

Consciousness and Metacognition (II)

Higher-Order Theories, Perceptual Reality Monitoring,
and a Challenge from Phenomenal Overflow

Plan for today

1. Reminder: higher-order theories of consciousness
2. Why we need a *discriminator*
3. Michel & Lau's Perceptual Reality Monitoring (PRM)
4. Adding a theory of qualitative character: Quality Space
5. Evidence and predictions
6. An indirect challenge from Block (and a reply from Beck)

The transitivity principle

Core intuition (Rosenthal).

A mental state is conscious only if one is in some way conscious *of* it.

- Distinguish *state* consciousness (a state's being conscious) from *transitive* consciousness (being conscious of something).
- Transitivity: a state is conscious only if its bearer is transitively conscious of it.
- This is the common starting point of higher-order (HO) theories.

Two motivations

From common sense. Armstrong's inattentive driver: long monotonous drive, "come to" and realise one has driven for miles without consciousness of the road. First-order perceptual processing seems to occur *without* consciousness of the perceiving.

From science. Blindsight, masked priming, and many other paradigms: sensory representations can be tokened without being conscious.

⇒ A first-order representation, however rich, is not by itself sufficient. Something extra is needed.

The higher-order hypothesis

Higher-order theorists propose that the “something extra” is a *higher-order representation*: a representation *about* the first-order state.

- HOT (Rosenthal): the higher-order state is a *thought*.
- HOP (Lycan, Carruthers): a higher-order *perception* or quasi-perceptual monitoring.
- HO-Misrepresentation (Brown): the HO state can be tokened even when the targeted FO state is not (“empty HO state”).
- Self-representational views (Kriegel): one and the same state represents the world *and* itself.

What HO theories typically claim *not* to do

Most HO theorists do not claim that HO mechanisms exhaust the qualitative character of experience.

- Qualia, on this picture, are typically reduced to *representational content* of the first-order state.
- What HO theories add is the explanation of why some such representations are *conscious* rather than unconscious.
- Two questions, often run together but conceptually distinct (Pautz 2010):
 - ▶ *The consciousness question*: why are some states conscious and not others?
 - ▶ *The quality question*: why does a conscious state feel the way it does?
- HO theories primarily target the first.

A standard worry: over-intellectualism

- HO theories seem to require sophisticated *thoughts* about one's own mental states.
- But infants, animals, distracted adults appear conscious without anything like reflective introspection.

Standard HO replies:

- The HO state need not be conceptually rich; non-conceptual or thin meta-representations may suffice.
- HO monitoring is typically *sub-personal* and automatic, not introspective.
- Distinguish consciousness from *introspection*: introspection is voluntary and effortful; HO monitoring is not.

Where Michel and Lau enter

Granting the HO framework, two questions remain wide open:

1. **Mechanism.** What exactly does the HO system *do*, and how could it be implemented?
2. **Function.** *Why* would a cognitive system have such a mechanism in the first place?

Lau's *Perceptual Reality Monitoring* (PRM) gives a unified answer.

Three reasons neuronal activity *cannot* be trusted at face value

Neurons fire for many reasons *other* than the actual presence of what they typically represent.

- **Noise.** Spontaneous firing, sometimes at rates comparable to stimulus-driven firing (e.g. face-selective neurons, Schwartz et al. 2021).
- **Working memory.** During a delay, identities and locations remain decodable from activity in sensory areas, with no external stimulus present.
- **Expectation.** Predictive coding: representations of expected stimuli become active *before* the stimulus is presented (Demarchi et al. 2019; Kok et al. 2017).

In each case: representational content tokened, corresponding feature *absent* from the world.

The functional need

Similar neuronal events should sometimes have *dissimilar* cognitive consequences.

Cognitive systems need a gateway. A mechanism that tells *cases where neuronal activity reliably indicates the world* apart from *cases where it does not*.

This gateway is the interface between perception and cognition.

PRM's central hypothesis. The brain implements such a mechanism: a *perceptual reality monitor*, or *discriminator*.

The output of the discriminator: pointers

The discriminator's output is a *pointer*.

- A pointer refers to a memory address where information is stored.
- It *refers without duplicating*. If today's date is at address 100, a pointer to 100 refers to the date without re-encoding it.
- In the brain: if activity at address 101 carries the FO representation *red dot*, the discriminator outputs a pointer to 101.
- Downstream systems can “dereference” the pointer to access the FO content.

The output is a higher-order representation

The pointer is paired with information about *reliability*.

In natural language, the content is something like:

“The signal at address 101 reflects how the world is right now.”

or, in a contrasting case:

“The signal at address 101 is internally generated.”

This is a meta-representation of the FO state. PRM is therefore a higher-order theory, but one that does not require duplicating FO content at the higher level.

Functional consequence: stubborn dispositions to believe

Once a pointer reaches the belief system, dereferencing it produces a *strong disposition* to form the corresponding perceptual belief.

- “I’m currently seeing a red dot”, not merely imagining or expecting one.
- These dispositions persist even when one has countervailing beliefs (you are told there is no red dot, yet it still looks red).
- This explains the *stubborn* quality of conscious experience: a headache cannot be reasoned away.

Conscious vs. unconscious states, on PRM

Conscious state: a representational state that the discriminator monitors and outputs a pointer to. The subject is aware of being in it; it provides strong dispositions to believe.

Unconscious state: a representational state to which the discriminator outputs no pointer. It can still influence behavior (priming, blindsight) but the subject is not aware of being in it.

PRM thus answers the *consciousness question*: a clean mechanistic distinction, and a functional role for consciousness as the gateway between perception and cognition.

The structuralist turn

PRM does not, on its own, answer the *quality question*: why does a conscious state feel the way it does? It is compatible with several theories of qualitative character. Lau favours pairing it with *Quality Space theory*.

Naive view. You discriminate red from green *because* they feel different.

Quality space (Clark, Rosenthal). They feel different *because* you can discriminate them. Had you not been able to discriminate them, they would not have felt different at all.

Qualitative characters are identified with *functional profiles* of discriminative dispositions.

Why this fits PRM

Each quality is a point in a space whose distances reflect discriminability. Red and pink are close; red and green are far; two qualities indiscriminable for a subject collapse to one point.

Recall: absolute firing rates are a poor cue to what is in the world. The discriminator tracks *relative* activity, effectively the sensitivity and discriminability of one's sensors.

⇒ The very mechanism the discriminator uses to do its reality-monitoring job *also* fixes the qualitative profile of the experience. Tracking discriminability is, on Quality Space, tracking how things look from the subject's perspective.

The colour-sensor robot

Imagine a robot with colour sensors, a discriminator, and a rational belief system.

- Noise sometimes spikes red-sensor activity. If the discriminator outputs a pointer to it, the robot *cannot help* believing there is something red, even if it knows there isn't. "I know there's nothing red, but I have the urge to believe there is."
- Suppose its red and blue sensors give similar signals. The discriminator outputs the same pointer for both. The robot reports: "I know you call them red and blue, but to me they're the same."

The robot has a *subjective perspective*: it answers colour questions from the structure of its own sensitivities, not the physics.

PFC as a candidate site for the discriminator

Multiple lines of evidence connect the prefrontal cortex to reality monitoring and metacognition.

- Paracingulate sulcus size correlates with reality-monitoring performance (Buda et al. 2011); its morphology distinguishes hallucinating from non-hallucinating patients (Garrison et al. 2015).
- Lateral PFC differentiates *perceived* items from items maintained in working memory (Mendoza-Halliday & Martinez-Trujillo 2017; Tanigawa et al. 2022).
- Note: PRM is *relationalist*. It predicts that whether the subject is conscious is decodable from PFC, but *not* that phenomenal content is.
- Consistent with the broader literature linking PFC activity to conscious vs. unconscious perception (Mashour 2020; Michel & Morales 2019).

Predictions: the discriminator can be tricked

When the discriminator's assumptions about noise are wrong, characteristic errors follow.

- **Blindsight.** After a V1 lesion, signal and noise activity drop. A discriminator that fails to recalibrate overestimates noise, and judges genuine signals as “unreliable”: residual visual ability without conscious experience (Lau 2008; Ko & Lau 2012).
- **Subjective inflation.** In the periphery, more false alarms and inflated confidence: the discriminator may underestimate peripheral noise (Rahnev et al. 2011; Odegaard et al. 2018).
- **Stable retinal images.** Stimuli that move with the eye look “not external”; the discriminator discards them and they fade from awareness.
- **Motion-induced blindness.** A stationary target on a moving background mimics a perceptual scotoma; the discriminator treats it as a defect of the visual system itself (New & Scholl 2008, 2018).

Repurposing the discriminator for explicit metacognition

Hypothesis: explicit confidence ratings re-use computations performed by the discriminator.

Prediction. Explicit metacognition should be *consciousness-selective*: signals invisible to the discriminator should be invisible to confidence judgements.

- Blindsight patients fail to bet appropriately on stimuli in their blind field (Persaud et al. 2007, 2011).
- Healthy subjects show metacognitive insensitivity for unconscious stimuli (Michel 2022).

The challenge in one sentence

HO theories make consciousness depend on *some kind of representation of one's own states*.

Block's recent work (on the perception/cognition border and on phenomenal overflow) presses on two questions:

1. Is conscious perception *nonconceptual*? If yes, can a higher-order *thought* or judgement be the right kind of mechanism?
2. Does phenomenal consciousness *overflow* cognitive access? If yes, does any HO architecture have the resources to capture what overflows?

These are *indirect* challenges: not arguments aimed at PRM by name, but pressures on the broader framework PRM inhabits.

Block's positive view (very briefly)

Block argues that perceptual and cognitive representations differ in *format*.

- Perception is **NNI**: *non-propositional, non-conceptual, iconic*.
- Cognition is paradigmatically **PCD**: propositional, conceptual, discursive.
- Iconicity = analog tracking and mirroring (mercury-thermometer style).
- Empirical leverage: visual adaptation, infant colour perception, object files.

The infant colour case

Block's key case study:

- Pre-11-month-old infants do not use colour to track object identity (no surprise when a red disc emerges as green from behind an occluder).
- But they fixate on a red disc against a green background; so they *see* colour.
- Block's conclusion: infants consciously perceive colour *without activating* colour concepts.

In the final chapter, Block recycles this case to argue that phenomenal consciousness *overflows* access:

"The child's nonconceptual color perception is conscious but not available to the global workspace because of the child's lack of color concepts."

Why this presses on HO theories, and on PRM

- If conscious perception is genuinely nonconceptual, the HO state cannot be a fully conceptual thought (HOT in Rosenthal's strict sense looks too rich).
- If phenomenal consciousness genuinely overflows access, there are conscious states with no corresponding HO uptake at all.

For PRM specifically:

- PRM's HO state is a *pointer* + reliability tag, not a fully conceptual thought, so nonconceptuality alone is not fatal.
- But PRM still requires that *some* discriminator output be tokened. If Block is right that some FO states are conscious without any such uptake, PRM faces a counterexample.

Beck's reply: a slide between two claims

Beck (2023) argues that Block's argument leans on a *slide*:

- Premise needed for nonconceptualism: infants *do not activate* colour concepts in object-individuation tasks.
- Premise needed for overflow: infants *lack* colour concepts altogether.

These are not the same. Failing to deploy a concept in one task does not show that the concept is missing.

Beck's analogy: you cannot establish phenomenal overflow of colour perception from a failure to activate colour concepts in object-individuation tasks. No more than you can establish phenomenal overflow of letter perception from the fact that one can stare at a Scrabble tray without noticing the seven-letter word the tiles make (Beck 2023).

An empirical wedge

- Infants can be trained in just two or three sessions to track colour for object-individuation purposes.
- If they had to construct colour concepts *from scratch*, this rapidity is hard to explain.
- Better explanation: they have the concepts but prioritise shape (a much better cue to object identity).

Upshot for HO / PRM. The infant case does not establish the strong overflow conclusion Block needs. The deeper question stays open: what *kind* of HO state is required, and can we tell empirically whether it is present?

Take-home

- HO theories: a mental state is conscious only if one is, in some sense, aware of being in it.
- PRM: that awareness is the output of a *discriminator* that monitors the reliability of one's own sensory signals, and outputs a pointer with a reliability tag.
- This explains both *what consciousness does* (gateway to belief) and *why conscious experience is stubborn*.
- Combined with Quality Space, it offers a candidate account of qualitative character.
- Indirect challenges from Block (nonconceptual perception, phenomenal overflow) press on HO views in general; but, as Beck shows, the strong overflow conclusion is not yet earned.

Selected references

- Beck, J. (2023). Review of N. Block, *The Border Between Seeing and Thinking*.
- Block, N. *The Border Between Seeing and Thinking*. OUP.
- Brown, R., Lau, H. & LeDoux, J. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*.
- Clark, A. (2000). *A Theory of Sentience*.
- Dijkstra, N. et al. (2022). Perceptual reality monitoring: a review.
- Lau, H. (2022). *In Consciousness We Trust*. OUP.
- Michel, M. (forthcoming). The perceptual reality monitoring theory. In Herzog, Schurger, Doerig (eds.), *Scientific Theories of Consciousness*, CUP.
- Rosenthal, D. (2005). *Consciousness and Mind*.
- Rosenthal, D. (2010). How to think about mental qualities.