Review of: "Growing Confidence and Remaining Uncertainty About Animal Consciousness"

Richard T. Born[1]

1 Harvard University

Potential competing interests: No potential competing interests to declare.

I enjoyed reading Prof. Irwin's article, and I'm grateful to him for writing it. I learned a lot—when I have some spare time, I'll spend much of it poring over many of the articles he cites.

I am a neurophysiologist who has spent a career studying basic mechanisms of visual perception in nonhuman primates. I rarely think about what I call "the C-word"—when I do, it's usually at the end of a week while enjoying a cigar and a glass of whiskey in front of my fire pit. I have largely avoided writing about C. ("consciousness") because I've always sensed that it was a bit of a trap . . . or an illusion. As my friend and colleague Gord Fishell says, "Consciousness and evolution are the realm of old men." But perhaps now that I'm over 60, I can indulge myself a bit.

I think I disagree with Prof. Irwin on his opening, perhaps central, claim: I don't think we're anywhere near a mechanistic understanding of C., though I agree with him that there might be some consensus on interesting places to look. When it comes to understanding nearly anything about how brains work, apart from basic reflexes, I think we are still more or less in the dark ages. Accordingly, I will not attempt to rebut his rebuttable assertions here—I am just not familiar enough with the large literature he cites—but I will offer some impressions of 1) where I think he gets it right, 2) where I think he misses the point, and, most importantly, 3) where I think there are opportunities for experimental approaches that might shed light on the problem.

First, I agree strongly with a critical role for ***attention*** in C.—Prof. Irwin lays this out nicely in one of his opening paragraphs. I would go even further and say that attention comprises a major part of what we call C. However, though Wm. James wrote that, "Everyone knows what attention is," it remains a great mystery as to how it is actually implemented in nervous systems. But here there is great hope for a mechanistic understanding. A reader interested in this would do well to begin with some of the work of John Maunsell (e.g., Ghose & Maunsell 2023), who has performed careful behavioral and neurophysiological studies in both monkeys and, more recently, mice. Given the groundwork he and others have laid, and the powerful armamentarium of tools available for circuit manipulation in rodents, I think this is now a tractable scientific question. And I predict that once we have worked out a detailed, circuit-level understanding of attention, we will be a long way towards understanding C.

Second, I believe the "mental unity" feature is a red herring—a kind of useful delusion. There are a number of lines of experiments that point to this, but the one that comes most readily to mind is those on "postdiction," the idea that much of what we think we perceive in real time is, in fact, a kind of just-so story made up after the fact. Interestingly, this kind of

thing takes place at multiple time scales, from 10s of milliseconds, in perceptual illusions like the flash-lag effect, to hours and days, in “hindsight bias.” For an overview of these and other related phenomena, I’d recommend an excellent article by Shin Shimojo (2014). Another interesting and vivid example comes from the experiments that Michael Gazzaniga and Roger Sperry did with split-brain patients (as described in Gazzaniga’s classic book, *The Social Brain*). In these patients, neurosurgeons had transected the corpus callosum (in an effort to treat intractable epilepsy), effectively disconnecting the half of the brain that has language and therefore “can talk” (usually the left hemisphere) from the non-verbal right hemisphere. In one telling experiment, Gazzaniga and his student, Joe LeDoux, *briefly* flashed two simple “conceptual problems”: one was a picture of a chicken claw in the right visual field that was “seen” by the verbal left hemisphere (LH); the non-verbal right hemisphere (RH) “saw” an image of a snowy scene. The patient responded by pointing to one of 8 possible images, continuously visible for perusal, that “matched” one of the flashed images. The patient’s right hand (controlled by the LH) pointed to a picture of a chicken’s head, and the left hand pointed to a snow shovel—both hemispheres solved the problem. The interesting part came when they asked the patient *why* he had made his choices. He replied, “Oh, that’s easy. The chicken claw goes with the chicken and you need a shovel to clean out the chicken shed.” (p. 72 of Gazzaniga 1985) Some set of circuits in the left hemisphere needed to account for what it observed while only being privy to the actual decision process local to its hemisphere. So it nailed the explanation for why it chose the chicken, but when it came to the shovel—a situation for which it “observed” only the choice behavior and not the process that led to it—it confabulated. Based on this and other experiments, Gazzaniga referred to this “set of circuits” as the “interpreter module” (or “I-module”), and I think this is a good metaphor for another critical element of C., which is what I’ll call an “internal model” or just “IM” for short. I think this is an important piece of the puzzle that is largely missing from Dr. Irwin’s discussion.

Internal models are a critical feature of the brains of nearly all animals, and I believe they are absolutely central to an understanding of C. This is actually a very old notion, going back to at least Hermann von Helmholtz, the formidable 19th c. German physicist and physiological psychologist, who provided an important framework for thinking about perception as “unconscious inference” (Helmholtz 1925). Because we and our sense organs are in nearly constant motion, we need IMs to predict the perceptual consequences of our own movements and thereby distinguish between sensations produced by things in the external world (“exafference”) and sensations produced by self-movement (“reafference”). To take a simple example, if I use my eyes to track a bird flying from left to right, the bird is stationary on my retina while the visual background is whizzing by in the opposite direction. Yet I have a vivid perception of the bird’s rightward motion and no sense that the background is moving. This is because the brain uses a so-called “forward model” to predict the sensory consequences of my own eye movement. The prediction is based on a copy of the motor command to move the eyes (a so-called “efference copy”), and if the command is issued but the actual movement is thwarted, say due to a paralysis of one of the eye muscles, the world suddenly appears to “fly by” (as described in Helmholtz 1925). These sorts of internal models are essential for perception, and their physiological bases have been studied in diverse animal species ranging from monkeys (e.g., Sommer & Wurtz 2002) to weakly electric fish (Kennedy et al. 2014) to fruit flies (Westeinde et al. 2024). Moreover, there are excellent computational models for how some of these IMs actually work, and these, combined with powerful new tools for hacking circuits, lead me to believe that we may be close to a detailed, mechanistic understanding of some of the simpler forms of IM.

But how does understanding internal models get us closer to understanding C.? My speculation here is that, at some point during the evolution of nervous systems, these kinds of IMs grew in their level of abstraction and computational power to eventually be useful for modeling the behavior of conspecifics (giving rise to what psychologists refer to as “theory of mind”)—critical for social interactions—and, ultimately, to model the behavior of the organism that owned the nervous system—what we call the “self.” By casting what Prof. Irwin refers to as the essential feature of the “sense of self” as an elaborated internal model, I believe we can gain experimental traction on this otherwise nebulous concept.

I fear that this “review” is rapidly approaching the length of the original article, so I will try to wrap things up. One thing I’ve learned from writing this is that I should probably bite the bullet, accept my status as an “old man,” and actually write a long-format piece in which I can spell out some of these ideas in more detail.

Let me end here by enthusiastically embracing a final one of Prof. Irwin’s “critical features” for C.: what he refers to as “mental causation.” I would again recast this somewhat using the title of another excellent article by my colleagues, Michael Shadlen and Roozbeh Kiani: “Consciousness as a Decision to Engage” (2011). By doing this, we, via Shadlen and Kiani, can embed the somewhat problematic concept of “mental causation” in the context of a rich theoretical, computational, and experimental literature on sensory decision-making and, again, give us the prospect of a mechanistic understanding.

My closing prediction is that it is only through a detailed, neuroscientific understanding of several key features of nervous systems, namely, 1) attention, 2) internal models, and 3) decision making, that we will ultimately come to an understanding (probably only “in principle”) of C. And, here, by “understanding” I mean what emerges from sound theory (informed by philosophers and psychologists), rich computational models, and lots and lots more experiments.

Two minor points in need of clarification:

‘. . . words of William James (1890), who further defined consciousness as "the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought."’ James was specifically referring to “attention” here, not “consciousness.” The full quote is “Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought.”

“. . . transcranial magnetic stimulation and high-density EEG can detect clear-cut changes in the ability of the thalamocortical system to integrate information when the level of consciousness fluctuates across the sleep-wake cycle (Massimini et al., 2009) . . .” TMS can’t “detect” anything, so I found this sentence confusing.

References:

Gazzaniga MS (1987) The Social Brain: Discovering the Networks of the Mind. Basic Books. ISBN 978-0-465-07850-9.

Ghosh S, Maunsell JHR (2023) Rodent attention: Probing the mouse mind with reverse correlation.*Curr Biol.* Sep 11;33(17):R916-R918. doi: 10.1016/j.cub.2023.07.015. PMID: 37699352.

Helmholtz HV (1925) Treatise on Physiological Optics. Optical Society of America.

Kennedy A, Wayne G, Kaifosh P, Alviña K, Abbott LF and Sawtell NB (2014) A temporal basis for predicting the sensory consequences of motor commands in an electric fish. *Nature neuroscience*, 17(3), pp.416-422.

Shadlen MN & Kiani R (2011). "Consciousness as a decision to engage," inCharacterizing Consciousness: From Cognition to the Clinic? Research and Perspectives in Neurosciences, eds S. Dehaene and Y. Christen (Berlin: Springer-Verlag), 27–46.

Shimojo S (2014) Postdiction: its implications on visual awareness, hindsight, and sense of agency.*Frontiers in psychology*, 5, p.196.

Sommer MA & Wurtz RH (2002) A pathway in primate brain for internal monitoring of movements.*Science*, 296(5572), pp.1480-1482.

Westeinde EA, Kellogg E, Dawson PM, Lu J, Hamburg L, Midler B, Druckmann S and Wilson RI (2024) Transforming a head direction signal into a goal-oriented steering command. *Nature*. https://doi.org/10.1038/s41586-024-07039-2