DOI: 10.1111/nous.12480

ARTICLE

Noûs

Disagreement & classification in comparative cognitive science

Alexandria Boyle

London School of Economics and Political Science

Correspondence
Alexandria Boyle, London School of Economics and Political Science.
Email: a.boyle2@lse.ac.uk

Funding information
UK Research and Innovation, Grant/Award Number: MR/W00741X/1

Abstract

Comparative cognitive science often involves asking questions like ‘Do nonhumans have C?’ where C is a capacity we take humans to have. These questions frequently generate unproductive disagreements, in which one party affirms and the other denies that nonhumans have the relevant capacity on the basis of the same evidence. I argue that these questions can be productively understood as questions about natural kinds: do nonhuman capacities fall into the same natural kinds as our own? Understanding such questions in this way has several advantages: it preserves the intuition that these are substantive empirical questions worth asking; it helps us to understand why they so frequently give rise to disagreements of the kind described; and it provides clues about how to diagnose and resolve them.

1 | INTRODUCTION

Many research programmes in comparative psychology focus on a question like, ‘Do nonhuman animals[1] have capacity C?’, where C is a capacity we take humans to have. For instance, there are research programmes investigating whether animals episodically remember the past and plan for the future, whether they engage in causal or numerical cognition, whether they recognise their own reflections, whether they are metacognitively aware of their own mental states, and whether they ascribe mental states to others. All of these research programmes begin with a capacity we take typical adult humans to have and investigate whether animals have that capacity.

[1] Hereafter, simply ‘animals’.

These comparative questions also arise in artificial intelligence (AI) research, though they do not structure AI research in quite the same way. AI researchers don't tend to begin with a question about whether artificial systems have the capacities we do. They are often motivated by more practical goals – to create machines that do certain things. But many achievements of AI invite description in terms of human capacities. For instance, it is tempting to say there are machines that see, understand language, make decisions, remember things and have goals. These ascriptions of human capacities to artificial systems properly invite scrutiny, and evaluating them involves asking the same sorts of questions that drive research in comparative psychology: do artificial systems really have the capacities we do?

In what follows, I'll use 'comparative cognitive science' as a catch-all term for both comparative psychology and AI, insofar as each involves asking these comparative questions. As I'll argue, these questions often generate disagreements in which one party affirms and the other denies that a nonhuman has a certain capacity, despite both parties having access to the same evidence. It is often unclear what is at stake in these disagreements, and how they should be arbitrated. As a result, they can become unproductive or stagnate. This feels unsatisfactory, since in most cases, 'do nonhumans have C?' is, *prima facie*, a substantive empirical question.

In this paper, I argue that we can make progress with these disputes by framing them in terms of natural kinds. That is, I propose that we understand the question 'Do nonhumans have the capacities we have?' as: 'Do the capacities of nonhumans fall into the same natural kinds as our own?'. Understanding comparative questions in this way has several advantages. First, it preserves the intuition that these are substantive empirical questions which are worth asking. Second, it helps to explain why these questions are liable to give rise to the kind of disagreement I describe. And third, it provides clues about how to diagnose these disagreements and resolve them.

On the pragmatic naturalist view of natural kinds I adopt (Magnus, 2012), different scientific domains operate with different natural kinds. As I argue, comparative cognitive science is home to multiple scientific domains, which classify capacities differently. So, comparative questions about whether nonhumans have the capacities we have cannot be addressed without a particular domain in view. Within a given domain, it may be a substantive, empirical question whether an animal has the capacities we do – and it may be a difficult one, in view of the uncertainty, indeterminacy and fuzziness characteristic of classification in the cognitive sciences. But, as I argue below, some disagreements in comparative cognitive science may result from attempts to address questions without any particular domain in view, whilst others may result from 'domain mismatch', in which the parties to a dispute are addressing a comparative question with reference to different domains. In this last case, the dispute may be verbal, but perhaps not trivially so: it may reflect deeper disagreement about what the goals of comparative cognitive science are, or ought to be.

I begin in §2 by fleshing out the type of disagreement I have in mind, and briefly discussing some tempting responses. In §3, I introduce the framework of natural kinds, and argue that comparative questions should be understood as questions about natural kinds. In §4, I show how this construal of comparative questions helps us to understand why they so frequently give rise to disagreements of various kinds, and in §5, I offer some suggestions for making progress.

2 DISAGREEMENTS

I begin by introducing the focal phenomenon in more detail, by way of a few examples.

First, consider the question whether rats have empathy – roughly, the capacity to recognise and share the emotional states of another creature. In support of the hypothesis that rats do experience empathy, it's been observed that rats will take action to release a conspecific from a restraining cage, at no material benefit to themselves and sometimes at a cost. The researchers concluded that 'rats free their cagemate in order to end distress, either their own or that of the trapped rat', and that this emotional motivation is 'arguably the rodent homolog of empathy' (Bartal et al., 2011, p. 1430). However, others – let's call them 'sceptics' – have argued against this, saying that there is no reason to think that rats are driven by 'the psychological goal of improving a groupmate's wellbeing' (Vasconcelos et al., 2012, p. 911). To be justified in ascribing empathy to rats, sceptics say, we would need to establish that rats' behaviour was not motivated by some other feature of the situation. This would mean, for instance, demonstrating that the rats only act when the trapped rat is distressed, and that their action is sensitive to the distress of the trapped conspecific – such that, for example, if moving away from the trapped rat would reduce its distress, the rats would do that instead. Someone in the first camp – call them 'optimists' – might respond that this sets too high a bar. Even in humans, it's not clear that empathic behaviour satisfies these requirements. For instance, one might feel moved to hug somebody whose distress will be made worse by physical contact. It's natural to think of this as motivated by empathy, despite its not being entirely sensitive to the distress of the other. Of course, a sceptic might disagree, arguing that in this situation we are not really motivated by empathy but are acting to improve our *own* emotional state. And an optimist might reply that this neglects the 'other-directedness' of such behaviours, which seem importantly different from other actions taken with a view to lessening one's own distress, like removing oneself from the situation.

This back and forth about what empathy is and how to demonstrate it in animals could go on. Different views about what empathy is unsurprisingly yield different judgments about whether rats are empathic: some make it 'easier' to establish empathy in rats, whilst others make it more 'difficult'. It's unclear how one ought to choose between rival views of empathy in this sort of case, and whether anything substantive hangs on the question. One might be tempted to abandon the question of whether rats, or other animals, are empathic altogether. But this feels unsatisfactory, since this seems like a substantive, important, empirical question.

This pattern of disagreement is replicated elsewhere in comparative psychology. There isn't the space to outline further cases in detail – but disagreements of this kind arguably arise, *inter alia*, in debates about whether animals have theory of mind (Heyes, 2015), engage in metacognition (Carruthers, 2008; Smith, 2009), use insight to solve problems (Shettleworth, 2012), or have episodic memory or episodic foresight (Suddendorf et al., 2009). I don't mean to suggest that *all* entrenched disagreements between comparative psychologists are of this sort; comparative psychologists certainly disagree about many other things. My suggestion is just that disagreements of the sort I've described are extremely common.

These disagreements also arise in connection with AI. Take AlphaGo, an AI system which defeated 18-time world Go champion Lee Sedol in 2016. In the second game of the match, AlphaGo made an unusual move – Move 37 – which commentators agreed no human Go player would have made. It later became clear that Move 37 was key to AlphaGo's establishing control of the board (Kohs, 2017). Move 37, along with some other distinctive patterns in AlphaGo's play, was widely characterised as a demonstration of creative intelligence in AI. But sceptics (e.g. du Sautoy, 2019) challenged this, arguing that AlphaGo displayed no more creative insight with Move 37 than 'when it played pedestrian moves' (Dixon, 2019). As Marta Halina argues, the way AlphaGo learns seems to support the optimist view. AlphaGo uses Monte Carlo tree search (MCTS) to construct a 'search tree' of possible moves, informed and constrained by its training, selecting the one with the highest expected value. MCTS can be viewed as analogous to mental simulation or 'scenario building', an important component of creative problem solving in biological systems (Halina, 2021, sec. 6.1). But the domain-specificity of AlphaGo's processes pushes in the direction of scepticism. AlphaGo 'lacks the input to know which actions might be rewarding in contexts other than a standard game of Go' – including minor variants on Go (Halina, 2021, p. 325). So, a sceptic might argue that AlphaGo is not creative because it lacks 'the capacity to solve novel problems through a domain general understanding of the world' (Halina, 2021, p. 326). An optimist, however, might deny that creativity must be domain general, saying that creative intelligence may be confined to a specific domain.

Once again, we might continue this exchange, trading accounts of what creativity requires. And again, it is unclear how one ought to arbitrate this dispute, or what really hangs on it – so one might be tempted to abandon the question of whether AlphaGo is creative altogether. But as before, this feels unsatisfactory. And as in comparative psychology, the same sort of disagreement threatens many AI research programmes: for instance, those which raise questions about whether chatbots 'understand' what they say (Weinberg et al., 2020), whether machine vision systems see as we do (Hendrycks & Gimpel, 2017), whether experience replay in machine learning systems involves anything like biological experience replay (Wittkuhn et al., 2021), and whether machine theory of mind really involves mindreading (Shevlin & Halina, 2019).

These disagreements follow a common pattern. In general terms:

1. Initial observations appear to support the claim that some nonhuman agent(s) has a capacity C.
2. Sceptics point out actual or epistemically possible differences between the nonhuman capacity and human manifestations of C, arguing that the nonhuman agent lacks C on this basis.
3. Whilst not denying those differences, optimists reply that this judgment relies on an unduly restrictive conception of C, noting that an alternative conception would support the claim that the nonhuman agent has C.
4. The debate becomes unproductive: it becomes unclear how the dispute might be settled in a principled way, and whether anything of substance hangs on it.

Disagreements of this sort are endemic in comparative cognitive science. My interest in this paper is in understanding the nature of these disputes, and in how to make progress with them.

Before moving on though, I want to anticipate two responses to these disputes one might immediately find tempting. Both are tempting for a reason, but neither provides an adequate account of (or response to) pervasive disagreement in comparative cognitive science. I don't propose to discuss these responses in detail – only to acknowledge what is attractive about them and indicate the respect in which I think they fall short. I substantiate these critical remarks later, after developing my positive view.

First, it's tempting to say that these disputes are merely verbal, and that since nothing hangs on verbal disputes, they should be abandoned. Let's say, following Chalmers (2011, p. 522), that a dispute over a sentence *S* is verbal 'when, for some expression *T* in *S*, the parties disagree about the meaning of *T*, and the dispute over *S* arises wholly in virtue of this disagreement regarding *T*.' Clearly, the disputes I've described are all verbal to some extent: they all involve a disagreement about the correct interpretation of a term. But I doubt that many of them are *wholly* verbal – that is, I doubt that disagreements over the meanings of terms explain these disputes without remainder.

One method for detecting and resolving verbal disputes is the elimination method, in which one bars the use of the contested term *T*, and tries to find a new sentence *S'* in the restricted vocabulary such that the parties disagree over *S'* and *S'* is part of their dispute over *S*. If there is an *S'*, then their dispute is at least partly non-verbal – they have a substantive disagreement about *S'*. If there is no such *S'*, then their dispute is wholly verbal; eliminating the term eliminates the disagreement (Chalmers, 2011, pp. 526–527). As I argue in more detail below, applying this method to disputes in comparative cognitive science may sometimes reveal them to be wholly verbal, but in many cases will reveal an underlying disagreement of a more substantive kind. Even where the dispute *is* wholly verbal, that there will be cases in which eliminating the contested term or abandoning the disagreement will not be the most productive course of action.

Second, some researchers have suggested that these disputes arise because we are asking the wrong questions. The suggestion is that questions like 'Do nonhumans have C?' demand 'all-or-nothing' answers when the reality may be that some aspects of a capacity and not others are present in nonhumans. As such, we should replace 'top-down' questions about capacities with a 'bottom-up' approach focusing on mechanisms (de Waal & Ferrari, 2010; Eaton et al., 2018; Shettleworth, 2007). On a bottom-up approach, rather than asking whether nonhumans have C, we instead break C down into its component mechanisms and investigate which of these *mechanisms* are found in nonhumans. We may find some of those mechanisms and not others, but the overarching question of whether C is present can be set aside.

There is a lot to agree with in these proposals. In particular, as I'll argue below, questions about the mechanisms underlying a capacity are often critical to determining whether nonhumans have it. Nevertheless, I suggest that we should not be too quick to set aside 'top-down' questions about capacities. As should become clear in what follows, questions about mechanisms and questions about capacities are companions in guilt: it can be as difficult to reach agreement on whether humans and nonhumans share a mechanism as it is to agree on whether they share a capacity. Since these disagreements will arise whether we take a top-down or bottom-up approach to comparative research, there is good reason to attempt to resolve them.

3 | NATURAL KINDS

In this section, I propose that the question 'Do nonhumans have capacity C?', where C is a capacity we take humans to have, should be understood as a question about whether nonhuman capacities fall into the same natural kinds as our own.

A kind is a group or category of things. To call a kind 'natural' is to claim that it corresponds to some real structure in the world, rather than merely reflecting human interests. Paradigm examples of natural kinds are things like **gold**, **polar bears**, and **electrons**.[2] Arbitrary groupings of things, like the category of **things currently on my desk** are not natural kinds; nor are less arbitrary groupings that merely reflect human conventions and interests, like **kitchen utensils**. The contrast class with 'natural' here is 'arbitrary' or 'conventional', rather than 'artificial': to call a kind 'natural' is not to say that its members are naturally occurring. Synthesised diamonds, for instance, are members of the natural kind **diamond**, despite being made in a lab. All this is relatively uncontroversial. Beyond this, what makes some kinds natural and what natural kinds are like are subjects of ongoing debate.

In this paper, I adopt P. D. Magnus' (2012) 'pragmatic naturalism' about natural kinds. Broadly speaking, I take this account to be motivated by a combination of three *prima facie* plausible ideas.

[2] Throughout this paper I use **bold** when denoting a kind, rather than the entities it comprises – such that, for instance, '**electrons**' refers to the kind and 'electrons' to the things.

The first is that the function of a theory of natural kinds is to make sense of how 'the world constrains our taxonomy' (Magnus, 2012, p. 4). The second is that scientific taxonomy is constrained by the world in many domains, including the special sciences: in all these domains, if scientific theorising is to be successful, taxonomies must correspond to genuine structure in the world. The third is that the worldly structures that constrain taxonomy in different scientific domains vary considerably. The picture that emerges is one on which natural kinds – the worldly structures that constrain scientific taxonomies – are metaphysically diverse, but are theoretically unified by their indispensability to successful science.

On pragmatic naturalism, then, natural kinds are the categories that are indispensable for successful science in a given scientific domain. That is, natural kinds are projectible categories which support our capacity to provide good explanations and predictions about the phenomena in a domain. To say that a kind is indispensable is to say that it is non-fungible: that no other category would support our ability to make successful predictions and explanations in the relevant domain to the same or a higher degree. Whether a category is a natural kind is therefore an empirical question. When we discover that a category once believed to be a natural kind in a domain is less effective than a newly discovered category at supporting prediction and explanation in that domain, we are discovering that it is not a natural kind in that domain after all.

What metaphysical features natural kinds have is also an empirical question, on this view. As such, pragmatic naturalism departs from a long tradition of giving accounts of natural kinds in terms of their metaphysical features. Among other things, in this tradition it has been claimed that natural kinds are characterised by intrinsic features which all members of a kind share; that natural kinds have sharp boundaries, such that it is always clear whether an individual is a member of a kind; and that natural kinds form a hierarchy, such that 'natural kinds are nested in one another like stacking dolls' (Magnus, 2012, p. 37). Pragmatic naturalism rejects these claims. On this view, there is no question of determining a priori that natural kinds have certain features: we find out what features natural kinds have by doing science successfully (Magnus, 2023). And when we look to the sciences, we find that success requires different things in different domains. Perhaps the most general thing we might say about the metaphysics of natural kinds, on this view, is that the members of a domain's natural kinds tend to exhibit clusters of properties which are theoretically significant in that domain. But what, if anything, underpins the clustering of a kind's members – what 'holds the kind together' (Magnus, 2014, p. 472) – appears to vary across domains. All of this is to say that the natural kinds of different domains might, metaphysically speaking, be quite different sorts of category.

The chemical elements may be 'essentialist kinds' – kinds whose members share an intrinsic property common to all and only members of the kind, which explains their shared surface features. On this view, what unifies the kind **gold** is that all and only instances of **gold** have atomic number 79, and this (perhaps together with other intrinsic, essential properties of gold) explains their surface properties. In the life sciences, universally shared underlying essences – and hence, essentialist kinds – are in short supply. Some natural kinds in the life sciences are thought to be homeostatic property cluster (HPC) kinds: that is, property clusters whose clustering is explained by a 'homeostatic mechanism' which causes the properties to cluster (Boyd, 1989, 1991). Some diseases are probably HPC kinds. **Chickenpox**, for instance, is characterised by a cluster of symptoms (a spotty rash which itches and blisters, a temperature, aches and pains, loss of appetite) whose co-occurrence is causally explained by a homeostatic mechanism (infection with varicella zoster virus, which spreads and replicates in the body in such a way as to cause the characteristic rash and trigger an immune response). Other biological natural kinds are historical: for instance, it might be that the cluster of properties characterising members of the species **chimpanzee** is to be explained by their shared ancestry,[3] or that the property cluster characterising members of the kind **vertebrate forelimb** is to be explained by their descent from the same ancestral structure.

Similarly, natural kinds in cognitive science might be a diverse bunch. As in the life sciences, essentialist kinds are probably thin on the ground here. But cognitive kinds might be unified in either of the other ways mentioned above. Some may be HPC kinds: characterised by a theoretically significant cluster of *functional* properties, whose clustering is explained in terms of underlying cognitive or neural mechanisms. Others may be historical or homologous kinds: characterised by a theoretically significant cluster of properties whose clustering in different individuals is explained by their shared evolutionary or developmental history. Since it's likely that both HPC and historical kinds are indispensable to cognitive science, I'll proceed with both views of cognitive kinds in mind.[4]

Aside from finding the view independently plausible, I adopt the framework of pragmatic naturalism in this paper for two reasons. The first is that, because pragmatic naturalism characterises natural kinds in epistemic rather than metaphysical terms, adopting it here enables us to sidestep several ongoing debates about natural kinds' metaphysical features. These debates, though important, are orthogonal to my concerns in this paper. A consequence of adopting this more inclusive epistemic account, I hope, is that the argument of this paper should not be hostage to the fortunes of any particular account of the metaphysics of cognitive kinds. The second is that this account provides the resources for making sense of pervasive, entrenched disagreements in comparative cognitive science. As I'll outline in the next section, the key to unlocking these disputes is pragmatic naturalism's construal of natural kinds as *domain relative*: the naturalness of a natural kind on this view is understood as its indispensability to scientific success *in a given scientific domain*. It is only right to acknowledge, though, that the two points that motivate the adoption of pragmatic naturalism in this context also make it somewhat controversial. So, whilst a full defence of pragmatic naturalism is well beyond the scope of this paper,[5] I should briefly address the concerns to which these two points might give rise.

First, as I've said, pragmatic naturalism departs from a tradition of characterising natural kinds in metaphysical terms, in favour of an account on which their 'naturalness' consists in their epistemic role in promoting successful science. Whilst I take this to be an advantage of the view, there is significant debate about whether an epistemic account of natural kinds can succeed. My inclination is to think that, given the apparent metaphysical disunity of natural kinds in various scientific domains, what these kinds have in common can only be their role in constraining successful science. But one might think that, rather than motivating an epistemic account, this provides a reason to retire the label 'natural kinds' altogether (e.g. Ludwig, 2018), or to restrict it to a more metaphysically unified type of category, such as essentialist kinds (e.g. Ellis, 2008) or substantial kind universals (Tahko, 2021). I have not much to say to this, except that the issue risks becoming terminological (see Magnus, 2012, p. 4). If the reader shares this concern, they are free to

[3] Whether such species in particular and historical kinds in general are a sub-category of HPC kinds is a matter for debate, which I sidestep here. See (Magnus, 2012) and (Slater, 2015) for opposing views.

[4] I use the term 'cognitive' in a very thin, general sense in this paper, derivative on its appearance in 'cognitive science'. That is to say, in talking about cognitive capacities, systems, kinds or homologies etc., I mean only to pick out the capacities, systems, kinds or homologies with which cognitive science is concerned. There are contested questions about what 'counts' as a cognitive phenomenon, what (if anything) unifies cognitive phenomena and whether cognitive phenomena themselves form a natural kind. These issues are orthogonal to my concerns in this paper, so I set them aside here. (See, e.g., Akagi, 2018; Ramsey, 2017 for discussion.)

[5] For that, I can do no better than to refer the reader to Magnus (2012).

substitute ‘*scientifically important categories*’ wherever I use ‘natural kinds’ in this paper. Still, one might think that pragmatic naturalism can’t be the whole story about natural kinds, even in this inclusive sense of the term. If the idea is that natural kinds are the ones that constrain taxonomy, one might reasonably think that a full account of requires explaining *how* they constrain successful science, in virtue of their metaphysical features (Kendig & Grey, 2021; Lemeire, 2020). If so, pragmatic naturalism must strike one as at best an incomplete account of natural kinds. That may be so, but it need not concern us here. The arguments of this paper do not depend upon pragmatic naturalism being the final word on natural kinds: they should be consistent with various ways of filling in the metaphysical details, so long as the core idea that natural kinds are domain-relative is preserved.

But this brings us to the second sticking point. Pragmatic naturalism denies what Magnus calls the ‘simpliciter assumption’ (2012, p. 39): that a natural kind must be a natural kind *simpliciter*, without reference to anything else. Instead, according to pragmatic naturalism, something can be a natural kind only for some scientific domain.[6] As I’ve said, this idea will be doing a lot of heavy lifting in what follows. But one might want to resist this claim.

One concern is that this domain-relative view of natural kinds entails that natural kinds are mind-dependent, such that what natural kinds there are depends in some problematic way on humans and their interests. This would be a problem, because the distinction between natural kinds and arbitrary or conventional kinds is supposed to be precisely that the natural kinds are ‘out there’ independently of us, whereas the non-natural ones merely reflect human eccentricities. Laura Franklin-Hall expresses this concern, writing that if something is a natural kind only relative to a given domain, then what the natural kinds are will ‘depend on what our particular epistemic projects [. . .] happen to be’ (2015, p. 940) – and that if hypothetical scientists were interested in different things, ‘there could be different natural kinds relative to these different investigators (or to the fields in which they work)’ (2015, p. 945).

This is not a consequence of the domain relativity of natural kinds, however. To see why, we first need to get clearer on what is meant by ‘scientific domain’. Following Magnus (2012, p. 44), I’ll take a scientific domain to be a collection of objects and phenomena. So, the domain of biology, let’s say, is the domain comprising living organisms, their vital processes and their interaction with their environment, whilst the domain of physics comprises energy, force, matter, their constituents, their relationship and their behaviour in space and time. To say that **cuttlefish** is a natural kind in biology but not in physics is to say that successful inductions and explanations about the first domain require classifying some things as cuttlefish, but successful inductions and explanations in the second domain do not require classifying anything as a cuttlefish. Assuming, for the sake of argument, that this is true, it is true independently of us. As it happens, we *are* interested in biology. But the domain of biology would exist, and **cuttlefish** would still be indispensable to successful science in that domain, even if we were not interested. As Magnus (2023, p. 13) puts it, ‘the categories are constraints on potential enquiry even if we do not conduct it.’ This is the sense in which this view of natural kinds is both pragmatic and naturalist. Our interests determine which domains we investigate, and hence which natural kinds we discover. But what the natural kinds *are* is independent of our interests – natural kinds are ‘out there’ to be discovered, not invented.

Still, one might be resistant to the idea that the *naturalness* of a kind could be domain relative. After all, I’ve said that on this view the natural kinds are what they are, independently of our interests, and that domains are independent of us too. So, rather than taking their claims to naturalness

[6] For other accounts of natural kinds which reject the simpliciter assumption, see Boyd (1989, 1991) and Slater (2015).

to be domain relative, one might wish to say instead that all natural kinds are absolutely natural, but that each is indispensable to explanation and prediction in some domains and not others (see, e.g., Godman, 2014). From the point of view of this paper, I think little hangs on this, so the reader is free to adopt the latter construal if they prefer. The disadvantage is simply that it reopens the question of what naturalness consists in. Recall that on pragmatic naturalism, naturalness consists precisely in indispensability to successful explanation and prediction *in a given domain*. So, to say that naturalness is domain invariant, but that indispensability to explanation and prediction domain relative, would be a contradiction in terms (see Magnus, 2018). The objection, I think, is a holdover from the assumption that the naturalness of natural kinds is supposed to be a matter of their metaphysical features, rather than their epistemological role – and, as I've said, there are reasons to be less than sanguine about the prospects for metaphysical accounts of natural kinds. But at any rate, if the reader prefers to reserve the term 'natural' for some other feature of natural kinds, they are free to paraphrase. The key claim is that indispensability to scientific explanation and prediction is always relative to some scientific domain.

I don't pretend to have provided anything like a knock-down argument for pragmatic naturalism, or surveyed all of the objections it might attract. Natural kinds are the subject of a lively ongoing debate in which relatively little is settled. Getting much further into the weeds of this debate would take us too far from the main business of this paper. But hopefully I have done enough to motivate the key elements of the view, to make clear what they do and do not entail, and to address some potential objections. What remains is to make good on the claim that this picture of natural kinds will help us to make sense of pervasive, entrenched disagreement in comparative cognitive science. To the extent that pragmatic naturalism does facilitate this, I take that to be a recommendation for the view.

Broadly speaking, then, my suggestion is that we understand comparative questions like 'Do nonhumans have capacity C?' as questions about cognitive kinds. For instance, when we ask whether animals have empathy, we are asking whether any animals have a capacity which – in common with the various instances of empathy so far observed in humans – is a member of **empathy**, where **empathy** is assumed to be a natural kind. Minimally, that amounts to hypothesising that the kind **empathy** is indispensable to successful prediction and/or explanation in the relevant domain and asking whether any nonhuman capacity qualifies as a member of that kind. This is a complicated question and would need some considerable pinning down before it could be answered. But it has the benefit of being, *prima facie*, a substantive empirical question, rather than a merely verbal one.

Understanding comparative questions in this way therefore has the twin advantages of preserving the intuition that something substantive is at stake when comparative cognitive scientists ask them, as well as making sense of why they so frequently do. On this view, comparative cognitive scientists are not simply engaged in a labelling exercise but in a classificatory one. When biologists ask whether an organism belongs to a particular species, this is a substantive question: it is about whether this way of classifying organisms reflects real causal-explanatory structure in the biological domain, and so supports better explanations and predictions in that domain than any alternative classification. So too, when comparative cognitive scientists ask whether the capacities of nonhuman agents fall into the same kinds as our own, they are asking a substantive question: which way of classifying nonhumans' capacities best reflects causal-explanatory structure in the cognitive domain, and so will help us to best explain and predict its phenomena?[7]

[7] For ease of presentation, I assume for the time being that there is such a thing as 'the biological domain' and 'the cognitive domain'. I revisit this assumption later.

There are good reasons for thinking that this is, at least approximately, how these questions are understood by comparative cognitive scientists. Whilst comparative psychologists don't typically use the language of cognitive kinds, the way they investigate comparative questions suggests they have in mind something like this. Their investigations typically proceed by looking for evidence of a nonhuman capacity with some high-level functional similarities to a human capacity of interest. If that evidence is forthcoming, scientists probe the nonhuman capacity to determine the extent to which it is functionally similar. If a similar cluster of functional properties is found, they investigate whether it shares a unifying basis with the relevant human capacity. That might mean investigating whether it is underpinned by similar mechanisms: 'functional similarity of input-output relations [...] guides mechanistic investigations at both behavioural and neural levels' (Shettleworth, 2012, p. 225). Or it might mean considering whether there is a shared history explaining the presence of this cluster across species.

Similarly, evaluations of the capacities of artificial agents typically focus not only on their functional properties but also their mechanistic underpinnings, suggesting something like an HPC view of cognitive kinds in artificial systems. For instance, Halina (2021) takes considerations of the mechanisms underpinning AlphaGo's outputs to be relevant to determining whether it exhibits creative intelligence. Similarly, adversarial examples in machine vision research are sometimes thought to make trouble for the view that the machines 'really' see, because they indicate deep dissimilarities between the mechanisms underlying biological and machine vision (e.g. Hendrycks & Gimpel, 2017; but see Firestone, 2020). One might also ask questions about the instantiation of historical cognitive kinds in artificial systems. For instance, one might investigate whether two artificial systems are operating with the same kind of representation by asking whether they have a similar learning history.

In short, comparative cognitive scientists appeal to the functional features of nonhuman capacities and their mechanistic and historical underpinnings when evaluating comparative questions, suggesting that these questions are taken to be about cognitive kinds. Having said that, to the extent that my proposal is not descriptively adequate as an account of how these questions are understood in comparative cognitive science, I intend it to be taken as a normative claim. In other words, this is how these questions *should* be understood – not only because on this reading they come out as substantive questions worth asking, but because this will help us to understand and make progress with the disagreements they generate. So, if it turns out that this is not how comparative cognitive scientists understand these questions, this should not be troubling. In that case, my proposal should be treated as a piece of conceptual engineering.

4 | MAKING SENSE OF DISAGREEMENT

Construing comparative questions as questions about cognitive kinds gives us several reasons to expect that they will give rise to disagreement.

First, recall that on the account of natural kinds I've recommended, the natural kinds of one scientific domain may differ from the natural kinds of another. This creates space for disagreement in comparative cognitive science by way of a phenomenon I'll call 'domain mismatch'. Domain mismatch occurs when something is a kind for more than one scientific domain, and those domains demand different ways of characterising and delineating the kind.

For example, the kind **polar bear** is a kind for at least two scientific domains: ecology and genetics. These domains are concerned with some of the same objects and phenomena, but do not entirely overlap. As a result, success in these domains requires different things. Ecologists are concerned with understanding the position of organisms within ecosystems. So, for ecologists, whether something should be counted as a polar bear is determined by whether it occupies the same ecological position as polar bears. Geneticists are concerned with the inheritance and distribution of traits in populations of individuals exchanging genetic material. For geneticists, whether something should be counted as a polar bear is determined by whether it is genetically related to other polar bears and can interbreed with other polar bears to produce fertile offspring. As Henry Taylor (2019) argues, these domains yield different verdicts about membership in the kind **polar bear** in the case of ABC island bears. ABC island bears are the result of interbreeding between polar bears and brown bears and can interbreed with both to produce fertile offspring. They have adaptations to their environment which are shared with brown bears but not polar bears; most obviously, they are brown. So, prototypical (white) polar bears and ABC island bears occupy distinct ecological positions. This yields the result that ABC island bears both are and are not polar bears – but there is no contradiction, since these apparently conflicting verdicts belong to different domains. ABC island bears are members of the kind **polar bear** in the domain of genetics; in the domain of ecology, they are not.

One might resist my framing of this case. I've said that **polar bear** is a kind in two domains and is individuated differently in each. But given that they are individuated differently in each, one might say that this can't be right: there are surely two *distinct* kinds in these two domains, both of which happen to be called 'polar bear'. On this view, we might say that the term 'polar bear' is ambiguous between these distinct kinds, rather than that the same kind is individuated differently.

There is clearly something right about this. Given that **polar bear** in ecology and **polar bear** in genetics have different extensions, they cannot be the same kind. But to describe them as distinct kinds which simply happen to be called 'polar bear' is inadequate. This is not a simple case of lexical ambiguity, in which a term 'happens' to have two unrelated meanings – as 'duck' denotes (among other things) both members of certain groups of waterfowl and the act of quickly lowering one's head or body so as to avoid a projectile. 'Polar bear' in ecology and 'polar bear' in genetics clearly stand in a closer relation than this.

One way to make out the idea that these domains individuate the same kind in different ways might be to say that one of the kinds is a sub-kind of the other, or that both are sub-kinds of a superordinate natural kind. But neither of these claims seems promising. Consider the idea that one is a sub-kind of the other. It may well be the case that **genetic polar bear** includes all the members of **ecological polar bear**. But, if that's true, it is only contingently so. Were a population of ecological polar bears to become reproductively isolated from and incapable of reproducing with the rest, they would still be ecological polar bears so long as they retained the same adaptations to the same sort of position in the same sort of ecosystem. That being the case, **ecological polar bear** is not clearly a sub-kind of **genetic polar bear**.[8] This also casts doubt on the idea that there is a natural kind subsuming both, since unless there is some further thing unifying genetic and ecological polar bears, any superordinate kind would be disjunctive – and disjunctive kinds tend not to be indispensable to successful inductions and explanations.[9]

A better strategy treats the genetic and ecological **polar bear** kinds as related because of their shared relationship to an *investigative* kind. An investigative kind is 'a group of things that are presumed to belong together due to some underlying mechanism or a structural property' (Brigandt, 2003, p. 1309). An investigative kind is consequently associated with the search for a

[8] See Brusse (2016, sec. 4.1.1) for a similar argument about the different kinds denoted by 'planet'.

[9] But see Currie (2016) for an argument that the disjunctive kind **species** is a natural kind in palaeontology.

corresponding *natural* kind – a natural kind which includes most or all members of the investigative kind, and whose unifying basis explains and vindicates the idea those individuals 'belong together'. If a natural kind emerges from our investigation, we identify the investigative kind with the natural kind. If no natural kind emerges, we may abandon the idea that these individuals form a kind. But it might be that several natural kinds emerge, each including most or all of the investigative kind's members, each explaining the original phenomenon to some degree, and each being such that, had it been the only natural kind to emerge from the investigation, we would have identified it with the investigative kind. In such circumstances, pluralism seems appropriate (Brusse, 2016).

This is how I propose we understand the polar bear situation. First, a phenomenon was observed: big, white bears living in the arctic. Because these bears seemed to 'belong together', we treated them as forming an investigative kind, **polar bear**, and began the search for a corresponding natural kind. The polar bear phenomenon appeared relevant to a number of scientific domains. Scientists in those domains discovered different natural kinds, each explaining the polar bear phenomenon to some degree: one explains it in terms of the bears' genetic features, another in terms of their ecological position. Their extensions differ, but both include the members of the original investigative kind **polar bear** – the big, white bears living in the arctic. Each is such that, had it been the only natural kind to emerge from the investigation, we'd have identified it with the investigative kind **polar bear**. But since there are two, we can't identify the investigative kind with either. In this situation, what seems right to say is that the investigation revealed that the term 'polar bear' to be polysemous, rather than ambiguous (Brusse, 2016): it picks out multiple natural kinds with overlapping extensions, each of which plays a role in explaining the polar bear phenomenon. This is what I mean by saying that the kind **polar bear** is individuated differently in different domains: it is an investigative kind taken up by several domains, which have found it to correspond to distinct natural kinds. In general, wherever an investigative kind is taken up in different scientific domains, domain mismatch may emerge.

Domain mismatch can generate verbal disputes, which can be diagnosed using the elimination method. Were an ecologist and a geneticist to disagree about whether ABC island bears are polar bears, for instance, eliminating the term 'polar bear' from their conversation might eliminate their disagreement, revealing their dispute to be verbal. But it is a further question whether ecologists and/or geneticists ought to stop using the term 'polar bear' in general. In fact, ecologists and geneticists seem quite capable of navigating the situation. Moreover, eliminating the term 'polar bear' would obscure the relationship between the genetic and ecological polar bear kind – that is, their shared relationship to a common investigative kind, in virtue of which both 'count' as polar bear kinds (see Brigandt, 2003, p. 1314). Relatedly, there may be some residual, non-verbal disagreement between the two scientists, which eliminating 'polar bear' would obscure: it could be that they disagree about which of these kinds best explains the polar bear phenomenon, or best corresponds to the investigative kind.

My suggestion is that domain mismatch arises systematically in comparative cognitive science. Comparative questions involve asking whether a cognitive kind **C**, familiar from human cognitive science, also has instances in nonhumans. Our account of **C** and our judgments about what counts as a member of **C** have so far been informed by what facilitates success in human cognitive science. But when we ask whether any nonhuman capacities are members of **C**, we are asking a question that belongs to a broader domain – one concerned with understanding not just human but also nonhuman minds. I take it that in most cases, it will be clear in advance that *if* any nonhumans have the capacity, there will be substantial differences between its human and nonhuman instances. These differences would likely preclude any putative nonhuman instances from falling into **C** as individuated by human cognitive science. So, when we ask whether any nonhuman capacity falls into **C**, we must implicitly be assuming a broader characterisation and delineation of the kind: effectively, proposing that **C**'s individuation in comparative cognitive science differs from its more familiar individuation in human cognitive science.

To take an example, consider mental simulation. It's a promising working hypothesis that **mental simulation** is a natural kind for the human cognitive sciences. Our understanding of mental simulation in humans is relatively good, so we're in a relatively good position to characterise **mental simulation** as a cognitive kind for the human cognitive sciences. Now, we might note that there are various sources of evidence for something *like* mental simulation in animals from a wide range of taxa, from corvids to cuttlefish, and in artificial agents like AlphaGo. So, we might ask: do any of these nonhuman capacities fall into the kind **mental simulation**? We might approach that question from the point of view of human cognitive science by asking: do these nonhuman capacities exhibit the functional features characteristic of mental simulation in humans, and are they underpinned by the same mechanisms as they are in humans? And would it promote success in human cognitive science to classify these nonhuman capacities as instances of mental simulation? I take it that the answer to these questions in at least some cases will almost certainly be 'no'. Here's just one reason: neither cuttlefish nor AlphaGo has a hippocampus, a brain area centrally involved in mental simulation in humans. So whatever mechanisms underpin their simulative capacities are likely to differ from the mechanisms underpinning mental simulation in humans in ways that matter when it comes to explaining and predicting human cognition. Corresponding to these mechanistic differences, there will probably be at least some differences between the functional profiles of these simulative capacities and our own. But this much would be obvious in advance to anyone asking the question whether these nonhuman capacities are instances of **mental simulation**. And I think that this demonstrates that in asking that question, we implicitly assuming a broader delineation of the kind.

The issue is not merely that in order to investigate a given capacity in nonhumans, we would need to measure and operationalise it differently than in humans – though of course this will be true in many cases, since the verbal measures and operationalisations used in human cognitive science will often be inapplicable. Rather, what I have in mind is that even if it were possible to apply the same measures to human and nonhuman subjects, we should nevertheless expect the capacities under investigation to vary across these populations, given the significant differences between them – in much the same way as we expect to find interspecific variations on anatomical phenomena like the heart (Boyle, 2022, p. 194). There is a more fundamental question here, about what it would *take* for different measures or operationalisations of a capacity to target the same phenomenon across populations whose members differ in myriad ways: under what circumstances do two capacities that differ belong to the same kind? This question can't be answered by deferring to the account of the relevant cognitive kind offered by human cognitive science, which nonhuman capacities are very likely not to satisfy. So, in asking this question, we must have a more inclusive delineation of the relevant capacity in view.

One might worry that the idea of domain mismatch puts pressure on the possibility of doing comparative cognitive science. If the kinds of comparative and human cognitive science inevitably differ, as I'm suggesting, how can we investigate whether animal and human capacities fall into the very same kinds, as I've proposed? There is no problem here, however, since – like polar bears – our capacities might fall into several related, overlapping but not co-extensive kinds. In short, I propose that in asking whether nonhumans have C, we are treating **C** as an investigative kind to which several natural kinds might correspond. That is, we are proposing that those instances of **C** observed in humans 'belong together' not only with one another, but with some *as yet unknown* but relevantly similar nonhuman capacities – and triggering the search for a natural kind which would both vindicate and explain that hypothesis. It may be that there is no such kind. By the same token, there might be several. But not just any natural kind is going to count as vindicating and explaining this hypothesis: to fit the bill, a kind would have to unify the familiar human capacity with a non-human capacity.

So, domain mismatch does not defeat the possibility of comparative cognitive science. But domain mismatch in comparative cognitive science is often more problematic than it was in the polar bear case. There, we said that **polar bear** belongs to two domains, and is individuated differently in each. But it is relatively clear that the relevant kinds in these two domains are both *polar bear kinds*. This is because the investigative kind sets a clear standard for what 'counts' as a polar bear kind: a natural kind which explains the apparent commonality between the big white bears in the arctic and has at least most of those bears as members. The genetic **polar bear** kind includes a few things we might not have anticipated – principally, ABC Island bears. But it is still a polar bear kind, since – like the ecological **polar bear** kind – it does what a polar bear kind is supposed to do: explains the commonality between big white bears in the arctic, and has those bears as members. Things are less clear in comparative cognitive science. In asking whether nonhumans have **C**, we treat **C** as an investigative kind with a broader membership than has previously been assumed and go looking both for evidence of this broader membership and the mechanisms that would explain it. To avoid changing the subject, any natural kind we uncover must at least include the human instances of **C** with which we are familiar. But it is difficult to specify in advance which *nonhuman* capacities should be included or excluded in any candidate **C** kind. In the absence of such a specification, disagreements over whether nonhuman capacities fall into the same kinds as our own – like **mental simulation** – seem inevitable.

In fact, the situation is more complicated than the discussion so far suggests. Until now, I have spoken as though comparative cognitive science were a single domain concerned with human and nonhuman cognitive systems. But just as biology cannot be productively treated as a single domain of inquiry concerned with 'living things', comparative cognitive science is an umbrella term for multiple domains. I don't mean only that comparative cognition and artificial intelligence are distinct domains, concerned with nonhuman animals and artificial agents respectively. Rather, I mean that even these two disciplines contain many domains of inquiry. We can see this if we ask *why* scientists ask a question like 'Do nonhumans have C?'. Their motivations for asking this question provide a clue as to which domain their question belongs to – that is, what collection of objects and phenomena is relevant to their question and constrains their answer to it.

Take empathy, for instance. We might ask whether animals have empathy for a number of reasons. For instance, we may be attempting to understand human empathy, its functions and dysfunctions. This might locate our question in the domain of biomedical science: roughly, the function and disfunction of cells, organs and systems in the human body. The right way of individuating **empathy** in this context will be the way that provides for the best explanations and inductions about the functions and disfunctions of human empathy. What **empathy** is here will be an empirical question – but we can make an educated guess that a nonhuman capacity will be a better candidate for inclusion in the kind to the extent that it shares similar biological bases with human empathy. Alternatively, we might be interested in the emergence of empathy in our evolutionary lineage. This would situate the domain of human cognitive evolution: roughly speaking, human cognitive capacities, their evolutionarily precursors, and the selection pressures that gave rise to them. Again, what **empathy** is in this domain is an empirical question, but we can say with some confidence that it will be a historical kind, and that animal capacities will be better candidates for inclusion in the kind to the extent that their bearers are more closely related to humans. Or again, we might be asking with a view to better understanding convergent cognitive evolution: roughly speaking, cognitive capacities and the selection pressures that give rise to them. In this domain, the right way of characterising **empathy** will be the one that affords the best explanations and predictions about the emergence of empathy in a wide range of taxa, and capacities will be better candidates for inclusion in the kind to the extent that they emerged in response to similar evolutionary pressures as human empathy. These will not be only domains in which we ask whether animals have the capacities we do. We might be interested in promoting animal welfare or conservation, explaining or predicting animal behaviour, better understanding animal communication, clarifying the evolutionary or functional relationships between capacities, and so on. Each of these motivations points in the direction of a different domain: a different collection of worldly phenomena constraining our taxonomical decisions. This is not to say that there are as many ways of individuating **empathy** as there are domains to which it is relevant, since some domains may individuate it in the same way. But it is to say that there will be some variation in the individuation of **empathy** across domains – and as a result, there is no canonical, domain-independent way of addressing the question ‘do nonhumans have empathy’.

Domain mismatch is not the only factor underlying disagreement in comparative cognitive science. Even once we have determined what domain our question belongs to and made our educated guesses about what sort of natural kind we might be dealing with there remain many questions about how that kind ought to be individuated, creating further space for disagreement.

For one thing, when we ask a comparative question like ‘Do nonhumans have C?’, it’s often the case that there are gaps in our understanding of **C** even as it manifests in humans. That is, we may not be certain what functional properties characterise **C**, what its underlying mechanisms are, how to investigate its evolutionary history, and even whether **C** *is* a cognitive kind. It might be that some of the functional features we take to be characteristic of C are only present in peripheral cases or in certain specific contexts, and shouldn’t be taken to play any role in individuating **C**. It might be that there is no homeostatic mechanism unifying the functional features we take to be characteristic of C, meaning that **C** is not an HPC kind. Or it might be that C is a recent human innovation rather than an earlier evolved capacity, meaning that **C** may not be a historical kind. This sort of uncertainty might be part of the motivation for looking for the capacity in nonhumans in the first place: if we find it in nonhumans or create it in artificial agents, this may shed new light on its functional profile, underlying mechanisms and/or evolutionary history and assist us in characterising it (or not) as a natural kind. As such, this type of uncertainty is often baked into the foundations of research programmes in comparative cognitive science.

To take an example of a research programme characterised by this kind of uncertainty, consider insight. Insight is often described as the sudden solving of a problem after an impasse (rather than through gradual trial and error), characteristically involving an ‘aha!’ moment. Comparative psychologists since Köhler (1925) have investigated whether animals have insight. Putative examples of the capacity have been found in animals, but critics have argued that these can be explained in terms of ‘low-level’ cognitive mechanisms, including mechanisms of associative learning – the implication being that *truly* insightful problem solving is a more complex affair, perhaps underpinned by a special problem-solving mechanism (Shettleworth, 2012), indicating a view of **insight** as an HPC kind.

But the failure so far to find evidence of this special problem-solving mechanism in animals doesn’t settle whether animals have insight, because there remains uncertainty about what insight is, if it exists at all. There are questions about the mechanisms underpinning insightful problem solving in humans, including whether there *is* a single cognitive process underpinning it. Some have argued that the 'aha!' moment is an affective reaction to the increase in fluency accompanying problem solving (Topolinski & Reber, 2010). The various problems that elicit this 'aha!' feeling might be solved with a range of different problem-solving processes, among which there may be various associative mechanisms. Deflationary explanations of apparently insightful behaviour in animals might reinforce the idea that it is not quite what we thought in humans. For instance, having argued that apparently insightful problem solving in pigeons could be explained in associative terms, Epstein (1985) applied the same approach to explain human behaviour in a classic test of insight. In short, considered as an investigative kind, **insight** begins with the observation that humans sometimes solve a problem suddenly after an impasse. Attempts to detect insight in animals have taken the form of looking for instances of sudden problem solving after an impasse in animals which might 'belong together' with these human instances and form a natural kind. But these investigations, together with investigations into those human instances, have yielded considerable doubt about whether there is anything significant unifying instances of sudden problem solving after an impasse, even in the original human case, and hence whether **insight** is indispensable to scientific success in any cognitive scientific domain.

In light of this, we might revise our views about **insight** in one of three ways. First, we might take an eliminativist approach. On this view, we would say that we used to think insight existed, but now we know it does not – like phlogiston or miasma. This would be a radically revisionary view, but we might be tempted to adopt it if we thought that 'insight' means something like 'special aha!-provoking problem-solving mechanism', and that we are so far from having discovered any instance of that kind of thing that we should eliminate all instances of that kind from our ontology. Second, we might adopt category dissolutionism about **insight**. On this view, we might say that our previous uses of the term 'insight' were picking out real things, but that those things are too dissimilar to form a cognitive kind. This view might be attractive if our interest in studying insight was understanding the mechanisms underpinning problem solving, and if it turned out that the problem-solving mechanisms underlying various putative examples of insight were too diverse to compose a natural kind. In this case, we would eliminate the kind **insight** from our theories, and instead investigate the distinct kinds of problem-solving mechanism we had discovered. Finally, we might take a revisionary realist approach to insight. On this view, we might say that insight *does* exist, but that it isn't what we thought it was. This view might be attractive if our interest in studying insight was in understanding the 'aha!' feeling which sometimes accompanies problem solving. In this case, we'd say that rather than being a special problem-solving mechanism, insight is an affective *response* to problem solving.[10] On this view, we'd retain token instances of insight in our ontology and retain the kind **insight** in our theories, but would adjust our account of what both insight and **insight** really are.

Since each of these approaches has its attractions, reasonable people might disagree about which route to take. Our verdicts on whether animals have insight turn on this choice. For an eliminativist or category dissolutionist, the question of whether animals have insight should be abandoned. In the eliminativist case, this is because we have determined that insight does not exist. In the dissolutionist case, this is because we have determined that **insight** is not a natural kind – classifying things as insight is not necessary for successful predictions or explanations in cognitive science. In either case, we might continue to study the mechanisms underlying problem solving in animals, but we will frame these investigations another way. A revisionary realist

[10] For more on the distinction between these three approaches, see (Ramsey, 2021).

would probably conclude that the evidence doesn't settle whether animals have insight, since our inquiries have been directed at detecting a special kind of problem solving, and not an affective response. To move forward, we would need to rethink our methods in light of our revised understanding of the phenomenon. Note, though, that on none of these approaches will it turn out that it was a mistake to ask whether animals had insight to begin with. Asking this question enabled us to investigate two hypotheses simultaneously – that **insight** is a cognitive kind, and that it has nonhuman instances – and yielded significant results. As such, we have been productively engaged in what David Colaço (2022, p. 98) calls 'kinding in progress' – a process wherein we group phenomena together whilst their nature is unknown, with a view to discovering which, if any, properties they share.

Whenever there is such uncertainty surrounding a phenomenon in comparative cognitive science, one can naturally expect some disagreement to arise. Looking for instances of a putative cognitive kind in nonhumans requires us to start with some account of the kind. As I've suggested, at this stage we will be dealing with an investigative kind: most likely, a phenomenon observed in humans, which we think may belong together with some nonhuman cognitive phenomena. Our characterisation of this investigative kind is necessarily provisional - a 'definition as hypothesis' (Colaço, 2022), where the hypothesis is that there is a cognitive kind with such-and-such properties which has both human and nonhuman instances. When evidence conflicts with this hypothesis, it is underdetermined whether the fault lies: that is, whether animals lack the capacity, or our provisional characterisation of the capacity was wrong. This epistemic situation is common in comparative cognitive science. As Kristin Andrews (2014, p. 15) puts it, 'since we are simultaneously investigating the nature of these phenomena and the nature of animal minds, there is no direct application of some well-established theory to these questions.' This leaves the way open for disagreements of the sort I've described, in which uncertainty about the nature of the capacity generates disagreements about whether animals have it, even where the domain of inquiry is clear.

Finally, even where we are reasonably confident about the domain our question belongs to, as well as about the nature of the relevant capacity, we ought nevertheless to anticipate disagreement when we ask whether a given nonhuman has a capacity of the same kind, because the boundaries of natural kinds are often unclear. Consider a case in which we are looking for a cognitive HPC kind. Whether a nonhuman capacity belongs to the kind will turn on whether it has the right functional properties, underpinned by the right mechanisms. But whether this is so in a given case may be very difficult to judge. It's one of the characteristic features of HPC kinds that their members can differ from one another in various ways. There are many differences between individual polar bears, and between my episodic memory capacity and yours. In some cases, it will be difficult to judge whether a candidate instance, which shares *some* properties characteristic of the kind to *some* degree is a member of the kind (Taylor, 2020).

We might hope to appeal to mechanisms to settle the question: if the individual's possession of the relevant properties is underpinned by the right kind of mechanism, then it is a member of the kind. But whether the mechanisms underpinning a capacity in humans and nonhumans are the same is itself a tricky question. As Carl Craver (2009) argues, there are different ways of characterising the mechanisms underpinning a cognitive kind. For example, we might characterise the mechanisms underpinning a kind like **episodic memory** in detailed neurobiological, mechanistic terms, or at a more abstract level in information processing terms (Boyle, 2022, pp. 197–198). A more abstract, high-level characterisation might allow some nonhuman mechanisms to 'count' as the same as the mechanisms underpinning episodic memory in humans which would not 'count' given a more concrete neurobiological characterisation. Again, the 'correct' way of characterising the underlying mechanism is likely to turn on the domain: at what level of organisation and in what level of detail one ought to specify these mechanisms turns on what one is trying to predict or explain. When it is unknown which way of individuating the underlying mechanisms is most productive, reasonable people might disagree.

Even where the mechanisms are relatively clearly individuated, there may still be fuzzy, liminal cases in which it is unclear, perhaps indeterminate, whether a nonhuman capacity is a member of a cognitive kind, creating more space for disagreement. In some cases, one might judge that 'neither theoretical nor methodological considerations assign the object being classified determinately to the kind or to its complement', and this indeterminacy 'could not be remedied without rendering the definitions unnatural in the sense of being scientifically misleading' (Boyd, 1991, p. 142). In others, it might be that theoretical considerations do motivate resolving the question one way or the other, or that empirical considerations can be brought to bear on whether one or the other resolution promotes success in this domain. But, again, these are things about which reasonable people may disagree.

5 | MAKING PROGRESS

I've argued that situating comparative questions in the framework of natural kinds reveals several sources of disagreement in comparative cognitive science. First, comparative questions give rise to domain mismatch. Comparative cognitive science is home to multiple scientific domains, where scientific success requires classifying things into different natural kinds. So, scientists may disagree on the individuation of a kind because they are operating in distinct domains, or because they disagree about what domain they are operating in, and what success in that domain demands. Second, comparative cognitive science involves investigating capacities whose existence, natural kind status and nature are uncertain even in the human case. Indeed, this uncertainty is often part of the motivation for investigating these capacities in animals. This creates further scope for disagreement about whether the kinds under discussion are real, and how they should be individuated. Finally, even where domain mismatch does not arise and we are relatively confident in the existence and individuation of the capacity we're looking for in the relevant domain, there may still be reasonable disagreements about whether a candidate capacity is an instance of the kind. This is because cognitive kinds are not tidy essentialist kinds but cluster kinds of one sort or another, which do not have sharp boundaries, and which may admit of ineliminably liminal cases.

It should now, I hope, be clear why neither of the tempting responses mentioned in §2 provides an adequate response to the pattern of disagreement so frequently illustrated in comparative cognitive science.

First, replacing top-down questions with bottom-up questions focussed on mechanisms is unlikely to foreclose the kind of disagreement under discussion. Recall that a bottom-up approach involves disarticulating a capacity into its component mechanisms and investigating which of those mechanisms is present in nonhumans. This is problematic for three reasons. First, disarticulating the capacity into its component mechanisms requires first having an account of what the capacity *is*, a question which is at the heart of many disagreements surrounding top-down questions – and which at least some of these research programmes are setting out to answer. Second, there is unlikely to be a canonical, domain-neutral disarticulation of a given capacity into its component mechanisms. Third, determining whether the component mechanisms are present in other species requires making judgments about the individuation of mechanisms across species, which are likely to be as complicated as judgments about the interspecific individuation of capacities.

Second, to dismiss these disputes as merely verbal oversimplifies matters – and to the extent that they are verbal, it would often be a mistake to abandon them or eliminate the terms involved. Recall our imagined geneticist and ecologist disagreeing over whether the ABC island bear is a polar bear. Their dispute is verbal, in a sense. The geneticist uses 'polar bear' to pick out **polar bear** as it figures in genetics, and the ecologist to pick out **polar bear** as it figures in ecology. Clarifying this eliminates the dispute. But it also illuminates matters in a way that simply eliminating the term 'polar bear' would not, by explaining how this disagreement arose: these two natural kinds, with their overlapping extensions, were discovered in complementary attempts to understand the polar bear phenomenon, and to explain and vindicate the postulation of **polar bear** as an investigative kind. As such, it is probably productive to retain the term 'polar bear' as a label for both. Moreover, clarifying this might reveal that their disagreement, such as it is, is not a disagreement about the meaning of the word 'polar bear', so much as it is a disagreement (or, at least, a lack of convergence) about what we are up to when we classify things as polar bears. Given that comparative cognitive science is home to multiple domains which individuate kinds differently, it seems highly likely that some disputes in comparative cognitive science will be like this – but not all. And whilst applying the elimination method in comparative cognitive science might be a good way to sniff out verbal disputes, it might not entirely illuminate what is going on, or provide much in the way of resolution.

Instead, when faced with a disagreement of the kind discussed here, I suggest that a useful first step is to begin asking questions about the domain of inquiry. Suppose we find that two researchers investigating whether nonhumans have a capacity C use the term 'C' in different ways, such that they are inclined to offer different answers to that question. We should understand this as a question about the natural kind **C**, which may be individuated differently in different domains. So, our first question ought to be: what domain are these scientists are operating within? We can get a clue about this by asking after their motivations: by finding out what they hope to achieve by asking this question, we can determine which collection of objects and phenomena constrain their answers.

It may be that neither scientist has a particular domain clearly in view. They simply wish to know whether nonhumans have C, *simpliciter*. On the view I've proposed, this is something like a category mistake. There is no such thing as C *simpliciter*, since our capacities fall into to multiple, overlapping natural kinds belonging to different scientific domains. In each of those domains it may be a substantive, open question what the relevant natural kinds are, but the domain of inquiry significantly constrains the answer to this question. Without a domain in view, it is at best indeterminate what 'C' refers to. In such cases, it may be difficult to determine whether there is a substantive disagreement lurking, or whether the disagreement is merely verbal – but it is certainly unlikely to be productive. Without more to constrain uses of the term 'C', the situation looks like a terminological free-for-all and is not clearly worth arbitrating. But the remedy is not (or not yet) to abandon either the term 'C' or the question whether nonhumans have C. Rather, it is to situate the question within a scientific domain.

Alternatively, our investigation might reveal that these researchers are approaching the question with different domains in view. In this case, they do not necessarily pick out the same natural kind by 'C': each means to pick out the kind **C** as it figures in their domain. These domains may individuate **C** in the same way, but they may not. The different domains impose substantial constraints on their respective uses of the term 'C', so this is not a terminological free-for-all. But it is possible that they are asking and answering quite different questions. If so, any dispute between them will be verbal – but this should be easy to spot, once the domains are properly specified. However, this would not yet show that the term 'C' should be eliminated from the lexicon of researchers working in either or both domains. By clarifying matters in this way, we might also have revealed that these two kinds are importantly connected: it may be that they have significantly overlapping extensions, because they have emerged from the attempts of two domains to vindicate and explain the very same investigative kind. In this case, the different uses of 'C' in these different domains will be a case of 'revealed polysemy' (Brusse, 2016, p. 105), rather than lexical ambiguity. Eliminating the term 'C' might obscure the complementary roles of these domains and their respective natural kinds in explaining a phenomenon of common interest. On the other hand, it might turn out that this sort of relationship between the two kinds does not exist, in which case the argument for calling this a wholly verbal dispute and eliminating the term 'C' is stronger.

Finally, it might turn out that both researchers are operating within the same domain, and both aim to pick out the very same natural kind by using the term 'C'. This shared aim imposes substantial constraints on their uses of the term. They are likely to be having a substantive dispute about the nature of the relevant natural kind – perhaps about the cluster of functional properties by which it's characterised, its underlying mechanisms or its selective history. In many cases, at least one of them will be wrong and their disagreement will be resolvable given the right empirical work. In other cases, there may be no empirical work which could resolve the disagreement – for instance, in cases where the disagreement is over a candidate instance which is an ineliminably borderline case. Here, their disagreement might be resolvable given the right theoretical work, but reasonable disagreement might persist. And there is another, perhaps more interesting case of within-domain disagreement, suggested by the example of **insight** above: it may be that they disagree on whether we ought to take a revisionary realist, kind dissolutionist or eliminativist approach to the category under discussion. Here, again, it might take considerable empirical and theoretical work to resolve their dispute.

6 | CONCLUSION

In comparative cognitive science, we often ask whether nonhumans have the capacities we do. These questions can give rise to disagreement, in which one party affirms and the other denies that nonhumans have a given capacity on the basis of the same evidence. These disagreements can quickly become unproductive. So, it is unsurprising that some researchers have called for a move away from questions of this kind. Against this, I've argued that we can make progress with these questions by construing them as questions about natural kinds. Doing so allows us to see that these are substantive empirical questions worth asking: they are attempts to discern real causal-explanatory structure in the world. It also demonstrates that we should expect disagreement to attend on these questions in virtue of both the multi-domain structure of comparative cognitive science and the nature of natural kinds.

The kinds of human cognitive science often provide a starting point for comparative cognitive science. But the natural kinds of one domain are not the natural kinds of another. So, comparative cognitive scientists can't simply lift cognitive kinds from human cognitive science and investigate whether they have non-human instances. Instead, progress with comparative cognitive science requires that we treat human capacities as belonging to investigative kinds: potentially more inclusive kinds into which as yet undiscovered nonhuman capacities might fall. This triggers a search for a correspondingly inclusive natural kind, whose individuation will be sensitive to the inductive and explanatory demands of this broader scientific domain. This search is not a straightforward matter, since comparative cognitive science is home to multiple domains, within which explanatory and inductive success may constrain cognitive taxonomy in different ways. This is not to say that anything goes: since natural kinds are the kinds that are indispensable to scientific success in a domain, plausibly there will be at most one way of characterising **C** such that it is a natural kind for a given domain – meaning that within a domain it should be possible to offer a fairly unequivocal verdict on whether nonhumans have C, borderline cases notwithstanding. But since success in different domains makes different demands, researchers in different domains may rightly offer different answers to questions like 'Do nonhumans have C?'. For this reason, it would be a mistake to expect domain invariant 'all or nothing' answers to questions about nonhumans' capacities. But within a suitably specified domain it is not a mistake to ask them.

As I've noted, one might be tempted to say that different domains should use different terminology to refer to the different cognitive kinds with which they're concerned. By using the same terminology, in addition to risking merely verbal disputes, they are at risk of miscommunication: a scientist working in one domain might draw inappropriate conclusions from research using the same terminology in a different domain. This is a natural thought, but a proliferation of terminology comes with its own challenges and risks. For instance, novel terminology may be difficult to adopt clearly and consistently, may not be as eye-catching or legible, and there may be disagreement about who gets to 'keep' the familiar terms. Moreover, as I've argued, in many cases eliminating the term risks obscuring the way in which the natural kinds of different domains are related. Where the natural kinds of different domains share a label, this may reflect polysemy, rather than ambiguity: cases in which distinct natural kinds emerge from the attempts of different domains to understand a common investigative kind. How to balance such challenges and risks is a judgment call (see Taylor & Vickers, 2017). But in any case, this terminological question is orthogonal to the more substantive suggestion made here: that comparative cognitive science is home to multiple domains of inquiry, and that our questions about nonhuman capacities, understood as putative cognitive kinds, can productively be addressed only with one of these domains in view.

ACKNOWLEDGEMENTS

I'd like to thank audiences at University College London, Sussex, Gottingen, Cambridge, the Institute of Philosophy and the London School of Economics, where early versions of this work were presented, for their insightful comments and feedback. Thanks also to Marta Halina, Henry Taylor, Liam Kofi Bright, Nick Shea, Mike Dacey, and two anonymous reviewers for detailed comments on earlier drafts which significantly improved the paper. Work on this paper has been supported by a UKRI Future Leaders Fellowship, Grant Ref: MR/W00741X/1. There are no data associated with this article.

ORCID

Alexandria Boyle https://orcid.org/0000-0001-8827-5479

REFERENCES

Akagi, M. (2018). Rethinking the problem of cognition. *Synthese*, *195*(8), 3547–3570. https://doi.org/10.1007/S11229-017-1383-2

Andrews, K. (2014). *The Animal Mind: An Introduction to the Philosophy of Animal Cognition*. Routledge.

Bartal, I. B.-A., Decety, J., & Mason, P. (2011). Empathy and pro-social behaviour in rats. *Science*, *334*, 1427–1430. https://doi.org/10.1126/science.1210789

Boyd, R. (1989). What realism implies and what it does not. *Dialectica*, *43*(1–2), 5–29. https://doi.org/10.1111/j.1746-8361.1989.tb00928.x

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, *61*(1/2), 127–148. https://doi.org/10.1007/BF00385837

Boyle, A. (2022). Episodic memory in animals: optimism, kind scepticism and pluralism. In A. Sant'Anna, C. J. McCarroll, & K. Michaelian (Eds.), *Current Controversies in Philosophy of Memory* (pp. 189–205). Routledge.

Brigandt, I. (2003). Species pluralism does not imply species eliminativism. *Philosophy of Science*, *70*(5), 1305–1316. https://doi.org/10.1086/377409/0

Brusse, C. (2016). Planets, pluralism, and conceptual lineage. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *53*, 93–106. https://doi.org/10.1016/J.SHPSB.2015.11.002

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind and Language*, *23*(1), 58–89. https://doi.org/10.1111/j.1468-0017.2007.00329.x

Chalmers, D. J. (2011). Verbal disputes. *Philosophical Review*, *120*(4), 515–566. https://doi.org/10.1215/00318108-1334478

Colaço, D. (2022). What counts as a memory? Definitions, hypotheses, and "kinding in progress." *Philosophy of Science*, *89*(1), 89–106. https://doi.org/10.1017/PSA.2021.14

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575–594. https://doi.org/10.1080/09515080903238930

Currie, A. (2016). The mystery of the triceratops's mother: How to be a realist about the species category. *Erkenntnis*, *81*(4), 795–816. https://doi.org/10.1007/s10670-015-9769-3

de Waal, F. B. M., & Ferrari, P. F. (2010). Towards a bottom-up perspective on animal and human cognition. *Trends in Cognitive Sciences*, *14*(5), 201–207. https://doi.org/10.1016/j.tics.2010.03.003

Dixon, B. (2019, May 16). *Why AI appears to create things*. https://mindmatters.ai/2019/05/why-ai-appears-to-create-things/

du Sautoy, M. (2019, May 11). Can AI ever be truly creative? *New Scientist*, 38–41.

Eaton, T., Hutton, R., Leete, J., Lieb, J., Robeson, A., & Vonk, J. (2018). Bottoms-up! Rejecting top-down human-centered approaches in comparative psychology. *International Journal of Comparative Psychology*, *31*, 1–19. https://doi.org/10.46867/ijcp.2018.31.01.11

Ellis, B. (2008). Essentialism and natural kinds. In M. Curd & S. Psillos (Eds.), *The Routledge Companion to Philosophy of Science* (pp. 139–148). Routledge. https://doi.org/10.4324/9780203000502-21

Epstein, R. (1985). Animal cognition as the praxist views it. *Neuroscience and Biobehavioral Reviews*, *9*(4), 623–630. https://doi.org/10.1016/0149-7634(85)90009-0

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117

Franklin-Hall, L. R. (2015). Natural kinds as categorical bottlenecks. *Philosophical Studies*, *172*(4), 925–948. https://doi.org/10.1007/S11098-014-0326-8/METRICS

Godman, M. (2014). Scientific Enquiry and Natural Kinds: From Planets to Mallards. *International Studies in the Philosophy of Science*, *28*(3), 343–346. https://doi.org/10.1080/02698595.2014.953346

Halina, M. (2021). Insightful artificial intelligence. *Mind and Language*, *36*(2), 315–329. https://doi.org/10.1111/mila.12321

Hendrycks, D., & Gimpel, K. (2017). Early methods for detecting adversarial images. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 1–9. https://doi.org/10.48550/arXiv.1608.00530

Heyes, C. M. (2015). Animal mindreading: what's the problem? *Psychonomic Bulletin & Review*, *22*(2), 313–327. https://doi.org/10.3758/s13423-014-0704-4

Kendig, C., & Grey, J. (2021). Can the epistemic value of natural kinds be explained independently of their metaphysics? *The British Journal for the Philosophy of Science*, *72*(2), 359–376. https://doi.org/10.1093/bjps/axz004

Köhler, W. (1925). *The Mentality of Apes*. Harcourt, Brace.

Kohs, G. (2017, September 29). *AlphaGo*. Submarine. www.alphagomovie.com

Lemeire, O. (2020). No purely epistemic theory can account for the naturalness of kinds. *Synthese*, *198*(12), 2907–2925. https://doi.org/10.1007/s11229-018-1806-8

Ludwig, D. (2018). Letting go of "natural kind": Toward a multidimensional framework of nonarbitrary classification. *Philosophy of Science*, *85*(1), 31–52. https://doi.org/10.1086/694835

Magnus, P. D. (2012). *Scientific Enquiry and Natural Kinds*. Palgrave Macmillan.

Magnus, P. D. (2014). NK≠HPC. *The Philosophical Quarterly*, *64*(256), 471–477. https://doi.org/10.1093/pq/pqu010

Magnus, P. D. (2018). How to be a realist about natural kinds. *Disputatio (Spain)*, *7*(8). https://doi.org/10.5281/ZENODO.2553734

Magnus, P. D. (2023). Scurvy and the ontology of natural kinds. *Philosophy of Science*, 1–17. https://doi.org/10.1017/PSA.2023.19

Ramsey, W. (2017). Must cognition be representational? *Synthese*, *194*(11), 4197–4214. https://doi.org/10.1007/S11229-014-0644-6

Ramsey, W. (2021). What eliminative materialism isn't. *Synthese*, *199*(3–4), 11707–11728. https://doi.org/10.1007/S11229-021-03309-Y

Shettleworth, S. J. (2007). Studying mental states is not a research program for comparative cognition. *Behavioral and Brain Sciences*, *30*(3), 332–333. https://doi.org/10.1017/S0140525X0700218X

Shettleworth, S. J. (2012). Do animals have insight, and what is insight anyway? *Canadian Journal of Experimental Psychology*, *66*(4), 217–226. https://doi.org/10.1037/a0030674

Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, *1*(4), 165–167. https://doi.org/10.1038/s42256-019-0039-y

Slater, M. H. (2015). Natural kindness. *British Journal for the Philosophy of Science*, *66*(2), 375–411. https://doi.org/10.1093/bjps/axt033

Smith, J. D. (2009). The study of animal metacognition. *Trends in Cognitive Sciences*, *13*(9), 389–396. https://doi.org/10.1016/j.tics.2009.06.009

Suddendorf, T., Addis, D. R., & Corballis, M. C. (2009). Mental time travel and the shaping of the human mind. *Experimental Brain Research*, *192*(3), 1317–1324. https://doi.org/10.1007/s00221-008-1491-9

Tahko, T. E. (2021). *Unity of Science*. Cambridge University Press. https://doi.org/10.1017/9781108581417

Taylor, H. (2019). Whales, fish and Alaskan bears: interest-relative taxonomy and kind pluralism in biology. *Synthese*, https://doi.org/10.1007/s11229-019-02284-9

Taylor, H. (2020). Emotions, concepts and the indeterminacy of natural kinds. *Synthese*, *197*(5), 2073–2093. https://doi.org/10.1007/s11229-018-1783-y

Taylor, H., & Vickers, P. (2017). Conceptual fragmentation and the rise of eliminativism. *European Journal for Philosophy of Science*, https://doi.org/10.1007/s13194-016-0136-2

Topolinski, S., & Reber, R. (2010). Gaining insight into the “Aha” experience. *Current Directions in Psychological Science*, *19*(6). https://doi.org/10.1177/0963721410388803

Vasconcelos, M., Hollis, K., Nowbahari, E., & Kacelnik, A. (2012). Pro-sociality without empathy. *Biology Letters*, *8*(6), 910–912. https://doi.org/10.1098/rsbl.2012.0554

Weinberg, J., Zimmermann, A., Chalmers, D., Askell, A., Montemayor, C., Khoo, J., Rini, R., Nguyen, C. T., Shevlin, H., & Vallor, S. (2020, July 30). *Philosophers on GPT-3 (updated with replies by GPT-3)*. https://dailynous.com/2020/07/30/philosophers-gpt-3/

Wittkuhn, L., Chien, S., Hall-McMaster, S., & Schuck, N. W. (2021). Replay in minds and machines. *Neuroscience and Biobehavioral Reviews*, *129*, 367–388. https://doi.org/10.1016/j.neubiorev.2021.08.002

How to cite this article: Boyle, A. (2023). Disagreement & classification in comparative cognitive science. *Noûs*, 1–23. https://doi.org/10.1111/nous.12480