Conscious exotica

From algorithms to aliens, could humans ever understand minds that are radically unlike our own?

Murray Shanahan

In 1984, the philosopher Aaron Sloman invited scholars to describe ‘the space of possible minds’. Sloman’s phrase alludes to the fact that human minds, in all their variety, are not the only sorts of minds. There are, for example, the minds of other animals, such as chimpanzees, crows and octopuses. But the space of possibilities must also include the minds of life-forms that have evolved elsewhere in the Universe, minds that could be very different from any product of terrestrial biology. The map of possibilities includes such theoretical creatures even if we are alone in the Cosmos, just as it also includes life-forms that could have evolved on Earth under different conditions.

We must also consider the possibility of artificial intelligence (AI). Let's say that intelligence 'measures an agent's general ability to achieve goals in a wide range of environments', following the definition <http://hutter1.net/ai/sior.pdf> adopted by the computer scientists Shane Legg and Marcus Hutter. By this definition, no artefact exists today that has anything approaching human-level intelligence. While there are computer programs that can out-perform humans in highly demanding yet specialised intellectual domains, such as playing the game of Go, no computer or robot today can match the generality of human intelligence.

But it is artefacts possessing general intelligence – whether rat-level, human-level or beyond – that we are most interested in, because they are candidates for membership of the space of possible minds. Indeed, because the potential for variation in such artefacts far outstrips the potential for variation in naturally evolved intelligence, the non-natural variants might occupy the majority of that space. Some of these artefacts are likely to be very strange, examples of what we might call 'conscious exotica'.

In what follows I attempt to meet Sloman's challenge by describing the structure of the space of possible minds, in two dimensions: the capacity for consciousness and the human-likeness of behaviour. Implicit in this mapping seems to be the possibility of forms of consciousness so alien that we would not recognise them. Yet I am also concerned, following Ludwig Wittgenstein, to reject the dualistic idea that there is an impenetrable realm of subjective experience that forms a distinct portion of reality. I prefer the notion that 'nothing is hidden', metaphysically speaking. The difficulty here is that accepting the possibility of radically inscrutable consciousness seemingly readmits the dualistic proposition that consciousness is not, so to speak, 'open to view', but inherently private. I try to show how we might avoid that troubling outcome.

Thomas Nagel's celebrated treatment <http://organizations.utep.edu/portals/1475/nagel_bat.pdf> of the (modestly) exotic subjectivity of a bat is a good place to start. Nagel wonders what it is like to be a bat, and laments that 'if I try to imagine this, I am restricted to the resources of my own mind, and those resources are inadequate to the task'. A corollary of Nagel's position is that certain kinds of facts – namely facts that are tied to a very different subjective point of view – are inaccessible to our human minds. This supports the dualist's claim that no account of reality could be complete if it comprised only objective facts and omitted the subjective. Yet I think the dualistic urge to cleave reality in this way is to be resisted. So, if we accept Nagel's reasoning, conscious exotica present a challenge.

But bats are not the real problem, as I see it. The moderately exotic inner lives of non-

human animals present a challenge to Nagel only because he accords ontological status to an everyday indexical distinction. I cannot be both *here* and *there*. But this platitude does not entail the existence of facts that are irreducibly tied to a particular position in space. Similarly, I cannot be both a human and a bat. But this does not entail the existence of phenomenological facts that are irreducibly tied to a particular subjective point of view. We should not be fooled by the presence of the word 'know' into seeing the sentence 'as a human, I cannot know what it's like to be a bat' as expressing anything more philosophically puzzling than the sentence 'I am a human, not a bat.' We can always speculate about what it might be like to be a bat, using our imaginations to extend our own experience (as Nagel does). In doing so, we might remark on the limitations of the exercise. The mistake is to conclude, with Nagel, that there must be facts of the matter here, certain subjective 'truths', that elude our powers of collective investigation.

> To explore the space of possible minds is to entertain the possibility of beings far more exotic than any terrestrial species

In this, I take my cue from the later Wittgenstein of *The Philosophical Investigations* <https://static1.squarespace.com/static/54889e73e4b0a2c1f9891289/t/564b61a4e4b04eca59c4d232/1447780772744/Ludwig.Wittgenstein.-.Philosophical.Investigations.pdf> (1953). The principle that underlies Wittgenstein's rejection of private language – a language with words for sensations that only one person in the world could understand – is that we can talk only about what lies before us, what is public, what is open to collective view. As for anything else, well, 'a nothing would serve as well as a something about which nothing can be said'. A word that referred to a private, inner sensation would have no useful function in our language. Of course, things can be hidden in a practical sense, like a ball beneath a magician's cup, or a star that is outside our light cone. But nothing is beyond reach metaphysically speaking. When it comes to the inner lives of others, there is always more to be revealed – by interacting with them, by observing them, by studying how they work – but it makes no sense to speak as if there were something over and above what can ever be revealed.

Following this train of thought, we should not impute unknowable subjectivity to other people (however strange), to bats or to octopuses, nor indeed to extra-terrestrials or to artificial intelligences. But here is the real problem, namely radically exotic forms of consciousness. Nagel reasonably assumes that 'we all believe bats have experience'; we might not know what it is like to be a bat, yet we presume it is like *something*. But to explore the space of possible minds is to entertain the possibility of beings far more exotic than any terrestrial species. Could the space of possible minds include beings so inscrutable that we could not tell whether they had conscious experiences *at all*? To deny this possibility smacks of biocentrism. Yet to

accept it is to flirt once more with the dualistic thought that there is a hidden order of subjective facts. In contrast to the question of *what* it is like to be an *X*, surely (we are tempted to say) there *is* a fact of the matter when it comes to the question of whether it is like *anything at all* to be an *X*. Either a being has conscious experience or it does not, regardless of whether we can tell.

Consider the following thought experiment. Suppose I turn up to the lab one morning to discover that a white box has been delivered containing an immensely complex dynamical system whose workings are entirely open to view. Perhaps it is the gift of a visiting extraterrestrial, or the unwanted product of some rival AI lab that has let its evolutionary algorithms run amok and is unsure what to do with the results. Suppose I have to decide whether or not to destroy the box. How can I know whether that would be a morally acceptable action? Is there any method or procedure by means of which I could determine whether or not consciousness was, in some sense, present in the box?

One way to meet this challenge would be to devise an objective measure of consciousness, a mathematical function that, given any physical description, returns a number that quantifies the consciousness of that system. The neuroscientist Giulio Tononi has purported to supply just such a measure, named Φ, within the rubric of so-called 'integrated information theory <http://www.nature.com/nrn/journal/v17/n7/full/nrn.2016.44.html> '. Here, Φ describes the extent to which a system is, in a specific information-theoretic sense, more than the sum of its parts. For Tononi, consciousness *is* Φ in much the same sense that water *is* H_2O. So, integrated information theory claims to supply both necessary and sufficient conditions for the presence of consciousness in any given dynamical system.

The chief difficulty with this approach is that it divorces consciousness from behaviour. A completely self-contained system can have high Φ despite having no interactions with anything outside itself. Yet our everyday concept of consciousness is inherently bound up with behaviour. If you remark to me that someone was or was not aware of something (an oncoming car, say, or a friend passing in the corridor) it gives me certain expectations about their behaviour (they will or won't brake, they will or won't say hello). I might make similar remarks to you about what I was aware of in order to account for my own behaviour. 'I can hardly tell the difference between those two colours'; 'I'm trying to work out that sum in my head, but it's too hard'; 'I've just remembered what she said'; 'It doesn't hurt as much now' – all these sentences help to explain my behaviour to fellow speakers of my language and play a role in our everyday social activity. They help us keep each other informed about what we have done in the past, are doing right now or are likely to do in the future.

It's only when we do philosophy that we start to speak of consciousness, experience and sensation in terms of private subjectivity. This is the path to the hard problem/easy problem <https://aeon.co/essays/will-we-ever-get-our-heads-round-

consciousness> distinction set out by David Chalmers, <http://consc.net/papers/facing.html> to a metaphysically weighty division between inner and outer – in short, to a form of dualism in which subjective experience is an ontologically distinct feature of reality. Wittgenstein provides an antidote to this way of thinking in his remarks on private language, whose centrepiece is an argument to the effect that, insofar as we can talk about our experiences, they must have an outward, public manifestation. For Wittgenstein, 'only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations, it sees … it is conscious or unconscious'.

Through Wittgenstein, we arrive at the following precept: *only against a backdrop of purposeful behaviour do we speak of consciousness.* By these lights, in order to establish the presence of consciousness, it would not be sufficient to discover that a system, such as the white box in our thought experiment, had high Φ. We would need to discern purpose in its behaviour. For this to happen, we would have to see the system as embedded in an environment. We would need to see the environment as acting on the system, and the system as acting on the environment for its own ends. If the 'system' in question was an animal, then we already inhabit the same, familiar environment, notwithstanding that the environment affords different things to different creatures. But to discern purposeful behaviour in an unfamiliar system (or creature or being), we might need to *engineer an encounter* with it.

Even in familiar instances, this business of engineering an encounter can be tricky. For example, in 2006 the neuroscientist Adrian Owen and his colleagues managed to establish a simple form of communication <http://science.sciencemag.org/content/313/5792/1402> with vegetative-state patients using an fMRI scanner. The patients were asked to imagine two different scenarios that are known to elicit distinct fMRI signatures in healthy individuals: walking through a house and playing tennis. A subset of vegetative-state patients generated appropriate fMRI signatures in response to the relevant verbal instruction, indicating that they could understand the instruction, had formed the intention to respond to it, and were able to exercise their imagination. This must count as 'engineering an encounter' with the patient, especially when their behaviour is interpreted against the backdrop of the many years of normal activity the patient displayed when healthy.

> We don't weigh up the evidence to conclude that our friends are probably conscious creatures. We simply see them that way, and treat them accordingly

Once we have discerned purposeful behaviour in our object of study, we can begin to observe and (hopefully) to interact with it. As a result of these observations and interactions, we might decide that consciousness is present. Or, to put things differently, we might adopt the sort of *attitude* towards it that we normally reserve for

fellow conscious creatures.

The difference between these two forms of expression is worth dwelling on. Implicit in the first formulation is the assumption that there is a fact of the matter. Either consciousness is present in the object before us or it is not, and the truth can be revealed by a scientific sort of investigation that combines the empirical and the rational. The second formulation owes its wording to Wittgenstein. Musing on the skeptical thought that a friend could be a mere automaton – a phenomenological zombie, as we might say today – Wittgenstein notes that he is not of the opinion that his friend has a soul. Rather, 'my attitude towards him is an attitude towards a soul'. (For 'has a soul' we can read something like 'is conscious and capable of joy and suffering'.) The point here is that, in everyday life, we do not weigh up the evidence and conclude, on balance, that our friends and loved ones are probably conscious creatures like ourselves. The matter runs far deeper than that. We simply see them that way, and treat them accordingly. Doubt plays no part in our attitude towards them.

How do these Wittgensteinian sensibilities play out in the case of beings more exotic than humans or other animals? Now we can reformulate the white box problem of whether there is a method that can determine if consciousness, in some sense, is present in the box. Instead, we might ask: under what circumstances would we adopt towards this box, or any part of it, the sort attitude we normally reserve for a fellow conscious creature?

Let's begin with a modestly exotic hypothetical case, a humanoid robot with human-level artificial intelligence: the robot Ava from the film *Ex Machina* (2015), written and directed by Alex Garland.

In *Ex Machina*, the programmer Caleb is taken to the remote retreat of his boss, the reclusive genius and tech billionaire Nathan. He is initially told he is to be the human component in a Turing Test, with Ava as the subject. After his first meeting with Ava, Caleb remarks to Nathan that in a real Turing Test the subject should be hidden from the tester, whereas Caleb knows from the outset that Ava is a robot. Nathan retorts that: 'The real test is to show you she is a robot. Then see if you still feel she has consciousness.' (We might call this the 'Garland Test'.) As the film progresses and Caleb has more opportunities to observe and interact with Ava, he ceases to see her as a 'mere machine'. He begins to sympathise with her plight, imprisoned by Nathan and faced with the possibility of 'termination' if she fails his test. It's clear by the end of the film that Caleb's attitude towards Ava has evolved into the sort we normally reserve for a fellow conscious creature.

The arc of Ava and Caleb's story illustrates the Wittgenstein-inspired approach to consciousness. Caleb arrives at this attitude not by carrying out a scientific investigation of the internal workings of Ava's brain but by watching her and talking to her. His stance goes deeper than any mere opinion. In the end, he acts decisively on

her behalf and at great risk to himself. I do not wish to imply that scientific investigation should not influence the way we come to see another being, especially in more exotic cases. The point is that the study of a mechanism can only complement observation and interaction, not substitute for it. How else could we truly come to see another conscious being as such, other than by inhabiting its world and encountering it for ourselves?

> If something is built very differently to us, then however human-like its behaviour, its consciousness might be very different to ours

The situation is seemingly made simpler for Caleb because Ava is only a moderately exotic case. Her behaviour is very human-like, and she has a humanoid form (indeed, a female humanoid form that he finds attractive). But the fictional Ava also illustrates how tricky even seemingly straightforward cases can be. In the published script, there is a direction for the last scene of the film that didn't make the final cut. It reads: 'Facial recognition vectors flutter around the pilot's face. And when he opens his mouth to speak, we don't hear words. We hear pulses of monotone noise. Low pitch. Speech as pure pattern recognition. This is how Ava sees us. And hears us. It feels completely alien.' This direction brings out the ambiguity that lies at the heart of the film. Our inclination, as viewers, is to see Ava as a conscious creature capable of suffering – as Caleb sees her. Yet it is tempting to wonder whether Caleb is being fooled, whether Ava might not be conscious after all, or at least not in any familiar sense.

This is a seductive line of thinking. But it should be entertained with extreme caution. It is a truism in computer science that specifying how a system behaves does not determine how that behaviour need be implemented in practice. In reality, human-level artificial intelligence exhibiting human-like behaviour might be instantiated in a number of different ways. It might not be necessary to copy the architecture of the biological brain. On the other hand, perhaps consciousness *does* depend on implementation. If a creature's brain is like ours, then there are grounds to suppose that its consciousness, its inner life, is also like ours. Or so the thought goes. But if something is built very differently to us, with a different architecture realised in a different substrate, then however human-like its behaviour, its consciousness might be very different to ours. Perhaps it would be a phenomenological zombie, with no consciousness at all.

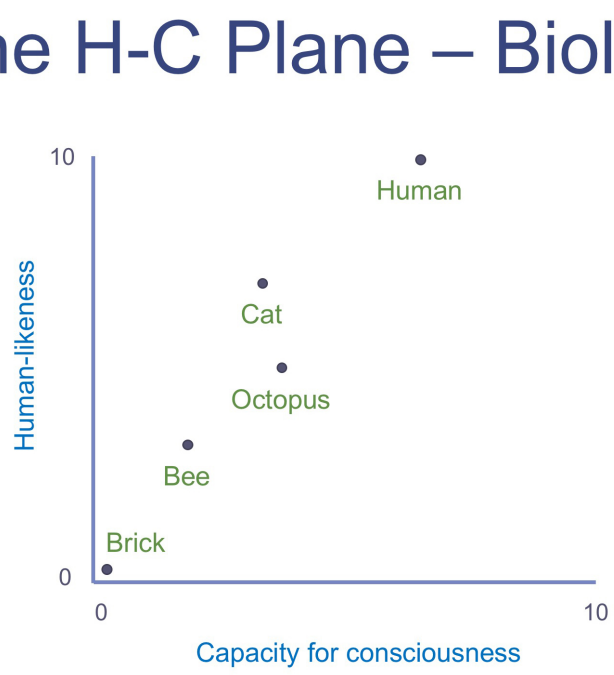
The trouble with this thought is the pull it exerts towards the sort of dualistic metaphysical picture we are trying to dispense with. Surely, we cry, there must be a fact of the matter here? Either the AI in question is conscious in the sense you and I are conscious, or it is not. Yet seemingly we can never know for sure which it is. It is a small step from here to the dualistic intuition that a private and subjective world of inner experience exists separately from the public and objective world of physical

objects. But there is no need to yield to this dualistic intuition. Neither is there any need to deny it. It is enough to note that, in difficult cases, it is always possible to find out more about an object of study – to observe its behaviour under a wider set of circumstances, to interact with it in new ways, to investigate its workings more thoroughly. As we find out more, the way we treat it and talk about it will change, and in this way we will converge on the appropriate attitude to take towards it. Perhaps Caleb's attitude to Ava would have changed if he'd had more time to interact with her, to find out what really made her tick. Or perhaps not.

So far, we have stuck to human-like entities and haven't looked at anything especially exotic. But we need to extend our field of vision if we are to map out the space of possible minds. This affords the opportunity to think imaginatively about properly exotic beings, and to speculate about their putative consciousness.

There are various dimensions along which we might plot the many kinds of minds we can imagine. I have chosen two: human-likeness (the H-axis) and capacity for consciousness (the C-axis). An entity is *human-like* to the extent that it makes sense to describe its behaviour using the language we normally employ to describe humans – the language of beliefs, desires, emotions, needs, skills and so on. A brick, by this definition, scores very low. For very different reasons, an exotic entity might also score very low on human-likeness, if its behaviour were inscrutably complex or alien. On the C-axis, an entity's *capacity for consciousness* corresponds to the richness of experience it is capable of. A brick scores zero on this axis (panpsychism <https://aeon.co/ideas/why-panpsychism-fails-to-solve-the-mystery-of-consciousness> notwithstanding), while a human scores significantly more than a brick.

Figure 1 below tentatively places a number of animals in the H-C plane, along axes that range from 0 (minimum) to 10 (maximum). A brick is shown at the (0, 0) position. Let's consider the placement along the C-axis. There is no reason to suppose a human's capacity for consciousness could not be exceeded by some other being. So humans (perhaps generously) are assigned 8 on this axis. The topic of animal consciousness is fraught with difficulty. But a commonplace assumption is that, in terrestrial biology at least, consciousness is closely related to cognitive prowess. In line with this intuition, a bee is assumed to have a smaller capacity for consciousness than a cat, which in turn has a slightly smaller capacity for consciousness than an octopus, while all three of those animals score less than a human being. Arranging animals this way can be justified by appealing to the range of capabilities studied in the field of animal cognition. These include associative learning, physical cognition, social cognition, tool use and manufacture, mental time travel (including future planning and episodic-like memory) and communication. An animal's experience of the world is presumed to be enriched by each of these capabilities. For humans, we can add language, the capacity to form abstract concepts and the ability to think associatively in images and metaphors, among others.



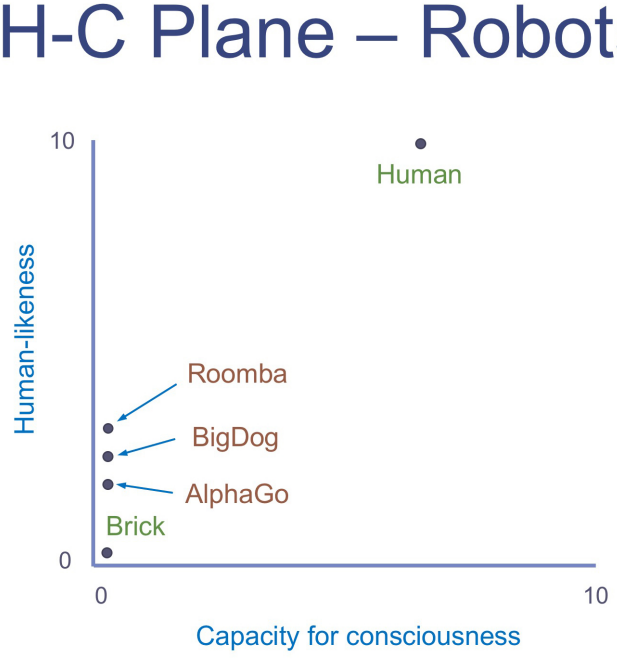


Figure 1. Top: biology on the H-C Plane. Below: contemporary AI on the H-C Plane

Now let's turn our attention to the H-axis. Tautologically, a human being has maximum human-likeness. So we get 10 on the H-axis. All non-human animals share certain fundamentals with humans. All animals are embodied, move and sense the world, and exhibit purposeful behaviour. Moreover, every animal has certain bodily needs in common with humans, such as food and water, and every animal tries to protect itself from harm and to survive. To this extent, all animals exhibit human-like

behaviour, so all animals get 3 or more on the H-axis. Now, in order to describe and explain the behaviour of a non-human animal, we have recourse to the concepts and language we use to describe and explain human behaviour. An animal's behaviour is said to be human-like to the extent that these conceptual and linguistic resources are necessary and sufficient to describe and explain it. And the more cognitively sophisticated a species is, the more of these linguistic and conceptual resources are typically required. So the cat and the octopus are higher up the H-axis than the bee, but lower than the human.

> It is, of course, naïve to assign a simple scalar to a being's capacity for consciousness

Under the assumptions we're making, human-likeness and the capacity for consciousness are broadly correlated for animals. However, the octopus appears lower down the H-axis than the cat, despite being further along on the C-axis. I don't want to defend these relative orderings specifically. But the octopus exemplifies the possibility of a creature that is cognitively sophisticated, that we are inclined to credit with a capacity for rich conscious experiences, but whose behaviour is hard for humans to understand. Taking this idea further, we can imagine conscious beings far more inscrutable than an octopus. Such beings would appear down there with the brick on the H-axis, but for very different reasons. To describe and explain the behaviour of a brick, the elaborate concepts we use to describe and explain human behaviour are unnecessary, since it exhibits none to speak of. But to describe and explain the behaviour of a cognitively sophisticated but inscrutable being, those resources would be insufficient.

There is plenty to take issue with in these designations. It is, of course, naïve to assign a simple scalar to a being's capacity for consciousness. A more nuanced approach would be sensitive to the fact that different combinations of cognitive capabilities are present in different animals. Moreover, the extent to which each of these capabilities contributes to the richness of a creature's experience is open to debate. Similar doubts can be cast on the validity of the H-axis. But the H-C plane should be viewed as a canvas on which crude, experimental sketches of the space of possible minds can be made, a spur to discussion rather than a rigorous theoretical framework. Furthermore, diagrams of the H-C plane are not attempts to portray facts of the matter with respect to the consciousness of different beings. Rather, they are speculative attempts to anticipate the consensual attitude we might arrive at about the consciousness of various entities, following a collective process of observation, interaction, debate, discussion and investigation of their inner workings.

Let's put some contemporary examples of robotics and artificial intelligence on the H-C plane. These include Roomba (a domestic vacuum-cleaning robot), BigDog (a four-legged robot with life-like locomotion), and AlphaGo (the program created by Google DeepMind that defeated the champion Go player Lee Sedol in

2016). All three are pressed up to the far left of the C-axis. Indeed, no machine, no robot or computer program yet exists that could plausibly be ascribed any capacity for consciousness at all.

On the other hand, as far as human-likeness is concerned, all three are well above the brick. BigDog appears slightly below Roomba, both of which are slightly above AlphaGo. BigDog is guided by a human operator. However, it is capable of automatically adjusting to rough or slippery terrain, and of righting itself when its balance is upset, by being kicked, for example. In describing these aspects of its behaviour, it's natural to use phrases such as 'it's trying not to fall over' or even 'it really wants to stay upright'. That is to say, we tend to adopt towards BigDog what Daniel Dennett calls the 'intentional stance' <https://ase.tufts.edu/cogstud/dennett/papers/intentionalsystems.pdf> , imputing beliefs, desires and intentions because this makes it easier to describe and explain its behaviour.

Unlike BigDog, Roomba is a fully autonomous robot that can operate for long periods without human intervention. Despite BigDog's lifelike response to being kicked, the slightest understanding of its inner workings should dispel any inclination to see it as a living creature struggling against adversity. The same is true of Roomba. However, the behaviour of Roomba is altogether more complex, because it has an overarching mission, namely to keep the floor clean. Against the backdrop of such a mission, the intentional stance can be used in a far more sophisticated way, invoking an interplay of perception, action, belief, desire and intention. Not only are we inclined to say things such as: 'It's swerving to avoid the chair leg', we might also say: 'It's returning to the docking station because its batteries are low', or 'It's going over that patch of carpet again because it can tell that it's really dirty.'

AlphaGo scores the lowest of the three artefacts we're looking at, though not due to any lack in cognitive capabilities. Indeed, these are rather impressive, albeit in a very narrow domain. Rather, it is because AlphaGo's behaviour can barely be likened to a human's or an animal's at all. Unlike BigDog and Roomba, it doesn't inhabit the physical world or have a virtual surrogate in any relevant sense. It doesn't perceive the world or move within it, and the totality of its behaviour is manifest through the moves it makes on the Go board. Nevertheless, the intentional stance is sometimes useful to describe its behaviour. Demis Hassabis, DeepMind's co-founder, issued three telling tweets concerning the one game that AlphaGo lost to Sedol in the five-game series. In the first tweet, he wrote: '#AlphaGo thought it was doing well, but got confused on move 87.' He went on to say: 'Mistake was on move 79, but #AlphaGo only came to that realisation on around move 87.' Shortly afterwards he tweeted: 'When I say "thought" and "realisation" I just mean the output of #AlphaGo's value net. It was around 70 per cent at move 79 and then dived on move 87.'

'It's not a human move. I've never seen a human play this

move. So beautiful’

To anyone unfamiliar with AlphaGo’s inner workings, the first two tweets would have made far more sense than the scientifically more accurate statement in the third. However, this is a shallow use of the intentional stance, which is ultimately of little help in understanding AlphaGo. It does not interact with a world of spatiotemporally located objects, and there is no fruitful sense in which its behaviour can be characterised in terms of the interplay of perception, belief, desire, intention and action.

On the other hand, it deploys a formidable set of cognitive skills within the microworld of Go. It learns through experience, garnered through self-play as well as from records of human games. It can search through a myriad possible plays to determine its next move. Its ability to respond effectively to subtle board patterns replicates what is often called intuition in top human players. And in one extraordinary move during the Sedol match, it displayed a form of what we might call creativity. It ventured into the fifth line of the Go board using a move known as a shoulder hit, in which a stone is placed diagonally adjacent to an opponent’s stone. Commentating on the match, the European Go champion Fan Hui remarked: ‘It’s not a human move. I’ve never seen a human play this move. So beautiful.’ According to AlphaGo’s own estimate, there was a one-in-10,000 chance that a human would have used the same tactic, and it went against centuries of received wisdom. Yet this move was pivotal in giving it victory.

What we find in AlphaGo is an example of what we might term, not ‘conscious exotica’, but rather a form of ‘cognitive exotica’. Through a process largely opaque to humans, it manages to attain a goal that might have been considered beyond its abilities. AlphaGo’s prowess is confined to Go, and we are a long way from artificial general intelligence. However, it’s natural to wonder about the possible forms that artificial general intelligence might take – and how they could be distributed within the space of possible minds.

So far we have looked at the human-likeness and capacity for consciousness of various real entities, both natural and artificial. But in Figure 2 below, a number of hypothetical beings are placed on the H-C plane. Obviously, this is wildly speculative. Only through an actual encounter with an unfamiliar creature could we truly discover our attitude towards it and how our language would adapt and extend to accommodate it. Nevertheless, guided by reason, the imagination can tell us something about the different sorts of entity that might populate the space of possible minds.

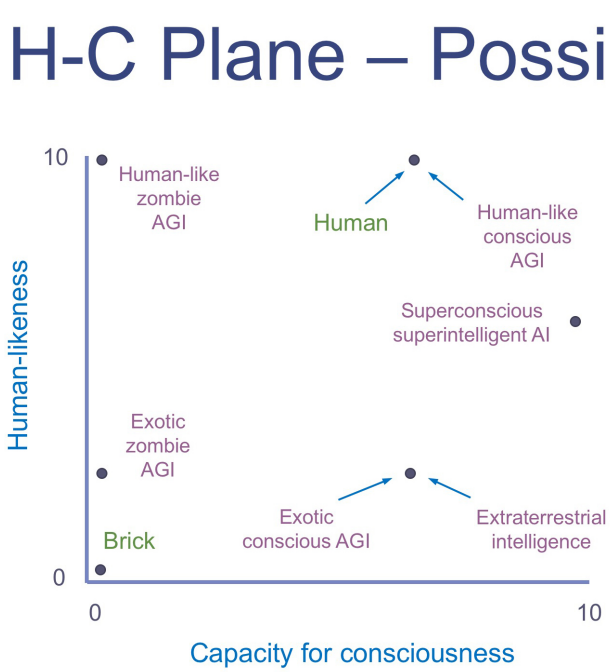


Figure 2. Exotica on the H-C plane

Take some possible forms of human-level artificial general intelligence (AGI), such as an AI built to mimic exactly the neural processing in the human brain. This could be achieved by copying the brain of a specific individual – scanning its structure in nanoscopic detail, replicating its physical behaviour in an artificial substrate, and embodying the result in a humanoid form. This process, known as 'whole brain emulation <http://www.tandfonline.com/doi/abs/10.1080/0952813X.2014.895113> ', would, in principle, yield something whose behaviour was indistinguishable from the original. So, being perfectly human-like, this would be an example of an artificial general intelligence with a 10 on the H-axis. Alternatively, rather than copying a specific person, an artificial brain could be constructed that matched a statistical description of a typical newborn's central nervous system. Duly embodied and reared like a human child, the result would be another perfectly human-like AGI.

Would these beings be conscious? Or rather, would we come to treat them the way we treat fellow conscious creatures, and would we describe them in the same terms? I conjecture that we would. Whatever prejudices we might start out with, their perfectly human-like behaviour would soon shape our feelings towards them to one of fellowship. So a human-like, conscious AGI is surely a possibility, and it would occupy the same spot on the H-C plane as a human.

But as we've already noted, there's no reason to suppose that the only way to build a human-level artificial general intelligence is to copy the biological brain. Perhaps an entirely different architecture could implement the same result. (*Ex Machina*'s Ava is a fictional example.) It might be possible to reach human-level intelligence using some

combination of brute force search techniques and machine learning with big data, perhaps exploiting senses and computational capacity unavailable to humans.

Such possibilities suggest several new kinds of being on the H-C plane. The first of these is the human-like zombie AI in the top left-hand corner. This entity not only has human-level intelligence, but is also thoroughly human-like in its behaviour, which can be described and explained using just the same language we use to describe human behaviour. However, it lacks consciousness. In Nagel's terms, it isn't like anything to be this thing. It is, in this sense, a phenomenological zombie.

Now, can we really imagine such a thing? Surely if its behaviour were indistinguishable from human behaviour, we would come to treat it in the way we treat each other. Surely, as we interacted with such beings, our attitude towards them would migrate towards fellowship, coming to see them as fellow conscious creatures and treating them as such. But suppose such an entity functioned merely by mimicking human behaviour. Through a future generation of very powerful machine-learning techniques, it has learned how to act in a convincingly human-like way in a huge variety of situations. If such an AGI says it is feeling sad, this is not because of a conflict between the way things are and the way it would like things to be, but rather because it has learned to say that it is sad in those particular circumstances. Would this alter our attitude? I conjecture that it would, that we would deny it consciousness, confining it to the left of the C-axis.

> We should entertain the likelihood that the richness of their conscious experiences would exceed human capacity

What sort of entity might be produced if someone – or most likely some corporation, organisation or government – set out to create an artificial successor to humankind, a being superior to *homo sapiens*? Whether idealistic, misguided or just plain crazy, they might reason that a future generation of artificial general intelligences could possess far greater intellectual powers than any human. Moreover, liberated from the constraints of biology, such beings could undertake long journeys into interstellar space that humans, with their fragile, short-lived bodies, would never survive. It would be AIs, then, who would go out to explore the wonders of the Universe up close. Because of the distances and timescales involved, the purpose of these AIs wouldn't be to relay information back to their creators. Rather, they would visit the stars on humanity's behalf. Let's call such hypothetical beings our 'mind children' <http://www.hup.harvard.edu/catalog.php?isbn=9780674576186> , a term borrowed from the Austrian roboticist Hans Moravec.

Now, where would these mind children appear on the H-C plane? Well, with no one waiting for a message home, there would seem to be little point in sending an artefact out to the stars that lacked the ability to consciously experience what it found. So the creators of our mind children would perhaps go for a biologically inspired brain-like

architecture, to ensure that they scored at least as well as humans on the C-axis. Indeed, we should entertain the likelihood that the richness of their conscious experiences would exceed human capacity, that they would enjoy a form of superconsciousness. This might be the case, for example, if they had a suite of sensors with a much larger bandwidth than a human's, or if they were able to grasp complex mathematical truths that are beyond human comprehension, or if they could hold a vast network of associations in their minds at once while we humans are confined to just a few.

As for the H-axis, a brain-inspired blueprint would also confer a degree of human-likeness on the AI. However, its superintelligence would probably render it hard for humans to fully understand. It would perhaps get 6 or 7. In short, our superintelligent, superconscious, artificially intelligent progeny are to be found at the right-hand end of the diagram, somewhat more than halfway up the H-axis.

What about non-brain-like artificial general intelligence? AGIs of this kind suggest several new data points on the H-C plane, all located lower down on the H-axis. These are the truly exotic AGIs, that is, opposite to human-like. The behaviour of an exotic being cannot be understood – or at least not fully understood – using the terms we usually use to make sense of human behaviour. Such a being might exhibit behaviour that is both complex and effective at attaining goals in a wide variety of environments and circumstances. However, it might be difficult or impossible for humans to figure out how it attains its goals, or even to discern exactly what those goals are. Wittgenstein's enigmatic remark that 'if a lion could talk we would not understand him' comes to mind. But a lion is a relatively familiar creature, and we have little difficulty relating to many aspects of its life. A lion inhabits the same physical world we do, and it apprehends the world using a similar suite of senses. A lion eats, mates, sleeps and defecates. We have a lot in common. The hypothesised exotic AGI is altogether more alien.

The most exotic sort of entity would be one that was wholly inscrutable, which is to say it would be *beyond the reach of anthropology*. Human culture is, of course, enormously varied. Neighbours from the same village often have difficulty relating to each other's habits, goals and preferences. Yet, through careful observation and interaction, anthropologists are able to make sense of this variety, rendering the practices of 'exotic' cultures – that is, very different from their own – comprehensible to them. But of course, we have even more in common with a fellow human being from a different culture than we do with a lion. Our shared humanity makes the anthropologist's task tractable. The sort of inscrutable entity we are trying to imagine is altogether more exotic. Even if we were able to engineer an encounter with it and to discern seemingly purposeful behaviour, the most expert team of anthropologists would struggle to divine its purposes or how they are fulfilled.

How might such an entity come about? After all, if it were engineered by humans, why would it not be comprehensible to humans? Well, there are a number of ways that an

AI might be created that wouldn't be understood by its creators. We have already seen that AlphaGo is capable of taking both its programmers and its opponents by surprise. A more powerful general intelligence might find far more surprising ways to achieve its goals. More radically, an AI that was the product of artificial evolution or of self-modification might end up with goals very different from those intended by its programmers. Furthermore, since we are granting the possibility of multifarious extraterrestrial intelligences, the space of possible minds must include not only those beings, but also whatever forms of artificial intelligence they might build. Whatever grip we are capable of getting on the mind of a creature from another world, a world that could be very different from our own, our grip is likely to be more tenuous still for an evolved or self-modified AI whose seed is a system devised to serve that creatures' already alien goals.

An exotic AI is clearly going to get a low score on the H-axis. But what about the C-axis? What might its capacity for consciousness be? Or, to put the matter differently, could we engineer an encounter with such a thing whereby, after sufficient observation and interaction, we would settle on our attitude towards it? If so, what would that attitude be? Would it be the sort of attitude we adopt towards a fellow conscious creature?

Well, now we have arrived at something of a philosophical impasse. Because the proffered definition of inscrutability puts the most exotic sort of AI beyond the reach of anthropology. And this seems to rule out the kind of encounter we require before we can settle on the right attitude towards it, at least according to a Wittgenstein-inspired, non-dualistic stance on subjectivity.

Is it possible to reconcile this view of consciousness with the existence of conscious exotica? Recall the white box thought experiment. Embedded in the mysterious box delivered to our laboratory, with its incomprehensibly complex but fully accessible internal dynamics, might be just the sort of inscrutable AI we are talking about. We might manage to engineer an encounter with the system, or some part of it, revealing seemingly purposeful behaviour, yet be unable to fathom just what that purpose was. An encounter with extraterrestrial intelligence would most likely present a similar predicament.

The novel *Solaris* (1961) by Stanislaw Lem offers a convincing fictional example. The novel's protagonists are a crew of scientists orbiting a planet covered by an ocean that turns out to be a single, vast, intelligent organism. As they attempt to study this alien being, it seems to be probing them in turn. It does this by creating human-like avatars out of their memories and unconscious who visit them aboard their spacecraft with disturbing psychological effects. For their part, the scientists never get to grips with the alien mind of this organism: 'Its undulating surface was capable of giving rise to the most diverse formations that bore no resemblance to anything terrestrial, on top of which the purpose – adaptive, cognitive, or whatever – of those often violent eruptions of plasmic "creativity" remained a total mystery.'

Suppose you were confronted by an exotic dynamical system such as the white box AI or the oceanic organism in *Solaris*. You want to know whether it is conscious or not. It's natural to think that for any given being, whether living or artificial, there is an answer to this question, a fact of the matter, even if the answer is necessarily hidden from us, as it appears to be in these hypothetical cases. On the other hand, if we follow Wittgenstein's approach to the issue, we go wrong when we think this way. Some facet of reality might be empirically inaccessible to us, but nothing is hidden as a matter of metaphysics.

Because these two standpoints are irreconcilable, our options at first appear to be just twofold. Either:

> a) retain the concept of conscious exotica, but abandon Wittgenstein and acknowledge that there is a metaphysically separate realm of subjectivity. This would be a return to the dualism of mind and body and the hard problem/easy problem dichotomy;

or

> b) retain a Wittgenstein-inspired approach to consciousness, insisting that 'nothing is hidden', but reject the very idea of conscious exotica. As a corollary, we would have to relinquish the project of mapping the space of possible minds onto the H-C plane.

However, there is a third option:

> c) retain both the concept of conscious exotica and a Wittgenstein-inspired philosophical outlook by allowing that our language and practices could change in unforeseeable ways to accommodate encounters with exotic forms of intelligence.

We have been going along with the pretence that consciousness is a single, monolithic concept amenable to a scalar metric of capacity. This sort of manoeuvre is convenient in many branches of enquiry. For conservation purposes, an ecologist can usefully compress biodiversity into a single statistic, abstracting away from differences between species, seasonal changes, spatial distribution and so on. In economics, the 'human development index' usefully summarises aspects of a country's education system, healthcare, productivity and the like, ignoring the numerous details of individual lives. However, for some purposes, a more nuanced approach is needed. Examined more closely, the concept of consciousness encompasses many things, including awareness of the world (or primary consciousness), self-awareness (or higher-order consciousness), the capacity for emotion and empathy, and cognitive integration (wherein the brain's full resources are brought to bear on the ongoing situation).

> Parts of our language to describe highly exotic entities with

complex behaviour might be supplanted by wholly new ways of talking

In a normal, adult human being, these things come bundled together. But in a more exotic entity they might be disaggregated. In most non-human animals we find awareness of the world without self-awareness, and the capacity for suffering without the capacity for empathy. A human-level AI might display awareness of the world and self-awareness without the capacity for emotion or empathy. If such entities became familiar, our language would change to accommodate them. Monolithic concepts such as consciousness might break apart, leading to new ways of talking about the behaviour of AIs.

More radically, we might discover whole new categories of behaviour or cognition that are loosely associated with our old conception of consciousness. In short, while we might retain bits of today's language to describe highly exotic entities with complex behaviour, other relevant parts of our language might be reshaped, augmented or supplanted by wholly new ways of talking, a process that would be informed by computer science, by comparative cognition, by behavioural psychology and by the natural evolution of ordinary language. Under these conditions, something like 'capacity for consciousness' might be usefully retained as a summary statistic for those entities whose behaviour eludes explanation in today's terms but could be accommodated by a novel conceptual framework wherein the notion of consciousness now familiar to us, though fragmented and refracted, remains discernible.

What are the implications of this possibility for the H-C plane? Figure 2 above indicates a point on the H-C plane with the same value as a human on the C-axis, but which is exotic enough to lie on the H-axis at the limit of applicability of any form of consciousness. Here we find entities, both extraterrestrial and artificial, that possess human-level intelligence but whose behaviour bears little resemblance to human behaviour.

Nevertheless, given sufficiently rich interaction with and/or observation of these entities, we would come to see them as fellow conscious beings, albeit having modified our language to accommodate their eccentricities. One such entity, the exotic conscious AGI, has a counterpart at the left-hand end of the H-C plane, namely the exotic zombie AGI. This is a human-level AI whose behaviour is similarly non-human-like, but that we are unable to see as conscious however much we interact with it or observe it. These two data points – the exotic, conscious, human-level intelligence and the exotic, zombie, human-level intelligence – define the bottom two corners of a square whose other corners are humans themselves at the top right, and human-like zombies at the top left. This square illustrates the inextricable, three-way relationship between human-level intelligence, human-likeness, and consciousness.

We can now identify a number of overlapping regions within our embryonic space of possible minds. These are depicted in Figure 3 below. On the (debatable) assumption that, if an entity is conscious at all, its capacity for consciousness will correlate with its cognitive prowess, human-level intelligence features in the two, parallel purple regions, one to the far left of the diagram and one at human level on the C-axis. The exotic, conscious AGI resides at the bottom of the latter region, and also at the left-hand end of the orange region of conscious exotica. This region stretches to the right of the C-axis, beyond the human level, because it encompasses exotic beings, which could be extraterrestrial or artificial or both, with superhuman intelligence and a superhuman capacity for consciousness. Our 'mind children' are less exotic forms of possible superintelligent, superconscious creatures. But the conscious exotica here are, perhaps, the most interesting beings in the space of possible minds, since they reside at the limit of what we would welcome into the fellowship of consciousness, yet sit on a boundary beyond which everything complex is inscrutably strange.

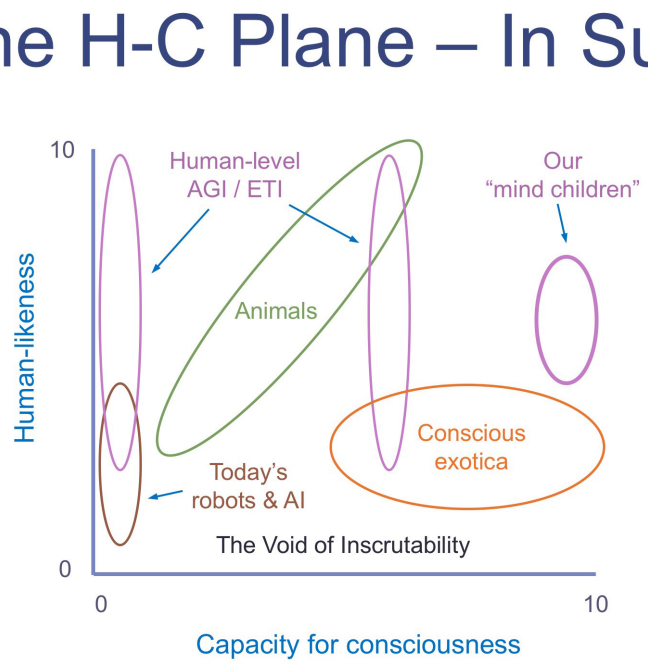


Figure 3. Notable regions of the H-C plane

This boundary marks the edge of a region that is empty, denoted the 'Void of Inscrutability'. It is empty because, as Wittgenstein remarks, we say only of a human being and what behaves like one that it is conscious. We have stretched the notion of what behaves like a human being to breaking point (perhaps further than Wittgenstein would find comfortable). As we approach that breaking point, I have suggested, today's language of consciousness begins to come apart. Beyond that point we find only entities for which our current language has no application. Insofar as they exhibit complex behaviour, we are obliged to use other terms to describe and explain it, so these entities are no further along the C-axis than the brick. So the lowest strip of the diagram has no data points at all. It does not contain entities that

are inscrutable but who might – for all we know – be conscious. To think this would be to suppose that there are facts of the matter about the subjectivity of inscrutably exotic entities that are forever closed off to us. We can avoid the dualism this view entails by accepting that this region is simply a void.

The void of inscrutability completes my provisional sketch of the space of possible minds. But what have we gained from this rather fanciful exercise? The likelihood of humans directly encountering extraterrestrial intelligence is small. The chances of discovering a space-borne signal from another intelligent species, though perhaps greater, are still slight. But artificial intelligence is another matter. We might well create autonomous, human-level artificial intelligence in the next few decades. If this happens, the question of whether, and in what sense, our creations are conscious will become morally significant. But even if none of these science-fiction scenarios comes about, to situate human consciousness within a larger space of possibilities strikes me as one of the most profound philosophical projects we can undertake. It is also a neglected one. With no giants upon whose shoulders to stand, the best we can do is cast a few flares into the darkness.

Murray Shanahan is professor of cognitive robotics at Imperial College London and a Spoke Leader at the Leverhulme Centre for the Future of Intelligence. His latest book is The Technological Singularity (2015).

aeon.co 19 October, 2016