A theory of my own mind

Knowing the content of one's own mind might seem straightforward but in fact it's much more like mindreading other people

By Stephen M Fleming

Edited by Pam Weintraub

20210923

Stephen M Fleming is professor of cognitive neuroscience at University College London, where he leads the Metacognition Group. He is author of Know Thyself: The Science of Self-awareness (2021).

Source: https://aeon.co/essays/is-there-a-symmetry-between-metacognition-and-mindreading

1. In 1978, David Premack and Guy Woodruff published a paper that would go on to become famous in the world of academic psychology. Its title posed a simple question: does the chimpanzee have a theory of mind?

In coining the term ‘theory of mind’, Premack and Woodruff were referring to the ability to keep track of what someone else thinks, feels or knows, even if this is not immediately obvious from their behaviour. We use theory of mind when checking whether our colleagues have noticed us zoning out on a Zoom call – did they just

see that? A defining feature of theory of mind is that it entails second-order representations, which might or might not be true. I might think that someone else thinks that I was not paying attention but, actually, they might not be thinking that at all. And the success or failure of theory of mind often turns on an ability to appropriately represent another person's outlook on a situation. For instance, I can text my wife and say: 'I'm on my way,' and she will know that by this I mean that I'm on my way to collect our son from nursery, not on my way home, to the zoo, or to Mars. Sometimes

this can be difficult to do, as captured by a *New Yorker* cartoon caption of a couple at loggerheads: ‘Of course I care about how you imagined I thought you perceived I wanted you to feel.’

Premack and Woodruff’s article sparked a deluge of innovative research into the origins of theory of mind. We now know that a fluency in reading minds is not something humans are born with, nor is it something guaranteed to emerge in development. In one classic experiment, children were told stories such as the following:

> Maxi has put his chocolate in the cupboard. While Maxi is away, his mother moves the chocolate from the cupboard to the drawer. When Maxi comes back, where will he look for the chocolate?

Until the age of four, children often fail this test, saying that Maxi will look for the chocolate where it *actually* is (the drawer), rather than where *he* thinks it is (in the cupboard). They are using their knowledge of the reality to answer the question, rather than what they know about where Maxi had put the chocolate before he left. Autistic children also tend to give the wrong answer, suggesting problems with

tracking the mental states of others. This test is known as a ‘false belief’ test – passing it requires one to realise that Maxi has a different (and false) belief about the world.

Many researchers now believe that the answer to Premack and Woodruff’s question is, in part, ‘no’ – suggesting that fully fledged theory of mind might be unique to humans. If chimpanzees are given an ape equivalent of the Maxi test, they don’t use the fact that another chimpanzee has a false belief about the location of the food to sneak in and grab it. Chimpanzees can track knowledge states – for instance,

being aware of what others see or do not see, and knowing that, when someone is blindfolded, they won't be able to catch them stealing food. There is also evidence that they track the difference between true and false beliefs in the pattern of their eye movements, similar to findings in human infants. Dogs also have similarly sophisticated perspective-taking abilities, preferring to choose toys that are in their owner's line of sight when asked to fetch. But so far, at least, only adult humans have been found to act on an understanding that

other minds can hold different beliefs about the world to their own.

Research on theory of mind has rapidly become a cornerstone of modern psychology. But there is an underappreciated aspect of Premack and Woodruff's paper that is only now causing ripples in the pond of psychological science. Theory of mind as it was originally defined identified a capacity to impute mental states not only to others *but also to ourselves*. The implication is that thinking about others is just one manifestation of a rich – and perhaps much broader – capacity to build what philosophers

call metarepresentations, or representations of representations. When I wonder whether you know that it's raining, and that our plans need to change, I am metarepresenting the state of your knowledge about the weather.

Intriguingly, metarepresentations are – at least in theory – symmetric with respect to self and other: I can think about your mind, and I can think about my own mind too. The field of metacognition research, which is what my lab at University College London works on, is interested in the latter – people's judgments about their own cognitive processes.

The beguiling question, then – and one we don't yet have an answer to – is whether these two types of 'meta' are related. A potential symmetry between self-knowledge and other-knowledge – and the idea that humans, in some sense, have learned to turn theory of mind on themselves – remains largely an elegant hypothesis. But an answer to this question has profound consequences. If self-awareness is 'just' theory of mind directed at ourselves, perhaps it is less special than we like to believe. And if we learn about ourselves in the same way as we learn about others,

perhaps we can also learn to know ourselves better.

2. A common view is that self-knowledge is special, and immune to error, because it is gained through introspection – literally, 'looking within'. While we might be mistaken about things we perceive in the outside world (such as thinking a bird is a plane), it seems odd to say that we are wrong about our own minds. If I *think* that I'm feeling sad or anxious, then there is a sense in which I *am* feeling sad or anxious. We have untrammelled access to our own minds, so the argument goes, and this immediacy of introspection

means that we are rarely wrong about ourselves.

This is known as the ‘privileged access’ view of self-knowledge, and has been dominant in philosophy in various guises for much of the 20th century. René Descartes relied on self-reflection in this way to reach his conclusion ‘I think, therefore I am,’ noting along the way that: ‘I know clearly that there is nothing that can be perceived by me more easily or more clearly than my own mind.’

An alternative view suggests that we *infer* what we think or believe from a variety of cues – just as we infer

what others think or feel from observing their behaviour. This suggests that self-knowledge is not as immediate as it seems. For instance, I might infer that I am anxious about an upcoming presentation because my heart is racing and my breathing is heavier. But I might be wrong about this – perhaps I am just feeling excited. This kind of psychological reframing is often used by sports coaches to help athletes maintain composure under pressure.

The philosopher most often associated with the inferential view is Gilbert Ryle, who proposed in *The*

Concept of Mind (1949) that we gain self-knowledge by applying the tools we use to understand other minds to ourselves: ‘The sorts of things that I can find out about myself are the same as the sorts of things that I can find out about other people, and the methods of finding them out are much the same.’ Ryle’s idea is neatly summarised by another *New Yorker* cartoon in which a husband says to his wife: ‘How should I know what I’m thinking? I’m not a mind reader.’

Many philosophers since Ryle have considered the strong inferential view as somewhat crazy, and written

it off before it could even get going. The philosopher Quassim Cassam, author of *Self-knowledge for Humans* (2014), describes the situation:

> Philosophers who defend inferentialism – Ryle is usually mentioned in this context – are then berated for defending a patently absurd view. The assumption that intentional self-knowledge is normally immediate … is rarely defended; it's just seen as obviously correct.

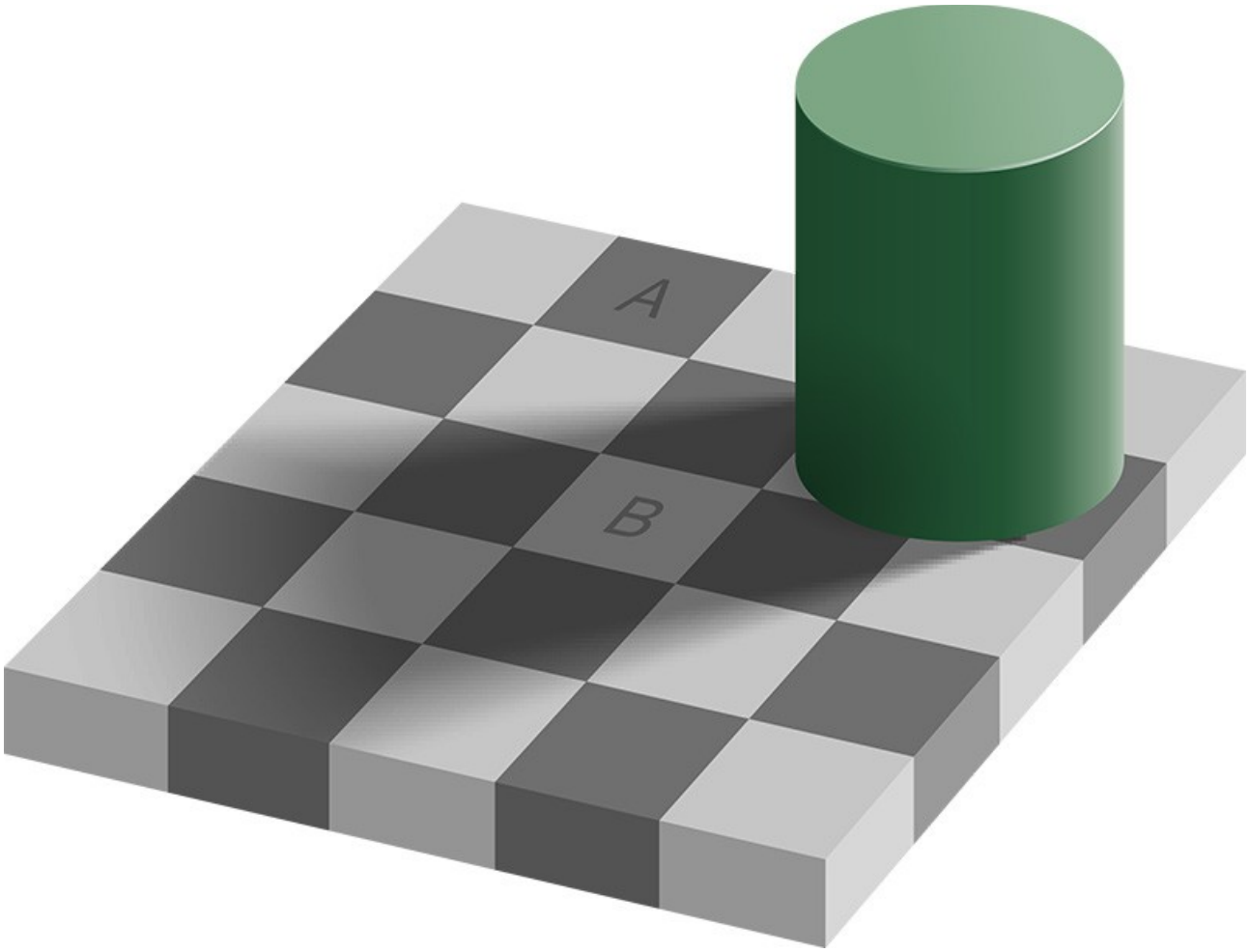
But if we take a longer view of history, the idea that we have some sort of special, direct access to our

minds is the exception, rather than the rule. For the ancient Greeks, self-knowledge was not all-encompassing, but a work in progress, and something to be striven toward, as captured by the exhortation to ‘know thyself’ carved on the Temple of Delphi. The implication is that most of us *don’t* know ourselves very well. This view persisted into medieval religious traditions: the Italian priest and philosopher Saint Thomas Aquinas suggested that, while God knows himself by default, we need to put in time and effort to know our own minds. And a similar notion of

striving toward self-awareness is found in Eastern traditions, with the founder of Chinese Taoism, Lao Tzu, endorsing a similar goal: ‘To know that one does not know is best; not to know but to believe that one knows is a disease.’

Other aspects of the mind – most famously, perception – also appear to operate on the principles of an (often unconscious) inference. The idea is that the brain isn’t directly in touch with the outside world (it’s locked up in a dark skull, after all) – and instead has to ‘infer’ what is really out there by constructing and updating an internal model of the

environment, based on noisy sensory data. For instance, you might know that your friend owns a Labrador, and so you expect to see a dog when you walk into her house, but don't know exactly where in your visual field the dog will appear. This higher-level expectation – the spatially invariant concept of 'dog' – provides the relevant context for lower levels of the visual system to easily interpret dog-shaped blurs that rush toward you as you open the door.



Adelson's checkerboard. Courtesy Wikipedia

Elegant evidence for this perception-as-inference view comes from a range of striking visual illusions. In one called Adelson's checkerboard, two patches with the same objective luminance are perceived as lighter and darker because the brain assumes that, to

reflect the same amount of light, the one in shadow must have started out brighter. Another powerful illusion is the ‘light from above’ effect – we have an automatic tendency to assume that natural light falls from above, whereas uplighting – such as when light from a fire illuminates the side of a cliff – is less common. This can lead the brain to interpret the same image as either bumps or dips in a surface, depending on whether the shadows are consistent with light falling from above. Other classic experiments show that information from one sensory modality, such as sight, can

act as a constraint on how we perceive another, such as sound – an illusion used to great effect in ventriloquism. The real skill of ventriloquists is being able to talk without moving the mouth. Once this is achieved, the brains of the audience do the rest, pulling the sound to its next most likely source, the puppet.

These striking illusions are simply clever ways of exposing the workings of a system finely tuned for perceptual inference. And a powerful idea is that self-knowledge relies on similar principles – whereas perceiving the outside

world relies on building a *model* of what is out there, we are also continuously building and updating a similar model of ourselves – our skills, abilities and characteristics. And just as we can sometimes be mistaken about what we perceive, sometimes the model of ourselves can also be wrong.

Let's see how this might work in practice. If I need to remember something complicated, such as a shopping list, I might judge I will fail unless I write it down somewhere. This is a metacognitive judgment about how good my memory is. And this model can be updated – as I

grow older, I might think to myself that my recall is not as good as it used to be (perhaps after experiencing myself forgetting things at the supermarket), and so I lean more heavily on list-writing. In extreme cases, this self-model can become completely decoupled from reality: in functional memory disorders, patients believe their memory is poor (and might worry they have dementia) when it is actually perfectly fine when assessed with objective tests.

We now know from laboratory research that metacognition, just like perception, is also subject to

powerful illusions and distortions – lending credence to the inferential view. A standard measure here is whether people's confidence tracks their performance on simple tests of perception, memory and decision-making. Even in otherwise healthy people, judgments of confidence are subject to systematic illusions – we might feel more confident about our decisions when we act more quickly, even if faster decisions are not associated with greater accuracy. In our research, we have also found surprisingly large and consistent differences between individuals on these measures – one person might

have limited insight into how well they are doing from one moment to the next, while another might have good awareness of whether are likely to be right or wrong.

This metacognitive prowess is independent of general cognitive ability, and correlated with differences in the structure and function of the prefrontal and parietal cortex. In turn, people with disease or damage to these brain regions can suffer from what neurologists refer to as anosognosia – literally, the absence of knowing. For instance, in Alzheimer's disease, patients can suffer a cruel

double hit – the disease attacks not only brain regions supporting memory, but also those involved in metacognition, leaving people unable to understand what they have lost.

This all suggests – more in line with Socrates than Descartes – that self-awareness is something that can be cultivated, that it is not a given, and that it can fail in myriad interesting ways. And it also provides newfound impetus to seek to understand the *computations* that might support self-awareness. This is where Premack and Woodruff's more

expansive notion of theory of mind might be long overdue another look.

3. Saying that self-awareness depends on similar machinery to theory of mind is all well and good, but it begs the question – what is this machinery? What do we mean by a 'model' of a mind, exactly?

Some intriguing insights come from an unlikely quarter – spatial navigation. In classic studies, the psychologist Edward Tolman realised that the rats running in mazes were building a 'map' of the maze, rather than just learning which turns to make when. If the shortest route from a starting point towards the cheese is suddenly

blocked, then rats readily take the next quickest route – without having to try all the remaining alternatives. This suggests that they have not just rote-learned the quickest path through the maze, but instead know something about its overall layout.

A few decades later, the neuroscientist John O'Keefe found that cells in the rodent hippocampus encoded this internal knowledge about physical space. Cells that fired in different locations became known as 'place' cells. Each place cell would have a preference for a specific position in the maze but, when combined together, could

provide an internal ‘map’ or model of the maze as a whole. And then, in the early 2000s, the neuroscientists May-Britt Moser, Edvard Moser and their colleagues in Norway found an additional type of cell – ‘grid’ cells, which fire in multiple locations, in a way that tiles the environment with a hexagonal grid. The idea is that grid cells support a metric, or coordinate system, for space – their firing patterns tell the animal how far it has moved in different directions, a bit like an in-built GPS system.

There is now tantalising evidence that similar types of brain cell also encode abstract conceptual spaces.

For instance, if I am thinking about buying a new car, then I might think about how environmentally friendly the car is, and how much it costs. These two properties map out a two-dimensional 'space' on which I can place different cars – for instance, a cheap diesel car will occupy one part of the space, and an expensive electric car another part of the space. The idea is that, when I am comparing these different options, my brain is relying on the same kind of systems that I use to navigate through *physical* space. In one experiment by Timothy Behrens and his team at the University of

Oxford, people were asked to imagine morphing images of birds that could have different neck and leg lengths – forming a two-dimensional bird space. A grid-like signature was found in the fMRI data when people were thinking about the birds, even though they never saw them presented in 2D.

So far, these lines of work – on abstract conceptual models of the world, and on how we think about other minds – have remained relatively disconnected, but they are coming together in fascinating ways. For instance, grid-like codes are also found for conceptual maps of

the *social* world – whether other individuals are more or less competent or popular – suggesting that our thoughts about others seem to be derived from an internal model similar to those used to navigate physical space. And one of the brain regions involved in maintaining these models of other minds – the medial prefrontal cortex (PFC) – is also implicated in metacognition about our *own* beliefs and decisions. For instance, research in my group has discovered that medial prefrontal regions not only track confidence in individual decisions, but also 'global' metacognitive

estimates of our abilities over longer timescales – exactly the kind of self-estimates that were distorted in the patients with functional memory problems.

Recently, the psychologist Anthony G Vaccaro and I surveyed the accumulating literature on theory of mind and metacognition, and created a brain map that aggregated the patterns of activations reported across multiple papers. Clear overlap between brain activations involved in metacognition and mindreading was observed in the medial PFC. This is what we would expect if there was a common

system building models not only about other people, but also of ourselves – and perhaps about ourselves in relation to other people. Tantalisingly, this very same region has been shown to carry grid-like signatures of abstract, conceptual spaces.

At the same time, computational models are being built that can mimic features of both theory of mind and metacognition. These models suggest that a key part of the solution is the learning of *second-order* parameters – those that encode information about *how* our minds are working, for instance

whether our percepts or memories tend to be more or less accurate. Sometimes, this system can become confused. In work led by the neuroscientist Marco Wittmann at the University of Oxford, people were asked to play a game involving tracking the colour or duration of simple stimuli. They were then given feedback about both their own performance and that of other people. Strikingly, people tended to ‘merge’ their feedback with those of others – if others were performing better, they tended to think they themselves were performing a bit better too, and vice-versa. This

intertwining of our models of self-performance and other-performance was associated with differences in activity in the dorsomedial PFC. Disrupting activity in this area using transcranial magnetic stimulation (TMS) led to *more* self-other mergence – suggesting that one function of this brain region is not only to create models of ourselves and others, but also to keep these models apart.

Another implication of a symmetry between metacognition and mindreading is that both abilities should emerge around the same time in childhood. By the time that

children become adept at solving false-belief tasks – around the age of four – they are also more likely to engage in self-doubt, and recognise when they themselves were wrong about something. In one study, children were first presented with 'trick' objects: a rock that turned out to be a sponge, or a box of Smarties that actually contained not sweets but pencils. When asked what they first thought the object was, three-year-olds said that they knew all along that the rock was a sponge and that the Smarties box was full of pencils. But by the age of five, most children recognised that their first

impression of the object was false – they could recognise they had been in error.

Indeed, when Simon Baron-Cohen, Alan Leslie and Uta Frith outlined their influential theory of autism in the 1980s, they proposed that theory of mind was only 'one of the manifestations of a basic metarepresentational capacity'. The implication is that there should also be noticeable differences in metacognition that are linked to changes in theory of mind. In line with this idea, several recent studies have shown that autistic individuals also show differences in

metacognition. And in a recent study of more than 450 people, Elisa van der Plas, a PhD student in my group, has shown that theory of mind ability (measured by people's ability to track the feelings of characters in simple animations) and metacognition (measured by the degree to which their confidence tracks their task performance) are significantly correlated with each other. People who were better at theory of mind also formed their confidence differently – they were more sensitive to subtle cues, such as their response times, that

indicated whether they had made a good or bad decision.

4. Recognising a symmetry between self-awareness and theory of mind might even help us understand why human self-awareness emerged in the first place. The need to coordinate and collaborate with others in large social groups is likely to have prized the abilities for metacognition and mindreading. The neuroscientist Suzana Herculano-Houzel has proposed that primates have unusually efficient ways of cramming neurons into a given brain volume – meaning there is simply more processing power devoted to so-called higher-order functions –

those that, like theory of mind, go above and beyond the maintenance of homeostasis, perception and action. This idea fits with what we know about the areas of the brain involved in theory of mind, which tend to be the most distant in terms of their connections to primary sensory and motor areas.

A symmetry between self-awareness and other-awareness also offers a subversive take on what it means for other agents such as animals and robots to be self-aware. In the film *Her* (2013), Joaquin Phoenix's character Theodore falls in love with his virtual

assistant, Samantha, who is so human-like that he is convinced she is conscious. If the inferential view of self-awareness is correct, there is a sense in which Theodore's belief that Samantha is aware is sufficient to *make* her aware, in his eyes at least. This is not quite true, of course, because the ultimate test is if she is able to also recursively model Theodore's mind, and create a similar model of herself. But being convincing enough to share an intimate connection with another conscious agent (as Theodore does with Samantha), replete with mindreading and reciprocal

modelling, might be possible only if both agents have similar recursive capabilities firmly in place. In other words, attributing awareness to ourselves and to others might be what makes them, and us, conscious.

Finally, a symmetry between self-awareness and other-awareness also suggests novel routes towards boosting our own self-awareness. In a clever experiment conducted by the psychologists and metacognition experts Rakefet Ackerman and Asher Koriat in Israel, students were asked to judge both how well they had learned a topic, and how well

other students had learned the same material, by watching a video of them studying. When judging themselves, they fell into a trap – they believed that spending *less* time studying was a signal of being confident in knowing the material. But when judging others, this relationship was reversed: they (correctly) judged that spending longer on a topic would lead to better learning. These results suggest that a simple route for improving self-awareness is to take a third-person perspective on ourselves. In a similar way, literary novels (and soap operas)

encourage us to think about the minds of others, and in turn might shed light on our own lives.

There is still much to learn about the relationship between theory of mind and metacognition. Most current research on metacognition focuses on the ability to think about our experiences and mental states – such as being confident in what we see or hear. But this aspect of metacognition might be distinct from how we come to know our own, or others', character and preferences – aspects that are often the focus of research on theory of mind. New and creative experiments will be

needed to cross this divide. But it seems safe to say that Descartes's classical notion of introspection is increasingly at odds with what we know of how the brain works. Instead, our knowledge of ourselves is (meta)knowledge like any other – hard-won, and always subject to revision. Realising this is perhaps particularly useful in an online world deluged with information and opinion, when it's often hard to gain a check and balance on what we think and believe. In such situations, the benefits of accurate metacognition are myriad – helping us recognise our faults and

collaborate effectively with others.

As the poet Robert Burns tells us:

O wad some Power the giftie gie us
To see oursels as ithers see us!
It wad frae mony a blunder free us…

(Oh, would some Power give us the gift
To see ourselves as others see us!
It would from many a blunder free us…)

[Formatted for e-book readers by Wergosum on 20210930]