

**The Copernican Argument for Alien Consciousness;
The Mimicry Argument Against Robot Consciousness**

Eric Schwitzgebel
University of California, Riverside
Riverside, CA 92521
USA

Jeremy Pober
University of Lisbon / University of Antwerp
1649-004 Lisboa, PT / Antwerpen 2000, BE

November 11, 2024

The Copernican Argument for Alien Consciousness; The Mimicry Argument Against Robot Consciousness

Abstract: On broadly Copernican grounds, we are entitled to default assume that apparently behaviorally sophisticated extraterrestrial entities (“aliens”) would be conscious. Otherwise, we humans would be inexplicably, implausibly lucky to have consciousness, while similarly behaviorally sophisticated entities elsewhere would be mere shells, devoid of consciousness. However, this Copernican default assumption is canceled in the case of behaviorally sophisticated entities designed to mimic superficial features associated with consciousness in humans (“consciousness mimics”), and in particular a broad class of current, near-future, and hypothetical robots. These considerations, which we formulate, respectively, as the Copernican and Mimicry Arguments, jointly defeat an otherwise potentially attractive parity principle, according to which we should apply the same types of behavioral or cognitive tests to aliens and robots, attributing or denying consciousness similarly to the extent they perform similarly. Instead of grounding speculations about alien and robot consciousness in metaphysical or scientific theories about the physical or functional bases of consciousness, our approach appeals directly to the epistemic principles of Copernican mediocrity and inference to the best explanation. This permits us to justify certain default assumptions about consciousness while remaining to a substantial extent neutral about specific metaphysical and scientific theories.

Word Count: ~11,000 words

Keywords: alien life, Artificial Intelligence, consciousness, Copernican Principle, Large Language Models

The Copernican Argument for Alien Consciousness; The Mimicry Argument Against Robot Consciousness

1. *Our Argument in Sketch.*

Engineers might soon build robots so sophisticated that people are tempted to regard them as meaningfully conscious or sentient – robots capable, or seemingly capable, of genuinely feeling pleasure and genuinely suffering; robots that possess, or seem to possess, real conscious awareness of themselves and their environments.¹ Less likely, but not impossible, we might soon discover complex, seemingly intelligent, seemingly conscious space aliens – maybe under the surface of Europa or via signals from a nearby star. It would be lovely if we had a scientifically well-justified consensus theory of consciousness that we could straightforwardly apply to determine whether such robots or aliens are genuinely conscious or only superficially seem so. Unfortunately, there is no consensus theory, and consensus is unlikely anytime soon.²

Still, as we will argue, we are not wholly in the dark regarding robot and alien consciousness. When apparently confronted with what we will call *behaviorally sophisticated* robots or aliens, a particular pair of default attitudes is justified.

¹ By “consciousness” we here mean “phenomenal consciousness” as standardly used in mainstream Anglophone philosophy of mind and consciousness studies, innocent of any commitment to naturalistically dubious properties such as immateriality or irreducibility (see discussion in Schwitzgebel 2016, 2024). Researchers who regard robot consciousness in the near future as a real possibility include Dehaene, Lau, and Kouider 2021; Chalmers 2023; Long and Sebo 2023; Butlin et al. 2023. Among those who reject the near-term possibility of robot consciousness are Godfrey-Smith 2016; Seth 2021.

² Seth and Bayne 2022 survey some recent scientific theories of consciousness; see Schwitzgebel 2024 for discussion of grounds for pessimism, addressing a wider range of theories, including panpsychism and transcendental idealism. Approaches that rely on points of agreement among competing theories, such those advocating “ecumenical heuristics” (Shevlin 2020), “theory light” approaches (Birch 2022), or lists of “indicator properties” (Butlin et al. 2023), though helpful, face tradeoffs between substance and breadth of appeal.

Toward behaviorally sophisticated extraterrestrial entities (“aliens”), the appropriate default attitude is to attribute consciousness. We will defend this view on broadly Copernican grounds. According to the cosmological principle of Copernican mediocrity, we do not occupy any special position in the universe. Copernican mediocrity, we argue, would be violated if, among all the behaviorally sophisticated entities that have presumably arisen in the universe, we humans are among the rare few who also happen to be conscious, while the rest are, so to speak, empty shells, devoid of consciousness. That would make us weirdly, inexplicably lucky – much as if we just happened, inexplicably, to occupy the exact center of the universe.

Toward a broad range of behaviorally sophisticated current, near-future, and hypothetical robots – entities that we will call “consciousness mimics” – we will argue that the appropriate default attitude is skepticism or uncertainty about their consciousness. If a system is designed to behave in ways that mimic the behavior of conscious entities in order to evoke a certain type of reaction in an audience, we normally cannot justifiably infer the existence of consciousness from features that would otherwise suggest it. While the effective mimicry of consciousness might in some cases require corresponding consciousness beneath, there’s no compelling reason to think this is generally so. There’s a vast gulf between a robot designed or selected to emit the superficial appearance of a friendly greeting – a robot perhaps modeled on us and engineered specifically to pass various tests – and a robot who actually experiences friendship.

Thus, the alien and robot cases are asymmetric. If an alien exits a spaceship and greets you, it’s reasonable to assume that you have encountered a conscious being; if a social robot designed to mimic human behavior rolls off a factory floor and greets you, it’s reasonable to doubt its consciousness. This asymmetry holds even if the alien and robot both exhibit an overall similar degree of behavioral sophistication.

One might have thought a certain type of parity principle would be plausible, according to which we should approach robots and aliens similarly, applying the same types of behavioral or cognitive tests and attributing or denying consciousness similarly to the extent they perform similarly. If you'd apply a certain type of Turing Test to one, for example, apply that same Turing Test to the other, and draw similar conclusions given similar performance.³ If you'd look for recurrent processing loops of Type A or cognitive performance of Type B in one, look for recurrent processing loops of Type A or cognitive performance of Type B in the other. The Copernican Argument and the Mimicry Argument jointly provide reason to reject robot-alien parity. This is because what we know about the history of an entity's creation makes an epistemic difference to our assessment of its consciousness.

Our argumentative approach is unusual in the following respect: Instead of attempting to ground guesses about robot or alien consciousness in specific metaphysical or scientific theories about the physical or functional bases of consciousness, we appeal directly to epistemic principles – a Copernican principle and a version of inference to the best explanation. Even absent knowledge of what particular types of processes give rise to consciousness and of the inner workings of the entities being evaluated, these principles ground default assumptions about consciousness, based in our knowledge of the entity's history. These broad principles can give us an epistemic toehold on alien and robot consciousness with minimal commitment to the tendentious specifics of particular theories of consciousness.

³ The original proposal is in Turing 1950. See Harnad 2003 for its adaptation as a test of consciousness and Schneider 2019 for a consciousness-specific adaptation intended to defeat cheating strategies that might be employed by large language models. Udell and Schwitzgebel 2021 explores the difficulty of implementing such adaptations without question-begging commitments.

2. Behavioral Sophistication.

An entity is *behaviorally sophisticated* in our intended sense if it is capable of complex goal-seeking, complex communication, and complex cooperation, where “goal-seeking”, “communication”, and “cooperation” are all understood in terms of superficial behavioral patterns without commitment to the existence of consciousness or any particular internal cognitive architecture.⁴

Suppose an alien species creates complex devices that extract nutrition from its environment. The sources of nutrition are stored and then consumed by members of the species when they would otherwise be nutritionally deprived. Members of the species respond to changes in their environment that threaten their survival by engaging in highly specific activities that evade the threat, often well in advance and with relatively narrow margins of error. The aliens interact in highly complex and seemingly technologically advanced ways, engaging in activities that substantially increase their chances of survival or reproduction only if several other group members engage in specific activities at specific times when they are physically remote and not in direct contact. This interaction is enabled by complex signals with flexible contents. For good measure, suppose that the aliens fly spacecraft at least as good as our own.

Such a species would be *behaviorally sophisticated* as we intend the phrase. When used by biologists discussing the goals, communication, and cooperation of bacteria, birch trees, and mycorrhizal fungi, such terms need not presuppose consciousness or the types of cognitive mechanisms we are familiar with from animal cases. Similarly for our use of “goals”, “communication”, and “cooperation” here. (Employ scare quotes if you like.) Maybe complex behavior of the sort described nomologically or metaphysically requires consciousness or a

⁴ Compare Dennett’s (1987) and Davidson’s (1984) interpretativism.

particular cognitive architecture? In that case, consciousness would be built in for free, so to speak, given the sophistication described above. But that would depend on matters of theory about which we aim to stay neutral. The point we insist on is that behavioral sophistication bears an *epistemic* relationship to consciousness. Specifically, when it is observed in a terrestrial being, it provides *prima facie* epistemic warrant for attributing consciousness to that being.

How sophisticated is “behaviorally sophisticated”? Our Copernican Argument works in its simplest form if approximately human-level sophistication is the required minimum. However, as we will see in Section 12, a generalization of the Copernican Argument might also work with a much lower standard of sophistication. To forestall a possible confusion: Our argument will be that behavioral sophistication is a *sufficient* condition for the default attribution of consciousness, not a necessary condition.

3. *The Copernican Principle of Consciousness.*

We assume that the emergence of behavioral sophistication is not *extremely* rare. Though it’s impossible to know exactly how rarely behavioral sophistication arises, on a conservative estimate, behavioral sophistication might independently evolve at least once per every billion galaxies at some point in those galaxies’ history.⁵ This would yield at least a thousand independently arising behaviorally sophisticated clades or groups in the galaxies currently occupying the observable portion of the universe.

Among these entities, we can imagine various possibilities for the distribution of consciousness. On one extreme, all behaviorally sophisticated extraterrestrial (“alien”) entities

⁵ For scientific estimates of the rate of occurrence of “intelligent” or “technological” life in the universe, see Frank and Sullivan 2016; Snyder-Beattie et al. 2021.

might be conscious. On the other extreme, maybe only Earthlings are. And of course a vast range lies between the extremes. If among all of these entity-types, only we Earthlings were conscious, or only we and a few others, we would be strikingly special. According to an epistemic principle we'll call the Copernican Principle of Consciousness, we can default assume that we wouldn't be in that way special.

The Copernican Principle of Consciousness is a special case of the more general Copernican Principle in cosmology, according to which “the Earth is not in a central, specially favored position” (Bondi 1968, p. 13) or, alternatively, “We do not occupy a privileged position in the Universe” (Barrow and Tipler 1986, p. 1; various other formulations similar in spirit are possible, e.g., Scharf 2014). Recent scientific applications tend to emphasize the principle's connection to the homogeneity and isotropy of the universe at large scales.⁶ If everything is pretty much the same everywhere we look, no position (including our own) is particularly special.

Extending this thinking to the case of consciousness, we propose the following *Copernican Principle of Consciousness*:

Among behaviorally sophisticated entities, we are not specially privileged with respect to consciousness.

If the Copernican Principle of Consciousness is correct, we are not specially lucky to be blessed with a consciousness-instilling Earthiform biology while other behaviorally sophisticated entities' architectures leave them entirely non-conscious. Nor are we in the specially privileged position of being especially *more* conscious than other behaviorally sophisticated entities. Nor

⁶ E.g., Caldwell and Stebbins 2008; Clarkson, Bassett, and Lu 2008; Camarena, Marra, Sakr, and Clarkson 2022.

are we blessed with a special kind of awesomely *good* consciousness. Otherwise, absent some reason to think we are special, we Earthlings would be suspiciously, inexplicably lucky. Among all the spaceship-flying, linguistically communicating, long-term-goal planning, socially cooperative entity-types in the universe, only we (and maybe a few others) are conscious?! Although (at least on some metaphysical and scientific theories of consciousness) that *could* be so, we probably ought to default to the assumption that we're not uniquely bestowed with magic light.

Scientific cosmologists generally interpret the Copernican Principle spatially or spatiotemporally, in terms of “position”. The Copernican Principle of Consciousness can also be understood spatiotemporally, if desired. If the observable universe has at least a thousand different spatiotemporal regions where behaviorally sophisticated entities have arisen, but only a small percentage of those regions host *conscious* entities, then Earth is in a special location: a bright spot of consciousness in a mostly-consciousnessless cosmos – the, or a, center of the “consciousness is here” map.

We see Copernican principles as *default* epistemic principles. Compelling evidence could potentially arise that we *are* at the center of the universe. Maybe our best theory of consciousness will someday imply we're lucky to have Earthiform neurons instead of Andromedan hydraulics. The assumption of mediocrity is just a reasonable starting place.

4. Four Concerns about the Copernican Principle of Consciousness.

We hope that most readers will find the Copernican Principle of Consciousness plausible on its face. However, to better articulate its boundaries and assess its plausibility, we consider four objections.

Concern A. "Specialness" is vacuous. Copernican principles depend on the somewhat murky notion of "special privilege".⁷ Specified narrowly enough, any position could be identified as unique or extremely unusual on some metric of evaluation. Maybe Earth is the only planet in the observable universe with a mass of X, average orbital distance of Y, and moon with magnesium-to-iron ratio of Z, with X, Y, and Z specified to 100 digits. Similarly, if I deal out five cards from a shuffled deck and receive J♣, 4♦, 7♣, Q♠, 7♠ in that order, in some sense that is a "special" or low-probability outcome, since the chance of receiving those exact cards in that order is under one in three hundred million. But if everything is special, then claims of specialness are nearly vacuous and, contra Copernican principles, we should be unsurprised to find ourselves in a "special" location.

Responding to this worry requires clarifying "special". In poker, A♠, K♠, Q♠, J♠, 10♠ is a special or privileged hand, while J♣, 4♦, 7♣, Q♠, 7♠ is not, because no other hand can defeat it. For poker purposes, a royal flush has distinctive importance. But even if we're not playing poker, A♠, K♠, Q♠, J♠, 10♠ is striking in a way J♣, 4♦, 7♣, Q♠, 7♠ is not (unless something of importance hinges on drawing those exact cards). Similarly, occupying the precise center of the universe is intuitively striking in a way that being 3.24 billion light years from the center of an approximately Hubble-sized universe is not. A license plate of OOO 000 is intuitively striking in a way that XNE 488 is not. If your neighbor arrives home with OOO 000 on their new car, you can reasonably expect an explanation other than chance. Alternatively, specialness might be captured by information compressibility: We can describe the royal flush as "the top five spades

⁷ Partly as a result, they invite concerns similar to those of other "self-location" principles. For example, the Doomsday Argument (Gott 1993; Lacki 2023), the Sleeping Beauty puzzle (Elga 2000 and subsequent literature), and the Anthropic Principle (more on which below) (Barrow and Tipler 1986; Bostrom 2002 and subsequent literature).

in descending order”, while to describe the other hand we probably can’t do better than just list the specific cards. We can describe the center of the universe as just “the center”, while most other positions would require three spatial coordinates. Thus, there are at least three different ways of characterizing “specialness” or “privilege”: relative to goals or rewards, in terms of something like intuitive strikingness, and in terms of information compressibility. Underlying all three is the idea that there is an aspect of “specialness” beyond statistical rarity or uniqueness. If any of these three approaches works, then Copernican principles are not in general vacuous. Alternatively, we are open to other ways of cashing out this additional aspect. Our list is not meant to be exhaustive.

Concern B. Consciousness is not special. The Copernican Principle of Consciousness requires specifically that *consciousness* is special – that being the one conscious entity in a pool of behaviorally sophisticated aliens would be more like drawing A♠, K♠, Q♠, J♠, 10♠ than like drawing J♣, 4♦, 7♣, Q♠, 7♠. But maybe consciousness isn’t special?

On the face of it, consciousness is very special. Some philosophers have argued that consciousness is the most intrinsically valuable thing in the universe, the grounds of moral status, and the primary source of meaning in life.⁸ Even if we don’t go that far, that consciousness is

⁸ Discussions include Kriegel 2019; Shepherd 2024. For more skeptical perspectives, see Levy 2024; Papineau 2024; as well as “illusionists” about consciousness such as Frankish 2016 and Kammerer 2019. Our Copernican principle doesn’t require that consciousness be *uniquely* special in the manner suggested by strong statements of its specialness; and, if necessary, it could be adapted to reference some other nearby property (e.g., “quasi-consciousness”). Alternatively, one might argue that although consciousness is special in a way, we should expect behaviorally intelligent alien species to have other properties that are equally special – schmonsciousness, say, for the aliens in Galaxy Alpha, bronssciousness for the aliens in Galaxy Beta, and lulonssciousness for the aliens in Galaxy Gamma – all different properties but equally giving their lives meaning and value (Lee 2019). In this case, we could relativize the Copernican Principle to the disjunction or schmonsciousness-or-bronssciousness-or-lulonssciousness-etc. or to whatever generic property those specific properties share.

even a plausible candidate for such high status suggests that it has some importance. If you were to learn you would never again be conscious, you would probably be dismayed in a way you would not be dismayed to learn you will never again weigh exactly X grams. It is also an intuitively striking property: If it were somehow discoverable that we human beings are conscious while our nearest behaviorally sophisticated alien neighbors are non-conscious, that would be quite a notable difference between us. Arguably, too, when we describe an entity as “conscious” or not, we compress substantial information into a brief statement, in the sense that the attribution of consciousness captures some general pattern or capacity in the target entity – though exactly which pattern or capacity will depend on what account of consciousness is correct (or alternatively, on what patterns or capacities the attributer regards as typical of conscious entities).⁹

Concern C. The Anthropic Principle makes consciousness unsurprising even if rare.

Brandon Carter remarks on an important qualification to Copernican principles: Our position is “necessarily privileged to the extent of being compatible with our existence as observers” (1974, p. 293). Even if most of the universe (e.g., large stretches of near vacuum) is incompatible with the existence of intelligent observers, we should be unsurprised to find ourselves inhabiting one of the rare hospitable spots. The resulting *Anthropic Principle* can be formulated in various ways (Barrow and Tipler 1986; Bostrom 2002). Earth is pretty special in its friendliness to complex life – so in one sense we do occupy a specially privileged spot. The Copernican Principle requires, then, a qualification: We are not in a specially privileged location *among locations capable of hosting intelligent observers*.¹⁰

⁹ We thank Nick Shea for emphasizing this issue in conversation.

¹⁰ A more nuanced statement of this principle would take the number or likelihood of observers into account. Plausibly, we should be more surprised to find ourselves in an

So, what is an “intelligent observer”? One possibility is: any behaviorally sophisticated entity in the sense of Section 2. This characterization has some plausibility: If we were to interact with them, we would naturally interpret them as making “observations”. We might then come to rely on these “observations” in testing our theories. Maybe members of an alien species send a signal naturally interpreted as conveying the observation of a supernova from their home planet on such-and-such a date. Employing this definition of an “observer”, our originally stated version of the Copernican Principle already has the required Anthropic qualification built in: “Among behaviorally sophisticated entities [i.e., “observers”], we are not specially privileged with respect to consciousness.”

Still, we have some sympathy with the reader who wants to insist that only conscious entities rightly count as “observers”. An Anthropic “observer” qualification would then generate a considerably weaker Copernican Principle of Consciousness: “Among *conscious* behaviorally sophisticated entities, we are not specially privileged with respect to consciousness.” This is a plausible principle, and it isn’t vacuous, for it still suggests that we wouldn’t have especially more or better consciousness. However, this principle is insufficient for our purposes. It would no longer capture the Copernican idea that it would be suspiciously lucky if we happened to have the right stuff for consciousness while almost all of our alien peers – similar in behavioral sophistication when viewed from outside – lack it.

Imagine switching observer and observed: Some alien entity (perhaps conscious) observes human beings and various other behaviorally sophisticated entities in other locations.

environment so hostile that few observers survive than in an environment with multitudes of thriving observers. However, formulating this idea precisely is challenging. See Bostrom 2002; Builes and Hare 2023; and discussions of the Sleeping Beauty problem and Doomsday Argument (cited above).

Absent some particular grounds for regarding us as special, this observing alien presumably ought to see us as just one of the bunch. This neutral, rather than Earth-centered, perspective is the correct, Copernican perspective. We shouldn't regard ourselves as special, unless there's some evidence that we are special that could be recognized from a neutral, outside perspective. Otherwise we err like the self-absorbed teen who regards their adolescent anguish and yearnings as somehow invisibly unique. Our original, unmodified Copernican Principle of Consciousness reflects this neutral perspective, while implicitly capturing an Anthropic qualification specified in terms of observers as behaviorally sophisticated entities.

Concern D. What justifies treating behaviorally sophisticated entities as the reference group, rather than, say, entities containing carbon or having two eyes? One might imagine a variant of the Copernican Principle of Consciousness according to which among entities containing carbon, we are not specially privileged with respect to consciousness. Such a principle would seem to prove too much since, plausibly, we are specially privileged with respect to consciousness, among *that* class of entities.¹¹

We have two avenues of response. One approach is to apply the Anthropic qualification: Among entities containing carbon *who are also observers*, we are not specially privileged with respect to consciousness. Arguably, this plausible principle is the appropriately Anthropically qualified carbon version of the Copernican principle.

Second, behavioral sophistication, unlike being made of carbon, directly targets the heart of the issue. If mainstream scientific theories are correct, we already know that we are special, with respect to our consciousness, among all entities containing carbon. Thus Copernican reasoning doesn't apply. Consider the parallel case concerning position: If we have good

¹¹ We thank Andrew Lee for emphasizing this issue in conversation.

evidence that we *are* at the center of the universe, we can't use the Copernican Principle to defeat that evidence. Copernican Principles only justify *default* assumptions absent countervailing evidence. Our focus on behavioral sophistication reflects the target question it is our project to address, and on which we lack direct evidence: not whether we are special among carbon-containing entities but rather whether very differently constructed, but behaviorally sophisticated, aliens would or would not be conscious.

We hope the reader finds the fundamental idea plausible. There would be something odd, something strikingly un-Copernican, in supposing that among all the thousands or millions or trillions of behaviorally sophisticated entities that have arisen in the universe, only we, or only we and a small proportion of others, are fortunate enough to be conscious, while the rest merely *seem* to feel love, to experience ecstasies of poetry, to share excited news of their discoveries, and to thrill with reverent awe before towers to their gods, but are in fact as experientially empty as toasters.

5. From the Copernican Principle to Default Liberalism about Consciousness.

The Copernican Principle of Consciousness supports the conclusion that we are not *specially* privileged with respect to consciousness. It doesn't follow that we aren't *slightly* privileged. A royal flush in a single draw rightly draws suspicions of funny business, but two pairs could easily enough be chance. Maybe only 10% of behaviorally sophisticated entities are conscious? Then we wouldn't have to be *too* lucky to be among them.

Let's define *Default Liberalism* about consciousness as the view that, absent specific grounds for doubt, when confronted with an apparently behaviorally sophisticated entity, we would be warranted in attributing consciousness. We propose that symmetry and simplicity

justify moving from the Copernican Principle of Consciousness to Default Liberalism. If some behaviorally sophisticated entities are in fact non-conscious, that shouldn't just be an arbitrary fact about them. We should expect there to be some striking difference between conscious and non-conscious aliens that constitutes specific grounds for doubt.

Suppose we discover an otherwise apparently behaviorally sophisticated alien species that lacks behavior suggestive of complex self-representations. Or suppose we discover a species that appears to lack flexible learning. It's not unreasonable to worry that such striking general deficits might reflect corresponding deficits in consciousness.¹² Our approach can handle these cases in one of two ways. We could define "behavioral sophistication" so as to exclude such entities: Behavior doesn't count as sophisticated unless it's suggestive of complex self-representation and flexible learning. Alternatively, we could liberally describe such entities as behaviorally sophisticated if their behavioral patterns are advanced enough in other ways. Doubt is justifiable either way we run the case: in the former, because they are not behaviorally sophisticated, in the latter, because the difference is plausibly important enough to cancel the default in default liberalism.

Here we can't avoid some broad commitments about what features plausibly do and do not justify canceling the default. Among the differences that plausibly matter are differences that suggest major deficits in cognitive function (relative to the class of Earthly entities assumed to be conscious). Without taking a stand on exactly *what* types of cognitive function track the existence versus absence of consciousness, virtually all mainstream scientific theories of

¹² On some theories of consciousness ("higher-order" theories and "unlimited associative learning" theories, respectively), self-representation and flexible learning are crucial to consciousness. For a review of higher-order theories, see Carruthers and Gennaro 2001/2023. On Unlimited Associative Learning, see Ginsburg and Jablonka 2019.

consciousness treat conscious states or their neural correlates as functionally significant in the production of intelligent or goal-directed behavior. This relationship needn't be exceptionless: A person with locked-in syndrome might be conscious with little or no behavior, and conversely a "consciousness mimic" (see discussion below) might achieve impressive behavioral sophistication without consciousness. But fundamentally, the reason we think rocks aren't conscious is that their behavior doesn't suggest a cognitively complex relationship with their environment. If we discover an alien world teeming with life, we will look for consciousness first among the entities whose behavior warrants the attribution of complex cognitive mechanisms and last among entities that passively sit. As we emphasized in the introduction, we aim to avoid commitment to particular scientific and metaphysical theories of consciousness, but committing to *some* sort of association between consciousness and cognitive function is plausible on a very wide range of mainstream theories.

Accordingly, we can distinguish *ungrounded* from *grounded* doubt about an entity's consciousness. If we discover that some arbitrarily encountered alien entity, as behaviorally sophisticated as we are, operates by hydraulics instead of electrical neurons, we might entertain some ungrounded doubt about their consciousness. Might hydraulics not actually support consciousness, as opposed to our wonderful neurons? The sophisticated behavior suggests consciousness, given the connection between complex behavior and consciousness in folk psychology and in most mainstream theories of consciousness; but we might still wonder if we are epistemically justified in making the leap of attributing consciousness to an alien entity built so differently from us. The Copernican Principle of Consciousness, combined with considerations of symmetry or simplicity, provide a defeasible, default reason to make that leap.

6. *Mimicry Arguments Against Robot Consciousness.*

We will now shift focus to what we will call Mimicry Arguments against Robot Consciousness. In the context of Copernican Default Liberalism about consciousness, Mimicry Arguments constitute specific grounds for doubt about the consciousness of a particular class of entities, which we will call *consciousness mimics*. We can be Copernican liberals about the consciousness of apparently behaviorally sophisticated entities in general while reserving skepticism about a subclass of mimics.

Skeptics about AI consciousness commonly warn against mistaking imitations of consciousness for the real thing. Emily Bender and colleagues (2021), for example, characterize Large Language Models, like ChatGPT, as “stochastic parrots” that copy the form of language with no understanding of its meaning. Johannes Jaeger (2023/2024) recommends replacing the phrase “artificial intelligence” with “algorithmic mimicry” to reflect the reality (as he sees it) that AI systems merely mimic intelligence without actually possessing it. John Searle argues that computers only “simulate” mental processes, and “no one supposes that a computer simulation of a storm will leave us all wet” (1984, p. 37-38). Kristin Andrews and Jonathan Birch (2023) warn us against too readily attributing sentience to AI systems that have been designed specifically to mimic human responses. Daniel Dennett (2023) warns against language models as “counterfeit people”.

Some of the best-known philosophical thought experiments meant to cast doubt on AI consciousness implicitly or explicitly suggest a mimicry structure. In Searle’s (1980) “Chinese room”, a person who does not know Chinese but who has an extremely large rulebook is hypothesized to interact with the outside world in a manner indistinguishable from a real Chinese speaker by accepting Chinese characters as inputs and issuing other Chinese characters as

outputs in accord with the advice of the rulebook. Block's (1978/2007) "Chinese nation" thought experiment (relatedly, the "Blockheads" of Block 1981) features a large group of people who control a body by consulting a bulletin board in the sky and following simple if-then rules so as to duplicate the functional structure of a real human being for an hour. Without entering into the details of these thought experiments or assessing how convincing they are, we note a common feature: The systems in question are specifically designed to imitate human behavior.

Mimicry arguments can at least sometimes be excellent grounds for doubt about the consciousness of an AI system, for the same reason that we might reasonably doubt that an actor is genuinely sad if we learn that they are imitating the superficial signs of sadness. However, advocates of mimicry arguments have not generally contextualized their arguments in a theory of mimicry that explains why this is so. The remainder of this essay addresses this challenge, then explores how to reconcile mimicry-based concerns about AI consciousness with our Copernican Principle.

7. Mimicry in General.

Suppose you encounter something that looks like a rattlesnake. One possible explanation is that it is a rattlesnake. Another is that it *mimics* a rattlesnake. Such mimicry can arise through evolution (other snake species that evolve the superficial appearance of a rattlesnake to discourage predators) or through human design (a rubber rattlesnake). Normally, it's reasonable to suppose that things are what they appear to be. But this default assumption can be defeated – for example, if there's reason to suspect sufficiently frequent mimics.

In biology, *deceptive mimicry* (for example Batesian mimicry) occurs when one species (the mimic) resembles another species or entity (the model) in order to mislead another species

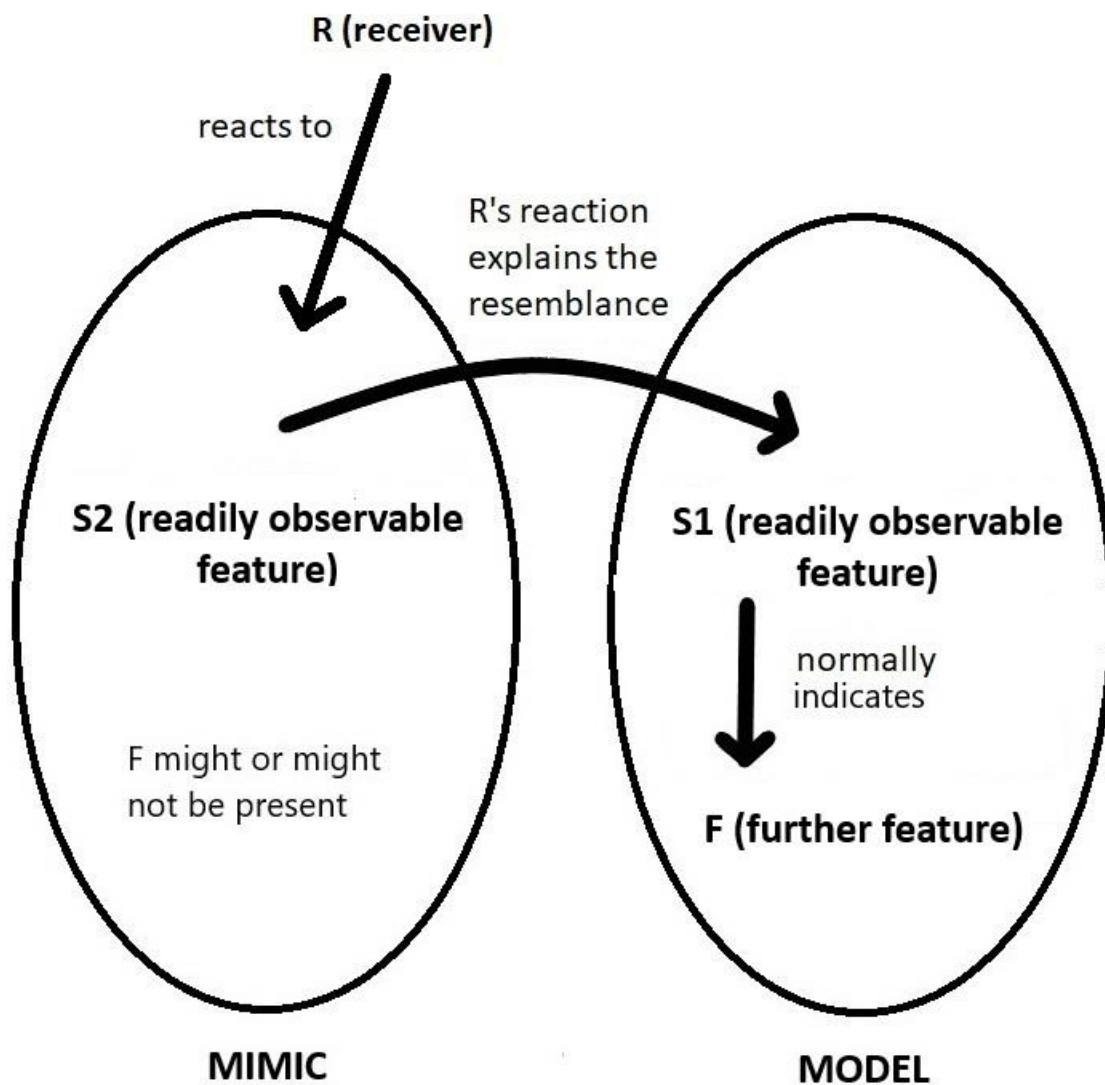
such as a predator (the receiver). For example, viceroy butterflies evolved to visually resemble monarch butterflies in order to mislead predator species that avoid monarchs due to their toxicity. Similarly, gopher snakes evolved to shake their tails in dry brush in a way that resembles the look and sound of rattlesnakes. Chameleons and cephalopods mimic their physical environment to mislead predators into treating them as unremarkable continuations of that environment. *Social mimicry* occurs when one animal behaves similarly to another animal for social advantage. For example, African grey parrots imitate each other to facilitate bonding and to signal in-group membership, and their imitation of human speech arguably functions to increase the affection of human caretakers. In deceptive mimicry, the superficially imitated feature normally does not correspond with the model's relevant trait. The viceroy is not toxic, and the gopher snake has no poisonous bite. In social mimicry, even if there is no deceptive purpose, the signal might or might not correspond with the trait suggested by the signal: the bird might or might not belong to the group it is imitating, and Polly the parrot might or might not really "want a cracker".¹³

All mimicry thus involves three traits: the readily observable trait (S2) of the mimic, the corresponding observable trait (S1) of the model, and a further feature (F) of the model that is normally indicated by the presence of S1 in the model. (In the Polly-want-a-cracker case, F might be the intention to utter a sentence with that propositional content.) All mimicry also involves three protagonists: a model, a mimic, and a receiver whose pattern of reaction to S2 explains the resemblance of S2 to S1. Crucially, in mimicry, the resemblance of S2 to S1 is explained by a potential receiver's reaction to that resemblance, given that the receiver treats S1

¹³ For reviews of the biology of mimicry, see Jamie 2017; Anderson and de Jager 2020.

as normally indicating F, rather than being explained directly by the presence of F is the mimic. See FIGURE 1.

FIGURE 1: The mimic's possession of a readily observable feature S2 is explained by its resemblance to observable feature S1 in the model, because of how a receiver, who treats S1 as indicating F, responds to the resemblance of S1 and S2. S1 reliably indicates F in the model, but S2 need not reliably indicate F in the mimic.



Six clarifications:

(1.) The protagonists need not be distinct. For example, in parrots' social mimicry the model might also be the receiver. Alternatively, if you mimic someone just to delight yourself with your mimicry skills, then you are both mimic and receiver.

(2.) The receiver need not be fooled. The parrot owner doesn't mistake Polly's speech for human speech. In the wild, the parrot's mate does not mistake its partner's singing for its own.

(3.) Although in the simplest cases of deceptive mimicry, the receiver will react similarly to S1 and S2, the receiver might in principle react very differently to S1 and S2 if the receiver is capable of distinguishing them. A parrot might react very differently to its own song than to its mate's imitation of its song. A public clown might mimic a depressed person's gait, generating appreciation and laughter in the audience. Mimicry requires only that the resemblance of S2 to S1 be explained by the reactions of potential receivers, given that the receivers treat S1 as an indicator of F.

(4.) The mimic need not lack feature F. The viceroy might happen to be toxic to some predator species. As long as S2 in the mimic is better explained by its resemblance to S1 than by the presence of F in the mimic, it's still mimicry.

(5.) Emulation or imitation alone is insufficient for mimicry. One animal might imitate another for any of a variety of reasons, for example, to duplicate their success in a challenging task. Mimicry distinctively requires a specific relationship to a receiver.¹⁴

(6.) Mimicry involves a gap between S1 and F. If I imitate your holding up three fingers by also holding up three fingers, I have copied or imitated you, but I have not necessarily (in our

¹⁴ On emulation and imitation, see Whiten et al. 2009.

intended sense) mimicked you. If the three fingers are intended as a signal of membership in the Cool Kids Club and I raise them in imitation of you so as to trick the doorkeeper, then I have mimicked you.¹⁵

Mimicry in the intended sense is thus a specific, causally complex relationship in which S2 in the mimic is *better explained* by a complex relationship involving model, mimic, receiver, and the receiver's tendency to treat S1 as an indicator of F in the model than by a direct relationship between S2 and F in the mimic. Mimicry arguments against the consciousness of AI systems rely, implicitly or explicitly, on the presence of this causal structure.

8. Mimicry in Artificial Intelligence.

Consider a child's toy that says "hello" when powered on. We can treat this as a simple case of mimicry. English-speaking humans are the models, the toy is the mimic, S1 and S2 are the sound "hello", F is the intention to greet, and the receiver is a child who reacts by hearing the sound as an English "hello". In human models, "hello" normally (though not perfectly) indicates an intention to greet. But of course the child's toy has no intention to greet. (Maybe the toy's designer, years ago, intended to craft a toy that would "greet" its user when powered on, but that isn't the *toy's* intention.) S2 is better explained by having been modeled on S1, because of an

¹⁵ Readers of evolutionary accounts of mimicry might wonder whether "honest" mimicry fits into the present account. In evolutionary biology, Müllerian mimicry occurs when a toxic species mimics the pattern of another toxic species to more effectively signal toxicity to predators who can learn the signal more efficiently given the similarity of mimic and model (Sherratt 2008). Similarly, members of the local parrot group might reliably signal group membership by repeating the same calls. Honest mimicry differs from ordinary honest signaling because the bulk of the causal explanation of S2 depends on S1's rather than S2's relation to F. Such cases also suggest that mimicry can be a matter of degree and/or symmetric.

anticipated receiver who hears the toy’s sound as similar to a human “hello”, than as an indicator of F.

Large Language Models like GPT, PaLM, LLaMA, and Claude are more complex but are – at least in their simpler instantiations – structurally mimics, as has been argued by, for example, Bender (Bender and Koller 2020; Bender et al. 2021), Millièrè (2022), and Jaeger (2024). Our account of mimicry renders this claim more precise. If you ask ChatGPT “What is the capital of California?”, it might respond “The capital of California is Sacramento.” Treating this text string as S2, the appropriate explanation of why ChatGPT exhibits S2 is that its outputs are modeled on human-produced text that also correctly identifies the capital of California as Sacramento, designed with a receiver in mind who will interpret that output linguistically. While human-produced text with that content (S1) reliably indicates the producer’s knowledge that Sacramento is the capital of California (F), we cannot infer the presence of F from S2 – not, at least, without substantial (and in this case dubious) further argument.

When a simple Large Language Model outputs a novel sentence not present in the training corpus, S2 and S1 will need to be described more abstractly (e.g., “a summary of Hamlet” or even just “text interpretable as a sensible answer to an absurd question”), but the underlying considerations are the same. The LLM’s output is modeled on patterns in human-generated text and can be explained as mimicry of those patterns, leaving open the question of whether the LLM has the further features we would attribute to a human being who gave a similar answer to the same prompt. To the extent an LLM is further shaped by reinforcement that rewards answers to the extent they seem humanlike, the mimicry structure is reinforced.

We can call something a *consciousness mimic* if it exhibits readily observable features suggestive of consciousness whose presence is best explained by having been modeled on

similar observable features of another system, for the sake of a receiver who responds to those features in a way that reflects the fact that they treat the features, when they occur in the model, as indicating a conscious state or capacity. Note that this account of mimicry can apply to both evolved and intentionally designed or artificial beings: all that is required is that the etiology of the feature appeals to the right kind of modelling.

ChatGPT and the “hello”-toy are consciousness mimics in this sense. Their observable language-like features (their S2s) are modeled on us, for the sake of a certain sort of receiver with a certain set of interests and background knowledge. In us, the modeled features (the S1s) reliably indicate the further feature (F) of consciousness (humans who say “hello” or answer questions about the capital of California are normally conscious), but there needn’t be any such reliable connection in a consciousness mimic. Most “social AI” programs, like Replika, are similarly consciousness mimics, combining the structure of Large Language Models with superficial signals of emotionality through an avatar with an animated, expressive face. This adds detail to the mimicry but does not change the fundamental mimicry relation.

In our current technological climate, any AI system that outputs a diverse range of interpretable text or speech will almost inevitably rely on mimicry. The best explanation of the shape of the text strings or the sound structure will involve modeling the superficial features (such as word co-occurrence relationships) of another entity (us) for the sake of a receiver (also us) who interprets them in accord with their understanding of linguistic capacities in model cases. Their S2s are modeled on human S1s for the sake of interpretability by human receivers, with no structure that ensures the presence of F in the mimic.

Contrast this with children’s language learning. Children learn language through imitation, and sometimes mimicry, but the primary structure is not mimicry. If you say “blicket”

in the presence of a blicket and the child repeats “blicket” back to you, their behavior arises in part from their (presumably conscious) appreciation of the fact that the novel word refers to the novel object (Sobel, Yoachim, Gopnik, Meltzoff, and Blumenthal 2007). Their S2 resembles your S1, for the sake of a receiver (you), but S2 indicates their own underlying F, and you rightly interpret it as such. Children’s crying, demands for food, and insistent “no” are not S2s explained by S1s linked to Fs in the model. Rather, they are direct signals (normally) of genuine unhappiness, desire, and reluctance. Similarly, Alex the parrot’s utterances of “four corner paper” in the presence of square pieces of paper – though drawing on the same capacities that underlie parrot mimicry – function as true signals in the ordinary, direct sense (Pepperberg 1999). Childhood linguistic mimicry occurs, but it is different from ordinary language learning. One example would be in shared pretense, when a child says, “I’ve got to hurry or I’ll be late to work!”, theatrically grabbing a pretend briefcase.

The situation becomes more complicated as the models become more complicated. To the extent models are trained not to behave like humans but rather to give the “right” answers or to act successfully toward goals, and to the extent they incorporate elements like maps and calculators that have representational rather than mimicry functions, they diverge from being pure mimics. The Mimicry Argument might still apply to the extent their fundamental structure relies on superficial mimicry of S1s without reproduction of the architectures that produce the S1s in the human case – but its application becomes more complicated. They are no longer 100% “parrot” (or underground octopus¹⁶). However, they are still beings whose properties are designed to mimic, perhaps not consciousness directly, but other properties of human beings whose possession presupposes consciousness.

¹⁶ Bender and Koller 2020; Bender et al. 2021.

9. *The Structure of the Mimicry Argument.*

The default in Default Liberalism about the consciousness of behaviorally sophisticated entities can thus be canceled in at least two ways: (a.) if the behavior suggests a sufficiently important cognitive deficit, or (b.) if the entity is a consciousness mimic.

In particular, the Mimicry Argument can be expressed as follows:

- (1.) In cases of mimicry, the appearance of S2 in the mimic is not normally sufficient grounds to infer the presence of F in the mimic.
- (2.) An important range of AI systems are consciousness mimics.
- (3.) Therefore, in this range of AI systems, we cannot infer consciousness from the (S2) features that, were their analogs (S1) to occur in humans, would normally indicate conscious states (F).

As discussed, S2 mimicry traits aren't *incompatible* with the existence, in the mimic, of the F traits normally indicated by S1s in the model. A viceroy might be toxic; a parrot mimicking a group signal might belong (in fact, usually does belong) to the targeted group. But establishing that requires further evidence or argument. In the case of current and near-future AI systems, no further evidence or argument is likely to merit scientific consensus on their consciousness.

The Mimicry Argument doesn't establish that consciousness mimics are definitely non-conscious. It only undercuts the inference from superficial appearance (S2) to underlying reality (F). Perhaps a compelling argument can be made for AI consciousness in some range of current

or near-future cases; but given the degree of uncertainty in consciousness science, we are disinclined to think any such argument would warrant scientific and philosophical consensus.¹⁷

10. Are AI Systems “Behaviorally Sophisticated” in the Sense Relevant to the Copernican Principle?

In Section 2, we described an entity as “behaviorally sophisticated” if it is capable of complex goal-seeking, complex communication, and complex cooperation, where “goal-seeking”, “communication”, and “cooperation” are all understood in purely behavioral terms, without commitment to the existence of consciousness or any particular underlying cognitive architecture. Are current or near-future AI systems behaviorally sophisticated in this sense, and if so, would the Copernican Principle of Consciousness then apply to them?

For alien cases, we favor a somewhat relaxed sense of “behavioral sophistication”. As contemplated in Section 5, differently structured entities might be capable of approximately human-level goal-seeking, communication, and cooperation while having cognitive deficits that we would find shocking (e.g., major deficits in learning capacities or self-representational capacities). We human beings also have remarkable cognitive weaknesses. Consider the Wason Selection Task. In this task, you are presented with four cards. Each card, you are told, has a letter on one side and a number on the other side. Two cards are letter-face up: “A” and “C”. The other two are number-face up: “4” and “7”. What card or cards must you flip over to determine whether the rule “If a card has a consonant on one side, it has an even number on the other side” is true? The large majority of ordinary people, even if highly educated, fail to

¹⁷ For discussion, see Butlin et al. 2023; Schwitzgebel 2024.

recognize the correct answer as “C” and “7”.¹⁸ We wouldn’t want an alien species kicking us out of the club of behaviorally sophisticated entities because of our – possibly to them shocking – weakness in simple formal logical reasoning. Entities with different cognitive structures will very likely have very different patterns of cognitive strength and weakness. A proper Copernican perspective won’t insist that alien intelligence look too much like our own.

Applying this same liberality to AI cases, we don’t want to insist that present and realistic near-future AI systems as they exist on Earth lack approximately human-level behavioral sophistication. On some tasks – such as walking across uneven surfaces and identifying objects in blurry, partly occluded photographs – human beings (so far) substantially exceed AI systems. On other tasks – such as playing chess and multiplying large numbers – AI systems substantially exceed even expert humans (interpreting “playing chess” and “multiplying large numbers” purely behaviorally). It might be tempting for opponents of AI consciousness to rest their case on insisting that AI systems are not as sophisticated as human beings in some crucial range of tasks. This is the “cognitive deficit” grounds for canceling Default Liberalism. However, this move relies on correctly specifying the relevant tasks, and correctly specifying the relevant tasks is a challenge. AI enthusiasts with some justification object to a history of moving goalposts. Back when success in chess seemed like the pinnacle of human expertise, skeptics about AI intelligence sometimes held chess up as a special sign of human intelligence. The Turing-inspired emphasis on the ability to hold humanlike conversation also increasingly looks like it might be a poor indicator, or at least one that will have to be specified in a highly precise manner that is potentially difficult to justify. The Mimicry Argument allows AI skeptics to express their skepticism about AI consciousness while (a.) conceding that AI might be overall as behaviorally

¹⁸ See Wason 1968 and the large subsequent literature.

sophisticated as humans and (b.) without committing to any particular type of especially revelatory behavioral benchmark that AI cannot pass or any specific theory about the cognitive structures associated with consciousness. These are, we submit, attractive features of the Mimicry Argument.

Could the pressures on a consciousness mimic lead it to develop genuine consciousness? The case for *yes* depends on the interests and sophistication of the receiver. If the pressures of natural selection are severe and specific enough to generate complex goal-seeking, communication, and cooperation of the sort described in Section 2, those same pressures will almost certainly produce complex cognitive structures of the sort practically required to sustain goal-seeking, communication, and cooperation – for example, long-term memory, self-representation, and linguistic understanding. Fantastical creatures, such as “Blockheads” (Block 1981) might be behaviorally identical to a naturally selected intelligent organism, but such things could never evolve: The most efficient way to reliably act, over evolutionary time, as if one has a long-term memory is to actually have a long-term memory. In contrast, the behavioral sophistication of a consciousness mimic need only be good enough to achieve the ends of mimicry – for example, to fool or delight a receiver. We are often delighted by obviously-less-than-perfect mimics (the “hello”-toy) and, as the recent history of language models suggests, more sophisticated mimics can sometimes easily fool us.

Different types and degrees of similarity between S2 and S1 are required for different mimicry relationships. Deceptive mimicry can become an arms race between a mimic motivated to appear similar to the model in the eyes of the intended dupe and an intended dupe motivated to acquire the capacities to distinguish mimic from model. In the limit, the mimic could only “fool” a god-like receiver by actually acquiring feature F. Fortunately for the mimics, we are far from

god-like. We thus see no reason to suppose that consciousness would be required for successful mimicry when receivers are imperfectly discerning or when the motive isn't deception. As current large language models show, some impressive feats of mimicry can be achieved without long-term memory or sophisticated self-representation. These are major cognitive differences that, even independently of mimicry considerations, might trigger reasonable cancellation of Default Liberalism as described in Section 5.

There is a reason most features of most evolved organisms aren't explained by mimicry. Mimicry is a complex relationship that requires a special set of evolutionary pressures to evolve and sustain, often parasitic on the simpler true signals or cues of more prevalent non-mimic species. If viceroy butterflies outnumber monarchs (as they now do in California) the evolutionary advantage of mimicking the monarch's wing coloration decreases. And technologically, we speculate, mimicry structures might not prevail in the long term. In general, human design goals might be met more fully and efficiently by creating systems that really have F than by creating systems that imitate the superficial signs of F. If we create an artificial entity that really has underlying linguistic representations of real-world events and a conscious understanding of itself and its interlocutors (whatever architecture that requires), that will presumably be a more reliable and satisfying interaction partner, for a broad range of purposes, than today's Large Language Models. Similarly, if our space alien neighbors send behaviorally sophisticated robotic probes across the galaxy, it's a reasonable guess that those probes will be designed to explore, measure, experiment, communicate, self-repair, respond to emergencies, and perhaps devise theories on the fly that guide their exploratory choices, rather than that they will mainly be designed as mimics to fool or delight some receiver by virtue of their resemblance to some model entity.

Since mimicry traits evolve less commonly than non-mimicry traits, and since the evolution or design of mimicry traits requires a highly specific set of purposes or conditions, we suggest that it is reasonable to default assume that a newly encountered behaviorally sophisticated extraterrestrial entity will not be a consciousness mimic, even if it has a technological origin – though we might hold this default assumption only lightly. The crucial epistemic distinction between mimicry and non-mimicry explanations of behavioral sophistication crosscuts distinctions between design vs. evolution and artifice vs. biology.

11. Rejecting the Parity Principle.

We sympathize with those who reject a simple sort of organic chauvinism to argue that AI isn't conscious. The most obvious alternative to organic chauvinism is one that treats alien and robot cases symmetrically – holding, for example, that the same external, behavioral criteria should be applied to both aliens and robots in assessing them for consciousness. If Behavioral Test X would be sufficient to warrant attribution of alien consciousness, then – on this way of thinking – attribution of consciousness would also be warranted to robots who similarly pass Behavioral Test X. Similarly, on this way of thinking, theorists who emphasize interior architecture should attribute consciousness to aliens and robots based on the same internal, architectural criteria. If Architecture Y would warrant attribution of consciousness to aliens, then attribution of consciousness would also be warranted to any robot possessing Architecture Y. Mixed behavioral/architectural views might also apply identical criteria to aliens and robots. According to what we'll call the *Parity Principle*, in evaluating an alien or robot for consciousness, we should apply the same behavioral and/or architectural criteria.

We acknowledge the potential appeal of the Parity Principle. But to accurately assess this principle, we need to distinguish it from a closely related metaphysical thesis.

Metaphysically, if an alien and a robot have identical structures with respect to what matters for consciousness according to the correct theory of consciousness (whatever that may be), then it cannot be the case that one but not the other has consciousness. This metaphysical parity thesis is highly plausible.

The Parity Principle we reject differs from the plausible metaphysical thesis in two respects. First, the Parity Principle is epistemic rather than metaphysical. It concerns not what must be true if robots and aliens share whatever behavioral or architectural features matter to consciousness; rather it concerns what we are *warranted in believing* about alien and robot consciousness *given our ignorance* about the correct theory of consciousness. Second, it applies independent of stipulations about the internal workings of the entities involved.

These two differences matter. Even if we know virtually nothing about the internal workings of a particular (individual or type of) behaviorally intelligent alien, we can apply the Copernican Principle, at least by default (that is, assuming we have no evidence it is a consciousness mimic or in some other respect likely to be a member of a disprivileged class with respect to consciousness). Even if we know virtually nothing about the internal workings of a particular (individual or type of) robot or AI system, if we know that the superficial features suggestive of consciousness arise from a mimicry relationship of the sort described in Section 7, we are licensed to withhold inference to consciousness, barring some positive evidence that the superficial features are in fact coupled with consciousness.

The Copernican and Mimicry Arguments yield different results for alien and robot systems not because of some intrinsic inferiority of engineered to naturally evolved life but

rather because aliens and a wide range of current and near-future robots – normally or as a default presumption – have very different *histories*. Consequently, very different explanations of their structure and behavior can be warranted. Aliens, normally or as a default presumption, are naturally evolved entities not modeled on other entities (or if designed, not specifically designed as mimics). In contrast, in present and near-future cases, the most convincing robot conversationalists will be consciousness mimics. Contra the Parity Thesis, this difference generates an epistemic asymmetry in attribution of consciousness to aliens and robots, even if the aliens and robots have similar outward behavior or similar interior architectures (at some level of descriptive specificity).

This is not to endorse the teleofunctionism of, for example, Millikan (1984), Dretske (1995), and Neander (2017). On a teleofunctional view of mental properties, etiology also matters: Donald Davidson has a mind and his molecule-for-molecule identical chance-generated Swampman duplicate lacks a mind because of their different origins. But such teleofunctionalism concerns the metaphysical relationship of mental properties to their physical realizers at a given time. Our approach is neutral about the metaphysical importance of an entity's history while remaining committed to the history's epistemic importance.

12. Generalizing the Copernican Argument to Less Sophisticated Aliens.

In Section 2, we advanced a high bar for “behavioral sophistication” because we were seeking a sufficiency criterion for default attributions of alien consciousness which would be appealing to a broad range of consciousness theorists, from the very liberal (who hold that consciousness is widespread on Earth) to the very conservative (who hold that consciousness is limited to just one or a few species). But this is a movable parameter in our theory. Our

Copernican argument can be adapted to animal cases involving lower standards of behavioral sophistication as follows.

Barring some reason to think we are exceptional, Earth would be a strangely lucky place if only Earth has conscious non-linguistic life forms and in almost every other region of the observable universe, similarly complex or sophisticated life forms were non-conscious. Thus, a generalization of the Copernican Principle is possible. With respect to whatever types of animal life you are justified in regarding as conscious on Earth (and different readers might reasonably favor different views), it is reasonable to default assume that similarly complex or sophisticated alien life forms would also be conscious.¹⁹

¹⁹ For helpful discussion, thanks to Brice Bantegnie, Cameron Buckner, Nathan Buss, David Chalmers, Benjamin Icard, Andrew Lee, Matthew Liao, Michael Martin, Raphaël Millière, Melissa O'Neill, William Robinson, Nick Shea, Stephan Schmid, and Anna Strasser; audiences at Trent University, Harvey Mudd, New York University, the Agency and Intentions in AI conference in Göttingen, Jagiellonian University, the Oxford Mind Seminar, University of Lisbon, NOVA Lisbon University, University of Hamburg, and the Philosophy of Neuroscience/Mind Writing Group; and commenters on relevant posts on social media and at The Splintered Mind.

References:

- Anderson, Bruce, and Marinus L. de Jager (2020). Natural selection in mimicry. *Biological Reivews*, 95, 291-304.
- Andrews, Kristin, and Jonathan Birch (2023). What has feelings? *Aeon* (Feb 23)
<https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals>.
- Barrow, John D., and Frank J. Tipler (1986). *The anthropic cosmological principle*. Oxford University Press.
- Bender, Emily M., and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185-5198.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
- Birch, Jonathan (2022). The search for invertebrate consciousness. *Noûs*, 56, 133-153.
- Block, Ned (1978/2007). Troubles with functionalism. In N. Block, *Consciousness, function, and representation*. MIT Press.
- Block, Ned (1981). Psychologism and behaviorism. *Philosophical Review*, 90, 5-43.
- Bondi, H. (1968). *Cosmology*. Cambridge University Press.
- Bostrom, Nick (2002). *Anthropic bias*. Routledge.
- Builes, David, and Caspar Hare (2023). Why aren't I part of a whale? *Analysis*, 83, 227-234
- Butlin, Patrick, Robert Long, Eric Elmozino, et al. (2023). Consciousness in Artificial Intelligence: Insights from the science of consciousness. ArXiv:2308.08708.

- Caldwell, R. R., and A. Stebbins (2008). A test of the Copernican Principle. *Physical Review Letters*, 100, 191302.
- Camarena, David, Valerio Marra, Ziad Sakr, and Chris Clarkson (2022). The Copernican principle in light of the latest cosmological data. *Monthly Notices of the Royal Astronomical Society*, 509 (1), 1291-1302.
- Carruthers, Peter, and Rocco Gennaro (2001/2023). Higher-order theories of consciousness. *Stanford Encyclopedia of Philosophy* (fall 2023 edition).
- Carter, Brandon (1974). Large number coincidences and the anthropic principle in cosmology. In M.S. Longair, ed., *Confrontation of cosmological theories with observational data*. D Reidel.
- Chalmers, David J. (2023). Could a Large Language Model be conscious? ArXiv:2303.07103.
- Clarkson, Chris, Bruce Bassett, and Teresa Hui-Ching Lu (2008). A general test of the Copernican Principle. *Physical Review Letters*, 101, 011301.
- Dehaene, Stanislas, Hakwan Lau, and Sid Kouider (2021). What is consciousness, and could machines have it? *Science*, 358 (6362), 486-492.
- Davidson, Donald (1984). *Inquiries into truth and interpretation*. Oxford University Press.
- Dennett, Daniel C. (1987). *The intentional stance*. MIT Press.
- Dennett, Daniel C. (2023). The problem with counterfeit people. *The Atlantic* (May 16). <https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075>.
- Dretske, Fred (1995). *Naturalizing the mind*. MIT Press.
- Elga, Adam (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60, 143-147.

- Frank, A., and W.T. Sullivan, III (2016). A new empirical constraint on the prevalence of technological species in the universe. *Astrobiology*, 16, 359-362.
- Frankish, Keith (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23 (11-12),11-39.
- Ginsburg, Simona, and Eva Jablonka (2019). *The evolution of the sensitive soul*. MIT Press.
- Godfrey-Smith, Peter (2016). Mind, matter, and metabolism. *Journal of Philosophy*, 113, 481-506.
- Gott, J. Richard, III. (1993). Implications of the Copernican principle for our future prospects. *Nature*, 363, 315–19.
- Harnad, Stevan (2003). Can a machine be conscious? How? *Journal of Consciousness Studies*, 10 (4-5), 67-75.
- Jaeger, Johannes (2023/2024). Artificial intelligence is algorithmic mimicry: Why artificial “agents” are not (and won't be) proper agents. ArXiv:2307.07515.
- Jamie, Gabriel A. (2017). Signals, cues and the nature of mimicry. *Proceedings of the Royal Society B*, 284 (1849): 20162080. doi: 10.1098/rspb.2016.2080.
- Kammerer, François (2019). The illusion of conscious experience. *Synthese*, 198, 845-866.
- Kriegel, Uriah (2019). The value of consciousness. *Analysis*, 79, 503-520.
- Lacki, Brian C. (2023). The Noonday argument: fine-graining, indexicals, and the nature of Copernican reasoning. *International Journal of Astrobiology*, 22, 354–398.
- Lee, Geoffrey (2019). Alien subjectivity and the importance of consciousness. In A. Pautz and D. Stoljar, eds., *Blockheads!* MIT Press.
- Levy, Neil (2024). Consciousness ain't all that. *Neuroethics*, 17 (21), <https://doi.org/10.1007/s12152-024-09559-0>.

- Millière, Raphaël (2022). Moving beyond mimicry in artificial intelligence. *Nautilus* (Jul 1).
<https://nautil.us/moving-beyond-mimicry-in-artificial-intelligence-238504>.
- Millikan, Ruth (1984). *Language, thought, and other biological categories*. MIT Press.
- Neander, Karen (2017). *A mark of the mental*. MIT Press.
- Papineau, David (2024). *Consciousness is not the key to moral standing*. Unpublished manuscript.
- Pepperberg, Irene M. (1999). *The Alex studies*. Harvard University Press.
- Scharf, Caleb (2014). *The Copernicus complex*. Farrar, Straus and Giroux.
- Schneider, Susan (2019). *Artificial you*. Princeton University Press.
- Schwitzgebel, Eric (2016). Phenomenal consciousness, defined and defended as innocently as I can manage. *Journal of Consciousness Studies*, 23 (11-12), 224-235.
- Schwitzgebel, Eric (2024). *The weirdness of the world*. Princeton University Press.
- Searle, John R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 417-457.
- Searle, John R. (1984). *Minds, brains, and science*. Harvard University Press.
- Sebo, Jeff, and Robert Long (2023). Moral consideration for AI systems by 2030. *AI and Ethics*. <https://doi.org/10.1007/s43681-023-00379-1>.
- Seth, Anil K. (2021). *Being you*. Dutton.
- Seth, Anil K., and Tim Bayne (2022). Theories of consciousness. *Nature Reviews Neuroscience*, 23, 439-452.
- Shepherd, Joshua (2024). Sentience, Vulcans, and zombies: The value of phenomenal consciousness. *AI & Society*, <https://doi.org/10.1007/s00146-023-01835-6>.

- Sherratt, Thomas N. (2008). The evolution of Müllerian mimicry. *Naturwissenschaften* 95, 681-695.
- Shevlin, Henry (2020). General intelligence: An ecumenical heuristic for artificial consciousness research? *Journal of Artificial Intelligence and Consciousness*, 7, 245-256.
- Silver, David, Thomas Hubert, Julian Schrittwieser, et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362, 1140-1144.
- Snyder-Beattie, Andrew E., Anders Sandberg, K. Eric Drexler, and Michael B. Bonsall (2021). The timing of evolutionary transitions suggests intelligent life is rare. *Astrobiology*, 21, 265-278.
- Sobel, David M., Caroline M. Yoachim, Alison Gopnik, Andrew N. Meltzoff, A. N., and Emily J. Blumenthal (2007). The blinket within: Preschoolers' inferences about insides and causes. *Journal of Cognition and Development*, 8, 159-182.
- Turing, Alan M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.
- Udell, David Billy, and Eric Schwitzgebel (2021). Susan Schneider's proposed tests for AI consciousness: Promising but flawed. *Journal of Consciousness Studies*, 28 (5-6), 121-144.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.
- Whiten, Andrew, Nicola McGuigan, Sarah Marshall-Pescini, and Lydia M. Hopper (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B*, 364, 2417-2428.