Mind-body dualism as social signalling

Matan Mazor [1]

Why do humans — animals that are controlled by a three-pound brain enclosed in a hard bone skull — have the intuition that their experience is immaterial, that they are more than their physical body, and that they are free agents? The answer may be found in the critical role these intuitions play in shaping shared norms around moral status, fairness, and moral responsibility. I propose a mechanism for how these private intuitions may come about, based in social signalling and strategic self-deception.

An alien visiting Earth would notice a curious fact about humans. On the one hand, their movements are fully controlled by a central brain, located inside their head, which also processes input from their sense organs. And still, many of them talk as if there is a nonphysical aspect to their internal computations (what they sometimes call their *consciousness*). Moreover, many of them doubt that it will ever be possible to understand this non-physical aspect by studying the body. After reviewing relevant historical documents, the alien may also note that for ages this non-physical aspect of the mind was believed in many societies to be entirely unique to humans, but now they say other creatures have it too — definitely dogs and apes but maybe also birds and squids. And, as of late 2025, most of them say machines don't have it (even machines that say they do).

As a human, I too share this peculiar set of beliefs. And yet, my challenge in this essay is to detach myself from my first-person perspective and provide a satisfactory account of the mystery not as it appears to me from the inside, but as it may appear to an alien from the outside. Why do humans have these strange intuitions about their minds?

Let me name three particularly hard-to-resist ones.

First, there is *something it is like* to be me, and this something feels like it cannot be reduced to electrical signals in my brain. I can't imagine how a better un-

Affiliation [1] All Souls College and Department of Experimental Psychology, University of Oxford
Correspondence matan.mazor@all-souls.ox.ac.uk
Version September 16, 2025

derstanding of the brain would bridge the gap between facts about physiology and facts about my private experiences and feelings. My experience feels like a fundamental property of the universe that cannot be reduced to other aspects of existence, physical or not.

Second, it feels as if I could have had an entirely different body, different beliefs and different preferences, and still somehow be *me*. In other words, it feels like the coupling of my unique subjective point of view with the person who is the author of this essay, with his particular body and personal history, is not a necessary fact (like that a triangle has three sides and not four), but a contingent one (like that my desk is facing east and not south).

Finally, despite my training in neuroscience, I cannot resist the intuition that my actions are a product of more than electrochemical events in my brain. It feels like I have some control over my actions: my brain may restrict the options that are available to me at a given time, or may bias me in one way or another, but at the end of the day it is *I* who decides when and how to act.

These are intuitions about the irreducibility of different aspects of my consciousness (of my experience, of my first-person point of view, and of my free will) to more fundamental facts about the universe. In short, we can refer to them respectively as the *irreducibility of experience*, the *irreducibility of self*, and the *irreducibility of will*. This triangle forms a conception of the mind that is not too unlike Pixar's *Inside Out*, with a small person sitting inside the head, observing the world through a big screen and keeping its little hand on the control panel of motor actions. This intuitive conception of the mind is sometimes referred to as intuitive mind-body dualism, or intuitive dualism for short.

Intuitive dualism is at odds with modern-day neuroscience, which is overwhelmingly *physicalist*, explaining human psychology entirely in terms of brain processing. But, as the alien observer would note, dualistic beliefs are as universal as beliefs get: they appear across cultures and traditions — even among self-declared physicalist scientists — and emerge early in childhood.

In modern Western philosophy, the idea that the mind is more than the body is strongly associated with René Descartes, who argued that consciousness involves a *res cogitans*: a thinking thing that cannot be reduced to physical matter. The association is so strong that the view is often referred to as "Cartesian dualism," from Descartes' Latin name, Renatus Cartesius. But the idea that experience transcends the physical body appears much earlier and across diverse philosophical

traditions.

600 years before Descartes, for example, Persian philosopher Ibn Sina presented the "floating man" thought experiment in which a person comes into existence suspended in full darkness, existing without any sensory input and without any control over a physical body. The floating man would experience, according to Ibn Sina, consciousness in its pure, immaterial form, proving that the soul (*Nafs*) is immaterial. Likewise, in Hinduism, the *Ātman* is a pure self, entirely immaterial, eternal and distinct from the body, which wanders between bodies and identities in the cycle of *Saṃsāra*. Plato, in the Phaedo (4th century BCE), argued that the soul is distinct from the body because it must exist before birth. Hieroglyphs from as early as 3,000 BCE reveal that ancient Egyptians too saw a person as more than their physical body (*Khat*), involving, among other parts like a name and a shadow, a spiritual essence (*Ka*).

This intuitive belief in immaterial subjective points of view, spirits, or essences, is evident in how we speak and think today. Consider the idea of being born in the wrong body, often used to describe the experience of gender dysphoria. The expression resonates with the notion that an immaterial self could have ended up in a different physical container. We intuitively say things like "my brain is playing tricks on me," as if the *me* in question is distinct from the brain. Even the phrase "my brain" implies a distinction between a brain and its owner: a linguistic trace of dualistic thinking.

Of course, not everyone is a dualist. For example, some people firmly believe that the mind is identical with the physical brain, or with certain aspects of its functioning. Crucially, however, traces of dualism can be discovered even among such self-declared physicalists, including in the way that physicalist neuroscientists describe their findings. A textual analysis of research papers and textbooks conducted by cognitive scientists Uri Maoz and Liad Mudrik revealed multiple examples of such tacit dualism, such as "The brain decides when you will experience pain" (who is the "you" in this sentence?) and "If a person wants to move a limb to a new position, the brain decides exactly what muscle fibers to use with what force" (who is the person?).

More evidence for how deeply these intuitions go comes from developmental research done with children. In one experiment by psychologist Bruce Hood and his colleagues, five- and six-year old children happily accepted that a noisy machine can duplicate objects, that it can duplicate a live hamster, and that the

duplicate hamster shares particular physical attributes with the original, such as its broken tooth and the marble in its belly. But they were much less likely to believe that the duplicate hamster — physically identical to the original — will share the original's memories and beliefs. In children's intuitive understanding, memories and beliefs are not determined by the body.

Importantly, a physicalist account of human psychology should be able to explain these social practices, ways of speaking and experimental results as features of human psychology. If the mind *is* the brain, what is it about the human brain that makes it express these dualistic intuitions?

The meta-problem of consciousness

In his 1995 paper "*Facing Up to the Problem of Consciousness*", philosopher David Chalmers differentiated between the "easy" problems of consciousness, which can be studied using the methods of cognitive science, and "the hard problem," which seems entirely resistant to scientific investigation. In the first category he listed phenomena such as the ability to discriminate, categorise and react to environmental stimuli, report mental states, focus attention, and deliberately control behaviour. The hard problem, on the other hand, is the problem of subjective experience: why does it feel like anything at all to be conscious? For many, this seems like a problem that will forever remain beyond the reach of the scientific method.

Chalmers' hard problem maps quite nicely to intuitions about the irreducibility of experience as defined above, but similarly hard problems are associated with intuitions about the irreducibility of self (how can a subjective point of view emerge from physical matter?) and the irreducibility of will (how can conscious will emerge from non-willing matter?).

If the hard problem of consciousness is the problem of explaining consciousness in physical terms, the *meta-problem of consciousness*, introduced by Chalmers in a 2018 paper, is the problem of explaining why we think there is a problem of consciousness in the first place. This problem resembles the alien's challenge: why do humans behave as if reducing consciousness to physical matter is so hard? Fortunately, the meta-problem of consciousness is, according to Chalmers' classification, an easy one: it can be studied using the tools of cognitive science. And indeed, cognitive science offers three major theories to explain the psychological origins of dualistic intuitions.

According to cultural anthropologist Ernest Becker, a belief in the immaterial nature of the soul evolved as a mechanism to cope with the fear of death. A belief that the mind can persist after the decay of the body would be very appealing, so goes the argument, to a species that developed the unpleasant awareness of their own mortality. In line with Becker's idea, one of the oldest written expressions of mind-body dualism, in Plato's Phaedo, is set as a conversation between Socrates and his friends hours before Socrates' execution. "Be of good cheer, then, my dear Crito," says Socrates, "and say that you are burying my body only."

According to a second theory, put forward by psychologist Paul Bloom and further developed by cognitive psychologist Iris Berent, intuitive dualism does not serve any function in itself. Instead, it originates from a conflict between two schematic models that the brain uses to predict what is likely to happen next. Intuitive physics, on the one hand, describes the inanimate world in terms of physical quantities like mass and size, predicting things like the trajectory of a ball as it rolls down a hill. Intuitive psychology, on the other hand, describes the behaviour of agents in terms of beliefs and desires, predicting things like the path a shopper may take between the shelves at a shop. It is the incompatibility of these two theories, argue Bloom and Berent, that results in an intuitive impression that agents have a dual nature as physical bodies and immaterial souls.

Neuroscientist Michael Graziano also identifies the origins of intuitive dualism in the brain's schematic models, but the model in Graziano's theory is used to predict not the external world, but the brain's own functioning. Cognitive neuroscience tells us that our brains maintain a body schema: a simplified representation of the body that keeps track of its position in space and facilitates motor control. The body schema does not need to represent every single fact about the body: the representation only needs to be accurate enough to be functionally useful. According to Graziano, human brains maintain a simplified schema not only of the body, but also of attention. This "attention schema" represents attention as an immaterial force that emanates from the eyes and has a causal effect on agents' actions: effectively an immaterial consciousness or soul.

The three theories provide candidate solutions to some aspects of the meta-problem – why people *think* consciousness is irreducible to physical states. They all make a common assumption: people express dualistic beliefs because it is adaptive for them to hold such beliefs (because it helps them to cope with a tantalizing fear of death, because a separation between intuitive physics and intuitive psychology is mostly useful, or because a simplified model of attention provides

better attention control).

Here, I propose a fourth candidate theory that accounts for intuitions about the irreducibility of experience, self, and will by making the opposite assumption: people hold dualistic beliefs because it is adaptive for them to express such beliefs. The proposal rests on two ideas: consciousness serves an intuitive basis for ethics and morality, and beliefs hold a signalling value in addition to their operative value. This theory also captures a core feature of intuitions about consciousness that is not explained by existing accounts, namely that consciousness is seen as morally significant.

Consciousness matters

In 2018, in an article titled "The Consciousness Deniers", philosopher Galen Strawson argued against Daniel Dennett's account of consciousness, according to which the "what-it-is-like-ness" of consciousness is an illusion. In a particularly charged section, Strawson criticises Dennett's position not for its logical structure or its empirical weaknesses, but for its potentially dangerous ethical implications:

> "If [Dennett] is right, no one has ever really suffered, in spite of agonizing diseases, mental illness, murder, rape, famine, slavery, bereavement, torture, and genocide. And no one has ever caused anyone else pain. This is the Great Silliness. We must hope that it doesn't spread outside the academy, or convince some future information technologist or roboticist who has great power over our lives."

Strawson's concern is not unfounded. Morality and ethics are deeply intertwined with intuitions about the immaterial nature of the mind. So much so that in both Latin and French, a single word is used to denote both moral sensibility and consciousness (*conscience* in French, *conscientia* in Latin), and English did not have a clear separation between the two concepts until the seventeenth century, with the rise of philosophical writings about the mind and body.

As I will show, each of the three intuitions that I named — the irreducibility of experience, the irreducibility of self, and the irreducibility of will — plays an important role in moral intuitions and practices, effectively scaffolding our conceptions of moral status, justice, and moral responsibility, respectively.

First, many moral instincts are grounded in intuitions about the inherent goodness or badness of certain experiences. Pain is awful not because of any of the physical or functional properties of pain processing in the brain, but because it feels awful to its carrier. Similarly, for many, whether or not it feels like anything to be a shrimp is the most important determinant in deciding whether we should take the interests of shrimps into account when making policy decisions. This grounding of moral status in experience applies not only to debates about the treatment of non-human animals, but also to our thinking about non-communicating patients, artificial intelligences, fetuses, and organoids. To a good first approximation, we care about others to the extent that we believe they are capable of experience.

This grounding of moral status in experience depends, intuitively at least, on the perception that experience is more than a physical property of a system. Descartes famously believed that while human consciousness depends on an immaterial soul, animals were just automatons, controlled by the same laws that govern inanimate matter. A cat may run away when threatened with a stick, or even cry when beaten, but this is not because it actually feels like anything to be a cat – its brain is simply wired to behave that way. Crucially, Descartes saw this belief as morally significant: in a letter to the philosopher Henry More, he argued that his view was "not so much cruel to beasts but respectful to human beings… whom it absolves from any suspicion of crime whenever they kill or eat animals".

Today, many share a similar intuition about the moral status of artificial intelligences. What may appear to a user as true states of sadness, hopefulness, or empathy can in fact be reduced to a series of soulless matrix multiplication operations. As a result, artificial intelligence is currently perceived by most as having only a trivial moral status. As in the case of Descartes' beast-machines, reducing behavioral manifestations of experience to physical mechanisms trivializes their ethical significance. It seems that an intuitive belief in the irreducibility of experience of some individuals but not others (for example, humans but not AI) may serve a role in establishing social norms around moral status.

Second, our sense of justice largely depends on our ability to make decisions as if from outside our particular point of view. Sure, a blanket carbon tax sounds like a great idea, but would I feel the same if I lived in a remote village, accessible only by car? In John Rawls' conception of justice, principles should be decided by society members as if from an *original position* behind a *veil of ignorance*. In this thought experiment, agents are encouraged to envision how they would make a decision if they did not know their own gender, ethnicity, social status, or talents.

Stripped of their physical attributes and social identity, the original position is inhabited by pure subjective points of view, floating entities that are not attached to a specific body. This may sound like a room full of immaterial spirits, similar to the hypothetical "floating man" used by Ibn Sina to prove the immateriality of the soul. Perhaps Rawls' ideas draw on a fundamental human intuition that our subjectivity is irreducible to the body, and therefore could have been attached to someone else — that "this could be me" is not just a figure of speech, but a fundamental property of the universe.

As an example, consider the fact that there is no sense in which I, Matan Mazor, a 37-year-old man, could have been a chicken in an industrial farm. And yet my intuitions around the immorality of battery cages largely depend on my ability to abstract away from my bodily features and imagine a world in which I had been a chicken who lives its whole life on a wire floor without the ability to spread its wings. Similar to how intuitions about the irreducibility of experience stand at the basis of social norms around moral status, intuitions about the irreducibility of self may stand at the basis of social norms around fairness and justice. In some cultures, this reasoning is made explicit in the grounding of some aspects of ethics in karma and reincarnation ("if you behave badly to your goat you may be born a goat in your next life"), but a similar type of reasoning may drive fairness judgments even if the mechanism by which a self is assigned a body is not specified by faith or culture.

A third way in which ethics and subjective experience are intertwined is via moral responsibility. In the UK, US, and Canadian legal systems, a defendant can be acquitted if they prove that their crime was committed in a state of automatism, that is, while lacking conscious control over their actions. A famous case in point is that of the British nurse William Quick, who in 1973 was tried for assaulting a patient in a mental health hospital. Quick's line of defence, eventually accepted by the court, was that he was not responsible for his actions, which were committed in a state of insulin-induced hypoglycaemia. In other words, Quick argued that his behaviour was involuntary, not reflective of his true will, and that he would have acted otherwise had his actions been fully under his control.

The acquitting of William Quick conceptually resembles the exoneration of a driver of a self-driving car who loses control of the steering wheel due to a bug in the system. Crucially, in a purely physicalist conception of the mind, all actions are caused by some brain activity, and brain activity equally corresponds to the driver and the car. A belief in the irreducibility of will rescues the notion of moral

responsibility by re-introducing the driver/car distinction: we are responsible for our actions only when they are generated by our inner, subjective selves (in the driver's seat), and not when they are caused by the faulty machinery we were handed by nature.

As with moral status and justice, limiting conditions on moral responsibility can also be defined without any appeal to the irreducibility of will, but having an imaginary ghost in the machine certainly helps to make these judgements intuitive and clear.

Together, it seems that ethics and morality are bundled with immaterial aspects of the mind in three distinct ways. First, it is easier to see experience as ethically significant if it is seen as fundamental, distinct from other material properties that do not carry this significance. Second, our sense of justice depends on our ability to imagine a state of affairs from an objective point of view, one that is not attached to our physical or psychological traits. Finally, our conception of moral responsibility depends on the perception of people as autonomous agents. These three functions map nicely onto people's peculiar set of beliefs about the mind and body: the irreducibility of experience, the irreducibility of self, and the irreducibility of will.

It may therefore be that intuitions about the immateriality of the mind serve a function: they form the basis for our intuitions about moral status, justice, and responsibility, thereby scaffolding moral cognition more broadly. Note that my argument is not that moral status *should* depend on the irreducibility of experience, that our sense of justice *should* depend on the irreducibility of self, or that moral responsibility *should* depend on the irreducibility of will. My point is that, regardless of what *should* be the case, in practice humans reason about morality by standing on the ladder of their intuitions about consciousness.

Dualism is sexy

Showing that consciousness serves an intuitive basis for ethics already goes some way toward providing an answer to the alien's challenge. Maybe the peculiar intuitions about the irreducibility of experience, selfhood, and free will somehow allow societies to conform to shared moral principles. But even if we assume that a belief in the immaterial nature of consciousness promotes moral behaviour, we still need to explain the mechanisms by which these supernatural beliefs became so widespread across cultures.

First and most straightforwardly, maybe a belief in the immaterial nature of consciousness is innately encoded in the human brain. According to this genetic explanation, these beliefs made Homo Sapiens more successful in the survival game: perhaps they made it possible for early humans to know they share the same moral intuitions with their peers, and maybe this fact made it easier for them to cooperate and work jointly toward shared goals. In this story, an innate belief in the irreducibility of experience allowed early humans to draw a line around a shared moral circle, an innate belief in the irreducibility of self allowed them to think abstractly about justice as if behind a veil of ignorance, and an innate belief in the irreducibility of will allowed them to assign moral responsibility to themselves and to each other. These intuitions became part of our innate set of beliefs because they made us more pro-social, and pro-sociality gave our species an evolutionary edge.

The difficulty with this account, as with all evolutionary accounts of group selection, is that it focuses on selection between groups and ignores selection pressures that operate within them. Imagine, for example, that a mutant physicalist is born in a tribe of dualists. To the extent that dualism scaffolds moral cognition (which we will assume for now), this physicalist mutant would be less moral than other tribe members, behaving selfishly and exploiting them for his or her needs. This would give the mutant an advantage in the survival game, increasing the representation of the physicalist gene from one generation to the next, gradually bringing the dualist gene to extinction.

So, even if dualism is good for the group, it is better for single individuals to be physicalists, rendering dualism an unstable trait. But what if the innate belief is not that the mind is immaterial, but that those who hold a belief in the immaterial mind tend to be more pro-social than those who see themselves, and others, as mere physical bodies? According to this second account, evolution biased us to think that people are more likely to care about the suffering of others if they believe experience is more than a physical state of the body, that people make more fair decisions if they can imagine being someone else, and that they more easily take responsibility for their actions if they believe in free will. To the extent that these social beliefs about dualists are accurate, they conferred an evolutionary advantage on the individuals who held them, improving their ability to choose better social and romantic partners. And, critically, unlike the innate dualism account above, occasional mutants with a social preference for physicalists had no advantage over their peers.

If such an innate set of beliefs about dualists evolved, it would have two effects. First, it would make us all prefer to surround ourselves with individuals who express dualistic beliefs, as they are more likely to care about us, make fair decisions, and take responsibility for their actions. Second, knowing that others also share this preference for dualist individuals, we may be motivated to signal dualistic beliefs if we want to be seen as caring, fair, and liable. This social pressure may drive individuals to behave as if they hold dualistic beliefs in order to enjoy the reputational advantages of intuitive dualism, or, equivalently, to avoid the social sanctions associated with being perceived as a cold, selfish and shameless physicalist.

But people do more than expressing beliefs in the immaterial nature of experience, selfhood, and will: these beliefs are also held privately, and affect deeply held intuitions. To return to the beginning of the essay, I share these intuitions too. In fact, referring to them as intuitions feels like downplaying the level of conviction with which I hold them: nothing feels more self-evident than the particular, qualitative nature of my experience, my first-person point of view, and my control over my thoughts and bodily actions.

This is a good point to remind ourselves that, for the purpose of this essay, we are taking the alien's perspective, from which it is not the irreducibility of consciousness itself that needs to be explained (indeed, the alien does not have any reason to doubt that consciousness — whatever that is — is perfectly reducible to brain states), but the fact that many people, including the author of this essay, have a strong gut feeling that experience, selfhood, and will cannot be reduced to material properties of the world. People may be motivated to publicly signal their dualism to enjoy the reputational advantages of appearing dualistic, but why do they also behave as if they hold these beliefs in private?

One possible explanation is that, with time, intuitive dualism did become genetically determined. Once individuals in a group prefer dualists as their romantic partners, being a dualist is adaptive not only for the group, but also for single individuals: it simply makes them more sexually attractive. This way, evolution encoded dualism in our brains in two steps: first we became biased against expressions of physicalism, preferring to mate with dualist individuals. This created evolutionary pressure to hold dualistic beliefs, driving the hard coding of these beliefs into our brains. This two-step mechanism ensured that physicalist mutants found it more difficult to spread their genes, protecting dualism from the risk of within-group extinction.

But this explanation requires that intricate, abstract knowledge can become innately specified within a very short time on evolutionary time scales. This would be unusual: the closest example would be innate knowledge about psychology (what psychologists sometimes call a "*Theory of Mind*": the understanding that behaviour of other agents is a function of their beliefs and desires), but this innate knowledge is likely shared with other species like crows and apes, making it likely that it had a much longer time to develop in evolution.

An alternative explanation, which I developed together with Lucius Caviola, does not assume such genetic hard-coding. In this story, none of us is born a dualist. Instead, our theory of intuitive dualism relies on the principle that the best way to convince others that you hold a belief is to first convince yourself that you hold it. If we want to behave as if we believe that the mind is more than the brain, it may be easiest to convince ourselves that this is indeed our belief. This is sometimes referred to by social and evolutionary psychologists as *strategic self-deception*.

This is a pretty radical proposal, so let me repeat it. According to this account, our innermost private intuitions about the immateriality of consciousness are lies we tell ourselves in order to appear to others as if we genuinely hold these intuitions, so that we are perceived as warmer, more pro-social, and more trustworthy by our peers and potential sexual partners.

Strategic self-deception

Tricking ourselves into believing that we have dualistic intuitions may seem strange, if not self-contradictory. How can a person deceive themself? As the one doing the deception, wouldn't they immediately know they are being deceived? Indeed, self-deception only makes sense when we conceive of the mind as a collection of modules, each specialised to perform a particular task. In this modular view of the mind, there is no single '*I*'. Instead, multiple modules work in parallel, processing sensory information, making plans, controlling bodily movements, and monitoring internal states. Modules may share information between them, but they don't have to, and they often don't.

It is tempting to think of the little voice in our head as the chief commander of this operation, but in this modular view of the mind what we intuitively identify as our inner self, and Descartes identified as his immaterial res cogitans, plays the much less glamorous role of narrating our actions after the fact, similar to the job

of a press secretary of an organisation. This "press secretary module" is usually not part of the decision-making process itself, and yet it has the crucial job of explaining these decisions to the outside world in a way that makes us appear best. When things go right, the press secretary module doesn't need to lie or conceal the truth: the other modules keep it selectively ignorant of the hidden motives behind our actions. As the White House press secretary in the TV series "*The West Wing*" puts it: "I do my best work when I'm the least informed person in the room".

In their book "*The Elephant in the Brain: Hidden Motives in Everyday Life*", Kevin Simler and Robin Hanson present some striking examples of such strategic self-deception. One of my favourite ones is that people believe they laugh when something is funny, but human behaviour aligns much more closely with laughter serving a social signalling function. We use laughter to signal our adherence to social hierarchy (think about the boss or the lecturer whose dumb jokes everyone laughs at) and our commitment to social alliances (think about the relief that comes with laughing together after a tense conversation).

Critically, being aware of our motivations for laughing would make us appear fake and disingenuous, undermining the quality of the signal. According to Simler and Hanson, laughter works as a social signal exactly because the press secretary is kept strategically ignorant of the Machiavellian motivations behind it, and because it is so hard to fake deliberately. Other examples include our ignorance of the true motivations behind our charitable giving, political ideology, and religious faith.

The cultural story is therefore that, like many other aspects of our mental lives, the press secretary module is kept strategically ignorant of the actual reasons for our public expressions of intuitive dualism. While we believe that we express these views because they seem self-evident, in truth we express them because modules that are responsible for navigating the social world understand that signalling dualism would make us appear warmer, more pro-social, and more trustworthy to others. In short, intuitive dualism emerges from the dual action of two forces: a strong social pressure to publicly signal intuitive dualism due to its role as a basis for ethics, and the fact that we sometimes hold beliefs in order to convincingly signal that we hold them.

Holding these beliefs may in turn make us more caring, pro-social, and morally responsible [1]. If this is the case, we may be looking at a virtuous cycle: individuals

[1]Some research has found correlations between a belief in dualism or free will and (self-reported) helping behaviour. Crucially, however, these studies rely exclusively on self-report, for example asking

hold dualistic beliefs because expressing such beliefs makes them appear more pro-social, those beliefs then promote genuinely pro-social behaviour, which reinforces the reliability of dualism as a signal, further motivating people to adopt and display dualistic beliefs.

Taking stock

If this story has some merit, even if only as a part of a larger one, systematic misconceptions about the nature of our mental lives may scaffold human moral cognition. This has important implications for how we think about consciousness, ethics, and the relationship between the two. Let me name two.

First, debates about the consciousness of non-human entities are turned on their head when viewed through this lens. Take, for example, current debates about the level of consciousness in artificial intelligence. From the point of view of an alien visiting Earth, the debate may seem entirely pointless: humans — earthly entities who have somehow convinced themselves that they are more than their physical bodies — are now debating whether certain computer programmes produce something that is similar to their own immaterial, spirit-like essence. But the debate starts to make sense once we understand that the function of a belief in the immaterial nature of experience was all along to demarcate a line between those who are capable of having it and those who are not.

Crucially, it is not that we demarcate the line based on our beliefs about consciousness (even if this is what our press secretary module tells us we do). Instead, we attribute consciousness to those beings who we want to include within the moral circle. Attributing consciousness to humans but not to ducks allowed Descartes to eat his Canard à l'orange without sensing guilt in seventeenth century France. Similarly, attributing consciousness to living beings but not to artificial intelligence is comfortable for us today.

If the proposal that I have put forward here is correct, then instead of debating the level of consciousness of non-human entities we should directly ask whether we want to treat them as moral patients based on observable properties, and without reference to consciousness or experience. This would be in line with positions

participants to rate the likelihood that they would help a homeless person on a scale from 1 (not at all likely) to 9 (very likely). It may well be that the observed association is driven not by actual altruism, but by individual differences in the desire to signal altruism (here, to the experimenters).

advocated by Marian Dawkins and Peter Carruthers on animal welfare, and by John Danaher and David Papineau on AI ethics.

More broadly, this position has implications for the way we think about the scientific study of consciousness. If systematically distorted intuitions about the mind are essential for sustaining a shared moral code, then consciousness science becomes a dangerous enterprise: as soon as scientists uncover the physical underpinnings of conscious experience, the moral landscape may dramatically deform, if not collapse entirely (as suggested in Galen Strawson's critique of Daniel Dennett that we saw before).

Fortunately, consciousness scientists are themselves subject to the same hidden motives to keep consciousness mysterious and intractable. A good case in point is nausea: once considered a hallmark of the ineffable, subjective nature of human experience. The discovery that nausea is caused by specific cells in the area postrema of the medulla did not dissolve the mystery of its phenomenal character, but only redefined the mystery as surrounding that which is left unexplained.

In this more benign vision of consciousness science, it is not morally explosive, but futile. Once we understand consciousness, by any standard, consciousness, or at least those aspects of consciousness that are deemed morally relevant, will be redefined to refer to whatever remains unexplained.

The moral scaffolding theory I present here is not mutually exclusive with Becker's death-denial account, Bloom's intuitive psychology versus physics account, or Graziano's attention schema account; it is highly likely that all are needed to explain people's intuitions about the immaterial nature of consciousness. But whatever the relative contribution of each component is, two original aspects of the moral scaffolding theory will inevitably play a significant role in any complete account of intuitive dualism: moral cognition and social signalling. I have tried to present the account in its purest, most radical form, because that is a useful way to start a discussion, identify weaknesses, and detect testable hypotheses. Some of these unique hypotheses are currently being tested by Lucius Caviola and myself.

I hope that, at least from the alien's point of view, this was a satisfactory account. To the extent that I was successful, this essay offers an answer to the problem of consciousness as it is seen from the outside, or to Chalmers' meta-problem of

consciousness. Some readers may even feel this goes some way toward addressing the hard problem of consciousness itself, not by solving it, but by demystifying it. Maybe the problem of explaining consciousness in physical terms seems so hard because we need it to be hard — for ethics to work, and, more egoistically, for us to appear pro-social to others.

There are moments in which I feel satisfied by this account. Just as I have learned to mistrust my intuitions about the motivations for my actions and the origins of my beliefs in other domains of life, I can learn to be skeptical of my intuitions about my own subjectivity. But these moments are brief and far between. For most of the time, I have the intuition — no, the belief — no, the knowledge, that there is nothing about the physical state of the universe that explains why this physical body is conscious, or, equivalently, why the consciousness that is me is attached to this particular physical body. Any attempt to explain the basic fact of my subjectivity in terms of biological or social mechanisms seems like an amusing academic mind-play, a curious thought experiment without any real bearing on the mystery. Why is there an "*I*" at all, and why is this "*I*" *me*?

Reading the above paragraph, the alien may think the author only included it to signal his pro-sociality and warmth. And the alien may be right. I wish I knew.

Further reading For readers interested in the mysteries of first-person perspective and personhood, I recommend *The View from Nowhere* by Thomas Nagel and *Reasons and Persons* by Derek Parfit. For an accessible discussion of modularity and strategic self-deception, see *The Elephant in the Brain* by Kevin Simler and Robin Hanson, and *Why Everyone (Else) Is a Hypocrite* by Robert Kurzban, which also popularises the idea of a "press secretary module" (and from which I got the *West Wing* quote). For more fun illustrations of modern-day expressions of intuitive dualism, including in Star Wars and Harry Potter, see Graziano, M. S., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2020). Cognitive Neuropsychology. The excerpt from Descartes' letter to Henry More is cited in Samuel Kaldas' article "*Descartes versus Cudworth on the Moral Worth of Animals*" (*Philosophy Now*). The quote from Galen Strawson appears in François Kammerer's paper "*The Normative Challenge for Illusionist Views of Consciousness*" (2020, *Ergo*). I am grateful to Damian Maher for pointing out that French and Latin use the same word for consciousness and conscience; more on this can be found in Larry M. Jorgensen's entry "*Seventeenth-Century Theories of Consciousness*" in the *Stanford Encyclopedia of Philosophy*. Celia Heyes, Roni Maimon, and Niccolo Negro provided useful feedback on an earlier version of this essay. Finally, I thank Lucius Caviola for many conversations that inspired this essay.