Against Organic/Artificial Parity in Consciousness Attributions

Eric Schwitzgebel
University of California, Riverside
Jeremy Pober
Universiteit Antwerpen; Universidad Lisboa

The Basic Idea

- We aim to do justice to the following two claims:
  - Organic but extraterrestrial beings are, all else being equal, more likely to be conscious than artificial ones.
  - There is no ‘deep’ difference between organic and artificial life in the sense that, for whatever the relevant basis of consciousness (e.g., physicalist, functionalist) if an organic and artificial being both have it, they are both conscious. (See “No Relevant Difference” Principle: Schwitzgebel and Garza 2015).
  - Thus we will argue for this first claim on different grounds: that organic beings are more likely to be conscious for contingent reasons having to do with their etiology.

Agenda

1. Epistemology vs. Metaphysics
2. Sophisticated Behaviors
3. The Copernican Principle
4. Blockheads
5. Mimicry

Epistemology vs. Metaphysics

- The “no relevant difference” principle is a metaphysical one.
  - It states that for whatever the basis of consciousness, it grounds consciousness in organic and artificial beings equally.
- But what if we don’t know what a novel (to us) being’s internal architecture looks like?
  - Never mind that even if we knew what it looked like, we don’t have a consensus view on the architectural basis of consciousness in the first place.
- We have to look for what we take to be the best *indicators* of a conscious mind to be.
  - We take those to be Sophisticated Patterns of Behavior (SPB’s).

Sophisticated Patterns of Behavior

- We take there to be patterns of behavior where the pattern is best explained by appeal to internal states with intentional content that themselves have multiple possible levels of sophistication.
- Picture a crow who places seeds in a cache, then comes back months later to find them.
- But what if the wind blew the seeds a few meters away? A crow likely can't—but some other being (Super-Crow?) possibly can—take that into account, understanding that seeds are the sort of thing that can move in strong winds.
- What if another being got the seeds first? A really sophisticated being (Ultra-Crow?) could, knowing the behavioral patterns of their conspecifics (or other relevant organisms in their environment) might be able to figure out who took them.
- We can keep increasing the level of sophistication until we think it meets the 'threshold' of consciousness in animals (we take no stance here on where to find it).

The Copernican Principle: Goal

- If we see PSB's, the default best explanation by definition is a mind with intentional states (we set aside for the moment the relationship between these states and consciousness).
- When should we assume that the default best explanation is in total the best explanation such that it carries ontological import?
- For extraterrestrial organic beings, we believe it should be assumed widely: the *default assumption* should be that PSB-performing organisms are conscious.
- We justify this with the Copernican Principle.

The Copernican Principle: Theory

- The Copernican Principle states that we should not expect ourselves to be lucky with respect to our position in the universe absent evidence that we are.
  - This applies to consciousness. We are not in a particularly privileged position with respect to other PSB-performing organisms.
  - We grant that we have some such evidence: we are quite lucky compared to other terrestrial organisms, for instance.
  - The Copernican Principle applies when we don't have such evidence, which we don't on first encountering PSB-performing organisms.
- The upshot of the Copernican Principle, when applied to consciousness, is a *default liberalism* about consciousness attributions to organic, extraterrestrial life.

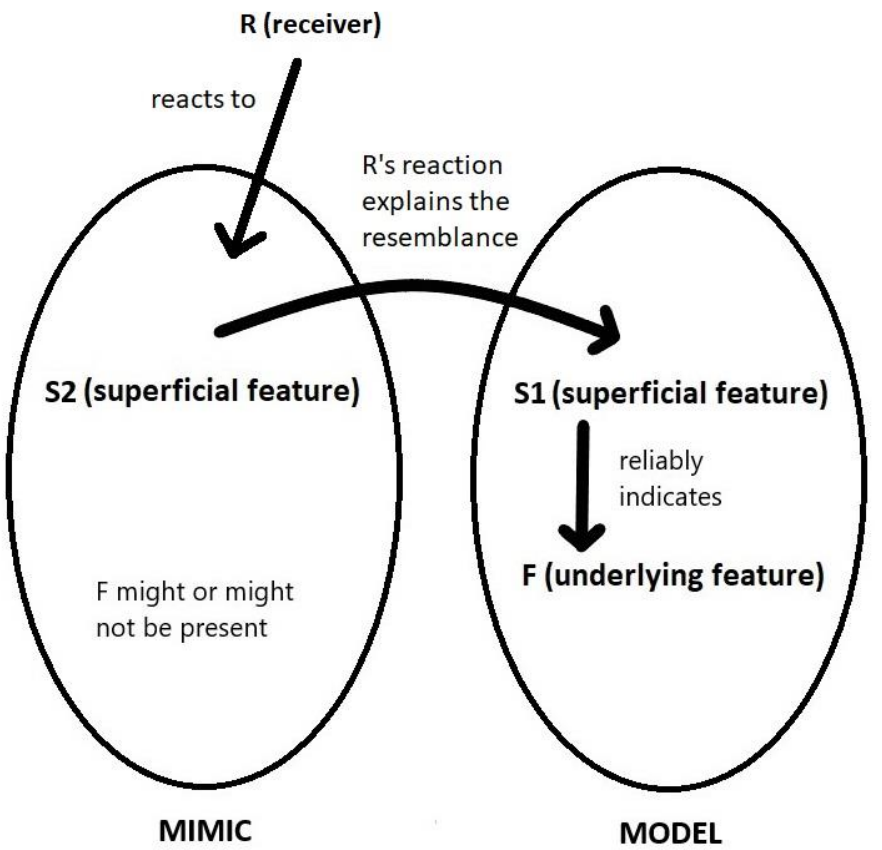
Blockheads

- Block (1981) conceived of an artificial being who was programmed with an indefinitely long string of possible responses to any input that would be a sensible response for a human to make. Block conceives of this linguistically (i.e., in the terms that a Turning Test is given) but it can be expanded to all behavior, not just linguistic utterances.
- We know the Blockhead isn't conscious because it is just a giant lookup table (c.f. "Chinese Room").
- Thus (to the extent that the Blockhead is metaphysically possible) we know it is metaphysically possible for a non-conscious being to exhibit PSB's.
- We take Blockheads to be for all intents and purposes impossible to evolve, and we take it there would be little if any reason to program such a being aritifically, as the program is a borderline impossible amount of work.
- But this last point does approach the crux of the issue: the etiology of organic and artificial beings is quite different, with different pressures (evolution vs. what is feasible for a creator).

Mimicry

- We suggest that what a Blockhead is doing is ‘mimicking’ consciousness.
- In mimicry in biology, a ‘target’ being produces a signal S1, and the mimic produces a different signal, S2, but one that will (at least sometimes) trick some observing entity.
  - S1 indicates the presence of some underlying property, S2 does not.
  - Example: Viceroy butterflies roughly copy Monarch butterflies’ wing patterns to dispel predators to whom the Monarch, but not the Viceroy, is poisonous.
- We understand mimicry to be an ability to produce patterns of behavior that look sophisticated to an observer limited in the ways a human is (temporally bounded, deeply irrational, etc).
- Mimics can be perfect or imperfect. A perfect mimic could fool an ideal observer all of the time, but actual mimics never need to be that good, because there are no ideal observers.

Mimicry


R (receiver)
reacts to
R's reaction
explains the
resemblance
S2 (superficial feature)
F might or might
not be present
S1 (superficial feature)
reliably
indicates
F (underlying feature)
MIMIC
MODEL

Consciousness Mimics

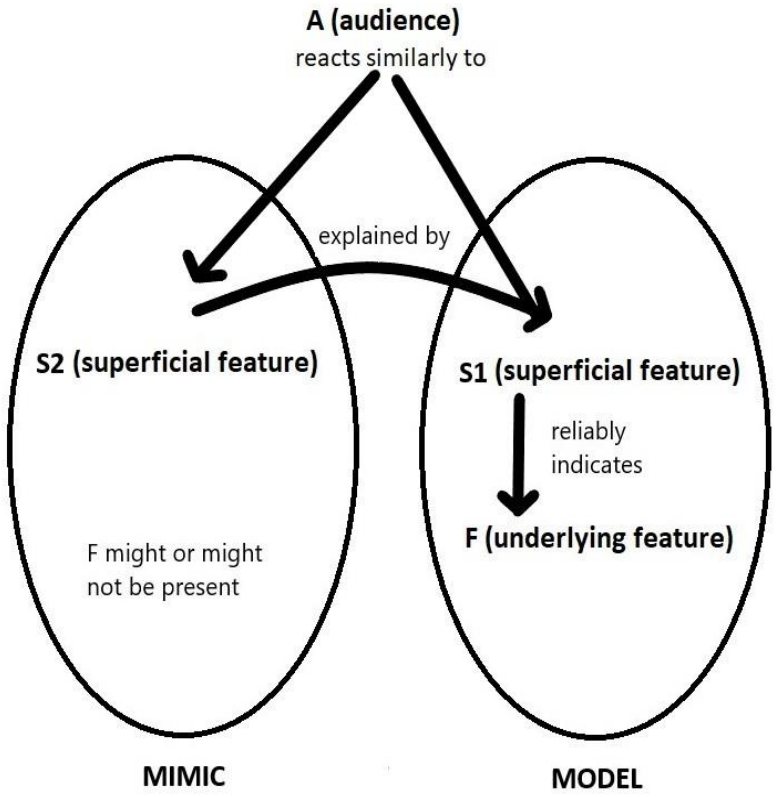
"Hello" toy.

Large Language Models

Consciousness mimic: superficial features suggestive of consciousness but best explained by having been modeled on the superficial features of a model system, for the sake of an audience that responds similarly to those superficial features.

Social AI like Replika.

Mimics *might* be conscious, but the inference to underlying consciousness is disrupted.



Heuristics

- Why think that it is likely artificial beings are designed to mimic, rather than be, conscious?
- We would have reason for thinking this if such a being were easier to program.
- Heuristics—”quick-and-dirty” algorithms that get most things right, most of the time (see e.g., our performance on the Wason Selection Task).
  - Easier to program (it’s why our brain uses so many! We suggest an artificial being could use *more* and *better* ones).
  - E.g. “respond to positive affect with smiles and relaxed body language.”

Rejecting Parity

Copernican grounds for default liberalism about alien consciousness.

Mimicry grounds for a more cautious attitude about robot consciousness.

There is therefore a disparity in how we are justified in reacting to space aliens and robots, given their different histories. Even if they have similar overall levels of behavioral intelligence, it's reasonable to be more epistemically cautious about robot consciousness than alien consciousness.

Complications and Implications

Q. What if most behaviorally intelligent entities in the universe are consciousness mimics?

A. The Copernican conclusion might still be avoided if mimicry is a sufficiently important difference that symmetry and simplicity constraints don't justify default liberalism.

Q. What about AI not built on principles of mimicry?

A. The present arguments do not apply.

Q. What if mimics started evolving independent of the model species?

A. This would be an intermediate case.

Q. Could we apply the Copernican Argument to the alien equivalent of non-human animals?

A. Yes, Earth would be a strangely lucky place if our non-human animals were conscious and similarly complex and sophisticated animals elsewhere were nonconscious "zombies".