DOI: 10.1111/phpr.12951

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Third-personal evidence for perceptual confidence

John Morrison

Barnard College, Columbia University

Correspondence
John Morrison, Barnard College, Columbia University.
Email: jmorrison@barnard.edu

Abstract

Perceptual Confidence is the view that our conscious perceptual experiences assign confidence. In previous papers, I motivated it using first-personal evidence (Morrison, 2016), and Jessie Munton motivated it using normative evidence (Munton, 2016). In this paper, I will consider the extent to which it is motivated by third-personal evidence. I will argue that the current evidence is supportive but not decisive. I will then describe experiments that might provide stronger evidence. I hope to thereby provide a roadmap for future research.

1 | INTRODUCTION

We are often uncertain about our environment. How do our conscious perceptual experiences contribute to that uncertainty? According to the standard view, our perceptual experiences just represent propositions—that a sign is blue, that a car turned 90 degrees. If we're uncertain about our environment, it's because of what happens later, at the level of belief. We might believe that our perceptual experience is unreliable, or that our background evidence supports another proposition. According to Perceptual Confidence, perceptual experiences can make a more direct contribution to our uncertainty: they can assign probabilities to propositions. Or, as I prefer to put it: they can assign *confidences* to propositions. Let's use examples to clarify what this means.

Color: Suppose you're walking on a trail that you believe is marked by green trail markers. At some point, you might see a marker in the distance and say, "That looks as though it could be blue." After you walk farther down the trail, you might say, "That looks as though it's probably blue." And then, after you walk even farther, you might say, "Uh-oh, that looks as though it's blue," or perhaps just, "Uh-oh, that's blue." All of these reports reflect your increasing confidence at the level of belief—what I call your "doxastic confidence." But they also seem to reflect your confidence at the level of perception—what I call your "perceptual confidence." Trail markers in the

distance don't just look blue or some other color. They sometimes look as though they're probably blue.

Direction: Suppose you're in the passenger seat of a car with your eyes closed. The driver suddenly accelerates and veers to the right. Immediately afterward, you might report equal confidence that the car turned 90 degrees and that the car turned 80 degrees, using gestures to indicate the relevant angles. Suppose the driver then turns again at the same angle, but at a much slower speed. You might report higher confidence that the car turned 90 degrees rather than 80 degrees. Once again, your reports don't seem to just reflect your doxastic confidence. They seem to reflect your perceptual confidence. We don't always feel like we're turning at a specific angle or a specific range of angles. We sometimes feel like we're more likely turning at some angles than others.

Location: Suppose you're at a rowdy party and hear a friend's voice from across the room, in the general direction of the kitchen. If asked, you might report slightly higher confidence that she's inside the kitchen, rather than slightly outside of it. But as the party starts to thin out, you might report increasing confidence that she's in the kitchen, until eventually you're sure that's where she is. As with the other examples, your reports seem to reflect your increasing perceptual confidence. We don't just hear sounds as coming from one location or range of locations. We sometimes hear them as more likely coming from some locations than others.

Flavor: Suppose you and two friends each ask for a coffee with a dash of sugar. You sip your coffee first and complain that it isn't even a little sweet. You then sip a friend's coffee and report that it might be a little sweet. Surprised by the discrepancy, you sip another friend's coffee and report that it's probably a little sweet, but you're still not certain. Once again, your reports seem to reflect your increasing levels of perceptual confidence. Beverages don't just taste sweet or fail to taste sweet. They sometimes taste as though they might be sweet, are probably sweet, etc.

We just considered examples in which perceptual confidence varies with distance, speed, noise, and intensity. There are many other causes of variation, and thus many other examples. For example, the relevant object might be small, partially occluded, blurred, moving, changing, or shaking. You might be intoxicated, tired, distracted, surprised, or confused. There will be more causes if perception represents propositions about the near future, such as whether a bowling ball will knock down the remaining pins (see, e.g., James, 1890, pp. 609–610). There will be still more causes if perception represents propositions about possible actions, such as whether an opponent's soccer ball is close enough to poke away (see, e.g., Gibson, 1979, Ch 8).

Perceptual Confidence leaves room for disagreement about how confidences are integrated into the propositional structure of experiences. On one view, an experience represents confidences as well as propositions. An experience might, for example, represent the ordered pair <.6, that the sign is blue>. On another view, confidences qualify the experience's relation to the proposition. An experience might, for example, represent-to-degree-.6 that the sign is blue. Saying that your experiences "assign" confidence is my way of remaining neutral.[1]

The examples above helped us get a grip on perceptual confidence from a first-personal perspective. We can also get a grip on it from a third-personal perspective. We might think of the brain as containing a stream of activity that begins with sensory input and ends with a behavioral output. Perceptual experience occurs somewhere in the middle. As I think Perceptual Confidence is most plausibly developed, there is confidence in the brain that precedes perceptual experience,

[1] For an overview of this disagreement, see my (2016, pp. 36–38) and (Gross, 2020). Vance (2020, pp. 387–389) argues that confidence belongs in the attitude. Shea (2020, Section 3) isn't convinced there's an attitude-content distinction for nonconceptual representations, but in general favors putting them in the content. See also (Moss, 2018, p. 91).

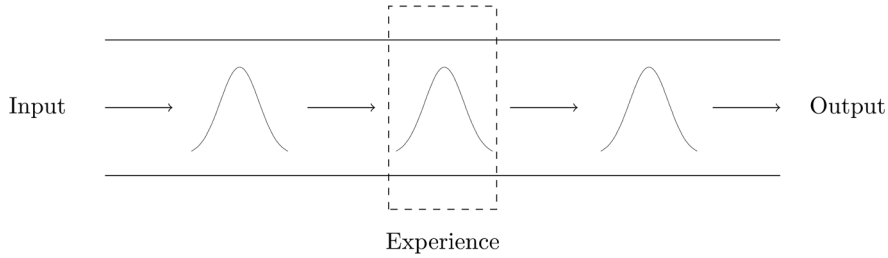

FIGURE 1 Third-personal perspective on Perceptual Confidence

is preserved through perceptual experience, and is available for whichever parts of the brain are ultimately responsible for behavior. Figure 1 helps convey this perspective.

The arrows depict the stream of activity through the brain, from earlier regions to later regions. If there's feedback from later regions to earlier regions, more arrows are needed. The normal distributions depict the assignment of confidence to a range of propositions (e.g., a proposition for each shade of blue and green), but the distribution needn't be normal. On the other extreme, confidence might be distributed over just two propositions (e.g., blue, not blue). The dotted line indicates the neural activity that gives rise to perceptual experience. There is disagreement about where, exactly, that neural activity falls in the stream of activity from earlier regions to later regions. There is also disagreement about the sense in which that activity *gives rise* to perceptual experience. Is it identity? Or merely necessary covariation? I'll return to both disagreements later.

Perceptual Confidence should interest philosophers because it bears on the content, phenomenology, and epistemology of perceptual experiences (see Morrison, 2016; Munton, 2016). It should also interest psychologists and neuroscientists. Perceptual decision-making is a central topic in psychology and neuroscience, but surprisingly little attention is paid to the role of perceptual experiences—that is, to the role of perceptual consciousness. If perceptual experiences merely carry forward representations of colors, directions, distances, and flavors, then uncertainty must have other sources, such as unconscious processes and background beliefs. But if Perceptual Confidence is true, perceptual experiences more directly introduce uncertainty into our decisions.

Perceptual Confidence might also challenge widespread methods for studying consciousness. When we ask subjects whether a stimulus is visible, it's natural to assume we're asking a straightforward yes-no question. But subjects sometimes report uncertainty (Green and Swets, 1966, Ch 2). How can they be uncertain? Perceptual Confidence provides a plausible explanation: they can have more or less perceptual confidence that the stimulus is present. In that case, when we ask whether a stimulus is visible, we are not asking a straightforward yes-no question. We are asking a gradable question. This might lead to subtler measures of whether a stimulus is visible (e.g., Sandberg & Overgaard, 2015). This might also lead to a new understanding of blindsight. Patients with blindsight say that they're blind, but perhaps they just have perceptual experiences that assign very low confidence (see Wu, 2018, Section 4.2). Similarly, when we look for priming effects of a stimulus, we standardly assume that the stimulus either was or was not perceived (for overview, see Kouider & Dehaene, 2007). But Perceptual Confidence implies that it can be perceived with more or less confidence, and stimuli perceived with low confidence might have diminished priming effects. In that case, stimuli without measurable priming effects might not be subliminal. They might just have been perceived with very low confidence.

Decision-making, blindsight, and priming are just three of the reasons why Perceptual Confidence should interest psychologists and neuroscientists. In the conclusion, I will list three more.

Why think that Perceptual Confidence is true? I previously argued (2016) that Perceptual Confidence best explains why our experiences sometimes *cause* varying degrees of doxastic confidence. My argument was first-personal, in that it relied on claims about which beliefs our experiences would produce in us if we believed whatever our experiences told us.

Jessie Munton subsequently argued (2016) that Perceptual Confidence best explains why our experiences sometimes *directly justify* varying degrees of doxastic confidence. Her argument was normative, in that it relied on claims about which beliefs our experiences directly justify and about how our experiences can justify them.[2,3]

The goal of this paper is to consider whether it's possible to give a third-personal argument for Perceptual Confidence—that is, an argument that draws primarily on scientific evidence.[4] I think it's important to consider whether there might be such an argument, because, like others, I'm inclined to give scientific evidence the most weight. Introspective judgments and normative intuitions are often divergent, and disagreements about them can be hard to resolve.

My conclusion will be that, while none of the existing scientific evidence is decisive, some of it offers Perceptual Confidence an intermediate level of support. I will also describe experiments that might provide more decisive evidence.

2 | CLARIFICATIONS

Because philosophers, psychologists, and neuroscientists often use the same words in different ways, some clarifications might be helpful.

As I use 'perceptual experience', it is a representation that's conscious, automatic, accessible, dissociable from doxastic states, directed towards nearby objects and properties, and fast enough that we can't detect any delay. My argument won't rely on most of these features. It'll just presuppose that perceptual experiences are conscious and directed at nearby objects and properties. Thus, you can still accept my conclusion even if your definition of 'perceptual experience' adds or subtracts other features.

[2] For criticisms of both first-personal and normative arguments, see (Denison 2017); (Block, 2018); (Cheng, 2018); (Gross, 2018); (Beck, 2020); (Nanay, 2020); (Byrne, 2021); (Raleigh & Vindrola, 2021).

[3] For other arguments, see (Moss, 2018, pp. 92–95) and (Laasik, 2020).

[4] Moss (2018, pp. 97–98) cites empirical evidence for probabilistic representations in cue combination. But she does not provide any evidence, empirical or otherwise, for thinking that the probabilistic representations are conscious. She merely points out that it's *possible* that the probabilistic representations are maintained until an action is selected, and that it's therefore *possible* that conscious perception includes a probabilistic representation that integrates multiple cues.Clark (2018, p. 84) argues that hierarchical predictive coding models of perception can explain perceptual confidences. But he doesn't present this as an argument for perceptual confidences, and for good reason. Not only is the empirical evidence for hierarchical predictive coding models controversial, but these models are consistent with alternative views, including several we'll consider. Thus, while predictive coding models are compatible with Perceptual Confidence, they do not motivate it. More generally, as Denison, Block, and Samaha (2022) point out, existing computational models of visual processing do not, by themselves, support Perceptual Confidence, because they don't indicate the role of consciousness.Vance (2020) argues that computational theories of perception have the resources to explain the clarity of our perceptual experiences. Their argument is primarily first-personal, because the clarity of our experiences is supposed to be something apparent by introspection.

As I use 'belief', it is a representation that is normally accessible for explicit reasoning and is often responsible for behavior. Compared to a perceptual experience, a belief falls later in the stream of activity and has a more direct relation to behavior. When we ask someone to tell us what they believe, we're using 'belief' in this way. Psychologists sometimes use 'belief' differently. They sometimes use it as a synonym for 'representation', and will thus ask, for example, what a part of the visual cortex believes. If you can hear 'belief' only in this way, feel free to instead use 'reportable representation'.

As I use 'confidence', it is the kind of probability that figures in decision-making. Probabilities of this kind are often called "subjective" to distinguish them from the objective probabilities that are independent of any decision-maker's perspective. Like all subjective probabilities, confidences have a loose but important connection to the axioms of probability theory (for more discussion, see Morrison, 2016, pp. 21, 34–35). One consequence is that, if your perceptual experience assigns confidence to two propositions, there must be a ratio that approximately describes the amount that it assigns to each. It thus isn't enough for perceptual experiences to involve an ungraded representation of uncertainty such as "possibly" or "maybe."

Given how I'm using 'confidence', Perceptual Confidence doesn't place any limitations on *which* propositions our experiences assign confidence to. But that's not to say that there aren't plausible limits. For example, I think it's implausible that our perceptual experiences assign confidence to the proposition that we made a correct decision ("decision confidence"). Even if our perceptual experiences are the result of earlier decisions in the brain, our experiences are not themselves *about* those decisions. We perceive trail markers, cars, friends, and cups of coffee, not our decisions about those objects. (In the psychological lingo: our experiences aren't "metacognitive.") In most cases, I think it's equally implausible that our experiences assign confidence to the propositions that a given feature would cause a given experience—for example, that a blue trail marker would cause this experience rather than another experience, that a green trail marker would cause this experience rather than another experience, and so on. We perceive trail markers, cars, friends, and cups of coffee, not the propensities of those objects to cause our current experience. Even if our perceptual experiences are the result of earlier assignments of confidence to propositions about such propensities, our experiences are not themselves about them. (In the decision theory lingo: our experiences don't seem to be about their own "likelihood.") For these reasons, I think Perceptual Confidence is most plausibly developed as the view that our perceptual experiences assign confidence to propositions about the properties and relations of objects in our immediate environment, such as the colors, motions, locations, and flavors of trail markers, cars, friends, and cups of coffee. And that's the version of the view that I'll explore.

Psychologists often use 'confidence' differently. They often use it as a synonym for 'decision confidence' (see Denison 2017; Morrison 2017). If you can hear 'confidence' only in this way, feel free to replace it with 'subjective probability'.

As I use 'represents' and 'assigns', a state represents that p only if p is then easily accessible for inference, i.e., to transitions at the computational level. Likewise, a state assigns confidence to p only if p and that confidence are then easily accessible for inference. I'm thus using 'assign' and 'represents' to help us understand the computational structure of the brain.

Of course, there is an important further question about what it is for something to be "easily accessible" for inference. This question will be central to my subsequent arguments, and I'll return to it later. There is also an important further question about what counts as a transition at the computational level. I'll say a bit more about this question later, but I hope everything I say will be compatible with any reasonable refinement of this concept. My strategy is to focus on states that contribute to behavior, because it's easier to reconstruct their computational role.

One final clarification: To say that a perceptual experience assigns confidence to a proposition is not to deny that assignments of confidence are fundamentally relative. For example, a perceptual experience might assign .75 confidence to the proposition that a sign is blue because, more fundamentally, it assigns three times more confidence to the proposition that the circle is blue than to the proposition that the circle is not blue.

3 | POST-PERCEPTUAL CONFIDENCE

There are many alternatives to Perceptual Confidence. I'm going to focus on what I take to be the most popular alternative: that our experiences represent one or more propositions but do not assign them confidence. I call this Post-Perceptual Confidence.[5]

To help introduce Post-Perceptual Confidence, let's return to our example involving the turning car. A proponent of Post-Perceptual Confidence might say that your kinesthetic experience represents that the car is turning at a specific angle (e.g., 90 degrees). If you *report* low confidence in that angle, they'll say it's just because you believe that your experience is unreliable. Perhaps you believe that experiences of this kind are in general unreliable (your "likelihood"), or perhaps you believe that this particular experience is unreliable because cars aren't likely to turn at that angle (your "prior"). Analogies might be helpful: Even if a car's speedometer says it's moving at exactly 5mph, you might report low confidence because you believe that the speedometer is in general unreliable, or because you believe that you rarely move at that speed. Even if a weatherman predicts rain, you might report uncertainty because you believe that the weatherman is in general unreliable, or because you believe that it almost never rains at this time of year. According to this proponent of Post-Perceptual Confidence, low confidence at the level of belief has a similar origin.

As an alternative, a proponent of Post-Perceptual Confidence might say that your kinesthetic experience represents that the car is turning at an angle between some minimum and maximum (e.g., between 80 and 100 degrees). If you *report* higher confidence that it is some angles rather than other angles in this range, they'll say it's because you believe that some angles were more likely to produce an experience of that range (your likelihood), or because of your prior beliefs about the angles of most intersections (your prior). Thus, once again, uncertainty at the level of belief is due to other beliefs.

For our purposes, it won't matter whether a proponent of Post-Perceptual Confidence says that your kinesthetic experience represents that the car is turning at a *specific* angle or within some *range* of angles. For our purposes, what's important is what they say about where confidence is assigned in the brain. With this issue in mind, let's introduce the two versions of Post-Perceptual Confidence that will be our main focus.

According to the first version, the processes that transform sensory inputs into behavioral outputs rely on assignments of confidence, but those distributions are *discarded* before perceptual experience. Perhaps you perceive that the car is turning at whichever angle was assigned the most confidence by those early perceptual processes. Or perhaps you perceive that the car is turning at an angle within some range, where the range includes all the angles assigned a sufficient amount of confidence by those early perceptual processes, such as all angles within one standard deviation of the mean. Confidence might then be reintroduced at the level of belief. Perhaps you assign less

[5] Proponents of Post-Perceptual Confidence include many of the authors listed in footnote 2. While I take it to be the default view among philosophers of perception, many haven't felt the need to explicitly endorse it.

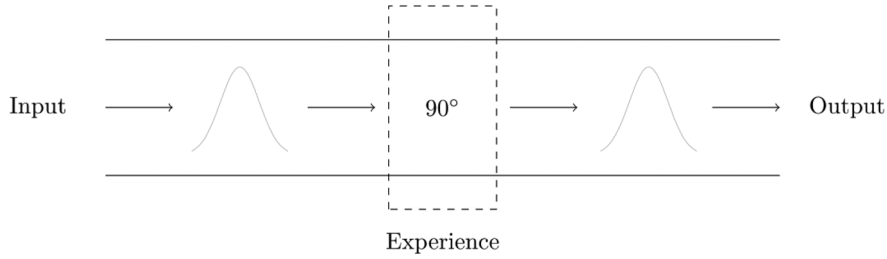

FIGURE 2 First version of Post-Perceptual Confidence

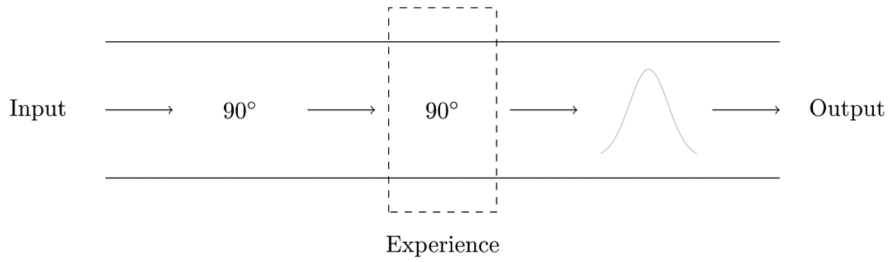

FIGURE 3 Second version of Post-Perceptual Confidence

than full confidence because of evidence that your perception is inaccurate or unreliable. Figure 2 depicts this first version.

According to the second version, the processes that transform sensory inputs into perceptual experiences represent angles without assigning them any confidence. Perhaps they represent only one angle, the angle that is later represented in experience. Figure 3 depicts this second version.

I'm going to focus on these versions of Post-Perceptual Confidence because I think they're the most appealing and, I suspect, also the most popular. What they have in common is that they imply that our behavior is informed by our perceptual experiences.

But there are other versions of Post-Perceptual Confidence. According to these other versions, our behavior results from a stream of activity that bypasses our perceptual experiences. According to one such version, assignments of confidence in early perceptual processing aren't *discarded*. They're instead *rerouted*. While perceptual experiences represent the relevant propositions, the assignment of confidence is made available to behavior through another channel. Our perceptual experiences are like non-probabilistic snapshots of the probabilities that often drive behavior, and we have access to both kinds of information when deciding how to behave. In many cases, we simply ignore our perceptual experience and rely entirely on the rerouted probabilistic information. Figure 4 depicts this third version.

According to another variant, confidences are assigned by the neural activity underlying our perceptual experiences, without being assigned by our perceptual experiences themselves. Our perceptual experiences emerge from activity that is probabilistic without themselves being probabilistic. Figure 5 depicts this fourth version.

What these last two versions have in common is that they imply that our behavior is informed by early, perceptual assignments of confidence that are not included in our perceptual experiences. I'll address these variants of Post-Perceptual Confidence later, towards the end of the paper. Until then, let's focus on the first two variations.

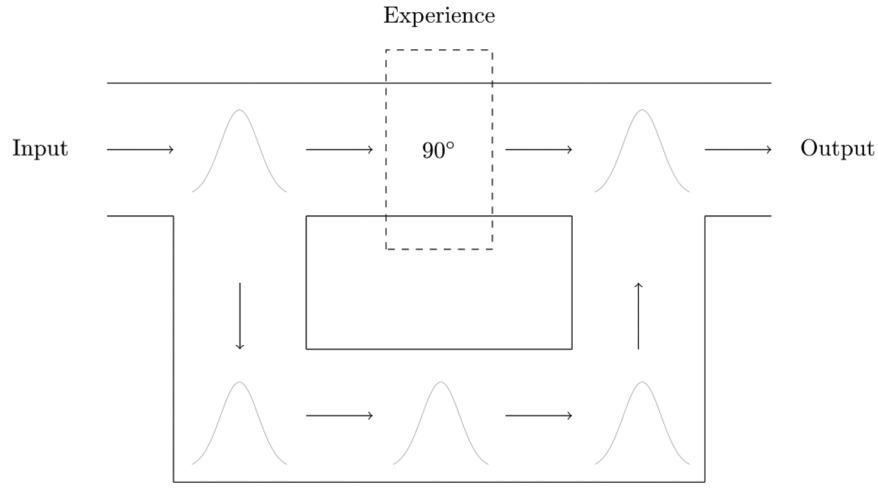

FIGURE 4 Third version of Post-Perceptual Confidence

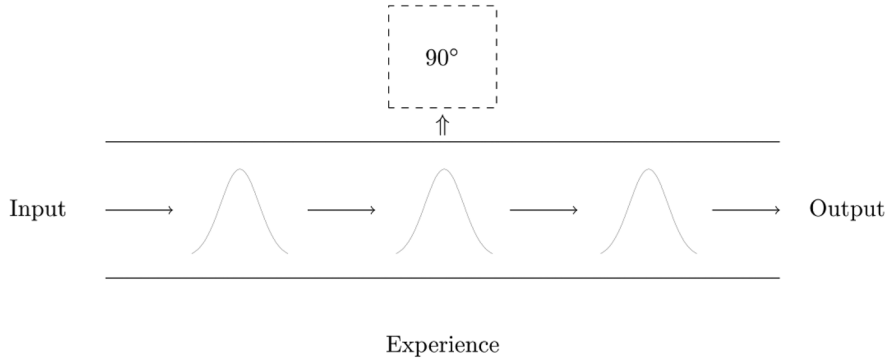

FIGURE 5 Fourth version of Post-Perceptual Confidence

4 | CONTINUITY ARGUMENT

How might we use third-personal evidence to motivate Perceptual Confidence over Post-Perceptual Confidence? I think the best strategy is to rely on what I'll call a 'continuity argument'. Suppose we can show that confidences that are assigned by early perceptual processing are also available for behavior. This is a reason to think that the relevant confidences are assigned by the perceptual experience itself. Why? If that confidence didn't arise until after the perceptual experience, we wouldn't expect to find it in the brain earlier than the perceptual experience. And if that confidence were discarded, we wouldn't expect to find it later. Once information is lost, it cannot be regained. In computer science, this is called the "data processing inequality."

According to the continuity argument, if there's evidence that (1) confidence was assigned before a perception experience and (2) that same confidence was assigned after that perceptual experience, the best explanation is that the same confidence is in the experience itself.

A figure might help convey the structure of this argument. (1) and (2) would establish the schema depicted in Figure 6

According to a continuity argument, the best explanation is the schema depicted in Figure 7.

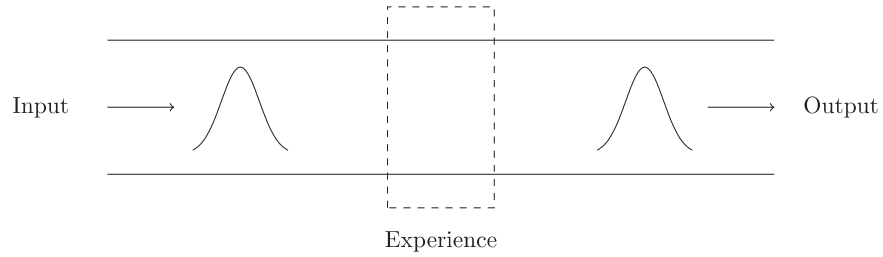

FIGURE 6 Premise of continuity argument

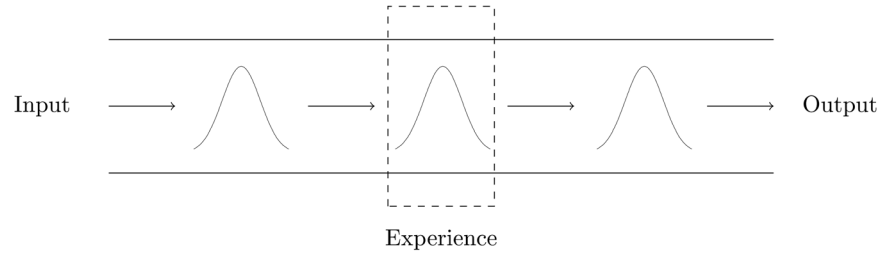

FIGURE 7 Conclusion of continuity argument

Strictly speaking, (1) and (2) are more demanding than what a continuity argument requires. With respect to (1), it would be enough for the confidence to be assigned by the neural activity that underlies the perceptual experience. It thus needn't occur *before* the perceptual experience. It could be simultaneous. With respect to (2), it would be enough for the confidence assigned after the perceptual experience to *draw on* the confidence assigned before it. It might take into account further background beliefs about the reliability of the process (a likelihood) or about the environment (a prior). One way to establish that the confidence assigned after the perceptual experience is drawing on the confidence assigned before it would be to show that pre-experience confidence and post-experience confidence vary together, trial by trial. But this kind of evidence would need to be handled with care, because we would then need to rule out the possibility that it is varying with changes in background beliefs rather than with changes in the pre-experience confidence. It would be better if we could find cases in which the same assignment can be found both before and after the perceptual experience.

Notably, continuity arguments do not take a stand on *how* confidence arises in the brain. Confidence could arise from feedforward processes, as suggested by our figures. But feedback from later regions could also play a role. All that's important for continuity arguments is that the confidence is assigned in regions that bookend whichever region gives rise to perceptual experience.

Also note that continuity arguments don't take a stand on the sense in which neural activity *gives rise* to perceptual experience. They are strongest if perceptual experience is identical to neural activity, so that these are just different ways of describing the same event. But, as we'll see, they can accommodate weaker assumptions as well.

Like most arguments to the best explanation, continuity arguments aren't decisive. Even if (1) and (2) are true, it's at least in principle possible for confidence to be assigned before and after the perceptual experience but not by the experience itself. For example, it's possible that it's a *coincidence* that the same confidence was assigned after the perceptual experience. It's also possible that the confidence assigned before the perceptual experience was used to generate a non-probabilistic

representation that later processes used to generate the same assignment of confidence. But even if these are possible explanations, they aren't the best explanation. Analogously, suppose two friends are standing side-by-side in front of you, and you tell a secret to the friend on the left. Suppose that, a minute later, your friend on the right repeats the secret back. It's possible that she correctly guessed the secret. But that seems unlikely, especially if she's able to reliably repeat back more secrets. It's also possible that your friend on the left didn't actually *say* the secret to your other friend, but rather told her something that let her infer it. But a simpler, and therefore better, explanation is that the first friend repeated the secret to the second friend.

What kind of evidence should we look for? It might be tempting to rely on behavioral evidence. After all, there *is* compelling evidence that our behavior on perceptual tasks often depends on assignments of confidence. In particular, our behavior is responsive to rewards and background information in a way that strongly suggests that we're relying on assignments of confidence. But this behavioral evidence *by itself* doesn't help us choose between Perceptual Confidence and the first version of Post-Perceptual Confidence, because it doesn't indicate *where* confidence arises and is discarded in the stream of activity. Confidence could be discarded *before* perceptual experience or arise only *after* perceptual experience. Thus, the behavioral evidence doesn't seem to help.

It might therefore be tempting to rely on *fMRI evidence* (e.g., van Bergen et al., 2015). But fMRI data has well-known shortcomings. Of particular concern to us is that each data point in fRMI data (a voxel) sums the activity of at least 10,000 neurons. Information that is distributed across tens of thousands of neurons might not yet be accessible for computation, and therefore not be "assigned" in our sense. Analogously, economic data about the Great Depression that is distributed across thousands of books and articles in a library is not yet easily accessible to an economist trying to identify the Depression's cause. She would first need to collect and synthesize the evidence. Moreover, fRMI data places a lot of weight on how information is spatially organized, and it's unclear when that's sufficient for the information to be accessible for computation. Just to be clear: I think that fMRI evidence is helpful and important. I just don't think that this evidence *by itself* is likely to provide us with the kind of evidence we would need to motivate Perceptual Confidence. Likewise for EEG and MEG evidence.

I'm thus going to focus on *single-unit* and *population* recordings of neural activity. This evidence has drawbacks as well, but I'm more optimistic that they can be overcome. I'll say more about these drawbacks later, when considering specific experiments. I first want to say more about the two premises of the continuity argument. Let's consider them one by one, underlining the terms most in need of clarification.

(1) Confidence was assigned before perceptual experience

To better understand what I mean by 'assign', consider the more familiar distinction between what's represented and what's merely implicit. *p* is represented only if *p* is then easily accessible for computation. *p* is *merely implicit* when *p* could ultimately be used to make decisions and guide behavior, but further inferences and background information are required. Think again of all the data about the Great Depression. It represents facts about the Great Depression without representing the cause of the Great Depression. Nonetheless, the cause might be implicit in the data, because a brilliant economist might be able to work it out. Likewise, the identity of a burglar might be implicit in a crime scene, the solution to a crossword might be implicit in the clues, and the theorems of geometry might be implicit in its axioms.

There's a similar distinction between confidence that's assigned and confidence that's merely implicit. The meteorologist's barometer represents a pressure without assigning any confidence to the proposition that it will rain. But confidence might still be implicit in its measurement, because a meteorologist could use that measurement to assign confidence to that proposition. Confidence that is merely implicit isn't yet "assigned" in our sense. To be assigned, it must be easily accessible for computation.

Why is this an important distinction? If the confidence is not easily accessible for computation, we might just be locating the non-probabilistic representations that the brain *later* uses to generate a probability distribution. We might just be locating the neural analog to barometer measurements and economic data. For the continuity argument to work, we need to show that the confidence itself, rather than the representations later used to generate it, is assigned both before and after the perceptual experience.

As our examples suggest, the distinction between what's represented/assigned and what's implicit doesn't just apply to the brain. It is perhaps sharpest when applied to sentences. The most famous example is from Grice (1961, p. 129). Suppose that an academic letter of recommendation merely reports, "Jones has beautiful handwriting." What's represented is a claim about the student's handwriting. What's implicit is that the student is unqualified. To extract this from what was said, we'd have to draw an inference that relies on background information, in this case information about the conventions of letters of recommendation. The distinction is thus between what is "right there," easily accessible in the sentence, and what requires "too many" additional inferences and/or "too much" background information.

This isn't a perfectly sharp distinction, because it's often unclear what counts as "too much" additional information and "too many" additional inferences. To see why, let's consider some hard cases. Suppose you say, "My cat is being spayed." What you said *presupposes* that your cat is female. Is it represented? It's not clear, because it's not clear whether it requires "too many" additional inferences and/or "too much" additional information. For our next example, suppose you say, "The tablecloth is scarlet." What you said *semantically entails* that the tablecloth is red. Is it represented? It's not clear, because, once again, it's not clear whether it requires "too many" additional inferences and/or "too much" additional information. Finally, suppose you say, "There are three rows with four dots each." What you said *mathematically entails* that there are twelve dots. Is it represented? This is a hard case. On the one hand, it's a simple inference, and it involves minimal background information. On the other hand, it does involve more inferential work than the information that there are three rows. (If you think this is explicit, is it explicit that there are an even number of dots? That there are fewer than fifteen dots?)

Thus, the distinction between what's represented/assigned and what's implicit is somewhat fuzzy. For brains, the distinction is even fuzzier. There are several reasons. First, we can't appeal to the intentions of a sender and a receiver, which at least helps constrain what might be represented by a sentence. Second, we can't appeal to conventions, because whereas there are conventions that link words to what they represent, there are no conventions that link neural activity to what it represents. Third, whereas it is clear which utterances and marks represent (e.g., "excellent handwriting" but not "e@n%$k"), it's not clear which aspects of neural activity represent. Is it spikes per second? Or do the intervals between the spikes matter too? And how many neurons should we consider at once? 10,000? 100? Just one? These are still very much open questions. Fourth, in linguistics there are straightforward tests, such as cancelability ("Jones has excellent handwriting, which isn't to say he's unqualified"), that cannot be applied to neurons.

As a result, whereas there are obvious and uncontroversial examples of linguistic representation which help us distinguish it from what's merely implicit, there are few examples of neural

representation which help us distinguish it from what's merely implicit. As a result, the distinction between what's represented in the brain and what's merely implicit is even fuzzier. The same goes for what's assigned.

How, then, can we use third-personal evidence to show that confidence is assigned in an early brain region? The best evidence would show that downstream neural areas are actually using it to perform probabilistic inferences. We could then be sure that the assignment of confidence was explicit. But we're a long way from collecting such evidence. Among other obstacles, it's not even clear which neural areas are performing probabilistic inferences. Even when we know that the brain must be performing probabilistic inferences, they can be hard to locate.

For now, I think we need to rely on behavioral decoders. What are behavioral decoders? They are functions that take neural activity as input and then output a behavioral prediction. For example, a behavioral decoder might take as input the activity of certain neurons in a subject's visual cortex, and then output a trial-by-trial prediction about whether that subject will report that the stimulus is tilted rightward rather than leftward. If a behavioral decoder can successfully predict a subject's behavior, it might seem reasonable to infer that the relevant variable is represented by that neural activity.

There's a helpful (if imperfect) analogy with what Quine (1960) called radical translation. In radical translation, the challenge is to assign contents to a person's utterances on the basis of what they're perceiving and their subsequent behavior. For example, suppose that when there are storm clouds approaching, a person usually says "blerg" and then reaches for her umbrella. A reasonable hypothesis is that "blerg" means that it's about to rain. Behavioral decoders are similar. If there is neural activity that, trial by trial, allows us to predict a person's trial-by-trial behavior, a reasonable hypothesis is that the relevant neurons are representing the variable that drives the person's trial-by-trial behavior. In essence, we're thinking of those neurons as speaking a language we don't yet understand.

Behavioral decoders and radical translation are both risky in an important respect: it's hard to locate the relevant vehicle in the chain of inference. For example, "blerg" might mean that it's raining, but it might also belong earlier or later in the chain of inference. It might mean that there are rain clouds approaching, or it might express the intention to reach for her umbrella. Behavioral decoders confront a similar difficulty. Suppose we show someone "2+2 = …." We might be able to predict their response ("4") using neural activity in the visual cortex. But that doesn't mean that the activity in the visual cortex is explicitly representing 4. It might just be representing the shapes that make up "2+2 = …." That is, it might just be representing the evidence that the brain reliably uses to generate the response. The decoder is able to predict the person's response just because there's a straightforward mapping from these shapes to the behavioral response, given that most people are good at simple arithmetic. The decoder could learn that mapping by learning to associate the input "2+2" with the output "4," or by learning the numerical value of "2" and then calculating the output. Likewise, suppose we show a chess grandmaster images of relatively straightforward board positions. A behavioral decoder might be able to use the activity in their visual cortex to predict the move they'll recommend, but that doesn't mean their move is already explicitly represented in the visual cortex. It's just that, given a representation of the board position, a decoder with a built-in chess calculator will be able to predict the grandmaster's choice.

What's the solution? We should look for behavioral decoders where the mapping from neural inputs to behavioral outputs is as simple as possible, and rely on as little extra information as possible. Otherwise, it will undercut our claim that the decoder is revealing what's *represented* or *assigned*. Ideally, the decoding should be something that the brain could do, which is to say

that it should rely on biologically plausible operations. We should also look for decoders that predict behavior as well as possible, because that will give us some indication that we've really identified the information driving the behavior, rather than just a subset of the information driving the behavior, or information about something else that just happens to correlate with it. If the information is already in a format from which the brain could easily extract whichever variable is driving its trial-by-trial behavior, and that tightly correlates with the behavior, it is reasonable to suppose that information is represented in that region.

There's a lot more to say about behavioral decoders and how they help us distinguish between what's represented and what's merely implicit. But these are tricky issues, hard to sort out in the abstract. Let's return to them while discussing specific experiments.

How can we be sure that the assignment of confidence occurs in activity that *precedes* perceptual experience? We often can't be certain, because, as I stressed before, we don't know exactly where perceptual experience arises in the brain. But we can be pretty sure it's not in anatomically early areas of the visual cortex, for example. So, any information we find there very likely precedes perceptual experience. Later regions are trickier and need to be considered on a case-by-case basis. In general, then, the earlier the region, the better.

Let's now consider the other premise:

(2) That same confidence was assigned after perceptual experience

It's much easier to show that confidence was assigned after the perceptual experience, because we can show that it has a direct connection to behavior. For example, as a direct measure of a subject's assignment of confidence to propositions about various orientations, we can simply *ask* her to report her subjective probability about those propositions. We can also ask her to make a decision about the stimulus's orientation, such as whether it's rightward or leftward, and then ask her how confident she is that she made a correct decision. As an indirect measure, we can provide asymmetrical rewards and see what effect that has on her decisions. We can also give her the opportunity to "opt out" of the task for smaller, guaranteed rewards of varying sizes. Many of these methods also work on monkeys, and in some cases on dolphins, birds, and rodents (for an overview, see Smith, 2009).

5 | FIRST EXPERIMENT

There's far less neural evidence than you might expect. In large part, this is because we don't yet have a basic understanding of how the brain computes. We're still sorting out basic questions. But there are three experiments that I think are worth considering in detail.

There are some notable similarities between these experiments: They all involve monkeys, because the relevant recordings are too invasive to make in humans. They all involve a decision between two categories, because it is relatively easy for monkeys to learn tasks involving that kind of decision. And they all involve stimulus direction, because it is relatively easy to find neurons that respond differentially to that kind of stimulus.

The first experiment is from Walker et al. (2020). The neural recordings are from V1, a macaque brain area with a straightforward and well-known human homologue. It's at the back of your head, near the top of your spine.

The stimuli were gratings with varying orientations. Stimuli were chosen from one of two categories ($C = 1$, $C = 2$). For each orientation (e.g., 20°), Figure 8 depicts the probability that a stimulus would have that orientation for each the category.

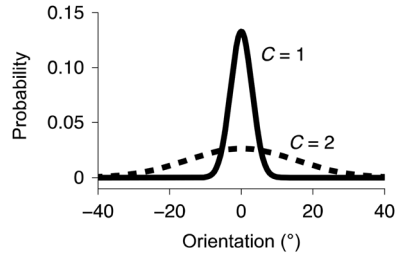

FIGURE 8 Probability of each orientation for each category. Adapted from Walker et al. (2020, Figure 2a).

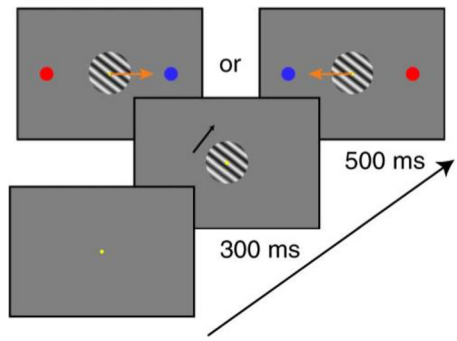

FIGURE 9 The monkeys task. Reprinted from Walker et al. (2020, Figure 2b).

Stimuli chosen from the first category (solid line) were more likely to have orientations between −5 degrees and 5 degrees, though would sometimes have lower or higher orientations. Stimuli chosen from the second category (dotted line) were less likely to have orientations between −5 degrees and 5 degrees as stimuli chosen from the first distribution, and were more likely to have lower or higher orientations.

During training, monkeys learned that there was an equal probability that the next stimulus was drawn from category one or category two (that is, they learned to use a flat prior over the two categories). On each trial, a monkey was shown one stimulus. The monkey's task was to indicate whether that stimulus was drawn from category one or category two. They indicated their decision by looking at a blue dot for category one and at a red dot for category two (see Figure 9). The locations of the red and blue dots were varied randomly across trials. Data was collected from two monkeys, and the total number of trials was around 300,000.

Before considering how moneys performed on this task, let's consider which approach is *optimal*. To choose between category one and category two, one must rely on neural responses in early sensory areas. That is, one must rely on one's "measurements." You might think that the optimal approach is to choose category one when a measurement represents the stimulus's orientation as falling between the intersections of the two probability distributions, and otherwise to choose category two. And if measurements of the stimuli's orientations were perfectly reliable, that would be the optimal approach. But measurements are never *perfectly* reliable—they are always noisy, which is to say that there is some degree of randomness in them. The same stimulus will produce different measurements on different occasions. For a given stimulus, there is a probability distribution over the measurements it produces. That distribution indicates the probability of getting

one's measurement from a stimulus with an orientation of −10, or an orientation of −8, and so on. These are objective rather than subjective probabilities, which is why I'm not calling them "confidences."

Given that the monkey's measurements of orientation are noisy, what's the optimal approach? It depends on the *amount* of noise. The greater the noise, the further out the criterion should be moved. Assuming that the probability distribution over possible measurements of orientation is normal (i.e., Gaussian), this means that as the variance of that distribution increases, the further out the criterion should be moved. This can be proven mathematically, but the intuition is clear enough: if there's a lot of noise in your measurements, and your measurement is just outside the intersection of the two distributions, it's more likely category one, because there are far more stimuli from category one with nearby orientations that could have produced that measurement.

Walker et al. manipulated measurement noise by lowering and raising the contrast of the stimulus. When the contrast of a stimulus is lower, there's more variation in a monkey's measurements of the stimulus's orientation; in particular the probability distribution over the monkey's measurements has a greater variance. They demonstrated that monkeys use the optimal approach in that, as measurement noise increased, each monkey's criterion moved outward (see their Figure 2d). Their data establishes that the monkeys relied on an estimate of measurement noise in their behavior. Because it's an estimate, it involves subjective probabilities and is therefore an assignment of confidence.

To show that this estimate was made in early visual processing, Walker et al. recorded from 96 neurons in V1. Their decoder then successfully predicted the monkeys' decisions as their criterion shifted trial by trial.

How successful was the decoder? It predicted the monkey's behavior in around 80% of low-contrast trials and around 90% of high-contrast trials (see their Extended Data Figure 2). Of particular interest: it was more successful than decoders that assumed a constant amount of measurement noise across all trials (what they call "Fixed Uncertainty Models"). That is, it predicted the monkey's behavior better than a decoder that did not vary its estimate of measurement noise. Importantly, it was able to predict the monkey's behavior across trials with the same stimulus, and thus with the same contrast. This is important, because it strongly suggests that they didn't just find activity that happens to correlate with the contrast of the stimulus and plays no role in the monkey's decision. They found the measurement that is driving the monkey's decision.

Does this establish that confidence is assigned *in V1*? As noted earlier, it depends on the decoder. Let's therefore pause for a moment to better understand this decoder. It was an artificial neural network consisting of an input layer, two hidden layers, and an output layer. Each node in the input layer was given the spike rates of a different weighted subset of the neurons. Each node then performed a nonlinear operation, outputting another number. Likewise, at the second level, each node took as input a weighted subset of the outputs of the nodes at the first layer and performed another nonlinear operation. Those outputs were then linearly combined and outputted at the final layer, generating an estimate of measurement noise. Walker et al. were able to use that estimate of measurement noise to predict the monkey's criterion and thus its behavior. There were between 400 and 1,000 nodes in each of the hidden layers, the exact number varying with contrast.

To what degree does this experiment support premises (1) and (2) in the continuity argument? Let's consider them separately.

(1) Confidence was assigned before perceptual experience.

Walker et al. found activity in an early visual area that contributes to the monkey's behavior on a task involving uncertainty. It is highly probable that the activity precedes perceptual experience. Representations of color and size constancies are an important part of our perceptual experiences, and they don't fully emerge until after V1 (Roe et al., 2012).

But there are three respects in which their experiment falls short of decisively supporting (1). First, they might not have decoded confidence. They manipulated the amount of noise by manipulating the contrast level. As a result, there was a correlation between the amount of noise and the contrast level. It was therefore possible to predict the monkey's estimate of noise, and thus its confidence, from its estimate of contrast. For this reason, Walker et al. might have just found the brain's estimate of contrast—that is, the evidence that it *later* used to assign confidence. Their decoder might have used this evidence to predict confidence that was not yet assigned. Alternatively, the monkey's decision might not have relied on an assignment of confidence at all. It might just have learned a direct mapping from estimates of orientation and contrast to the two categories. Notably, the monkey's estimate of contrast can vary even when the actual contrast is fixed. This might explain why Walker et al. were able to predict the monkey's decision across trials with the same actual contrast.

How could this issue be resolved? The experiment could be modified so that other variables, such as distance, size, and duration, are all used to manipulate the amount of noise. If we are still able to predict the monkey's behavior, that would demonstrate that we did not just find an estimate of contrast, because the monkey's estimate of contrast by itself would no longer be enough to predict its behavior. If the activity in V1 responds to contrast but not the other variables, we might try recording from downstream areas in the visual cortex. Thus, future evidence might more effectively support (1).

Second, even if they decoded confidence, it was not the right kind of confidence. One kind of confidence indicates the probability that one's measurement would have been caused by an orientation of −10, −8, and so on. This would be an estimate of measurement of noise. Bayesians would call it a "likelihood" over orientations. Another kind of confidence indicates the probability that the stimulus has an orientation of −10, −8, and so on. Bayesians would call this a "posterior" over orientations. We're looking for this second kind of confidence, because the confidences assigned by our perceptual experiences seem more like a posterior than a likelihood. Which kind of confidence did Walker et al. decode?

They decoded a likelihood over orientations. Notably, in order to train their decoder, they used that likelihood to calculate a posterior over orientations. Their calculation assumed the "true" prior. They then adjusted their decoder to bring the resulting posterior closer to the "true" posterior (see Methods, Full-likelihood decoder). In a sense, then, they also decoded a posterior over orientations. But this was just for the purposes of training their decoder. They did not assume that the monkey's brain calculated it. The monkey's decision might have resulted from a posterior over categories that was calculated directly from a likelihood over orientations and a prior over categories, and thus without calculating a posterior over orientations.

How could we decode a posterior over orientations? We should train the decoder without assuming any particular prior, including the "true" prior. Otherwise, we might have just found a likelihood over orientations. We should also use the decoder to predict the monkey's decision on tasks with different categories. Otherwise, we might have just found a posterior over categories. It we cannot find a posterior over orientations in V1, we might try recording from downstream areas in the visual cortex. Thus, once again, future evidence might more effectively support (1).

Third, they might not have decoded an *assignment* of confidence, because the confidence might not yet be accessible for inference. As noted earlier, whether a behavioral decoder gives us

sufficient evidence that confidence is assigned depends on whether that decoder relies on too much background information or on too many inferences. What about Walker et al.'s decoder?

It didn't rely on any background information. It was trained using a subset of the data from the experiment, and then tested against the remaining data. Thus, their decoder doesn't rely on "too much" background information.

It's less clear whether their decoder depends on "too many" inferences. There are only two hidden levels in the neural network, with between 400 and 1,000 nodes at each level. This is much simpler than the network used by the facial recognition software on your computer. The number of nodes might sound like a lot, but it shouldn't cause concern. Keep in mind that the brain has roughly 100 billion neurons. Any information that could be extracted using fewer than 1,000 neurons should count as easily accessible for the brain. Also, note that their decoder made a prediction on the basis of only 96 neurons. That's like trying to predict the outcome of a presidential election on the basis of polling only one city block. To extrapolate to the country as a whole, you would need to rely on a complicated algorithm that took into account gender, age, income, etc. With data from more people, including people from other parts of the country, you might be able to use a simpler algorithm. Likewise, with data from more neurons in V1, Walker et al. might have been able to use a simpler decoder. Finally, keep in mind that while *we* need the neural network to transform the information into a format that *we* can use to predict the monkey's criterion, the brain might not. The relevant information might already be in a format that the brain can use to adjust the criterion.

Still, a simpler decoder with fewer levels and nodes would be better. As is, we should be slightly worried that it would take the brain too many inferences to mirror the computations of the neural network. Even two hidden layers can implement functions that move the output far beyond its inputs. In fact, with enough nodes, they can implement *any* continuous mathematical function, no matter how complex (Hornik 1991).

Thus, while Walker et al. have given us some initial evidence in support of (1), further experiments and analyses would be helpful.

(2) That same confidence was assigned after perceptual experience.

Walker et al. provide no evidence that the same confidence was assigned *after* the monkey's perceptual experience. But the experiment could be modified so that it might provide such evidence. For example, we could ask the monkeys to make a relatively simple "bet" that would indicate their confidence in their decision, and see how well their bets correlated with the confidence decoded from V1. If there's a sufficiently high correlation, that would give us evidence that the confidence assigned in V1 is still accessible after their perceptual experience. Thus, with respect to this premise too, future evidence might provide more effective support.

In the meantime, it's worth noting that there is an *indirect* reason to suspect that this confidence is still accessible. Suppose, for the sake of argument, that it was discarded *before* a monkey's perceptual experience. Given the results of the experiment, the monkey's decision about whether the stimulus belongs to category one or category two must also have occurred before the monkey's perceptual experience, because that decision took their confidence into account. In that case, the monkeys must have *perceived* that the stimulus belongs to category one or category two. It's hard to know what monkey experiences are like, but that's not what *our* experiences are like. We perceive the orientation of the stimulus. We then decide *after* our perceptual experience whether it belongs to category one or category two. For us, these categories are introduced in cognition, not

in perception. If that's true of the monkey's experience as well, then the confidence must have been accessible after its experience.

Thus, while Walker et al. didn't support (2), future experiments might, and there is an indirect reason to think that they will.

6 | SECOND EXPERIMENT

The second experiment is from Fetsch et al. (2011). They recorded from the dorsal medial superior temporal area (MSTd), a region adjacent to the visual cortex. Its closest homologue in the human brain is a subdivision of the complex of regions known as MT+.

There's an important difference between this second experiment and the first experiment: In the first experiment, the monkey's decision was based on only one kind of measurement, namely its visual measurement. In the second experiment, the monkey's decision is based on measurements of two kinds, namely both its visual measurement and its vestibular measurement. This is called "cue combination." There are many examples from everyday life. For example: we decide whether a doorknob is circular by looking at it and feeling it; we decide whether a classmate is talking by looking at her and listening to her; we decide whether kimchi is spoiled by smelling it and tasting it; we decide whether the boardwalk is slanted by looking at it and balancing on it. In ordinary cases, these are independent measurement of the same stimulus—for example, of the same doorknob shape. In such cases, our measurements correct and amplify each other. But it's interesting to consider what happens when, unbeknownst to us, they are measurements of discrepant stimuli—for example, when we are looking at and balancing on floors with different slants.

Perhaps surprisingly, humans and other animals have been shown to integrate our measurements in a way that's *near-optimal*. What does that mean? The optimal way to integrate measurements is to weight each of them by an accurate estimate of their relative noise and then take the average. For example, suppose that, on a given kind of trial, a subject's visual measurement is significantly noisier than her vestibular measurement, so that the relative noise in her visual measurement is .6 and the relative noise in her vestibular measurement is .4. The optimal way for the subject to integrate these two measurements is to weight her visual measurement by .6 and her vestibular measurement by .4 and then take the average. This is a paradigmatic example of a probabilistic inference, because it relies on an estimate of noise, in particular an estimate of the variance of the probability distribution over measurements of that kind of stimulus. It is thus an application of Bayes's Theorem. In a number of different experiments, the performance of humans and other animals has been shown to be near-optimal in that is approximates the performance of a subject using the optimal approach (e.g., Young, Landy, & Maloney, 1993; Ernst & Banks, 2002; Hillis et al. 2004).

In this experiment, monkeys were placed on a moving platform with a monitor. After a forward motion, they were asked whether the motion was slightly leftward or slightly rightward. The platform gave them a vestibular measurement of the direction of their motion. The screen gave them a visual measurement of the direction of their motion. In many trials, their actual motion and the visual motion on the screen were consistent. But on some trials, they were discrepant. In particular, on some trials they were shown visual motion in a slightly different direction.

In earlier work, this group established that monkeys behaved near optimally on this task (Fetsch et al., 2009). By default, monkeys place far more weight on their visual measurements, because visual measurements are far less noisy. But when the monkey's visual measurements became

noisier, they put proportionally less weight on those measurements. In these experiments, they didn't add noise to the vestibular measurements, though this could in principle be done (e.g., by vibrating the platform while it's moving).

How did they add noise to the monkey's visual measurements? By adjusting the coherence of the dots used to display the direction of motion. In particular, the monkeys were shown visual stimuli at 60% and 16% coherence. At 60% coherence, 60% of the dots were moving in the same direction, while 40% of the dots were moving randomly. At 16% coherence, 16% of the dots were moving in the same direction, while 84% of the dots were moving randomly. These coherences were chosen because at 60% coherence, monkeys gave more weight to their visual measurements, and at 16%, they gave more weight to their vestibular measurements. The researchers didn't add noise to the monkeys' vestibular measurements, but there was always some noise, in part because the platform's motion wasn't perfectly smooth. The behavioral data (from Fetsch, 2009) establishes that monkeys relied on an estimate of measurement noise in their behavior. Otherwise, their behavior wouldn't have been near-optimal.

In an attempt to find where these measurements are weighted and combined in the brain, the researchers recorded from MSTd. Unlike in the last experiment, they individually recorded each neuron's response to a stimulus, and then later combined all their recordings to compute the average neural activity of all the neurons in response to that kind of stimulus. They recorded from 108 neurons in total: 60 neurons in one monkey and 48 neurons in a second monkey.

They found neurons that were, to varying degrees, responsive to both visual and vestibular stimuli as well as their relative amounts of noise. Their decoder used the activity of these neurons to predict each monkey's behavior. They found that the amount of actual weight the monkeys placed on their vestibular information was, on average, very close to the amount of weight predicted by the decoder on the basis of the activity of the MSTd neurons (see their Figures 2a,b and 6e,g).

What kind of decoder did Fetsch et al. use? Unlike Walker et al., they didn't use a neural network. They instead calculated the Bayesian maximum a posteriori (MAP) of direction based on the following: the assumption that leftward and rightward were equally likely on each new trial (i.e., a flat prior over categories); information about the tuning curves of all the neurons; information about the modality of the input to those neurons (visual, vestibular, both); and the actual coherence of the visual stimuli (16%, 60%) (see "Likelihood-Based Decoder: Assumptions and Caveats" in the online supplemental material.)

Does this experiment support the two premises of the continuity argument?

(1) Confidence was assigned before perceptual experience.

In at least one respect, the findings of Fetsch et al. better support (1) than the previous study. Recall that Walker et al. might have just decoded a representation of contrast that was later used to assign confidence. Fetsch et al. better supports (1) in this respect. They decoded a MAP estimate from activity that *already* reflected a weighting of both measurements by the amount of noise in them. For example, on trials in which dot coherence was higher (60%), the activity of the relevant neurons already reflected that more weight was placed on the monkeys' visual measurement than their vestibular measurement. This seems to establish that confidence must have been assigned, because a MAP estimate is a paradigmatic example of a probabilistic computation and, by definition, confidence is assigned when it is accessible for probabilistic computation. Even if the visual system initially just represented dot coherence, those representations seem to have already been used to assign confidence.

For the same reason, we don't need to worry about the details of their decoder. A simple decoder is important when searching for representations that are easily accessible for probabilistic inference. The decoder's simplicity is evidence that the representations are easily accessible. When searching for representations that are the *result* of a probabilistic inference, the complexity of the decoder isn't important. Even a complex decoder can be evidence that the inference has already taken place.

There's an important limitation on Fetsch et al.'s decoder that's worth mentioning, however. Fetsch et al. recorded from one neuron on each trial. They repeated the same stimulus many times in order to record how each neuron responded to that stimulus. Their decoder tried to predict the monkey's behavior using an *aggregate* of all these recordings. As a result, their decoder was limited in an important way: it couldn't rely on *correlations* between the responses of different neurons on the same trial. The brain, however, might be relying on such correlations (Walker et al raise this worry on p. 127 of their 2018 preprint). Fortunately, this doesn't undermine the support this experiment provides for (1). Even if their experiment didn't provide insight into the activity responsible for the probabilistic inference, it still established that the inference occurred, and that's what's important.

But there are four other respects in which this study falls short of decisively supporting (1). First, we're looking for an assignment of confidence. While they *decoded* an assignment of confidence, they only verified one feature of the decoded confidence: whether more confidence was assigned to heading directions greater than 0° or less than 0°. But that could be decoded even if there is no assignment of confidence. It's possible, for example, that MSTd just includes a representation of an estimate (say 0–5°) and that their decoder transforms this representation into an assignment of confidence only to collapse it back to the original estimate (0–5°). In that case, there is no reason to think there really is confidence in MSTd. The decoded confidence might just be an artifact of their decoder, and thus might not correspond to anything probabilistic in MSTd. Further experiments might help better support (1) by verifying features of the assignment of confidence that couldn't be derived from a mere estimate.

Second, we're looking for an assignment of confidence *over heading directions* (e.g., 0–5°, 5–10°). They decoded the *mean* and *variance* of a distribution of confidence over heading directions. Whether that's enough for MSTd to assign confidence over heading directions (e.g., 0–5°, 5–10°) depends on whether the confidence assigned to those directions is easily accessible for computation. This is a tricky issue, because it depends on the *format* of the mean and variance estimates. As an illustration, consider the number *four hundred and forty-one*. If it were written in base ten (*441*), further inferences would be required to determine if it is divisible by seven. But when it is written in base seven (*1200*), we know immediately that it is divisible by seven, because the last digit is *0*. Likewise, depending on the format of the estimate of the mean and variance, it might be immediately apparent how much confidence is assigned to each range of orientations, or figuring that out might require a lot more computation. Further experiments might help better support (1) by verifying the confidences assigned over other heading directions (e.g., 5–10°).

Third, while MAP estimation is a paradigmatic example of a probabilistic computation, it is hard to empirically distinguish MAP estimation from other, non-probabilistic computations, such as loss minimization (for details, see Lippl et al., manuscript). Further experiments might better support (1) by demonstrating that the relevant representations can also be used in marginalization, change of variables, or another probabilistic computation that is distinguishable from non-probabilistic alternatives.

Fourth, they might not have decoded confidence that was assigned *before* perceptual experience. Whereas it should be uncontroversial that activity in V1 occurs before perceptual experience,

MSTd is later and therefore less clear-cut. Nonetheless, it is adjacent to and receives direct input from the visual cortex. Moreover, there is an abundance of evidence that its immediate predecessor, MT/V5, contributes to perceptual experience (for an overview, see Block, 2005, p. 46). So, while recordings from V1 would have been preferable, this should not be a source of serious concern.

(2) That same confidence was assigned after perceptual experience.

Fetsch et al. established that the activity in MSTd resulted from a weighted average of the monkey's visual and vestibular measurements. They did not establish that the monkey subsequently had access to the weights placed on each measurement—that is, to the estimates of their relative reliabilities. They thus didn't rule out the possibility that the assignment of confidence was *discarded* before the monkey's perceptual experience (as in our second figure). Their experiment therefore didn't provide any evidence for (2). But future experiments might be able to support it. For example, we could use one of the experimental paradigms already mentioned to probe the monkey's confidence confidence in their combined estimate. If they have access to the pre-experiential assignment of confidence, we should expect, on average, lower confidence in response to discrepant stimuli.

Thus, while Fetsch et al. didn't decisively establish (1) and (2), future experiments might.

7 | THIRD EXPERIMENT

The third experiment is from Kiani and Shadlen (2009) and builds on earlier experiments from Shadlen's lab (summarized in Gold & Shadlen, 2007). They recorded from each monkey's lateral interparietal cortex (LIP). The homologue in humans is somewhere in the posterior parietal cortex. It projects to the motor system that controls eye movements.

In this task, monkeys were shown a large number of dots, some of which were moving coherently rightward or leftward, and the rest of which were moving randomly. The monkeys were prompted to indicate whether the coherent motion was rightward or leftward. They indicated their choice by looking at a target to the right (to indicate rightward motion) or left (to indicate leftward motion).

Approximately 70 neurons in LIP were selected, because of their responsiveness to either leftward or rightward motion. Neurons of both types seemed to "accumulate" evidence in that they gradually increased their activity the longer they were exposed to motion of a certain type. The activity of "rightward" neurons was inversely correlated with the activity of "leftward" neurons. Figure 10 simulates the total activity of rightward neurons on a trial. The line at the top indicates the point at which that activity suffices for a decision that the motion is rightward.

Figure 11 simulates the total activity of rightward neurons on another trial.:

Why doesn't the activity increase by a constant amount? Sometimes the neurons receive a lot of evidence of rightward motion. Sometimes they receive only a little evidence. And sometimes they receive conflicting evidence. The amount and type of evidence depends on measurement noise as well as the behavior of the dots that happen to be in the receptive field of each neuron.

In other work by this group, they were able to predict a lot about each monkey's decision on the basis of this neural activity. In particular, they were able to predict the *accuracy* of the monkey's decision as well as the *timing* of the monkey's decision (Roitman & Shadlen, 2002). Moreover, they used the simplest possible decoder: a *threshold* that predicted the monkey would decide that the motion was rightward or leftward very soon after the activity crossed that threshold. Given that

FIGURE 10 Activity of rightward neurons on a trial.

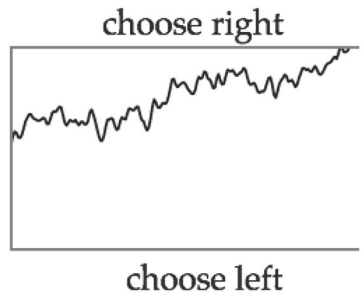

FIGURE 11 Activity of rightward neurons on a different trial.

their decoder is so simple, it's reasonable to assume that the brain is using this very mechanism to make its decision.

A further manipulation allowed them to also predict the monkey's *confidence* in their decision. In the earlier experiments, Shadlen and colleagues asked monkeys to choose between two options: that the coherent motion is rightward, and that the coherent motion is leftward. Kiani and Shadlen gave monkeys an "opt out" option. It didn't offer as great a (juice) reward as correctly choosing the direction of overall motion, but it was a sure thing, and thus a good decision if they were uncertain about the overall motion. They found that the activity in LIP predicted how the monkey would respond when forced to make a decision before activity reached the threshold. Basically, monkeys would "opt out" when the activity above wasn't close enough to the threshold (also taking into account the amount of time the monkey had been looking at the stimulus). Significantly, this information let the researchers predict when the monkey would opt out, even when comparing stimuli of equal difficulty. This suggests that they found the activity that is responsible for the decision, rather than activity that merely correlates with a feature of the stimulus.

Like Fetsch et al., they recorded from one neuron at a time, and then aggregated their recordings to estimate the activity of all 76 neurons to a given stimulus. Unlike Fetsch et al., they also showed that even a single neuron was enough to predict the monkey's decision, albeit weakly (see their Figure 3).

Let's consider the two premises of the continuity argument.

(1) Confidence was assigned before perceptual experience.

In several respects, Kiani and Shadlen better support (1) than either of the previous studies. To start, their behavioral decoder was the simplest possible: it was just a threshold of activity. As

such, it didn't rely on any background information and didn't require any inferences. The decoder would predict an immanent decision as soon as the activity in the recorded neurons surpassed the threshold. We might think of it as a neural network containing just two nodes, one for leftward and one for rightward, each performing a simple step function. In contrast, Walker et al.'s neural network contained thousands of nodes, and Fetsch et al.'s decoder relied on sophisticated mathematical computations.

Further, Kiani and Shadlen found confidence, not just the evidence that later regions might use to assign confidence. In particular, they found neurons *performing* a probabilistic inference. As noted in our discussion of cue combination, combining evidence that is weighted by reliability is a paradigmatic example of a probabilistic inference. Evidence accumulation does just that. The only notable differences are that the cues are combined gradually over time and the number of cues is much greater, given that each measurement is treated as an independent cue. These differences in no way make evidence accumulation a less paradigmatic example of probabilistic inference.

Finally, Kiani and Shadlen found the right kind of confidence, namely confidence assigned to the direction of dot motion. In particular, on other experiments, they found that the threshold adjusted with the prior (Hanks et al., 2011). For example, the monkey relied on a lower threshold for deciding that the motion is rightward when there were proportionally more trials with rightward motion. The threshold also lowered gradually within a trial, in effect placing more weight on the prior when the measurements alone weren't enough to a decision. This was accomplished by gradually increasing the baseline activity of the relevant neurons. Thus, they found confidence assigned to the direction of dot motion, a posterior.

All that being said, there is an important and perhaps obvious respect in which they do not provide decisive evidence for (1): they do not provide evidence that confidence is assigned *before* perceptual experience. LIP is directly involved with behavior, specifically eye movements. In this task, the monkey is moving its eyes to indicate its decision. As LIP is a premotor region, it is natural to assume its activity is a consequence of the monkey's perceptual experience, and thus not itself responsible for that experience. In support of this assumption, consider another experiment from the same lab. Kira et al. (2015) used an experiment similar to the experiment we just considered, except that different shapes—triangles, pentagons, etc.—gave the monkey different amounts of evidence in favor of selecting one target rather than the other. The lab found similar ramping activity in LIP. But when we look at the shapes, we just see the shapes, not their evidential relation to the targets. It's natural to assume that the monkeys have similar experiences. Thus, whereas activity in V1 *unquestionably* precedes perceptual experience and activity in MSTd *plausibly* precedes perceptual experience, putting one's faith in LIP can seem like a desperate gamble.

But there is a case to be made using first-personal evidence. When we look at a dot motion stimulus, we eventually seem to *perceive* the motion as rightward or leftward. In some cases, the overall motion of the dots snaps into view. In other cases, our decision feels like a guess, but a guess rooted in our perceptual experience—we seem to be reiterating its uncertainty. There are intermediate cases too, cases in which the overall direction of dot motion doesn't snap into view but feels like more than a guess. This at least suggests that confidence is assigned during our perceptual experience. If the representation of overall motion didn't occur until after our perceptual experience, we wouldn't see any direction of motion. We would just see dots, and the decision about their overall direction of motion would be more cognitive, like the decision that a piece of furniture is expensive, or that a knight is pinned by a bishop. If the representation was sometimes in the perceptual experience and sometimes after it, there wouldn't be intermediate cases.

Given that the assignment of confidence seems to be in our perceptual experience, why do we find it in LIP? One explanation is that LIP is drawing on activity that occurred earlier. There is

tentative evidence of this. In one study, chemical inactivation rendered LIP useless, and task performance wasn't significantly disrupted (Katz et al., 2016).[6] Perhaps this earlier activity is enough to run a continuity argument.

But there's another explanation. Maybe we're wrong to assume that perceptual experience occurs *before* LIP. Perhaps our perceptual experience arises, in part, from activity in LIP. According to enactive views of perceptual experience, the function of perceptual experience is to enable us to choose between actions (see, e.g., Shadlen et al., 2008). If these views are correct, we should expect the neural correlates of consciousness to be in regions of the brain where decisions between possible actions are made, and LIP seems to be such an area. In that case, the results described in this section would provide direct support for Perceptual Confidence without any need for a continuity argument. They would directly show that the neural activity responsible for perceptual experience assigns confidence.

(2) That same confidence was assigned after perceptual experience.

The monkeys' behavior on the opt-out task indicates that they have access to the confidence assigned in LIP, so this experiment gives us extremely strong evidence for (2).

Let's step back. We considered three experiments in support of the continuity argument's two premises, (1) and (2). None of the experiments decisively support both premises. But this isn't due to a systematic obstacle. It is instead due to a haphazard assortment of gaps in the experiments, many idiosyncratic to one particular experiment. Future experiments might be able to fill in the gaps, and we were able to describe what many of those experiments might look like. We might even be able to stitch together the strengths of all three experiments into a more decisive experiment. What might such an experiment look like? Like all three experiments, it might involve a task in which the monkey's decision takes into account the reliability of its own measurements. Unlike in the first experiment but like in the second and third experiments, we would look for the activity in the brain that reflects the output of a probabilistic inference, and not just the representations later used in a probabilistic inference. Unlike in the first and second experiments but like in the third experiment, we would look for behavior (such as "opt out") that indicates that the monkey has access to an assignment of confidence and not merely a point estimate. Like in the first and second experiments but unlike in the third experiment, we would look for that activity in areas that more people will agree precede the monkey's perceptual experiences. As technology improves and we're able to simultaneously record from more neurons in more regions of the brain, such an experiment might become easier to design and execute.

There's a second respect in which our experiments weren't decisive. Recall that the continuity argument is an argument to the best explanation, not a logical deduction. As such, its premises do not logically entail its conclusion. Most significantly, its premises leave open the possibility that there is an assignment of confidence before and after perceptual experience, but experience itself is non-probabilistic. Perhaps our experiences represent the possibility assigned the most confidence by earlier processes along with enough clues for later processes to reconstruct that distribution. The relevant clues might include contrast, blur, and distance. In that case, our

[6] This isn't decisive evidence, because the chemical inactivation might have changed the way the monkey made its decision. Also, Zhou and Freedman (2019) provide tentative counterevidence. In a slightly more complicated version of the task, they found that inactivating LIP degraded sensory and motor aspects of decision-making, with the greatest effects on the sensory aspect.

perceptual experience would include these clues, and thereby allow later processes to assign the same confidence, without our experiences themselves assigning confidence.

To use third-personal evidence to decisively rule out this possibility, we would need to identify the activity that gives rise to perceptual experiences and then establish that it assigns confidence. That's unlikely to happen in the near future, given that we know so little about how the brain gives rise to consciousness. In the meantime, we could study the size of the divergence between the distribution before experience and after experience. The smaller the distortion, the less likely it is that the distribution was discarded and then reconstructed, because that would introduce new sources of error, widening the distortion.

There are also two methodological reasons to place less credence in this alternative explanation. First, it would be an odd and inefficient way for the brain to process representations. In particular, if a representation is required by later processes, it would be odd for the brain to discard it before it reaches those processes, and inefficient to require those processes to reconstruct it on the basis of non-probabilistic clues. It would be like trying to run a company in which people speak to each other only in riddles. While the brain is odd and inefficient in many respects, it's many successes should lead us to assume, at least as a default, that it doesn't have such a structure. Second, if this kind of possibility were given as much credence in the absence of confirming evidence, it would be hard to make progress in neuroscience. We couldn't make inferences about the flow of representations through the brain, given that there are often intermediate areas that might be discarding and then reconstructing those representations. We would need to wait for detailed whole-brain recordings so that we could follow the representation from one region to the next. But that's too demanding. Even without such recordings, we seem able to make the reasonable assumption that, when we find a representation in two separate but connected regions, the representation was transmitted from one to the other without being discarded and reconstructed.

Stepping back even further, our discussion supports four conclusions. In order of increasing strength, they are:

- At present, there is no decisive third-personal evidence for Perceptual Confidence.
- In principle, there might be such evidence.
- In the near future, we might be able to collect such evidence.
- In the meantime, our evidence gives Perceptual Confidence an intermediate level of support.

8 | OTHER VARIANTS

We focused on variants of Post-Perceptual Confidence in which early assignments of confidence are *discarded* before perceptual experience. We focused on these variants because they seem to be the most credible. To see why, let's consider variants in which early assignments of confidence aren't discarded and instead contribute to behavior in a way that's independent of perceptual experience. While the reasons to reject these variants are less empirical than the reasons to reject the previous variants, I think they are nonetheless more compelling.

According to the first of these variants, assignments of confidence in early perceptual processing are *rerouted* around our perceptual experience. While our perceptual experience just represents one possibility, or range of possibilities, the confidence is made available to behavior through another channel (see Figure 4).

There is precedent for this view. The ventral stream is often said to include conscious, object-centered representations, while the dorsal stream is said to include unconscious, viewer-centered

representations (Goodale & Milner, 1992). Perhaps there is likewise a stream for conscious, non-probabilistic representations and another stream for unconscious, probabilistic representations.

There are two problems. First, if there were two channels, we would expect to find examples of double dissociation. By interfering with the lower channel, it would be possible to temporarily or permanently interrupt a person's access to the assignment of confidence, without any impact on their perceptual experience. They would report no change in their perceptual experience, but their performance on perceptual tasks requiring access to the confidences would deteriorate. Likewise, by interfering with the upper channel, it would also be possible to temporarily or permanently interrupt a person's perceptual experience without interrupting their access to the assignment of confidence. The result would be what we might call "probabilistic blindsight," because they would perform just as well on many tasks, despite a lack of perceptual experience. There is evidence of such dissociations of the ventral and dorsal streams (see again Goodale & Milner, 1992). As far as I know, however, there are no documented cases of either dissociation with respect to confidence. Moreover, we would expect the two channels to normally output *different* estimates, due to noise in both channels. But, once again, I'm not aware of any empirical evidence to that effect. In general, I think that we should avoid postulating multiple channels until there's evidence of them, or at least a reason to expect them.

Second, in the examples I described, it doesn't seem like the assignment of confidence is coming from an unconscious, perceptual source. In blindsight, people say things that suggest the information seems to be coming through another channel. To them, it feels like an *urge* that's independent of their perceptual experience. The familiar examples I listed at the start aren't like that. If you're asked to decide the color of the trail marker, the angle of your turn, the location of your friend, or the taste of your coffee, your decision seems to draw on your conscious perceptions. If your consciousness were taken away, it's hard to imagine how you could come to the same decision in the same way. When we're asked to make a decision about a stimulus that's fast, masked, or unattended, it is somewhat plausible that consciousness doesn't play a role in decision-making. But in ordinary cases, that's hard to accept. If the assignment of confidence seems to originate before our beliefs, it doesn't seem to come from out of nowhere. It seems to come from our perceptual experience. I thus think that the introspective, first-personal evidence strongly counts against this variant of Post-Perceptual Experience.

It might be possible to refine this view so that it avoids some of these problems. Gross and Flombaum suggest that perceptual experiences might be samples from an underlying, unconscious probabilistic representation (see Gross & Flombaum, 2017, p. 384; Gross, 2018, Section 5a). That would explain why we can't interfere with the unconscious assignment of confidence without disrupting the perceptual experience. But probabilistic blindsight should still be possible, because it should still be possible to interfere with the perceptual experience without interfering with the unconscious assignment of confidence. The uncertainty would also be coming from an unconscious perceptual source, rather than just our perceptual experiences and/or our beliefs, and that's not how it seems.

Siegel (2020) suggests a different refinement. She suggests that the perceptual experience represents the *mean* of the confidence assignment and the unconscious process represents its *variance*. That would give perceptual experiences a central role in perceptual decision-making. But we would still expect it to be possible to interfere with the unconscious process without any effect on the perceptual experience. Our performance on tasks requiring access to the confidences would deteriorate, because we couldn't access the variance, while our perceptual experiences remained the same. We would also expect it to be possible to interfere with the perceptual experience without disrupting the unconscious representation of confidence. When presented with a grating at

low contrast, we couldn't report its orientation. We might even deny seeing it. But we could still indicate our level of confidence, because we would still have access to the variance. As far as I'm aware, there is no evidence that any of this is possible.

According to a second variant, we should think of our perceptual experiences as *emerging from* the relevant neural activity without being identical to it. Perhaps the confidences assigned by that neural activity are like the colors and shapes of those neurons—they aren't among the properties shared by the neural activity and the experiences that emerges from them (see Figure 5). Because there's only one channel, this variant doesn't have the problems of the last variant.

But it has its own problems. If perceptual experiences have too little in common with their underlying neural activity—perhaps because they merely correlate with it—they might not be causes of behavior. It would be the activity of the neurons, rather than the representations in the perceptual experience, that cause the monkey to select one target rather than another. (This is the familiar "exclusion problem" for dualism, see Robb and Heil 2021, Section 6.2.) On the other hand, if perceptual experiences have a lot in common with their underlying neural activity—enough for them to not count as the same event—it is unclear why perceptual experiences wouldn't assign confidence. Even if the underlying neural activity doesn't share its size, shape, and speed with the perceptual experience, it does share its representational properties. For example, if the neural activity represents a color, shape, face, distance, or direction, so does the perceptual experience. Why wouldn't it also share its confidence? It seems arbitrary to insist that representational properties in general are shared between neural activity and perceptual experiences, but not confidence.

There are still other variants of Post-Perceptual Confidence.[7] There might also be refinements of these variants that avoid or minimize the problems I listed. But I hope this is enough to explain why we focused on the variants that we did.

9 | CONCLUSION: POTENTIAL CONSEQUENCES FOR PSYCHOLOGY AND NEUROSCIENCE

We already listed some of Perceptual Confidence's consequences for psychology and neuroscience. Let's conclude by listing others.

First, Perceptual Confidence might help psychologists and neuroscientists identify the neural activity that gives rise to consciousness. Given Perceptual Confidence, we should look for regions capable of assigning confidence. Learning more about the kinds of neural activity capable of assigning confidence could significantly narrow our search.

Second, Perceptual Confidence might challenge popular methods for studying metacognition. In experiments involving categorization that allow opting out, humans, monkeys, birds, dolphins, and rodents opt out more often when the stimulus is at the border of two categories and they are less reliable at categorizing it. Many take this as evidence that humans, monkeys, etc., are

[7] Siegel (2020) suggests a variant I don't consider. She suggests that perceptual experiences represent a point value and merely *dispose* us to form probabilistic representations with that point value as the mean. This disposition is supposed to vary independently of perceptual experience, including its phenomenology and what it represents. I'm skeptical of this variant. One could similarly suggest that our perceptual experiences merely dispose us to form beliefs about colors and do not represent them. More generally, one could deny that perceptual experiences represent anything, and instead merely dispose us to form certain beliefs. Some have endorsed this view. But most don't. If this response isn't acceptable in general but is acceptable for confidence, we need to be given a reason. I can't think of one. (For relevant discussion, see my 2016, p. 31–32, 35–36.)

capable of representing their own reliability, a kind of metacognition (e.g., Smith, 2009). But, given Perceptual Confidence, they might just be relying on their perceptual confidence. That is, they might just be relying on a first-order, perceptual assignment of low confidence to both categories, rather than a higher-order, cognitive judgment about their own reliability. Metacognition might be less central to decision-making than is often supposed.

Third, Perceptual Confidence might generalize beyond perception to imagistic, or perception-like, memories. As an illustration, try to remember the color of the tablecloth at breakfast this morning. A natural assumption is that you either remember the color of the tablecloth or you don't, just as a painting of the tablecloth either includes a color or it doesn't. But there's another possibility: your memory might assign confidence to a range of propositions, including the proposition that it was spinach green and the proposition that it was olive green. Perhaps your memory assigns more confidence to the first proposition than the second. If perceptions can be uncertain, perhaps imagistic memories can too. In some cases, forgetting might result from a gradual dilution of confidence rather than incremental deletion of detail.

This view of memory might have implications for its role in decision-making. A well-known reason to rely on multiple memories is that it provides better estimates (Hertwig & Erev, 2009; Gershman & Daw, 2017). For example, suppose you want to order a slice of pie at your favorite café. To choose the most delicious pie, you might rely on memories of past slices. Relying on multiple memories would be a good way to compensate for random variation between those slices. For example, it would help compensate for the fact that, when you ate a slice of the blueberry pie on Tuesday, there were an unusually large number of blueberries in it. If memories can be uncertain, there's another reason why relying on multiple memories in decision-making might be helpful: it can compensate for the uncertainty within each memory. By combining multiple, uncertain memories of slices from the same pie, you would end up with a more certain overall estimate of the pie's deliciousness. Analogously, by combining the predictions of multiple uncertain meteorologists, each relying on independent evidence, you can end up with high confidence that it will rain. Memories might have a similar role in decision-making.

Perceptual Confidence thereby provides a new and interesting framework for future research in the mind sciences.[8]

REFERENCES

Beck, J. (2020). On perceptual confidence and 'completely trusting your experience'. *Analytic Philosophy*, 61, 174–188. https://doi.org/10.1111/phib.12151

van Bergen, R.; Ma, W.; Pratte, M.; & Jehee, J. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, 18, 1728–1730. https://doi.org/10.1038/nn.4150

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Science*, 9, 46–52. https://doi.org/10.1016/j.tics.2004.12.006

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B*. https://doi.org/10.1098/rstb.2017.0341

[8] I am grateful for feedback from Weiji Ma's lab meeting at NYU in December 2018, the Eastern APA in January 2019, Steve Fleming's lab meeting at UCL in May 2019, a conference at Ruhr-Universität Bochum in July 2019, a conference at Cambridge University in August 2019, the Inter-American Congress in Bogotá in October 2019, and an interdisciplinary conference at the Israel Institute for Advanced Studies in January 2020. I am also grateful for feedback from Jake Beck, Mazviita Chirimuuta, Raphael Gerraty, Niko Kriegeskorte, Ian Phillips, Susanna Siegel, and Jonathan Cohen and his reading group at UCSD. Special thanks to Weiji Ma, Mike Shadlen, and especially Chris Fetsch for reviewing my discussions of their experiments. Finally, I would like to thank all of my friends and mentors at Columbia's Zuckerman Mind Brain Behavior Institute. Without their help, I couldn't have written this paper.

Byrne, Alex (2021). Perception and probability. *Philosophy and Phenomenological Research*, 104, 343–363. https://doi.org/10.1111/phpr.12768

Cheng, T. (2018). Post-perceptual confidence and supervaluative matching profile. *Inquiry*, 65, 249–277. https://doi.org/10.1080/0020174X.2018.1562370

Clark, A. (2018). Beyond the 'Bayesian blur': predicting processing and the nature of subjective experience. *Journal of Consciousness Studies*, 25, 71–87.

Denison, R. (2017). Precision, not confidence, describes the uncertainty of perceptual experience. *Analytic Philosophy*, 58, 58–70. https://doi.org/10.1111/phib.12092

Denison, R.; Block, N.; & Samaha, J. (2022). What do models of visual perception tell us about visual phenomenology? In F. De Brigard & W. Sinnott-Armstrong (Eds.), *Neuroscience and Philosophy* (pp. 241–283). Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/12611.003.0014

Ernst, M. & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415, 429–433. https://doi.org/10.1038/415429a

Fetsch, C.; Pouget, A.; DeAngelis, G.; & Angelaki, D. (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, 15, 146–154. https://doi.org/10.1038/nn.2983

Fetsch, C.; Turner, A.; DeAngelis, G.; & Angelaki, D. (2009), Dynamic re-weighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience*, 29, 15601–15612. https://doi.org/10.1523/JNEUROSCI.2574-09.2009

Gershman, S. & Daw, N. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual Review of Psychology*, 68, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin Harcourt.

Gold, J. & Shadlen, M. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Goodale, M. & Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15, 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Green, D. & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Grice, P. (1961). The Causal Theory of Perception. *Aristotelian Society Supplementary Volume*, 35, 121–168. https://doi.org/10.1093/aristoteliansupp/35.1.121

Gross, S. (2018). Perceptual consciousness and cognitive access from the perspective of capacity unlimited working memory. *Philosophical Transactions of the Royal Society B*, 373. https://doi.org/10.1098/rstb.2017.0343

Gross, S. (2020). Probabilistic representations in perception: are there any, and what would they be? *Mind & Language*, 35, 377–389. https://doi.org/10.1111/mila.12280

Gross, S. & Flombaum, J. (2017). Does perceptual consciousness overflow cognitive access? The challenge from probabilistic, hierarchical processes. *Mind & Language*, 32, 358–391. https://doi.org/10.1111/mila.12144

Hanks, T.; Mazurek, M.; Kiani, R.; Hopp, E.; & Shadlen, M. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience*, 31, 6339–6352. https://doi.org/10.1523/JNEUROSCI.5613-10.2011

Hertwig, R. & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13, 517–523. https://doi.org/10.1016/j.tics.2009.09.004

Hillis, J.; Watt, S.; Landy, M.; & Banks, M. (2004). Slant from texture and disparity cues: optimal cue combination. *Journal of Vision*, 4, 967–992. https://doi.org/10.1167/4.12.1

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4, 251–257. https://doi.org/10.1016/0893-6080(91)90009-T

James, W. (1890). *Principles of Psychology*, Volume I. New York: Dover Books, 1950.

Katz, L.; Yates, J.; Pillow, J.; & Huk, A. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, 535, 285–288. https://doi.org/10.1038/nature18617

Kiani, R. & Shadlen, M. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324, 759–764. https://doi.org/10.1126/science.1169405

Kira, S.; Yang, T.; & Shadlen, M. (2015). A Neural Implementation of Wald's Sequential Probability Ratio Test. *Neuron*, 85, 861–873. https://doi.org/10.1016/j.neuron.2015.01.007

Kouider, S. & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 857–875. https://doi.org/10.1098/rstb.2007.2093

Laasik, K. (2021). Perceptual confidence: A Husserlian take. *European Journal of Philosophy*, 29, 354–364. doi: https://doi.org/10.1111/ejop.12580

Lippl, S.; Gerraty, R.; Morrison, J.; & Kriegeskorte, N. (manuscript). Source invariance and probabilistic transfer: a testable theory of probabilistic representation.

Morrison, J. (2016). Perceptual confidence. *Analytic Philosophy*, 57, 15–48. https://doi.org/10.1111/phib.12077

Morrison, J. (2017). Perceptual confidence and categorization. *Analytic Philosophy*, 58, 71–85. https://doi.org/10.1111/phib.12094

Moss, S. (2018). *Probabilistic Knowledge*. Oxford: Oxford University Press.

Munton, J. (2016). Visual confidences and direct perceptual justification. *Philosophical Topics*, 44, 301–326. https://doi.org/10.5840/philtopics201644225

Nanay, B. (2020). Perceiving indeterminately. *Thought*, 9, 160–166. https://doi.org/10.1002/tht3.454

Quine, W. (1960). *Word and Object*. Cambridge, MA: The MIT Press.

Raleigh, T. & Vindrola, F. (2021). Perceptual experience and degrees of belief. *The Philosophical Quarterly*, 71, 378–406. https://doi.org/10.1093/pq/pqaa047

Robb, D. & Heil, J. (2021). Mental causation. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2021/entries/mental-causation/

Roe, A.; Chelazzi, L.; Connor, C.; Conway, B.; Fujita, I.; Gallant, J.; Lu, H.; & Vanduffel, W. (2012). Toward a unified theory of visual area V4. *Neuron*, 74, 12–29. https://doi.org/10.1016/j.neuron.2012.03.011

Sandberg, K. & Overgaard, M. (2015). Using the perceptual awareness scale (PAS). In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research* (pp. 181–195). Oxford: Oxford University Press.

Shadlen, M.; Kiani, R.; Hanks, T.; & Churchland, A. (2008). Neurobiology of decision making: an intentional framework. In C. Engel & W. Singer (Eds.), Better Than Conscious? Decision Making, the Human Mind, and Implications *for Institutions* (pp. 71–101). Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/9780262195805.003.0004

Shea, N. (2020). Representation in cognitive science: replies. *Mind & Language*, 35, 402–412. https://doi.org/10.1111/mila.12285

Siegel, S. (2020). Explaining uncertainty. *Mind & Language*, 37, 134–158. https://doi.org/10.1111/mila.12348

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

Vance, J. (2020). Precision and perceptual clarity. *Australasian Journal of Philosophy*, 99, 379–395. https://doi.org/10.1080/00048402.2020.1767663

Walker, E.; Cotton, R.; Ma, W.; & Tolias, A. (2018). A neural code for probabilistic computation in visual cortex. *bioRxiv*. https://doi.org/10.1038/s41593-019-0554-5

Walker, E.; Cotton, R.; Ma, W.; & Tolias, A. (2020). A neural code for probabilistic computation in visual cortex. *Nature Neuroscience*, 23, 122–129. https://doi.org/10.1038/s41593-019-0554-5

Wu, W. (2018). The neuroscience of consciousness. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2018/entries/consciousness-neuroscience/

Young, M.; Landy, M.; & Maloney, L. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research*, 33, 2685–2696. https://doi.org/10.1016/0042-6989(93)90228-O

Zhou, Y. & Freedman, D. (2019). Posterior parietal cortex plays a causal role in perceptual and categorical decisions. *Science*, 365, 180–185. https://doi.org/10.1126/science.aaw834

How to cite this article: Morrison, J. (2023). Third-personal evidence for perceptual confidence. *Philosophy and Phenomenological Research*, 1-30. https://doi.org/10.1111/phpr.12951