Invited Prize Paper

QJEP

Quarterly Journal of Experimental Psychology
2023, Vol. 76(8) 1711–1723

DOI: 10.1177/17470218231164502
qjep.sagepub.com
SAGE

Consciousness, (meta)cognition, and culture

Chris D Frith[1,2]

Abstract

Our conscious experience is determined by a combination of top-down processes (e.g., prior beliefs) and bottom-up processes (e.g., sensations). The balance between these two processes depends on estimates of their reliability (precision), so that the estimate considered more reliable is given more weight. We can modify these estimates at the metacognitive level, changing the relative weights of priors and sensations. This enables us, for example, to direct our attention to weak stimuli. But there is a cost to this malleability. For example, excessive weighting of top-down processes, as in schizophrenia, can lead to perceiving things that are not there and believing things that are not true. It is only at the top of the brain's cognitive hierarchy that metacognitive control becomes conscious. At this level, our beliefs concern complex, abstract entities with which we have limited direct experience. Estimates of the precision of such beliefs are more uncertain and more malleable. However, at this level, we do not need to rely on our own limited experience. We can rely instead on the experiences of others. Explicit metacognition plays a unique role, enabling us to share our experiences. We acquire our beliefs about the world from our immediate social group and from our wider culture. And the same sources provide us with better estimates of the precision of these beliefs. Our confidence in our high-level beliefs is heavily influenced by culture at the expense of direct experience.

Keywords
Precision; schizophrenia; habit; top-down control

Received: 30 September 2022; revised: 28 December 2022; accepted: 9 January 2023

Remembering Bartlett

When I began to prepare my talk, I decided that I ought to learn something about Bartlett so that I could bring him into his lecture. I remembered that I had a copy of his book *Remembering* (Bartlett, 1932) which I had bought for 10 shillings when I was a student in Cambridge. I was rather surprised by what I found. Sir Frederic[1] Bartlett is renowned as the founder of the sort of hard-line experimental psychology that continues to be promoted by the Experimental Psychology Society (EPS) and published in the *Quarterly Journal of Experimental Psychology* (QJEP). And yet his book is subtitled *A Study in Experimental and Social Psychology*, and there is a Part II which is entitled *Remembering as a Study in Social Psychology.* On page 242 he says, " . . . everything in psychology belongs to social psychology . . ." So, what became of the social in the hard-line experimental psychology he promoted in Cambridge?

I soon discovered that I was not the first person to ask this question (see, for example, Costall, 1992; Why British psychology is not social). Indeed, it seems that even Bartlett himself was not entirely happy with the way British psychology developed. Beate Hermelin complained to him about how tedious EPS meetings could be. "Oh yes," Bartlett confided to her, "It's all gone wrong. I wish I'd written novels instead" (Costall, 2009).

Bartlett's promotion of a limited form of hard-line experimental psychology was probably aimed at getting this new discipline recognised as a natural science like physiology and to make a clear distinction from philosophy. At that time, the social phenomena associated with psychology

[1]Wellcome Centre for Human Neuroimaging, University College London, London, UK
[2]Institute of Philosophy, School of Advanced Study, University of London, London, UK

Corresponding author:
Chris D Frith, Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3AR, UK.
Email: c.frith@ucl.ac.uk

seemed simply too ill defined and too difficult to study. To some extent, my own career has exemplified the slow and painful process of regaining Bartlett's insight that everything in psychology belongs to social psychology.

Subjective experiences (consciousness) and the brain

When I was a student, I was not actually reading Bartlett. Like everyone else in the 1960s, I was reading Aldous Huxley's *The Doors of Perception*, where Huxley suggests that, by taking psychedelic drugs, we can see the world as it really is, unmediated by words and symbols. I was also reading the books of Philip K Dick which frequently explore the idea that there is something hidden behind our everyday reality (e.g., *The Penultimate Truth*).

At the very beginning of his book, Huxley (1959) notes that "these changes [produced by mescalin] are similar to those which occur in that most characteristic plague of the twentieth century, schizophrenia. Is the mental disorder due to a chemical disorder?" (p. 3). So, do patients with schizophrenia have better access to a hidden reality than the rest of us? After graduating, I was trained in clinical psychology. Perhaps I thought that, by studying patients, I could find out about this hidden reality? To jump ahead a bit, I should probably reveal that I have not discovered this hidden reality. It is not so much that there is another hidden reality, but rather that reality is well hidden.

Hallucinations and delusions are the symptoms that most usually lead to a diagnosis of schizophrenia. Hallucinations are false perceptions—perceptions in the absence of sensory input. These usually involve hearing voices, rather than the brightly coloured visual experiences associated with psychedelic drugs. Delusions are false beliefs about the world. However, the subjective experiences reported by patients are not always easy to classify as either hallucinations or delusions. Here are some examples of the "first-rank" symptoms that are particularly associated with schizophrenia (Schneider, 1959; Soares-Weiser et al., 2015).

Thought Broadcasting: "It was like my thoughts (were) shouted out"

Thought Insertion: "Thoughts come into my head like 'Kill god'. It's just like my mind working, but it isn't. They come from this chap, Chris. They're his thoughts."

(Both from Leff, 1982)

Delusions of Control: "My fingers pick up the pen, but I don't control them."

(From Mellor, 1970)

The distinction between perceptions (hallucinations) and beliefs (delusions) is not always very clear. For example, patients with delusions of control do not simply believe that someone is controlling their actions. They have an experience—it feels to them as if their actions are being controlled.

In 1975, I joined a Medical Research Council research unit dedicated largely to the study of schizophrenia. Here, I learned of the important distinction made by psychiatrists between signs and symptoms. Symptoms, such as *hallucinations* or *delusions*, are subjective experiences reported by the patient. In contrast, signs, such as *incoherence of speech* or *catatonia*, are objectively observable behaviour. The key point I take from this distinction is that symptoms, such as hallucinations and delusions, are subjective experiences. In this sense, schizophrenia is a disorder of consciousness (Frith, 1979).

The treatment of choice for schizophrenia then, and still today, is antipsychotic drugs (Haddad & Correll, 2018). In the 1970s, it had been shown that there was a strong correlation between the clinical effectiveness of these drugs and their ability to block dopamine receptors (Seeman & Lee, 1975). I was involved in an elegant study of this effect using the antipsychotic drug flupenthixol (Johnstone et al., 1978). This drug exists in two isomeric forms. Both are psychoactive, but only one form blocks dopamine receptors. We found that it was only this α-isomer that reduced the severity of psychiatric ratings as assessed by clinicians. But we also noted that this dopamine blocking version of the drug reduced the severity of symptoms, such as hallucinations and delusions, but not behavioural signs such as poverty of speech or motor retardation.

Here, I saw a connection between a rather basic aspect of brain function and subjective experiences. This was probably the moment when I realised that, by studying psychology, we might be finding out how the brain works. Of course, I was not alone in noting this link between mind and brain. This time was the heyday of cognitive neuropsychology, and I was fascinated by the case studies of neurological patients being reported by neuropsychologists, such as Elizabeth Warrington and Tim Shallice. For example, the demonstration that brain damage can lead to a loss of understanding of specific semantic categories (Warrington & Shallice, 1984). At the time this seemed very surprising, but it is now a well-established phenomenon (Martin, 2007).

The problem with positive symptoms

I find it straightforward enough to accept that damage to a brain region can knock out some aspect of subjective experience (as, for example, when damage to V4 in the extrastriate visual cortex can eliminate the experience of colour; Zeki, 1990). But hallucinations are more difficult to understand. Rather than a lack of subjective experience, we have experiences that are abnormal by their presence (positive symptoms). What mechanism can explain the emergence of these "interloping experiences"? One suggestion, associated with Hughlings Jackson (Berrios, 1985), is that a normal function of healthy brain tissue is released by the removal of "top to bottom" inhibition. An example of this mechanism in action is the "utilisation behaviour" associated with frontal lobe lesions (Lhermitte, 1983). Such patients are unable to resist grasping an object placed in front of them, for example, putting on a pair of glasses on top of another pair. Here, the frontal damage removes the mechanism which inhibits the tendency for actions to be automatically elicited by environmental stimuli.

This example reminds us of the importance of top-down processes in cognition. A top-down process emerges from a higher level in the neural hierarchy and modifies the way lower levels can operate, changing the way an action is performed or a stimulus is perceived. For example, dorsolateral prefrontal cortex (dlPFC) has an important role making sure that only the actions appropriate to the task in hand will be performed. As a result, relevant actions are primed and irrelevant ones are suppressed. I have suggested that what dlPFC is doing in this case is "sculpting the response space" (Frith, 2000; see also Miller & Cohen, 2001).

The importance of such top-down processes in perception was recognised by Hermann von Helmholtz. He pointed out that the sensory stimulus alone is not enough for us to know what is out there in the world (Helmholtz, 1867/1948). We need to infer what is out there on the basis of sensory evidence (bottom-up) and prior expectations about what is likely (top-down). In his book *Remembering*, Bartlett was mainly concerned with demonstrating the constructive processes involved in memory, but he also recognised the importance of such processes for perception (". . . even the most elementary looking perceptual processes can be shown . . . to have the character of inferential construction," p. 33).

Once we recognise the importance of top-down processes in perception, positive symptoms, such as hallucinations, cease to be quite so mysterious. What we perceive is biased by our prior expectations and these expectations can sometimes be sufficient to determine what we perceive (Corlett et al., 2019). This is the case for visual imagery, for example, which is associated with increased top-down connectivity from frontal and parietal regions to occipital cortex (Dijkstra et al., 2017). Both visual imagery and visual perception are associated with activity in early visual areas, and if that signal is sufficiently strong and precise, an internally generated signal can be mistaken for reality (Dijkstra & Fleming, 2021). This is an example of a hallucination.

Simple hallucinations of this sort can readily be created in people and even in mice (Schmack et al., 2021). In these experiments, mice (or people) had to detect very weak auditory signals buried in noise. If they report hearing a tone with high confidence when none was actually present, this is considered to be a hallucination. Such false alarms are not infrequent when the signal is difficult to detect. Their frequency is higher when the proportion of signals to non-signals is higher, indicating an effect of expectations. It is also higher in people who are prone to hallucinate.

In a Bayesian framework (see, for example, Knill & Pouget, 2004), the sensory signal is treated as evidence, while the prior expectation is a hypothesis about the state of the world. The Bayes' equation indicates whether the new evidence is sufficient to change our perception of the world. This process operates in the same way on perceptions as it does on beliefs, providing a unified account of hallucination and delusions (Fletcher & Frith, 2009). In either case, the final content of the perception or the belief depends on the balance between new evidence and prior expectations. For people with paranoid delusions, the prior belief that certain others are against them is so strong that any evidence to the contrary is ignored or explained away.

The importance of precision in top-down control

The balance between the effects of new evidence and prior belief is determined by their relative reliability or precision. Precision is simply the inverse of variance. Our brain gives more weight to a sensory channel (or representation) with high precision (low variance) than it does to one with low precision (high variance). We see this process of precision weighting at work when the brain combines information from different senses. In one such experiment (Ernst & Banks, 2002), people estimated the width of a bar using both touch and vision. In normal circumstances, visual information dominates the estimation because the visual signal is much more precise than the tactile signal. However, if sufficient noise is added to the visual signal, then touch dominates. There is an intermediate point where vision and touch are combined in a statistically optimum way, based on their precision, to achieve a better estimate than with either sense on its own.

The use of precision to optimise sensory integration happens automatically and is stimulus-driven (bottom-up). However, there are many circumstances in which this automatic weighting of sensory signals by their precision is not compatible with the current goal. For example, our task might be to respond to touch while ignoring vision, requiring suppression of the automatic response to the more salient signal. This requires a *selective* form of attention which is voluntary and depends on behavioural goals.

The (top-down) control signals that modulate the behaviour of sensory systems in the case of selective attention emerge from regions in the frontal and parietal cortex (Yantis, 2008). But what form does this modulation take? One solution is to alter our prior beliefs about estimates of precision to suit our goals. This enables us to behave *as if* the sensory channel we want to attend to has high precision (see Figure 1). In other words, "We weight sensory inputs from different sensory channels in proportion to their precision, given our goals" (Mirza et al., 2019). This is an example of precision control (see, for example, Limanowski, 2022).

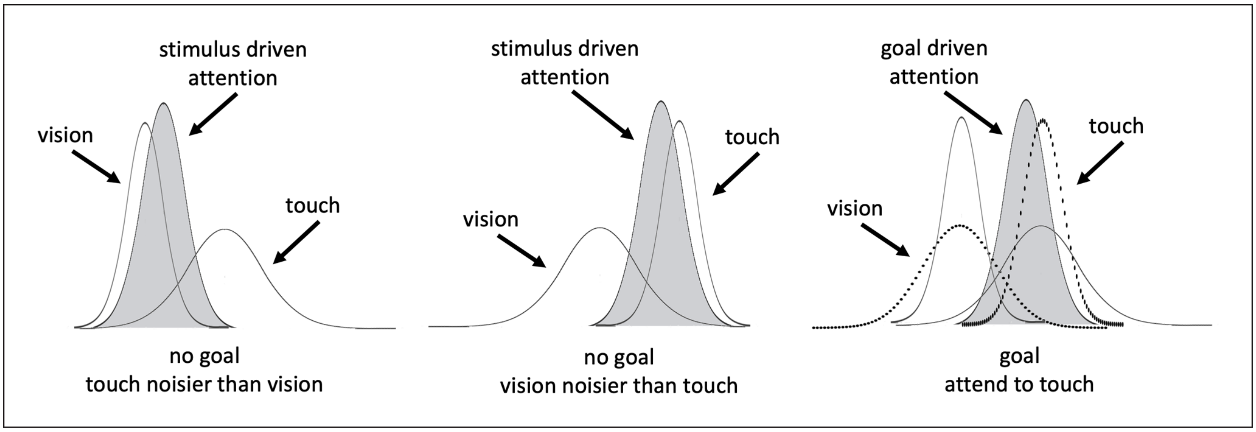
stimulus driven attention
vision
touch
no goal
touch noisier than vision
stimulus driven attention
touch
vision
no goal
vision noisier than touch
goal driven attention
touch
vision
goal
attend to touch

Figure 1. Precision control.
The narrower the distribution, the more precise the signal. Left: Precision of vision is greater than that of touch. So, we automatically attended to vision. Middle: Precision of touch is greater than that of vision. So, we automatically attended to touch. Right: The goal is to attend to touch when it is objectively less precise. So, we behave as if touch was more precise (dotted lines).

From these considerations, we recognise that the balance between evidence and prior expectations has a role in both automatic, unconscious processes (e.g., multisensory integration) and deliberate, conscious processes (e.g., selective attention). However, the symptoms of schizophrenia seem to arise largely from problems occurring with conscious processes. In some cases, in the early stages, there is no problem with behaviour, but only with subjective experience. For example, I know of a patient who had a successful career in accountancy, despite severe psychotic delusions. These delusions remained encapsulated and did not interfere with day-to-day work. Problems with precision control at the automatic level would lead to much greater impairments than those typically associated with schizophrenia. And, where problems with behaviour are observed, they typically arise from problems with selective attention. Patients have difficulty in attributing salience on the basis of their current goals (Kapur, 2003). If we are to understand these symptoms, we first need to understand the nature of these deliberate, conscious control processes.

Consciousness and top-down precision control

My interest in the scientific study of consciousness was sparked by a paper by Tim Shallice (1972) in which he outlined a possible function for consciousness.[2] In this article, he made a distinction between low-level competing action systems (automatic processes) and a high-level selector system which coordinates the operations of the low-level processes. “The selector input selects which action system is to be dominant, sets the goal of the action system, and is itself preserved in memory.” Shallice suggested that it is this selector input which corresponds to the concept of consciousness (see also Norman & Shallice, 1986).

This description of the selector system (or *supervisory attentional system*) has many similarities with the idea of working memory, a concept which continues to play a major role in approaches to the study of consciousness. Some have equated the contents of consciousness with the contents of working memory (e.g., Baddeley, 1992). However, working memory also involves many unconscious processes, leading Baars and Franklin (2003) to equate the contents of consciousness with a global workspace, a key sub-component of working memory. In the global workspace, information provided by low-level systems is broadcast and can be shared (Baars, 1988; Dehaene et al., 1998). This enables the selector system to coordinate the operations of the low-level systems.

But how is this coordination achieved? Here again, I suggest that precision plays a critical role. We need to take account of the precision of the representations of the low-level systems that are broadcast in the global work space (Shea & Frith, 2019). This is because we need to consider relative precision when choosing one low-level action system over another and when we want to combine low-level systems in an optimal manner.

The selector system is an example of a metacognitive process because it applies cognitive operations to low-level cognitive processes. The selector system monitors what is going on at the lower level and, when necessary, modifies and controls what is going on at that level. But this is a special kind of metacognitive process because its operations form an important part of our conscious experience. To capture this feature, I refer to it as *explicit metacognition*.

For the selector system to operate, it needs to receive signals about how the low-level systems are working in the form of metacognitive parameters, such as processing speed, and accuracy. And some estimate of the precision of these parameters is crucial for knowing which should be

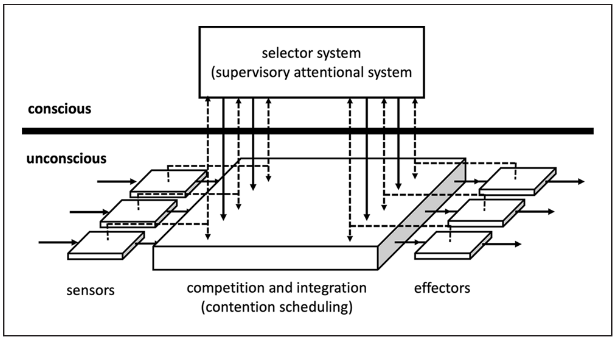

Figure 2. Supervisory attentional system.
The selector system monitors and controls the low-level action systems. Dotted lines—feedback and monitoring, solid lines—control (after Norman & Shallice, 1986).

relied on and to achieve optimum integration. The selector system also needs to send control signals that alter the functioning of the low-level systems (see Figure 2). Here again, estimating the precision of these signals is important for optimising control.

Thus, precision control has an important function at the top as well as at lower levels of control. But, as we move up the cognitive hierarchy, precision becomes increasingly more difficult to estimate (Yon & Frith, 2021). This is for two reasons. First, at the bottom of the hierarchy, we have very accurate estimates of the precision of, for example, a single sensory channel. And, by combining signals about the same event or object from two sensory channels, we can increase the precision still further (Ernst & Banks, 2002). In contrast, at the top of the hierarchy, precision estimates will be based on evidence about many unrelated and more abstract factors. For example, before deciding to perform a task, I would need to combine estimates about the intrinsic difficulty of the task with how tired I am. With combinations from such different sources, the final precision estimate will be decreased rather than increased. As a result, noise in the estimates at these high levels will be greater than in estimates at lower levels. Second, at the bottom of the hierarchy, we have a wealth of past experience with, for example, reaching and grasping, providing sufficient information for a good estimate of precision. In contrast, at the top of the hierarchy, we find processes concerned with much more abstract concepts, relating to complex activities that are less frequently performed. So, there is much less information available for characterising the shape of the distribution needed to estimate precision. Furthermore, the increasing number of different inputs at these higher levels presents a challenge, with ideal statistical solutions that are too complex to compute.

Consider, for example, the problem of being faced with the last biscuit on the plate. At the lower levels of control, I know exactly how hungry I am, and I have computed the optimal way of reaching and grasping the biscuit. But, at the higher levels, I need to have some idea of whether it is socially appropriate to take the last biscuit. Who else is present? Will my reputation suffer? Is there some way I make my action seem amusing? The precision estimates needed to resolve the competition between these different factors will be difficult to compute.

How can we get around this problem? One way is to use quick and dirty approximations that overcome the computational problems. We can rely on such heuristics as they will work most of the time (Gigerenzer & Todd, 1999). Many of these heuristics are acquired socially. We can also take advantage of the experience of others, and of our culture more generally, to acquire more information relevant to the situation. For example, the strong belief that I am a member of a group "that never takes the last biscuit" would lead to the feeling that reaching for last biscuit was wrong and this feeling could override all other factors.

Applying precision control to explicit metacognitive signals

The subjective feelings associated with explicit metacognitive signals tell us something about the workings of the low-level cognitive systems, but these signals, conscious though they are, can be vague. Thomas Metzinger (2006) has described them as *thin and evasive*. As a result, we do not always interpret these signals correctly. But there is also an advantage to this vagueness. We can more easily use top-down control to change the way we interpret these signals.

Here are some examples of how these signals can be misinterpreted. The feeling of fluency is an example of a metacognitive signal that tells us something about the workings of the lower levels of control. This feeling occurs for perception and for action. It signals how quickly and easily on object was perceived or an action selected.

In a classic experiment (Kunst-Wilson & Zajonc, 1980), participants were shown a sequence of irregular polygons which they had to try to remember. They were then shown a polygon they had seen before, paired with a novel one. When asked which one they had *seen before*, they performed no better than chance. But when asked which one they *preferred*, they reliably chose the one they had seen before. Subsequent work has shown that this effect is due to perceptual fluency. An item is perceived more fluently when it has been seen before, and the more fluently an item is perceived, the more it is liked (Reber et al., 1998). The experience of perceptual fluency is used as a marker of liking, but not as a marker of familiarity. Fluency is a useful marker for liking because we are likely to have looked more frequently at things which we like. But this heuristic may be misleading. For example, it may prevent us from using the feeling as a marker of familiarity.

Action selection fluency is another metacognitive signal that can be misinterpreted. As with perceptual fluency, action selection fluency is greater for actions we have performed before. We use it as a marker of how much we are in control of what we are doing. Here again, this is reasonable because it is practice with actions that gives us more accuracy and fluency. When we are in control of our actions, we are less likely to make mistakes. But this feeling of fluency can be misleading. By priming actions, we can make them feel more fluent. And this feeling makes us believe we are more in control of actions, even when the priming has made our performance worse (Chambon & Haggard, 2012).

The feeling of confidence is another important explicit metacognitive signal. We use it as a marker of the accuracy of our decisions. If I feel confident in my decision, then I expect that it is correct. If someone else expresses confidence in their decision, then I expect that they made the right choice. But this feeling of confidence can be manipulated to gain advantages.

For example, expressions of confidence have an important role in coordinating decision making when people are working together. When two people make decisions together, a useful heuristic is to follow the choice of the more confident person on a trial-by-trial basis (Bahrami et al., 2010; Koriat, 2012). This strategy enables the joint decisions to be better than the better person working alone.

The strategy assumes that the choice of the more confident person is more likely to be right.

However, this will not work if one of the partners is, for example, persistently over-confident, because his confidently expressed wrong decisions would be given too much weight. We can easily be misled by the confidence expressed by others.

To overcome this problem, when people work together, they need to align their expressions of confidence. On their own, two people may express their confidence (on a 6-point scale) systematically high (4–7) or systematically low (2–4), but when these same two people need to work together, their reports of confidence become aligned (Bang et al., 2017). When people express their confidence to each other in words, they spontaneously develop a verbal scale for expressing levels of confidence (e.g., *sure, almost sure, not sure, very unsure*). The more rapidly they can agree on the appropriate vocabulary for such scales, the greater the advantage of working together (Fusaroli et al., 2012). The development of scales like this, on the fly, helps to achieve alignment of expressions of confidence.

These results show that there is a difference between private and public representations of confidence. The confidence I feel need not be the same as the confidence that I express. This distinction between private and public representations is also observed in brain activity. Representations of the private feeling of confidence are observed in medial prefrontal cortex, while representations of public expressions of confidence are observed in lateral frontal pole (Bang et al., 2020), a brain region with no analogue on the monkey brain (Semendeferi et al., 2001).

The ability to express our confidence decoupled from our private feeling is an advantage in cooperative situations because we can align our expression of confidence and make better group decisions. Our expressions of confidence can also be manipulated to gain an advantage in competitive situations. When two advisors are competing for the attention of a client, it pays the advisor currently being ignored to express higher levels of confidence (Hertz et al., 2017). If the advice turns out to be wrong, there is nothing to lose, but, if it is right, he may gain the attention of the client.

We can think of confidence in terms of precision. When we have very low confidence, we feel that we are just guessing. The precision of our estimate of the answer is very low. When we are confident, our estimate of the precision is high. We can use top-down precision control to adjust our expressions of confidence to suit the social situation.

So, at least in the case of confidence, there may be advantages to the thin and evasive nature of this metacognitive signal. The expression and perhaps even the feeling of confidence can be manipulated depending on the social context. Cecilia Heyes and colleagues (2020) suggest that, precisely because they can be discussed with others, we learn from others how to experience, interpret, and express these explicit metacognitive signals.

Applying precision control to prediction errors

The most basic message that can be given by a metacognitive signal, whether conscious or not, is that there is something wrong. This signal typically takes the form of a prediction error. The world is not how we thought it was and we need a different strategy. Here again, precision plays an important role. If the prediction error has a low precision, it should be ignored. But, in some circumstances, even precise prediction errors can be uninformative and should be ignored.

Consider the simple game of hide-and-seek. Our opponent can hide behind the tree or behind the wall. Our opponent might be a very simple agent that behaves randomly, but with a bias. Such an agent might hide behind the tree 80% of the time and we will learn to expect to find it behind the tree. But, occasional, when we look behind the tree, it is not there. This will generate a prediction error. In response to this error, a simple model-free learning device will down-weight the value of looking behind the tree even though such prediction errors are not informative. The opponent is not more likely to be behind the wall on the next trial. Such prediction errors should be treated as noise, but for this to happen, the learner has to believe that the errors are noise.

Such beliefs emerge in reversal-learning tasks. Here, after a variable number of trials, the opponent changes its bias. Now it hides behind the wall 80% of time. After experiencing a number of such reversals, a more sophisticated model-based learner will realise that the opponent can be in two states: mostly hiding behind the wall or mostly hiding behind the tree. While the opponent remains in one of these states, the prediction errors should be treated as noise and ignored. But when a change of state occurs, the prediction errors become important because they indicate the change of state (Soltani & Izquierdo, 2019).

Precision control enables us to take advantage of this model of our opponent. When we believe the world is in a stable state, we can down-weight the precision of the prediction errors and upweight the precision of the belief that he will hide behind the wall. This will maintain the belief that he is behind the wall in spite of the prediction errors. But, if we believe that state is about to change (perhaps it is a long time since the last reversal), we will upweight the precision of the prediction errors and down-weight the precision of the belief.

High-level beliefs are sticky

This analysis of the hide-and-seek game reveals a hierarchy of beliefs (or priors) that are relevant for playing the game. At the bottom level, we have beliefs about the values of the two actions, i.e., beliefs about what I should best do next (*look behind wall* or *look behind tree*). This level of learning takes no account of the opponent, but only considers the value of my actions. At a higher level, we have beliefs about the state of our opponent. These are beliefs about what he will do next (*hide behind wall or hide behind tree*). These beliefs determine the values for my two actions. But there is a higher level still. Here, we have beliefs about the nature of our opponent (*simple, sophisticated*). This belief determines how I expect him to change in response to my choices (see, for example, Hampton et al., 2008). This enables me to determine what he will do next and hence determines the values of my actions.

Jean Daunizeau and his colleagues (Devaine et al., 2014) created a series of artificial agents that could play hide-and-seek at the different levels of this hierarchy. At the lowest level, these agents behaved randomly, but with a bias. At the highest level, the agents computed what their opponent thought they were going to do next. People played against these agents in two contexts. In one context, they believed that they were playing against a one-armed bandit. In the other context, they believed that they were playing against a person. This framing created a belief about the sophistication of their opponent. This high-level belief determined the strategies that the players used.

When they were told they were playing against another human, players typically adopted a high-level strategy assuming that their opponent was trying to discover their strategy. Using this strategy, they were able to beat the low-level agents and break even with the most sophisticated agents. When, however, they believed that they were playing against a robot, they typically employed a simple win-stay, lose-shift strategy. This is a strategy with minimal cognitive load. All you have to take into account is what happened on the last trial, did you win or lose? This is equivalent to model-free learning with a very high learning rate. This strategy works well against an agent with a random bias and even against a simple model-free learning agent as long as the learning rate is relatively slow. But it fails against more sophisticated agents. As a result, people lost when they thought they were playing against robots which were actually sophisticated agents. The actual behaviour of the agent was not enough to alter their belief that robots cannot be sophisticated.

There are several studies suggesting that, at this high level of learning, the incoming evidence is down-weighted relative to the prior belief about the situation. For example, people activate the medial prefrontal cortex (part of the brain's mentalising system) when they think they are competing with a person rather than a computer, even though the behaviour of the opponent is the same (Gallagher et al., 2002; Rilling et al., 2004). An even more compelling result comes from Chris Miall's lab. Point-light animations were developed in which a dot moved up and down tracing an arm movement. In one case, the movements were created by a person and were roughly sinusoidal. In the other case, the movements were square wave, resembling the jerky movements made by a primitive robot. When people observe the movements of another person, this will interfere with their own movements (Kilner et al., 2003), and the same effect occurred when watching the moving dots. However, the interference was related to the belief about who was making the movements, rather than the form of the movements (Stanley et al., 2007). Jerky movements would cause interference, if the observer believed that they were being made by a human. In this paradigm, activation of the brain's mentalising system was also determined by the belief about the source of the movements, rather than the form of the movement (Stanley et al., 2010).

High-level beliefs are social

The delusions associated with schizophrenia are typically very sticky. For example, according to the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM*-5), these delusions "are not amenable to change in light of conflicting evidence" (American Psychiatric Association, 2013). This is what we would expect, given that these beliefs lie at the top of the processing hierarchy. But there is another striking feature of these delusions: They are typically intensely social in nature. It is other people who are hearing my thoughts (*thought broadcasting*), controlling my actions (*delusions of control*), or working in league against me (*paranoia*).

But why should beliefs at the top of the hierarchy be so intensely social? I believe the reason concerns the nature of the top. After all, what is at the top of the hierarchy? Where does top-down control come from when we reach this highest level? The answer turns out to be straightforward as well as surprising. At this level, top-down control comes from *other people* (Roepstorff & Frith, 2004). As participant in experiments, we attend to a less salient stimulus because the experimenter tells us to. We believe we are playing against a person, rather than a computer, because the experimenter tells us so. There can also be reciprocal effects on the experimenter. For instance, participants can refuse to obey the instruction or ask for further explanations. It is typically mutual interactions with other people that change the priors at the top of the hierarchy and exert precision control.

Consider the task of writing a paper for a journal.[3] We are intending to write the phrase, "To boldly go where no other species has been." At the bottom of the hierarchy of control needed to achieve this task—the sub-personal level—we find the largely automatic processes of pressing keys and monitoring the letters that appear on the screen. If there is a key press error (*bodly* for *boldly)*, we briefly slow down. At a higher level—the personal level—we need a bit more awareness. We notice that we have written "go wear" instead of "go where." We advise ourselves to concentrate harder. But there is an even higher, extra-personal level. We think our sentence is fine, but the editor disagrees. Split infinitives are not allowed. It should read "to go boldly." We must change our priors and conform with the house style.

Cultural transmission and cultural priors

Learning to conform to the house style is an example of cultural transmission. We have learned the *right way* to express ourselves. We have internalised the cultural prior that one does not split infinitives. But how does this cultural transmission work? There are two critical processes. First, a private cognitive representation must be converted into a public form that can be broadcast to others. Second, there must be an equivalent reverse process through which a public broadcast can alter private cognitive processes (Sperber, 1996).

I have already discussed an example of the first of these processes. This process is involved when we report our confidence in a decision. A private message emerges from the depths of cognitive activity suggesting how likely we are to have made the right decision and we can report this feeling of confidence to others (e.g., Fusaroli et al., 2012). This report has an important role in decision making, particularly if we are working with others. If our confidence is low, we may pause to collect more information before acting. But, maybe, we remain overconfident and continue to respond too quickly. Our partner can change such behaviour by the command, "Don't be so impulsive." Here, we have an example of the second process underlying cultural transmission. This is supra-personal control (Shea et al., 2014). Messages from others can alter the way our private cognitive processes work. These interactions at the top of the hierarchy of control create and maintain cultural priors.

One of Bartlett's most well-known studies used the "method of serial reproduction." A participant was asked to draw from memory an obscure figure (e.g., the Egyptian hieroglyph *mulak OWL*). A second participant drew this reproduction from memory and so on (Bartlett, 1932, p. 180). As the series continued, the drawings became less and less like the original. The 10th person in the series drew a cat. Bartlett (p. 178) explained the process as follows:

> . . . whenever material visually presented purports to be representative of some common object, but contains certain features which are unfamiliar in the community to which the material is introduced, these features invariably suffer transformation in the direction of the familiar.

Using Bayesian terminology, I would rephrase this idea as follows: When reproducing an image from memory, the participant combines the evidence from memory of what has just been seen with a prior expectation of what the image might be. Each time the reproduction will move away from the original and towards the prior (*the familiar*). At the end of this process, we will arrive at the cultural prior (a cat) shared by all the participants. In a footnote (p. 181), Bartlett comments: "It may be interesting to note that in another series from the same starting point this design has again developed into a cat by the time the 17th reproduction is reached."

Precision control from outside the brain

But how do instructions have their effect? My suggestion is that instructions, and culture, work, usually via language, to apply precision control at the top level of the hierarchy. This idea is supported by studies of trust games (Berg et al., 1995). In these games, one player transfers money to a partner in the hope that this person will return the money with interest. To succeed in this game, you need to learn to distinguish between those who can be trusted and those who will just take your money. In most studies, this learning occurs slowly during direct interactions. However, you can also learn very quickly whom to trust through information from the experimenter or from gossiping with other players. Such instructions change your behaviour. You invest in those you are told are trustworthy and pay less attention to their actual behaviour (Sommerfeld et al., 2007).

The same process can be observed in the brain. When you learn about people by direct interaction, prediction errors can be detected in the striatum. In other words, if the person gives back more money than you expect (positive prediction error), there is a brief increase in activity, and you increase your prior belief as to the trustworthiness of this person (King-Casas et al., 2005). The opposite happens if you get back less money than you expect (negative prediction error). Here, we see the interplay between prior expectations (trustworthiness) and evidence (what the person actually does).

In the early stages of learning, the precision of the prior belief is much lower than the precision of the evidence, as we don't yet know how trustworthy the people are. We must attend closely to their behaviour. This pattern changes if we are told, in advance, how trustworthy the various people are. Now, the precision of our prior belief is high, higher than the precision of the evidence that we collect on each trial. In other words, we no longer need to attend closely to the behaviour of our partners because we know precisely how trustworthy they are. Still, the responses of our partner can vary from trial to trial, and they do not always return the money. This results in a prediction error. However, we treat it as irrelevant noise. Remarkably, such prediction errors no longer elicit increased activity in the striatum (Delgado et al., 2005; Fouragnan et al., 2013). This is an example of precision control. The instruction about trustworthiness has altered the balance between prior belief and evidence by changing their relative precision.[4]

Earlier in this essay, I emphasised that top-level priors can be sticky and resist modulation. But this stickiness only relates to bottom-up effects of evidence. We now see that top-level priors can be very quickly changed by top-down messages from other people. There are good reasons for this asymmetry. Top-level priors concern complex, abstract concepts, such as trustworthiness. Evidence for such concepts is difficult to collect and needs much experience. It takes a long time to learn such things directly by trial and error (Yon & Frith, 2021). We can get more precise priors from other people who have had more experience. We can get even better estimates from our cultural milieu because this encompasses the experience of many people over a long time. As a result, outside influences can come to dominate over direct experience.

Free will as a cultural prior

In this final section, I will speculate on how one particularly contentious cultural prior, the idea of free will, gets into the brain and changes our behaviour. Free will is not so much an idea, as an experience. At least as far back as Epicurus (~300 BC), this experience is considered to have two components: the feeling of being in control of my actions and the feeling that I could have done otherwise. Taking together, these feelings create a sense of causal agency; it is me causing this outcome, and, if I had done something else, this outcome would not have occurred. The belief that I could have done something different is a fundamental component of free will and can lead to intense feelings of regret (Frith & Metzinger, 2016).

Epicurus believed that our sense of agency interacts with our culture to create a feeling of responsibility (Bobzien, 2006). In his time, as in most societies today, children were rewarded or punished for their actions from quite a young age. As they grow up embedded in a culture where it is assumed that they are in control of their actions and can choose to do otherwise, children learn that they will be held responsible for their actions. They rapidly learn that the excuse "it was an accident" can reduce punishment. As the result of such upbringing, we accept responsibility for our actions, and we link this commitment to the vivid experience of being in control of our actions, i.e., having free will. Most people believe that there is a strong link between moral responsibility and free will (Nahmias et al., 2005).

Our culture also induces us to believe that acting responsibly is hard work. At least as far back as Plato, many philosophers have proposed that selfishness is a basic human urge that constantly needs to be overcome through reason. So, constant self-control is required. This account of how to maintain responsible action is an example of explicit (conscious) metacognition. We must constantly monitor our actions and, if they seem aimed at selfish outcomes, we should exert top-down control to suppress them. This is free will in action. Furthermore, this characterisation suggests that free will is important. Without it, we would all give in to our selfish urges with disastrous consequence for social cohesion (Smilansky, 2002). Is this a testable hypothesis? If people could be persuaded that free will is an illusion, would their behaviour change?

A series of experiments tested this idea by presenting people with statements such as "most rational people now recognise that free will is an illusion" citing Francis Crick as a leading example of a rational person (Crick, 1994). Bad effects on behaviour were observed in these studies. Participants who were persuaded that free will was an illusion became less prosocial. They showed increased aggression and reduced helping behaviour (Baumeister et al., 2009). They were also more likely to cheat in exams (Vohs & Schooler, 2008).

I found the results of these studies rather depressing. They suggest that the belief in free will is indeed a good thing, but also that this belief can all too easily be changed. My depression was lifted, however, when I learned that more recent studies have not been able to replicate any of these results (Genschow et al., 2023). Perhaps this high-level prior is not so easy to change. But how does this observation fit with the idea that high-level priors can

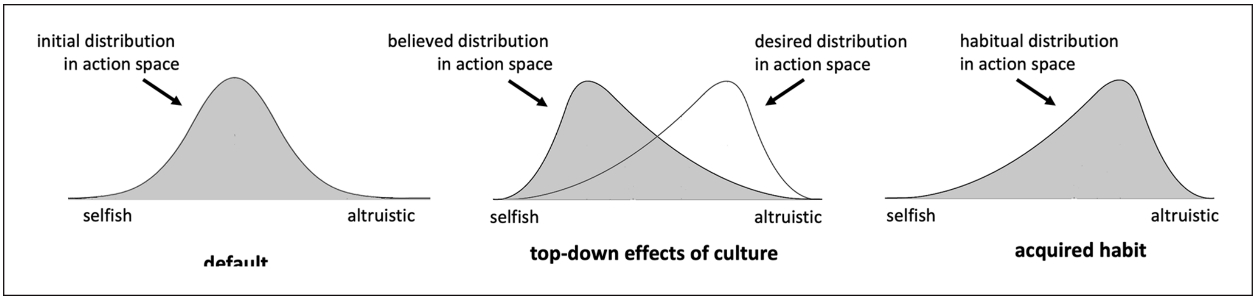
initial distribution in action space
selfish
altruistic
default
believed distribution in action space
desired distribution in action space
selfish
altruistic
top-down effects of culture
habitual distribution in action space
selfish
altruistic
acquired habit

Figure 3. Culture creates habits.
One dimension of a potential action space varies from selfish to altruistic. On the left, we see the default action space. In the middle, our culture determines the believed action space (biased towards selfish actions) and the desired action space (biased towards altruistic actions). On the right, top-down control processes have created habits that implement the desired action space.

readily be changed by information from external sources? One possibility is that, although the prior can be changed, this does not necessarily lead to a change in behaviour or may take time to do so.

One problem with this account of free will in action is that it would be cognitively very taxing. You would need to constantly monitor your behaviour to avoid any selfish urges. But this problem is not unique to free will. It applies to any task which requires us to monitor and control low-level processes. Fortunately, there is an important feature of brain function designed to deal with this problem; with practice, cognitive processes cease to be controlled and become automatic (Shiffrin & Dumais, 1981). In the brain this change is accompanied by a reduction of activity in frontal cortex presumably because monitoring and control is no longer needed (Jenkins et al., 1994). The need for high-level control is reduced because habits of action have developed which are consistent with the high-level prior. As a result, in the specific case of the responsible behaviour associated with free will, selfish behaviour has ceased to be the default behaviour and altruistic behaviour has become habitual.[5] This combination of a high-level prior and a low-level habit is difficult to dislodge.

This account of responsible behaviour emerging from the combination of a high-level prior and a low-level habit has parallels with a philosophical approach to the problem of justice. Our commitment to justice and associated legal systems also enhances social cohesion, in this case through the resolution of conflicts. But we don't constantly think about the need to resolve conflicts. We are committed to justice for its own sake (Rawls, 1971/1999, p. 416).

This high-level prior about justice has been fixed in most of us from earliest childhood. At a lower level, the behaviour that is consistent with this high-level prior has become automatic and engrained. Behaving according to the tenets of justice has become habitual for us. Our commitment to justice would not be so robust if we always had to think in detail about the psychological origins of this commitment (Rawls, 1971/1999, p. 451). Here again, this combination of a high-level prior and a low-level habit are difficult to dislodge. This "intuitive and inflexible" behaviour (Bernhard & Cushman, 2022) also gives a signal of our moral status to others (Critcher et al., 2012). People who make moral decisions quickly are perceived to have higher moral standards.

Figure 3 presents a simplified illustration of how culture might install norms of responsible behaviour. It shows a mental quality space for action, analogous to a perceptual quality space (Lau et al., 2022), in which one dimension runs from selfish to altruistic actions. In the initial, default state, most actions are neutral regarding selfishness. But, in our culture, we have come to believe that an essential selfishness skews our action space away from altruistic behaviour (see, for example, Hobbes, 1651, chapter XIII, of the Natural Condition of Mankind as Concerning Their Felicity and Misery). In consequence, we believe that top-down control is needed to create the desired action space in which altruistic actions are more likely. Continued application of top-down control generates habits of action creating an action space in which altruistic acts are indeed more likely.

Conclusion: explicit metacognition and culture

Explicit metacognition lies at the top of a hierarchy of control. Messages from below indicate how our cognitive system is working. But this evidence is "thin and evasive," and so their interpretation is largely determined by the culture in which we are imbedded (Heyes et al., 2020). In addition, the attributes of the priors at the top of the hierarchy are difficult to estimate from our own direct experience (Yon & Frith, 2021). Hence, it is our interactions with other people which have the greatest effect on the priors at the top of the hierarchy.

In the case of schizophrenia, this link between metacognition and culture seems to be broken. High-level priors are no longer properly constrained by the beliefs of other people and by culture more generally. As a result, beliefs are particularly difficult to change and will increasingly diverge from consensus. In time, habits of thought will develop which are difficult to dislodge, even if the priors at the top level are "normalised."

Near the end of "Remembering," Bartlett points to one of the many difficulties that need to be overcome if we are to develop a truly social psychology:

> Yet, however much agreement there may be as to the fact of social constructiveness, we know almost nothing of its exact mechanism . . . These all constitute important sociological and psychological problems which will provide a great field for future research. (Bartlett, 1932, p. 280)

In this essay, I have made some speculations about how "social constructiveness" might work. I suggest that social constructs impinge on the brain via the processes of explicit metacognition that lie at the top of our brain's hierarchy of control. Our prior expectations at this level of control are malleable and largely determined by our culture. However, we know almost nothing about the physiological mechanisms that enable these high-level priors to influence low-level cognitive processes and, hence, our behaviour. Exploring these mechanisms will not only increase our understanding of how culture gets into the brain (Frith & Frith, 2022) but may also throw some light on the nature of schizophrenia.

Acknowledgements

I thank the Experimental Psychology Society for commissioning this lecture. I am grateful to Andreas Roepstorff for directing me to Alan Costall's work on Bartlett. Preparation and revision of the paper gained much from Nick Shea and friends at the Institute of Philosophy and Steve Fleming and friends in the Consciousness Club at UCL. This work was presented at 49th Bartlett lecture (on 6 January 2022).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Chris D Frith https://orcid.org/0000-0002-8665-0690

Notes

[1] He doesn't have a "k".

[2] Note that this article was published in the time prior to Crick and Koch (1992) when, supposedly, no one was studying consciousness scientifically.

[3] In the days before MS-Word™ did all these levels of monitoring for you.

[4] A similar effect is associated with instructed fear conditioning (Lindström et al., 2019).

[5] This idea is prefigured in the writings of Aristotle. In the Nicomachean Ethics, he suggested that we acquire a virtuous character by habituation of the passions.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Baars, B. J., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences*, *7*, 166–172.

Baddeley, A. (1992). Consciousness and working memory. *Consciousness and Cognition*, *1*, 3–6.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*, 1081–1085.

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*, Article 0117.

Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private–public mappings in human prefrontal cortex. *eLife*, *9*, Article e56477.

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge University Press.

Baumeister, R. F., Masicampo, E. J., & Dewall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin*, *35*, 260–268.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142.

Bernhard, R. M., & Cushman, F. (2022). Extortion, intuition, and the dark side of reciprocity. *Cognition*, *228*, Article 105215.

Berrios, G. E. (1985). Positive and negative symptoms and Jackson: A conceptual history. *Archives of General Psychiatry*, *42*, 95–97.

Bobzien, S. (2006). Moral responsibility and moral development in Epicurus's philosophy. In B. Reis (Ed.), *The virtuous life in Greek ethics* (pp. 206–299). Cambridge University Press.

Chambon, V., & Haggard, P. (2012). Sense of control depends on fluency of action selection, not motor performance. *Cognition*, *125*, 441–451.

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences*, *23*, 114–127.

Costall, A. (1992). Why British psychology is not social: Frederic Bartlett's promotion of the new academic discipline. *Canadian Psychology/Psychologie Canadienne*, *33*, 633–639.

Costall, A. (2009). Frederic Bartlett and the idea of an historical psychology. *Ethnographic Studies*, *11*, 24–38.

Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. Scribner.

Crick, F., & Koch, C. (1992). The problem of consciousness. *Scientific American*, *267*, 152–159.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2012). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*, 308–315.

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 14529–14534.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*, 1611–1618.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLOS Computational Biology*, *10*, Article e1003992.

Dijkstra, N., & Fleming, S. M. (2021). Fundamental constraints on distinguishing reality from imagination. *PsyArXiv*. https://doi.org/10.31234/osf.io/bw872

Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M. A. J., & Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, *7*, Article 5677.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*, 48–58.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience*, *33*, 3602–3611.

Frith, C. D. (1979). Consciousness, information processing and schizophrenia. *British Journal of Psychiatry*, *134*, 225–235.

Frith, C. D. (2000). The role of dorsolateral prefrontal cortex in the selection of action as revealed by functional imaging. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes* (pp. 549–565). MIT Press.

Frith, C. D., & Frith, U. (2022). The mystery of the brain-culture interface. *Trends in Cognitive Sciences*, *26*, 1023–1025.

Frith, C. D., & Metzinger, T. (2016). What's the use of consciousness? In A. K. Engel, K. Friston & D. Kragic (Eds.), *Where's the action? The pragmatic turn in cognitive science* (pp. 193–214). MIT Press.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylen, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science*, *23*, 931–939.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, *16*, 814–821.

Genschow, O., Cracco, E., Schneider, J., Protzko, J., Wisniewski, D., Brass, M., & Schooler, J. (2023). Manipulating belief in free will and its downstream consequences: A meta-analysis. *Personality and Social Psychology Review*, *27*, 52–82.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Haddad, P. M., & Correll, C. U. (2018). The acute efficacy of antipsychotics in schizophrenia: A review of recent meta-analyses. *Therapeutic Advances in Psychopharmacology*, *8*, 303–318.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6741–6746.

Helmholtz, H.v. (1867/1948). Concerning the perceptions in general. In W. Dennis (Ed.), *Readings in the history of psychology* (pp. 214–230). New York: Appleton-Century-Crofts.

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, *8*, Article 2191.

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in Cognitive Sciences*, *24*, 349–362.

Hobbes, T. (1651). *Leviathan, or, the matter, form, and power of a common-wealth*. Andrew Crooke.

Huxley, A. (1959). *The doors of perception & heaven and hell*. Penguin Books.

Jenkins, I. H., Brooks, D. J., Nixon, P. D., Frackowiak, R. S., & Passingham, R. E. (1994). Motor sequence learning: A study with positron emission tomography. *The Journal of Neuroscience*, *14*, 3775–3790.

Johnstone, E. C., Crow, T. J., Frith, C. D., Carney, M. W., & Price, J. S. (1978). Mechanism of the antipsychotic effect in the treatment of acute schizophrenia. *Lancet*, *1*, 848–851.

Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry*, *160*, 13–23.

Kilner, J. M., Paulignan, Y., & Blakemore, S. J. (2003). An interference effect of observed biological movement on action. *Current Biology*, *13*, 522–525.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*, 78–83.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719.

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*, 360–362.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*, 557–558.

Lau, H., Michel, M., LeDoux, J. E., & Fleming, S. M. (2022). The mnemonic basis of subjective experience. *Nature Reviews Psychology*, *1*, 479–488.

Leff, J. P. (1982). Acute syndromes of schizophrenia. In J. K. Wing & L. Wing (Eds.), *Handbook of psychiatry: Psychoses of uncertain aetiology* (pp. 8–16). Cambridge University Press.

Lhermitte, F. (1983). "Utilization behaviour" and its relation to lesions of the frontal lobes. *Brain*, *106*, 237–255.

Limanowski, J. (2022). Precision control for a flexible body representation. *Neuroscience & Biobehavioral Reviews*, *134*, Article 104401.

Lindström, B., Golkar, A., Jangard, S., Tobler, P. N., & Olsson, A. (2019). Social threat learning transfers to decision making in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 4732–4737.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–48.

Mellor, C. S. (1970). First rank symptoms of schizophrenia. *British Journal of Psychiatry*, *117*, 15–23.

Metzinger, T. (2006). Conscious volition and mental representation: Toward a more fine-grained analysis. In N. Sebanz & W. Prinz (Eds.), *Disorders of volition* (pp. 19–48). Bradford Books, MIT Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Mirza, M. B., Adams, R. A., Friston, K., & Parr, T. (2019). Introducing a Bayesian model of selective attention based on active inference. *Scientific Reports*, *9*, Article 13915.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*, 561–584.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self regulation: Advances in research* (pp. 1–18). Plenum Press.

Rawls, J. (1999). *A theory of justice* (Revised ed.). Harvard University Press. (Original work published 1971)

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*, 45–48.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, *22*, 1694–1703.

Roepstorff, A., & Frith, C. (2004). What's at the top in the top-down control of action? Script-sharing and "top-top" control of action in cognitive experiments. *Psychological Research*, *68*, 189–198.

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, *372*, Article eabf4740.

Schneider, K. (1959). *Clinical psychopathology*. Grune & Stratton.

Seeman, P., & Lee, T. (1975). Antipsychotic drugs: Direct correlation between clinical potency and presynaptic action on dopamine neurons. *Science*, *188*, 1217–1219.

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., & Van Hoesen, G. W. (2001). Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Biological Anthropology*, *114*, 224–241.

Shallice, T. (1972). Dual functions of consciousness. *Psychological Review*, *79*, 383–393.

Shea, N. J., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*, 186–193.

Shea, N. J., & Frith, C. D. (2019). The global workspace needs metacognition. *Trends in Cognitive Sciences*, *23*, P560–P571.

Shiffrin, R. M., & Dumais, S. T. (1981). The development of automatism. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 111–140). Lawrence Erlbaum.

Smilansky, S. (2002). Free will, fundamental dualism, and the centrality of illusion. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 249–505). Oxford University Press.

Soares-Weiser, K., Maayan, N., Bergman, H., Davenport, C., Kirkham, A. J., Grabowski, S., & Adams, C. E. (2015). First rank symptoms for schizophrenia. *Cochrane Database of Systematic Reviews, 2015*, Article CD010653.

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*, 635–644.

Sommerfeld, R. D., Krambeck, H. J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 17435–17440.

Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Wiley-Blackwell.

Stanley, J., Gowen, E., & Miall, R. C. (2007). Effects of agency on movement interference during observation of a moving dot stimulus. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 915–926.

Stanley, J., Gowen, E., & Miall, R. C. (2010). How instructions modify perception: An fMRI study investigating brain areas involved in attributing human agency. *NeuroImage*, *52*, 389–400.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*, 49–54.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*, 829–853.

Yantis, S. (2008). The neural basis of selective attention: Cortical sources and targets of attentional modulation. *Current Directions in Psychological Science*, *17*, 86–90.

Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology*, *31*, R1026–R1032.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain*, *113*(Pt. 6), 1721–1777.