mental traits as the known-about or model species (often humans), at least if the model species is the one with the most complex known version of the trait. Rather, we should have a slight bias towards their being simpler. This is because we need to account for the fact that there will be some creatures without any version of these traits, and we have reason to suspect different creatures reaching different points along that scale.

Overall, then, use of phylogenetic considerations will be most effective the more we know not just about the phylogenetic relationships between target and model species (especially how recent their most recent common ancestor was), but when we are entitled to more confidence about the course of evolution of the particular traits under consideration. In particular, it will be helpful if we have evidence for slow evolution of that trait, independent evidence for homology in relevant brain areas, and evidence for the likely course evolution took to get to the traits current form. Knowing about the trait in detail both at a neural and computational level, and knowing about its status in multiple species, will help with this.

3.5.2 Functional Similarity

As a crude slogan, the phylogenetic approach predicts that animals which are more closely related are more likely to have similar traits. Close ancestry is not the only evolutionary determinant of similarity in traits, however. There are many examples of convergent evolution—where some trait has evolved independently in different lineages. As Darwin put it,

> "in nearly the same way as two men have sometimes independently hit on the very same invention, so natural selection, working for the good of each being and taking advantage of analogous variations, has sometimes modified in very nearly the same manner two parts in two organic beings, which owe but little of their structure in common to inheritance from the same ancestor".[108]

This is not just a quirk that happens every now and then. There are many examples of convergent evolution.[109] The *Origin of Species* documents the repeated evolution of a certain body shape in cetaceans and fish. Darwin at one point thought that different groups of Galapagos finches were closely related to English finches, grosbeaks, blackbirds and wrens respectively due to similarities in their traits with the English birds and dissimilarities with each

[108] Darwin (1859: 193).
[109] See Losos 2017, who catalogues many more examples in detail and puts them into theoretical context.

other, as opposed to being related to each other, until John Gould discovered all the finches were in fact from common stock.[110] There are many other cases where extremely similar-looking and -behaving creatures have turned out to be extremely distantly related, such as New World and Old World porcupines, whose last common ancestor did not have their most distinctive features of quills/bristles.[111] Sometimes similarities and differences in phenotypic traits are so striking that genetic analysis has completely overturned opinions about likely phylogenetic relationships. For example, in 2013 it was discovered that different populations of beaked sea-snake, long thought to be one species widely distributed geographically, are in fact are very genetically different - and each population was much closer genetically to its local rival species of sea snakes, who are behaviourally and anatomically very different.[112] The Long Term Evolution Experiment mentioned above, meanwhile, has found many adaptations that have arisen in all its populations, often in the same order and with similar timing.

Convergent evolution has almost certainly happened in the case of cognition too. Notoriously, corvids are much better at a range of behavioural tests of various aspects of cognition than not only most birds but most mammals too. Corvids seem to have developed similar cognitive systems to primates convergently, even at the neural level.[113] Meanwhile, Octopuses and cuttlefish seem to have developed cognition and brains far more sophisticated than anything at all closely related to them, completely independently of mammals and birds.[114]

---

[110] Losos (2017: 11f.).

[111] See Losos (2017: 10f.) for illustrations emphasizing the similarity of these species.

[112] Ukuwela et al 2013.

[113] Van Horik et al 2012; Clayton 2015; Güntürkün & Bugnyar 2016; Olkowicz et al 2016.

[114] Godfrey-Smith 2016

We should not overstate the extent of convergent evolution. Evolution is not entirely determined to repeat itself down to the last detail, and there is debate about how far—and how frequently—it does so.[115] But there is enough convergent evolution around to have two effects.

First, it should sometimes worry us about how reliable a guide any form of cladistic parsimony will be if it implies that closeness in phylogenetic relations tracks similarity in traits. Particularly worrying are the cases of adaptive radiation, where selection pressures seem to override phylogenetic similarity entirely. These are cases, like in the Galapagos finches and the sea snakes above, where a population of closely phylogenetically related individuals reliably splits into the same quite phenotypically different populations, adapted to fulfil slightly different niches, sometimes to the extent that the phenotypically similar but phylogenetically distinct populations look like phylogenetically unified species.

But secondly, from the perspective of looking for considerations that can inform our attribution of mental traits to animals, these sorts of case can be seen as presenting opportunities, rather than just as undermining the use of evolutionary dependence in certain cases. Our fundamental problem is not to use just cladistic parsimony and evolutionary dependence to shift our credences about what mental traits animals are likely to have: it is to use whatever information we have available to set those credences. Cases of convergent evolution can be thought of as giving another set of considerations which can be used to shift our credences besides detailed behavioural and neural observations: considerations about ecological niche and lifestyle. Because we are working within a framework of graded credences rather than an

[115] For example, Gould 1989 emphasizes cases where there is no convergence and the role of chance in natural selection; Conway Morris 2003 emphasizes cases where there is convergence; Losos 2017 tries to chart a middle ground.

accept/abstain/reject framework, we are in a position to combine these different sorts of considerations.

Even for distantly related animals, we can assign higher probabilities to distributions of traits postulating certain kinds of similarity in cognitive mechanisms between species facing similar problems to be solved. As with phylogenetic considerations, versions of this sort of reasoning are widely appealed to, at least implicitly, by many scientists and philosophers. Here is a typical case:

> "The honey bee particularly is held up as an insect with cognitive capacities that rival those of many mammals. Without consideration of the underlying mechanisms, this may seem like no more than a curiosity. The systems that underlie these abilities were shaped by evolutionary pressures similar to those that shaped the mammalian midbrain. The insect brain does a similar sort of modeling, for the same reasons, in a similar way. That is strong evidence that the insect brain has the capacity to support subjective experience."[116]

Barron and Klein are here arguing that bees have subjective experience on the grounds that they have brain structures which are *functionally* similar to the brain structures they (somewhat idiosyncratically) think support subjective experience in mammals (the midbrain). And this functional similarity is explicitly understood at least partly in terms of selection pressures.

What determines the strength of evidence provided by considerations of this sort? Convergence is more likely given similarities in the environment of the animal and the relevant aspects of the role the animal plays in that environment, in addition to relevant similarities in relevant body plans, mental architecture and sensory and motor capabilities, which will determine the options available to natural selection in terms of possible solutions to an

[116] Barron & Klein (2016: 4905). Note that if this passage is read as extrapolating from the honey bee brain to the insect brain generally (rather than using 'the insect brain' to mean *at least one insect brain*), we also have an appeal to phylogenetic considerations here.

environmental problem. These similarities will only be directly relevant to systems that might have the uses in question: Both squirrels and scrub jays cache food, but they differ in flying, climbing trees etc. and so have different relations to predators, the environment around them. So they may have similarly structured memory for locations etc., but very different motor and action-planning systems.

Evidence for such strong selection pressures will need to start with direct evidence that the species in question do share many of the relevant similarities, based on ethological observation, anatomy, existing literature on these species' cognition and on the trait in question. But it will also need to be established that these similarities are not just flukes, and are due to similar selection pressures (whether because these similarities are selected for by the same pressures directly, or are side-effects of similar adaptations to similar selection pressures). Evidence for adaptations can include surprising, unique predictions of optimality-based models being borne out, but direct evidence for selection will be more compelling: this will include evidence that variations in the allegedly selected trait have a genetic basis; evidence that the trait actually does improve reproductive success, and a mechanistic account of how it does so; evidence that the environment in which selection is supposed to have occurred had the selection pressure; and ideally experimental manipulation of the environment/trait shows differences in reproductive success.[117] Genetic evidence can also provide evidence of comparable traits following similar paths. One important example of this is involved in temporal coordination: one variant of the protein cryptochrome entrains the central circadian clock to light in *Drosophila*, and while the homologue of *that* variant of cryptochrome in mice does not, but a

[117] Cf. Lloyd (2005: 4ff.) for discussion.

different variant of cryptochrome does.[118] In the case of evidence for convergent evolution in cognition in corvids and primates, we have several pieces of evidence of this kind, such as surprising similarities in brain structure and success and failure on similar tasks.[119] But our evidence of how these cognitive capacities evolved largely rests on extrapolating from what we know about related contemporary species.

Like for phylogenetic considerations, the more we know about the evolutionary processes we are implicitly appealing to in these cross-species extrapolations, the stronger the evidence that adaptation-based extrapolations will provide. And like in the case of phylogenetic considerations, such knowledge will need to be achieved in the case of cognition in a back-and-forth dialogue with direct evidence for which contemporary species have which mental traits, for as many species as possible, as well as models of how such traits could have arisen. Furthermore, like in the case of phylogenetic considerations, we will be on stronger ground with weaker hypotheses: in this case hypotheses which posit only disjunctions of several of the possible solutions to adaptive problems rather than focusing on just one (at least, where several are in fact available). But unlike in the case of phylogenetic considerations, stronger selection pressures and resulting rapid evolution will strengthen, rather than weaken, adaptation-based extrapolations, as they mean we can be more confident that both animals' lineages have in fact responded to their shared selection pressures in similar ways.

Having discussed the conditions under which both kinds of evolutionary extrapolation between species' minds are stronger or weaker, I will now briefly survey some general upshots.

---

[118] See Bechtel & Richardson (2010: xlii) for discussion of this case in the context of model organisms.

[119] Güntürkün & Bugnyar 2016.

3.5.3 Evolutionary Inferences Between Different Species and Humans

In both sorts of evolution-based extrapolation between species, the extrapolation will be weak without evidence for how the traits evolved and how they are likely to have evolved across multiple lineages. One upshot of this is that inferences from just one species will be much weaker than inferences from multiple species. And given the other considerations that determine the strength of both kinds of evolution-based cross-species extrapolation, inferences from humans in particular are likely to be weak. Humans have faced strong, and rather different, selection pressures, to any other surviving lineage, for considerable periods of time. Brown 2019 argues at length that human minds are likely to be unlike other animals' minds in many respects. She emphasizes that there is some evidence that we were selected to deal with variability and patchy resources (often with difficult to process foods), whereas chimpanzee environments reliably feature the same easy to digest foods; because humans have much longer childhoods relative to our overall lifespan during which we learn extensively (which is known to correlate with general intelligence, social complexity, and foraging complexity), and because we are tolerant of their conspecifics (including of juveniles) and display a high level of sociality compared to other primates. All this makes for strong selection pressures and hence rapid evolution, undermining phylogenetic-based extrapolation to or from human minds, and an unusual set of selection pressures relative to other species, undermining adaptation-based extrapolation to or from human mind.

However, we should put a few qualifications on Brown's skepticism here. One point is that she herself is making claims about the evolutionary history of the human mind. And while she provides considerably stronger evidence for her claims than is often provided in this context,

views on which there is a great deal more continuity between humans and other animals than she suggests remain live options.

Secondly, the point that we should only focus on *relevant* similarities in selection pressures will be important in some cases. For example, we know that the visual systems of humans and monkeys are similar in many ways, and that extrapolation between the two is often justified. This is partly because the selective pressures on our visual system probably do not depend heavily on humans' need for flexibility, social cognition, and our ability to spend a lot of time learning. Rather they likely relate to quickly and accurately forming representations of features of the immediate environment like objects' shapes, locations and colours. This point will be important in the case of temporal cognition: while evolutionary considerations will not provide strong evidence for extrapolating from complex forms of human temporal cognition to other animals, temporal representation in perception may well generalize better.

Furthermore, her argument apply just to humans, not to evolutionary considerations generally: where we are able to gather information about temporal cognition in a range of species, and especially species who have either close living relatives or corresponding species subject to very similar, strong selection pressures which we can test, evolutionary considerations will be more useful. After all, the phylogenetic-based considerations need not just be applied at the level of species: at the extreme, exactly the same sorts of reasoning can be used to extrapolate between different lineages within species, and in this case they look much stronger (though by no means infallible: pure chihuahuas are in some respects rather different to pure great danes—precisely because of strong (artificial) selection pressures leading to rapid changes since their last common ancestor). This sort of reasoning, for example, is presumably key to any extrapolations from a small number of animals studied in a lab to the capacities their species as a whole.

Indeed, this point suggests another advantage of the general credence-based framework. It does not require choosing the appropriate level of relationship between two animals for us to extrapolate between them. Whereas some principles, like Birch's precautionary principles, require us to choose e.g. whether evidence about animals in the same order is relevant as opposed to just animals in the same phylum/genus/family/etc., our framework can allow that closer relationships will generally provide stronger evidence.

3.5.4 General Limitations on Evolutionary Reasoning

Much of the foregoing discussion in §3 has suggested that in many situations evolutionary reasoning, despite being intuitively appealing and providing good evidence in some situations, will turn out to provide only weak evidence. This point corroborates the proposal made in chapter 1 to generally focus more on issues of cognitive significance, which are more easily testable and often more fundamental than evolutionary claims. Especially in advance of direct evidence about the path evolution has taken and the selection pressures involved, any use of evolutionary claims will often be based on large doses of speculation—about selection pressures, the precursors to the trait of interest, and factors affecting speed like rate of mutation. As I have emphasized, this speculation is not necessarily bad: using evolutionary speculation to constrain speculation about who has which traits now, and vice versa, can help generate novel and helpful hypotheses which can be tested, pushing us towards greater evidence for an integrated picture of both evolution and the current distribution of traits. But it does need to be given only the weight it deserves.

We need to be particularly wary of overconfidence in adaptationist hypotheses in the case of politically and morally charged traits—which, as we said above, animal mental traits often

are.[120] We should be especially careful in these contexts to remain open to alternative possibilities to our favoured hypotheses about adaptation: possibilities including other selective pressures, and the trait in question arising as a byproduct of selection for other traits.[121]

3.5.5 What are Traits?

One final issue with evolutionary reasoning in our context is that we need to be careful about the kind of traits we are reasoning about. As noted in Chapter 1, I have been using the term 'trait' as a catch-all to cover systems, processes, states and so forth. In an evolutionary context, calling something a trait can often imply a heavier duty notion. In particular, for evolution by natural selection to operate on a trait, we need that trait to be inherited. This inheritance may not be based directly in genes: we can have extended notions of inheritance like being likely to arise during development given a reliably reproduced developmental niche.[122] But there must be inheritance in some sense. This means that these evolution-based inferences will be much weaker for 'traits' in my broad sense when they depend on highly contingent learning rather than inheritance. It may be that many of our ordinary psychological kinds, including sophisticated kinds of temporal representation, will turn out to be like this, in which case evolutionary reasoning will be weaker where they are concerned (this issue will arise in Chapter 4).

Related to this issue of individuating traits is the issue of how far we can talk about *homologies* in purely cognitive cases, as I did above. The main point of the notion of homology is to allow a robust sense in which two organisms can have the very same feature, despite that

[120] Sahlins 1976; Lloyd 2005; Buller 2005.

[121] Gould & Lewontin 1979.

[122] Stotz 2010.

feature taking on a different form and playing a different functional role. With physical body parts, this notion is often illuminating and allows for various kinds of extrapolation between species, as when we can trace the bones of a bat's wing to the bones of a hand. And it can be studied through analyzing the topology of these structures, looking at their embryological development from the same foetal structures, and looking at their genetics.

In the case of anatomically defined brain structures, all this clearly applies: we can look for homologous structures to the hippocampus in different creatures in exactly the same sense that we can look for homologous structures to the hand. We can also study homologues of traits which involve both psychological and bodily elements, like emotions, or orgasm.[123] In the case of traits defined purely cognitively, in a way which allows for multiple realizability in the brain, things are more complex. We can make sense of the same abstractly defined cognitive architecture or computation being used for different purposes in different species, for example if it is used in different domains (social cognition in one species, foraging in another), and we can make sense of looking for corresponding sub-computations in one of these systems because we have found them contributing to the overall computation in the other. And we can make sense of their following similar developmental paths up to a point before diverging. But we cannot look for genes unless we tie the computation to a particular brain area, and development of the same computational architecture might well arise through very different paths in a way that development of the same bodily structure would not. Two alternatives are available here: tying cognitive traits more closely to their neural implementation and sacrificing the possibility of multiple realizability to allow for more robust evolutionary reasoning, or maintaining a

[123] For emotions, see Griffiths 1997; Clark 2010. For orgasm, see Lloyd 2005 and Pavličev & Wagner 2016.

commitment to multiple realizability at the expense of robust evolutionary reasoning. Chapter 5 argues for an approach to individuating mental traits for cross-species comparisons which implies the latter of these two options.

3.6 Upshots for Animal Representations of Time

The main lessons of our discussion of how we should assess different kinds of evidence for mental traits in animals generally are as follows. We should not think in terms of conditions for full belief or disbelief: in the case of animal minds, we will rarely have enough evidence for either. Being rigorous scientists and philosophers does not mean believing that animals lack mental traits when we lack decisive evidence for their doing so—or even suspending belief (depending on what this latter position comes to). Instead, it means carefully thinking through what evidence we do have and how it should affect our credences in the range of possibilities which could be at play: most notably, that the target animal has the trait in question, that it has a version of that trait/a trait from a broader class to which our trait belongs, or a precursor or modification of our trait; that it has some of the components of our trait, including non-representational or non-psychological components, and that it completely lacks the trait. The most relevant kind of evidence for possession of the trait will be direct evidence that this animal has all of the core features of that trait. Next will come behaviours which can be explained by the relevant hypothesis (or are predicted by it with high probability) whilst not being easily explained (or being predicted with low probability) by rival hypotheses. Next will come more circumstantial kinds of evidence, relating to other features which are correlated with the trait in question but not essential to it, and evolutionary considerations. We need to be particularly careful with evolutionary considerations, but there are some conditions under which they will provide non-trivial amounts of evidence. When we have reason to believe strong (weak) selection pressures are at play, extrapolations based on animals with relevantly similar selection

pressures (close phylogenetic relationships) to the target animal will be more important; but it is often possible and desirable to find independent evidence about both the selection pressures that have shaped the trait in question and the different forms its precursors have likely taken, both of which can improve such inferences. When we have evidence for the state of the trait in question in multiple different animals, we will be better able to infer to further animals. And often we will only be able to support weak hypotheses with evolutionary considerations, rather than very strong hypotheses specifying many details of the traits of interest.

What does all this imply for the question of which animals have temporal representation?

First, it undercuts arguments against temporal representation which move from providing an anti-representational alternative explanation of animal behaviour to positively denying that animals have temporal representation. The anti-representational alternatives are hypotheses alongside the representational ones, and it is not obvious which we should accord more credence to, let alone that we should accord a high credence to one and a low credence to the other.

Arguments that we should adopt the anti-representational hypotheses due to 'parsimony' also look suspect.[124] Our discussion of cladistic parsimony showed both that we should not always favour parsimonious hypotheses, and that denying that animals have various kinds of representation is often not parsimonious in the most relevant sense anyway, as it typically commits one to assuming that there are big jumps in evolution.

Chapter 2 argued that the core of temporal representation should be seen as not just sensitivity to time and resulting successful coordination, but flexible use of such sensitivity for computation. This sort of flexible use is strongly suggested by the evidence that scrub jays

[124] For example, Hoerl 2008—especially the arguments he attributes to Bennett and Smith; Hoerl & McCormack 2019.

combine such sensitivity with other information to produce behaviour that is suited to a wide range of different scenarios, and by evidence that monkeys can remember, extend, and reverse sequences, and even respond to the ordinal positions of individual items from those sequences. In both cases, there are limitations to the evidence such as the domain specificity of scrub jay when-memories. But positing at least representations which are ambiguous between representing duration and something more domain-specific like freshness seems well-supported.

In many other animals, the right kind of flexible use of temporal sensitivity and coordination has not been directly shown, or has not been directly shown extensively. This does not mean that we should believe they do not have temporal representation. It shows what is at stake when we consider whether other evidence can be wielded to support belief that they nonetheless do have temporal representations: if they do, then they are capable of such flexibility, but have not shown it in experimental conditions yet.

Nearly all animals show temporal coordination of many kinds. Many animals behave in ways which would only be predicted by temporal representation or fairly elaborate anti-temporal-representational sensitivities, such as conditioning to respond to durations in multiple sense modalities and domains, and to produce durations in different sorts of actions.[125] In these cases, it is reasonable to be pretty confident that there is temporal representation, but nowhere near 100%: there should be at least some credence accorded to the elaborate anti-temporal-representational hypotheses. The details will depend on the case, of course, but evolutionary considerations may suggest that the anti-representational hypotheses should receive a sizeable chunk of credence in many cases like this.

[125] Gallistel 1990, Chs. 7-9.

A very reasonable story about the evolution of temporal representation, given my understanding of it, is that it is built on mere sensitivity, which is then harnessed in ambiguous ways for representations indeterminate between temporal and something else, before their use becomes more specialized and tailored to temporal representation specifically. If this is what happened, then we should expect many animals to be left with temporal sensitivities which have not become temporal representational. However, the warnings above about relying on evolutionary considerations apply here: I have not modelled the strength of selection pressures for temporal sensitivity and temporal representation, let alone found direct evidence to support estimates of the speed of evolution, the exact selection pressures involved etc. So this sort of consideration should only operate in any strong way when there is very little else shaping our distribution of credences between the representational and anti-representational hypotheses for different animals. Furthermore, one could come up with other plausible stories about the evolution of temporal representation, such as its being exapted from other systems such as systems for number representation; and depending on how such stories were fleshed out, they could either result in higher or lower credences for the temporal representational vs. anti-temporal-representational hypotheses.

On the other hand, we do have evidence of complex temporal sensitivity and conditioning to durations (whether representational or not) in many different animals, from different lineages. So here phylogenetic considerations are fairly secure in supporting the view that this sort of possibly representational sensitivity is extremely widespread, at least in mammals and birds. Furthermore, both scrub jays and rhesus monkeys tested have much closer relatives (other corvids and other old-world monkeys), with much more similar minds, than humans do, so using phylogenetic considerations to be relatively confident that these broader groups of animals have

temporal representation is also reasonable. And the fact that we find good evidence of temporal representation in both primates and corvids, given that these seem to be convergent in many other ways, means that they should mutually reinforce our confidence in each having temporal representation.

Does this mean we should also extend the findings in scrub jays to assume that there will be similar abilities in other food caching animals, such as squirrels, chickadees, nuthatches etc., given that they may face similar selection pressures? There is a serious problem with this move: even if the other food cachers have a selection pressure in favour of sensitivity to how long ago food was cached, it is not clear that they face the same further selection pressures as corvids for integrating that information with a wide range of other sorts of information (especially social information, given that jays have sophisticated social cognition).

A final upshot is that these ways of thinking about evidence suggest new ways of testing for temporal representation in different species. Besides the ways suggested by the previous chapter—testing for different kinds of flexibility—the considerations here suggest modelling both selection pressures facing these animals and possible evolutionary paths to their contemporary kinds of temporal flexibility, and testing any ideas the modelling suggests. For example, what is the optimal way to integrate different kinds of information about time with other information about food caches and with other sorts of information, for different animals? What sorts of temporal sensitivity could be present in ancestors of scrub jays or monkeys that could be harnessed by evolution for their current systems, and how would such information go?

3.7 Conclusion

The overall takeaway from all this seems to be that we should be confident that at least very sophisticated forms of temporal *coordination* are extremely widespread in the animal kingdom, and that at least some animals have full temporal representation, whilst treating the

idea that temporal representation is extremely widespread as a live option. So we have an answer to the constitutive question about temporal representation, and a tentative answer to the epistemic question. But what about the cognitive significance question?

The kind of temporal representation discussed so far does not seem to be overwhelmingly cognitively significant. It is no doubt adaptive and impressive that scrub jays can represent their caches in the way that they do, but this in itself does not seem to mark out the scrub jay's mind as different in kind to other animals' minds. It does not, for example, seem (yet) like a more important feat, one which opens up a wider range of further capacities, than representing the kind of food in the cache or its nutritional value. What is going on here? Why did the claims about the importance of temporal cognition surveyed in Chapter 1 often seem like they were onto something, if temporal representation does not seem to be any more significant than other kinds of representation? The answer will lie in more detailed discussion of what can be done with temporal representation, and with specific kinds of temporal representation. That will be the topic of the next chapter.

Chapter 4: Temporal Frameworks

4.1 Introduction

The last two chapters discussed the questions of what it takes to represent time at all, and whether non-human animals do so. We have seen that while demonstrating temporal representation (or any specific hypothesis about representation) in the sense of ruling out every other potential hypothesis is difficult, it is reasonable to put a serious chunk of credence in the hypothesis that certain forms of temporal representation, such as (potentially ambiguous) representation of short durations in perception, are widespread. But now we can return to the issue raised in Chapter 1: how significant is such representation? One lesson of Chapter 2 was that many of the jobs any form of temporal representation does *could* be done through the right kinds of architectural constraints. So, for example, one might argue that representation of duration is highly significant because it allows for the estimation of rates of change, and many variables' rates of change are important. But we also saw in Chapter 2 that rates of change can be estimated and responded to without representing duration or any other temporal feature at all.

Furthermore, it is not at all clear that perceptual representation of short durations would impress people like Avicenna, Burns, Nietzsche and the rest who claimed that animals are *in some sense* stuck in the present. It would be reasonable to think that perceptual representation of short durations is like perceptual representation of any other magnitude—weight, light intensity, or nutritional value—more abstract perhaps, and so more widely applicable, but not a serious candidate for setting apart kinds of mind.

The reason it is at all plausible to make the sorts of claims about the cognitive significance of temporal representation canvassed in Chapter 1, is that those claims can be construed as being about special kinds of temporal representation, not temporal representation

simpliciter. A binary distinction, such as Hoerl & McCormack 2019's division between Temporal Reasoning Systems and Temporal Updating Systems, is too simple to capture the important variety in potential kinds of temporal representation systems.

Most participants in the literature recognise that there are important distinctions between different kinds of temporal representation in this context, whether this is a distinction between having an 'extended present' and representing a more distant past and future,[126] or distinguishing between genuinely tensed thought and representation of current properties with temporal aspects, like 'completed'.[127] Such distinctions can be easily multiplied further. This raises a question: which types of temporal representation are cognitively significant, and why?[128]

I will give reasons to suspect that three distinct sorts of temporal representation might be particularly significant, for quite different reasons: representations of frameworks of times; narrative representations of structures of specific events; and representations of essentially dynamical kinds of entity that include information about the typical temporal evolution of those entities. All of these could be significant for relatively simpler minds, because they help cope with limits of reinforcement learning, especially in its ability to deal with non-Markovian dynamics (to be explained below), and to allow for rationality which goes beyond maximizing the discounted sum of expected utility.

[126] Nelson & Fivush (2004: 500).
[127] Hoerl & McCormack 2011.
[128] Or, to be more precise, which if any types of temporal representation are both (1) plausibly available to technologically and socially unsophisticated humans and (2) cognitively significant, and why? If we want to understand the diversity in animal minds, including the distinction between humans and other species, it will not be helpful to focus on kinds of temporal understanding that rely on complex scientific theories and access to advanced technology. The understanding one gets of time in the context of the theory of general relativity may well be significantly different to earlier ways of understanding of time. But few humans have this kind of understanding, and it is obviously not required for any of the important cognitive achievements of humans or other animals before Einstein.

I will proceed by first (in §4.2) laying out a series of seemingly orthogonal distinctions which potentially hold between different types of temporal representation. I will then articulate promising accounts of why sophisticated temporal representation might be cognitively significant, and in each case consider what types of temporal representation would be required to play the envisaged role: which of §4.2's distinctions are actually important.

In §4.3, I consider the idea that temporal representation could serve as a common currency to serve domain-general intelligence, but ultimately reject the idea that temporal representation is necessary for this role. §4.4 introduces an important limitation of standard reinforcement learning algorithms like those considered in Chapter 2: they assume that the world has the Markov Property. §4.5 shows that one way of partially overcoming these limitations is to represent essentially dynamical entities. §4.6 shows why the representation of narratives might also help a limited mind partially grasp complex, non-Markovian dynamics, in a complementary way: it can help a limited mind grasp different aspects of complexity to the representation of essentially dynamical entities. But I also show that narrative understanding can help overcome a different limitation of reinforcement learning: alleged limitations to RL's conception of diachronic rationality. One upshot of this discussion, combined with the earlier discussion about the role of narratives in understanding, will be a way of rethinking preferences for narratively structured lives, along with the ethical upshots some philosophers have drawn from these. Finally, §4.7 shows just how little we know about what kinds of sophistication in temporal representation different animals have, and suggests ways of developing the study of these issues in light of the distinctions and suggestions in this chapter.

4.2 Distinctions Between Temporal Representation Systems

There are many distinctions one can draw between temporal representations systems. I will begin with some distinctions which will help our discussion for clarity's sake and have

occupied much discussion in the literature, but will not turn out to be crucial to cognitive significance—distinctions between kinds of system; different temporal relations which might be represented; different temporal relata (including indexical or non-indexical relata and generic or particular relata); and differences in format. With these distinctions in hand, we will be in a position to appreciate more fully the further distinctions which turn out to be more cognitively significant, between different kinds of complex structures of events or times (especially framework and narrative structures), and more atomized representations.

Firstly, we can make general distinctions between kinds of system in the case of temporal processing. For example, if we think that there is a robust distinction between perception and cognition, temporal representation might appear in perception even in an animal which does not have cognition, or there might be an animal which can represent time cognitively but relies on mere temporal updating for perceptual systems.[129] This is not the only cut we can make on the basis of kinds of system. Hoerl and McCormack 2019 tie their distinction between temporal reasoning and temporal updating to more general views about there being two kinds of (non-perceptual) cognition: one deliberative and slow, and the other fast, less controlled and more prone to error.[130] Or we might distinguish between systems using temporal *concepts* and those using more primitive kinds of temporal representation, particularly if we think that durations are represented in an analogue format, and that analogue representations cannot be

[129] There is considerable debate over how robust the distinction between perception and cognition is, and how to draw it. See e.g. Fodor 1983; Pylyshyn 1999; Camp 2009; Block 2014; Burge 2014; Lupyan & Clark 2015; Echeverri 2016; Firestone & Scholl 2016; Drayson 2017; Siegel 2017; Beck 2018; Phillips 2019; Montague (forthcoming); and Quilty-Dunn (forthcoming). Much of the philosophical literature on time in the mind concerns the phenomenal character of perceptual experience, rather than focussing on its role in cognition more broadly (e.g. much of Montemayor 2013; Merino-Rajme 2014; Phillips 2014 and the references therein).

[130] There are many accounts of 'dual systems' of cognition along these lines, including Sloman 1996; Gendler 2008; Frankish 2010; and Kahneman 2011.

conceptual.[131] Each of these distinctions is controversial, but we need not take a stand on them here: the proposals below assume only that temporal representations are available to certain sorts of computation. On some views, the computations discussed below might imply that the representations in question must be cognitive rather than perceptual, or system 1 rather than system 2, or conceptual rather than non-conceptual; but exactly how we draw these distinctions will not be important to the proposals themselves.

The most common distinction in the literature between systems using temporal representation is on the basis of the kinds of temporal relations represented.[132] Typically, three kinds of relation are distinguished: the temporal order of two events (the before/after relation); the duration or temporal distance between two events; and the phase of an event relative to a regularly repeating cycle. A rabbit might be represented as leaving its warren *after* the birds started singing. The rabbit leaving its warren and the onset of singing might be represented as 30 minutes apart. Or the rabbit might be represented as leaving its warren at 06:00 each day, as part of a cycle which also features the birds singing at 05:30 each day. There may well be systems which are capable of some of these sorts of representation but not others, even though there are a priori relationships between these relationships such that a full understanding of time integrates all of these different sorts of representation smoothly.

A further sort of temporal representation which may be possible without being capable of representing any of these other relations is what Hoerl and McCormack call applying 'aspectual

[131] Beck 2012. For debate over how to draw the conceptual/non-conceptual distinction, and whether there is non-conceptual content, see Evans 1982; Peacocke 1983, 1992, 2001; Crane 1992; McDowell 1994; Stalnaker 1998; Bermudez 2007; Heck 2007; Roskies 2008; Camp 2009; Schellenberg 2013; and Gauker 2017.

[132] For example, versions of these distinctions are emphasized in the discussions of temporal representation in Gallistel 1990; Friedman 1993; Campbell 1994, 1997, 2006; Roberts 2002; Burge 2010a; Montemayor 2013; and Hoerl and McCormack 2011.

notions': notions like 'ongoing', 'completed', and 'yet to come'. Young children and some animals may treat these as properties which they automatically apply through temporal updating in certain circumstances, without understanding how they can be cashed out in terms of events having happened or being due to happen.

We can further distinguish, within systems capable of representing some durations or some cycle-phases, between systems that have different ranges (length of durations or cycles) and levels of precision (how finely they subdivide durations or cycles). Human perceptual systems and animal and infant cognitive systems could well feature multiple distinct subsystems specialized for different scales, with quite different mechanisms dedicated to precise measurement of very short durations and others dedicated to less precise measurement of much longer durations.[133]

Besides the sorts of relations different systems can represent, we can also distinguish between systems on the basis of what they can represent those relations as being between—or how they represent those relata. Most importantly, they could represent these relations as holding between an *indexically* picked out event or time and some other event or time, or they could be capable of entirely *non-indexical* temporal representations. And they could represent relations holding between *particular* times or events, or they could represent more *generic* patterns which hold between all events satisfying certain descriptions. Of these distinctions, the generic/particular distinction will turn out to be far more important elsewhere in this dissertation, so it is worth expounding in greater detail.

We can represent particular events, such as the rabbits' going outside at 06:00 on Tuesday June 14th 2020. We can also represent generic, repeatable event-types, such as the

[133] Viera (2019: 33ff.) summarizes evidence for this claim.

rabbits' going outside, which may occur at 06:00 every day, or on many occasions, or typically. This is not a distinction which arises only in the case of repeating cycles: we can also represent that the rabbits always or typically go outside 30 minutes after the birds start singing.[134] Representing generic events is more immediately useful than representing particular events, or at least particular past events. When it comes to decision-making, it is useful to have access to general patterns that will predict the effects of different actions, and it is not obvious what good it would be to represent events which have already happened and cannot be changed.[135] Furthermore, it is easier to create neural networks that learn general patterns through repeated exposure to those patterns, than networks which learn about particular events (which in practice requires learning from a single exposure).

There is evidence that children learn *scripts*—generic sequences of events as they typically happen—around 2 years old, long before they are able to remember particular events (or unique sequences).[136] They can learn that bath-time follows the routine of running the water, getting undressed, getting in, washing, playing with a rubber duck, getting out, getting toweled down, and getting dressed again, long before they can recount an unusual event that happened at bath time. These scripts allow young children to retrieve fragments of specific past experiences

[134] This point is obscured in some presentations (e.g. Campbell 1994: 38), which deal with the two issues of repeatability: cycles vs. linear relations, and particular vs. general, together.

[135] See §§5.4-5 below.

[136] The notion of a script was developed by Schank & Abelson 1977 and adapted to developmental psychology by Nelson 1978, 1986, Nelson & Fivush 2004, and also discussed in detail by McCormack & Hoerl 1999, 2017, Hoerl & McCormack 2011, Campbell 2006. There is some discrepancy between these different authors over what these scripts involve. Hoerl & McCormack (2011: 447) interpret Nelson 1999 as saying that learning a script is not about coming to *represent* a sequence, but rather learning to, in the right order, think of those events; while Campbell 2006 not only thinks that sequences are represented, but also that they are represented in a way which provides a local framework which can be applied to the particular current situation in a way which allows for slotting in unique events.

as early as 2-3 years old, despite doing so rather differently to adults. For example, when questioned about a camping holiday, they will report events like eating dinner then going to bed then waking up, rather than unique features of the holiday.

We have already touched on a further source of distinctions about temporal representation: format. This distinction relates to the way that temporal features are represented, and specifically to ways in which features of the vehicles of a representation have an impact on its potential uses. One much-discussed cluster of distinctions here relates to differences between image-like and sentence-like representations, and, to some extent hybrid or in-between forms of representation like maps and graphs.[137] These distinctions have been recently used both in the context of distinguishing between perception and cognition, and making sense of cognition in non-linguistic animals.[138]

For our purposes, the most important potential difference between images and sentences is that some kinds of non-linguistic formats represent a large number of items and several of their relations at once, in a single surveyable structure. Contrast the efficiency in representing spatial relations (and, as a result, the kinds of operations which are easy to perform), on a *map* of Wales, on the one hand, with the lengthy, cumbersome *list of sentences* which would be required to specify the same information (information like 'Cardiff, Swansea, and Newport are all on the South coast, with Swansea the furthest West and Newport the furthest East of the three. The Brecon Beacons stretch from West of an imaginary line extended North from Swansea, to East of an imaginary line extended North from Cardiff…'). It will be much easier to plan a route or

[137] Goodman 1968; Haugeland 1981; Block 1983; Peacocke 1986; Lopes 1996; Wollheim 1998; Hopkins 1998; Kitcher & Varzi 2000; Kosslyn et al 2006; Camp 2007; Fodor 2007; Rescorla 2009a; Beck 2012; Kulvicki 2013; Giardino & Greenberg 2015; Pearson et al 2015; and Qusilty-Dunn (forthcoming).

[138] Fodor 2007; Camp 2007, 2009a, 2009b; Carey 2009; Rescorla 2009a, 2009b; Burge 2010b; Gauker 2011, 2017b; Block 2014; Quilty-Dunn 2016, (forthcoming); Boyle 2019b; Clarke (forthcoming).

appreciate the relative positions of the different locations using the map. This seems to be at least in part because in representing the locations of the cities in a map-like way, one *thereby* represents their relations. Using a map makes it is impossible or at least much harder than linguistic representations to represent just one feature, without taking a stance on a number of others (e.g. representing the relations of Cardiff to Swansea and Swansea to the Brecon Beacons, while remaining neutral on Cardiff's relation to the Brecon Beacons).[139]

This seems comparable to the differences between a representation of the events of Welsh history on a timeline diagram as opposed to a list of sentences. Marking a line with Llywelyn ap Gruffydd's defeat by Edward I and subsequent rebellion, Owain Glyndŵr's rebellion, the Chartist Uprising of 1839 and the Rebecca Riots, the Industrial Revolution etc. helps us see immediately which events came first, which are simultaneous with one another, which are far apart or close together, which lasted a long time, and so on. It is plausible that being able to represent events in a unified, surveyable structure like this is extremely significant: it allows for organization of events and comprehension of overall patterns. And it will turn out below that we can develop this idea in several ways which do capture cognitively significant kinds of temporal representation.

However, it will be useful to have a more detailed characterization of what it is that makes such forms of representation distinctive. Call representations like timelines and maps 'holistic', and the contrasting kind of representation, exemplified by sentences, 'atomized'. How should we characterize the difference between holistic and atomized representations? What does it mean for a map to be a single surveyable structure?

[139] Cf. Gallistel & King 2009: 208. Related ideas can be found in the literature on cognitive maps in animals, e.g. Tolman et al 1946; Tolman 1948; Olten 1979; Gould 1986; Menzel et al 2005.

One tempting way of going here is to say that maps represent all these features *simultaneously*. This suggests a distinction which certainly can be applied to temporal representations: whether the representations are diachronic or synchronic in format. A synchronic representation of a set of features would be such that all those features could be read from a time-slice of the representation, whereas a diachronic representation unfolds in time, and only represents some features at any moment. A strongly diachronic representation would use temporal properties to represent, much as a map or graph uses spatial relations: in this case there would be some features represented which would not be possible to extract from any single time slice of the representation. One can use a diachronic representation to represent time, just as maps can use spatial dimensions to represent space. For example, a regular cycle in the brain could be used to represent an external cycle (a diachronic representation of a temporal property); or that cycle might instead be represented by a static image, with space being used to represent time.[140] Even when temporal relations are used to represent temporal relations, it is plausible that temporal relations are sometimes used to represent *different* temporal relations in the external world, such as when some temporal pattern is recapitulated in the brain at one hundred times the speed, or in reverse.

This suggests that an important kind of temporal representation is a timeline, construed as a kind of representation which simultaneously represents the temporal relations between a number of different events.[141] However, this is to be misled by the idea of simultaneous representation. One might think that simultaneous representation of the relations between

[140] Cf. Campbell (1997: 112)'s distinction between 'rhythm' and 'cycle'.

[141] To think that human cognition *uses* such timelines, we do not need to think that we *store* all episodic memories in a way that immediately places them on such a timeline: it is plausible that we occasionally or even frequently (Friedman 1993: 58f.) have to do some reconstruction to figure out when exactly a remembered event occurred; but this reconstruction is often precisely a computation with the goal of placing the event on a timeline.

multiple events—of a complex sequence or cycle in its entirety—would allow for many kinds of operation to be carried out more efficiently, for temporal patterns to be spotted which otherwise would go unnoticed, and so forth. But this is not the case. For a diachronic representation of the entire sequence or cycle, especially if it can be played quickly, played only in part, or played in reverse, as need be, could allow all the same operations just as efficiently for practical purposes. Furthermore, temporal properties of neural firing, such as spikes per second, are generally taken to be the most plausible implementations of representations of many variables, not just temporal ones, and this is not limiting in the way one might expect if literally *simultaneously* representing different properties were so important.

Instead of looking at whether the representational vehicle is diachronic, I suggest that there is a different, more important sense in which a representation might include information about many different items and features 'at once'. We can recast this thought as about representations which are members of a kind of representation (whether it be maps, timelines, video clips, or representations in the brain which are analogous to any of these) whose instantiations systematically always include information about many different items and features *in one representation*. Unlike the simultaneity-based proposal, this suggestion is neutral on how individual representations are individuated. They will have to be individuated somehow—a set of sentences will need to count as multiple representations rather than just one—but exactly how they are individuated will depend on the kind of representation in question; and many representations will be diachronically extended. One important constraint on the individuation of representations will need to be that we should not require a great deal of extra processing to read all of the parts of the representation: we should be able to distinguish between a system that makes all the features it conveys explicit, and one which leaves them to be reconstructed from a

few explicitly represented features. But the way to make this distinction will not be in terms of *simultaneous* representation of all these features.

The characterization of holistic forms of representation covers a number of different kinds of complex structured representations of temporal sequences, cycles, and so forth. Besides all the distinctions we drew above between temporal representations in general, however, new and even more important distinctions arise specifically for holistic temporal representation.

In this context of holistic structures, it starts to be useful to distinguish between structures consisting of *times* and structures consisting of *events*—things that happen at times. If one represents a dimension of times, one might slot events into many of these times, but between these events there might be many times that one does represent, without representing them as filled by events (this is not to say that one would be representing these as times at which nothing happens; just that one is not committed about what happened at those times). Or one might represent variables as following continuous functions of time, without segmenting the time-courses of these variables into events. At the other extreme, one could represent a complex sequence or web of events, compactly representing all their before and after relations in a holistic representation, without representing any metric information about *times* at all.

Holistic structures could also differ in how they represent both events and times, for example in incorporating hierarchy, with holistic structures consisting of holistic structures at more precise scales. Rather than using, as it were, a smooth line to represent time, subjects may well represent the time of day and the month in which some event occurred separately, integrating these into a broader holistic representation when necessary but with the system sometimes breaking down at each level independently of the other.[142]

[142] Friedman (1993: 55).

A further important distinction is between a full framework and a smaller structure. We can think of this as a matter of *comprehensiveness*. At one end of this scale we have just a small structure of events, perhaps with the temporal relations between these events and the ability to slot further events into that small structure: most scripts and narratives will fall into this category. We can imagine a creature (indeed it has been claimed that human infants are like this) [143] who represents many such small structures but is unable to relate them to one another. For such a creature, there are just temporal 'islands'.[144] At the other end of the comprehensiveness scale would be a full framework of times, such that *all* events could be slotted into it.

A complete framework would specify some kind of geometry: presumably either a linear ordering (finite in neither, one or both directions), a branching structure, or a huge cycle of eternal recurrence.[145] Given the shape and extent of historical disagreement about what the overall geometry of time is (such as heated debates in Islamic and Christian Mediaeval Philosophy about whether the universe has a beginning, whether it is infinitely divisible, whether the future is fundamentally different to the past etc.), it seems unlikely, contra Campbell 1997, that we have a folk understanding of time which is strongly committed on these issues. Hence it is unlikely that humans had a fully comprehensive holistic representation of time before achieving theoretical and possibly technological sophistication. However, it is plausible that they had something more comprehensive than islands: structures relating some islands but not all to one another, perhaps.

---

[143] Friedman (2005: 151); Campbell (2006: 11); Hoerl & McCormack 2011.

[144] This last clause is important: we should allow that we sometimes do not have all the information we need to place an event precisely on a framework, but the framework might still be comprehensive in that we at least implicitly appreciate that we could place it on the framework given more information—especially if we can specify or recognize what further information would help.

[145] Compare Campbell (1997: 105).

Closely related to this sort of comprehensiveness, but important to distinguish from it, is *domain generality*. Time is unusual in that it appears in many domains: the mental, the physical, and the social; science and the arts. Fully understanding time would involve being able to relate events from different domains to one another in the same holistic temporal framework, and to at least appreciate the possibility of doing this for any set of events taken from any domain. Again, this sort of comprehensiveness could come in degrees, from the completely domain-specific system, to a system which can integrate the temporal properties of events from several domains but not all, to the fully domain-general system.

When we have comprehensive, domain-general holistic representations of many times, it feels natural to talk of representing temporal frameworks. Such frameworks will turn out to be cognitively significant. But these are not the only important kind of structured, holistic temporal representation. *Narratives* will often be quite narrow both in comprehensiveness and domain-generality, relating only a few choice events in one unified structure.

At a minimum, a narrative will be a structure consisting of a number of related events (or, for a generic, script-like narrative, event-types). It need not be a single sequence: it could be a tree-like structure of sequences converging on the same point, as in a love story with both partners' lives up until they meet, or a story about a band of adventurers who split up to pursue the same goal independently. Narratives are not just sets of events or even sets of sequences of events; they should have some additional kind of unity. Annals which simply recount all the important events which happened each year, organizing events only by when they happened without linking related events across years, will not count, and neither will a sequence of unrelated events. What does this unity come to? What more is there to narrative? A number of answers have been proposed. I will draw heavily on the discussion in Velleman 2003 in laying

out some of the most plausible, to give an option-space for understanding the most significant features of narrative-style representations in the mind in §4.5.

One idea is that narratives involve applying an already familiar script to organize a sequence of particular events, picking out some as important on the basis of a more generic narrative-type.[146] This would fit with a view of unification as assimilating data to a more general, repeating pattern, and explanation as unification. But it does not tell us what makes a script, a generic narrative-type, itself count as narrative: it is not plausible that any sequence would do. A similar problem applies to the idea that narratives put events into a broader temporal context, placing them in relation to other important events (especially autobiographical narratives which relate events to other major events in a life, which can be construed as a broader narrative).[147] It may be that narratives often do this, but it would be helpful to know why, and it is unclear that being placed in context is distinctive of narratives.

One distinctive structural feature of narratives might be that they have a clear climax, terminus, or endpoint: an event which all the other events in the narrative contribute to in some way (mostly via other events in the narrative).[148] This provides a certain sort of unity, but narratives might require more structure than this, such as emphasizing the most significant contributors to the final outcome.

One way of cashing this idea of a climax out is to emphasize that narratives often involve some kind of teleology. They often involve some combination of agents pursuing goals (perhaps including the motivating events which led to their forming that goal) and either achieving or

[146] Schank 1990.

[147] Nelson & Fivush (2004: 494, 500).

[148] Velleman 2003 finds versions of this idea in Mink 1968 and Gallie 1964, as well as in various literary critics.

failing to achieve them, and/or the universe in some sense bringing things into alignment with how they are supposed to be. It is striking that infants seem to at first organize collections of memories differently—around things like sharing objects rather than sharing goals.[149] However, it is not obvious that *all* mature narratives involve a straightforward teleological structure. This is particularly so in history, where we might tell a narrative of events leading up to the outbreak of a war or some economic crash, without even implicating that this outcome was anyone's (even the universe's) goal.

In any case, both the teleological view and the climax view imply a further condition, which is at least initially extremely plausible as a necessary condition on narratives: that there must be causal relations here, with the earlier events within each strand contributing to the later ones. Velleman finds versions of this view in thinkers from Aristotle to Carroll, but criticizes it on the grounds that there can be cases which count as narratives thanks to a kind of fittingness, without genuine causal connections between the events in question. His two examples here are stories involving twins separated at birth who are eventually reunited through a mere fluke, and Aristotle's case of the story that someone killed Mitys, only to later see a statue of Mitys and be killed by that very statue falling on him.[150]

Rather than treating these as degenerate cases, Velleman proposes an alternative account of what unifies the events of a narrative, drawing on further remarks in Aristotle: narratives produce an 'emotional cadence' in the audience that follows a typical temporal pattern of emotions. Velleman thinks that emotions involve 'biologically programmed' affect programs,

[149] Ratner et al 1986, Pillemer et al 1994.

[150] Aristotle, *Poetics* 9.1452b6-9.

which specify a particular routine or way for the emotion to unfold over time.[151] These specify characteristic elicitation conditions (such as unexpected positively or negatively valenced events), followed by physiological symptoms and distinctive feelings, reflexive behavioural symptoms and so on. Crucially emotions will either develop, decay or morph into other emotions according to these patterns, depending on what happens, as when fear morphs into relief, anger or grief depending on whether the apparent danger that caused it comes to fruition and whether this is due to some blameworthy action. A narrative with the right sort of sequence—which provokes an appropriate emotional cadence—unifies its events and renders them intelligible. It does so in the sense that it helps us survey the actions and events at a single glance, due to assimilating them to familiar scenarios, namely familiar sequences of *emotional responses.*[152] It does not convey new objective understanding of how historical events came about but a subjective understanding of how to feel about them, although we sometimes have a kind of projective error when we think that the sense in which events are thus 'explained' means we've understood what happened, as can be seen by thinking about the statue of Mitys case.[153]

Within the class of narratives, a particular class of narratives might be thought particularly likely to be cognitively significant: autobiographical narratives. Nelson and Fivush 2004 (alongside many others) emphasize the idea that we often organize many or all our memories into a personal narrative, a life story. For this, they claim, we need personal and cultural knowledge about how memories are organized into the story of a life (or a chapter of one's life): which events are considered important and why, which sorts of stories are typical,

[151] Velleman (2003: 13).

[152] Velleman (2003: 19).

[153] Velleman (2003: 20f.). Nelson & Fivush 2004; 494 make a similar claim, that narratives capture evaluative information, and are taken to having meaning or a broader moral.

and what a self is for that culture. For Nelson and Fivush, we actively go about performing this kind of organization, and the stories we tell ourselves and others not only shape how we structure our memories, but also shape our actions, as we try to live out certain narratives.

So we can make a huge number of potentially important distinctions between systems of temporal representation. Our task now will be to consider ways in which temporal cognition could be cognitively significant (in minds without understanding of human culture and technology), and to see which of these distinctions between types of temporal representation turn out to be playing an important role in this cognitive significance.

4.3 Cross-Domain Integration and Cross-System Coordination

To get at our first potential cognitively significant role for sophisticated kinds of temporal representation in technologically and scientifically unsophisticated creatures, it will be useful to consider one important role temporal technology seems to have played in sophisticated human societies historically. Boerner & Severgnini 2019 have analysed an impressive set of data relating to the effects of the introduction of mechanical clocks in Late Mediaeval Europe. Before this period, people had to rely on sun- or water-clocks to tell time. These were either not reliable, precise, and accurate enough, or not easily accessible, so did not play any role in everyday economic and social activities. Events like the opening and closing of markets were coordinated based on publicly available but imprecise measures like the Sun rising, setting or reaching its zenith. When mechanical clocks were built, however, they were often at the top of high towers, and made a noise every hour. People did not immediately take advantage of the opportunities afforded by such technology (which was generally built more for prestige than with a view to shaping business practices). But over the course of several centuries, the clocks started being used for coordinating activities, to the extent that new activities became possible, such as

having multiple meetings in one morning. Boerner and Severgnini show that a town's building a clock had significant effects on economic growth decades later.

Just as public clocks can coordinate economic activity, temporal representation could be used in the mind to coordinate and integrate the activities of different systems. It could be used for low-level coordination, such as coordination between different sense modalities and action systems, all dealing with slightly different lags relative to the world. More generally, shared temporal frameworks could function as common currency for relating events that are dealt with by different systems. Time is even more domain-general than space: even mental states have temporal properties, whereas they do not have spatial properties (although the brain-states that realize them, of course, do). Coming to link, say, certain kinds of mental events (like intoxication) and certain external stimuli (like fermented fruit), might turn on being able to spot a temporal pattern of one of these sorts of events reliably preceding another by a certain duration. Being able to conceive of such linkages playing out in time, meanwhile, might be crucial to being able to make sense of the idea that they might be related.

Generality and bridging gaps between domains is seen in many traditions as core to general intelligence, so if temporal representation is unusually important to this ability, it is significant indeed. For example, one of Fodor 1983's two defining properties of 'central cognition' (genuine intelligence, as opposed to domain-specific, informationally encapsulated and inflexible modules) is being what he calls 'isotropic'.[154] Isotropy means that everything one

[154] The other, which he calls 'Quineanism', relates to one's degree of belief in any given proposition being sensitive to properties of the overall structure of the totality of one's beliefs, like simplicity. Fodor (1983: 107f.).

"knows is, in principle, relevant to determining what else [one] ought to believe. In principle, our botany constrains our astronomy, if only we could think of ways to make them connect".[155]

A quite different tradition sees a related kind of generality as a key mark of genuine concepts and understanding, as opposed to kinds of representation resting on less intelligent processes.[156] This is the tradition centering on Evans' Generality constraint. There are various formulations of this constraint; one version requires that one can grasp thoughts combining the thought-based equivalents of any n-place predicate and any n-tuple of singular terms in one's genuine conceptual repertoire, provided they are in appropriate categories.[157] Violating this constraint would mean being unable to grasp a thought that is (i) syntactically well-formed, (ii) consists of concepts one has and which (iii) should be relatable to each other: for example, possessing the concepts SHARK and SLEEP, being able to grasp other thoughts using these concepts, like SHARKS EAT and BABIES SLEEP, but being unable to grasp the thought SHARKS SLEEP. If temporal representation acts as a common currency between domains, it could be crucial to grasping some thoughts like this. Being able to figure out what it would mean for sharks to sleep would involve being able to apply what one knows about the domains they belong to (say *ordinary activities* and *underwater creatures*), which means having a common point of reference. This could be provided by appreciating that sleeping is an activity on

[155] Fodor (1983: 105).

[156] Concern with this as a mark of genuine concepts dominates discussion in, Peacocke 1992; Camp 2004, 2007, 2009; Heck 2007; Bermudez 2007; Carruthers 2009; and Beck 2012.

[157] Evans' original versions can be found in Evans (1982: 75, 100ff.). The qualification about belonging to appropriate categories is intended to avoid requiring that one grasp syntactically well-formed thoughts like COLOURLESS GREEN IDEAS SLEEP FURIOUSLY. Such thoughts would, the idea goes, involve category mistakes an attempt to combine e.g. colour concepts like GREEN with nouns concepts like IDEA belonging to a category (maybe *mental* states) which does not allow for combination with colour concepts. Camp 2004 argues that this constraint is unnecessary and that we should adopt a less restricted version of the generality constraint.

a regular (usually circadian) cycle, which consists of certain patterns of awareness, levels of activity etc.: the key will be mapping a temporal pattern onto possible activities of sharks. Of course, temporal representation will not be the whole story in meeting the generality constraint, and in *some* cases, it may turn out not be part of the story at all; but it will provide a key part of the story in *many* cases.

How sophisticated does temporal representation need to be to act as a common currency? The answer seems to be: Not very, but greater sophistication may help in special cases. All that is required for a common currency is representations of the same temporal feature, in a domain-general way. This could mean plotting all events on an utterly comprehensive temporal framework. But representations of cycles or durations which could be coordinated would also work for features like the sleep patterns of sharks. We can even see temporal representations playing the role of a common currency in integrating different sense modalities and action representations, suggesting temporal representation need not always be conceptual to play this role, although providing for the kind of integration of cognition that makes for meeting the Generality Constraint would require more than non-conceptual representation of time. Narrative structure, meanwhile, seems irrelevant. As for representing particular events, this does not seem to be especially helpful: relations can just as easily be mapped between domains using shared script or cycle structures.[158]

Mapping different domains to one another in a systematic way would, however, benefit from a certain amount of comprehensiveness in temporal frameworks: shared landmarks and shared units (or the ability to systematically translate) across systems or structures of representations would be important, and an understanding of, for example, how repeating cycles

[158] Cf. Campbell (2006: 7).

can be mapped onto linear structures, would be of help. With a full framework, extending indefinitely, with a metric structure of times rather than just orders of events, one could gain full generality: a framework into which *any* event from *any* domain could be placed.

4.4 Non-Markovian Dynamics

Chapter 2 showed that Reinforcement Learning (RL) algorithms can, under certain assumptions, reliably produce the behaviour that a diachronically rational agent would produce, and can do so without representing time at all. A natural place to look for cognitively significant kinds of temporal representation would be in overcoming the limitations of RL. In particular, certain kinds of temporal representation might be needed to produce diachronically rational behaviour in environments where the Markov property fails (we will consider other limitations of RL in the course of this chapter).

An environment has the Markov Property iff, for all t, the only way its state at time t - 2 affects its state at time t is via its effect on the environment's state at time t - 1.[159] In the context of RL, where we model the environment as developing probabilistically on the basis of past states and actions, the environment has the Markov property iff $p(s_{t+1},r_t|s_t,a_t) = p(s_{t+1},r_t|s_t, s_{t-1}\ldots.a_t, a_{t-1}...)$. This would fail if, for example, an action taken now triggers an unobserved process whose effects only become apparent three periods in the future.

Note that the issue here is whether the world *as it is modeled* by the subject is Markovian: an unobserved process will mean that the state of the world *as modelled* will not depend on the immediately preceding state of the world *as modelled*. In such a case, someone who knew about

[159] This is slightly simplified for ease of comprehension: instead of just t-2, we care about t-n for all n>1; and while this formulation uses causal language, the important issue has to do with conditional probabilities rather than causation. The statement of the Markov Property in the context of RL avoids these limitations.

the unobserved variable may be able to describe the underlying world in a completely Markovian way, where the world in each period was entirely determined by the world (*including the unobserved variable*) in the preceding period. But if the only way to learn about the hidden variable is to spot the delayed effect of the observed events two periods ago, the fact that the underlying dynamics of the environment are Markovian will not do much good to the system which does not represent the hidden variable. Such a system would only able to spot effects of the variables it does represent on those same variables just one period ahead.

The RL models discussed in ch. 1 assume the dynamics of the environment are exhausted by p(s`,r|s,a). This is crucial to procedures like simply updating a state's value estimate every time it is visited, on the basis of observed effects, rather than taking into account when it is being visited or what else has happened so far. If the environment has the Markov Property and remains stable, the value of the state will not depend on what happened one period ago, or two periods ago, or before that: its value will be the same whenever it is visited.

The Markov Property almost certainly does fail in many real-world cases, and sometimes in extreme ways. This problem is not unique to RL or machine learning: we will also see a version of the same issue crop up in Campbell's framework in the next section. So if certain kinds of temporal representation enable learning models of the environment (or ways of assigning values to states or actions) without assuming the Markov Property, they could be highly significant.

What would a non-Markovian replacement look like? At a schematic level, it seems that any algorithm which could learn non-Markovian dynamics would need to be able to zoom out from individual event-pairs to see overall temporal patterns in the data. However, it is not clear

what such algorithms might look like in detail.[160] We do not have good models of general, plausible learning mechanisms which avoid this assumption, so it is not clear what kinds of temporal representation would be required. We do have sophisticated ways of analyzing time-series data and building models on that basis, and these do require sophisticated temporal frameworks, but unlike RL and other kinds of learning (including the relatively sophisticated sort of learning described in Ch. 5), it is not clear how they could be implemented at all efficiently in the brain. Furthermore, we have very little systematic knowledge of how different RL algorithms perform under such conditions, or how to fix them. Standard tools for assessing algorithms become extremely difficult to apply without assuming the Markov Property. For example, most convergence proofs showing properties of the value function or policy an algorithm converges on in the long run rely on assuming the Markov Property, and without it the maths often becomes intractable.

Unfortunately, because it is hard to study RL-style algorithms in such circumstances, it is difficult to say which kinds of temporal representation could help and how. It is possible that explicit representation of time could be useful in simple heuristics to supplement more traditional RL algorithms, and these might only require island-representations or narratives. But it might be that full temporal reasoning, over a synchronically represented, comprehensive, unifying framework, might be required.

We can say something more informative about more specific kinds of understanding. These do not use completely general algorithms like we are searching for with RL and general intelligence. Instead, they use certain kinds of simplification that make the problems posed by non-Markovian dynamics more tractable.

[160] Various ideas have been proposed (Sutton & Barto 2018: 464ff.), but all are either very demanding in terms of data, or only work under other restrictive assumptions.

4.5 Essentially Dynamical Entities

Many kinds of entity are defined in part by their dynamics. Understanding and dealing effectively with these entities requires representing a certain kind of sophistication in temporal representation which also provides opportunities for overcoming some of the most pressing non-Markovian features of animals' environments . I will get at these points by first surveying proposals by Campbell, and by Hoerl and McCormack, about the importance of temporal frameworks for understanding objects and causal relations, and then drawing out what is right in their discussions, and showing how it can be generalized.

I will be extracting lessons from Campbell and Hoerl and McCormack's discussions rather than engaging in close exegesis. Whilst both Campbell and McCormack and Hoerl are in some sense trying to capture why temporal frameworks are important, they are engaged in a slightly different project to giving an account of the cognitive significance of temporal frameworks. They are more interested in the constitutive question of what it takes to count as representing temporal frameworks at all. While Campbell 1994 describes himself as concerned with the 'causal significance' of a temporal framework, he thinks this 'gives meaning' or 'assigns physical meaning' to the framework and takes himself to be in dialogue with pragmatists and empiricists about our concept of time. Hoerl and McCormack, meanwhile, often seem to be engaged in either the constitutive or epistemic projects. However, their suggestions can be carried over to the project here.[161]

[161] Indeed, they look more plausible as suggestions about cognitive significance than as constitutive claims, where they are arguably excessively demanding. The suggestion that having a certain kind of understanding of causation is *constitutively necessary* for having the temporal frameworks is much logically stronger than the claim that having temporal frameworks is i*mportant* in part because they can play a crucial role in the relevant sort of causal understanding: the latter view allows that one could have temporal frameworks without actualizing their full cognitive significance.

Before explaining his view of temporal frameworks, Campbell 1994 gives an account of why *spatial* frameworks are significant. Spatial frameworks are used, Campbell thinks, to understand how objects maintain identity through movements in space, and to appreciate the 'internal causal connectedness' of objects.[162] An object is internally causally connected when its condition at one time causally depends on its condition at earlier times (in addition to its interactions with external things)—even if this just means there is something about the object which means it will reliably maintain stability of shape and location.[163] Coming to understand informative identities with respect to physical things (e.g. realizing that Mt. Afla just is Mt. Ateb, viewed from a different location), requires the ability to use a spatial framework to track and reidentify objects based on understanding their internal causal connectedness.

Campbell gives closely related accounts of the significance of representations of temporal order, the past tense and the direction of time. The significance of these sorts of temporal representation comes from the idea that part of what it is to be a causal relation is to be temporally asymmetric. Causes must temporally precede effects. One cannot affect the past.[164] Representing the order of events allows one to (tacitly or explicitly) follow this temporal priority principle, when constructing narratives or even just assigning causes to events.[165] The

[162] Campbell (1994: 32).

[163] Campbell (1994: 27).

[164] Campbell (1994: 64-71, 1997: 107f., 2006: 8ff.). The idea that representation of temporal order is related to appreciating that causes must precede effects traces to Kant: for discussion see Hoerl & McCormack (2011: 449ff.). One might quibble with the idea that causation is temporally asymmetric in this way, especially through appeal to features of modern physics. However, even if causation turns out not to be asymmetric, treating it as such does seem to be a very good idea for most contexts which animals (including most humans in history) who are not in a position to appreciate the relevant physics will find themselves in.

[165] Hoerl and McCormack 2011: 441 survey conflicting evidence about the extent to which three year old humans' judgements of what caused what respect this principle. Roughly, they seem to respect it in some circumstances, but it can be overridden, for example if they are shown evidence that A, not B, causes C, before encountering a particular instance where B but not A precedes C.

latter activity is particularly important to Campbell, who thinks that it is crucial to understanding the self that we understand the it as a common cause of numerous effects and construct narratives which imply this.[166] Campbell seems to think that causal understanding obeying this principle (and hence also representation of temporal order) will be very widespread, as he thinks it is necessary to engage in any complex action.[167]

However, while explicitly using such principles might be required for explicit reasoning about which of a variety of potential causes and effects is which, such principles need not be explicitly represented for ordinary interactions with the world. The kinds of RL algorithms discussed in Chapter 2 do not involve representation of either temporal order or causation, yet produce effective, complex action. Even when they are based on models of the world, these may use probability distributions rather than causal nets. And it seems one can learn a non-Markovian model of the world, which includes causation and implicitly follows the principle that causes must be prior to effects in all kinds of ways, without representing temporal order. Most naturally, this could be done through having two architecturally defined slots corresponding to the present and the previous period (as in the learning systems discussed in Chapter 2), where part of their functional role is that the 'is caused by' relation can only be applied from the occupant of slot 1 to the occupant of slot 2, and not vice versa.

[166] There is a further strand of Campbell's ideas here which will not be useful to discuss in detail because it depends on so many of his other views: For Campbell, the self is a particularly important special case of an object which develops over time due to complex causal interactions. Campbell 1997: 107f. argues that understanding the self as a common cause of many effects at different particular times is required for providing an empirical grounding for the 'transparent unity of the self', i.e. the assumption that all first-person memories can refer to the same self. He thinks one needs such a grounding to count as grasping the use of 'I' as a referring term, and hence that one only counts as doing so if one constructs self-narratives which at least tacitly respect various principles about how the self persists as a spatiotemporally continuous unit and interacts with other objects.

[167] Campbell (1994: 54).

Campbell has a separate account of the significance of representing *particular past events* within a comprehensive temporal framework.[168] Representing particular past events allows a "narrative grasp of a collection of remembered events", by which Campbell means a grasp of the mass of the events' causal relations to one another: the most minimal version of this seems to be representing "a process that persists over time and whose later stages causally depend on its earlier ones".[169] Physical objects undergo numerous interacting processes whilst persisting. Understanding such causal networks and sustained processes requires representing particular times within a comprehensive framework, according to Campbell, because the state of a physical object at later stages will bear the mark of earlier interactions, and so it will respond differently to the same causal influences depending on its particular history. Even if the object is undergoing a repeating cycle of causal influences, it will react differently to different occurrences of the same phase, as the object's properties gradually change. For example, even properties like how a physical object will react to being pushed will depend on properties subjet to change, like whether it is at the top or bottom of a slope, its shape, weight and texture. So even appreciating the object's interactions particular times in small temporal islands rather than a comprehensive framework will not do: the same objects will appear in multiple islands, and the object's behaviour in later islands will depend on its interactions in earlier islands. The islands themselves need to be placed within a linear framework of particular times.[170]

Campbell thinks that it is possible to somewhat successfully interact with such objects without a reflective grasp of the relevant causal networks and how they have evolved over time.

---

[168] Campbell (1994: 54).

[169] Campbell (1994: 59)

[170] Campbell 2006 brings out this last point more explicitly than Campbell 1994.

There can be non-reflective 'working concepts' of some of such objects' properties, which allow for very limited sorts of reasoning about how they will respond if manipulated in certain ways. But one will not be able to consider the full range of counterfactuals and have a full grasp of why they hold without understanding the temporal-causal network.

Hoerl & McCormack (2011: 455) criticize Campbell's suggestions here. They are not convinced that a *linear* ordering of time is needed for appreciating the dependence of objects in one island on their interactions in earlier islands. They point out that one could grasp the role these objects have in transmitting causal influence over time through representing aspectual properties like *chewed* or *squashed*. One would not need to represent the particular event in which the object was chewed or squashed in order to understand how it will behave now, and how this is different to occasions when it was not chewed or squashed. Rather, one just needs to understand that it has a different property now to on those occasions.

Hoerl and McCormack suggest that representation of a linear ordering of times will only be important in cases where the order in which multiple past events happened to the object makes a difference to how the object will behave now. This will sometimes require representation of order rather than just relying on aspectual properties, especially if the subject is not presented with events in the order in which they happen but has to infer the order from causal effects. An example of this (not theirs, although similar in essentials) would be where a subject has to infer who ate a slice of cake most recently from the state of the cake, knowing that one of the people would leave it in a messy state with crumbs everywhere, and the other would clear up after the other and leave it in a neat state.

There are a few worries one might have with these proposals. First, while Campbell does distinguish his account of the significance of representing temporal order from his account of the

significance of representing particular events in a linear framework, some of the distinctions above cast doubt on exactly what is needed for the sorts of understanding he and Hoerl and McCormack discuss. Whilst both Campbell and Hoerl and McCormack seem to commit to full comprehensiveness—a subject representing a linear structure incorporating all events—being required for understanding either objects' changes over time or cases where the order of events matters causally, only a certain degree of comprehensiveness seems to be required. The sequence of events that can be represented needs to be at least as long as the chains of causal interaction one seeks to understand, and needs to preserve an ordering over these events. However, to do all this the subject need not take a position on, for example, whether the structure of time will ultimately be a much longer cycle rather than a linear structure; and it does not need to incorporate islands from outside the chain in question. Relatedly, it is not clear why the events need to be represented as particular: generic scripts featuring multiple instances of the same event kind, and stretching as long as the interaction in question, seem to be sufficient for their purposes. What's more, there does not need to be representation of *times* as opposed to *events* for either of their proposals to work.

More importantly, these suggestions so far look like mere special cases of the problem of non-Markovian dynamics, without a new solution. Hoerl and McCormack point out that unlike Campbell's, their proposal just has to do with the order of events mattering, rather than having anything in particular to do with *objects* and their persistence. But we can generalize further: the root of the issue about objects for Campbell seems to just be that both objects and the causal webs they are in do not obey the Markov property: their causal dispositions and powers depend on their history, stretching back beyond the immediate past. And Hoerl and McCormack's cases, too (of outcomes depending on the order in which various events happened in ways which

require representing that order to predict the outcome), concern events from several time periods ago (the first event in the sequence) having effects which are not screened off by the most recent events. By restricting the general problem of non-Markovian dynamics to these special cases, Campbell and Hoerl and McCormack seem to be underplaying the significance of sophisticated temporal representation.

One response to this worry is that focusing on these special cases points to a way of providing traction in learning about at least some non-Markovian cases. §4.4's issue was that we do not have a good account of general learning algorithms for learning non-Markovian dynamics efficiently. Perhaps, restricting attention to the kinds of case Campbell and Hoerl and McCormack are interested in, rather than dealing with all non-Markovian cases at once, is useful. It would be useful if we do have efficient learning algorithms for these sorts of problems. This would provide the makings of a good account of a distinctive reason why the kinds of temporal representation appealed to by such learning are cognitively significant.

Neither Campbell nor Hoerl and McCormack gives or even points the way to such learning procedures: at best, we get from them claims that certain kinds of temporal representation *will be required* for understanding in these cases, much like the claims we were able to make about non-Markovian cases in general in the previous section. But there is a way of drawing on Campbell's ideas about objects and causation to do so.

My account along these lines will go as follows (I will summarize it in this paragraph and explain the different parts in turn): Many kinds of entity (using this word as a maximally general term to include objects, properties, and relations) are defined in part by their (sometimes non-Markovian) dynamics. Such kinds include *physical object* and *causal relation* but also more specific kinds like *person*. Given holistic, structured temporal representations, it is possible to

have cognitive templates or prototypes corresponding to these kinds of entity, which can be applied in tractable ways to understand and deal effectively with certain instances of non-Markovian dynamics. This could make these sorts of temporal representation deeply significant, as a way of efficiently supplementing and overcoming the limitations of more general learning algorithms to understand a wide range of important temporal patterns.

Many kinds of entities are defined in part by their dynamics. Campbell's case of physical objects—defined by maintaining some of their properties in a certain way over time, and tracing a continuous spatial path through time, while changing many of their properties as a result of their causal interactions—provides one such example. His approach to causal relations—defined in part by their temporal asymmetry—provides another. At this level of very general categories, we might also point to categories like *process* or *rate of change*. But more interesting for our purposes, given that it is often more plausible that animals represent concrete, specific categories than extremely abstract, theoretical ones are categories like *organism* (or if this sounds too scientific, *living thing*), *self*, and *social hierarchy*.

In the literature on temporal representation, it is standard when introducing the idea of repeating cycles to note that all *organisms* go through numerous regular cycles (sleep-wake cycles, reproductive cycles, maintaining a heartbeat and regular breathing etc.). But this is not always linked to the point that one of the defining features of being a living organism is having a metabolism. Organisms are mortal: they decay and die. But at a shorter timescale, we are remarkable in our ability to maintain a stable state: to maintain key variables (glucose levels, oxygen levels, temperature etc.) within a certain range throughout our bodies. One of the most efficient ways to do *this* is to go through regular cycles allowing for fluctuation around ideal levels that can be fine-tuned, adding necessities in response to shortages and subtracting away

unwanted in response to excess — like having the heart pumping blood around the body to maintain oxygen levels. And because of our tendency to maintain ourselves, we go through regular dynamics when damaged, repairing ourselves in ways that inanimate objects would not. Part of what unifies us as individuals is having metabolisms which maintain a certain bodily structure. It is also part of being an organism that we go through regular cycles on a longer scale: life-cycles of maturation and growth, then reproduction following more or less stereotyped patterns for our species. All these temporal patterns are not just widespread: they follow from the fact that constitutively, life involves tendencies towards sustaining and reproducing. Understanding what organisms are requires complex temporal representation.

Minds are also partly defined by their dynamics. Nearly all kinds of mind involve learning, at least in the sense of long-lasting changes in the mind in response to particular stimuli. Perhaps it is possible to have minds with no learning, and just action-guided perception according to fixed patterns. But even such simple minds will need to reliably display temporal patterns of adaptation to a stable scenario, and responses to changes in stimuli with characteristic sequences of actions like information-seeking and avoidance or mitigation of harmful stimuli or pursuit of useful ones. Furthermore, even simple biologically embedded minds will also reliably undergo maturation throughout the lifecycle, with preferences and priorities changing depending on the organism's status with respect to reproduction.

Different kinds of social groups will also be defined partly by their dynamics, as dominance hierarchies, kinship and friendship relations move in and out of equilibria and undergo predictable changes and long-term trends in response to members' maturation, reproduction and aging, periods of competition and coalition-building.

Even the ecosystems which animals inhabit, or the different regions they move between, will be partly defined by their dynamics. Rather than being defined by static properties, they will be defined by reliable patterns of seasonal variations in weather and fire, regular cycles of population dynamics and associated patterns in all the variables affected by species going through cycles of birth and death, migration, predation. Like individual organisms and social systems, they will go through periods of static equilibrium or following stable cycles, interrupted by periods of unpredictability, and will undergo broad trends over time.

This suggests that many of the entities that are most important to animals' survival and flourishing are essentially dynamical entities. But this does not yet show that sophisticated types of temporal representation might be significant in technologically and scientifically unsophisticated animals. One might accept that what it is to be an organism, a social system of a certain kind or a local ecosystem is to be an entity with dynamics of a certain kind, while also maintaining that animals can interact successfully with such entities without representing their dynamic features as such. After all, one might also think that one can only understand what it is to be an organism if one understands the theory of evolution by natural selection, but clearly it is possible to represent, learn about and deal effectively with organisms without *that* theoretical understanding. Could a simple-minded organism not get by in interacting with essentially dynamical entities without representing them as essentially dynamical, through some combination of non-representational temporal updating and simplified representations of the dynamical objects' static features?

It certainly is possible to interact successfully with at least some essentially dynamical entities without sophisticated temporal representation. To take the examples of physical objects and causation, one could in both cases have mere temporal updating systems with architectures

set up so as to respect some of their dynamical properties, as was said above. In the case of objects, this could mean tracking these objects in ways which implicitly respect their persisting in shape and following a continuous path through space. In the case of causation, this could mean using the kind of system discussed briefly earlier in this chapter, which had different slots corresponding to different times and only treated the occupants of one of these spots as candidates for causes of the occupants of the other, and not vice versa.

There are two ways in which sophisticated kinds of temporal representation could nonetheless be significant because of dynamically complex objects, however. First, mere temporal updating systems have their limits in terms of flexibility, as discussed in Chapter 2. Understanding essentially dynamical objects more deeply, because one understands their dynamics, could help with one's interactions with them in all kinds of complex, novel scenarios the architecture of the updating system has not been set up to deal with in advance, in integrating the dynamics of different objects, dealing with changes in dynamics and so forth. In general, high-level understanding of some phenomenon will have its benefits, whether that understanding draws on high-level temporal representation or some other kind of sophisticated representation.

But, second, we can say something more interesting in these cases, which relates specifically to *temporal* representation. Because so many important categories of objects are essentially dynamical, having dynamical models of some of these objects could solve some of the key problems generated by non-Markovian dynamics in important classes of special cases, and hence enable a way to get around the limitations of RL and similar kinds of learning. By fitting entities encountered in the world into a limited number of dynamical categories, one can massively narrow the hypothesis space about the dynamics of the entities one encounters, effectively collapsing all possible dynamical interactions into a few possibilities defined by a few

temporal templates associated with different kinds of objects. If the dynamical categories one has are well-chosen—if they capture dynamical patterns which do occur in the world a great deal—then models of the world based on them could do better than Markovian models.

Given the complexity of things like organisms, social systems, minds and ecosystems, combined with their complex dynamical patterns of growth and maturation, disorder and returns to equilibrium, they will have many important non-Markovian effects from the perspective of a creature which is not capable of representing all their hidden variables.

For example, all the systems which switch between states of equilibrium and disorder, such as organisms repairing themselves and returning to normal functioning after bodily damage, or social systems going through periods of upheaval as some individuals try to seize power, then settling into a stable hierarchical order, might see residual effects from the period of disorder play out several periods after stable order has been, to outward appearances, restored. Even as the new stable state is maintained and it appears that each period's state predicts the next period's state will remain the same, processes might be playing out behind the scenes (relating to grudges or distrust from the period of disorder and power struggle in the case of social systems, or to the need to replenish reserves of fat etc. in the case of healing from bodily damage), which result in slight departures from stability several periods into the future. The mind will provide many other examples of non-Markovian dynamics, especially if the mind being modelled is capable of forming plans or intentions to do something only upon some signal, like the dawn, and then between the forming of the plan and the signal carries on as before.

Having an understanding that bodies tend to use resources to restore equilibrium over time, or that minds tend to form plans in response to certain conditions and hence change what their behavioural responses to stimuli far down the road, can help one model what their

behaviour will be. Without any sophisticated temporal representation at all, these non-Markovian patterns would not be spotted. With a sophisticated temporal framework, the patterns would be there but hard to spot amidst all the possible temporal patterns in a mass of unstructured data. But with good temporal templates associated with the relevant kinds of entity—a template for organisms saying they will tend to expend resources to return themselves to equilibrium in response to damage, or a template giving the dynamics of intentions and plan-fulfillment—such patterns could be spotted. They would either be incorporated within the templates (they would be typical behaviour for objects of that kind), or only slight modifications or extensions on them. The templates would give patterns of expected dynamic behaviour to give order to the incredibly complex evolution over time of the objects in question.

This is much like how the perceptual system being primed by evolution in combination with perceptual learning to form representations of objects with certain sorts of typical shapes (concave rather than convex, bounded, with edges etc.) is crucial to our making sense of the huge range of possible interpretations of incoming sensory stimulation. One need not learn from scratch in every situation that there are physical objects with certain shape, colour, and textural properties: one can extract these properties from sensory stimulation very quickly because of biases in what to expect. Likewise for extracting objects' longer-scale dynamic properties from interactions with them on numerous occasions. One is faced with an intractable problem: learning about which of a huge number of possible dynamics, including many non-Markovian interactions, these objects exhibit. And one can cut this hypothesis space down by representing or tacitly appreciating the typical dynamics of that object, and applying this representation or tacit appreciation to a representation of that object's behaviour over time.

Campbell's points about objects' internal causal connectedness and persistence through change, and about causal relations being temporally asymmetric, can be treated as the sorts of information that would be included in such templates. Now, the category *physical object* will not include a huge amount of information about dynamics, as it is such an abstract category that it includes many sorts of objects with very different dynamics—animate and inanimate objects; rocks, pieces of ice, and electric food processors. And in general, more specific categories (*organism*, *animal*, *bird* etc.) could contain more information about dynamics. We can also expect that, like with other properties, represented sub-categories will sometimes inherit information from their represented super-categories. When one learns about a new species, one might import all the expectations one has about animals, or organisms, or physical objects. Furthermore, we can expect that when encountering a new kind of object one can be directed to look for certain dynamic patterns to see if it should be brought under a general class: does it go through life stages and repair itself from damage like an organism, for example?

Exactly which kinds of sophistication in temporal representation are required for this sort of understanding? To form predictions about dynamics on the basis of templates and check if those predictions are born out, subjects would need to represent temporal patterns in unified, holistic representations of the objects' behaviour at multiple times. However, these representations would not need to be especially comprehensive: they would not need to extend in range (or connect different islands) beyond the length of the temporal patterns predicted by the templates; and they certainly would not need to involve representation of the overall structure of time (whether it has a beginning, branches etc.). It will be the case, however, that a greater range will lead to longer range patterns being representable. Likewise, the amount of precision required, and whether subjects would need to represent cycles, durations, or sequences, seems to

depend on the distinctive features of dynamics of the entity in question: for example, representation of durations and cycles would not be useful in understanding entities which always evolve through a certain sequence of stages, but take unpredictable durations to do so. We do not have a clear break here between creatures with and without temporal frameworks, but rather degrees of sophistication, leading to degrees of understanding of complexity.

Finally, narratives, at least if these are understood as involving more than sequences of events, such as teleological structure or causal links between every event and a climax, seem not to be immediately irrelevant to this style of understanding. Frameworks are cognitively significant here, whilst narratives are not. This is not to say that narrative representations are insignificant *tout court*, however. They may have a different use.

4.6 Narrative Understanding

Once we distinguish different kinds of sophistication in temporal representation, we can see that understanding dynamically complex entities does not seem to involve narratives. Nonetheless, narrative understanding is nonetheless important and distinctive. There is a reason we often think about aspects of our lives in terms of narratives. As we have seen, our understanding of the world can go beyond that provided by Markovian models, but it is challenging to give any general learning algorithm for constructing alternatives. Instead, we can find strategies which work in many cases and turn on temporal representation of various kinds. In the previous section we saw how templates for the dynamics of certain kinds of objects could help, whilst not coming close to a general learning algorithm. These can be supplemented by the use of narrative representations, which capture other aspects of the world's temporal complexity.

There are two ways narrative representations could help here, corresponding to two different kinds of narrative representation: skeletal scripts can be used for capturing coarse-

grained but oft-repeated patterns. More specific historical understanding can be used for capturing more intricate mechanisms.

Scripts, as we discussed above, represent typical, generic sequences of events. Why would these help with non-Markovian interactions? If the entire script is represented (as opposed to McCormack's suggestion discussed above about infants' scripts really amounting to just a temporal updating system), they span multiple events in a sequence. This implies they can capture the effects of the early parts of the sequence on the later parts, even if those effects do not go via the events represented as in between. For example, a script for bath time might start with getting out pyjamas, getting undressed, running the water and so on, and end with getting into the pyjamas. This script might enable a child to appreciate the connection between getting out the pyjamas and their being available later, even if the child would forget about getting the pyjamas out if they pursued a similarly complex and time-consuming but *unscripted* sequence of events between getting them out and needing them to be available.

Learning scripts avoids the issues of intractability of general non-Markovian, because in learning a script one is not building a general model capturing a wide range of circumstances. One need only consider a relatively narrow path through the space of possible states of the world, and one can remain neutral on what will happen on other such paths. So, one need only collect evidence and devote cognitive resources to considering that narrowly defined sequence.

This strategy of pursuing only a narrow path through state space to provide tractability is taken to a greater extreme by a less generic use of narratives, which we might call *historical understanding*. Here we have representation not of a skeletal script for a typical, oft-repeated sequence of events, but of a detailed series of events leading to one specific outcome.

This sort of understanding will be particularly useful where the individual events comprising the narrative and their local effects are relatively well understood, but their longer-range effects, and how the different sorts of events fit together to produce an emergent phenomenon, are not. In such cases, representing all the events in the sequence in a holistic representation can make the larger patterns clearer, and can allow for reconstructing the causal path from the earlier events to the later events—even if that path involves skipped, non-remembered or non-observed steps, so would not be learned by a learning algorithm assuming the Markov Property.

This strategy for understanding a special case of a very complex sequence of events avoids the intractable demands that would be imposed by trying to understand such complex events in more general terms. And it does so in a way which mirrors a strategy which has been much studied in philosophy of science recently: studying how specific kinds of mechanism, defined by complex arrangements of better understood components, give rise to specific phenomena, as an alternative to trying to find general laws. This mechanistic strategy is thought to be common in science, especially in the life sciences and special sciences.[171] Just as we might focus on how one sort of mechanism (a very specific combination of physical parts including activities) gives rise to one phenomenon, rather than looking for general laws, we might focus on how one specific sort of narrative structure (very specific sequence of events) gives rise to one phenomenon. This coheres well with the idea that narratives have a climactic structure, with multiple strands of events leading to one specific event—the phenomenon to be explained.

[171] E.g. Machamer et al 2000; Robins & Craver 2009; Bechtel & Richardson 2010.

Would historical understanding actually be useful to animals? Is it not bound to specific past events which are unlikely to be repeated in the future? We can separate out the issue of being about the past from the issue about being particular here.

The terminology 'historical understanding' does come from the parallels between using these kinds of narratives to understand past series of events. But we can also understand sequences of events in the same way—using a holistic representation of a sequence (or set of sequences) of specific events, leading to one specific event—when events in the sequence are not in the past. Such understanding could also be applied to simulated sequences of events for the purposes of counterfactual understanding or future prediction. The exercise of constructing such representations could be a good means of understanding how events that have already happened and events which might happen might interact to produce certain outcomes. Such an activity could be used for prediction or planning. This is parallel to the activity of simulating or constructing mechanisms to predict and explain the phenomena they will give rise to, rather than just studying them once they have already given rise to a phenomenon of interest.

Even when used purely historically, such representations could be useful in the future, if they allow subjects to extract projectible patterns that can be applied elsewhere, especially on the basis of understanding how very specific combinations of types of events, in very specific orders, give rise to certain emergent phenomena. What would be extracted would not be laws, of course, and might be reliably useful only in quite similar scenarios. But understanding how the events in question gave rise to the outcome would allow the subject to have at least some ability to assess how projectible any predictions they make would be to new cases.

The kind of representations characteristic of 'historical understanding' are about more specific kinds of events than the more skeletal scripts, so we would not expect them to be

repeated in all details. But this does not mean they relate to *particular* events as opposed to event-types. Rather, they are specific in that they are represented in great detail, with so many features that they are unlikely to be exactly repeated. This is very different to being in principle unrepeatable, like particular events: they are multiply instantiable, even if they would not be expected to in fact turn out to be multiply instantiated.

It turns out to be quite difficult to specify a positive role for representing *particular* past events in this sense: we will return to this issue in Chapter 5. But we can see how representing *specific* events within an historical representation could be useful. This is because what can be learned from such representations is how events with specific features give rise to larger scale phenomena or surprising kinds of events, on the basis of well-understood, generalizable interactions between the small-scale events. Often these same interactions will not turn out to depend on all the specific features represented in the historical representation: part of generating understanding from that representation will be figuring out exactly which features of the specific events gave rise to the climax; and once this has been extracted, the lesson can be generalized appropriately (if not formulated as a universal law). A good example of this sort of phenomenon would be studying a specific chess game, really learning about and thinking through each move and how they gave rise to a certain result, and thereby extracting a more general strategy or kind of sequence of moves, which can be applied in many other chess games on the basis of the understanding thereby acquired.

It is difficult to estimate in advance just how significant either form of narrative representation (the script or the historical representation) is, especially in simple, technologically and theoretically unsophisticated minds: These ideas would need to be tested in simulations of

learning in dynamically complex environments. But we can at least see the potential for an important role of narratives in expanding learning beyond the limitations of Markovian models.

What sort of sophistication in temporal representation do these sorts of narrative representation require? Clearly, they involve holistic representation of structures of events, and we saw that there could well be something right in the idea that narratives should involve climactic, causal structure: historical understanding particularly involves focus on the complex web of *causes* leading to the production of *one specific phenomenon*. We have said that these can be generic rather than particular events. What about the other ideas about narratives, such as that they involve teleological structure or an emotional cadence, or about other sorts of sophistication in temporal representation, such as comprehensiveness in temporal frameworks, or domain generality?

These properties do not seem so important. Teleological structure does not seem necessary: everything I have said about historical understanding could be applied to understanding the events that led to an entirely unplanned phenomenon such as a rock-fall or the break-out of war or the break-down of social order due to miscalculation. Events with a teleological structure will be an important special case, as often it is important to understand how individuals achieved their goals; they just do not exhaust the relevant sorts of case here. Comprehensiveness in temporal frameworks and domain generality are not required either: the whole point is to just focus, in detail, on a relatively limited cluster of events, rather than a much wider framework. Large frameworks are important for formulating laws, not so much for specific understanding.

There is another potential use of narrative representation which could use emotional cadences in conjunction with these ideas about historical understanding, however. Narrative

representation could be used to overcome what at least at first glance seems like a very different limitation of RL.

We have been focusing on one sort of limitation of RL: its tendency to assume Markovian dynamics. This limitation is well-recognised and increasingly studied by computer scientists. However, there is a further limitation which we can extract from philosophical discussions of limitations the picture of rationality assumed by (amongst others) RL researchers. We discussed how RL can reliably produce the behaviour that rational reflection would produce. However, this is only true if we accept that rational reflection would produce the behaviour that maximizes streams of expected discounted utility given constraints. One might think there is more to genuine rationality than this, for two related reasons.

First, one might think that genuine rationality requires more than just producing the right behaviour. We might think, for example, that genuine rationality requires producing the right behaviour on the basis of a reflective understanding that it *is* the rational behaviour, which might entail a reflective understanding and endorsement of norms of certain kinds. I will not focus on this issue: it is an instance of a more general issue of self-reflectiveness and conceptual grasp and their role in rationality, which arises for all rational-like behaviour and cognition in animals, and the specific case of temporal cognition and diachronic rationality does not illuminate it.

Second, one might think that there are things which people value—and perhaps which all rational agents should at least decide whether or not they value—which cannot be captured by this framework. One such value would relate to the overall structure of one's life in time. Perhaps, when reflecting about rationality, one should not simply try to maximize expected discounted utility, but one should also reflect on and form preferences about the

narrative shape of one's life as a whole. This might matter to living a life which is *meaningful*. I will focus on the development of this idea in Velleman (1991). [172]

There are some features of the temporal structure of rewards over time which standard utility-maximization frameworks do capture. Temporal discounting (valuing rewards now rather than in the distant future) is built into such frameworks (and into many RL algorithms) with an explicit parameter. Consumption smoothing—a preference for having consumption of goods distributed evenly over time rather than living in dire poverty at some moments and extreme wealth at others, and the corresponding behaviour of borrowing during times of low income and saving during the good times—is implied by standard assumptions of diminishing marginal utility of goods (i.e. the assumption that even if every extra bit of consumption increases utility, the same increase in consumption has more of an effect on utility when one has very little to begin with compared to when one already has a lot).[173]

However, there seem to be cases where what real humans want is not smooth consumption, but rather a more elahorate shape to their life. For example, if we have to choose between success early in life followed by a slow decline, and success late in life preceded by a steady increase, many of us would choose the latter, even if they have the same overall sum of utility. These intuitions get stronger if we flesh out the cases (as Velleman does) in terms of a life of a troubled youth followed by self-improvement and a peaceful middle age, compared to a precocious youth followed by failures and subsequent misery. Another sort of case which brings out the relevant intuitions is the case of a life with a huge amount of utility at just one moment: if this spike is big enough, it should be enough to 'save' a life of much lower utility at all other

[172] See also Campbell (1994: 1, 63) and various other authors discussed in Velleman 1991.

[173] Friedman 1956.

times (and likewise a very sharp downward spike could 'spoil' a life of consistently fairly high utility).[174] None of this is predicted by standard expected utility theory.

Slote 1982's explanation of these intuitions is that we have a brute preference for more utility later in life than at the beginning. Velleman claims instead that the fundamental issue here is the contributions events in one's life make to its narrative structure, which he thinks determines the 'meaning', and hence the value of both the life overall and the individual events that compose it. The desire for a certain sort of narrative structure cannot be reduced to a desire for any particular pattern of utility over time, he thinks. Other features of events besides their utility will matter to the life's narrative, and different patterns, involving improvement or deterioration, variation and intensity or consistency, might be consistent with equally desirable good narratives. A consequence of this view is that one's whole life should not be evaluated by even a weighted sum of welfare at individual moments.[175] Although he does not seem to have any specific sorts of algorithm such as RL in mind, Velleman explicitly infers from this that "a life's value can never be computed by an algorithm applied to bare amounts of momentary well-being, or even to ordered sequences of such amounts, in abstraction from the narrative significance of the events with which they are associated".[176]

In favour of the view that narrative structure is important, Velleman points out that the intuitions that there is something insufficient with the standard sums of utility framework are

---

[174] Velleman (1991: 51).

[175] Velleman also emphasizes that neither can welfare at individual moments be derived by decomposing the value of the whole life they comprise, given his view. In both cases, Velleman's official view is not that one *must* not think that the value of one's life is equivalent to the sum of individual moments' welfare, and *must* evaluate one's whole life as a narrative, but that the latter sort of procedure is intuitively plausible and cannot be ruled out *a priori*. This would be enough to undermine RL algorithms, which have no room at all for such considerations.

[176] Velleman (1991: 60).

much stronger when the cases are fleshed out, especially if they are fleshed out with narrative elements. If you just described the distribution over time in mathematical or graphical terms, you would not get strong, stable preferences such as strong preferences for later rewards. Postponing rewards is not inherently good. Rather, you get the intuitions strongly when you describe the cases fully, e.g. as the *culmination* of a slow ascent or the earlier rewards being a *prelude* to sudden decline, or as *fleeting good luck*, and especially the later rewards being the result of *drawing lessons from one's misfortune*.[177] Velleman's explanation of these effects on intuitions is that learning from one's misfortune, not just having good luck after bad luck, changes the *meaning* of the misfortune and the later success. Meaning depends on the part the momentary well-being plays in an overall trend and specific narrative relations: on living out a story like a story of efforts rewarded rather than wasted.[178] Furthermore, this meaning is perspectival, and the perspective focusing on an individual moment is different to the perspective focusing on a whole life, just like perspective focusing solely on financial well-being is limited.[179]

We do seem to have the intuitive reactions to the kinds of cases Velleman is concerned with which he describes; and it is plausible that this is in part because we value living out a story. However, one might have two worries about this idea. Firstly, reducing a real person's life to one or even a small collection of narratives seems simplistic: any story you tell about someone's life will—even if that life is unusually unified—distort it and leave many great, many banal, and many painful actions, projects, and experiences out. Secondly, one might wonder

[177] Velleman (1991: 54).

[178] Velleman (1991: 53ff).

[179] Velleman (1991: 64ff.).

how deep these preferences for a certain narrative structure to one's life really go: one can get into a frame of mind where such preferences seem like he results of rather sentimentally taking the idea of someone writing a book about one's life with oneself as a hero too seriously, rather than the sorts of preferences we should be treating as core to rationality. Caring deeply about narrative structure can seem rather like making one's ambition in life having one's obituary printed in *The Economist*, as opposed to doing something intrinsically important and using whether *The Economist* would want to write one's obituary as a heuristic for that intrinsic importance. Or it might seem to involve the immaturity exhibited by Catherine Morland in *Northanger Abbey*, who misunderstands the events in her life and their significance because she interprets them through the frame of seeing herself as a heroine of a romantic novel.[180]

One way of responding to this sort of worry is to adopt a slightly different attitude toward the importance of thinking of a life as a narrative, drawing on the above account of narrative's role in understanding. Just as Velleman gives a deeper explanation of the preferences about temporal distribution of utility in terms of narratives than Slote, who treats such preferences as brute, we can give a deeper explanation of the preference for a life fitting a narrative, which Velleman treats as brute.

Narratives can be used to partially understand very complex domains featuring huge numbers of interactions between smaller-scale, better understood events giving rise to larger scale phenomena. Entire lives are very complex domains featuring huge numbers of interactions

[180] A closely related issue is how culturally specific preferences about the narrative structure of one's life are. It would be suggestive if consuming large numbers of novels, films, or folk stories of a certain sort leads people to care more about the narrative structure of their lives. We do know that there is variation in both gender and culture in the kinds of narrative individuals tell when describing their own experiences and lives (Nelson & Fivush 2004, Wang 2011). If the strength of preferences about narrative structure is easy to manipulate and highly contextual, this might be (defeasible) reason to suspect that leaving them out of an account of rationality would not be such a deep omission.

between smaller-scale, better understood events giving rise to larger scale phenomena. We pursue many small and large projects in parallel. We pursue projects in different domains, over different timescales. These projects sometimes interact in complex, hidden and delayed (hence non-Markovian) ways with one another, with particular actions we take, and with small and large events which happen to us. Fully understanding a life, in a way which is presumably required for fully grasping its meaning or having considered preferences about how it all should fit together, would be beyond our ken even if we could remember and can represent every project, action, or event. Grasping a whole life and pulling out its important aspects is difficult; narratives simplify certain strands within a life so that they can be grasped more fully, and so that very specific patterns can be extracted which may apply to other parts of that life. The feeling of meaningfulness from a series of events fitting a narrative form comes from grasping that series of events more fully and being in a better position to treat those events as meaningful.

There are two ways this sort of view could be developed. One could think of meaning as an emergent, objective phenomenon, which the subject of the meaningful life might only *grasp* partially or even not at all; on this view, narratives could provide a kind of partial grasp of the meaning of a life. Or one could think of *meaning* as depending on the subject of a life thinking of some pattern in their life as meaningful: Velleman (1991: 70f.) explicitly adopts a version of this view. On this approach, narratives could provide a kind of partial grasp of the complex patterns involved in one's life, which would otherwise be missed, and thereby allows one to have preferences about these complex patterns.

Can this sort of view explain everything Velleman can explain? Why, on this view, would we prefer rewarded struggles to dumb luck or to early success followed by failure? The answer will have something to do with preferences for more robust patterns, for the underlying

mechanisms involved in rewarded struggles which are inactive, ineffective or non-existent in cases of dumb luck. A narrative about someone being rewarded for their struggle will help one extract certain patterns in the particular events making up that struggle and gradual success: patterns relating to an individual's character and habits, and their consistency in certain kinds of choices. These character traits might be valued in themselves, or valued because they make the success non-lucky, and we tend to value non-lucky success in general, not just in narrative contexts. We prefer to know things rather than to be in Gettier cases of stopped clocks.

One feature of the ideas about historical narrative understanding above was that such understanding focuses in the first instance on understanding the events that led to one specific phenomenon. How does this square with the idea of understanding lives in narrative terms? Surely, we do not think that the narrative should end with a climax (it can, of course, in a heroic death; but need not). This points the way to what I think is an advantage over Velleman's view: on the view that the importance of a narrative understanding of a life is important because it provides understanding, we can expect that a single life can be captured by multiple narratives, emphasizing different climaxes, different projects fulfilled or frustrated, different strands in a person's life. Some of these narratives may end part-way through someone's life but capture some of the most important sources of meaningfulness in that person's life. We do not have to think of the life as a whole as fitting into one unified narrative. Velleman does not explicitly commit to the view that we should, but he does think that the subject's perspective on their life as a whole is particularly important to that life as a whole's meaning, and ties this perspective to narratives. And the view he ends up with does intuitively seem simplistic precisely because it does not make clear why having multiple narratives to one's life might be a good thing.

One might wonder what any of this has to do with *cognitive* significance. Velleman's view is about ethical significance and perhaps a certain kind of rationality: he is not making claims about how the mind works or what its capacities are. The answer is two-fold. First, this is a concrete case where an account of the cognitive significance of some capacity—the account of the significance of narrative representation in terms of its contribution to a certain sort of understanding—can help in grounding accounts of other sorts of significance, thereby illustrating the significance of cognitive significance that is a running theme of this dissertation. Second, it is often thought (especially by advocatess of Bayesian frameworks) that showing that RL and other learning algorithms produce results which are in some sense rational sheds light on how they actually work and why the mind uses them. A natural suggestion if one has this attitude to computational explanation would be to try to build algorithms which in the long run will produce rational behaviour of a kind which matches the view of rationality Velleman or I have proposed. My view of narratives' role in rationality in accounting for understanding of one's life as a whole and hence allowing assessment of its meaning is particularly suggestive of how one might go about doing this: instead of simply having an objective function of maximizing discounted rewards, such an algorithm would need to form representations of possible life-courses and estimate the value of these using different narratives. It is not impossible that such an approach would help model actual decision-making in complex choice scenarios.

What kinds of temporal representation are required for narrative-based rationality? Since the idea is based around the sorts of processing which historical understanding can produce, it will require at least as much sophistication as that sort of understanding: holistic representation of sequences of events, their causal structure and climactic structure in giving rise to a phenomenon of interest. Some of the other proposed features of narratives seem more relevant

to the specific project of understanding one's life for the purpose of assessing its meaning than they do to narrative understanding of complex patterns generally. In particular, narratives involving a teleological structure will be particularly relevant: some of the most important features of a life to its meaningfulness will be the pursuit of intentional projects. Velleman (2003)'s ideas about narratives having an emotional cadence could also be argued to be relevant here, at least on a certain sort of view of the emotions. If one takes the (controversial) view that emotional responses are a good way of determining whether something is meaningful in the relevant sense, and what its meaning is, then narratives which are well-suited to provoking emotional responses will be particularly relevant to understanding meaningful parts of one's life.

Once again, we have a divergence between narratives and more comprehensive temporal frameworks in terms of the reasons for their cognitive significance. Understanding one's entire life, mapped out on a comprehensive framework, is an intractable problem, which is why narratives become important. This implies that a comprehensive framework is not required. There will need to be a certain amount of range involved: at least some of these narratives will stretch over one's entire life: but one need not represent the whole structure of time to do this. As for domain generality, representing times as opposed to events, or even particular as opposed to specific but multiply instantiable events, these do not seem to be required for a narrative approach to understanding one's life.

An upshot of this divergence between temporal frameworks and narrative understanding is that it becomes somewhat more plausible that animals might be able to have the relevant narrative understanding, at least on a small scale, understanding simple sequences of events and their significance. If the whole point of using narratives to capture the meaning of one's life is human limitations—an angel, after all, could simply appreciate the entire pattern of all the events

in a life, without resorting to narratives to extract the most important parts—then it is wrong to rule out different non-zero levels of sophistication here. We will discuss the plausibility of animals having sophisticated kinds of temporal representation in the next section, but before we do so it is worth considering the stakes. Velleman explicitly assumes that non-human animals (he singles out cows) are not capable of understanding the kinds of narratives of a whole life he is interested in, and explicitly argues from this assumption to the conclusion that while cows have momentary welfare, it does not make sense to attribute good or bad *lives* to cows.[181] And from this, he concludes that

> "in relation to an animal's interests, as I have now described them, the traditional Epicurean arguments about death are correct. That is, there is no moment at which a cow can be badly off because of death, since (as Lucretius would put it) where death is, the cow isn't...a person can care about what his life story is like, and a premature death can spoil the story of his life Hence death can harm a person but it cannot harm a cow."[182]

There are various places one might challenge this argument, such as the move from thinking that cows cannot think about the narrative structure of their life to the view that the narrative structure of their life does not matter *tout court* (Velleman defends this move at some length), or the further move to claiming that death is not a bad for them.[183] But the view of the importance of narrative structure here provides a different reply: animals *could* have at least some degree of the sort of understanding of broader features of their temporally extended lives that actually matters. They could have, and think about, somewhat sustained projects, without needing to grasp their entire lives as one narrative. They could use simpler kinds of narrative to achieve this, with fewer events and simpler causal structures than those humans represent. Or

[181] Velleman (1991: 68ff.).

[182] Velleman (1991: 71).

[183] Korsgaard 2018 Ch. 2 provides a detailed account of how animals with limited cognitive capacities can be subjects of a temporally extended life, in a sense relevant to broadly Kantian ethics.

they could use aspectual representations to sustain ongoing projects without understanding their temporal dynamics, representing e.g. that they have the property of being a mother (having-had-offspring), would like to be raising their offspring, and hoping that the offspring continues to have the property of being alive into whatever future they can foresee. None of this requires understanding their *entire* lives as a *single* narrative, let alone comprehensive temporal frameworks, or an understanding of the nature of time. But humans who do try to bring their entire lives under *one* narrative are probably making a mistake, diminishing the richness which could be in their lives, even if they cannot quite grasp how all the different strands fit together.

4.7 Do Non-Human Animals Have Sophisticated Kinds of Temporal Representation?

The above proposals about the cognitive significance of different kinds of temporal representation could be tested. There would be two main stages of testing: using simulations to study the conditions under which the different kinds of temporal representation take on the roles described in the text; and behavioural and neural tests to see if these sorts of processes are implemented in different animals. But can we say anything based on *current* evidence or a priori considerations to suggest whether different animals have these kinds of representation? Here I will show that two tempting arguments, one purporting to show that holistic, comprehensive temporal frameworks are widespread, and the other purporting to show that holistic temporal representations (including frameworks and narratives) are unique to humans, are both inconclusive; in Chapter 7 I will consider some strongly suggestive neural evidence for limited kinds of narrative representation in several animals.

First, one way of arguing in favour of widespread representation of temporal frameworks is to claim that they are necessary for all or nearly all other kinds of representation. These arguments will often rely on a claim to the effect that the kind of coordination required for fundamental operations like binding different features into unified representations of objects,

coordinating action and perceptual representations, or distinguishing external objects from internal features, requires representing all these different features and actions with spatiotemporal (or at least temporal) coordinates.[184] All versions of arguments like this which I know of either use 'representation' in a deflationary way, or fall foul of the sorts of considerations in Chapter 2. That is, the functions they ascribe to temporal representation *could* be performed through such representation, but it is also relatively easy to specify ways in which they could also be performed by inflexible, anti-representational forms of temporal coordination. As Chapter 3 showed, this does not mean that such considerations provide *no* evidence for temporal representation, but the evidence is pretty weak, as there is little to favour these sorts of hypothesis over anti-representational hypotheses, and plenty of reason to suspect that the kind of flexibility with respect to time which the representational hypotheses posit is not present in all the animals with other kinds of flexibility.

Rather than going through such arguments and their pitfalls in detail, which would largely involve recapitulating points from Chapter 2, I will consider a stronger argument for holistic temporal framework representations specifically in scrub jays, which faces similar problems but not as decisively.[185] This argument starts from the findings discussed in Chapter 2 of scrub jays' sensitivity to duration since they cached various bits of food, and the considerations discussed there for why we should think that such sensitivity is representational. The question now is what form that representation takes: Is it a representation of a duration, representing a cache as THREE DAYS OLD? Or do scrub jays represent a holistic

[184] Montemayor (2010: 5ff.), Russell & Hanna (2012: 32ff.). As is explicit in Russell & Hanna 2012, this sort of argument has roots in Kant (as do many of Campbell's arguments above): see Guyer 2007 Ch. 2 for a survey of Kant's arguments along these lines.

[185] Gallistel & King (2009: 156f., 213ff., 266f.).

framework of times, and represent each cache with its coordinate in that framework—something more like 3PM, SEPTEMBER 13TH 1972 than like 3 DAYS AGO?

One reason to suspect that they might have framework representations is that this might make for far more efficient computations and storage. This is because jays have such memories for numerous caches (at least tens, but the numbers could well be in the thousands).[186] It seems inefficient to require a new clock for each cache-memory—a stopwatch that is started when I bury a nut, then another stopwatch for the next nut, then another stopwatch for a worm, and so on. But duration representations that are not based on a holistic temporal framework seem to require such a proliferation of stopwatches. The framework-based strategy, by contrasts, allows the temporal coordinates of each event to be stored. Given just the temporal coordinate, when the jay needs to know the duration since caching, the jay can do the simple computation of subtracting the cache coordinate's time value from the present moment's time value.

There are two problems with this argument, both relating to its use of the notion of efficiency. First of all, Gallistel and King give only intuitive reasons to think this approach is more efficient than alternatives, but do not explain in detail why this should be. It is intuitive that each individual clock would require extra resources. But it is not clear just how many extra resources they would require than the extra resources required to develop and maintain a coordinate system. And it is not clear that the alternative really does need to posit numerous individual clocks in the sense that uses more resources. For example, an alternative implementation would be to have one regular cyclical process which sweeps through each of these representations every period, updating each representation's duration-estimate by +1

[186] Gallistel & King (2009: 217, 268).

automatically. Perhaps this would be considerably more inefficient than the coordinate system, but it does not seem safe to assume this without some sort of empirically based estimate.

The second problem is that it is not clear how far we should be assuming scrub jays will use the most efficient computations for this task. Perhaps there are very strong selection pressures favouring efficiency here; but perhaps the relevant selection pressures are not strong enough to have reached the optimal solution yet. Or perhaps these selection pressures drove the scrub jays to get as efficient as they could at using their existing style of clocks, but then they got stuck at a local maximum, with no way of being selected to get to the system of temporal coordinates without a new *kind* of machinery being added, a new kind of machinery that would require an extremely rare set of mutations. We do not know.

Given all this, it seems unreasonable to take the scrub jay evidence to constitute strong evidence of a holistic temporal framework representation. We should certainly have higher credence for scrub jays having temporal frameworks than many other animals, but we should not be anywhere near to a credence of one.

The best arguments to speak positively *against* animals having holistic forms of temporal representation go via the claim that such representations would require language, and the assumption that non-human animals do not have language.

Why think that holistic forms of temporal representation require language? One source of evidence for this claim is evidence that for humans to learn to represent certain kinds of narrative and temporal frameworks, we need at least some language in place. Nelson & Fivush 2004 argue this case in detail. Their general view is that narrative understanding and temporal understanding generally emerge gradually as part of a broader suite of capacities which are all mutually reinforcing (with the development of any one of these capacities leading to

developments in others). Such capacities include autobiographical memory, a concept of self, mind-reading, causal understanding, language, and culturally specific beliefs about topics, like which features are important to remember and the relationship between self and society. They present evidence that in the first year, infants start remembering and imitating only individual actions, before gradually getting better at recalling longer action sequences and then (in the second year) learning script representations but without being able to remember novel experiences. Around 18 months old, infants begin to refer to the past verbally, but these references are only fleeting & fragmentary, usually just referring to completed actions or familiar routines. Crucially, early verbal discussions of the past and of narratives typically occur during conversation with adults, who offer extensive framing and help with complex temporal relationships including tense. Children gradually improve over the next few years on both the complexity of the temporal relations they can talk about (including more complex, unique narratives, tense, and ultimately time lines), and their ability to do so unprompted.

Nelson and Fivush (2004: 496ff.) give a few reasons to think language is crucial to the process of developing complex forms of temporal representation. First, 2-5-year-olds' memory for features of events, including of temporal features like order, is improved by greater vocabulary and conversations about those features with adults, especially when adults frame the event linguistically at the time of encoding (as opposed to their vocabulary and which features are asked about at the time of retrieval). Second, evidence of understanding of complex narratives (narratives with more complex actions, temporal markers, explicit links to a wider context, an evaluative stance on what occurred, and the ability to deal with backwards sequences) only arises when preschoolers are able to linguistically express and discuss these complexities, with the degree of complexity in their re-enactments never outstripping the degree

of complexity in their linguistic recountings of narratives. Finally, differences in how parents linguistically engage with children determine how good those children are at constructing narratives, and the details they include (which shows systematic variation both by culture and gender, as well as with parents' individual styles).

One worry for several of these phenomena is that the direction of causation is often very hard to tease out. Nelson and Fivush recognize this with their claim that multiple capacities here are mutually reinforcing. But in several of the studies they cite, yet another possibility is that some other variable (such as general intelligence) explains parallel developments in *both* temporal representation and language abilities, rather than one of these explaining the other. To answer this sort of objection, one could use all manner of statistical tricks and careful experiments, but most important will be getting a clearer idea of the reason why temporal understanding is linked to language.

Nelson & Fivush (2004: 494) suggest three such reasons. Firstly, language might provide the organizational and evaluative forms characteristic of autobiographical memory. This seems to mean that it provides a format of representation, or representation of a certain kind of structure, that allows for certain kinds of abstraction and comprehensive representations.[187] Secondly, language allows for dialogues, which helps children develop skills in forming organized representations of past experiences by prompting them in the right ways when they fail. Finally, these dialogues not only prompt children in organizing specific experience, but help facilitate an emerging awareness that memories are representations of past

[187] One way to supplement Nelson and Fivush's ideas here would be to appeal to Carey (2009)'s development of the idea of 'Quinean Bootstrapping', in which linguistically provided structures are crucial to learning about things like a comprehensive number-line (as opposed to only being able to count up to a definite limit).

events which can be evaluated from multiple subjective perspectives, by virtue of showing the adult's alternative perspectives.

However, there is a basic problem with wielding developmental evidence as an argument that language is necessary for sophisticated temporal representation. This is that the particular developmental path humans need to follow to get to temporal representation may be a contingent feature of humans, which cannot be extrapolated to other animals. As discussed in §3.5.3, we need to be extremely careful about extrapolating from humans to other animals, especially for cognitive (as opposed to perceptual) traits and especially those relating to flexibility, as humans may well have faced rather different selection pressures to other animals. And this is in large part because humans have a lengthier juvenile period, and hence learn things differently to other animals.

In the case of sophisticated temporal representation, while according to Nelson and Fivush humans are not able to represent complex sequences until the age when language starts developing, we have independent evidence that at least some non-linguistic creatures represent complex sequences,[188] and care about other times and are capable of appreciating that their future perspectives will be different to their current perspectives.[189] We also know that representation of complex structures in a holistic rather than item-by-item, relation-by-relation fashion does not require language in other animals, given what we know about how baboons represent their social structures.[190] Several of the studies cited about remembering more given

[188] See the evidence relating to the simultaneous chain paradigm surveyed in §2.7 above, and the evidence for representation of sequences by hippocampal replay in §7.5 below.

[189] Correia et al 2007; De Kort et al 2007; Shettleworth 2007; and Clayton et al 2008 show that scrub jays cache food now in a way which depends on future rather than current preferences, for example taking into account that even though they have just had many peanuts so don't want any more peanuts now, they will like peanuts again tomorrow.

[190] Cheney & Seyfarth 2007, Camp 2009.

better linguistic abilities suggest not that linguistic abilities improve structured representations or temporal representations specifically, but rather improve memory generally; but we know animals can remember many things, in complex structures, without language. And we know that there are other cases, like the ability to read human emotional expressions in subtle ways, where human learning depends on or is greatly enhanced by social, linguistic scaffolding, but animals are still able to learn the capacity in question without this scaffolding. We can imagine all kinds of reasons why humans would rely on linguistically-based structures and have lost non-linguistic structures that other animals use during the course of evolution. For example, language may allow for more general representations, or make it easier to learn and re-shape new formats of representation, or easier to integrate with external, cultural-technological means of measuring time. All of this means that even if it is true that humans can only achieve sophisticated kinds of temporal representation given language, this should only raise our credence in non-linguistic animals' lacking sophisticated kinds of temporal representation a little.

The upshot of all this is that the correct attitude to adopt towards the claim that animals do not have temporal frameworks—and towards the claim that they do—is (for now) one of extreme uncertainty. Numerous, quite different hypotheses remain live options. Each has some weak considerations in its favour. None merits high credence.

4.8 Conclusion

There do seem to be kinds of sophistication in temporal representation which are highly significant. Both rest on holistic structures, and both are significant not because they allow a full understanding of the dynamics of the world, which are so complex that they are well out of reach of most, perhaps all animals (perhaps some humans are an exception but even this is unclear).

Rather, they allow for ways to cope with that complexity more effectively. Representation of somewhat comprehensive temporal frameworks allows for the representation of essentially dynamical kinds of entities and hence for picking up on and dealing with some broad dynamical patterns, including non-Markovian dynamics. Narrative structures allow for understanding specific (though not necessarily particular) events in ways which can help with learning new kinds of dynamical entities or otherwise being able to model phenomena which arise in incomprehensibly (to the animal) complex ways from better understood local dynamics. Both sorts of complex representation come in degrees. And both are in principle available to non-linguistic creatures, despite there being no strong evidence yet either for or against such creatures possessing them.

All this suggests a number of further issues. One is how unique these issues are to time: do they arise for holistic representations of other sorts of frameworks? We can also represent spatial frameworks, frameworks of numbers, colour wheels, and (in our understanding of modality) structures of possible worlds. Would these sorts of frameworks, and especially analogues of narrative structures or essentially dynamical entities, be cognitively significant for similar reasons to the temporal representations discussed here? One reason to think that a great deal of the temporal understanding described in this chapter is unique to time is the close connection between time and causation we have repeatedly come across in this chapter. However, these issues would merit further discussion, as would the issue of how the ontogeny of these different sorts of systematic understanding relate to one another. We have already touched on the idea of Quinean Bootstrapping, and linguistically provided structures being used to provide scaffolding for learning temporal structures. But it is also possible that temporal structures are used as scaffolding for learning other forms of framework, such as systematic

thinking about modality (note, for example, that we very easily slip between ‘must’ and ‘always’).

However, this dissertation will continue to pursue the question of the cognitive significance of temporal cognition of different forms. I now move from discussion of representation of time (its nature and varieties, its cognitive significance, and evidence for its instantiation in different animals), to discussion of episodic memory. Many have claimed that this too depends on temporal representation, although drawing on Chapter 2’s points about anti-representational temporal coordination, I will argue against this. I will, however, argue that like sophisticated forms of temporal representation, episodic memory is cognitively significant, could come in many varieties, and could well be instantiated in many animals. Furthermore, unlike narrative representation and representation of dynamically complex entities, the significance of episodic memory could involve a capacity to underpin completely general, unlimited forms of learning.

Chapter 5: The Significance of Episodic Memory

5.1 Introduction

I remember the day I adopted my dog. I don't know whether he remembers it, but I would like to know. I am not alone in this. A sizeable, sophisticated experimental literature aims to determine which species have episodic memory—roughly, the kind of memory we have when we remember events as we experienced them, as opposed to recalling facts.[191] However, this literature is dogged by a problem. Whenever researchers find that some species has a system with some of the features of episodic memory—such as recalling when, where, and when some event occurred—the question arises as to whether this is enough. Is this genuine episodic memory, or merely episodic-*like* memory—memory which shares a few features with episodic memory but is fundamentally different?[192]

We should not think that a system constitutes episodic memory only if it has *all* the features that human episodic memory has, any more than we should think that birds' eyes are not genuine eyes because they detect a different range of frequencies to humans'. Rather, the issue is whether animals have systems with the *core* features of episodic memory. The problem is that thinkers have not pursued any systematic way of determining which features are core to episodic memory, and which are merely associated with the human version of episodic memory. So we face questions like: To count as having episodic memory, must dogs have a specific kind of consciousness? Mental imagery? A certain neural structure? And it is unclear not just what the answers to these questions are, but what considerations should be used to assess them: intuitions? Introspection? Or some other method?

---

191 Chapter 7 will review the main strands of evidence for episodic memory in animals in more detail.

192 See discussion in e.g. Clayton & Dickinson 1998; Tulving 2004; Suddendorf & Corballis 2007; Allen & Fortin 2013.

Disputes with this general form arise for the attribution of any mental or social kind to other species. For example: what does *language* require? Flexible communication? Communication of a structured message? With a subject and a predicate? Recursion? More specific grammatical features of human natural language?[193]

In this chapter, I address the issue of which features are core to episodic memory, and to do so, I defend a method which can be used to answer other questions of this kind.

I argue that we should treat a set of features as core to a mental kind if they combine in a *cognitively significant* way: roughly, if together (and only together), they make a big difference to how the mind works. Drawing on ideas from artificial intelligence, I show that a particular set of episodic memory's features meets this description, and so should be thought of as episodic memory's core: imagistic representation, of a particular past event, on the basis of a certain kind of memory trace. Episodic memory defined by these core features could play a key role in a form of learning which is unconstrained in an important sense—it provides the animal with the ability to continue improving its model of its environment without limit. This account of episodic memory and the methodology used to defend it offer useful ways of redirecting many research questions in philosophy and psychology about the nature of episodic memory, its role in learning, and its evolutionary function.

§5.2 explains how cognitive significance can be used to individuate kinds of mental states, while §5.7 compares this approach to other approaches to individuating mental states, once the account of episodic memory and its significance is in hand. §§5.3-5.6 apply the framework to the case of episodic memory. §5.3 intuitively motivates a preliminary definition of

[193] See e.g. Chomsky 1980, Hauser et al 2002, Slobodchikoff et al 2009, Watumull et al 2014. De Waal 2016 describes numerous examples of animals being shown to have something resembling an important human mental trait prompting ensuing debates about whether these are '*really'* the same trait.

episodic memory; §5.4 lays out challenges for showing that this picks out a cognitively significant kind; §5.5 illustrates one way of answering that challenge; §5.6 answer natural objections to the account.

5.2 Individuating Psychological Kinds for Cognitive Significance

Sometimes it is not clear which features of a psychological trait are core, and which are merely features of that kind's manifestation in a particular species. In such cases, I advocate searching for cognitively significant combinations of features. §§5.3-5.6 will give a detailed illustration of the approach in the case of episodic memory, but we can first sketch a general procedure.

One approach to deciding which features are core to a trait would be to rely on intuitions. We do want to end up with a notion that has some correspondence to pre-theoretical intuitions. We do not want to change the subject. But relying on intuitions *alone* tends (at least in the case of episodic memory) to give rise to a stalemate in the face of conflicting intuitions, and does not explain either *why* just *these* features are so important, or why these features hang together to define a kind.

Instead, we can start with a collection of intuitive candidate core features, and consider the cognitive significance of these features individually and in different combinations. This involves studying what different kinds of mind could do with and without the relevant traits. To study such questions, empirical research will be relevant—where we can find actual examples of minds with the right combinations of traits. But simulation and modelling will have a large role to play, both for guiding and interpreting the empirical research, and for studying a wide range of relevant possibilities in a carefully controlled way.

At the end of this process, we will find that certain features are individually very cognitively significant—no matter what other features they combine with, any trait with that

feature will be cognitively significant. In such a case, we could define a psychological kind which has just one core feature, but given that we already have that feature in our ontology, we do not seem to add anything to our understanding by doing so. By contrast, we may find that certain traits are cognitively significant in a way that is due to a certain *combination* of features. In such a case, we have a kind with multiple core features which hang together in an interesting way.

Given a set of core features, we can make cross-species comparisons: instead of asking just 'do corvids have episodic memory?', we can ask if they have a trait that instantiates the specific set of episodic memory's features we have found to be distinctively significant. And corvids may have a version of episodic memory with that same core but whose non-core features are very different.

One might worry here because cognitive significance is defined relative to a kind of mind. In our consideration of how different features play out, which sorts of minds should we be considering adding them to? Will this method result in species-relative psychological kinds which are defined very differently (with corvid-episodic-memory being defined differently to human-episodic memory, for example)? And would this not make useful cross-species comparisons impossible?

The best approach in the face of such issues is to explore the cognitive significance of proposed collections of features in a range of minds—human-like minds, corvid-like minds, minds somewhere in between. We might find a certain kind of robustness, where a certain collection of features is always cognitively significant. If so, we have a further reason to treat this as an interesting psychological kind. If, by contrast, we find that one combination is significant in a corvid-like mind, and a different combination is significant in a human-like mind,

that seems like good reason to think that there are two kinds worth studying. Admitting multiple kinds, however, does not make interspecies comparison impossible: we can still compare both which species have traits with the relevant combinations of features. It might turn out that some species have episodic memory—a trait with all the core features of human episodic memory—but in a context which means that episodic memory is not significant in their mind in the way that it is for humans. Thereby finding that they have one of the key components of an important broader capacity or set of capacities would be interesting, both in itself and for our understanding of the evolutionary history of that broader capacity, as it would show the possibility of that broader capacity evolving via first evolving the component in question as an intermediate stage.

We have an approach to individuating psychological kinds which is guaranteed to carve the mind into traits which can make a big difference to how the mind works. And given the considerations above, these kinds will also be useful for other kinds of theorizing, such as tracing the evolutionary history of the mind and animal ethics. This approach provides for kinds at the appropriate level of abstraction to allow for cross-species comparison. And it leads to extremely fruitful lines of investigation. It brings into focus important issues like what exactly can be done with a trait of a certain kind and how. And it provides a systematic framework which both relates different lines of simulation-based and empirical research to show how they jointly shed light on these questions, and suggests which new variations on such research would be useful. This will become clearer in light of the more detailed application of this framework to episodic memory in the following sections.

5.3 Episodic Memory as Imagistic Memory of Particular Past Events

We can now apply this approach to episodic memory. I will begin with intuitively motivating a pair of core features (I will add another feature, memory traces, in §5.5): representation of particular past events, and an imagistic style of representation.

It is common to contrast episodic with semantic memory.[194] There is an intuitive distinction here, often marked in language by a 'that'-clause, e.g. in 'remembering my party' and 'remembering that my party was on a Tuesday'. It is controversial exactly what the fundamental difference is here, but one attractive option is that episodic memory constitutively involves mental imagery of the event the memory is about.[195]

Episodic memory is not, however, the only form of long-term memory which involves imagery. Representing *particular* events in an imagistic way contrasts with imagistic memories for *generic* event-kinds. A memory of what it is like to ride a roller-coaster may involve imagery of the different motions one undergoes, what it looks like from the top, and the sound of screaming riders into a single unified representation that is rather like a representation of a particular event.[196] But it may nonetheless be a memory of what it is like in general, based on amalgamating annual trips to the theme park, rather than of what it was like on a single occasion. Call this 'generic imagistic memory'. Generic imagistic memory may well share many neural mechanisms with episodic memory, and individuals may sometimes introspectively confuse the

[194] The 'semantic'/'episodic' distinction was introduced to contemporary psychology by Tulving 1972, though related distinctions were earlier made with different terminology by Bradley 1899, Russell 1921, Broad 1925, Bergson 1991, and many others - See Brewer 1996 for discussion of some of the history and alternative terminology here. There are a number of other phenomena that can be called 'memory', but which will not concern us here. Tulving distinguished semantic and episodic memory from 'procedural memory', something close to philosophers' 'knowledge how' (Ryle 1949). Furthermore, we will be focused here on long-term memory, rather than short term forms of memory.

[195] Many historical figures, surveyed in Brewer 1996: 23, and contemporary thinkers, such as Hoerl 2001; Schacter et al 2007, 2012; Hassabis et al 2007; Buckner & Carroll 2007; Boyer 2008; Rubin & Umanath 2015; Michaelian 2016; Hopkins 2018, have argued on the basis of correlations, intuition, or phenomenology, that imagery is important to episodic memory, but none has provided a cognitive significance based argument.

[196] This terminology helps us get a handle on one intuitive case of generic imagistic memory, but it should not be taken to imply that imagistic memory can only be about conscious experiences.

two, but the two are functionally very different, with different conditions for encoding and retrieval, and different kinds of cognitive significance.[197]

Intuitively, then, episodic memory has at least two core features: imagistic representation, and being about a particular event. Do these features combine in a cognitively significant manner?

5.4 Challenges to the Significance of Episodic Memory

It might seem obvious that episodic memory is cognitively significant. However, many of the cognitively important roles we might point to for episodic memory are roles which either semantic memory about a particular event, e.g. memory that my party (a specific event) was on a Tuesday, or generic imagistic memory, could very easily play instead.[198] Thus, it is non-trivial to say why combining even the two features presented so far would be significant. This section illustrates this by showing why some of the most natural suggestions for accounts of the significance of episodic memory fail.

It is initially plausible that episodic memory could be transformative because it tells us about episodes from our past. However, semantic memory does this too. Indeed, semantic memory may do so more usefully by abstracting away details which are unlikely to matter in the future.

To respond to this point, we should not simply appeal to things that imagery can enable in general. For example, it is plausible that imagery has special force for emotions—imagining

[197] Similar distinctions can be found in philosophers like Campbell 1994; Burge 2011; and in the psychological literature on script memory discussed in Chapter 2. Some, including Rubin & Umanath 2015, explicitly do not require episodic memory to relate to specific events.

[198] Many of these issues also arise for the evolutionary significance of episodic memory, which has been much discussed recently, e.g. by Suddendorf & Corballis 1997, 2007; Klein et al. 2002; Buckner & Carroll 2007; Schacter et al 2007, 2011; Boyer 2008, 2009; Rasmussen & Berntsen 2009; Allen & Fortin 2013; Templer & Hampton 2013; De Brigard 2014; Michaelian 2016; Mahr & Csibra 2018; Mar & Spreng 2018; and Rau & Botterill 2018.

your spouse cheating has more impact than thinking about it as a possibility in a non-imagistic way.[199] And imagery allows planning of detailed movements by representing the spatial layout of a scenario in the same format that the senses will present it during action.[200] But episodic memory seems less relevant than generic imagistic memory.

Instead, we can return to the idea of episodic memory telling us about the world, and ask why telling us about particular events could be significant for some forms of learning, then see if imagery has anything distinctive to add to this story. It is plausible that, given that generic imagistic memory and semantic memory for particular events abstract from various details, remembering individual events should be useful for learning by providing something like raw data. We will see that this intuition has something right about it, but to appreciate why, it will be useful to recast these ideas in a framework where we think about learning about the world as constructing a statistical model.

In this framework, the basic challenge for explaining the importance of particularity is that a model of the world should not include everything the animal has experienced. Typically, only a few general facts will be useful, and most details about past events can usually be ignored. For example, suppose you want to collect nectar effectively. A good way to do this might be to learn average nectar levels at different locations, throwing away a great deal of other information about those locations (the fact that you once heard a sparrow chirp while you were there, say) as irrelevant.

Perhaps remembering particular events is nonetheless useful for learning the relevant averages? You could visit a location *t* times and use your remembered, particular nectar level Y_i from each visit *i* to calculate the mean $\overline{Y_t}$:

---

[199] Boyer 2008.

[200] Pacherie & Haggard 2010.

(Eq. 1) $\overline{Y_t} = \frac{\sum_{i=1}^{t} Y_i}{t}$

However, an alternative method would be to update $\overline{Y_t}$ each visit, throwing away the particular data that went into your running average as soon as it is incorporated into that average, setting $\overline{Y_1} = Y_1$ then updating this initial estimate using the following formula:[201]

(Eq. 2) $\overline{Y_t} = \frac{(t-1)\overline{Y_{t-1}} + Y_t}{t}$

(Eq. 1) and (Eq. 2) are guaranteed to give the same results, but (Eq. 2) does not require memory for specific visits. The estimate $\overline{Y_{t-1}}$ already incorporates Y_{t-1}, Y_{t-2} and so on. Although it initially looked like memory for particular events might be useful here, it turns out not to be — if anything, it is an inefficient waste of storage resources.

To see a use for storage of particular events, we need to complicate the story a little. Notice that (Eq. 2) is only useful if you have been updating an estimate $\overline{Y_{t-1}}$. If you suddenly need to estimate that mean for the first time, (Eq. 2) will be useless; and if you have been throwing away the data, you will be stuck.[202] But if you *have* been storing the particular nectar levels from each occasion, you could use (Eq. 1). There is a general lesson here which we will return to: storing particular data points (which episodic memory, among other kinds of memory, does) allows us to perform unanticipated operations on our data, whereas collapsing data into generic memory is more efficient, but only for anticipated computations.[203]

[201] See Sutton & Barto 2018 for many examples like this.

[202] This is a slight over-simplification: as Nagy & Orban 2016 point out, you can *estimate* what your data was using your current model, and use this to fit a new model; it's just that your new model is unlikely to look better than your old model under this procedure, as you will have smoothed out the parts of the data that would have favoured the new model.

[203] Authors from Klein et al 2002 to Mar & Spreng 2018 have suggested that episodic memory could be used in unanticipated calculations without linking this idea to generalized learning like the below. Nagy & Orban 2016 do make this link, but do not discuss imagery.

Nothing has been said yet about imagery. Calculating means unexpectedly can be done with mere lists of semantically represented data-points. §5 shows why imagery is nonetheless relevant to such learning.

5.5 Indefinitely Complex Models

Creating a model that allows prediction of some target variable is not limited to estimating a mean. Multiple variables can be used. We can think of this in terms of regression analysis.[204] A regression model is an equation giving our target variable *Y* as a function of n variables $X_1,..., X_n$ and at least n+1 coefficient parameters $a_0,...,a_n$. The simplest case will look like this:

(Eq. 3) $Y = a_0 + a_1X_1$

Here, Y depends on just one variable. This might be a model that predicts nectar levels solely on the basis of distance from a point at the centre of the garden.

Given a regression equation like this, we can fit it to the data. That is, observed combinations of values of *Y* and $X_1,..., X_n$ on particular occasions can be used to estimate the values of the parameters $a_0...a_n$ for the environment we are in, using methods that are similar in spirit to (Eq. 1) or (Eq. 2).

We might want to complicate our model of the world in a variety of ways. We might want to add extra variables. Perhaps both location and time of day are relevant to nectar levels:

(Eq. 4) $Y = a_0 + a_1X_1 + a_2X_2$

[204] This particular formal framework is used largely because it allows a more intuitive grasp of the issue of increasing complexity and the relevance of imagistic representation of particular events to that issue. The same underlying point would hold in slightly different frameworks, e.g. using a Fourier basis instead of linear regression. Indeed, Nagy & Orban 2016 develop similar ideas in the context of a Mixture of Gaussians framework.

We might want to capture non-linear effects. Perhaps the influence of light levels is not constant, but instead small changes in these variables have tiny effects, whilst larger changes have disproportionately larger effects:

(Eq. 5) $Y = a_0 + a_1X_1 + a_2X_2 + a_3X_1^2$

We might want to include interaction effects capturing how one variable's effect on *Y* is partly mediated by the value of another. Perhaps the *influence* of location on nectar level is larger at higher light levels:

(Eq. 6) $Y = a_0 + a_1X_1 + a_2X_2 + a_3X_1^2 + a_4X_1X_2$

Complicating the model can increase predictive power, assuming that the world is more complex than our current model allows—but at a cost. It requires extra computation; and it requires stretching the limited data we have to estimate more parameters. Adding extra parameters too freely introduces the potential for overfitting—introducing so many parameters that the equation that results hews very closely to the specific, noisy patterns in the data we happen to have rather than picking out the broader patterns that are likely to generalize beyond the dataset.

The net benefits of additional complexity therefore increase with the amount of data available. As we gain the data to fit it properly, a more complex model may capture more of the complexity of the world, without leading us astray in the way that it would without enough data to go on. This means that the form of our optimal model, given our experiences, will change. There are a few ways of dealing with this.

We could ignore this change and simply fit one model, gradually improving parameter estimates using extra data but not changing the structure of the model itself. This would be relatively simple to implement. But we would be blocked from ever deeply improving their

model. They would be guaranteed to remain eternally blind to any structure in the data that they did not hypothesize from the outset.

We could fit multiple models in parallel from the beginning—one which takes into account light levels, another with non-linearities, yet another with rival nectar-lovers. This would require a great deal of computation at every stage and so would presumably be extremely costly, especially with a large number of such models. We could easily end up updating values for thousands of parameters in thousands of models.

Ideally, we could pursue a different strategy: start with just one or two simple models and gradually increase their complexity, flexibly adding and deleting variables and testing the resulting models for improvements in performance on their predecessors, abandoning them if needs be. This kind of learning would be extremely cognitively significant, as it would be in principle capable of capturing indefinite amounts of complexity in the environment, and even with realistic amounts of data it could reach a better model than the rival approaches relatively quickly. It could also be completely domain-general, linking any variables into one big model if appropriate. This sort of flexible learning of indefinitely complex patterns involving any combinations of variables could be crucial to coming to understand deep features of a complex environment.

This is where episodic memory comes in. This optimal strategy requires memory of particular data points. This is for essentially the same reason as in §4: the strategy involves estimating parameters that have not been estimated in any form already. If we throw away all the data except for the parameter estimates of our current model, i.e. keep just a generic representation, we won't be able to fit a new, fancier model that we'd like to try out.

5.5.1 Imagery

Why might imagery be important to this strategy? In principle, this sort of model-building could be done with semantic memories of the values of all the relevant variables at particular times. However, an imagistic-representation-based system could be so much more efficient at this sort of learning that, for any plausible neurobiology, it is capable of learning much more complex models. My argument for this claim depends on the thesis that imagistic representations are particularly well-suited to representing and storing the simultaneous values of a large number of variables compactly. This could be defended on different views of imagery. I will show how one conception of imagery supports it: offline use of perceptual systems.[205]

By 'the offline use of perceptual systems', I mean making use of systems that are specialized for automatically processing sensory inputs, because they are so specialized, but in a context where they are not processing sensory inputs (or not doing so in the usual way).[206] Note that imagery is not here defined in phenomenological terms: indeed, this view of imagery is compatible with the possibility of imagery occurring unconsciously Nonetheless, this activity

[205] Imagistic representation' is defined as offline use of perceptual systems for the purposes of this chapter, but this leaves open the best definition(s) of 'imagery' or 'imagistic' more generally, and the claim that imagery allows for compact storage of many features of particular episodes could be argued for under other views, including views on which such terms capture multiple phenomena (Kind 2013). But the definition adopted here is not arbitrary. There is some evidence for imagistic representation in my sense playing a role in many instances of sensory imagination and in episodic memory. There is considerable evidence from neuroimaging and neuropsychology for visual (and other sensory) areas being used in both perception and mental imagery, surveyed in Kosslyn et al 2006 and Pearson et al 2015, and in episodic memory, for which see Wheeler et al 2000; Vaidya et al 2002; Gottfried et al 2004; Buckner & Carroll 2007; and Rubin & Umanath 2015. Human subjects with limited sensory imagination tend to also have various limits to their episodic memory, whether they always had limited sensory imagination (Zeman et al 2015), or lost visual imagery due to lesions (Greenberg & Rubin 2003), and subjective ratings of vividness of imagery associated with episodic memory are correlated with memory accuracy (Neisser and Harsch 1992, Brewer 1988: 68). There is also evidence for a related kind of offline processing in the hippocampus, an area heavily implicated in human episodic memory (e.g. Hassabis et al 2007; Ólafsdóttir et al 2018), which we will discuss in more detail in Chapter 7.

[206] Related notions of 'simulation' and 'offline' use of psychological systems are developed in more detail by Nichols et al 1996, Currie and Ravenscroft 2002, Goldman 2006, and Carruthers 2015 in other contexts.

may form part of the explanation for the perception-like phenomenology of mental imagery and episodic memory, when it is conscious.[207]

Perceptual systems are specialized for representing multiple features of the scenario immediately around the subject. They are honed to start with sparse data such as limited retinal stimulation and quickly capture the shapes, spatial relations and locations of different objects, pick up on important properties like faces or animacy, and combine these into an integrated, coherent whole. This specialization could be achieved through some combination of natural selection and perceptual learning. Either way, perceptual systems will embed a great deal of information about the regularities in the subject's environment: that certain patterns of light and shade tend to mean a certain 3D shape; that it is likely that those two protruding shapes with an occluding object separating them (but aligned just so) are really parts of one object; that most bananas share a distinctive shape and colour-pattern (and likewise for faces, guns, and tables).

This specialization can be exploited by offline use of those systems. If information about a scenario is fed to these systems in the right way, these heuristics can be used to construct a filled-out representation of that scenario which respects the environmental regularities embedded in the system. This allows for a kind of data compression. If one stores the right kinds of information, one can store only a relatively small amount of information about an event (saving on memory resources), yet still have the ability to reconstruct a rich representation of that event, using these reliable perceptual heuristics.

For example, for Bob to store the details of how Alice's face looked on a particular occasion, there would be no need to actually explicitly record, as it were, every pixel. Instead, suppose that while spending time with Alice, Bob's visual face-recognition sub-system has

[207] Neither is this notion of imagery tied to vision — we will discuss this point further in §6.3.

developed a scheme for Alice's face. This scheme could be used for quickly recognizing Alice, for filling out the shape and other low-level features of her face when he only looks at her briefly, in poor lighting conditions etc., and for prompting a second look when the pattern of lower-level features exhibited on some occasion is Alice-like but does not quite fit with her regular appearance. Suppose Bob also has a sub-system for recognizing people's emotions from facial expressions. Then all that need be stored, to be able to recall a good approximation of Alice's face on a particular occasion, is two pieces of information: that it was Alice, and that she looked angry. By running his facial processing systems offline on these two pieces of information, Bob could reconstruct many details of Alice's face — the space between her eyes, the way her mouth went taut and her eyes widened (down to details of shape that are difficult to capture in language). And precious storage capacity could be used for any occasion-specific details — that Alice was (atypically for her) wearing sunglasses, and had an unusual mark on her chin. Having space to store such aberrant details is crucial to the kind of learning we are interested in: they will often be the details which are unexplained noise for Bob's existing models and relevant according to potential new models.[208]

For a fixed storage capacity, this will allow far more events to be stored with rich amounts of detail available at retrieval. This, in turn, gives us the thesis about imagistic

[208] Related ideas about data compression via storing general patterns along with the unpredictable divergences from those patterns are widely used in computer science, and have long been proposed to be important to the brain generally, at least since Barlow 1961. The idea that episodic memory retrieval involves some form of pattern completion from sparse code is also suggested by Rolls & Treves 1994 and Cheng et al 2016, although their models focus on pattern completion in the hippocampus rather than perceptual systems. It is not straightforward to directly test the claim that using offline perceptual systems allows for more effective data compression than non-imagistic methods, absent running detailed simulations, but there is evidence that subjective ratings of vividness of imagery in episodic memory, thought to be related to the number of details it represents, are correlated with memory accuracy (Brewer 1988: 68), suggesting that representations which use imagery to a greater degree may accurately represent a greater number of variables.

representation we need: that it allows for compact storage of rich information about particular events.

Without such information about the simultaneous values of multiple variables, for multiple particular events, potential new models could only be considered against future data. The subject would have to wait for just the right combinations of variables to co-occur to test proposed new models. This would be considerably slower than being able to at least preliminarily test potential models against previously encountered relevant scenarios. Therefore, having imagistic representations of particular past events massively expands the range of models that can be learned in a reasonable amount of time, compared to a system which only has access to non-imagistic representation and generic imagistic representation.

5.5.2 Memory Traces

This account of the significance of episodic memory implies a further essential feature: imagistic representations will need to be reconstructed from memory traces. A key part of the account of imagistic representation's importance was that there is information stored in such a way that it can be used to reconstruct a rich imagistic representation, and that this information should reliably derive from a single past event rather than averaging over multiple events. Otherwise it would be less useful as data for novel, complicated models. The most natural way to have simulation-relevant information available would be to store it in a compressed format, matched to the available reconstruction systems: i.e., as what we might call readable memory traces. This gives good reason to include such memory traces in the core of episodic memory, and to rule out other imaginative simulations of particular past events (such as reading a book about the Battle of Blenheim and then carefully imagining being there — a process which does not take advantage of the potential for compression in episodic memory).

This implication firmly sets the account apart from some other recent simulationist accounts of episodic memory, especially Michaelian 2016, on which any simulation of a past event, even if based purely on a combination of general knowledge and information derived through testimony, would count as episodic memory. However, it does not commit to some of the controversial doctrines associated with the terminology 'memory trace'. For example, memory traces in this sense can degrade, and could sometimes have information added to them during access, provided they remain tied to single events and sufficiently reliable. And positing memory traces in this sense is in no way in tension with the view that fallible reconstruction takes place during episodic memory retrieval.

5.5.3 General Learning, Narratives, and Essentially Dynamical Entities

The sort of general learning based on episodic memory envisaged by this chapter would go well beyond the kind of learning that uses representation of narratives or essentially dynamical entities discussed in Chapter 4. There, we discussed strategies for learning about intractably complex domains (especially intractably complex dynamical, non-Markovian domains), which operated by abandoning the hope of fully general learning. Instead they imposed simplifications on the world (the use of templates for essentially dynamical entities) or only hoped to understand very limited parts of it (narrative-based understanding). There are no such limitations here. This kind of learning is entirely general: as no assumptions have been made about the complex forms of model such a procedure could learn, it could learn complex, non-Markovian and general models of the world, at least if enough of the right kind of data were stored. The 'right kind of data' here would include allowing access to episodes' temporal coordinates and/or internal temporal structures at retrieval, but from the point of view of this sort of learning, this would just be another important piece of data about episodes.

This does not mean that a creature who could learn in this way would have no need for narrative representations or representations of essentially dynamical entities: the individual's lifetime might not be long enough for such learning to produce a fully general morel of the most complex dynamical entities, and up to that point simpler models would be useful. What's more, even for a creature in possession of a fully general model, simpler models might be useful for making quick, unimportant decisions with less cognitive effort.

5.6 Objections and Avenues for Future Research

Several objections to this proposal lurk. Answering them shows how the approach can lead to a rich project of computational modelling.

At first glance, learning using episodic memory in the way suggested would benefit from remembering *all* the details of all events ever encountered. Fitting complicated regressions would ideally be done with as much data as possible. And yet we do not seem to episodically remember every detail of every event in our life. We do throw away a lot of data. Is this a problem for a view which says episodic memory is important because it is used for such processes?

No. Even if a system is not optimized for performing a certain role, it may nevertheless perform that role well enough to be significant because of it. We can go further than just making this general point: episodic memories may be formed by a process that trades off costs of storage with some heuristic-based estimate of how useful different details or entire events are likely to be later.[209] Modelling the relevant costs and benefits and procedures for estimating them would be an interesting project. It would need to take into account factors such as the costs of explicitly storing certain kinds of variable, the reliability of reconstruction of some variables on the basis of

[209] Nagy & Orban 2016 model something very much like this using a Bayesian measure of surprise.

others under different conditions, and any biases the system has about which variables will be most important to learn about.

One might also worry that human episodic memory is too unreliable to be used as data for any sort of useful model-building. Psychologists have determined a number of ways of predictably inducing subjects to make mistakes when relying on their episodic memories, from changing small details to confabulating entire events.[210] If subjects routinely make such mistakes with episodic memory, it is hard to see how it could be useful as a means for even preliminary testing of complicated models.

Episodic memory may not be as unreliable as it is often portrayed to be. Although there are some well-known cases of subjects misremembering key details in high-stakes scenarios (Neisser 1981), most of the best-known experimental effects only occur in unnatural conditions, given certain kinds of prompting. It may be best to think of them on the model of visual illusions: there are many well-known cases where we can reliably induce subjects to make mistakes when relying on their vision, but this does not show that vision is unreliable most of the time.[211] Indeed, it must be at least somewhat reliable given that we commonly can accurately recall details of events, relying on it in our daily lives to remember what time to meet people for dinner etc..[212] And somewhat reliable most of the time may be enough for the processes outlined above to be useful. Again, there is a fruitful question for computational modelling here, namely

[210] Reviewed in Roediger 1996, Loftus 2005.

[211] Michaelian 2016 Ch. 7 emphasizes this point; it is at least implicit in e.g. Roediger 1996. The idea that episodic memory is often reliable about the past is compatible with the view that we frequently get the nature of episodic memory wrong when introspecting, e.g. with our tendency to underestimate the degree of simulation involved (Michaelian 2018).

[212] Several of the key adherents to the view that episodic memory evolved for some purpose other than veridicality, who emphasize simulation and distortions, nonetheless emphasize that episodic memory is often veridical, e.g. Mahr & Csibra 2018, Schacter et al 2018.

determining just how much unreliability can be tolerated and under what conditions. We should not assume in advance of such work that episodic memory is reliable enough for the kind of learning proposed above, but neither should we assume that it is not.

A more sophisticated version of this objection would point out that even if episodic memory is not too unreliable, the effects reviewed in Loftus (2005) show that it involves reconstruction of a particular kind—reconstruction influenced by the beliefs or model of the environment we already have (indeed, the ability to explain such effects is one of the major motivations for the simulation view for some of its proponents, such as Schacter et al 2011, De Brigard 2014, Michaelian 2016). Does this not pose a problem for using episodic memory to develop those very models? Would they not just generate data that confirms themselves and distort data from the world so it is in line with their own predictions, making it seem like they cannot be improved even when they are wrong?

There are reasons to think we may be able to get around this problem. The reconstructed episodic memory at the moment of recall will be based on existing models *in addition to event-specific information*. Event-specific information, even filtered, implies that we do not simply have a model confirming that its own predictions about a scenario are borne out by its own predictions about that scenario.

This would be especially true if the process of selecting which events to store uses an error signal at the time of encoding, specifically choosing to store events which the current model predicted poorly. In general, these encoding-selection processes will be crucial to the functioning of episodic memory, and modelling which strategies work well and testing these models against human and animal data, could be an extremely fruitful avenue for further research.

In fact, there is already an exciting literature developing in artificial intelligence, trying to incorporate something called 'episodic memory' into various learning algorithms.[213] The ideas in this chapter offer a systematic way to use this sort of modelling to understand episodic memory generally as opposed to just its instantiation in particular minds, as we have seen. This chapter also offers a clearer way of thinking about episodic memory for those engaged in computer science. This is because talk of 'episodic memory' in AI tends to focus on just one feature of episodic memory—storage of particular data points instead of generalizations. Nagy & Orban (2016: 2699) do ask a very similar question to the cognitive significance question: "what is the benefit of devoting precious mental resources to…storing rich snapshots of actual experience…?" But their model does not explicitly have any special role for imagery. Ideas about episodic memory developed in the context of reinforcement learning also usually have little to say about the imagistic nature of episodic memory — they usually only require storing sequences of states, actions and rewards.[214] More attention to the different options here would lead to a better understanding of what is distinctive of episodic memory, but also to better understanding of which options are likely to work best for engineering purposes.

5.7 Alternatives to the Cognitive Significance Framework

This chapter has advocated finding the core features of episodic memory through a process of starting with intuitive features of episodic memory, then looking for combinations of those features which are cognitively significant compared to what those features can achieve

[213] E.g. Lengyel & Dayan 2007, Gershman & Daw 2017, Lin et al 2018. Bornstein & Pickard 2020 explicitly draw on the role episodic memory can play in reinforcement learning to define its core features (for them, one-shot acquisition and being pervasively associative).

[214] That is, there is no special role for imagery in their models – their discussions do describe episodic memory with phrases like "connection between many aspects of that event, including multiple sensory dimensions" Gershman & Daw (2017: 110), but this talk is not straightforwardly connected to the actual algorithms they propose.

individually. What other ways of individuating episodic memory are available, without relying solely on intuitions (an approach which quickly leads to stalemate)?

A popular alternative is to look for the system in the brain that is involved in tasks that intuitively involve episodic memory, and to identify episodic memory with whatever this system does. However, this approach faces a number of difficulties. First, there is a kind of circularity in defining the tasks that involve episodic memory: it can always be disputed exactly which tasks do involve episodic memory, qua its most essential features. Furthermore, the account is liable to give kinds at the wrong level of abstraction: specifically *human* neural systems. Unless it is supplemented by psychological-level considerations about what role episodic memory plays in the mind, it will miss psychologically interesting, multiply realizable kinds—and, as a result, it will miss high-level similarities between different species where they implement interestingly similar cognitive roles with different brain structures.[215] Relatedly, where the relevant brain system is a region like the hippocampus—complex, connected to many other regions—it is likely that it is recruited in several functionally distinctive processes.

A more subtle approach is looking for a homeostatic property cluster—a collection of features which regularly co-occur, thanks to a uniform underlying mechanism.[216] Cheng and Werning 2016 apply this approach to episodic memory. On their account, the relevant cluster of properties includes features like having a content that consists of a sequence of events; the

[215] For example, Buckner and Carroll 2007 claimed that it was unlikely that scrub jays have episodic memory on the grounds that they do not have a six-layered cortex, which seems to be empirically important to human episodic memory. But we now know that birds have an area, the nidopallium caudolaterale, which does roughly the job that the prefrontal cortex performs for primates, with sub-areas roughly corresponding to the six layers in primate cortex (Güntürkün & Bugnyar 2016).

[216] This applies a well-known approach to natural kinds from Boyd 1989. The underlying mechanism is crucial here: unlike e.g. Templer & Hampton 2013's approach of looking for "a constellation of mnemonic functions, rather than a single entity", the approach identifies a reason why there should be clustering, which allows us to distinguish some features as more important than others

accuracy of the content; and the content being based (via a memory trace) on an earlier experientially represented. They think that there are states with some but not all of these properties (e.g. states which involve simulation of an event based on a memory trace but which are not accurate), but they emphasize that these are not central cases of episodic memories. They argue that the uniform mechanism underlying this cluster is hippocampal replay (a phenomenon we will discuss in more detail in Chapter 7).

This account faces the problem that hippocampal replay is probably used in a range of other functions besides episodic memory, as we will see in in Chapter 7.

But the greatest worry for the approach is that it is unclear if there really is a cluster of regularly co-occurring properties at all distinctive of episodic memory. Several but not all of the features identified by Cheng and Werning occur in generic imagistic memory arguably at least as frequently as Cheng and Werning's 'cluster'. One major advantage of the Homeostatic Property Cluster view of natural kinds is that it avoids strict necessary and sufficient conditions and allows a large range of borderline cases. But if there are too many exceptions and borderline cases, there is no cluster at all.

The cognitive-significance-based account can be seen as a very different version of a homeostatic property cluster account of episodic memories, which inherits its advantages without its defects. This is because we can think of the cognitive significance approach as looking for collections of features which could underlie very distinctive clusters in the overall capacities of the mind. Here the underlying mechanisms would be the imagistic representations of particular events in the case of episodic memory, and imagistic representations of event-kinds in the case of generic imagistic memory. The clusters of properties they give rise to would be the very different learning mechanisms and resulting capacities they support. There would be no problem for this

account with the fact that there are only small differences between the structure of generic imagistic and episodic memory: they would be different kinds because of the very different phenomena they support, much like how carbon allotropes have many similarities in their local properties, but nonetheless give rise to very different kinds of material like diamond and graphite when we zoom out.

An approach closely related to the HPC view would claim that we should aim for natural kinds in the mind that can be seen to be *homologous* between different species. As we discussed in Chapter 3, it does sometimes make sense to describe mental traits like emotions as homologies, and when we can do so, this can point to an underlying reason justifying extrapolations between different species. And (with the HPC approach) one of the defining features of natural kinds is supposed to be their justifying extrapolations on the basis of underlying shared mechanisms.

However, as we also saw in Chapter 3, phylogenetic- and homology-based extrapolations are often weak in the case of the mind, especially for more cognitive, flexible traits, where strong selection pressures are more likely to have overridden similarities due to shared descent than for other traits. Furthermore, such an approach, unlike the approach here, will be likely to miss on surprising, deeply important shared similarities in abstract computational forms, where those similarities are due to convergent evolution. Of course, finding homologous systems to the systems implementing episodic memory in humans could be one important source of evidence for episodic memory in other animals, and we will discuss some lines of evidence of this sort in Chapter 7; but this does not mean that episodic memory should be *defined* in terms of homology.

5.8 Conclusion

Episodic memory is cognitively significant, if its core is imagistic representation of particular events based on a certain kind of memory trace (while autonoesis does not contribute

to the distinctive significance of episodic memory). Cognitive significance is a good way of individuating mental traits, especially for the purposes of cross-species comparison. Therefore, for the purposes of cross-species comparison, the core of episodic memory is imagistic representation of particular events based on a certain kind of memory trace. I showed that this proposal suggests many interesting empirical and computational modelling projects as well as useful ways of reframing projects already being carried out in the AI literature and the interdisciplinary literature on the evolutionary function of episodic memory. Episodic memory has the potential to transform an animal's mind, and thinking about it from this perspective has the potential to transform how we approach the cognitive science of episodic memory.

The next two chapters will flesh out this proposal in more detail. Chapter 6 will discuss the various forms episodic memory could take while still retaining its three core features, thanks to those three core features taking different forms, and thanks to additional peripheral features. It will also wield the cognitive significance framework to go beyond the positive account of episodic memory's core features, and argue more explicitly that features like consciousness and meta-representation, which many have assumed to be essential to episodic memory, really are peripheral. All this will put us in a better position to interpret the existing lines of evidence for episodic memory in animals, in Chapter 7.

Chapter 6: Varieties of Episodic Memory

6.1 Introduction

Chapter 5 argued that episodic memory has three core features in combination: *memory traces*, being used to construct and *imagistic* representation of a *particular past event*. It argued that these features are core because they jointly make for a kind of memory that is suited to learning indefinitely complex models. This is not to say that any animal who has episodic memory will be able to learn indefinitely complex models: the extent to which they do so will depend on further background conditions like memory capacity, and actually using episodic memory in the right sort of algorithm. But they will in an important sense have the potential to learn general models.

One feature of the cognitive significance approach to individuating traits is that it allows that these traits can be instantiated in many varieties. Traits' core features can come in different forms, and they can be combined with different peripheral (non-core) features. This point is crucial to understanding the distribution of these traits in the animal kingdom, as it allows that different animals may have very different varieties of the same trait.

This chapter explores different possible varieties of Episodic Memory. I first explore different varieties of each of the three core features: memory traces (§6.2), imagery (§6.3), and representation of past events (§6.4). Doing so will serve to further clarify what it takes to count as having these different features, especially representation of past events, which was left largely unanalyzed in Chapter 5.

I also develop the use of cognitive significance in individuating traits, by showing that certain additional features do *not* combine with the three core features of my account in particularly significant ways, and that this supports their not being included as core features of episodic memory. Including internal temporal structure to an episode (§6.5), putting the episode

into a broader temporal context (§6.6), consciousness (§6.7), and metacognition (§6.7), are all features of one variety of episodic memory: *human* (perhaps just contemporary Western human) episodic memory. And all are likely to be significant in their own right. But none is significant specifically in combination with the core features of episodic memory. Animals without these features, therefore, may still have episodic memory: they just have a different form of episodic memory to humans. Indeed, it may be that *humans* sometimes have unconscious episodic memory, and episodic memories of particular past events which do not represent that they are particular or that they are in the past. As so often, studying the varieties a mental trait could take in different species opens up possibilities which have not yet been systematically searched for of forms that the trait might take in humans. Denying that these features are core to episodic memory runs sharply against a great deal of the literature, which typically assumes (or gives unsuccessful arguments for) these features (often under the terms of art 'autonoesis' and 'chronesthesia') are essential to episodic memory, and often uses this view as part of an argument that animals either are not known to have episodic memory, or even are known not to have episodic memory.

6.2 Varieties of Memory Trace

As mentioned in §5.5.2, my account makes very few demands on memory traces; so they can come in many varieties. All the account requires is information stored in such a way that it can be used to reconstruct a rich imagistic representation, and that this information should reliably derive from a single past event rather than averaging over multiple events. The account is neutral on where in the brain these traces are: they could be in the hippocampus or could start in the hippocampus before being transferred to the cortex, for example.[217] They could involve a

[217] See Barry & Maguire 2018 for discussion of different options here.

unique code, in a dedicated neural system, or they could be stored much like other memories. They could involve explicit representations which can be read and used by other systems in contexts besides episodic memory retrieval, or they could involve a mere information-carrying code which can only be read by the hippocampus in conjunction with other systems, and only in the context of episodic memory retrieval. If representational, they could use a format involving language-like sentences, analogue forms of representation including map- or image-like representations, or a high-dimensional vector.

What exactly is required of memory traces if they are to play the role demanded by my account of Episodic Memory will vary according to other features of the specific system under consideration. The requirements on memory traces will be determined in tandem with the systems available for reconstruction based on those memory traces. As we will see in §6.7, if a system uses metacognition in different kinds of reasoning in the reconstructive process of getting from a memory trace to a full-fledged simulation of the remembered event, then there will be fewer requirements on memory traces. In such a system, traces will need to relate to a particular past event, but they may not need to have any special features relative to other stored information. Michaelian 2016 argues for an extreme version of this view, where the use of metacognition and a wide range of reconstructive processes give him the resources to allow for episodic memory with no dedicated memory traces at all. Instead, Michaelian thinks that any semantic-format memories relating to one event (or appearing to relate to one event) could be used. At the other extreme, we will see that a system without metacognition will require memory traces to be embedded in a system with a rather distinctive architecture. Intuitions about memory traces certainly favour their being in a dedicated system rather than their being distributed, as this makes for a straightforward account of what distinguishes memory of an

event from mere informed imagination of that event: memory derives from an experience via an uninterrupted memory trace, whereas any representation of the event which derives from some other sources, such as testimony, should not count as memory.[218] And much of the distinctive advantage of compressed storage which derives from the imagistic nature of episodic memory will be lost in the more flexible sort of system. But a version of episodic memory could operate in this way and still count at least as a borderline case, and we can allow within core cases of episodic memory for systems with different levels of distinctiveness in their memory traces compensated for by different levels of sophistication in the trace-reading systems.

Alongside variety in what form memory traces take, there can be huge variety in other features of their functional role, including the processes used to select which bits of information to store (e.g. do these processes have a list of features of an event that they always store? Or do they use some process to determine the most efficient thing to store on a particular occasion, based on how reconstructive processes will work, or do they do something in between?), the processes used to encode and consolidate these traces, and the way these traces are organized and accessed (e.g. are they arranged by topic or by chronology, in the sense that accessing one memory primes 'nearby' memories in this organization to be accessed next?). Again, it would be perfectly possible to find different animals instantiating different aspects of this variety, and we should not assume that finding one of these varieties in humans means that other animals either will have that same variety, or must if they are to count as having episodic memory.

[218] See Robins 2016 and Martin & Deutscher 1966 for more detailed discussion bringing out these intuitions.

6.3 Varieties of Imagistic Representation

The key role played by imagistic representation in the significance of episodic memory is its allowing for compact storage of many features of a specific event. §5.5.1 explained why imagistic representation could do this if imagistic representation consists in offline use of perceptual systems.[219] This view allows for variety in imagistic representation corresponding to the offline use of different sensory modalities. But as I will show in this section, this view also allows a natural generalization of the notion of imagistic representation, which provides for even greater variety in episodic memory, based around offline use of certain kinds of non-perceptual systems.

The point that episodic memory need not be based around vision, or around human perceptual modalities (whether this means the traditional five, or some other categorization) is important in the context of thinking about animal minds. Many animals have different sensory modalities to humans: sometimes they lack or only have limited versions of human vision, for example. But many have versions of our senses which are far more expansive in various senses, whether that means seeing colours in a wider range of the electromagnetic spectrum, smelling so much more acutely and with so much more sophisticated categorization that we are baffled by abilities that turn out to be based on their olfaction, or using sound for echolocation. And several animals have senses which humans lack entirely, like the many fish with electrosensory organs.

[219] As noted in §5.5.1, this is far from the only account of imagery on the market. Other accounts of imagery might still support the idea of compact storage of many features of a specific event, and allow for a different variety of episodic memory systems. For example, if one thinks that imagistic representation is to be defined not in terms of the systems used, but the format of the representations involved, one might still be able to adopt a version of my view of episodic memory. For it is not implausible that image-like format allows for compact representation of relational information, as touched on in the discussion of map-like representation in Chapter 4. If one represents Cardiff's and Swansea's positions on a map or an aerial photograph, one thereby represents the spatial relations between Cardiff and Swansea, without any further calculation or extra representations being required. If one adopted a view of episodic memory based around *that* account of imagery, it would allow for variety in episodic memory corresponding to different image-like formats (maps, pictures, and diagrams all admit of many sub-kinds). One advantage of the account of imagistic representation adopted in the text is that it enables the more interesting kinds of variety involving different sensory modalities and non-sensory 'imagistic representation'.

It would be entirely unsurprising if many animals have imagistic memories based around a very different set of sensory modalities to humans.

But we can generalize further. The key features of perceptual systems which allowed them to be used for compressed episodic memories were their embedding heuristics for quickly and tolerably accurately getting from an underdetermined stimulus to a fleshed out representation of many features of the world immediately around the subject. But other systems operate similarly: systems which are relatively fast, relatively unsophisticated, and relatively specialized, when compared to more domain-general, cognitively intensive central cognition.[220] Motor systems and perhaps emotions both include subsystems matching this sort of description. They both seem to be specialized for performing particular information-processing tasks, such as transforming representations of where medium-sized objects around the perceiver are now and a representation of where the subject might want them to be, into a series of motor instructions, or picking up on threats in the immediate environment.

Episodic memories related to actions and to evaluation could well make use of such systems as part of retrieval. This is somewhat controversial in the case of the emotions: there is controversy about how to think about emotions, and empirical reasons suggesting humans cannot run emotions offline,[221] and disagreement about the role emotions play in human episodic memory.[222] But we can at least understand how such systems might be used for reconstructing compact representations in principle. Rather than representing all the features of a scenario, one

[220] This characterization has affinities with Fodor (1983)'s notion of a 'module', but it does not presuppose some of the stronger theses associated with modularity as Fodor thinks about it. For example, it is not required that peripheral systems be innate or hardwired, and it is not required that they be 'cognitively impenetrable' i.e. that they function in a way that is impossible for central systems to influence.

[221] Currie and Ravenscroft (2002: 189ff.) claim emotions cannot be simulated; Goldman (2006: 47) claims they can.

[222] Rubin & Siegler 2004; Richards & Gross 2000.

could store the menacing character of someone's look or the dangerous way a tree was swaying, and use one's templates for dangerousness embedded in a fear system to reconstruct details. And likewise, rather than remembering all the precise movements one went through via remembering their sensory and proprioceptive upshots, one could just remember an action representation of jumping from branch to branch in a tree, in addition to the layout of the tree, and reconstruct all the sensory upshots using the motor system.

6.4 Representation of Particular Events

Much more needs to be said here about the representation of particular events than was said about imagistic representation and about memory traces, as we have so far left some key questions unanswered about what it takes to represent a particular event. What is it to represent a particular event? What does this require? And what makes it the case that a subject represents one particular event rather than another?

It is crucial to return to the distinction touched on in §4.2, between two kinds of uniqueness: being *specific*, and being *particular*. Specificity is about how many features something has to be to count as a member of a type. Specific types of event will involve so many features in combination that they are unlikely to ever be repeated. The more specific a type of event—the more features it includes—the less likely it is to have more than one instance. For particular events, however, repetition—or, rather, multiple insatiability—is metaphysically impossible, not just unlikely. By comparison, specific types of *object* can be distinguished on the same lines: we can imagine science fiction cases in which two cloned particulars are alike in extremely specific detail, but still numerically distinct particulars. Of course, any particular event is also an instance of an extremely specific event-kind. And one might have a metaphysics where particulars just are maximally specific sets of properties. But in

the context of what individuals can *represent*, the two come apart. It is not feasible to represent a maximally specific set of properties, property by property; and one can represent a particular event without representing enough of its features for those features to be uniquely identifying, as in THAT CRICKET MATCH.

Intuitively, generic imagistic memory can be just as specific as episodic memory: imagistic memory of some type of event one has experienced many times in a very set routine, such as making tea in the morning, may be extremely specific, including many details in rich vividness, while episodic memory can be rather vague about many details, especially for events that are more distant in the past. But which feature—specificity, or particularity—is important to the story about significance I told above?

Specificity is important: the whole point of using imagery is to be able to represent multiple features of a single event, the value of multiple variables at the same moment in time. But particularity is crucial. Important and unique episodes may be important to remember with as much detail as possible, even if not all that much detail *is* possible to remember. Episodic memories are important in providing the right kind of variance for testing potential models, and this means remembering unusual events even when one is only aware of a few features of those events to begin with, as when things going bump in the night, and one might only have been aware of a few features in a sleepy, low-attention haze, in poor lighting conditions etc.; or when the events happened so long ago that everything but the most distinctive parts have been lost to degradation. More importantly still, if a specific event-kind repeats, this is important information to know rather than to gloss over, especially when doing any kind of statistics responsive to frequencies.

What does it take to represent particular events? We need to distinguish two feats here: Representing particular events, and representing particular events *as particular events, based on some kind of grasp of particularity or their location in a linear temporal framework.* The importance of particularity in our account of significance did not seem to rely on the latter ability, which at least at first glance sounds like something only available to some humans. But something like the latter ability may be thought to be required for the former ability. And this may make it seem suspect that animals could possibly represent particular events at all.

It is reasonable to think that an important part of the identity conditions for events are spatiotemporal coordinates, and that an account of what makes something a particular event will cite these, combined with a specific event-type. It might seem reasonable to then infer from this point that one needs to represent temporal coordinates to represent an event as particular. Something like this line of thinking seems to be driving Campbell (2006: 5) claiming it is 'questionable' whether episodic memory is possible without a conception of there being a particular time at which the past event occurred. Campbell and McCormack & Hoerl 1999 both offer accounts of what it is to represent particular past events roughly along these lines, but differ over details: McCormack and Hoerl claim that the key is 'decentring' from a perspective bound to now: having a conception of different temporal perspectives (perspectives from which a certain moment counts as the present), and being able to reason, for example, that a current ongoing event will be in the past from the perspective of subsequent days. Campbell 2006 criticizes this view on the grounds that it is not demanding enough. He thinks one could appreciate different temporal perspectives, whilst thinking that they are tied to different phases of a large cycle (which will be repeated), rather than appreciating them as different unique points in linear time. One could, for example, not have a sense that time extends beyond a weekly cycle,

and distinguish 2pm on Monday from 2pm on Tuesday, and understand that at each of these times one will use 'now' to refer to a different 2pm, but be incapable of distinguishing 2pm on Monday 1st from 2pm on Monday 7th. Campbell 2006's positive alternative involves understanding that the same object will behave differently in different occurrences of the same script and in different scripts, because of what has happened to it before.[223]

Both Hoerl and McCormack and Campbell are radically overintellectualizing here, at least when it comes to the kind of representation of particular events required for episodic memory. A system with the following functional architecture would give us everything we want for episodic memory, as far as representing past events goes: Suppose memory traces that are formed, updated and used such that:

(i) The only way to create a new memory trace is in response to experience of an event.

(ii) The only way to add a feature to the memory trace is on the basis of that experience, or on the basis of finding, during an episodic reconstruction of that event from this very memory trace, that it probably had a certain further feature.

(iii) Episodic reconstructions of an event should directly draw on exactly one memory trace (although it may also draw indirectly on others, by drawing on general semantic or generic imagistic memories that may in turn derive from other episodic memory traces).

[223] Compare Campbell (1994: 39ff.)'s account of representing particular past times, and Hoerl 2008, Hoerl & McCormack 2018. It should be noted that it is not entirely clear how Campbell's alternative to Hoerl and McCormack's proposal is supposed to get round his own objection: if one had a long enough cycle or script, one could incorporate the object's different interactions on different occasions. And one could in principle do so for any object: if one believed in eternal recurrence, in the sense of believing not just that events will repeat in detail, but that the structure of time in the universe is cyclical, one would think that *all* events involving causal objects developing over time could be placed on a long enough cycle and repeat. Representing events as non-repeating in such a way as to completely rule out such possibilities seems to require having detailed beliefs about the nature of time that go well beyond those of an ordinary human.

(iv) Use of episodic reconstructions based on memory traces should treat them as single events—as data and not as models. That is, they should be used for estimating statistics like averages, regression coefficients or other parameters, by doing calculations over multiple such reconstructions.

(v) The way such averaging operations work should be such that:

(v.a) When drawing on multiple memory traces to calculate such statistics, the same episodic reconstruction should not be used multiple times, and different reconstructions based on the same trace should not be used in the same calculation.

(v.b) Two memory traces can both be used even if they represent all the same features of their respective events.

We could complicate rules for this in all kinds of ways, e.g. allowing for the discovery that two traces derive from the same event on the basis of complex causal reasoning which respects the fact that qualitative identity is not sufficient for numerical identity of particular events, and allowing their contents to be merged under that situation. But (i)-(v) seems to be sufficient for a system to be representing particular events in the relevant sense. In this sort of system, an episodic memory trace will have a privileged causal link to a particular event, and will play a special role in computation because it links to that particular event. Of course, in some cases these rules can break down, or be followed only partially, resulting in cases of failed reference and indeterminacy. Even whilst following these rules, there may be cases where the modification of memory traces allowed by (ii) could both introduce all kinds of errors and distortions relative to the original event, and allow in more and more information from specific general knowledge based on a small number of other events until it becomes indeterminate

which event is being referred to.[224] But if the system functions this way normally, it will be enough to allow for the sorts of learning described in Chapter 5 to get by and this will be because we have a privileged link to particular, not just specific, events.[225]

As with other features, representation of particular events can take many forms, and it may be that the way it works in humans (and in many other animals) is far more sophisticated, and far more intellectualized, than the simple functional architecture of (i)-(v). The point of the cognitive significance framework is not to insist that episodic memory only ever has the minimal features it endorses as core, but that there are many potential varieties of episodic memory, including some rather minimal ones.[226]

---

[224] It is also possible, as in Chapter 4, to envisage cases where a functional role is indeterminate: it has some features which make it describable as representing particular events and other which do not, or it operates like (i)-(v) some of the time but diverges at others. An episodic-memory-like system based on a borderline case of representing particular past events would be a borderline case of episodic memory (and it might well not work very well). This is very different to an alternative sort of case, where we have a system which does not follow anything like rules (i)-(v), but maintains highly specific representations formed on the basis of past experiences. Such a system might be such that, were it to encounter two experiences with very similar features, it would simply merge them, and would treat them as one data point rather than two in any computations done over these experiences. But it might also be such that it represents events in such fine-grained detail—so specifically—that it never encounters any such pairs of events in practice. Such a system might in practice behave much like an episodic memory system, but should not be counted as one, given its very different counterfactual behaviour.

[225] Michaelian (2016: 112f.) endorses an alternative account on which "either the intention of the subject himself or the "intention" of his episodic construction system...understood as shorthand for talk of the system responding to given retrieval cues provided by either the agent or his environment" determines that a particular event is being represented, and which particular event that is: he is forced into something like this view by dropping the requirement of memory traces as a necessary condition on episodic memory. If this proposal works, it simply provides for another alternative implementation of episodic memory (albeit an unusual borderline case without memory traces, which may not operate very well at providing data for statistical models). If it does not, this is not a problem for my view either given that I am happy to endorse the possibility of memory traces.

[226] One upshot of the fact that representing particular past events can occur via an architecture like (i)-(iv), and indeed of the importance of particular, as opposed to specific, events at all in episodic memory, is that there is a strong disanalogy between episodic memory on the one hand and both generic imagistic memory and future-directed simulation on the other, as they will not use memory traces according to the same implicit rules, even if they use memory traces. Debus (2014) argues that we should think of remembering the past and imagining the future as different kinds, on the grounds that the past-directed remembering partly on the grounds that past-directed remembering is of particular events; Michaelian (2016: 116f.) objects that this does not correspond to a psychologically real difference. By contrast, the difference between remembering particular events and generic (even if specific) events clearly is a psychologically real difference on my view, as it involves using information quite differently.

However, one might wonder whether the above is really an account of the representation of particular past events. I have focused on the representation of *particular* past events, but have not said much about representing them as *past* or as *events*. Perhaps these require something more sophisticated?

Plausibly an architecture like (i)-(v) takes care of pastness: if memory traces are only formed as a causal result of an experience which is caused by an event, then absent backwards causation they will only be formed in response to past events. And this, rather than the representation of these events as past and a grasp of what that means, seems to be what is required of episodic memory to play a role in learning.

The case of events is trickier. It is tempting to say that crucial to the individuation conditions for all events are their temporal boundaries and internal temporal relations: an event's start and end, the unfolding of sub-events and processes within that event. Could a creature count as representing events without being appropriately responsive to events' temporal boundaries and internal temporal relations? And could they be so responsive without some further important feature involving temporal representation?

The brain does seem to segment our experience into temporally structured events.[227] This shows up in a number of effects in memory. It is generally easier to recall information from event boundaries, and to recall events that are clearly bounded (with consistency within the event and a sharp contrast to whatever is happening before and after the event). In the primacy effect, we are more accurate in judging the time of occurrence of items presented near the beginning rather than the middle of a stimulus list, suggesting the whole list is chunked into one event with

[227] Keven 2016, 2018 emphasizes this literature in his constitutive account of episodic memory.

the beginning serving as a landmark.[228] Memory for order of events is better for information from the same kind of stimuli (e.g. faces or objects rather than a mix of the two), suggesting it is easier to chunk coherent experiences into one event.[229] Items appearing either side of an event boundary are less bound together/less likely to be integrated into a unified memory, while those appearing within one event are more likely to be bound. Estimates of duration are affected by this segmentation, with items spanning boundaries are remembered as happening further apart.[230]

There are a number of factors known to influence where the brain places the boundaries between events, including large changes in perceived physical features and in perceptual features, motion, shifts in spatial context, and more cognitively defined information based on narratives and scripts.[231] Segmentation can be spontaneous and does not require attention,[232] but attention can influence these processes, and top-down processing can override some effects of event segmentation, possibly through hippocampal processing by linking events across context shifts into a unified structure, even across long temporal gaps.[233]

Many of the effects related to event segmentation can be tied to stability of patterns in the hippocampus and to interactions between hippocampus and areas like the Prefrontal Cortex: areas heavily implicated in episodic memory. There are different patterns within and between these areas for tasks asking for information about the order of sub-events within one unified

[228] Friedman (1993: 54) for references and discussion.

[229] DuBrow & Davachi 2014.

[230] Ezzyat & Davachi 2014; Lositsky et al 2016.

[231] See Zacks & Tversky 2001 and Clewett et al 2019 for reviews.

[232] Zacks et al 2017.

[233] DuBrow & Davachi 2014; Cai et al 2016; Chen et al 2016.

event, vs. across event boundaries, and for familiar vs. novel sequences.[234] This also seems to be true in rodents.[235]

This all suggests that the brain does represent temporally structured events, which include ordered sequences of sub-events. Does this suggest that even minimal episodic memory would require at least some sophistication in temporal representation, with representation of sequences?

Things are not so clear. Once again, we need to ask whether what happens in humans needs to happen in any system where episodic memory plays its distinctive role. And just counting as representing events, for those purposes, does not seem to require representing the temporal structure of events. Rather, what is really required is just representation of the values of a number of different variables at one time. That time need not be temporally extended to encompass multiple sub-events: it could be more like a snapshot, and still provide data concerning the relationship between multiple variables. Some might want to insist at this point that this would not really be representing an 'event' or an 'episode'. But this terminology does not matter here: what matters is what is required for the kind of cognitively significant learning I have proposed for episodic memory. And if one wants to call *that* 'a particular instantiation of the (more or less) simultaneous values of multiple variables' rather than an 'event' or an 'episode', that will not affect the substance of the proposal.

One might, however, suspect that the kind of learning discussed in Chapter 5 not relying on representing the temporal structure of events points not to the representation of internal temporal structure not mattering, but to the paucity of that view of significance. Perhaps internal

[234] DuBrow & Davachi 2014, 2016; Ezzyat & Davachi, 2014.

[235] Guise & Shapiro 2017

temporal structure combines in a cognitively significant way with other features of episodic memory, and so will turn out to be a neglected core feature of episodic memory after all.

6.5 Internal Temporal Structure

Several philosophers have suggested that representing temporal structure internal to an episode is core to episodic memory: episodes should consist of sequences of sub-events, perhaps linked together into some kind of narrative structure (as we saw in §4.2, this could mean a lot of different things) or at least unified by event segmentation processes and/or experienced as a unity.

Two main motivations have been offered for this view.

One is Cheng and Werning (2015, 2016)'s suggestion that hippocampal replay could be a mechanism that explains a cluster of related phenomena at a higher level, combined with their (plausible) interpretation of hippocampal replay as representing sequences. I have already discussed the shortcomings of Cheng and Werning's view here in §5.7 and will not rehash them here. I will return to discussion of the general evidential value of hippocampal replay in this context in Chapter 7.

Russell & Hanna (2012), meanwhile, argue that internal temporal structure is necessary for different features to be bound together into a synthetic unity, which they think is necessary for experience. This view depends on broader Kantian commitments which are related to some of Campbell's views about temporal representation which I have rejected above.

A more promising approach to arguing that internal temporal structure is important to episodic memory would be to try to develop a view which combines the ideas about narrative representations from Chapter 4 and the general-purpose learning from Chapter 5. However, as I noted in §5.5.3, this sort of combination, whilst possible, does not promise to especially

significantly add to what either kind of representation can do on its own. Adding representation of sequences within episodes to episodic memories does mean that episodic memory could be used to learn local temporal patterns and their causal relevance more effectively using general learning; but this is no different in kind to adding representation of sounds and becoming able to spot patterns in sounds using general learning.

Narrative-based learning as discussed in chapter 4 would benefit from episodic memory. It benefits from having many details about every event within its narrative, as its whole point is to study how large-scale patterns arise from very complex, detailed interactions of specific (perhaps particular) events. However, it would not especially benefit from the entire narrative being represented by episodic memory: if anything this would be rather limiting, as it would also aim to include many other kinds of information in the narrative which are not best represented episodically, such as information about broader trends which are not part of the individual's immediate surroundings. Rather, it benefits from placing episodic memories *within* a narrative, within a temporal structure *external* to the represented event. Perhaps representation of the broader temporal context of an episodic memory, either with narratives or temporal frameworks, combines in a significant way with episodic memory's other features?

6.6 Broader Temporal Context

Given the discussion in Chapter 4, there are two natural ways in which we might try to develop the idea that representation of an event's broader temporal context combines in a cognitively significant manner with core features of episodic memory: episodic memory's appearing as part of a narrative structure (specifically a narrative about oneself, an autobiography), and placing an episode within a holistic, comprehensive framework of times.

6.6.1 Autobiography

There is a large literature on 'autobiographical memory'. Sometimes this term is used as a synonym for episodic memory, but often it is used to include semantic memories about one's life.[236] Certainly, if one were to tell the story of one's life, it would not just consist a series of events, but also information about one's location, relationship status, job etc. at various times and perhaps also how one's personality, projects, and interests changed over time. And the sensory details of key events would often be their less important features. This should suggest that episodic memory is not especially useful in this context.

Nonetheless, the discussion of narrative understanding in Chapter 4 suggests a distinctive use for compressed information about a wealth of different features, of the kind imagery could provide, in narrative understanding. If such understanding can help us get to grips with complex patterns at least in a very local, constrained case, then having a great deal of information with respect to each event in the narrative would be a boon. And we have argued that the combination of memory traces and imagistic representation is especially suitable for such compressed representation. The most relevant kinds of narrative understanding here might not be autobiographical in the sense of covering one's whole life, but in the sense of covering a few (usually temporally local) events, such as events relating to one project one pursued.

One question here will be how far this sort of activity depends on representations of *particular*, as opposed to *specific*, events. Insofar as the aim is to get round the limitations of general learning to pursue learning which is still to some extent generalizable, it would seem that merely specific, not particular, events might be more worth learning about.

[236] Surveyed in Rubin & Umanath (2015: 15).

Another question will be just how much of a boon imagistic representation offers here, and just how significant this kind of narrative understanding is compared to the sorts of general learning discussed in chapter 5. These ideas merit further exploration, but the prospects for finding a combination of features here so significant as to overturn Chapter 5's account of episodic memory seem slim.

6.6.2 Temporal Framework

The idea that episodic memory needs to be embedded in a temporal framework—that episodic memories must include representation of when they occurred—seems to be weaker. The best case for it would be a way of showing that when episodic memory includes information about the temporal coordinates of each remembered episode, general learning can be used to reach models of systems that display non-Markovian dynamics. However, a few points can be made against this idea. First, for really intractably complex non-Markovian systems which cannot be dealt with using the non-episodic-memory based strategies of representing dynamically complex entities and narratives, it is unclear if general learning will actually uncover the non-Markovian dynamics unless fed a huge amount of data, and especially an impractically large amount of data for an individual's lifetime. Furthermore, it is not clear if, in the context of powerful general learning, non-Markovian dynamics will be such an issue any more. Recall that plausible examples of violations of the Markov property were often due to Markovian processes unfolding in unobserved variables. With general learning, one might be able to simply learn about those unobserved variables instead of modelling the non-Markovian process in observed variables that results. And it may be able to do this without representation of a temporal framework: as more complex models are developed, extra variables will be uncovered. Finally, for the remaining cases, where there are genuine non-Markovian processes,

but they are neither intractable nor solvable through modelling an unobserved process, general learning on representations of time may well help. But these are just special cases, and representation of time does not seem to be doing special or widespread work here which it would not be doing for many other variables.

Furthermore, it is not entirely clear if even human memories are all arranged straightforwardly on a temporal framework. We are capable of placing our memories on a temporal framework, or at least appreciating that they are events which could be so placed, given more information. But making representation of a temporal framework essential to episodic memory seems to require more than that. Episodic memories do prime temporally nearby memories; in free recall (when asked to talk about events without being guided by specific instructions), individuals do often recount events in forwards or backwards chronological order; and providing recall cues in forward or backward order leads to more accurate and detailed recall.[237] Furthermore, there is some evidence that the explanation of these findings (as well as findings related to event segmentation's effects on duration estimates) lies in a signal in the hippocampus which slowly evolves (although more quickly at event boundaries), so that at any given time it is in a state so specific as to in practice be unique to that time (but similar to nearby times' signals). This signal may be used as a temporal context signal when associated with new memories, and associations formed with that signal may help create associations with temporally nearby memories etc.[238] However, we often do not know the times our episodic memories occurred, or we know the time of day they occurred but not when exactly in our lifetimes, and in

[237] Discussion and references in Friedman 1993 and Aronowitz 2018.

[238] Discussion and references in Clewett et al (2019: 162, 165, 174).

general having a representation of when exactly the target event of an episodic memory occurred seems to be something which happens only for some episodic memories.[239]

In all, then, we do not seem to get any distinctive cognitive significance from combining the representation of time with other features of episodic memory: the best case for such distinctive significance relates to narratives that embed episodic memory alongside other sorts of representation, in a very specific sort of understanding; but the significance of that sort of understanding seems to be dwarfed by the kind of general understanding that can be achieved without temporal representation. Temporal representation is not implied by the core features of episodic memory, and it should not be counted as a core feature in its own right.

6.7 Consciousness, Metacognition, and Autonoesis

The most popular alleged additional core feature of episodic memory, however, is arguably not temporal representation of any kind, but autonoesis. 'Autonoesis' was coined by Tulving 1985, to capture a kind of conscious awareness that contrasts with 'noetic awareness' (associated with semantic memory). Tulving describes autonoesis somewhat differently in different places. I will focus on the two features which appear most prominently in the literature: that autonoesis is a phenomenological feature; and that it involves self-representation. I treat each separately, asking if either combines with other features of episodic memory in a distinctive, cognitively significant way.

Does retrieving an episodic memory entail that there is something it is like — a phenomenal character — for the subject to be retrieving that memory? Many either endorse or

[239] Friedman 1993; Rubin & Umanath 2015.

express sympathy with the idea,[240] although not all are convinced.[241] There are good reasons to be skeptical of any such proposal.[242]

Firstly, we should be wary of prematurely concluding phenomenal features are essential to any mental state. Compare perception. It was traditionally thought that perception is necessarily conscious.[243] However, various cases were found where subjects' perceptual systems seem to be processing stimulus-information, where this impacts aspects of their behaviour similarly to when conscious, but where subjects linguistically report not seeing the stimulus. While discussion continues about how to interpret such findings, the dominant view is

[240] Tulving 2005, Burge 2011, Markowitsch & Staniloiu 2011, Russell & Hanna 2012, Klein 2015. Clayton & Dickinson 1998 and others hesitant about describing animal memories as more than 'episodic-like' explain their hesitancy via uncertainty about animals' phenomenology. E.g. "The subjective experiences that accompany episodic recall require the re-experience of the past event and it also involves in oneself who travels back to a point in time and, therefore, is able to have a subjective sense of past, present and future time (chronesthesia). The definition in terms of these phenomenological constructs makes it extremely difficult, if not impossible, to demonstrate this type of memory in nonverbal species because there are no agreed upon non-linguistic behavioural markers of conscious experience." Martin-Ordas et al 2010: 331-332; cf. Tulving 2001: 1512, Eichenbaum et al 2010: 2281, Klein 2015: 23f., Fugazza et al 2016. Michaelian 2016 (e.g. pp. 13, 117f., 208, 232). and Cheng et al 2016 assume that consciousness is typical of episodic memory in humans although not strictly necessary, and suggest a number of possible roles it might play in the functioning of the system, , including one discussed below.

[241] Henke 2010, Templer & Hampton 2013. Boyle 2019 usefully shows how some of the alleged phenomenological features of episodic memory can be thought of in information-processing terms, although she presents this as showing 'impure phenomenology' rather than as an alternative to phenomenological approaches.

[242] The view here is also at odds with Keven 2016 and Mahr & Csibra 2018's proposal that both 'event memory' and 'episodic memory' be used to pick out distinctive kinds, where the former is something like my 'episodic memory', and the latter builds in many more features: their 'episodic memory' does not pick out a cognitively significant combination of features. Rubin & Umanath 2015 advocate replacing 'episodic memory' with 'event memory', defined in even more minimal terms than my 'episodic memory', whilst keeping a number of distinctions between event memories orthogonal to one another. This misses interesting interactions between some of these dimensions.

[243] For a particularly explicit example, see: "the fact that I am seeing something now, is obviously related to the fact that I am conscious now in a peculiar manner. It not only entails the fact that I am conscious now (for from the fact that I am seeing something it follows that I am conscious: I could not have been seeing anything, unless I had been conscious, though I might quite well have been conscious without seeing anything) but it also is a fact, with regard to a specific way (or mode) of being conscious, to the effect that I am conscious in that way: in the same sense in which the proposition (with regard to any particular thing) "This is red" both entails the proposition (with regard to the same thing) "This is coloured," and is also a proposition, with regard to a specific way of being coloured, to the effect that that thing is coloured in that way." Moore (1925). For the view that consciousness is generally extremely widely spread in the mind, see James (1889).

that these are cases of unconscious perception.[244] Dehaene & Naccache 2001 survey evidence for the unconscious occurrence of not just perceptual but also motor, cognitive, and emotional processes, finding relatively few processes that seem to require consciousness.

Thus, foreclosing the possibility of unconscious episodic memory by definition is suspect. Indeed, Henke 2010, Hannula & Greene 2012, and Olsen et al 2012 have already found evidence of hippocampus-based memories, encoded on the basis of a single event, having an impact on behavior, despite not being explicitly reportable.

Sensory or quasi-sensory (as well as emotional or quasi-emotional) phenomenology is a major part of the phenomenology of episodic memory. Rubin et al 2003 develop a similar position, supporting it with their empirical finding that subjects' ratings of the degree to which they recollected/re-lived their memories correlated with their self-reports about sensory imagery and emotions. However, unlike the Controlled Imagination View they think that the phenomenology of re-living itself, as opposed to just the processes that explain it, is essential to episodic memory. Several philosophers have made distinctions between kinds of memory by appealing to sensory imagery - see Brewer 1996: 23 for a survey - often due to introspective considerations like those in this section.

The cognitive significance framework provides a more systematic way to determine if phenomenology should be considered core to episodic memory. Phenomenal consciousness' contribution to cognitive significance will depend on our theory of phenomenal consciousness. If consciousness is entirely epiphenomenal, then it will come out as cognitively insignificant. But there are many theories of phenomenal consciousness which do not treat it as epiphenomenal. If one's theory of consciousness is a global broadcasting theory along the lines

[244] For both sides of the debate, see Peters et al 2016.

of Dehaene & Naccache 2001's, for example, capacities which depend on many subsystems of the brain coordinating and projecting information to one another may turn out to require phenomenal consciousness. It might be possible to argue on such a view that episodic memory requires just the kind of global broadcasting that is distinctive of phenomenal consciousness. It would need to be shown, however, that alternative means of achieving coordination would not work. In the absence of such a positive case, we should not deem phenomenal consciousness core to episodic memory.[245]

The other oft-discussed feature of autonoesis is that it is a variety of self- or meta-representation, representing something like THIS IS BASED ON MY EXPERIENCE OF A PAST EVENT.[246] Does this feature combine in a distinctive, cognitively significant way with other components of episodic memory?

Combining in a distinctive way is important. Everyone can agree that self-representation is cognitively significant in its own right. Redshaw 2014 argues that meta-representation is essential to episodic memory, plausibly claiming it allows more insightful, flexible use of episodic memories. But metacognition is required for insightful, flexible use of every mental process, for the same reasons as for episodic memory; and the distinctive, significant kind of

[245] Carruthers 2015: 77ff. argues that episodic memory does involve global broadcasting, because he believes that it involves running the sensory systems offline, and also believes in a model on which this can only be done by the mechanism he thinks is identical with consciousness - frontal areas recruiting sensory areas through attention. However, there are three reasons to be wary of this conclusion on current evidence. First, I do not take it to be empirically established that this is the only way sensory areas can be used offline, even in humans - the cognitive controller may be less sophisticated, without the ability to broadcast *globally* - it may be connected only to areas where the memory trace is stored (presumably hippocampus) and to parts of the sensory areas. Secondly, one might deny that frontal areas recruiting sensory areas is sufficient for consciousness, in the sense of consciousness in which one is interested. Thirdly, one might think that episodic memory can appear in less sophisticated forms, with simpler forms of cognitive control of sensory areas that are not sufficient for consciousness, in other animals.

[246] Autonoesis is understood in metarepresentational terms by Perner 2001; Dokic 2001; Templer & Hampton 2013; Redshaw 2014; Michaelian 2016; Mahr & Csibra 2018; and Fernández 2019, although motivations and exact meta-representational contents vary.

learning involving episodic memory discussed above does not require the relevant kind of flexibility.

Several authors argue that meta-representation is required for episodic memory to work, because without it the mind would confuse different cases of imagery — like imagination of counterfactual scenarios and imagination-based future planning — with episodic memory.[247] These processes will use overlapping mechanisms (offline perceptual machinery and simulation mechanisms in the hippocampus) for representing scenarios. But such representations should be put to different uses. Confusing merely imagining your spouse's being unfaithful with remembering it is not advisable. To avoid such confusions, the argument goes, each simulation of a scenario should include a further content specifying its status: in the case of memory, something like THIS IS BASED ON MY EXPERIENCE OF A PAST EVENT.

If this kind of self-representation is crucial to episodic memory's not being confused with other kinds of offline simulation (and hence to its cognitive significance), it should be included as core to episodic memory. However, meta-representation is not the only way to avoid such confusion.

We can think of the problem whose solution allegedly necessitates meta-representation as a problem of ensuring information flows down the right streams at the right times. Each instance of a simulation will need to be driven by the right sorts of input and give the right sorts of output. Episodic memory will need to be driven only by information from a memory trace of a past event and general knowledge about how the world is, not by pretense or information specific to other events; and it will need to send the results of its simulation to systems for learning models

[247] Versions of this idea appear in Tulving 2005; Goldman 2006; Buckner & Carroll 2007; Klein 2014, 2015; Michaelian 2016.

about how the world actually is, rather than systems dedicated to mind-reading others, considering how the world might turn out to be, how the animal would like the world to turn out to be. Simulations for counterfactual purposes should take in constraints specifying the counterfactual scenario to be simulated, and should not send their results directly to systems for learning about the actual world.

An alternative way of achieving the right flow of information would be the following. Suppose the mind's architecture is set up to have a few modes of operation, corresponding to memory, counterfactuals etc., each consisting in a certain pattern of excitation and inhibition of connections to and from the simulation systems. Switches between modes could arise spontaneously and automatically in response to triggers like the system for developing new models of the world becoming active, or the animal's facing a complicated decision scenario requiring planning. Given architecturally specified patterns of excitation and inhibition with automatic triggers, none of this need be governed by computations over meta-representations.

On a deflationary use of representational language, such modes of operation could be described as 'representing that the simulation represents a self-involving event'. But this description would not add explanatory value, and could misleadingly suggest a system that computes over meta-representations, as it would imply on a more robust understanding of representation. The system is no more doing that than a system that performs addition thereby represents itself as enacting a commutative and associative operation that is one of the fundamental operations of arithmetic.

My aim in this section has not been to decisively refute all possible arguments for autonoesis's being core to episodic memory. Autonoesis may yet turn out to be crucial to the

kind of learning discussed above, whether for a better version of the canvassed reasons, or for some other reason not discussed here. But I have shown that some of the initial appeal of treating autonoesis as core to episodic memory is misleading, and how such debates should proceed on the cognitive significance framework.

6.8 Conclusion

Episodic memory could come in many varieties, and we should be live to the possibility that different species will have different varieties. To count as having episodic memory, a creature will need to use imagistic representation of particular past events on the basis of a memory trace, but there are many ways to do this, and many additional features which might be added to a system which does this. It could be based on different sensory modalities or other weakly modular systems being run offline. It could be based on different kinds of memory trace with corresponding forms of reconstruction. It could be based on intellectualized or architecturally-based means of representing particular past events. It could include various forms of temporal structure within its representations of events —but it need not—, and the events it represents could be placed in a broader temporal context, such as a narrative or temporal framework—but need not be. And it could be conscious or unconscious. These possibilities just scratch the surface of the kinds of variety episodic memory could exhibit in different species.

But what is the evidence for episodic memory in different species? And how do all these distinctions and constitutive accounts affect our interpretation of that evidence? That will be the topic of Chapter 7.

Chapter 7: Which Animals Have Episodic Memory? The Evidence

7.1 Introduction

We now have an account of what it would take to count as having episodic memory, and why this would be important. But this raises the question: which (if any) animals have episodic memory so-defined? What light can our account (in combination with the attitude towards different kinds of evidence for mental states in animals defended in chapter 2) shed on the evidence the literature already offers as evidence for or against animals having episodic memory? And does it suggest new lines of empirical investigation?

As mentioned above, there is a sizeable literature purporting to speak to the question of which (if any) animals have episodic memory. Much of this interest was thanks not to an explicit account of the cognitive significance of episodic memory, but to pronouncements by Tulving and others that episodic memory is unique to humans. Whenever it is prominently claimed that some capacity is unique to humans, sooner or later a flock of studies arrives purporting to show a version of that capacity in other animals. And one of Tulving's relatively early books opened as follows:

> "Remembering past events is a universally familiar experience. It is also a uniquely human one. As far as we know, members of no other species possess quite the same ability to experience again now, in a different situation and perhaps in a different form, happenings from the past, and know that the experience refers to an event that occurred in another time and in another place."[248]

Tulving 2005 argues for this position at more length, as do Suddendorf and Corballis 1997, 2007. They offer two main sorts of consideration for this position. Firstly, they point out

[248] Tulving (1983: 1).

(as Tulving does in the above quote) that there is very little evidence that animals do have episodic memory. This is rather a weak consideration, unless one also shows that one would have had strong evidence of episodic memory if animals did have it. But these authors do not argue for this conditional: on the contrary, they criticize various experimental paradigms as *not* being such as to provide evidence of episodic memory even if animals pass them. Their second, somewhat stronger consideration, draws on a very demanding conception of episodic memory as requiring sophisticated versions of features like autonoesis and chronesthesia (or conceptually self-aware 'mental time travel'), and argues that animals are unlikely to have these features. However, Chapters 5-6 have argued that it is a mistake to think that autonoesis and even temporal representation are essential to episodic memory.

The failure of some arguments that only humans have episodic memory, of course, does not imply that animals do have episodic memory. And all the evidence that has been collected on that question in response to Tulving's challenge deserves to be interpreted in light of the constitutive account of episodic memory I have interpreted.

I will discuss three main lines of evidence, and how we should think about them in light of my account of episodic memory. §7.2 will discuss the most popular behavioural paradigm for studying episodic memory in animals: showing that animals can remember what, where, and when some event (typically a food caching event) happened, perhaps alongside some other information. §7.3 will discuss experiments based around the idea of incidental encoding: memory for features of events which subjects could not know they would be asked about. §7.5 discusses neural evidence relating to the replay of sequences of events in the hippocampus. I will also briefly discuss additional lines of evidence in the literature in §7.6 In each case, I will show that the paradigms in question do not offer direct evidence for episodic memory, at best only

providing direct evidence for one of the core features of episodic memory. However, in line with the points made in Chapter 3, this does not mean that these paradigms offer no evidence—or even only weak evidence—for episodic memory. They do mark serious progress on the issue. Furthermore, I will show how my account of episodic memory offers at least broad-brush recommendations for ways that such paradigms could be developed to test more directly for episodic memory: I will lay these out in §7.4 and as part of my discussion of hippocampal replay in §7.5.

7.2 WWW-Memory

The best-developed line of behavioural evidence on animal episodic memory focuses on showing that various animals can form what I'll call 'WWW-memories'. It aims to show that a creature has representations of particular events that integrate information about a number of features of those events. In particular, researchers try to show that creatures represent *w*hat kind of event happened; *w*hen it happened (typically this means *how long ago* it happened, though some studies focus instead on other temporal properties like *what time of day* it happened); and *w*here it happened. We discussed details of the original version of this sort of experiment—Clayton & Dickinson 1998's work on scrub jays and the subsequent work on scrub jays in their lab—in Chapter 2, so I will not rehash those details here.

The Clayton lab's scrub jay studies inspired studies using structurally similar designs to find evidence for WWW-memories in a wide range of other animals, including other corvids like magpies;[249] other food-caching birds like Black-capped chickadees;[250] rodents such as rats and

[249] Zinkivskay et al 2009.

[250] Feeney et al 2009.

meadow voles;[251] Yucatan minipigs;[252] dogs;[253] Rhesus monkeys;[254] great apes;[255] and even the cuttlefish, a mollusc whose closest surviving relative is the octopus.[256] This sort of behavioural paradigm can also be combined with neurobiological work, and some have started to do this, especially in rats.[257]

It is a live issue how exactly WWW-memory relates to human episodic memory. Some in the literature are happy to refer to WWW-memory as 'episodic memory', but many scientists are cautious, referring to WWW-memory only as 'episodic-like' memory. There is a widely shared sense that there may well be differences between episodic memory and WWW-memory, but uncertainty and disagreement about what exactly those differences are. It is therefore common to use terms like 'episodic-like memory' or 'proto-episodic memory' to hedge about how WWW-memory relates to episodic memory.[258]

In humans, WWW-memory is not sufficient for episodic memory. One can have a semantic memory of the WWW-features of an event, without ever reconstructing the event (whether with offline peripheral or other systems). For example, most of us have semantic

[251] Zhou & Crystal 2009; Ferkin et al 2008.

[252] Kouwenberg et al 2009.

[253] Kaminski et al 2008.

[254] Hoffman et al 2009. With a slightly different set-up, Hampton et al 2005 found that Rhesus monkeys failed a www-task. It is an important question what the relevant difference between the cases might be (assuming this divergence is not just a result of noise)

[255] Again, there are mixed results here, with Martin-Ordas et al 2010 finding positive results and Dekleva et al 2011 finding negative results.

[256] Jozet-Alvez et al 2013.

[257] e.g. Barker et al 2017.

[258] e.g. Clayton & Dickinson 1998, Allen and Fortin 2013.

memories for what, where, and when concerning historical events and events of family or personal significance such as our own births.[259]

I argued in Chapter 6 that 'when' information is not necessary for episodic memory either; and similar arguments could be given for 'what' information. Neither of these is required for the distinctive, cognitively significant role episodic memory can play. And in any case, most of us humans do seem to have episodic memories that include only rather scant where-information (especially with events from childhood), or rather indeterminate when-information that perhaps only narrows down the episode's time to within a decade or so.

Failing to be necessary or sufficient for episodic memory does not mean that WWW-memory is irrelevant, however. WWW-information may well turn out to be typical to episodic memory in some or many species. Humans often have some memories we can not locate, but it has been known for centuries that memory has a special relationship to space in the sense that one way of enhancing memory for items is to use the method of loci—to intentionally form associations between those items and locations on a map. §6.6 showed that episodic memory based learning could well be used in conjunction with autobiographical representations or temporal frameworks as a special case to learn various temporal patterns. And WWW-information is a good way of guaranteeing particularity: it is plausible that events are at least partly individuated in terms of their spatio-temporal locations and kinds, so a system's always representing these features in enough detail would be one way of ensuring that its representations are always about particular events. It is just not the only form such a system could take. All this means that having WWW-memory *could* turn out to be correlated with having episodic

[259] Templer and Hampton 2013. Easton et al 2012 find evidence that when humans are given WWW-tests comparable to those given to animals, we tend to solve them without relying on episodic memory.

memory. *If* that turns out to be true, then evidence for www-memory would also be evidence for episodic memory, at least for animals that are relevantly like humans (i.e. such that this correlation is likely to hold in their case).

What does it mean to be relevantly like humans here? A natural response is to appeal to evolutionary considerations—some combination of phylogenetic relationships and similarities in selection pressures. As we saw in Chapter 3, this is a dangerous game, but there are a few advantages in this case. First, we have independent evidence that homologous structures are used in human episodic memory and avian www-memory (though not in cuttlefish). We know that the hippocampus is homologous in birds and mammals, and connects to many homologous areas.[260] And we know that the hippocampus is involved in human episodic memory and seems to both be involved in memories for food caches (i.e. disrupting the hippocampus disrupts the ability to recover caches) and be under selective pressure for food caching (i.e. variation in hippocampus size across species is more powerfully predicted by whether a species caches food than other variables) in birds.[261] Furthermore, other features of corvid cognition and brain structure are known to resemble primate cognition in ways suggesting convergent evolution, so the cognitive background into which these kinds of memory fit might be similar.[262] Finally, whereas we saw in Chapter 3 that evolutionary inferences are particularly dicey in cases where only a small number of species is known about, and especially for cases where the only species involved are humans and one other species, in this case www-memory (if not episodic memory) is known about in a wide range of species from quite different lineages.

[260] Reiner et al 2004.

[261] Sherry & Vaccarino 1989; Sherry et al 1989; Garamszegi & Eens 2004; Lucas et al 2004; Sherry 2011.

[262] Clayton 2015; Güntürkün & Bugnyar 2016.

All this might be taken to support attributing episodic memory to scrub jays. But this would be too quick. First, selection pressures are clearly not identical for humans and scrub jays. Food caching is not as much of a big deal in primates as in most of the species tested for www-memory. And this could be important: food caching animals could have much more specialized, much less flexible, systems of memory than primates if their memory is strongly dedicated to the narrow task of remembering things about stores of food.

But there is also a more important limitation, which will come up throughout this chapter every time the involvement of the hippocampus is brought up as evidence that an animal has episodic memory. It is extremely unlikely that the hippocampus' function is *just* episodic memory. It almost certainly is involved in many other processes, even within humans. It is involved in many other processes involving imagistic cognition and reconstruction (or construction) of scenarios, such as imagining the future, counterfactual imagination, and generic imagistic memory.[263] It seems to be heavily implicated in navigation and representation of spatial maps, especially in other animals,[264] although Eichenbaum 2017 argues that it has a more general function than this, relating to the representation of relations generally (especially in abstract domains and time), rather than just spatial relations. And its more historically distant functions may have specifically involved navigation based on olfaction. [265] Indeed, as we will see in §7.5, the very same hippocampal cells, doing the very same sorts of things at a local level, may well be implementing different overall functions in different contexts. So inferring from the

[263] Hassabis et al 2007; Schacter et al 2007, 2011, 2018; De Brigard 2014; Michaelian 2016; Barry & Maguire 2019.

[264] Moser et al 2014. We will discuss some of this evidence in §7.5.

[265] Jacobs 2012.

hippocampus' involvement in a process to that process being or involving episodic memory is not a solid inference.

Thus, there seem to be differences, at least sometimes, between WWW-memory in humans and episodic memory in humans; and evolutionary considerations do not mean that we can rely on limited similarities between bird WWW-memory and human episodic memory to close this gap. We do not yet have a full understanding of just what the difference between WWW-memory and human episodic memory is, however.

The biggest divergence is given by the core components of episodic memory which we have not discussed yet in this section: its use of imagistic representation for reconstruction from a memory trace. Is there any connection between WWW-memory and this?

One could argue that images, or representations in offline perceptual systems (or, with §6.3, weakly modular systems such as motor systems), will include a great deal of spatial information. Perception typically represents how things are *around the perceiver*, filling out the space around the perceiver from a spatial perspective; and imagistic cognition can be expected to inherit this much spatial content. So the where component has at least *some* connection to imagery; but it is clearly a weak connection, as we can represent locations in many other ways. However, the emphasis on just three features, one of which (*What*) is categorical rather than based around rich, specific features, is at odds with the important role imagery plays in representing a large number of features in a coherent whole.

7.3 Incidental Encoding

A different behavioural paradigm has more obvious connections to the style of learning algorithm I discussed in Chapter 5, but as currently implemented falls well short of providing good evidence of episodic memory. Zentall (2005, 2006, 2013) has argued that a good form of

evidence for episodic memory is animals' ability to answer an unexpected question about an event. This is sometimes described as showing 'incidental encoding': storing information without knowing that it has to be remembered later. Zentall and colleagues have found ways of showing that pigeons can do this,[266] and versions of the paradigm have also been applied successfully in rats and dogs.[267]

Zentall's rationale for this approach is this:

> "Events that one considers important are often encoded purposefully as a semantic memory. For example, one may attempt to encode the name of a person one has just met because one may want to recall that memory at a later time. However, it is unlikely that other, incidental aspects of that encounter (e.g., what they were wearing) would be purposefully committed to memory."[268]

This rationale is unconvincing. It is hardly plausible that all non-episodic forms of memory require purposefully committing something to memory. It is extremely common to pick up on statistical regularities involuntarily through automatic learning processes. Even for Zentall's case of learning someone's name, we do sometimes come to recall the names of certain people without either deciding to do so or having an episodic memory involving their name: for example, if one is a student in a large class and another student on the other side of the class (who one does not ever expect to meet) keeps being called on.

This problem with Zentall's motivation shows up in how he implements his test, and in particular with the features he chooses to test for. They have no relationship to anything genuinely distinctive of episodic memory, such as imagery. Instead, his unexpected 'question' is about the last action the pigeon performed, the locations they most recently pecked, and whether

[266] See Zentall 2013 for details.

[267] Zhou et al 2012 for rats, Fugazza et al 2016 for dogs.

[268] Zentall (2013: 579).

they have recently been fed (Fugazza et al 2016's test on dogs, meanwhile, involves remembering a command). As a result, it is very plausible that the information he tests for was encoded non-episodically, and these experiments provide only weak evidence for episodic memory.

There is, however, something importantly right about the general idea of testing for incidental encoding. Incidental encoding clearly can be linked to Chapter 5's ideas about imagery storing many features of an event, and the idea that we need to do so because we may unexpectedly need to run a computation on some previously insignificant-seeming feature. And linking to these ideas about imagery was precisely the main limitation of the WWW-memory paradigm. This suggests that the incidental encoding paradigm, asking about the right sorts of features (preferably genuinely obscure perceptual features), while harder to implement experimentally, could provide a meaningful piece of evidence for episodic memory in animals, alongside other paradigms. Indeed, both Crystal 2010 and Zentall 2013 recommend integrating tests of incidental encoding with other paradigms. But what other sorts of test does the constitutive account of episodic memory suggest?

7.4 Possible Direct Tests for Core Features of Episodic Memory

While we can draw on our constitutive account of episodic memory to guide empirical research (and, as Chapter 5 advocated, vice versa: there should be an iterative process of refining the constitutive account and empirical evidence in light of one another), we should not aim to develop one single paradigm which will function as *the* test of episodic memory. Neither should we expect to find evidence which is entirely unambiguous—which is entirely incompatible with all alternative explanations besides the animal in question having episodic memory. Rather, we should aim for converging lines of evidence from multiple paradigms which all have some

connection to the constitutive account of episodic memory (even if that is just a known correlation with some of the features it proposes). The goal, as Chapter 3 made clear, is to find ways of raising credences in the correct view and lowering them in alternatives.

Furthermore, we should not expect the same paradigms to work for every species. We need to take into account the differing preferences, habits and capabilities of creatures with very different lifestyles, as otherwise we will see many spurious 'negative' results where animals 'fail' tests due not to lacking the capacity but due to some of these other features, or simply finding no clear results at all.[269] So expertise with particular species will be required to develop versions of any of the tests proposed below for actual studies: I simply include suggestions meant to stimulate those with the relevant expertise and to illustrate the general direction a cognitive-significance-based constitutive account of episodic memory can be taken.

The offline use of peripheral systems may seem hard to test, but it has already been tested in other contexts in the case of vision. There are many behavioural measures of visual imagery in humans, several of which are often thought to pick up specifically on the visual system being used offline. Many rely on interviews or questionnaires carried out in language, and are hence not suited to the study of non-human animals.[270] Other tasks, however, can be adapted to non-linguistic creatures. One of the most famous sorts of test for mental imagery has already been adapted for animals, albeit on a small scale. These are mental rotation tasks.

[269] e Waal 2016 outlines numerous examples of this happening in the animal cognition literature. The participants in the www-literature are sensitive to this point, taking advantage of natural behaviours like food caching in some species and different behaviours—like idiosyncratic features of the meadow vole's reproductive cycle (Ferkin et al 2008)—in others.

[270] E.g. the Vividness of Visual Imagery Questionnaire (Marks 1973), or the spontaneous use of imagery scale (Reisberg et al 2003).

When asked to make judgements about visually presented figures requiring that we consider what they would look like if rotated, human reaction time linearly increases depending on how much rotation would be required.[271] Exactly how to interpret this finding is highly controversial. It is interpreted by some as suggesting that imagination uses an analogue or image-like representational format,[272] though others vehemently deny this.[273] One version of the 'image' interpretation says that the brain's way of solving the task is to use the visual system offline, to construct a series of images corresponding to what the subject would see as the image was gradually rotated—and there is now a wealth of neuropsychological and imaging evidence suggesting that this is in fact what happens.[274]

Suppose one buys this sort of interpretation of mental rotation tasks. There are various options for developing tasks along these lines without requiring language. One can show a shape, then after a delay give subjects a choice between rotated versions of that shape and its mirror image, rewarding them for selecting the original shape. Or one can show two rotated shapes at once and reward subjects for selecting correctly whether the images match or not. Tests of this sort have been done on a small number of animals from a handful of species, including pigeons—whose reaction time does *not* seem to depend on the amount of rotation,[275] a California sea lion, who *did* show similar reaction time effects to humans,[276] and rhesus monkeys, where

[271] Shepard & Metzler 1971, where the task is to decide whether two figures are rotated or rotated and reflected versions of each other. There are various complications to and developments on the basic finding - see Shepard and Cooper 1982 for discussion of many.

[272] E.g. Shepard and Metzler 1971 and Mauck and Denhardt 1997 describe things this way.

[273] E.g. Pylyshyn 2002.

[274] Georgopoulos et al 1989; Kosslyn et al 1999, 2006; Slotnick et al 2005; Pearson et al 2015.

[275] Hollard & Delius 1982.

[276] Mauck & Denhardt 1997.

findings differed between the three individuals tested.[277] The most important point for our purposes, however, is not specific to mental rotation tasks. The compelling feature of mental rotation tasks in providing evidence of use of a perceptual system offline is (in addition to lesion data and neural data suggesting visual areas are involved in mental rotation tasks) that subject performance shows a distinctive signature which the online system would also show when performing an equivalent task online. In this case, the signature is reaction times, which in turn reflect the way the visual system deals with motion in space. But in general, any weakly modular system being run offline would show quirks in reaction times, error rates etc. that it also showed in its online functioning, and these quirks would not typically be predicted by alternative models of task performance.[278]

Another classic kind of behavioural evidence that can suggest the use of particular systems offline is distinctive patterns of breakdown or disruption of performance due to particular kinds of load. In humans, articulatory suppression—repeating some word like 'the' whilst trying to do a task, typically reduces performance in verbal short-term memory tasks, but does not reduce performance in visual/spatial tasks, whereas the opposite effect can be produced by having subjects to tap on the table in a figure of 8 pattern.[279] It may also be possible to produce system-specific disruptions by making animals perform multiple tasks at once, or by using certain sorts of masking or bombarding them with lots of irrelevant stimulation in the relevant modality, and to thereby disrupt performance only (or, more likely, especially) on tasks hypothesized to involve using a specific system offline.

---

[277] Köhler et al 2005.

[278] See Parsons 1994; Johnson 2000 for reaction time results similar in character to visual mental rotation experiments.

[279] See Salway & Logie 1995 for discussion of these sorts of strategies and their histories.

We can also have neural evidence for the offline use of weakly modular systems, provided we know either a distinctive region, circuit or processing signature of the weakly modular system. Kosslyn and others have provided considerable evidence that visual cortex is used in visual imagery, for example, including evidence that damaging visual areas can impair visual imagery, as well as fmri and other imaging evidence suggesting activation of visual areas during visual imagery tasks, and decoding analyses of such imagery suggesting that similar patterns of activation occur in visual areas during seeing stimuli and during imagining those stimuli.[280]

For direct evidence for episodic memory, we need not just evidence for imagistic representation (offline systems), but also for its being used in the right way—to represent a particular past event, on the basis of a memory trace. This will require pursuing tests for the different components in the same tasks, and tests for interactions between these components. Perhaps it would be possible to design behavioural tasks requiring the combination of evidence for the use of imagery (including incidental encoding of obscure perceptual features and mental rotation) and for the representation of particular past events (such as www-tasks, or tasks which require keeping multiple qualitatively identical events distinct). An example which clearly wouldn't work as an experimental design in practice, but does clearly illustrate the kind of thing I have in mind, would be a www-task combined with mental rotation: one could teach subjects that e.g. the symbol _|- signals reward in this location, provided it is at least 6 hours since last checking that location, then in test rotate the symbol and see how long their response takes.

[280] E.g. Wheeler et al 2000, Slotnick et al. 2005, and see Goldman 2006 ch. 7. See Georgopoulos et al. 1989 for single-unit recordings in rhesus monkeys suggesting motor areas behaves similarly during imagined and actual actions in these sorts of tasks.

Most convincing, of course, would be being able to manipulate performance on www-tasks by specifically manipulating components—especially with selective lesions/optogenetics etc. or alternatively with more behavioural system-specific distractors as discussed above. Neural evidence will also be essential to seeing exactly how the different components and their interactions are implemented in particular tasks. To appreciate how neural evidence could be used in combination with behavioural evidence to test for episodic memory, it will be important to have in hand an understanding of the most prominent sort of evidence for episodic memory in animals in the last decade: hippocampal replay.

7.5 Hippocampal Replay

Although Suddendorf and Corballis (1997, 2007) argue that episodic memory is unique to humans, Corballis 2013 claims that hippocampal replay provides evidence for episodic memory in animals (although Suddendorf 2013 disagrees). As discussed in Chapter 5, Cheng et al 2016 suggest that hippocampal replay is the essential core of episodic memory, a mechanism underlying a homeostatic property cluster. What is hippocampal replay, and how important is it to episodic memory?

The rat hippocampus (especially areas CA1 and CA3) has place cells.[281] These are cells which fire when the subject is at a certain location, so that by analyzing recordings from multiple

[281] Most of the results I describe in what follows are from electrophysiology in rodents, except where noted.

place cells, neuroscientists can decode the subject's position at any given time.[282] There is strong evidence that place cells are involved in navigation in some way.[283]

Place cells do not just fire in response to the subject's current location, however. Sometimes numerous place cells are active in a burst, a Sharp-Wave Ripple (SWR). These occur in several contexts. They were first discovered in slow wave sleep and resting, but have also been found when the animal is standing still and eating or performing some other simple task, and, most suggestively of all, while paused at decision points.[284] A similar phenomenon to SWRs, theta sequences, is found while the animal is moving.[285]

The striking thing about SWRs and theta sequences is that they are not just random bursts of place cells: they seem to correspond to paths through space, at least some of the time. This is most clearly seen in data from animals recorded whilst moving on a track with only relatively simple routes available, to reduce the number of locations and routes that might be encoded. Using this strategy, SWRs have been found corresponding to routes on the track, during sleep following the training,[286] awake but paused (at a decision point or at the end of the route),[287] and in a rest box outside the track.[288] Sometimes these are paths through space the

[282] There are some complications to this picture, for example space may not be all that can be decoded from these neurons (e.g. Eichenbaum 2017); there are also closely related cells coding for other aspects of space, like grid cells and head direction cells (Moser et al 2014); some neurons do not have unimodal place fields but fire in multiple locations; which place they code for changes over time, with the overlap in response of these populations changing radically in just 30 days (Ziv et al 2013) and even varying within one session by task (for references, see Eichenbaum 2017: 1789f.).

[283] E.g. O'Keefe & Speakman 1987

[284] Foster & Wilson 2006; Diba & Buszaki 2007.

[285] Feng et al 2015; Kay et al 2019.

[286] Lee & Wilson 2002.

[287] Foster & Wilson 2006.

[288] Karlsson & Frank 2009.

animal has previously traversed, either tracing the past route in the order it occurred or in reverse.[289] It may also be that SWRs can construct new trajectories the animal has never taken before, especially in novel environments.[290] They can start from the position the animal is currently in and extend from there (either backwards into the just travelled route or forwards into the possible next route), especially when the animal is paused in a maze, but (especially in sleep, whilst resting etc.) they will often start somewhere else.[291]

SWRs in the hippocampus are known to affect other areas of the brain. Logothetis et al 2012 combine electrophysiological recording from hippocampus place cells with whole-brain fmri in primates, and find that SWRs are associated with increases in activity in many other areas (especially more cortical, cognition-associated areas like Entorhinal Cortex and PFC, rather than areas associated with action and motor skills, like the Cerebellum and Basal Ganglia, which are downmodulated). Jadhav et al 2016 also find hippocampal SWRs to be associated with (and, for the most part—with interesting exceptions involving the auditory cortex—to precede) increased firing in cortical areas. Both Euston et al 2007 and Peyrache et al 2009 find not just hippocampal SWRs, but also that these are associated with replay of previous patterns in Prefrontal Cortex during sleep. Most suggestively for our purposes, Ji and Wilson 2007 found that V1 activity during sleep could be matched to the trajectories followed by SWRs in hippocampus and to V1 activity while following those same trajectories earlier.[292]

---

[289] Wilson & McNaughton 1994, Lee & Wilson 2002, Foster & Wilson 2006; Diba & Buzsaki 2007.

[290] Gupta et al 2010; Dragoi & Tonegawa 2011; Ólafsdóttir et al 2015. See Silva et al 2015 for a critique.

[291] Gupta et al 2010, Karlsson & Frank 2009. Theta sequences—which occur whilst the animal is engaged in action, unlike SWRs—seem to typically start a few steps back from where the animal is now and run to slightly into a possible future trajectory (Feng et al 2015).

[292] Rothschild et al 2017 find similar results for the auditory cortex.

There is considerable debate about the possible function(s) this activity plays. Interactions between hippocampus and cortex are implicated in many processes in humans, and there are not only multiple distinctive functions suggested by SWRs, but also evidence favouring multiple functions.[293] They could be used as holistic representations of sequences for all kinds of purposes (see Chapter 4), including narrative understanding of causal sequences, and could be used to simulate sequences for all kinds of purposes: counterfactual reasoning and causal understanding, prediction etc. The main proposals in the literature, however, relate to either planning, memory consolidation, or memory retrieval.

SWRs could well be used for planning and deciding between routes through space.[294] This could be the case whether they involve replay of memories of taking those routes before, or involve 'preplay' of unvisited routes. We could also imagine them having either direction of fit—they could be used either to simulate the outcome a given choice would have before committing to that choice, or to represent an intended path in order to derive further actions—and we could imagine reverse replay starting from a goal and proceeding backwards to figure out possible routes to that goal—although the option of simulating paths to discover their consequences is generally the default view in neuroscience.

There are various pieces of evidence suggesting SWRs are used for planning and decision-making, at least sometimes. For example, Jadhav et al 2012 used electrical stimulation to disrupt SWRs in animals while they were learning a spatial alternation task (a maze where the

[293] Ólafsdóttir et al 2018 argues that SWRs probably shift their function depending on the context and tasks required of them: it is particularly plausible, for example, that SWRs in sleep are used for memory consolidation or retrieval, while SWRs while paused at a choice point in a maze are used for planning. There is a great deal of suggestive evidence about different kinds of SWRs occurring in different contexts which supports this view, e.g. when rats stop for reward at the end of a track, SWRs tend to be in the reverse direction, but when at a choice point they tend to be in the forward direction. (Diba & Buzsaki 2007).

[294] Diba & Buzsaki 2007; Jadhav et al 2012; Pfeiffer & Foster 2013; Singer et al 2013.

animal has to choose between different prongs). This made for slower learning and worse performance. And Wu et al 2017 found evidence for replay of paths that previously led to punishment causing animals to avoid those paths.

There are a number of ways in which SWRs might be used in learning and memory consolidation. One is if there is something right about the Systems Consolidation view of memory consolidation.[295] On this view, the hippocampus forms connections very quickly but is unstable, whilst the cortex forms connections more slowly but is stable in the connections it does form,[296] so memory consolidation is about transferring traces from the hippocampus to the cortex. The Systems Consolidation view was initially motivated by retrograde amnesia: cases where human episodic memory for distant events is preserved but memory for recent events is disrupted by lesions to the hippocampus.[297] It is possible that hippocampal SWRs are repeatedly replaying memory traces (especially during sleep) to provide the cortex with more training and thereby consolidate more stable connections. But replay could also be used for memory consolidation on other views of consolidation, for example if SWRs are used as part of a process of fitting the individual event into a broader theoretical context and this helps with consolidation. Either way, there is evidence for a connection between SWRs and memory recall. The number of ripples increases after learning something new, and predicts memory recall in rats and humans.[298] Using electrical stimulation to suppress SWRs which occur during sleep

[295] Squire & Alvarez 1995.

[296] This might be because of network properties, or (as Barry & Maguire 2018 emphasize) because of differences in cell types and neurogenesis.

[297] There is controversy over this phenomenon, as hippocampus does seem to still be used in at least some very old memories; but this may be because while the trace has been transferred to the cortex, hippocampus is used as part of the retrieval and reconstruction process (Barry & Maguire 2018 fig. 2).

[298] O'Neill et al 2010.

following training impairs memory and slows down learning.[299] And Maingret et al 2016 were able to *enhance* recall through manipulating the coordination of PFC activity and hippocampal SWRs during sleep. It is also plausible that SWRs are specifically useful for learning or consolidating representations of (generic or particular) sequences, and in associated processes of event segmentation.[300]

Should we think of SWRs in rodents (or any other animal) as providing reason to think these animals have episodic memory? It is notable that several of the effects described above did not have to do with episodic memory, instead relating to prediction, or to other kinds of memory (it is not just episodic memories that need to be consolidated, for example: this will also be true for any memory, and especially memories formed very quickly).

One reason to think SWRs are evidence for episodic memory is that they occur in the hippocampus, and the hippocampus is important to human episodic memory. §7.2 gave reasons to think this is a pretty feeble reason generally, but here it is even weaker. This is because it is hard to directly observe whether humans even have SWRs without invasive surgery, let alone whether SWRs play the same cognitive roles. There is evidence for SWRs in monkeys.[301] But in humans, most evidence is based on analyzing EEG data rather than direct recording.[302] This evidence strongly suggests that there are distinctive ripple-like signals in the hippocampus e.g. at

[299] Girardeau et al 2009.

[300] Clewett et al 2019; Ólafsdóttir et al 2018.

[301] Logothetis et al 2012.

[302] Surveyed in Clewett et al (2019: 170).

event boundaries, and these do impact memory consolidation,[303] but this does not establish that these ripples consist of place cells firing in order.

A better reason to think of SWRs as evidence for episodic memory is their simulation-like character: they do seem to involve operating a specialized system offline in my sense. Place cells are being used to represent other positions, in virtue of their specialization for normally having the function of representing the subject's *current* position.[304] If this were the extent of the connection to imagistic representation, this would not be overwhelmingly interesting. While this does seem to be a case of operating a system offline, a system dedicated to representing just the space around the subject would not on its own allow for a great deal of compression and reconstruction of a wide range of features of a scenario: perhaps it would allow us to get away without explicitly storing spatial relations between different landmarks, but that would be about it. However, there are strong reasons to suspect that place cells could play a key role in compressing and reconstructing a great deal of other information besides location and spatial information. First, there is the live empirical possibility suggested by Eichenbaum (2017) and others that location is just the easiest kind of information for neuroscientists to test for and decode, and that these cells are actually specialized to represent many other abstract 'spaces' alongside location. Even if this turns out to be false, however, we know that place cells are relatively directly connected to sensory areas, and in particular (from the Ji and Wilson 2007 and Rothschild et al 2017 studies cited above) that hippocampal SWRs are connected to sweeps

[303] Axmacher et al 2008.

[304] I continue to use 'offline' here in the same way as introduced in §5.5.1 and generalized in §6.3: making use of systems that are specialized for a certain kind of computational role, because they are so specialized, but without their playing that computational role in the usual way. In the context of SWRs, 'offline' is sometimes used slightly differently, to contrast the occurrence of SWRs during sleep and rest with their occurrence during action and decision-making (the latter get counted as online in that sense but offline in mine).

through offline states in sensory areas. Another suggestive feature is that many of the contexts where these events occur are intuitively contexts where humans often use conscious mental imagery to plan, imagistically remember, dream or daydream, all activities thought to implicate the hippocampus in humans.

One reason to nevertheless be suspicious of the imagistic nature of SWRs is their speed. They typically only last less than half a second for an entire sequence, often running at about twenty times the speed of the events encoded.[305] Episodic memory, meanwhile, though it can involve a zoom through events, often involves lingering on particular events at their actual speed (think of remembering a climactic part of a musical performance, where tempo is important) or sometimes even slowed down (this is often how people report remembering traumatic (or just significant) but fast events: seeing the bullet or the ball travelling through the air in slow motion). This issue of speed is not a decisive problem for the advocate of the view that SWRs are tied to the imagery involved in episodic memory—there are open empirical possibilities that would assuage these worries. It may be that fast hippocampal SWRs are just what we have picked up on experimentally, and that they are either connected to occasional slower, less obvious (and hence thus far undetected) patterns of hippocampal activity that sometime plays out events at slower speeds in just the same way that SWRs play them out at high speeds; or they may cause slower ripples in other areas, with each millisecond of SWR in the hippocampus translating into many more milliseconds of more detailed sensory activity. Or (given the discussion in §6.7), it may be that SWRs implement episodic memories and offline processing generally, and that most episodic memories are unconscious: the episodic memory

[305] Ólafsdóttir et al 2018: R38.

retrieval events that humans find introspectable may be *atypical* of episodic memory generally in being slow, contra views that treat conscious episodic memories as the central case.

Another worry for the view that SWRs suggest the kind of simulation involved in episodic memory is how far SWRs are actually involved in reconstruction of events from memory traces, in retrieval events that are then used for broader computations, and how far SWRs are just about memory *consolidation*. If SWRs are just used for consolidation, this would explain their fast time-scale but also undermine the view that they are clearly connected to the kind of imagistic representation most important to episodic memory and its potential distinctive role: imagistic reconstruction of a scenario from a memory trace at retrieval. Better cases for SWRs being used for this kind of activity are the cases where SWRs arise at decision-making points and guide behaviour, such as in the SWRs corresponding to paths the animal had previously found to be aversive in Wu et al 2017.

However, in cases of SWRs being used at decision points, we need to ask whether this activity represents a particular past event, rather than a generic, multiply instantiable event-kind or a possible future event. Even if the activity is in some sense derived from past experiences with this particular path rather than being an extrapolation based on general knowledge of the landscape,[306] this does not make it a representation of those particular past experiences, any more than if you imagine your route to work this needs to be an episodic memory of a particular occasion on which you took that route. Testing for the use of representations of particular past events is tricky: it requires setting up situations where repeatability of a specific event-kind

[306] This, not memory vs. planning, is the topic of debate between Gupta et al 2010; Dragoi & Tonegawa 2011; Ólafsdóttir et al 2015, and Silva et al 2015, about whether SWRs can sometimes be 'preplay' events or if they must always involve 'replay'.

matters. But finding ways of doing so will be crucial to any use of SWRs as direct evidence of episodic memory.

Panoz-Brown et al. 2018 do discuss evidence that rat hippocampus is used in what they call an 'episodic memory task', which may be thought to suggest that hippocampal activity in the rat *is* connected to particular past events.[307] This task involves presenting subjects with a list of 5-12 trial-unique odours in a distinctive encoding context. Subjects are then (either immediately or one hour later) given forced choice between different odours from the list and have to pick the item that appeared at a certain ordinal position on their list (either 2nd-last or 4th-last item, depending on the rat). In some versions of this task they are given a different odour-based task during the interval between exposure to the stimuli and test, and different strengths of odours are used to ensure that ordinal position rather than dodour decay is used to distinguish odours from different positions in the list. Panoz-Brown et al found that using designer drugs to suppress the hippocampus impaired performance on this task specifically, and not on tasks of odour discrimination, judging which odours were familiar etc. They give five explicit reasons to think this is 'replayed' episodic memory. First, it is hippocampal dependent—but we have already criticized the probative value of this consideration above. Second, these involve trial unique lists. The problem here is again the ambiguity in 'unique': we need not interpret the rats as forming a representation of a particular, unrepeatable event, just a specific, unrepeated one. Their other considerations are also not unique to episodic relative to various forms of semantic and generic memory: the task requires representation of ordinal position within the list rather than familiarity, a 'long retention interval' (although at one hour,

[307] They also describe it as involving 'replay' but do not tie this explicitly to SWRs.

this could be disputed: it is not long enough to be consolidated in sleep, for example), and the memory is resistant to interference from memory of other odours.

To sum up the value of SWRs as evidence for episodic memory in animals, then, they could form a valuable part of an argument for imagistic representation in episodic memory, but more evidence would be needed to have any confidence that they are used in imagistic representation of particular past events, in a way which is available to other computations such as learning algorithms in the way episodic memory needs to be to have a distinctive role.

There is a different way of interpreting the evidence this provides, however, which is relevant to the earlier parts of this dissertation: SWRs do seem to be strong evidence for representation of temporally ordered sequences.[308] They do seem to allow for at least some flexible computations in virtue of order (including running orders in reverse), and are available to a number of different processes. Furthermore, they could count as *holistic* representations of sequences, rather than just pair-by-pair representations, provided they can be used to get at relations between different sub-events: this is one important upshot of §4.2's discussion of the difference between requiring *simultaneous* representation of a structure and *holistic* representation of that structure, as although the different sub-events of a sequence will not be represented simultaneously by SWRs, this is less important for computational purposes than their all being represented within one unified, structured representation. As we have seen, holistic representation of sequences could be important in its own right and irrespective of particularity or episodic memory, especially if supplemented by climactic causal structure and other features

[308] Cf. Shea (2018: 115, 136), although Shea's discussion of SWRs is mainly focused on showing that place cells represent location in a distinctive way, given his view of representation.

typical of narratives. We do not have evidence that SWRs have these sorts of features, but again it would be an interesting avenue for further exploration.

7.6 Other Correlates

Many other lines of evidence for episodic memory in non-human animals have been proposed. Most provide rather good evidence for properties which are associated with episodic memory in humans, but either are also possessed by many other forms of mental state even in humans, or could well come apart from episodic memory in other species. I will provide three examples to illustrate the general weaknesses that tend to recur in this sort of evidence.

For example, Eichenbaum et al 2010 emphasize certain properties of animals' Receiver Operating Characteristic curves (a somewhat technical notion from Signal Detection Theory), which do suggest that rats' hippocampal memory for recently present items goes beyond those items merely being familiar, and involves some kind of representation of those items, but does not narrow down whether these representations are imagistic, or whether they represent particular occurrences of these items' being presented, or just that they were presented at least once in the recent past.

Meanwhile, a range of work has looked for future-directed mental time travel in animals. This is relevant to episodic memory in a few ways. First, it suggests representations of time—or at least a degree of decoupling from the present—that *could* be used to ensure representation of particular past events (although in this regard it is even less direct evidence for episodic memory than www-memory). Second, insofar as it establishes future-directed *imagistic* thought, it shows that animals can use such representations in at least one context. Third, the ability to simulate future scenarios is associated with episodic memory in humans, so this plays at least the (rather weak) role of a feature known to correlate with episodic memory in one

species. Scrub-jays' forward-looking behaviour (what, where, how much they cache and re-cache) has also been shown to depend on interesting variables, in particular predicted future preferences (*what* food they will want when they come to recover caches, even if that is different to the food they most prefer to eat now),[309] which *locations* they will be able to access,[310] and who observes them during caching.[311]

But as with WWW-memory research, there has been little focus on how far these representations are imagistic. Researchers have generally focused not on the imagistic representation of the future, but on animals' ability to take decisions now based on their future preferences, even when those will conflict with their current preferences,[312] although there has also been some work looking directly for a relationship between abilities in memory tasks and in future planning tasks.[313] This focus has been because Tulving 2004 and Suddendorf and Corballis 1997 proposed the 'Bischof-Köhler hypothesis' according to which animals can not decouple future planning from present preferences. This sort of idea looks at least somewhat motivated on the view of episodic memory and future-directed mental time travel as involving metacognition, but seems less relevant in light of the view of episodic memory defended in Chapters 5-6.

More evidence for simulation-based, or at least reconstructive, memory, might be had from evidence that animals can be induced to make mistakes in reconstructing past

[309] Clayton et al 2005; Raby et al 2007; Correia et al 2007; Clayton et al 2008.

[310] Raby et al 2007; de Kort et al 2007.

[311] Clayton et al 2006; Dally et al 2006.

[312] Mulcahy & Call 2006; Correia et al 2007; Bräuer & Call 2015.

[313] Pastalkova et al 2008.

events. Millin & Riccio 2019 survey evidence that is suggestive of this idea. However, the paradigms they describe involve much the same weaknesses as other paradigms discussed in this chapter when it comes to construing them as more than suggestive, indirect evidence for episodic memory: it is not clear that the memories involved should be construed as memories of particular events (after all, humans can be induced to make all kinds of errors in generic imagistic memory and semantic memory), and it is debatable exactly how the manipulated reconstruction works and whether it involves imagery.

7.7 Conclusion

Do rats, scrub jays, or other non-human animals have episodic memory? The lines of evidence surveyed in this chapter certainly are evidence favouring the view that they do. WWW-memory would be one way of remembering particular past events (at least where the 'when' and 'where' components are not simply any spatiotemporal information such as time of day, but kinds of information specific enough to guarantee a particular event). Evidence for incidental encoding evidence for memory distortions and confabulation could be evidence for the involvement of simulation to reconstruct rich details, as could the use of SWRs at decision-points. But we could gather much stronger evidence through carefully modifying and combining these existing paradigms, to take into account the core features of episodic memory in chapter 5: by trying to find evidence that reconstruction is based at least in part on imagistic processes operating on memory traces, and by looking for clearer evidence that these processes are reconstructing particular (as opposed to just specific) events. This all suggests that there is something right about the hedging and talk of 'episodic-like' memory in these contexts: the experimental paradigms used so far are compatible with other, non-episodic forms of memory. But the issue is not consciousness or autonoesis, and contra many who say that it is, the missing

features are very much testable.[314] Rather, the issue is connecting the tests we have to the core features of episodic memory, and to correlates which are likely to hold across species.

[314] Indeed, the notion that consciousness is untestable in animals is debatable. Firstly, if one really thinks episodic memory has to be conscious, then one need not treat lack of direct evidence for consciousness as showing the deficiency of paradigms demonstrating what otherwise looks very much like memory in animals. Instead, one can treat such paradigms as providing evidence for episodic memory, and therefore also for consciousness (both Gennaro 2009: 187ff. and Andrews 2015: 75ff. take this line). Furthermore, it is a substantive position that needs to be argued for that consciousness is untestable except via linguistic introspective report, and there may turn out to be many ways of testing directly for consciousness or at least providing evidence for it in non-linguistic creatures (see e.g. Barron & Klein 2016; Tye 2016; Allen-Hermanson 2017; Prinz 2017; Godfrey-Smith 2019 for discussion).

Chapter 8: Conclusions

What, then, are the main lessons of this dissertation?

Thinking in terms of cognitive significance is fruitful for understanding both animal and human minds, and for generating new research. It can reframe unhelpful questions about the difference between humans and animals. It can get at the most interesting and tractable parts of many questions framed in terms of evolutionary function whilst avoiding their epistemic problems. It can help us in answering constitutive questions about which features are core to multiply realizable mental traits.

The claims surveyed in the introduction to the effect that animals are stuck in time were almost certainly wrong. Many animals can plan and have attitudes towards events at least some distance into the future, despite not having explicit temporal representation. Relatedly, we need to make many distinctions between varieties of temporal sophistication which are typically glossed over by the idea that animals are stuck in time and its different expressions. Different kinds of temporal representation are often independent of both one another and episodic memory: each of these traits' significance does not depend on the other traits, and neither do these traits strongly constitutively depend on one another. However, there was an important grain of truth in the idea that a *Big Difference Between Humans and Other Animals* relates to time: episodic memory could be hugely cognitively significant, and narrative representation and representation of essentially dynamical entities could well be too.

All this raises many questions, and suggests lines for further combined philosophical, computational, behavioural, and neural research. I have pointed to many places where speculative ideas about significance could be tested and refined through running more detailed simulations, especially in the cases of episodic memory and narrative understanding. And I have argued that much of the behavioural and neural evidence, whilst important, interesting and of

some value to the questions of which animals have episodic memory, temporal representation and the rest, could be developed in ways which are much more directly tied to the essential cores of the capacities in question. But there are also questions raised which go beyond those raised so far. For example, how much of what I have said transfers to other abstract and theoretically important domains, such as space, number, probability, and modality? How special is time? Many of the issues discussed in this dissertation do have time-unique features, such as the issue of non-Markovian dynamics causing complexity that cannot be dealt with in a tractable, feasible way. But issues of capacities based on representation as opposed to carefully designed architectures, holistic structured representations, and abstract, non-sensory features of the world, arise in many of these other cases.

One theme of this dissertation has been the importance of huge variety and complexity in the world, and more and less fruitful, flexible and tractable ways of coming to some degree of understanding of that complexity (whether that involves learning about the complexity with a general-purpose learning algorithm, simplifying it as with representations of essentially dynamical entities, or simplifying it by only focusing on special cases, as with narrative understanding). The minds of different species constitute one domain where humans struggle to cope with the complexity involved. I hope to have shed light on some of that complexity, by pointing to many varieties of possible minds, and by suggesting ways of fruitfully organizing and simplifying our approach to that complexity, by focusing on questions of cognitive significance.

References

Allen, T.A., & Fortin, N.J. (2013). The Evolution of Episodic Memory. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 10379-10386.

Allen-Hermanson, S. (2005). Morgan's Canon Revisited. *Philosophy of Science, 72*, 608-631.

Allen-Hermanson, S. (2017). So That's What It's Like! In K. Andrews & J. Beck (Eds.), *The Routledge Handbook of Philosophy of Animal Minds* (pp. 157-164). New York, NY: Routledge.

Altschul, D., Jensen, G., & Terrace, H. S. (2017). Perceptual Category Learning of Photographic and Painterly Stimuli in Rhesus Macaques (*Macaca mulatta*) and Humans. *PLoS ONE.*

Alwishah, A. (2016). Avicenna on Animal Self-Awareness, Cognition and Identity. *Arabic Sciences and Philosophy, 26*, 73-96.

Andrews, K. (2009). Politics or Metaphysics? On Attributing Psychological Properties to Animals *Biology and Philosophy, 24*(1), 51-63.

Andrews, K. (2015). *The Animal Mind: An Introduction to the Philosophy of Animal Cognition.* New York, NY: Routledge.

Aronowitz, S. (2018). Retrieval is Central to the Distinctive Function of Episodic Memory. *Behavioral and Brain Sciences, 41*, e2

Atance, C.M., & O'Neill, D.K. (2001). Episodic Future Thinking. *Trends in Cognitive Science, 12*, 533–539.

Axmacher, N., Elger, C.E., and Fell, J. (2008). Ripples in the Medial Temporal Lobe are Relevant for Human Memory Consolidation. *Brain, 131*, 1806–1817.

Barlow, H. (1961). Possible Principles Underlying the Transformation of Sensory Messages. In W. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barron, A. & Klein, C. (2016). What Insects Can Tell Us About the Origins of Consciousness. *Proceedings of the National Academy of Sciences, 113*(18), 4900-4908.

Barry, D.N., & Maguire, E.A. (2019). Remote Memory and the Hippocampus: A Constructive Critique. *Trends in Cognitive Sciences, 23*(2), 128-142.

Bartlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge: Cambridge University Press.

Bauer, P.J. & Leventon, J.S. (2013). Memory for One-Time Experiences in the Second Year of Life: Implications for the Status of Episodic Memory. *Infancy, 18*(5), 755-781.

Bechtel, W., & Richardson, R.C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (2nd. Ed.). Cambridge, MA: MIT Press.

Beck, J. (2012). The Generality Constraint and the Structure of Thought. *Mind, 121*(483), 563-600.

Beck, J. (2013). Why We Can't Say What Animals Think. *Philosophical Psychology, 26*(4), 520-546.

Beck, J. (2018). Marking the Perception–Cognition Boundary: The Criterion of Stimulus-Dependence. *Australasian Journal of Philosophy, 96*(2), 319-334.

Bennett, C. & Lindaker Lindsey S. (1992). Some Notes on the Physiological and Behavioral Ontogeny of Okapi (*Okapia johnstoni*) Calves. *Zoo Biology, 11*, 433-442.

Bergson, H. (1991). *Matter and Memory* (N.M. Paul & W. Scott Palmer, Trans.). New York, NY: Zone Books

Bermúdez, J. L. (2007). What is at Stake in the Debate on Nonconceptual Content? *Philosophical Perspectives, 21*(1), 55-72.

Birch, J. (2017). Animal Sentience and the Precautionary Principle. *Animal Sentience*, 2:16(1).

Bischof-Köhler, D. (2000). *Kinder auf Zeitreise: Theory of Mind, Zeitverständnis und Handlunsorganisation.* Bern: Huber.

Block, N. (1983). Mental Pictures and Cognitive Science. *Philosophical Review*, *92*(4), 499--542.

Block, N. (2014). Seeing-As in the Light of Vision Science. *Philosophy and Phenomenological Research, 89*(1), 560-572.

Boerner, L. & Severgnini, B. (2019). Time for Growth. *King's Business School Working Paper, No. 2019/4*.

Bornstein, A. M. & Pickard, H. (2020). 'Chasing the First High': Memory Sampling in Drug Choice. *Neuropsychopharmacology*, Advance online publication.

Boyd, R. (1989). What Realism Implies and What it Does Not. *Dialectica*, *43*(1-2), 5–29.

Boyer, P. (2008). Evolutionary Economics of Mental Time Travel? *Trends in Cognitive Sciences, 12*(6), 219-224.

Boyer, P. (2009). What are Memories For? Functions of Recall in Cognition and Culture. In P. Boyer & J. Wertsch (Eds.), *Memory in Mind and Culture* (pp. 3-28). Cambridge: Cambridge University Press.

Boyle, A. (2019a). The Impure Phenomenology of Episodic Memory. *Mind and Language,* Advance online publication.

Boyle, A. (2019b). Mapping the Minds of Others. *Review of Philosophy and Psychology, 10*(4), 747-767.

Bradley, F.H. (1899). Some Remarks on Memory and Inference. *Mind, 8,* 145-166.

Brannon, E.M. & Terrace, H.S. (1998). Ordering of the Numerosities 1 to 9 by Monkeys. *Science, 282,* 746-749.

Brannon, E.M. & Terrace, H.S. (2000). Representation of the Numerosities 1-9 by Rhesus Macaques (*Macaca mulatta*). *Journal of Experimental Psychology: Animal Behavior Processes, 26*(1), 31-49.

Bräuer, J. & Call, J. (2015). Apes Produce Tools for Future Use. *American Journal of Primatology, 77*, 254-263.

Brewer, W.F. (1988). Memory for randomly sampled autobiographical events. In U. Neisser & E. Winograd (Eds.), *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*, Emory Symposia in Cognition (pp. 21–90). Cambridge: Cambridge University Press.

Brewer, W.F. (1996). What is Recollective Memory? In D. C. Rubin (Ed.), *Remembering Our Past: Studies in Autobiographical Memory* (pp. 21-90) Cambridge: Cambridge University Press.

Broad, C.D. (1925). *The Mind and Its Place in Nature.* New York, NY: Routledge & Kegan Paul.

Brown, G.D.A., Preece, T., & Hulme, C. (2000). Oscillator-Based Memory for Serial Order. *Psychological Review, 107*(1), 127-181.

Brown, R. & Kulik, J. (1977). Flashbulb Memories. *Cognition, (5)*1, 73-99.

Brown, R.L. (2019). Infer with Care: A Critique of the Argument from Animals. *Mind and Language, 34*(1), 21-36.

Buchak, L. (2014). Belief, Credence, and Norms. *Philosophical Studies*, *169*, 285-311.

Buckner, C. (2013). Morgan's Canon, Meet Hume's Dictum: Avoiding Anthropofabulation in Cross-Species Comparisons. *Biology and Philosophy, 28,* 853-871.

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

Buller, D.J. (2005). *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature.* Cambridge, MA: MIT Press.

Burge, T. (2010a). *Origins of Objectivity.* Oxford: Oxford University Press.

Burge, T. (2010b). Steps toward Origins of Propositional Thought. *Disputatio, 4*(29), 39 - 67.

Burge, T. (2011). Self and Self-Understanding. *Journal of Philosophy, 108*(6-7), 287-383.

Burge, T. (2014). Reply to Block: Adaptation and the Upper Border of Perception. *Philosophy and Phenomenological Research, 89*(3), 573-583.

Burns, R. (1786). To a Mouse, on Turning Her Up in Her Nest, with the Plough. In *Poems, Chiefly in the Scottish Dialect* (pp. 138-140). Kilmarnock: John Wilson.

Byrne, R.W., & Whiten, A. (Eds.) (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford: Clarendon Press.

Cai, D.J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Song, W., Wei, B., Veshkini, M., La-Vu, M., Lou, J., Flores, S.E., Kim, I., Sano, Y., Zhou, M., Baumgaertel, K., Lavi, A., Kamata, M., Tuszynski, M., Mayford, M., Golshani, P., Silva, A.J. (2016). A Shared Neural Ensemble Links Distinct Contextual Memories Encoded Close in Time. *Nature, 534*(7605), 115–118.

Calcott, B. & Sterelny, K. (2011). Introduction: A Dynamic View of Evolution. In B. Calcott & K. Sterelny (Eds.) *The Major Transitions in Evolution Revisited.* Cambridge, MA: MIT Press

Camp, E. (2004). The Generality Constraint and Categorial Restrictions. *Philosophical Quarterly, 54*(215), 209-231.

Camp, E. (2007). Thinking with Maps. *Philosophical Perspectives, 21*, 145-182.

Camp, E. (2009a). Putting Thoughts to Work: Concepts, Systematicity, and Stimulus-Independence. *Philosophy and Phenomenological Research, 78*(2), 275-311.

Camp, E. (2009b). A Language of Baboon Thought. In R. W. Lurz (Ed.), *The Philosophy of Animal Minds* (pp. 108-127). Cambridge: Cambridge University Press.

Campbell, J. (1994). *Past, Space, and Self.* Cambridge, MA: MIT Press.

Campbell, J. (1997). The Structure of Time in Autobiographical Memory. *European Journal of Philosophy, 5*(2), 105-118.

Campbell, J. (2006). Ordinary Thinking about Time. In M. Stöltzner & F. Stadler (Eds.), *Time and History: Proceedings of the 28. International Ludwig Wittgenstein Symposium, Kirchberg Am Wechsel, Austria 2005* (pp. 1-12). Berlin: De Gruyter.

Cantlon, J. F. & Brannon, E.M. (2006). Shared System for Ordering Small and Large Numbers in Monkeys and Humans. *Psychological Science, 17*(5), 401-406.

Carey, S. (2009). *The Origin of Concepts.* Oxford: Oxford University Press.

Carroll, N. (2001). *Beyond Aesthetics: Philosophical Essays*. Cambridge: Cambridge University Press.

Carruthers, P. (2006). *The Architecture of The Mind: Massive Modularity and the Flexibility of Thought.* Oxford: Oxford University Press.

Carruthers, P. (2009). Invertebrate Concepts Confront the Generality Constraint (and Win). In R. W. Lurz (ed.) *The Philosophy of Animal Minds (*pp. 89-107). Cambridge: Cambridge University Press.

Carruthers, P. (2015). *The Centered Mind: What the Science of Working Memory Shows Us About The Nature Of Human Thought.* Oxford: Oxford University Press.

Carruthers, P. (2018). Episodic Memory isn't Essentially Autonoetic. *Behavioral and Brain Sciences, 41*, e6.

Chen, J., Honey, C., Simony, E., Arcaro, M., Norman, K., & Hasson, U. (2016). Accessing Real-Life Episodic Information from Minutes Versus Hours Earlier Modulates Hippocampal and High-Order Cortical Dynamics. *Cerebral Cortex, 26*(8), 3428–3441.

Chen, S., Swartz, K.B. & Terrace, H.S. (1997). Knowledge of the Ordinal Position of List Items in Rhesus Monkeys. *Psychological Science, 8*(2), 80-86.

Cheney, D. and Seyfarth, R. (1990). *How Monkeys See the World.* Chicago, IL: University of Chicago Press.

Cheng, S. & Werning, M. (2016). What is Episodic Memory if it is a Natural Kind? *Synthese, 193*, 1345-1385.

Cheng, S., Werning, M., & Suddendorf, T. (2016). Dissociating Memory Traces and Scenario Construction in Mental Time Travel. *Neuroscience and Biobehavioral Reviews, 60,* 82-89.

Chirimuuta, M. (2020). Charting the Heraclitean Brain: Perspectivism and Simplification in Models of the Motor Cortex. In C.D. McCoy & M. Massimi (Eds.), *Understanding Perspectivism: Scientific Challenges and Methodological Prospects* (pp. 141-159). New York, NY: Routledge.

Chomsky, N. (1980). Human Language and Other Semiotic Systems. In T. A. Sebeok & J. Umiker-Sebeok (Eds.), *Speaking of Apes: A Critical Anthology of Two-Way Communication with Man* (pp. 429-440). New York, NY: Plenum Press.

Clark, A. (1998). Time and the Mind. *Journal of Philosophy, 95*(7), 354-376.

Clark, A., & Toribio, J. (1994). Doing Without Representing? *Synthese, 101*, 401-431.

Clark, J.A. (2010) Relations of Homology Between Higher Cognitive Emotions and Basic Emotions. *Biology and Philosophy, 25*, 75-94.

Clarke, S. (forthcoming). Beyond the Icon: Core Cognition and the Bounds of Perception. *Mind and Language*, Advance online publication.

Clatterbuck, H. (2016). Darwin, Hume, Morgan and the *Verae Causae* of Psychology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 60*, 1-14.

Clayton, N. (2015). Ways of Thinking: From Crows to Children and Back Again. *The Quarterly Journal of Experimental Psychology, 68*(2), 208-241.

Clayton, N. & Dickinson, A. (1998). Episodic-Like Memory During Cache Recovery by Scrub Jays. *Nature, 395*(6699), 272-274.

Clayton, N. & Dickinson, A. (1999). Scrub Jays (*Aphelocoma coerulescens*) Remember the Relative Time of Caching as Well as the Location and Content of their Caches. *Journal of Comparative Psychology, 113*, 403–416.

Clayton, N., Yu, K., & Dickinson, A. (2003). Interacting Cache Memories. *Journal of Experimental Psychology: Animal Behavior Processes, 29*, 14–22.

Clayton, N., Dally, J., Gilbert, J., & Dickinson, A. (2005). Food Caching by Western Scrub-Jays (*Aphelocoma californica*) Is Sensitive to the Conditions at Recovery. *Journal of Experimental Psychology: Animal Behavior Processes, 31*(2), 115-124.

Clayton, N., Emery, N., & Dickinson, A. (2006). The Rationality of Animal Memory. In S. Hurley & M. Nudds (Eds.), *Rational Animals?* Oxford: Oxford University Press.

Clayton, N. S., Correia, S. P. C., Raby, C.R., Alexis, D.M., Emery, N.J. & Dickinson, A. (2008). Response to Suddendorf & Corballis (2008): in Defence of Animal Foresight. *Animal Behaviour, 76,* e9-e11.

Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending Time in the Brain: How Event Memories are Constructed from Experience. *Hippocampus, 29*, 162–183.

Conway Morris, S. (2003). *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge: Cambridge University Press.

Corballis, M. (2013). Mental Time Travel: A Case for Evolutionary Continuity. *Trends in Cognitive Sciences, 17*(1), 5-6.

Correia, S., Dickinson, A., & Clayton, N.S. (2007). Western Scrub-Jays Anticipate Future Needs Independently of Their Current Motivational State. *Current Biology, 17*, 856-861.

Crane, T. (1992). The Nonconceptual Content of Experience. In T. Crane (ed.), *The Contents of Experience: Essays on Perception* (pp. 136-157). Cambridge: Cambridge University Press.

Csibra, G., & Gergely, G. (2011). Natural Pedagogy as Evolutionary Adaptation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 366*(1567), 1149–1157.

Crystal, J. D. (2010). Episodic-Like Memory in Animals. *Behavioural Brain Research, 215*, 235-243.

Currie, G. and Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford: Oxford University Press.

Dally, J., Emery, N., & Clayton, N. (2006). Food-Caching Western Scrub-Jays Keep Track of Who Was Watching When. *Scienc*e, *New Series, 312*(5780), 1662-1665.

D'Amato, M.R. & Colombo, M. (1988). Representation of Serial Order in Monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes, 14*(2), 131-139.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection, Or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Darwin, C. (1889). *The Descent of Man, and Selection in Relation to Sex* (Revised Ed.). New York, NY: D. Appleton & Co.

Davidson, D. (1975). Thought and Talk. In S. Guttenplan (Ed.), *Mind and Language* (pp. 7-23). Oxford: Oxford University Press.

Davidson, D. (1982). Rational Animals. *Dialectica*, *36*(4), 317-327.

De Brigard, F. (2014). Is Memory for Remembering? Recollection as a Form of Episodic Hypothetical Thinking. *Synthese, 191*, 155-185.

De Kort, S.R., Correia, S.P.C., Alexis, D.M., Dickinson, A., & Clayton, N.S. (2007). The Control of Food Caching by Western Scrub-Jays: Where and When to Cache. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(4), 361-370.

De Waal, F. (1999). Anthropomorphism and Anthropodenial: Consistency in Our Thinking about Humans and Other Animals. *Philosophical Topics, 27*(1), 255-280.

De Waal, F. (2016). *Are We Smart Enough to Know How Smart Animals Are?* New York, NY: W. W. Norton & Co.

De Waal, F., & Ferrari, P.F. (2010). Towards a Bottom-Up Perspective on Animal and Human Cognition. *Trends in Cognitive Sciences, 14*(5), 201-207.

Debus, D. (2014). "Mental time travel": Remembering the Past, Imagining the Future, and the Particularity of Events. *Review of Philosophy and Psychology, 5*(3), 333–350.

Deese, J. (1959). On the Prediction of Occurrence of Particular Verbal Intrusions in Immediate Recall. *Journal of Experimental Psychology, 58*, 17-22.

Dehaene, S., & Naccache, L. (2001). Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition, 79,* 1-37.

Dekleva, M., Dufour, V., de Vries, H., Spruijt, B.M., & Sterck, E.H.M. (2011). Chimpanzees (Pan troglodytes) Fail a What-Where-When Task but Find Rewards by Using a Location-Based Association Strategy. *PLoS ONE, 6*(2), e16593.

Dennett, D. (1976). Conditions of Personhood. In A.O. Rorty (Ed.), *The Identities of Persons.* Berkeley, CA: University of California Press.

Dennett, D. (1996). *Kinds of Minds.* New York, NY: Basic Books.

Descartes, R. (1646). Letter to the Marquess of Newcastle, 23/11/1646 (AT iv.569; Am vii. 222). In R. Descartes (1970), *Philosophical Letters* (A. Kenny, Trans.). Minneapolis: University of Minnesota Press (pp. 204-208).

Descartes, R. (1911). *The Philosophical Works of Descartes, Vol. 1* (E. Haldane & G.R.T. Ross, Trans.). Cambridge: Cambridge University Press.

Diba, K., and Buzsaki, G. (2007). Forward and Reverse Hippocampal Place-Cell Sequences During Ripples. *Nature Neuroscience, 10*, 1241–1242.

Dokic, J. (2001). Is Memory Purely Preservative? In C. Hoerl & T. McCormack (Eds.), *Time and Memory: Issues in Philosophy and Psychology* (pp. 213–32). Oxford: Oxford University Press.

Dolan, R.J. & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron, 80,* 312-325.

Dragoi, G., & Tonegawa, S. (2011). Preplay of Future Place Cell Sequences by Hippocampal Cellular Assemblies. *Nature, 469*, 397–401.

Drayson, Z. (2017). Modularity and the Predictive Mind. In T. Metzinger and W. Wiese (Eds.), *Philosophy and Predictive Processing*. MIND Group, Frankfurt am Main.

Dretske, F.I. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Drucker, C.B. & Brannon, E.M. (2014). Rhesus Monkeys (*Macaca mulatta*) Map Number onto Space. *Cognition, 132*, 57-67.

DuBrow, S., & Davachi, L. (2014). Temporal Memory is Shaped by Encoding Stability and Intervening Item Reactivation. *Journal of Neuroscience, 34*(42), 13998–14005.

DuBrow, S., & Davachi, L. (2016). Temporal Binding Within and Across Events. *Neurobiology of Learning and Memory, 134*, 107–114.

Dunbar, R.I.M. & Shultz, S. (2007). Evolution in the Social Brain. *Science, 317*, 1344-1347.

Easton, A., Webster, L., & Eacott, M. (2012). The Episodic Nature of Episodic-Like Memories. *Learning and Memory, 19*, 146-150.

Echeverri, S. (2016). Indexing the World? Visual Tracking, Modularity, and the Perception–Cognition Interface. *British Journal for the Philosophy of Science, 67*(1), 215-245.

Eichenbaum, H. (2017). The Role of the Hippocampus in Navigation is Memory. *Journal of Neurophysiology, 117*(4), 1785–1796.

Eichenbaum, H., Fortin, N., Sauvage, M., Robitsek, R.J., & Farovik, A. (2010). An Animal Model of Amnesia That Uses Receiver Operating Characteristics (ROC) Analysis to Distinguish Recollection from Familiarity Deficits in Recognition Memory. *Neuropsychologia, 48*, 2281-2289.

Emery, N. & Clayton, N. (2001). Effects of Experience and Social Context on Prospective Caching Strategies in Scrub Jays. *Nature, 414*, 443–446.

Euston, D.R., Tatsuno, M., & McNaughton, B.L. (2007) Fast-Forward Playback of Recent Memory Sequences in Prefrontal Cortex During Sleep. *Science, 318*, 1147–1150 80.

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

Ezzyat, Y., & Davachi, L. (2014). Similarity Breeds Proximity: Pattern Similarity Within and Across Contexts is Related to Later Mnemonic Judgments of Temporal Proximity. *Neuron, 81*(5), 1179–1189.

Feeney, M.C., Roberts, W.A., & Sherry, D.F. (2009). Memory for What, Where, and When in the Black-Capped Chickadee (*Poecile atricapillus*). *Animal Cognition, 12,* 767-777.

Feng, T., Silva, D., Foster, D.J. (2015). Dissociation Between the Experience-Dependent Development of Hippocampal Theta Sequences and Single-Trial Phase Precession. *Journal of Neuroscience*, *35*(12), 4890-4902.

Ferkin, M.H., Combs, A., delBarco-Trillo, J., Pierce, A.A., & Franklin, S. (2008). Meadow Voles, *Microtus pennsylvanicus*, have the Capacity to Recall the 'What', 'Where', and 'When' of a Single Past Event. *Animal Cognition, 11,* 147-159.

Fernández, J. (2019). *Memory: A Self-Referential Account.* Oxford: Oxford University Press.

Firestone, C. & Scholl, B.J. (2016). Cognition Does Not Affect Perception: Evaluating the Evidence for "Top-Down" Effects. *Behavioral and Brain Sciences, 39*, 1-72.

Fitzpatrick, S. (2008). Doing Away with Morgan's Canon. *Mind & Language, 23*(2), 224-246.

Fodor, J.A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. (2007). The Revenge of the Given. In B.P. McLaughlin & J.D. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind* (pp. 105-116). Hoboken, NJ: Wiley-Blackwell.

Foster, D.J., and Wilson, M.A. (2006). Reverse Replay of Behavioural Sequences in Hippocampal Place Cells During the Awake State. *Nature, 440*, 680–683.

Frankfurt, H.G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy, 68*(1), 5-20.

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass, 5*(10), 914-926.

Friedman, M. (1956). *A Theory of the Consumption Function*. Princeton, NJ: Princeton University Press.

Friedman, W. (1993). Memory for the Time of Past Events. *Psychological Bulletin 113*(1), 44-66.

Friedman, W. (2005). Developmental and Cognitive Perspectives on Humans' Sense of the Times of Past and Future Events. *Learning and Motivation, 36,* 145-158.

Fugazza, C., Pogány, Á., & Miklósi, Á. (2016). Recall of Others' Actions after Incidental Encoding Reveals Episodic-like Memory in Dogs. *Current Biology, 26*(23), 3209-3213.

Gaesser, B. (2013). Constructing Memory, Imagination, and Empathy: A Cognitive Neuroscience Perspective. *Frontiers in Psychology, 3*(576), 1-6.

Gallie, W.B. (1964). *Philosophy and the Historical Understanding*. New York, NY: Schocken Books.

Gallistel, CsR. (1990). *The Organization of Learning.* Cambridge, MA: MIT Press.

Gallistel, C.R. & King, A.P. (2010). *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience.* Hoboken, NJ: Wiley-Blackwell.

Garamszegi, L.Z., & Eens, M. (2004). The Evolution of Hippocampus Volume and Brain Size in Relation to Food Hoarding in Birds. *Ecology Letters, 7,* 1216–1224.

Gauker, C. (2011). *Words and Images: An Essay on the Origin of Ideas*. Oxford: Oxford University Press.

Gauker, C. (2017a). Three Kinds of Nonconceptual Seeing-as. *Review of Philosophy and Psychology, 8*, 763-779.

Gauker, C. (2017b). Visual Imagery in the Thought of Monkeys and Apes. In K. Andrews & J. Beck (Eds.), *The Routledge Handbook of Philosophy of Animal Minds* (pp. 25-33). New York, NY: Routledge.

Gendler, T.S. (2008). Alief and Belief. *Journal of Philosophy, 105*(10), 634-663.

Gennaro, R.J. (2009). Animals, Consciousness, and I-Thoughts. In R. Lurz (ed.) *The Philosophy of Animal Minds* (pp. 184-200). Cambridge: Cambridge University Press.

Georgopoulos, A.P., Lurito, J.T., Petrides, M., Schwartz, A.B., & Massey, J.T. (1989). Mental Rotation of the Neural Population Vector. *Science,* 234–36.

Gershman, S.J. & Daw, N.D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology, 68*, 101-128.

Giardino, V. & Greenberg, G. (2015). Introduction: Varieties of Iconicity. *Review of Philosophy and Psychology, 6*(1), 1-25.

Gill, F.B. (1988). Trapline Foraging by Hermit Hummingbirds: Competition for an Undefended, Renewable Resource. *Ecology, 69*(6), 1933-1942.

Girardeau, G., Benchenane, K., Wiener, S.I., Buzsaki, G., and Zugaro, M.B. (2009). Selective Suppression of Hippocampal Ripples Impairs Spatial Memory. *Nature Neuroscience, 12*, 1222–1223.

Godfrey-Smith, P. (2009). *Darwinian Populations and Natural Selection.* Oxford: Oxford University Press.

Godfrey-Smith, P. (2016). *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness.* New York, NY: Farrar, Straus & Giroux.

Godfrey-Smith, P. (2018). Towers and Trees in Cognitive Evolution. In B. Huebner (Ed.) *The Philosophy of Daniel Dennett* (pp. 225-253). Oxford: Oxford University Press.

Godfrey-Smith, P. (2019). Evolving Across the Explanatory Gap. *Philosophy, Theory, and Practice in Biology, 11*.

Goldman, A.I. (2006). *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading.* Oxford: Oxford University Press.

Goodman, N. (1968). *Languages of Art*. Indianapolis, IN: Bobbs-Merrill.

Gottfried, J.A., Smith, A.P.R., Rugg, M.D., & Dolan, R. (2004). Remembrance of Odors Past: Human Olfactory Cortex in Cross-Modal Recognition Memory. *Neuron, 42*(4), 687-695.

Gould, J. (1986). The Locale Map of Bees: do Insects have Cognitive Maps? *Science, 232*, 861-863.

Gould, S.J. (1989). *Wonderful Life: The Burgess Shale and the Nature of History*. New York, NY: W.W. Norton.

Gould, S.J. & Lewontin, R.C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 205*(116), 581-598.

Greenberg, D.L., & Rubin, D.C. (2003). The Neuropsychology of Autobiographical Memory. *Cortex, 39,* 687-728.

Griffiths, P.E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago, IL: University of Chicago Press.

Gross, C.G. (1998). Galen and the Squealing Pig. *The Neuroscientist, 4*, 216-221

Gross, S. (2017) Perception and the Origins of Temporal Representation. *Pacific Philosophical Quarterly, 98*(S1) 275-292.

Güntürkün, O., & Bugnyar, T. (2016). Cognition without Cortex. *Trends in Cognitive Sciences, 20*(4), 291-303.

Gupta, A.S., van der Meer, M.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal Replay is Not a Simple Function of Experience. *Neuron, 65*, 695–705.

Guyer, P. (2007). *Kant*. New York, NY: Routledge.

Haith, M. (1998). Who Put the Cog in Infant Cognition? Is Rich Interpretation Too Costly? *Infant Behavior & Development, (21)*2, 167-179.

Hampton, R.R., Hampstead, B.M., & Murray, E.A. (2005). Rhesus Monkeys (*Macaca mulatta*) Demonstrate Robust Memory for What and Where, but not When, in an Open-Field Test of Memory. *Learning and Motivation, 36,* 245-259.

Hassabis, D., Kumaran, D., & Maguire, E.A. (2007). Using Imagination to Understand the Neural Basis of Episodic Memory. *Journal of Neuroscience, 27*(52), 14365-14374.

Haugeland, J. (1981). Analog and analog. *Philosophical Topics, 12*(1), 213-226.

Hauser, M.D., Chomsky, N., & Fitch, W.T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science, 298*(5598), 1569-1579.

Hedges, S.B. (2002). The Origin and Evolution of Model Organisms. *Nature Reviews Genetics, 3*, 838–849.

Heck, R. (2007). Are There Different Kinds of Content? in McLaughlin, B. and Cohen, J. (Eds.) *Contemporary Debates in Philosophy of Mind* (pp. 117-138). Hoboken, NJ: Wiley-Blackwell.

Heider, F. & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology, 57*(2), 243-259.

Henke, K. (2010). A Model for Memory Systems Based on Processing Modes Rather than Consciousness. *Nature Reviews Neuroscience, 11*(7), 523-532.

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Harvard University Press.

Hobbes, T. (1991). *De Homine* (B. Gert, Trans). In *Man and Citizen* (pp. 37-43). Indianapolis, IN: Hackett.

Hoerl, C. (2001). The Phenomenology of Episodic Recall. In C. Hoerl & T. McCormack (Eds.), *Time and Memory: Issues in Philosophy and Psychology* (pp. 315–336). Oxford: Oxford University Press.

Hoerl, C. (2008). On Being Stuck in Time. *Phenomenology and the Cognitive Sciences,* 7, 485-500.

Hoerl, C., & McCormack, T. (2011). Time in Cognitive Development. in C. Callender (Ed.). *The Oxford Handbook of Philosophy of Time* (pp. 439-459). Oxford: Oxford University Press.

Hoerl, C., & McCormack, T. (2018). Animal Minds in Time: The Question of Episodic Memory. In K. Andrews & J. Beck (Eds.), *The Routledge Handbook of Philosophy of Animal Minds* (pp. 56-64). New York, NY: Routledge.

Hoerl, C. & McCormack, T. (2019). Thinking in and about Time: A Dual Systems Perspective on Temporal Cognition. *Behavioral and Brain Sciences, 42*(e244), 1-77.

Hoffman, M., Beran, M., & Washburn, D. (2009). Memory for 'What', 'Where', and 'When' Information in Rhesus Monkeys (*Macaca mulatta*). *Journal of Experimental Psychology: Animal Behavior Processes, 35*(2), 143-152.

Hollard, V.D., & Delius, J.D. (1982). Rotational Invariance in Visual Pattern Recognition by Pigeons and Humans. *Science, 218*(4574), 804-806.

Hopkins, R. (1998). *Picture, Image and Experience: A Philosophical Inquiry*. Cambridge University Press.

Hopkins, R. (2018). Imagining the Past: On the Nature of Episodic Memory. In F. Macpherson & F. Dorsch (Eds.), *Perceptual Imagination and Perceptual Memory.* Oxford: Oxford University Press.

Horowitz, A. (2009). Disambiguating the 'Guilty Look': Salient Prompts to a Familiar Dog Behaviour. *Behavioural Processes, 81*, 447-452.

Horowitz, A.C., & Bekoff, M. (2007). Naturalizing Anthropomorphism: Behavioral Prompts to our Humanizing of Animals. *Anthrozoös*, *20*(1), 23-35.

Horowitz, A. (2016). *Being a Dog: Following the Dog into a World of Smell*. New York, NY: Scribner.

Hume, D. (1998). *An Enquiry Concerning Human Understanding: A Critical Edition*. Oxford: Oxford University Press.

Jacobs, L. (2012). From Chemotaxis to the Cognitive Map: The Function of Olfaction. *Proceedings of the National Academy of Sciences 109, supp. 1,* 10693-10700.

Jadhav, S.P., Kemere, C., German, P.W., and Frank, L.M. (2012). Awake Hippocampal Sharp-Wave Ripples Support Spatial Memory. *Science, 336*, 1454–1458.

Jadhav, S.P., Rothschild, G., Roumis, D.K., and Frank, L.M. (2016). Coordinated Excitation and Inhibition of Prefrontal Ensembles During Awake Hippocampal Sharp-Wave Ripple Events. *Neuron, 90*, 113–127.

James, W. (1890). *The Principles of Psychology.* New York, NY: Henry Holt & Co.

James, W. (1896). The Will to Believe. In S.M. Cahn (Ed.), *The Will to Believe: And Other Essays in Popular Philosophy* (pp. 1-15). New York: Longmans, Green, & Co.

Jensen, G., Altschul, D., Danly, E., & Terrace, H. (2013). Transfer of a Serial Representation between Two Distinct Tasks by Rhesus Macaques. *PLoS ONE, 8*(7).

Ji, D., and Wilson, M.A. (2007). Coordinated Memory Replay in the Visual Cortex and Hippocampus During Sleep. *Nature Neuroscience, 10*, 100–107.

Johnson, S.H. (2000) Thinking Ahead: The Case for Motor Imagery in Prospective Judgements of Prehension. *Cognition, 74*, 33-70.

Jozet-Alves, C., Bertin, M., & Clayton, N. (2013). Evidence of Episodic-Like Memory in Cuttlefish. *Current Biology, 23*(23), R1033-R1035.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kail, P.J.E. (2007). *Projection and Realism in Hume's Philosophy.* Oxford: Oxford University Press.

Kaminski, J., Fischer, J., & Call, J. (2008). Prospective Object Search in Dogs: Mixed Evidence for Knowledge of *What* and *Where*. *Animal Cognition, 11,* 367-371.

Kaplan, D.M. (2018). A Bridge Too Far? Inference and Extrapolation from Model Organisms in Neuroscience. In K. Andrews & Beck, J. (Eds.), *The Routledge Handbook of Philosophy of Animal Minds* (pp. 448-457). New York, NY: Routledge.

Karlsson, M.P., & Frank, L.M. (2009). Awake Replay of Remote Experiences in the Hippocampus. *Nature Neuroscience, 12*, 913–918.

Kay, K., Chung, J.E., Sosa, M., Schor, J.S., Karlsson, M.P., Larkin, M.C., Liu, D.F., Frank, L.M. (2019). Constant Sub-Second Cycling Between Representations of Possible Futures in the Hippocampus. *bioRxiv* 528976

Keeley, B. (2004). Anthropomorphism, Primatomorphism, Mammalomorphism: Understanding Cross-Species Comparisons. *Biology and Philosophy, 19,* 521-540.

Keven, N. (2016). Events, Narratives and Memory. *Synthese, 193*(8), 2497-2517

Keven, N. (2018). Carving Event and Episodic Memory at their Joints. *Behavioral and Brain Sciences, 41*, e19.

Kind, A. (2013). The Heterogeneity of the Imagination. *Erkenntnis, 78*, 141-159.

Kitcher, P. & Varzi, A. C. (2000). Some Pictures Are Worth $2^{\aleph_0}$ Sentences. *Philosophy, 75*(3), 377-381.

Klein, S. B. (2014). Autonoesis and Belief in a Personal Past: An Evolutionary Theory of Episodic Memory Indices. *Review of Philosophy and Psychology, 5*, 427-447.

Klein, S. B. (2015). What memory is. *WIREs Cognitive Science, 6*(1), 1-38.

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the Evolution of Memory: Multiple Systems, Multiple Functions. *Psychological Review, 109*(2), 306-329.

Köhler, W. (1925). *The Mentality of Apes* (2nd Ed.) (E. Winter, Trans.). New York, NY: Harcourt, Brace & Co., Inc.

Köhler, C., Hoffmann, K.P., Dehnhardt, G., & Mauck, B. (2005). Mental Rotation and Rotational Invariance in the Rhesus Monkey (*Macaca mulatta*). *Brain, Behavior and Evolution, 66,* 158-166.

Korsgaard, C. (2018) *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: Oxford University Press.

Kosslyn, S.M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J.P., Thompson, W.L., Ganis, G., Sukel, K.E., & Alpert, N.M. (1999) The Role of Area 17 in Visual Imagery: Convergent Evidence from PET and rTMS. *Science, New Series, 284*(5411), 167-170.

Kosslyn, S.M., Thompson, W.L., & Ganis, G. (2006). *The Case for Mental Imagery.* Oxford: Oxford University Press.

Kouwenberg, A.-L., Walsh, C.J., Morgan, B.E., & Martin, G.M. (2009). Episodic-Like Memory in Crossbred Yucatan Minipigs (*Sus scrofa*). *Applied Animal Behaviour Science, 117,* 165-172.

Kulvicki, J.V. (2013). *Images*. New York, NY: Routledge.

Langdon, J. (2015, October 1). Sir Edward Heath: One Nation Tory's Political Legacy. Retrieved from https://www.bbc.com/news/uk-politics-33958116

Lee, A.K., & Wilson, M.A. (2002). Memory of Sequential Experience in the Hippocampus During Slow Wave Sleep. *Neuron, 36*, 1183–1194.

Lewis, D. (1973). *Counterfactuals*. Hoboken, NJ: Wiley-Blackwell.

Lin, Z., Zhao, T., Yang, G., & Zhang, L. (2018). Episodic Memory Deep Q-Networks. *ArXiv, abs/1805.07603*.

Lloyd, E.A. (2005). *The Case of the Female Orgasm: Bias in the Science of Evolution*. Cambridge, MA: Harvard University Press.

Locke, J. (1689). *An Essay Concerning Human Understanding*. Oxford University Press.

Loftus, E.F. (2005). Planting Misinformation in the Human Mind: A 30-Year Investigation of the Malleability of Memory. *Learning and Memory, 12*(4), 361-366.

Logothetis, N.K., Eschenko, O., Murayama, Y., Augath, M., Steudel, T., Evrard, H.C., Besserve, M., & Oeltermann, A. (2012). Hippocampal-cortical Interaction During Periods of Subcortical Silence. *Nature, 491*, 547–553.

Lopes, D. (1996). *Understanding Pictures*. Oxford: Oxford University Press.

Lositsky, O., Chen, J., Toker, D., Honey, C.J., Shvartsman, M., Poppenk, J.L. Hasson, U., & Norman, K. A. (2016). Neural Pattern Change During Encoding of a Narrative Predicts Retrospective Duration Estimates. *eLife, 5*, e16070.

Losos, J.B. (2017). *Improbable Destinies: Fate, Chance, and the Future of Evolution.* New York, NY: Riverhead Books.

Lucas, J.R., Brodin, A., de Kort, S.R., & Clayton, N.S. (2004). Does Hippocampal Size Correlate with the Degree of Caching Specialization? *Proceedings: Biological Sciences, 271*(1556), 2423–2429.

Ludvig, E.A., Bellemare, M.G., & Pearson, K.G. (2011). A Primer on Reinforcement Learning in the Brain: Psychological, Computational, and Neural Perspectives. In E. Alonso & E. Mondragón (Eds.), *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications.* Hershey, PA: Medical Information Science Reference.

Lupyan, G. & Clark, A. (2015). Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Current Directions in Psychological Science, 24*(4), 279-284.

Lyon, P. (2015). The Cognitive Cell: Bacterial Behavior Reconsidered. *Frontiers in Microbiology, 6*, Art. 264.

Machamer, P.K., Darden, L., & Craver, C.F. (2000). Thinking About Mechanisms. *Philosophy of Science*, *67*(1), 1-25.

MacKinnon, R., Cohen, S.L., Kuo, A., Lee, A., & Chait, B.T. (1998). Structural Conservation in Prokaryotic and Eukaryotic Potassium Channels. *Science, 280*(5360), 106–109.

Mahr, J.B. & Csibra, G. (2018). Why do We Remember? The Communicative Function of Episodic Memory. *Behavioral and Brain Sciences, 41*, e1.

Maingret, N., Girardeau, G., Todorova, R., Goutierre, M., & Zugaro, M. (2016). Hippocampo-Cortical Coupling Mediates Memory Consolidation During Sleep. *Nature Neuroscience, 19*(7), 959–964.

Manger, P., Cort, J., Ebrahim, N., Goodman, A., Henning, J., Karolia, M., Rodrigues, S.-L., & Štrkalj, G. (2008). Is 21st Century Neuroscience Too Focussed on the Rat/Mouse Model of Brain Function and Dysfunction? *Frontiers in Neuroanatomy, 2*, Art. 5.

Mar, R. A. & Spreng, R.N. (2018). Episodic Memory Solves Both Social and Nonsocial Problems, and Evolved to Fulfill Many Different Functions. *Behavioral and Brain Sciences, 41*, e20.

Markowitsch, H. J., & Staniloiu, A. (2011). Memory, Autonoetic Consciousness, and the Self. *Consciousness and Cognition, 20*(1), 16-39.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.

Martin, C.B., & Deutscher, M. (1966). Remembering. *Philosophical Review, 75*(2), 161-96.

Martin, M. (1992). Sight and Touch. In T. Crane (Ed.), *The Contents of Experience.* Cambridge: Cambridge University Press.

Martin-Ordas, G., Haun, D., Colmenares, F., & Call, J. (2010). Keeping Track of Time: Evidence for Episodic-Like Memory in Great Apes. *Animal Cognition, 13*, 331-340.

Mauck, B. & Denhardt, G. (1997). Mental Rotation in a California Sea Lion (*Zalophus Californianus*). *The Journal of Experimental Biology, 200,* 1309-1316.

Matsuzawa, T. & Kawai, N. (2000). Numerical Memory Span in a Chimpanzee. *Nature, 403(*6765), 39–40

Maynard Smith, J., & Szathmáry, E. (1995). *The Major Transitions in Evolution.* Oxford: Oxford University Press.

McClelland, J.L., Botvinick, M.M., Noelle, D.C., Plaut, D.C., Rogers, T.T., Seidenberg, M.S., & Smith, L.B. (2010). Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Cognition. *Trends in Cognitive Sciences, 14*(8), 348-356.

McCormack T. (2001). Attributing Episodic Memory to Animals and Children. In C. Hoerl & T. McCormack (Eds.), *Time and Memory: Issues in Philosophy and Psychology* (pp. 285–313). Oxford: Oxford University Press.

McCormack, T. & Hoerl, C. (1999). Memory and Temporal Perspective: The Role of Temporal Frameworks in Memory Development. *Developmental Review, 19*, 154–182.

McCormack, T. & Hoerl, C. (2017). The Development of Temporal Concepts: Learning to Locate Events in Time. *Timing and Time Perception, 5*(3-4), 297-327.

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

Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., Bundrock, G., Hülse, S., Plümpe, T., Schaupp, F., Schüttler, E., Stach, S., Stindt, J., Stollhoff, N., Watzl, S. (2005). Honey Bees Navigate According to a Map-like Spatial Memory. *Proceedings of the National Academy of Sciences, 102*(8), 3040-3045.

Mercado III., E., Murray, S.O., Uyeyama, R.K., Pack, A.A., and Herman, L.M. (1998). Memory for Recent Actions in the Bottlenosed Dolphin (Tursiops truncatus): Repetition of Arbitrary Behaviors Using an Abstract Rule. *Animal Learning and Behavior, 26*, 210-218.

Merino-Rajme, C. (2014). A Quantum Theory of Felt Duration. *Analytic Philosophy, 55*(3), 239-275.

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

Michod, R.E. (1999). *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality.* Princeton, NJ: Princeton University Press.

Midgley, M. (1973). The Concept of Beastliness: Philosophy, Ethics and Animal Behaviour. *Philosophy*, *48*(184), 111-135.

Millikan, R.G. (1984). *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Millin, P.M., & Riccio, D.C. (2019). False Memory in Nonhuman Animals. *Learning and Memory, 26*, 381-386

Mink, L.O. (1968). Philosophical Analysis and Historical Understanding. *Review of Metaphysics, 21*(4), 667 - 698.

Montague, M. (forthcoming). The Sense/Cognition Distinction. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Montemayor, C. (2013). *Minding Time: A Philosophical and Theoretical Approach to the Psychology of Time.* Leiden: Brill.

Moore, G.E. (1925). In Defence of Common Sense. In J. H. Muirhead (Ed.), *Contemporary British Philosophy.* London: George Allen & Unwin Ltd.

Morgan, C.L. (1904). *An Introduction to Comparative Psychology* (2nd Ed.). London: The Walter Scott Publishing Co.

Morgan, C.L. (1930). Autobiography of C. Lloyd Morgan. In C. Murchison (Ed.), *History of Psychology in Autobiography Vol. 2* (pp. 237-264). Worcester, MA: Clark University Press.

Morgan, D. (2019). *Temporal* Indexicals are Essential. *Analysis, 79*, 452–61.

Moser, E., Roudi, Y., Witter, M.P., Kentros, C., Bonhoeffer, T., & Moser, M.-B. (2014). Grid Cells and Cortical Representation. *Nature Reviews Neuroscience, 15,* 466-481.

Mulcahy, N. & Call, J. (2006). Apes Save Tools for Future Use. *Science, 312*(5776), 1038-1040.

Nagy, D.G., & Orbán, G. (2016). Episodic Memory as a Prerequisite for Online Updates of Model Structure. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.). *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2699–704). Available at: https://arxiv.org/pdf/1806.07990.pdf

Neander, K. (2017). *A Mark of the Mental.* Cambridge, MA: MIT Press.

Neisser, U. (1981). John Dean's Memory: A Case Study. *Cognition, 9*(1), 1-22.

Neisser, U. & Harsch, N. (1992). Phantom Flashbulbs: False Recollections of Hearing News about Challenger. In E. Winograd & U. Neisser (Eds.), *Affect and Accuracy in Recall: Studies of 'Flashbulb Memories'* (pp. 9-31). Cambridge: Cambridge University Press.

Nelson, K. (1978). How Young Children Represent Knowledge of their World in and out of Language. In R.S. Siegler (Ed.), *Children's Thinking: What Develops?* (pp. 225–273). Hillsdale, NJ: Erlbaum.

Nelson, K. (1986). *Event Knowledge: Structure and Function in Development*. Hillsdale, NJ: Erlbaum.

Nelson, K. & Fivush, R. (2004). The Emergence of Autobiographical Memory: A Social Cultural Developmental Theory. *Psychological Review, 111*(2), 486–511.

Newman, E.J., & Lindsay, D.S. (2009). False Memories: What the Hell are They For? *Applied Cognitive Psychology, 23* 1105-1121.

Nichols, S., Stich, S., Leslie, A., & Klein, D. (1996). Varieties of Off-line Simulation. In P. Carruthers & P.K. Smith (Eds.), *Theories of Theories of Mind* (pp. 39-74). Cambridge: Cambridge University Press.

Nietzsche, F. (1997). *Untimely Meditations* (R.J. Hollingdale, Trans.). Cambridge: Cambridge University Press.

Ohshiba, N. (1997). Memorization of Serial Items by Japanese Monkeys, a Chimpanzee, and Humans. *Japanese Psychological Research, 39*(3), 236-252.

O'Keefe, J., & Speakman, A. (1987). Single Unit Activity in the Rat Hippocampus During a Spatial Memory Task. *Experimental Brain Research, 68*, 1–27.

Ólafsdóttir, H.F., Barry, C., Saleem, A.B., Hassabis, D., & Spiers, H.J. (2015). Hippocampal Place Cells Construct Reward Related Sequences through Unexplored Space. *Elife, 4*, e06063.

Ólafsdóttir, H.F., Bush, D., & Barry, C. (2018). The Role of Hippocampal Replay in Memory and Planning. *Current Biology 28*(1), R37-50.

Olkowicz, S., Kocourek, M., Lučan, R.K., Porteš, M., Fitch, W.T., Herculano-Houzel, S., & Němec, P. (2016). Birds have Primate-Like Numbers of Neurons in the Forebrain. *Proceedings of The National Academy of Sciences of the United States of America, 113*(26), 7255-60.

Olton, D. S. (1979). Mazes, Maps, and Memory. *American Psychologist, 34*(7), 583-596.

O'Neill, J., Pleydell-Bouverie, B., Dupret, D., and Csicsvari, J. (2010). Play It Again: Reactivation of Waking Experience and Memory. *Trends in Neuroscience, 33*, 220–229.

Orlov, T., Yakovlev, V., Hochstein, S., & Zohary, E. (2000). Macaque Monkeys Categorize Images by their Ordinal Number. *Nature, 404*(6773), 77-80.

Pacherie, E. & Haggard, P. (2010). What are Intentions? In W. Sinnott-Armstrong (Ed.), *Conscious Will and Responsibility. A Tribute to Benjamin Libet* (pp. 70-84), Oxford: Oxford University Press.

Panoz-Brown, D., Iyer, V., Carey, L.M., Sluka, C.M., Rajic, G., Kestenman, J., Gentry, M., Brotheridge, S., Somekh, I., Corbin, H.E., Tucker, K.G., Almeida, B., Hex, S. B., Garcia, K.D., Hohmann, A.G., & Crystal, J.D. (2018). Replay of Episodic Memories in the Rat. *Current Biology, 28*(10) 1628-1634.

Parsons, L.M. (1994) Temporal and Kinematic Properties of Motor Behavior Reflected in Mentally Simulated Action. *Journal of Experimental Psychology: Human Perception and Performance, 20*(4), 709-730.

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsáki, G. (2008). Internally Generated Cell Assembly Sequences in the Rat Hippocampus. *Science, New Series, 321*(5894), 1322-1327.

Pavličev, M. & Wagner, G. (2016). The Evolutionary Origin of Female Orgasm. *Journal of Experimental Zoology (Molecular and Developmental Evolution), 326B*, 326-337.

Peacocke, C. (1983). *Sense and Content.* Oxford: Oxford University Press.

Peacocke, C. (1986). Analogue Content. *Proceedings of the Aristotelian Society, 60*, 1-17.

Peacocke, C. (1992). *A Study of Concepts.* Cambridge, MA: MIT Press.

Peacocke, C. (2001). Does Perception Have a Nonconceptual Content? *The Journal of Philosophy, 98*(5), 239-264.

Peacocke, C. (2014). *Mirror of the World: Subjects, Consciousness, and Self-Consciousness*. Oxford: Oxford University Press.

Pearson, J., Naselaris, T., Holmes, E.A., & Kosslyn, S.M. (2015). Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends in Cognitive Sciences, 19*(10), 590-602.

Peebles, D. & Cooper, R.P. (Eds.) (2015). Thirty Years after Marr's Vision: Levels of Analysis in Cognitive Science. *Topics in Cognitive Science, 7*(2), 185-381.

Penn, D.C. (2012). How Folk Psychology Ruined Comparative Psychology: And How Scrub Jays Can Save It. In R. Menzel & J. Fischer (Eds.), *Animal Thinking: Contemporary Issues in Comparative Cognition*, (254-266). Cambridge, MA: MIT Press.

Perner, J. (2001). Episodic Memory: Essential Distinctions and Developmental Implications. In C. Moore & K. Lemmon (Eds.), *the Self in Time* (181-202). Mahwah, NJ: Lawrence Erlbaum.

Peters, M.A.K., Kentridge, R.W., Phillips, I., & Block, N. (2017). Does Unconscious Perception Really Exist? Continuing the ASSC20 Debate. *Neuroscience of Consciousness, 3*(1), 1-11.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I., & Battaglia, F.P. (2009). Replay of Rule-Learning Related Neural Patterns in the Prefrontal Cortex During Sleep. *Nature Neuroscience, 12*, 919–926.

Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal Place-Cell Sequences Depict Future Paths to Remembered Goals. *Nature, 497*, 74–79.

Phillips, B. (2019). The Shifting Border Between Perception and Cognition. *Noûs, 53*(2), 316-346.

Phillips, I. (2014). Experience of and in Time. *Philosophy Compass, 9*(2), 131-144.

Pillemer, D.B., Picariello, M.L., & Pruett, J.C. (1994). Very Long-Term Memories of a Salient Preschool Event. *Applied Cognitive Psychology, 8*(2), 95–106.

Povinelli, D.J., & Vonk, J. (2004). We Don't Need a Microscope to Explore the Chimpanzee's Mind. *Mind and Language, 19*(1), 1-28.

Prinz, J. (2017). Attention, Working Memory, and Animal Consciousness. In K. Andrews & J. Beck (Eds.), *The Routledge Handbook of Philosophy of Animal Minds* (pp. 185-195). New York, NY: Routledge.

Preuss, T.M. (2000). Taking the Measure of Diversity: Comparative Alternatives to the Model-Animal Paradigm in Cortical Neuroscience. *Brain, Behavior and Evolution, 55*(6), 287–299.

Pylyshyn, Z. (1999). Is Vision Continuous with Cognition? The Case for Cognitive Impenetrability of Visual Perception. *Behavioral and Brain Sciences, 22*(3), 341-365.

Quilty-Dunn, J. (2016). Iconicity and the Format of Perception. *Journal of Consciousness Studies, 23,*(3-4), 255-263.

Quilty-Dunn, J. (forthcoming). Perceptual Pluralism. *Noûs*. Advance online publication.

Raby, C., Alexis, D., Dickinson, A., & Clayton, N. (2007). Planning for the future by Western Scrub-Jays. *Nature, 445*, 919–921.

Raby, C.R., & Clayton, N.S. (2012). Episodic Memory and Planning. In J. Vonk & T.K. Shackelford (Eds.), *The Oxford Handbook of Comparative Evolutionary Psychology* (pp. 217-235). Oxford: Oxford University Press.

Radford, C. (1966). Knowledge — By Examples. *Analysis, 27*(1), 1-11.

Rasmussen, A.S., & Berntsen, D. (2009). The Possible Functions of Involuntary Autobiographical Memories. *Applied Cognitive Psychology, 23*(8), 1137-1152.

Ratner, H.H., Smith, B.S., & Dion, S.A. (1986). Development of Memory for Events. *Journal of Experimental Child Psychology, 41*(3), 411–428.

Rau, P. & Botterill, G. (2018). Enhanced Action Control as a Prior Function of Episodic Memory. *Behavioral and Brain Sciences, 41,* e27.

Reiner, A., Perkel, D.J., Bruce, L.L., Butler, A.B., Csillag, A., Kuenzel, W., Medina, L., Paxinos, G., Shimizu, T., Striedter, G., Wild, M., Ball, G.F., Durand, S., Güntürkün, O., Lee, D.W., Mello, C.V., Powers, A., White, S.A., Hough, G., Kubikova, L., Smulders, T.V., Wada, K., Dugas-Ford, J., Husband, S., Yamamoto, K., Yu, J., Siang, C., & Jarvis, E.D. (2004). Avian Brain Nomenclature Forum. Revised Nomenclature for Avian Telencephalon and Some Related Brainstem Nuclei. *Journal of Computational Neurology*, *473*(3), 377-414.

Rescorla, M. (2009a). Cognitive Maps and the Language of Thought. B*ritish Journal for the Philosophy of Science, 60*(2), 377-407.

Rescorla, M. (2009b). Chrysippus' Dog as a Case Study in Non-Linguistic Cognition. In R.W. Lurz (Ed.), *The Philosophy of Animal Minds* (pp. 52-71). Cambridge: Cambridge University Press.

Richards, J.M. & Gross, J.J. (2000). Emotion Regulation and Memory: The Cognitive Costs of Keeping One's Cool. *Journal of Personality and Social Psychology, 79*(3), 410-424.

Roberts, W. (2002). Are Animals Stuck in Time? *Psychological Bulletin, 128*(3), 473-489.

Robins, S. (2016). Representing the Past: Memory Traces and the Causal Theory of Memory. *Philosophical Studies, 173*, 2993-3013.

Robins, S., & Craver, C. (2009). Biological Clocks: Explaining with Models of Mechanisms. In J. Bickle (Ed.) *The Oxford Handbook of Philosophy of Neuroscience* (pp. 41-67). Oxford: Oxford University Press.

Roediger III, H.L. (1996). Memory Illusions. *Journal of Memory and Language, 35*(2), 76-100.

Romanes, G.J. (1883). *Mental Evolution in Animals*. London: Kegan, Paul, Trench, & Co.

Romanes, G.J. (1892). *Animal Intelligence*. New York, NY: D. Appleton & Co.

Rosenbaum, D.A., Gong, L., & Potts, C.A. (2014). Pre-Crastination: Hastening Subgoal Completion at the Expense of Extra Physical Effort. *Psychological Science, 25*(7), 1487-1496.

Roskies, A.L. (2008). A New Argument for Nonconceptual Content. *Philosophy and Phenomenological Research, 76*, 633–659.

Rothschild, G., Eban, E., & Frank, L.M. (2017). A Cortical-Hippocampal-Cortical Loop of Information Processing During Memory Consolidation. *Nature Neuroscience, 20*, 251–259.

Rubin, D.C., Schrauf, R.W., & Greenberg, D.L. (2003). Belief and Recollection of Autobiographical Memories. *Memory and Cognition, 31*(6), 887-901.

Rubin, D.C., & Siegler, I.C. (2004). Facets of Personality and the Phenomenology of Autobiographical Memory. *Applied Cognitive Psychology, 18,* 913-930.

Rubin, D.C. & Umanath, S. (2015). Event Memory: A Theory of Memory for Laboratory, Autobiographical, and Fictional Events. *Psychological Review, 122*(1), 1-23.

Russell, B. (1921). *The Analysis of Mind.* London: George Allen & Unwin Ltd.

Russell, J., & Hanna, R. (2012). A Minimalist Approach to the Development of Episodic Memory. *Mind and Language, 27*(1), 29-54.

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

Sahlins, M. (1976). *The Use and Abuse of Biology: An Anthropological Critique of Sociobiology*. Ann Arbor, MI: University of Michigan Press.

Schacter, D.L., Addis, D.R., & Buckner, R.L. (2007). Remembering the Past to Imagine the Future: The Prospective Brain. *Nature Reviews Neuroscience, 8*(9), 657–661.

Schacter, D.L., Guerin, S.A., & St. Jacques, P.L. (2011). Memory Distortion: An Adaptive Perspective. *Trends in Cognitive Sciences, 15*(10), 467-474.

Schacter, D.L., Carpenter, A.C., Devitt, A., Roberts, R.P., & Addis, D.R. (2018). Constructive Episodic Simulation, Flexible Recombination, and Memory Errors. *Behavioral and Brain Sciences, 41,* e32.

Schank, R.C. (1990). *Tell Me A Story: A New Look at Real and Artificial Memory*. Evanston, IL: Northwestern University Press.

Schank, R.C., & Abelson, R.P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Mahwah, NJ: Lawrence Erlbaum.

Schellenberg, S. (2013). A Trilemma About Mental Content. In J. Schear (Ed.), *Mind, Reason, and Being-in-the-World* (272-282). New York, NY: Routledge.

Schultz, W., Dayan, P., & Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. *Science, 275*(5306), 1593-1599.

Schnabel, J. (2008). Neuroscience: Standard Model. *Nature News, 454*, 682–685.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.

Shepard, R.N. & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science, New Series, 171*(3972), 701-701.

Shepard, R.N. & Cooper, L.N. (1982). *Mental Images and their Transformations.* Cambridge, MA: MIT Press.

Sherry, D.F. (2011). The Hippocampus of Food-Storing Birds. *Brain, Behavior and Evolution, 78*, 133–135.

Sherry, D.F., & Vaccarino, A.L. (1989). Hippocampus and Memory for Food Caches in Black-Capped Chickadees. *Behavioral Neuroscience, 103*, 308–318.

Sherry, D.F., Vaccarino, A.L., Buckenham, K., & Herz, R.S. (1989). The Hippocampal Complex of Food-Storing Birds. *Brain, Behavior and Evolution, 34*, 308–317.

Shettleworth, S. (2007). Planning for Breakfast. *Nature, 445*, 825-826.

Shettleworth, S.J. (2010). Clever Animals and Killjoy Explanations in Comparative Psychology. *Trends in Cognitive Sciences, 14*(11), 477-481.

Siegel, S. (2017). *The Rationality of Perception*. Oxford: Oxford University Press.

Silva, D., Feng, T., and Foster, D.J. (2015). Trajectory Events Across Hippocampal Place Cells Require Previous Experience. *Nature Neuroscience, 18*, 1772–1779.

Simon, H.A. (1996). *Sciences of the Artificial* (3rd Ed.). Cambridge, MA: MIT Press.

Singer, A.C., Carr, M.F., Karlsson, M.P., and Frank, L.M. (2013). Hippocampal Sharp Wave Ripple Activity Predicts Correct Decisions During the Initial Learning of an Alternation Task. *Neuron, 7*7, 1163–1173.

Singer, P. (2009). *Animal Liberation: The Definitive Classic of the Animal Movement*. New York, NY: HarperCollins.

Slobodchikoff, C. N., Perla, B., & Verdolin, J.L. (2009). *Prairie Dogs: Communication and Community in an Animal Society*. Cambridge, MA: Harvard University Press.

Sloman, S.A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, *119*(1), 3-22.

Slote, M. (1982). Goods and Lives. *Pacific Philosophical Quarterly, 63*(4), 311.

Slotnick, S.D., Thompson, W.L., & Kosslyn, S.M. (2005). Visual Mental Imagery Induces Retinotopically Organized Activation of Early Visual Areas. *Cerebral Cortex, 15*, 1570-1583.

Smeets, T., Merckelbach, H., Horselenberg, R., & Jelicic, M. (2005). Trying to Recollect Past Events: Confidence, Beliefs, and Memories. *Clinical Psychology Review, 25*(7), 917-934.

Smith, A.P.R., Henson, R.N.A., Dolan, R.J., & Rugg, M.D. (2004). fMRI Correlates of the Episodic Retrieval of Emotional Contexts. *NeuroImage, 22,* 868-878.

Smith, B.R., Piel, A.K., & Candland, D.K. (2003). Numerity of a Socially Housed Hamadryas Baboon (Papio hamadryas) and a Socially Housed Squirrel Monkey (Saimiri sciureus). *Journal of Comparative Psychology, 117*(2), 217–225.

Sober, E. (1990). Let's Razor Ockham's Razor. In D. Knowles (Ed.), *Explanation and Its Limits* (pp. 73–94). Cambridge: Cambridge University Press.

Sober, E. (2005). Comparative Psychology Meets Evolutionary Biology: Morgan's Canon and Cladistic Parsimony. in L. Daston & G. Mitman (Eds.), *Thinking with Animals: New Perspectives on Anthropomorphism* (pp. 85-99). New York, NY: Columbia University Press.

Sober, E. (2015). *Ockham's Razors: A User's Manual*. Cambridge: Cambridge University Press.

Squire, L.R., & Alvarez, P. (1995). Retrograde Amnesia and Memory Consolidation: A Neurobiological Perspective. *Current Opinion in Neurobiology, 5*(2), 169-177.

Stalnaker, R. (1998). What Might Nonconceptual Content Be? *Philosophical Issues 9*, 339-352.

Steel, D. (2008). *Across the Boundaries: Extrapolation in Biology and Social Science.* Oxford: Oxford University Press.

Steinbeck, J. (2001). *Novels 1942-1952*. New York, NY: Library of America.

Sterelny, K. (2012). *The Evolved Apprentice*. Cambridge, MA: MIT Press.

Steward, H. (2018). Morgan's Canon: Animal Psychology in the Twentieth Century and Beyond. in P. Adamson & G.F. Edwards (Eds.), *Animals.* Oxford: Oxford University Press.

Stotz, K. (2010). Human Nature and Cognitive-Developmental Niche Construction. *Phenomenology and the Cognitive Sciences, 9*(4), 483-501.

Suddendorf, T., (2013). Mental Time Travel: Continuities and Discontinuities. *Trends in Cognitive Sciences, 17*(4), 151–152.

Suddendorf, T. & Corballis, M.C. (1997). Mental Time Travel and the Evolution of the Human Mind. *Genetic Social and General Psychology Monographs, 123*(2),133-167. Available at: http://cogprints.org/725/

Suddendorf, T. & Corballis, M.C. (2007). The Evolution of Foresight: What is Mental Time Travel, and is it Unique to Humans? B*ehavioral and Brain Sciences, 30*(3), 299–351.

Suddendorf, T., & Redshaw, J. (2013). The Development of Mental Scenario Building and Episodic Foresight. *Annals of the New York Academy of Sciences, 1296*, 135-153.

Sutton, R.S., & Barto, A.G. (2018). *Reinforcement Learning: An Introduction* (2nd Ed.). Cambridge, MA: MIT Press.

Templer, V.L. & Hampton, R.R. (2013). Episodic Memory in Nonhuman Animals. *Current Biology, 23*(17), R801-R806.

Terrace, H., Son, L.K., & Brannon, E.M. (2003). Serial Expertise of Rhesus Macaques. *Psychological Science, 14*, 66–73.

Terrace, H. (2005). The Simultaneous Chain: A New Approach to Serial Learning. *Trends in Cognitive Sciences, 9*(4), 202-210.

Tolman, E.C. (1948). Cognitive Maps in Rats and Men. *Psychological Review*, *55*, 189-208.

Tolman, E.C., Ritchie, B.G., & Kalish, D. (1946). Studies in Spatial Learning: I. Orientation and the Short-Cut. *Journal of Experimental Psychology, 36*, 13-24.

Treves, A. & Rolls, E.T. (1994). Computational Analysis of the Role of the Hippocampus in Memory. *Hippocampus, 4*/3, 374-391.

Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381–403), New York, NY: Academic Press.

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

Tulving, E. (1985). Memory and Consciousness. *Canadian Psychology, 26*(1), 1-12.

Tulving, E. (2005). Episodic Memory and Autonoesis: Uniquely Human? In H. Terrace & J. Metcalfe (Eds.), *The Missing Link in Cognition: Origins of Self-Reflective Consciousness* (pp. 3-56). Oxford: Oxford University Press.

Tye, M. (2016). *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* Oxford: Oxford University Press.

Ukuwela, K.D.B., de Silva, A., Mumpuni, Fry, B.G., Lee, M.S.Y., & Sanders, K.L. (2013). Molecular Evidence that the Deadliest Sea Snake *Enhydrina schistosa* (Elapidae: Hydrophiinae) Consists of Two Convergent Species. *Molecular Phylogenetics and Evolution, 66*, 262-269.

Vaidya, C.J., Zhao, M., Desmond, J.E., & Gabrieli, J.D. (2002). Evidence for Cortical Encoding Specificity in Episodic Memory: Memory-Induced Re-Activation of Picture Processing Areas. *Neuropsychologia, 40*(12), 2136-2143.

van Duijn, M., Keijzer, F., & Franken, D. (2006). Principles of Minimal Cognition: Casting Cognition as Sensorimotor Coordination. *Adaptive Behavior, 14*(2), 157-170.

van Gelder, T. (1995). What Might Cognition be if not Computation? *The Journal of Philosophy, 97*, 345–381.

van Horik, J. O., Clayton, N. S. & Emery, N. J. (2012). Convergent Evolution of Cognition in Corvids, Apes and Other Animals. In T. K. Shackelford & J. Vonk (Eds.), *The Oxford Handbook of Comparative Evolutionary Psychology* (pp. 80–101). Oxford: Oxford University Press.

Velleman, J.D. (1991). Well-Being and Time. *Pacific Philosophical Quarterly, 72,* 48-77.

Velleman, J.D. (2003). Narrative Explanation. *The Philosophical Review, 112*(1), 1-25

Viera, G.A. (2019). The Fragmentary Model of Temporal Experience and the Mirroring Constraint. *Philosophical Studies, 176*(1), 21-44.

Warhol, A., & Benirschke, K. (1986) *Vanishing Animals*. New York, NY: Springer.

Watumull, J., Hauser, M.D., Roberts, I.G., & Hornstein, N. (2014). On Recursion. *Frontiers in Psychology, 4*, 1017.

Weber, M. (2004). *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press

Wesley, J. (1868). *The Poetical Works of John and Charles Wesley: Reprinted from the Originals, with the Last Corrections of the Authors; Together with the Poems of Charles Wesley Not Before Published*. London: Wesleyan-Methodist Conference Office.

Wheeler, M.E., Petersen, S.E. & Buckner, R.L. (2000). Memory's Echo: Vivid Remembering Reactivates Sensory-Specific Cortex. *Proceedings of the National Academy of Sciences of the United States of America, 97*(20), 11125-11129.

Williamson, T. (1994). *Vagueness*. New York, NY: Routledge.

Wilson, R.A. (2008). The Drink You Have When You're Not Having A Drink. *Mind and Language, 23*(3), 273-283.

Wilson, M.A., & McNaughton, B.L. (1994). Reactivation of Hippocampal Ensemble Memories During Sleep. *Science, 265*, 676–679.

Wittgenstein, L. (1989). *Lectures on Philosophical Psychology, 1946-1947.* Chicago, IL: University of Chicago Press.

Wittgenstein, L. (2009). *Philosophical Investigations* (4th Ed.) (G.E.M. Anscombe, P.M.S. Hacker & J. Schulte, Trans.). Hoboken, NJ: Wiley-Blackwell.

Wollheim, R. (1998). On Pictorial Representation. *Journal of Aesthetics and Art Criticism, 56*(3), 217-226.

Wozniak, R.H. (1997). Conwy Lloyd Morgan, Mental Evolution, and *The Introduction to Comparative Psychology*. Available at http://www.brynmawr.edu/psychology/rwozniak/morgan.html

Wu, C.-T., Haggerty, D., Kemere, C., & Ji, D. (2017). Hippocampal Awake Replay in Fear Memory Retrieval. *Nature Neuroscience, 20*, 571–580.

Wynne, C.D.L. (2004). The Perils of Anthropomorphism. *Nature, 428,* 606.

Zacks, J., & Tversky, B. (2001). Event Structure in Perception and Conception. *Psychological Bulletin, 127*. 3-21.

Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., & Reynolds, J.R. (2007). Event Perception: A Mind-Brain Perspective. *Psychological Bulletin, 133*(2), 273–293.

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives Without Imagery—Congenital Aphantasia. *Cortex, 73,* 378-380.

Zentall, T.R. (2005). Animals May Not be Stuck in Time. *Learning and Motivation, 36*, 208–225.

Zentall, T.R. (2006). Mental Time Travel in Animals: a Challenging Question. *Behavioural Processes, 72*, 173–183.

Zentall, T.R. (2013). Animals Represent the Past and the Future. *Evolutionary Psychology, 11*(3), 573-590.

Zhou, W. & Crystal, J. (2009). Evidence for Remembering When Events Occurred in a Rodent Model of Episodic Memory. *Proceedings of the National Academy of Sciences, 106*(23), 9525-9529.

Zhou, W., Hohmann, A.G., and Crystal, J.D. (2012). Rats answer an Unexpected Question After Incidental Encoding. *Current Biology, 22*, 1149-1153.

Zinkivskay, A., Nazir, F., & Smulders, T.V. (2009). *What-Where-When* Memory in Magpies (*Pica pica*). *Animal Cognition, 12*, 119-125.

Ziv, Y., Burns, L., Cocker, E., Hamel, E.O., Ghosh, K.K., Kitch, L.J., El Gamal, A., & Schnitzer, M.J. (2013). Long-term Dynamics of CA1 Hippocampal Place Codes. *Nature Neuroscience,* 16, 264–266.