Author-final pre-print, for *Mind & Language*
Please cite published version

Concept-Metacognition

Abstract

Concepts are our tools for thinking. They enable us to engage in explicit reasoning about things in the world. Like physical tools, they can be more or less good, given the ways we use them – more or less dependable for categorisation, learning, induction, action-planning, and so on. Do concept users appreciate, explicitly or implicitly, that concepts vary in dependability? Do they feel that some concepts are in some way defective? If so, we metacognize our concepts. One example that has been studied is a person's judgement about how well they have learnt a new category. There are many other forms that concept-metacognition could take. This paper offers a preliminary taxonomy of different forms of metacognition directed at concepts. It suggests that concept-metacognition may affect the way one concept from a range of candidates is selected for use, and the way a concept is relied on in reasoning. Concept-metacognition may also play a pivotal role in the social process of constructing concepts, in replacing the old and constructing the new tools for thinking.

(1) Introduction

Surveying current psychological research for areas of excitement, metacognition would come high up the list. Concepts have had their own time in the limelight but concept research is now in a more steady phase. The metacognitive standpoint, which has been so fruitfully applied to memory, is increasingly being extended to other cognitive processes, but little so far to concepts. Yet concepts remain one of our core psychological capacities, organising much of our knowledge and underpinning the remarkable power of human reasoning. It will surely be useful, then, to study the metacognitive aspects of concepts. But what could metacognition applied to concepts be? How would it operate?

We use concepts to store and organise our knowledge. Encountering one member of a category we can learn what that thing is like and what to do with it. Concepts allow us to carry that knowledge forward, shaping our expectations about future encounters with things of the same kind. Concepts can be more or less suited to doing that job. Indeed, a concept can be more or less good-for-purpose with respect to every purpose for which we rely on it.

We reflect on this when we have explicit debates about which concepts to use and which to abandon. The concept INNATENESS has been much-debated in philosophy and cognitive science, with many researchers arguing that it should be abandoned because it confuses more than it elucidates (Griffiths 2002, Mameli 2008). Such reflection is also common outside academia. We debate whether we ought to use racial concepts—debates that are only partly to do with whether these concepts correspond to any real categories. The question is whether, given the ways they are used, we ought to rely on racial concepts in our thinking, revise them, or abandon them altogether (Haslanger 2000, Mallon 2006). These are explicit metacognitive judgements about concepts.

The aim of this paper is to initiate a debate about how metacognition applies to concepts: what forms it takes and whether it is important. Are the judgements about concepts just mentioned a special feature of theoretical discourse or can everyone answer metacognitive questions about their concepts? Presuming that they can, a set of experimental questions immediately arises about how stable and systematic these judgements are, and what they are based on. A theoretical question then follows about what these judgements are supposed to track, opening up a further empirical question about how reliable they are. The aim of this paper is to set out a tentative agenda for philosophical and empirical research into these questions.

At this early stage of the investigation a key task is to define terms and point to potentially interesting phenomena. If we follow standard ways of characterising a concept as a psychological structure in an individual thinker (section 2), each such psychological structure may be defective or dependable. Section 3 formulates more precisely the hypothesis that some kind of metacognitive assessment of such facts is at work in our mental life – that there is metacognition about concepts. Section 4

summarises the small store of existing research on the question and section 5 offers a preliminary taxonomy of the forms that concept-metacognition might take. Section 6 suggests ways of investigating the role of concept-metacognition in the way thinkers deploy and rely upon concepts. Section 7 argues that an understanding of concept-metacognition may have interesting wider applications.

(2) Concepts; Metacognition

What are concepts, such that they could be the target of one or more kinds of metacognition? I take concepts to be mental representations that are exercised in cognitive processes like categorising, reasoning and recalling a semantic memory. They are unsaturated representations, representing things like individuals and properties, and combining to form saturated, truth-evaluable representations. Concepts are most obviously at work when engaged in conscious deliberate thought. When I'm planning for the future I use concepts like CAREER and FAMILY; when I'm planning for dinner it's more mundane concepts like TOMATOES and PASTA that figure in my stream of thought. Concepts may also figure in stored memories and/or standing beliefs, but our focus will be on concepts as they are found in occurrent thoughts.

Concepts are used for categorisation—when we apply a concept to an item—but not all acts of categorisation involve concepts.[1] Phonemes are perceived categorically, carving the continuous variations in the incoming stream of sound into more coarse-grained classes with somewhat sharp boundaries between them. There are pairs of sounds near the border where acoustic similarity between pairs in different categories is greater than the similarity between other pairs that fall into the same category. The representations responsible for this kind of categorical perception will not be concepts unless they also figure as recombinable constituents of thought, which typically they do not. Many people do of course have concepts of phonemes, but applying a concept to auditory experience is something additional to categorical perception of the phoneme.

We can distinguish between object-level and meta-level contents. Object-level contents are about aspects of the non-psychological world. Meta-level contents are about a thinker's psychological states and processes. Cognitive psychology has mostly studied concepts with object-level contents (DOG, CHAIR, CHESS). The distinction between the object level and the meta level can also be applied to psychological processes. Many are directed at the non-psychological world, but some are directed at other psychological processes, both monitoring how they unfold and controlling their operation (Nelson and Narens 1990). These are metacognitive.

The majority of metacognition research to date has concerned memory. How do people judge how well they have learnt the material that they have been asked to

[1] Nor are all concepts categorical, e.g. LENGTH. Non-categorical quantitative concepts are particularly prominent in scientific theorising.

memorise? For instance, how do students determine whether they have learned some material well enough to take an exam? During recall, you may be unable to remember a name but think that you would recognise it in a forced-choice test. How do people make these judgements and how accurate are they? Confidence has also been a major focus of metacognition research. This extends beyond memory to processes including perception and action execution. For any judgement or decision a person makes, we can ask how confident they are about it. Confidence in perceptual decision-making has been the particular focus of a flourishing research programme in recent years.

To date there has been almost no research into metacognition about concepts. The question is not whether concepts can have meta-level contents. Obviously they can. The concept BELIEF – the topic of such fruitful work in developmental psychology – has meta-level content. The question is whether there is a species of metacognition that is concerned with or directed at concepts as such.

Superficially it might seem that metacognition of perceptual decision-making would qualify. Psychologists studying perceptual decision-making typically ask subjects to categorise stimuli. Are the dots moving mainly to the right or to the left? They then elicit a confidence judgement about that decision. That is not, however, a judgement about a concept. It is about a particular occasion—how likely is it that I correctly categorised this stimulus?—rather than about the usefulness of the perceptual category in general, or the reliability of the subject's disposition to categorise by motion direction. The fruitfulness of studying metacognition of perceptual decisions is an inspiration but it doesn't offer an existing example of metacognition being applied to concepts.

Concepts are deployed when we categorise things in deciding how to act in the moment, and categorisation for current action has been the focus of much psychological research, however future-directed uses of concepts are equally important. Learning and reasoning are future-directed in the relevant way. When I observe something new about tomatoes, storing that information with my TOMATO concept will allow me to make use of it on a future occasion. When I'm planning what to do, I go through a chain of reasoning using concepts like CAREER and FAMILY to arrive at a future-directed plan of action.

A central example is category-based induction. I observe a hard shiny object of a certain size and shape and categorise it under my HAZELNUT concept. I then pick it up and discover it rattles when shaken. I crack it and discover how it tastes when eaten. These two new pieces of information are stored with my HAZELNUT concept. When I encounter another object of this kind my concept now tells me something I can do with this object and gives me an expectation of how it will taste if I eat it. Those two inductive inferences are underpinned by the nature of the category, *hazelnut*. They would have been much less reliable if I had categorised the initial object under my PHYSICAL OBJECT concept and made the same induction. Then I would have projected the expectation about rattling and taste far too widely. These properties do not carry over

to physical objects in general. For this kind of category-based induction, HAZELNUT is a more dependable concept than PHYSICAL OBJECT.

So concepts are relied on in category-based induction and can be more or less dependable for that purpose. Do concept users register those differences in some way, even if only approximately? We can ask the same question for other ways of using concepts, for other ways in which we rely on them. Much of our concept-based thinking occurs when we are communicating with others. That is one of the main ways we learn, so as to encode new information with our concepts. Here too concepts generate of expectations: expectations of what people will accept and reject, and of what they will say next. Concepts thereby form the basis of fluid communication: of understanding others and making oneself understood. Concepts can be more or less suited to this purpose as well. When a thinker has encoded little information in a concept (e.g. through ignorance), or when they have encoded very idiosyncratic information, that concept is likely to be an unreliable tool on which to base communication. This is another important dimension of variation—of fitness for purpose—that may be registered in some way by concept users.

(3) The Hypothesis

Research on 'metacognition' covers a wide range. We should not presuppose from the outset that it corresponds to a single underlying cognitive process or function (in the way visual working memory, say, is a single cognitive process). To cover the range of phenomena psychologists and philosophers using the term have been interested in, it will be useful to work with an inclusive definition of metacognition as:

> 'the set of capacities through which an operating subsystem is evaluated or represented by another subsystem in a context-sensitive way' (Proust 2013, p. 4)

'Thinking about thinking' is sometimes used as shorthand. However, the targets of metacognition are wider: not just thoughts, but also perception, action and other cognitive processes. Nor need the metacognitive state itself be a thought. It could be a feeling or other experience, or a non-conceptual representation of some other kind.

Applying explicit mental concepts like belief and desire to oneself is an obvious example of metacognition. So are self-applications of other psychological concepts, for example thinking that *I can see p clearly*, or *I am uncertain about q*. Not all cases need be so explicit, however. The definition above has the merit of extending to cases of what Proust has called 'procedural' metacognition, for instance a feeling of knowing or a feeling that an answer is on the 'tip of the tongue'. Such feelings may play a role when subjects regulate their cognitive processes, deciding what to think about or when to give up, without the thinker having to exercise or even possess epistemic concepts like KNOWING and LEARNING. When a fact is recalled quickly, the experience of fluent recall can increase the thinker's confidence in the resulting judgement.

Procedural metacognition can be based on experiences that are not affective and so are not readily classified as feelings, a sense of dependability for example (cp. sense of agency). It could also be based on nonconceptual representations that play a role in regulating psychological processing without themselves being experienced. Procedural metacognition is contrasted with analytic metacognition, where people draw on background beliefs or mini theories to form metacognitive judgements. The belief that I will remember a text better if I've read it twice is a piece of analytic metacognition.

The definition of metacognition above readily extends to concepts. Concept use is a potential target for evaluation and/or representation by another subsystem, and the metacognition involved could be analytic or procedural.

A form of metacognition that has been studied extensively is confidence—the certainty or uncertainty associated with a decision (or with a belief, percept, etc.). Certainty or uncertainty affects the way an item of information is relied on in cognitive processes, for example: when weighing and integrating different sources of perceptual information about the size of an object (Ernst and Banks 2002), when adjusting expectations on the basis of feedback in the course of reinforcement learning (Meyniel and Dehaene 2017), and when different agents pool information in order to take a joint decision (Bahrami, et al. 2010, Shea, et al. 2014). Certainty measures are also likely to be involved when model-based and model-free reinforcement learning systems compete for control of an action (Lee, et al. 2014, Donoso, et al. 2014). Certainty measures can usefully play these computational roles when they at least roughly reflect the reliability of a representation, that is, the probability that it is correct. Models of perceptual decision-making suggest that certainty measures are associated both with percepts (e.g. a visual experience of an array of contrast gratings) and with subsequent perceptual decisions taken on the basis of those percepts (e.g. the decision that the second-presented array contains an odd-one-out) (Fleming, et al. 2010).[2]

The definition of metacognition is wide enough that it potentially extends to quite low-level processes. The weighting of haptic against visual information described by Ernst & Banks probably takes place outside the subject's awareness. Thus, the kind of metacognition that does not require possession of psychological concepts is unlikely to be distinctively human, and could in fact extend to some psychological processes that are widely shared in the animal kingdom (Shea 2014). Nevertheless, paradigmatic cases even of procedural metacognition do involve consciousness. They are cases where conscious thoughts or feelings are playing the regulative role. Similarly, paradigmatic cases of concept-metacognition will be thoughts or feelings of which the subject is aware and which play a role in controlling voluntary action.

[2] (Pouget, et al. 2016) suggest that the term 'confidence' should be reserved for representing the probability that a decision is correct, with the term 'certainty' playing a more inclusive role so as to cover all other probability distributions over sensory and cognitive variables (such as the percept in this case).

The hypothesis I want to consider is that metacognition of some kind is directed at concepts.

Hypothesis
There is metacognition directed at concepts

It is in fact easy to find, in theoretical or reflective discourse, judgements about certain concepts being reliable or, more often, defective. I already mentioned the debates about innateness. These are explicitly conceived by many as debates about whether INNATENESS is a good concept to use (Mameli and Bateson 2006). Similarly, some have argued that our everyday concept of consciousness is irredeemably confused and ought to be abandoned (Churchland 1983). In psychology similar debates have led to the demise of concepts like EGO and ID in scientific theorising; current debates about REPRESSED MEMORY will determine its fate. A famous example from the history of science is the way the concepts of space and time were replaced by the concept SPACE-TIME. How stable are these judgements and are they based on some kind of systematic process that monitors concepts?

Some of these judgements are at the object level. The discovery that there is no phlogiston hastens the abandonment of the concept (recently somewhat revived as a philosophical example). The observation that there is no such thing as a witch is taken to be an argument against relying on the concept. Questions about whether X is a natural kind are at the object level, as are questions about essences. Answers are still likely to have global effects on how we use the concept of X. Although they could remain at the object level, participants in these debates often explicitly reflect on the goodness of the concept. That is especially useful if we want to examine whether we ought to be using a concept without presupposing that it refers, or by raising concerns that do not just turn on the nature of the property referred to (e.g. whether it has an essence or not).

Often people's reservations about a concept are broadly epistemic: that it doesn't succeed in picking out anything in the world or that the conceptions associated with the concept are seriously confused. However, concerns also extend to the normative. Many think we ought to stop using racial concepts irrespective of whether, despite the common misconceptions encoded in the concepts for instance about essence, they do in fact succeed in referring (Mallon 2006). A contemporary argument about gender concerns how we ought go about conceiving of gender (Godman 2018, Haslanger 2000). That is a claim at the meta-level, about which concepts we should use. Although my main focus here is on the reliability or dependability of concepts, normatively-driven metacognition about concepts is also very interesting, and it is not known whether ordinary thinkers treat these concerns differently. The moral opprobrium associated with using a racial concept may contribute to thinkers having a general sense that the concept is undependable or to-be-avoided, without their distinguishing between epistemic and normative defects.

A more everyday use of concept-metacognition is in guiding learning. I realise I don't know much about deep convolutional neural networks. That motivates me to learn more. Explicit judgements of one's own knowledge or ignorance are of course metacognitive, but it is not obvious that these judgements concern one's concepts, rather than the subject matter itself. Sometimes we do guide our learning by explicitly conceptual judgements: 'I just don't think I've understood the concept of model-based reinforcement learning yet', 'understanding the concepts of soundness and validity, and the difference between them, was the most useful thing I learnt in first year logic.' The judgement that one grasps a concept poorly will make one reluctant to rely on it and keen to learn more. The judgement that a concept is useful and dependable, and that one understands it well, motivates one to rely on it. However, these kinds of outcomes can equally be the result of object-level judgements like 'consciousness is not a natural kind', or judgements at the meta level that do not concern concepts as such like 'I need to learn more about episodic memory'. So we can't infer the existence of concept-metacognition simply from observing that subjects seem reluctant to rely on some concepts, or from the fact that they guide their learning in a way that enriches some concepts rather than others. Nor can we rule it out, especially once we realise that thinkers can direct nonconceptual, procedural metacognition at their concepts without exercising or having the concept of a concept.

Where there is an explicit judgement it is straightforward to see whether the judgement is metacognitive or not, and whether it concerns concepts as such. It is much harder to tell in cases of procedural metacognition. A thinker can engage in procedural metacognition without using or possessing any concepts of mental states or cognitive processes. My preferred way of understanding procedural metacognition is as a species of non-conceptual representation. What would make it count as metacognitive is that its content concerns another representation or cognitive subsystem. When the representation is not explicit and conceptually-structured, it is much harder to assess the circumstances in which metacognitive contents exist.

A nonconceptual representation can have metacognitive content because of the role it plays in cognitive control, in the way epistemic and cognitive processes unfold. Feelings of uncertainty drive actions that are epistemic rather than directly instrumental: moving to a new location to get a better view, manipulating an object to learn more about it, asking someone for information. Uncertainty also tends to lead to broadly epistemic mental actions like thinking for longer about a problem or engaging in a strategy to recall a memory. Conversely, feelings of certainty can make the subject stop exploring and thinking, and move to taking a decision (Ackerman and Thompson 2017). A 'sense' or feeling that plays these roles would seem to be about certainty or uncertainty, even if it is nonconceptual; playing these roles may even be part of what constitutes the representation as having a metacognitive content.

There will often be a rival object-level story. For example, psychological processes that weigh different sources of information (Ernst and Banks 2002) could just be integrating object-level probabilities, something like degrees of belief: visual information is a probability distribution over potential values for the location of the

object, haptic information is a different probability distribution, and in integrating them the mean of the sharper distribution carries more weight. In many cases involving epistemic actions and control of cognitive processing it is hard to give an object-level explanation that is not implausibly byzantine (Carruthers 2008). A metacognitive story may also be more consistent with what is known about the time course and processing steps involved (Proust 2009). The content question is admittedly less straightforward in the procedural case but that is not an argument against the existence of nonconceptual metacognitive contents.

Some aspects of the dependability of a concept are clearly epistemic, relating directly to the thinker's status as a knower. One is whether the information encoded with a concept is correct or incorrect. Another is whether the concept is empty. A concept like WEED, that is not empty and is perfectly useful for some purposes, may not be very dependable for induction. Furthermore, a thinker's assessment or sense of the 'goodness' of a concept may extend beyond the epistemic. Debates about whether to use racial concepts are like that when the concepts are presupposed to be non-empty and to encode some true statistical generalisations. Broader normative considerations are in play. Conversely, empty concepts may have considerable utility, for example when talking about a fiction, or when describing others who don't know them to be empty (e.g. PHLOGISTON, when talking about scientists of the past). Concepts that are initially empty may also be very useful because they form the basis for children to acquire new, difficult concepts (Carey 2009). The question is not just whether the concept is defective in some narrowly epistemic sense, but whether the concept is fit for purpose given the various ways it is relied on in thought, communication and action. The Hypothesis is that thinkers keep track of some of these qualities of at least some concepts, either separately, or in some compendious overall sense of how much one ought to depend on the concept.

(4) Relevant Existing Research

We are interested in metacognition that attaches to a concept itself, rather than just to a particular judgement made using the concept, so it should have some kind of global effect on the way the concept is used. This excludes one obvious case, what might be called 'confidence-in-categorisation'. People can be asked to categorise some items (e.g. is chess a game?), as in standard concepts research, and then asked about their confidence in the answer they just gave. That is a judgement about a particular categorisation decision, not about the concept in general. It will be affected by the particular features of the question, for example whether the answer is already stored with the concept (part of semantic memory) or whether it has to be worked out. If there is some kind of metacognitive certainty or uncertainty associated with a concept then we should expect it to modulate judgements of confidence-in-categorisation, and do so in a relatively uniform way, but judgements of confidence-in-categorisation do not on their own tell us that there is anything metacognitive going on at the level of the concept.

The existing literature has looked at one kind of metacognitive judgement which is made at the level of the concept: a category learning judgement. Subjects are trained to classify items into one or more novel categories. They are then told that they will be given some novel items to categorise and asked how accurate they think they will be. For example, people learn six novel bird families by looking at images of birds (jays, orioles, etc.) and are then asked, for each novel category, how accurate they think they will be at categorising novel items (Jacoby, et al. 2010). That is a judgement at the level of a (newly-acquired) concept. People are also asked, of an item they have categorised, how confident they are that they have categorised it correctly (a confidence-in-categorisation). Whereas a category learning judgement (CLJ) does concern the concept, and so qualifies as a case of concept-metacognition, a confidence-in-categorisation judgement about an individual item does not.

CLJs show the same broad pattern as many other kinds of metacognitive judgement: they are roughly reliable (Jacoby, et al. 2010, cf. Rawson, et al. 2015, Hartwig and Dunlosky 2017), both in terms of bias and calibration; but they can be readily pushed around by factors that are only very indirect proxies for reliability, for instance repetition (Wahlheim, et al. 2012, Doyle and Hourihan 2016), and massed as opposed to spaced training (Eglington and Kang 2017, Kornell and Bjork 2008). People do seem to know that testing is a more effective way to learn a category than simple repetition, even when the testing does not involve feedback (Jacoby, et al. 2010). One would think people would use their CLJs in deciding what to study more of, but it is not clear that they do (Morehead, et al. 2017). More broadly, it is not clear that CLJs reflect some information that is permanently stored with the concept rather than being a more transient judgement about the learning episode the subject has just engaged in.

CLJs are clearly relevant to our Hypothesis. They are one form of metacognition about a concept. But they are only probing one very particular aspect of a thinker's grasp of a concept, and of its utility, namely how accurate the thinker expects to be in categorising things under the concept. CLJs do not interrogate broader issues about how good a tool a concept is, given the way it is used: how much information the user encodes with the concept, how good it is as a basis for communicating with others, or how well the category referred to supports inductive inferences, for example.

Existing metacognition research gives an indication of what concept-metacognition is likely to be based on. Sources of metacognitive information can be divided into two types: directly accessed and inferred (Schwartz 1994). An example of direct access, in the case of a perceptual decision, would be if confidence in a decision were just based directly on the quantity of evidence for and against the decision. In fact, other factors are involved in shaping the confidence judgement (Palser, et al. 2018, Bang, et al. 2017), but the certainty associated with a percept may be based directly on the quantity of evidence suggesting the world is as it is represented to be (Aitchison, et al. 2015, Zylberberg, et al. 2012). Other sources of information are indirect: cues and heuristics that indicate whether a memory is likely to be reliable, for example (Koriat 2012). For instance, when it takes a long time to come to an answer the subject may feel or infer that the answer is less likely to be correct.

Cues and heuristics that drive other forms of metacognition are also likely to operate on concept-metacognition. Fluency is one promising candidate. People track the speed, and perhaps also the ease, of perceptual processing. Fluent processing is taken to indicate familiarity. If people are asked to remember a word list, the fluency of test items can be manipulated by presenting an unconscious prime immediately before some of the words. This induces an illusory feeling of familiarity (Jacoby and Whitehouse 1989). Conceptual fluency can have a similar effect. Miller, et al. (2008) gave people lists of words to be memorised and then gave them test sentences. The task was to say whether the final word in the sentence had appeared on the list. Conceptual fluency was induced by putting the word in a predictive sentence: 'The stormy sea rocked the BOAT'. Conceptual processing of the final word is facilitated compared with a more neutral sentence: 'He saved up his money and bought a BOAT'. Again, conceptual fluency induced an illusion of familiarity.

A classic electrophysiological measure of this kind of conceptual fluency is the N400 ERP component. An N400 response is elicited by an unexpected word at the end of a sentence, for example, 'He took a sip from the transmitter' (Kutas and Hillyard 1980). The less probable the word, the larger the amplitude of the N400 (Kutas and Hillyard 1984). A signature of this kind, of the fluency of conceptual processing, is a good candidate to play a metacognitive role. It suggests that psychological processes may be automatically generating a sense of the plausibility of meanings being processed. That in itself would be a useful form of metacognition. It would be interesting to discover whether metacognitive ratings of concepts modulate the N400 response.

(5) Forms of Concept-Metacognition

Fundamental to the utility of a concept is the fact that it encodes information that allows us to form expectations. So an important dimension of assessment is how reliable this process is. When I learn information on one occasion, carry it forward in a concept, and that generates an expectation on a future occasion, how likely is it that my expectation will be met? Those could be expectations about what people will say, in the context of communication, or direct expectations about things in the world, as when I expect that a hazelnut will rattle when shaken.

We can define the reliability of a concept relative to its role in generating expectations. Roughly, a reliable concept is one that can be used to generate a large number and variety of correct expectations. On occasions when the thinker would be disposed to deploy the concept, how often would the expectations she then forms be correct? To make this precise we would need to pin down the dispositions to deploy a concept and to form expectations, and specify the circumstances that would count towards correctness and incorrectness of the expectations. Forming a large variety of correct expectations counts in favour of reliability, and incorrect expectations count against. To get a measure of metacognitive accuracy we would have to be able to assess a thinker's metacognitive ratings of their concepts against the ground truth about the

reliability of those concepts. That is complex, but possible at least to estimate. For example we could sample the expectations a thinker forms when they categorise an item under a concept C, and the inductions they are disposed to make on the basis of C, and ask how many more are correct than incorrect.[3]

If I store the expectation that hazelnuts have such-and-such taste, then my expectation will go wrong when I encounter a rotten hazelnut. It will also go wrong when I miscategorise an acorn as a hazelnut. So if I am disposed to be inaccurate when I categorise objects under HAZELNUT, that counts against the reliability of my HAZELNUT concept. The broad notion of reliability will encompass both sources of error. Relatedly, concepts should not produce contradictory categorisations. If I am disposed to categorise a tree I encounter both under ELM and under BEECH, and my concepts tell me it cannot be both, that reflects badly on those two items in my conceptual repertoire.

A thinker's concept may be deficient due to that individual's ignorance. They may only have a very partial grasp of the category in question. Ask a non-specialist and he would probably agree that he doesn't know much about Higgs bosons. He may have a concept, HIGGS BOSON, that encodes some information, e.g. that the particle was first detected at the Large Hadron Collider, but his HIGGS BOSON concept won't be very reliable, in his hands, because he doesn't know much about Higgs bosons. This should be distinguished from another kind of case where there is just not that much be known about the category, or where the category is an unreliable basis for induction. Compare MOTORIST with PRIEST. People might feel that it is more informative to learn that someone is a priest than that they are a motorist, and that PRIEST is also a more dependable basis for making new inductions; correlatively that learning that someone is a motorist will tell you less about what other properties they might have. This difference need not be a matter of ignorance: there may be just less to learn about the characteristics of motorists than of priests.

So there are at least three potential sources of unreliability of an individual's concept of X: how much information they encode about Xs, how accurately they can categorise instances, and how dependable X is as a basis for forming expectations. Concept users might distinguish between these different sources of unreliability, perhaps having a separate sense of the extent of their understanding, of how accurate they are at categorisation, and of the dependability of the category for forming expectations. Or thinkers might roll together all three factors into a single estimate of the relative reliability or unreliability of their concept. Generally, we might expect that a strong sense of understanding would go with a strong sense of dependability, but these could also dissociate in certain domains. Concepts of categories studied by science, GLAUCOMA say, may be rated by an individual as a dependable category, but a category about which one knows little oneself. One might then be inclined to defer to expert knowledge. Religious concepts may show the same pattern. Sperber (2010) has argued that religious believers have a low sense of understanding coupled with a

[3] How we should weight incorrect against correct expectations might depend on the domain.

strong sense of the dependability of concepts like GOD (with his divine omnipresence).[4]

So a first step in investigating these possibilities empirically will be to study the kind of metacognitive assessment or assessments people make in relation to their stock of concepts and to see if they distinguish between different aspects of reliability. Then we need to understand what cognitive role is played by each. For example, does feedback that lowers a thinker's confidence in a concept thereby reduce their confidence in all the information they encode with the concept? Is a concept associated with more than one metacognitive parameter, each with its own cognitive role? Is there a metacognitive parameter that affects which concept a thinker will choose to deploy? An object typically falls under more than one concept. Is there something metacognitive that affects how it will be categorised; and, when learning something about the object, which concept that new information will be stored with?

In other domains researchers measure how accurate people's metacognitive judgements are: their bias and calibration. To do that with concept-metacognition, we first need to define what reliability consists in. The discussion above characterises reliability in broad terms. To pin down something more definite we first need to know what kinds of concept-metacognition thinkers go in for and what cognitive role these assessments play.

We also want to know how thinkers form these metacognitive assessments. Fluency seems to have a pervasive influence on metacognition so we would expect it to play a role here. We saw above that learning a new category from massed as opposed to inter-spaced examples makes people's category learning judgements (CLJs) more confident. Does that have a long-term effect on how the concept is used? Does it have an effect on other kinds of assessment of a concept, for instance whether a thinker is disposed to depend on it for induction? And do other fluency-related manipulations affect these stored metacognitive parameters?

Communication is central to concept use, even more so than with perceptual and perhaps mnemonic processes. So we might expect communicative cues to be affected by and in turn to affect concept-metacognition. Scare quotes are often a written cue that the writer is reluctant to place much weight on a concept; air quotes and tone of voice can play a similar role in oral communication. A low metacognitive rating may make a speaker reluctant to rely on a concept in what they say at all. Conversely, when feedback from a communicative encounter is negative, that might redound negatively to the concepts relied on. If I start getting quizzical looks and blank stares when I talk about the SUPEREGO, that might reduce my sense of the reliability of that concept.

[4] Thanks to Joëlle Proust for the suggestion. There may also be individual differences in the relation between sense of understanding and sense of dependability (if both senses exist), with a spectrum of cases, or pathological cases, where people have a high sense of dependability across the board despite knowing that they only understand many concepts partially. (Thanks to Francesca Happé for this thought.)

Correspondingly, when everything goes smoothly, the concepts I rely on get a clean bill of health. Their metacognitive rating might be increased as a result.

Another source of variation that arises for concepts is coherence, in addition to informativeness, which we discussed above. A concept that coheres with many other concepts in a thinker's repertoire might thereby get a higher rating of dependability or understanding than one which encodes a more isolated body of information.

Does the idea of concept-metacognition apply to different theories of concepts? I have presented the idea by reference to concepts thought of as mental representations possessed by individual thinkers. Other ways of theorising concepts include: as abilities, as public or shared entities, and as abstract objects. Abilities have a basis in the thinker's psychology so it is straightforward to extend the idea to them: thinkers may go in for metacognitive assessment of their possession or exercise of abilities to categorise, induce, and so on. I'm sceptical of the idea of there being "the" (public) concept of an X, except perhaps derivatively from generalising over individual concept users. But if there is such an entity, then thinkers could have views about how dependable it is and how well they understand it. If instead concepts are abstract objects, then something must be said about how thinkers grasp that abstract object so that it affects their thought and action. Concept-metacognition might then exist relative to the psychological processes involved in grasping and using a concept.

The speculations in this section derive plausibility from everyday concept use. One can presumably get some inclination of what is going on when we use concepts by reflecting on everyday concept use by oneself and others. However, these are empirical questions. Psychological research is clearly needed if we are to get a better understanding of the existence, variety and role of concept-metacognition.

Metacognition of concepts falls within the broader class of metacognition of thoughts and thought processes. Here there is a large body of empirical research, as mentioned above, for example on metacognition of semantic memory. If there is indeed a form of metacognition more specifically targeted on concepts then we should expect that to have a systematic effect on thought. For example, a concept that carries a low sense of dependability should systematically decrease the thinker's confidence in any belief involving that concept. We would then predict that a modulation to the sense of concept-dependability should produce a corresponding modulation in a whole suite of beliefs.

(6) Role in Conceptual Processing

If systematic concept-metacognition is found to exist, it will doubtless have a role to play in various processes that have already been studied by concepts researchers. One obvious place to look is work on conceptual hierarchy and the existence of a preferred 'basic' level of categorisation (Rosch, et al. 1976, Rosch 1978). As mentioned above, a thinker will typically have more than one concept under which they could categorise

a given object. Which do they use? Existing research has established that in very many domains there is a preferred level of categorisation. Asked to categorise pictures of animals people will typically use BIRD rather than SPARROW for an item that falls under both concepts. This is the basic level.

There are various theories as to why there should be a preferred level of categorisation. It may be the level in a hierarchy where there is an optimal trade-off between generality and informativeness or inductive power (Murphy 2004). We need a category that is specific enough that there is a lot of information to encode about items in the category. But we also want generality so we can generalise anything we learn to many new instances. Understanding why there should be a preferred basic level does not tell us, however, how thinkers register that fact. Do they perform a calculation on the fly or do concepts carry around some parameter that affects whether they are selected for categorisation and so privileges one concept when several are available? If a concept does encode some form of metacognitive assessment, as suggested in the last section, that parameter is a good candidate for this role. That would then predict that concept-users' relative ratings of the goodness of a concept for induction, communication and so on would predict which concepts are at the basic level for them.

Expertise affects where the basic level is found in a hierarchy (Tanaka and Taylor 1991, Johnson and Mervis 1997). Encoding more information makes the basic level move lower down in the hierarchy. For example, a bird expert will typically say 'sparrow' where a non-expert would say 'bird'. If concept-metacognition is the psychological marker of what a concept-user takes to be the basic level, it too should be influenced by expertise in this way. In the last section I distinguished dependability of a concept from the thinker's individual understanding of it. Dependability looks to continue to increase as we move down the hierarchy. Categorising an item as a sparrow tells us more specific things about it than categorising it as a bird or an animal does. A new observed feature is more likely to apply to all sparrows than to apply to all birds (because of the class inclusion). However, a thinker's sense of understanding may decrease once they descend below the basic level in their hierarchy. They may know little more about a sparrow than that it is a small bird. So we might expect the basic level to correspond with the conceptual level at which a thinker rates their own individual understanding to be greatest.

A second phenomenon of interest here is developmental change. Younger children tend to use perceptual similarity to generalise from one sample to another. There is then a developmental change: children begin to generalise based on underlying features of a category, where different members of a category may not share surface features (Sloutsky 2010, Keil and Batterman 1984). It would be interesting to know whether this developmental step is accompanied by any metacognitive change. If a rating as good or otherwise for induction is a ubiquitous feature of concepts, then we might find that a metacognitive appreciation accompanies the developmental step. The child realises in some way that a concept they have, DOG say, is a better basis for generalisation than surface similarity. On the other hand, we could find that

metacognitive assessment of concepts is initially absent—that there are children who have and use concepts but without having a concept-level certainty-related parameter that has a global effect on use of a concept.

Concepts researchers once debated whether prototypes, exemplars or mini-theories offered the best explanation of categorisation and other ways in which people use concepts (Murphy 2004). Now many researchers agree that all three kinds of psychological structure exist (Kruschke 2005, Ashby and O'Brien 2005). Machery (2009) relied on these results to argue that 'concept' does not pick out a natural kind and we should stop using the term in serious theorising. Some react to these results by advocating pluralism (Weiskopf 2009). A thinker has several different concepts of the same category: a stored prototype, a set of exemplars, and a mini-theory. Others argue that these different psychological structures are connected in a hybrid representation stored in semantic memory (Vicente and Martínez Manrique 2016), and that psychological structures of different kinds can be active together in the same task.

Assays of concept-metacognition could help to arbitrate this debate. If there is a global confidence parameter that regulates all deployments of a concept, then pluralism predicts that there will be independent parameters for each structure (prototype, exemplars, theory), whereas hybridism predicts that there will be one metacognitive setting that operates in a unified way on all these psychological structures. An intervention on concept-dependability administered while activating a prototype say, should have an effect that carries over directly to the concept-dependability operating when using exemplars or a mini-theory. If concept-metacognition lends support to hybridism over pluralism, then metacognition will be playing an important unifying role. Psychological processes that have looked heterogeneous when studying categorisation in isolation would then in fact be connected by a metacognitive process that operates on those structures in a unified way. That would also furnish another argument against Machery's concept eliminativism.

Laurence Barsalou's influential 'enactivism' is the claim that reactivation of perceptual and motor representations, and interaction with affective states and states of the body, are all heavily involved in episodes of conceptual thought (Barsalou 1999, Prinz 2002). Enactivism is often presented as a radically new proposal. It certainly differs in important ways from accounts of concept processing based on prototypes, exemplars or theories. If true, enactivism about concepts would have repercussions for the way we should understand concept-metacognition. Procedural metacognition is a natural ally of enactivism. If to think with a concept is just to reactivate a sensorimotor process then it would be unsurprising that aspects of that reactivation process, like speed and fluency, should have systematic effects on confidence. However, for there to be metacognition at the level of the concept we would have to find more than that. Feelings of fluency or judgements of confidence or familiarity about an occasion of concept use need not reflect a systematic feature that operates across the concept. We are interested in a parameter that operates in a relatively uniform way across uses of a concept. It could for example be associated with Barsalou's 'simulator' – the

structure responsible for patterns of context-sensitive co-activation between sensory, motoric and affective states. Whether something metacognitive is a feature of 'enactive' conceptual thinking is not known.

Should concept-metacognition figure in the way concepts are individuated? An account of concept individuation says what it is to be an instance of the same concept again. Suppose a concept is stored with a rating of reliability or dependability that plays an important role in how the concept is used. That seems to argue for concept-metacognition being amongst the material that counts as part of a concept and is relied on for individuation (in the 'conceptual core' of Laurence and Margolis (1999), rather than amongst Camp (2015)'s 'characterisations'). However, this move would undermine one of the things we want to do with concept-metacognition, namely to track how the metacognitive rating or ratings associated with a concept can change as a result of experience and thereby alter the way the concept is relied upon. So we need to allow that the same concept can have different metacognitive ratings in a given thinker at different times. So we don't want sameness of a metacognitive parameter to be necessary for two tokens to count as exercises of the same concept.

That should not undermine the importance of concept-metacognition. We should not expect the material that figures in concept individuation to explain everything about the way a concept is used. On any theory, information and other material stored with a concept will be important for psychological explanation. Indeed, for a concept atomist (Fodor 1998), most of what we say and do relying on a concept is explained by pieces of information stored using the concept (for psychologists this is stored 'knowledge', for Fodor these are stored sentences in the language of thought); the same is true if enactive dispositions to produce perceptual imagery or affective states are associated with a concept. We reason our way to conclusions about whales using the information stored with our WHALE concept. We expect a surface categorised as GRASS to have a certain texture because of a sensorimotor expectation stored with our GRASS concept. On any view not all information of this kind is individuative of the concept. Similarly, concept-metacognition can account for robust patterns in the way concepts are deployed and relied upon, without that committing us to its playing a role in concept individuation.

(7) Wider Applications

The idea of concept-metacognition has wider applications. One that is particularly relevant to philosophy is its potential role in generating intuitions about claims and arguments. If it exists, it seems likely that the sense of reliability or unreliability associated with a concept will affect the intuitive plausibility of arguments that use the concept. Consider an episode of reasoning involving the concept SHRUB (deductive or inductive). Compare it with an inference with exactly the same form that instead relies on the concept AMPHIBIAN. It seems likely that the second inference will appear more plausible than the first just in virtue of relying on a more dependable concept.

Having our intuitions about arguments shaped in this way would make sense. There is a rough parallel with confidence in a judgement. When a thinker judges that p, one useful epistemic property is if judgements of that type are reliable: the probability is high that, when the thinker is disposed to judge that p, p is the case. That is an externalist species of justification that the thinker does not have access to directly. Confidence is an internal index of reliability: high confidence should indicate that the thinker's judgement that p is reliable. Confidence is a good index to the extent that high vs. low confidence tracks correctness vs. incorrectness in the judgement. Concept-metacognition can work in the same way. The thinker needs concepts that are dependable for induction and the other uses for which she relies on them. A certainty parameter attached to a concept can act as an internal index of reliability. It is a good index to the extent that concept reliability tends to be rated high for concepts that really are reliable and low for those that are not.

If this is right, then the low reliability I associate with INNATENESS will reduce the intuitive plausibility of any claim I consider that involves the concept of innateness. There low reliability flows from the category being defective. A sense of unreliability may also make me reluctant to use the concept in the first place. A sense of ignorance can have the same effect. My feeling that I don't really understand the concept HIGGS BOSON reduces my confidence in the few things I encode in that concept. I do know that the Higgs boson was first discovered in Geneva, but even that simple claim feels a bit shaky because I have a background sense of not really knowing what I'm thinking about. Compare the claim that water has been found on Mars. My source of knowledge is similar, but WATER and MARS have less uncertainty and so the claim seems more plausible.

Often philosophical disagreements can be traced to differences in background theoretical assumptions. Differences in concept-metacognition may be another important source. If so, two philosophers could agree about relevant background assumptions, and agree about the soundness and validity of an argument, while one finds the argument intuitively plausible and the other not. Maybe the argument concerns zombies and one rates their ZOMBIE concept as less reliable than the other does. Their difference of opinion could then be located in different assessments about the reliability of the concepts involved. Concept-metacognition thus offers us a novel way to understand philosophical disagreement. We might be able to test for this via disjunction introduction. People are often more reluctant to accept a disjunction than they are to accept one of its disjuncts (which is irrational, from a truth conditional perspective). Offer our philosophical interlocutors an argument that introduces the potentially unreliable concept: 'Snow is white; therefore, snow is white or zombies are conscious.' Suppose they agree about the truth value of the second disjunct, so it is not intuitions about truth that are driving any intuited difference. If the two sense a difference in the force of the argument, that could plausibly be traced to having a different sense of the reliability of the concepts involved (ZOMBIE and/or CONSCIOUS).

The second application I want to discuss concerns concept construction. Constructing the concepts with which we think is rarely an individual endeavour. It is something we

do along with others. Contemporary philosophy of science has drawn a similar lesson about scientific epistemology. Scientific discoveries are rarely individual achievements. Progress occurs through a social process by which hypotheses are generated and subjected to collective sceptical scrutiny. That is equally true in philosophy. Progress depends on researchers trying out ideas and others building on them, with each step subject to analysis and debate. In both science and philosophy, a central plank of theory-construction is to create the concepts to theorise with. Researchers have to continually assess which concepts they can rely on and which should be changed or abandoned. Concept-metacognition probably plays a key role in these decisions.

I highlighted the way researchers use signals like scare quotes or tone of voice to signal that they are dubious about relying on a concept like INNATENESS. That in turn affects whether others rely on the concept. Sometimes explicit disclaimers are also given. These signs surely contribute to the collective sense that some concepts are a more secure basis for theorising than others, for example that WORKING MEMORY is better than REPRESSED MEMORY. Individuals exercise a concept-level analogue of epistemic vigilance (Sperber, et al. 2010). The result is that collective concept construction proceeds in useful and rational ways.

Or that it can do. If the sources of potential dissenting voices and disagreement are removed, then this process will instead proceed in unhelpful ways. Repetition in an 'echo chamber' of people selected for the similarity of their views might act to increase everyone's confidence ratings in a set of concepts in a way that completely loses touch with the actual reliability of those concepts. That is to say, concept-metacognition should offer a window into cases where social concept construction goes awry as well as those where it succeeds.

(8) Conclusion

This paper has set out to flesh out the hypothesis that there is metacognition directed at concepts. Existing research suggests this is a promising line of enquiry. It may take the form of a body of information we store about the reliability of a concept, or may be 'procedural', consisting of a sense, feeling or nonconceptual representation. We can in principle distinguish between the thinker's individual level of understanding and the dependability of a category itself for forming expectations through induction and reasoning. Concept users may keep track of this difference, and also of their accuracy in categorising instances under the concept, or they may run these things together.

If some form of concept-metacognition is a widespread feature of human cognition then it may help explain various phenomena: developmental transitions, what makes a taxonomic level psychologically 'basic' for a thinker, and may forge a connection between different kinds of knowledge structure (prototypes, exemplars, theories, enactive simulations). It is also likely to have wider applications, for example in diagnosing the reason for disagreements about philosophical arguments (and

arguments in other areas), and in understanding the social process of collective concept construction, both in science and in everyday life. As has proven the case in other areas of cognition, digging down into the metacognitive processes that operate over concepts will tell us important new things about the nature of our concepts, and is likely to shed light on the way thought and reasoning work more generally.

Acknowledgements

For discussion and comments the author would like to thank Steve Fleming, James Hampton, Jake Quilty-Dunn, Joulia Smortchkova, Sapphira Thorne, and audiences at the Institute of Philosophy lab meeting, the Oxford mind work-in-progress group, and the 2018 *Mind & Language* workshop on metacognition. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 681422 (MetCogCon).

References

Ackerman, R., and V. A. Thompson. 2017. "Meta-Reasoning: Monitoring and Control of Thinking and Reasoning," *Trends Cogn Sci*, **21**: 607-17.

Aitchison, Laurence, Dan Bang, Bahador Bahrami, and Peter E Latham. 2015. "Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making," *PLOS Computational Biology*, **11**: e1004519.

Ashby, F Gregory, and Jeffrey B O'Brien. 2005. "Category Learning and Multiple Memory Systems," *Trends in Cognitive Sciences*, **9**: 83-89.

Bahrami, Bahador, Karsten Olsen, Peter E. Latham, Andreas Roepstorff, Geraint Rees, and Chris D. Frith. 2010. "Optimally Interacting Minds," *Science*, **329**: 1081-85.

Bang, Dan, Laurence Aitchison, Rani Moran, Santiago Herce Castanon, Banafsheh Rafiee, Ali Mahmoodi, Jennifer YF Lau, Peter E Latham, Bahador Bahrami, and Christopher Summerfield. 2017. "Confidence Matching in Group Decision-Making," *Nature Human Behaviour*, **1**: 0117.

Barsalou, L.W. 1999. "Perceptual Symbol Systems," *Behavioral and Brain Sciences*, **22**: 577-660.

Camp, Elisabeth. 2015. "Logical Concepts and Associative Characterizations". In Margolis and Laurence, eds, *Conceptual Mind: New Directions in the Study of Concepts.* London / Cambridge MA: MIT Press.

Carey, Susan. 2009. *The Origin of Concepts*. Oxford: O.U.P.

Carruthers, Peter. 2008. "Meta-Cognition in Animals: A Skeptical Look," *Mind & Language*, **23**: 58-89.

Churchland, Patricia Smith. 1983. "Consciousness: The Transmutation of a Concept," *Pacific Philosophical Quarterly*, **64**: 80-95.

Donoso, Mael, Anne G. E. Collins, and Etienne Koechlin. 2014. "Foundations of Human Reasoning in the Prefrontal Cortex," *Science*, **344**: 1481-86.

Doyle, Mario E., and Kathleen L. Hourihan. 2016. "Metacognitive Monitoring During Category Learning: How Success Affects Future Behaviour," *Memory*, **24**: 1197-207.

Eglington, Luke G, and Sean HK Kang. 2017. "Interleaved Presentation Benefits Science Category Learning," *Journal of Applied Research in Memory and Cognition*, **6**: 475-85.

Ernst, M. O., and M. S. Banks. 2002. "Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion," *Nature*, **415**: 429-33.

Fleming, Stephen M, R. S. Weil, Z. Nagy, R. J. Dolan, and G. Rees. 2010. "Relating Introspective Accuracy to Individual Differences in Brain Structure," *Science*, **329**: 1541-3.

Fodor, Jerry A. 1998. *Concepts: Where Cognitive Science Went Wrong*. New York: OUP.

Godman, Marion. 2018. "Gender as a Historical Kind: A Tale of Two Genders?," *Biology & Philosophy*, **33**: 21.

Griffiths, Paul E. 2002. "What Is Innateness?," *The Monist*, **85**: 70-85.

Hartwig, Marissa K., and John Dunlosky. 2017. "Category Learning Judgments in the Classroom: Can Students Judge How Well They Know Course Topics?," *Contemporary Educational Psychology*, **49**: 80-90.

Haslanger, Sally. 2000. "Gender and Race:(What) Are They?(What) Do We Want Them to Be?," *Nous*, **34**: 31-55.

Jacoby, Larry L, and Kevin Whitehouse. 1989. "An Illusion of Memory: False Recognition Influenced by Unconscious Perception," *Journal of Experimental psychology: General*, **118**: 126.

Jacoby, Larry L., Christopher N. Wahlheim, and Jennifer H. Coane. 2010. "Test-Enhanced Learning of Natural Concepts: Effects on Recognition Memory, Classification, and Metacognition," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **36**: 1441.

Johnson, Kathy E, and Carolyn B Mervis. 1997. "Effects of Varying Levels of Expertise on the Basic Level of Categorization," *Journal of Experimental psychology: General*, **126**: 248.

Keil, Frank C., and Nancy Batterman. 1984. "A Characteristic-to-Defining Shift in the Development of Word Meaning," *Journal of verbal learning and verbal behavior*, **23**: 221-36.

Koriat, Asher. 2012. "The Self-Consistency Model of Subjective Confidence.," *Psychological Review*, **119**: 80-113.

Kornell, Nate, and Robert A Bjork. 2008. "Learning Concepts and Categories: Is Spacing the "Enemy of Induction"?," *Psychological Science*, **19**: 585-92.

Kruschke, John K. 2005. "Category Learning". In Lamberts and Goldstone, eds, *The Handbook of Cognition*. London: Sage, 183-201.

Kutas, Marta, and Steven A Hillyard. 1980. "Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity," *Science*, **207**: 203-05.

———. 1984. "Brain Potentials During Reading Reflect Word Expectancy and Semantic Association," *Nature*, **307**: 161.

Laurence, Stephen, and Eric Margolis. 1999. "Concepts and Cognitive Science". In Laurence and Margolis, eds, *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Lee, Sang Wan, Shinsuke Shimojo, and John P. O'Doherty. 2014. "Neural Computations Underlying Arbitration between Model-Based and Model-Free Learning," *Neuron*, **81**: 687-99.

Machery, Edouard. 2009. *Doing without Concepts*. Oxford University Press New York.

Mallon, R. 2006. "'Race': Normative, Not Metaphysical or Semantic," *Ethics*, **116**: 525-51.

Mameli, Matteo, and P Bateson. 2006. "Innateness and the Sciences," *Biology and Philosophy*, **22**: 155-88.

Mameli, Matteo. 2008. "On Innateness: The Clutter Hypothesis and the Cluster Hypothesis," *Journal of Philosophy*, **55**: 719-36.

Meyniel, Florent, and Stanislas Dehaene. 2017. "Brain Networks for Confidence Weighting and Hierarchical Inference During Probabilistic Learning," *Proceedings of the National Academy of Sciences*, **114**: E3859-E68.

Miller, Jeremy K., Marianne E. Lloyd, and Deanne L. Westerman. 2008. "When Does Modality Matter? Perceptual Versus Conceptual Fluency-Based Illusions in Recognition Memory," *Journal of Memory and Language*, **58**: 1080-94.

Morehead, Kayla, John Dunlosky, and Nathaniel L Foster. 2017. "Do People Use Category-Learning Judgments to Regulate Their Learning of Natural Categories?," *Memory & Cognition*, **45**: 1253-69.

Murphy, G. L. 2004. *The Big Book of Concepts*. London / Cambridge, MA: MIT Press.

Nelson, Thomas O., and Louis Narens. 1990. "Metamemory: A Theoretical Framework and New Findings," *The psychology of learning and motivation*, **26**: 125-41.

Palser, E. R., A. Fotopoulou, and J. M. Kilner. 2018. "Altering Movement Parameters Disrupts Metacognitive Accuracy," *Conscious Cogn*, **57**: 33-40.

Pouget, A., J. Drugowitsch, and A. Kepecs. 2016. "Confidence and Certainty: Distinct Probabilistic Quantities for Different Goals," *Nat Neurosci*, **19**: 366-74.

Prinz, Jesse. 2002. *Furnishing the Mind*. Cambridge, MA: MIT Press.

Proust, J. 2009. "The Representational Basis of Brute Metacognition: A Proposal". In Lurz, ed, *The Philosophy of Animal Minds: New Essays on Animal Thought and Consciousness*. Cambridge: C.U.P.

———. 2013. *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford University Press.

Rawson, Katherine A., Ruthann C. Thomas, and Larry L. Jacoby. 2015. "The Power of Examples: Illustrative Examples Enhance Conceptual Learning of Declarative Concepts," *Educ Psychol Rev*, **27**: 483–504.

Rosch, E., C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. 1976. "Basic Objects in Natural Categories," *Cognitive Psychology*, **8**.

Rosch, Eleanor. 1978. "Principles of Categorisation". In Rosch and Lloyd, eds, *Cognition and Categorisation*. Hillsdale, NJ: Laurence Erlbaum.

Schwartz, Bennettl. 1994. "Sources of Information in Metamemory: Judgments of Learning and Feelings of Knowing," *Psychonomic Bulletin & Review*, **1**: 357-75.

Shea, Nicholas. 2014. "Reward Prediction Error Signals Are Meta-Representational," *Nous*, **48**: 314-41.

Shea, Nicholas, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D. Frith. 2014. "Supra-Personal Cognitive Control and Metacognition," *Trends in Cognitive Sciences*, **18**: 186-93.

Sloutsky, Vladimir M. 2010. "From Perceptual Categories to Concepts: What Develops?," *Cognitive Science*, **34**: 1244-86.

Sperber, Dan. 2010. "The Guru Effect," *Review of philosophy and psychology*, **1**: 583-92.

Sperber, Dan, Fabrice Clement, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. "Epistemic Vigilance," *Mind & Language*, **25**: 359-93.

Tanaka, James W, and Marjorie Taylor. 1991. "Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder?," *Cognitive Psychology*, **23**: 457-82.

Vicente, Agustín, and Fernando Martínez Manrique. 2016. "The Big Concepts Paper: A Defence of Hybridism," *The British Journal for the Philosophy of Science*, **67**: 59-88.

Wahlheim, Christopher N., Bridgid Finn, and Larry L. Jacoby. 2012. "Metacognitive Judgments of Repetition and Variability Effects in Natural Concept Learning: Evidence for Variability Neglect," *Memory and Cognition*, **40**: 703-16.

Weiskopf, Daniel A. 2009. "Atomism, Pluralism, and Conceptual Content," *Philosophy and Phenomenological Research*, **74**: 131-63.

Zylberberg, Ariel, Pablo Barttfeld, and Mariano Sigman. 2012. "The Construction of Confidence in a Perceptual Decision," *Frontiers in integrative neuroscience*, **6**: 79.