FEATURE

The good, the bad and the intentional

Dan Jones on the often surprising part played by moral judgements in our 'folk psychology'

In 2003, when George Bush and Tony Blair inaugurated the ongoing war in Iraq, both men surely knew that civilian deaths would be one of the costs of the military engagement. But did Bush or Blair *intentionally cause* these deaths?

How you answer this question is likely to turn on your moral stance towards the war – whether you see it as an immoral violation of international law, or a liberating intervention in an oppressive regime. That's the message emerging from research in the burgeoning field of experimental philosophy, which applies empirical methods to age-old questions such as how the beliefs, desires and intentions behind certain actions affect how others view these actions.

A cornerstone question for the field is how the capacity for assessing whether an action was morally permissible relates to the capacity for making other, non-moral, judgements, such as who did what to whom and when, and whether someone did something intentionally. 'The standard view was that there was a one-way relationship between the two domains,' says Joshua Knobe, associate professor of philosophy at Princeton University. 'On this view, we answer these non-moral questions first, and then work out whether a particular action was morally good or bad, praiseworthy or blameworthy.'

Rather than consulting his own philosophical intuitions, Knobe set out to find out how ordinary people think about intentional action. In a study published in 2003, Knobe presented passers-by in a Manhattan park with the following scenario. The CEO of a company is sitting in his office when his Vice President of R&D comes in and says, 'We are thinking of starting a new programme. It will help us increase profits, but it will also harm the environment.' The CEO responds that he doesn't care about harming the environment and just wants to make as much profit as possible. The programme is carried out, profits are made and the environment is harmed.

Did the CEO intentionally harm the environment? The vast majority of people

Did they intentionally cause civilian deaths?

Knobe quizzed – 82 per cent – said he did. But what if the scenario is changed such that the word 'harm' is replaced with 'help'? In this case the CEO doesn't care about helping the environment, and still just wants to make a profit – and his actions result in both outcomes. Now faced with the question 'Did the CEO intentionally help the environment?', just 23 per cent of Knobe's participants said 'yes' (Knobe, 2003a).

This asymmetry in responses between the 'harm' and 'help' scenarios, now known as the Knobe effect, provides a direct challenge to the idea of a one-way flow of judgements from the factual or non-moral domain to the moral sphere. 'These data show that the process is actually much more complex,' argues Knobe. Instead, the moral character of an action's consequences also seems to influence how non-moral aspects of the action – in this case, whether someone did something intentionally or not – are judged.

The Knobe effect echoes findings from previous work on how people think about intentions, causation and blame. Back in 1992 Mark Alicke published his findings that people were more likely to blame a speeding driver for causing an accident when he was rushing home to hide a cocaine vial from his parents than when he wanted to hide an anniversary present from them.

Knobe's results also build on earlier work on the everyday or 'folk' concept of intentional action. When asked what actions count as intentional, most people mention various mental states of those performing them. But mental states aren't all that count. A person who desires, intends or even believes that they will win the lottery, and whose numbers actually come up in the draw, is not normally judged to have won intentionally, as they didn't have any causal role in producing the outcome, however much they wanted it. Skill in achieving a desired end, together with a causal relationship with the outcome, also matters (Malle & Knobe, 1997).

Imagine Jake, an aspiring marksman, participating in a shooting competition in which he has to hit a bullseye to win. Jake is not, however a very good shot. Yet

references

Adams, F. & Steadman, A. (2004a). Intentional action and moral considerations. *Analysis, 64*, 268–276.

Adams, F. & Steadman, A. (2004b). Intentional action in ordinary language. *Analysis, 64*, 173–181.

Alicke, M.D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368–378.

Cushman, F. (2008). Crime and punishment. *Cognition 108*, 353–380.

Cushman, F., Knobe, J. & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108*, 281–289.

Kliemann, D., Young, L., Scholz, J. & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologica, 46*, 2949–2957.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–194.

Knobe, J. (2003b). Intentional action in folk psychology. *Philosophical Psychology, 16*(2), 309–324.

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis, 64*, 181–187.

Knobe, J. (2006). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences, 9*, 357–359.

Malle, B.F. & Knobe, J. (1997). The folk concept of intentional action. *Journal of Experimental Social Psychology, 72*, 288–304.

Saxe, R. & Kanwisher, N. (2003). People thinking about people. *NeuroImage, 19*, 1835–1842.

Saxe, R., Xiao, D-K., Kovacs, G. et al. (2004). A region of right posterior temporal sulcus responds to observed intentional actions. *Neuropsychologica, 42*, 1435–1446.

he raises his rifle, pulls the trigger…and gets a bullseye. Did Jake hit it intentionally? In Knobe's study most people (72 per cent) said 'no', even though he was striving for this result; there's too much luck involved for it to be genuinely intentional (Knobe, 2003b). But now consider an analogous scenario. Jed plans to shoot his aunt and cash in early on his inheritance, but he's a terrible shot and is likely to miss. In the event, he lines up his sights – badly – but jerks at the moment of pulling the trigger, which corrects his shot, resulting in one dead aunt. Did Jed intentionally kill her? As you probably expect, most people studied claim that Jed did, indeed, intentionally murder his unwitting relative (Knobe, 2003b).

From theory of mind to morality, and back again

Our skill in mental simulation, or theory of mind (ToM), allows us to represent and reason about the beliefs, desires and intentions of others, and the link with morality is obvious: when we judge that someone desired, intended and was causally responsible for some particular outcome, we naturally hold that person more morally responsible for that outcome than if it was accidental. What is intriguing about the Knobe effect is that it suggests that the arrow also sometimes points the other way (Knobe, 2006). When thinking about the CEO's actions, the goodness or badness of the consequences determine for the most part whether the harm or help to the environment is deemed intentional. 'ToM is suffused with moral considerations at a very fundamental level,' says Knobe.

One possible explanation for the Knobe effect is that the negative emotional reaction generated by immoral actions seeps into ToM, and skews its judgements. This explanation, however, has been ruled out by the fact that people with damage to the ventromedial prefrontal cortex, which is crucial for processing such emotional responses, still show the same effect (Young et al., 2006).

For Knobe, these and similar studies suggest that evaluative considerations – whether something has morally good or bad consequences – genuinely factor into whether we construe actions as intentional or not. Such moral judgements, in Knobe's view, affect the way we deploy the very concept of intentional action.

Others disagree. Fred Adams and Annie Steadman, for example, have argued that a better interpretation of the Knobe effect draws on conversational pragmatics, or the way we use language to convey our attitudes and judgements (Adams & Steadman, 2004a, 2004b). On this view, people are loath to say that the CEO did not harm the environment intentionally, or that Jed did not intentionally kill his aunt, as that would seem to absolve them of the blame we would like to accord them for their actions. We want to hold them to account, and so we say that they did these things intentionally, as this is frequently a prerequisite for moral condemnation.

Knobe is not convinced, and points to unpublished studies by Tiziana Zalla, of the Institut Nicod, Paris, on people with

Intentional action, ToM and the brain

In the modern era, new psychological insights are frequently followed up with studies employing the tools of neuroscience. Rebecca Saxe of Massachusetts Institute of Technology is one of the leading researchers currently exploring the neurological basis of ToM, and the brain systems that link analyses of intentional action with morality.

In a series of papers, published with a variety of collaborators, Saxe has begun to piece together a picture of what goes on in the brain during the representation of, and reasoning about, intentional actions. Together, these amount to more than just a series of pretty coloured maps of the brain; they also point to the component functions that underlie the capacity to think about other people's thoughts, reason about intentional actions, and generate moral judgements based on these factors.

Saxe's studies indicate that the superior temporal sulcus is selectively recruited in representing observed intentional actions, over and above reacting to bodily movements *per se* (Saxe et al., 2004). Separate studies have implicated other brain regions, including areas of the prefrontal cortex, precuneus, and the temporo-parietal junction, as subserving various ToM tasks. The right temporo-parietal junction (RTPJ) in particular seems to be both selective for representing the mental states of others (Saxe & Kanwisher, 2003), and is also implicated in linking ToM with morality (Young et al., 2007; Young & Saxe, 2008).

The logic of the experiments Saxe and colleagues have used is complex, as is interpreting their results. The take-home message, however, is more straightforward, as Saxe explains. 'Brain regions that are used for thinking about people's beliefs in non-moral contexts are also used for thinking about belief in moral contexts, but we can divide thinking about beliefs into two parts,' says Saxe. 'One is figuring out what people believe – creating a mental picture of what they're thinking – and the other is using that information to make a moral judgement.' Saxe and Liane Young have couched this as the difference between encoding and integrating beliefs.

'The encoding response is insensitive to all kinds of things you might expect it to react to, such as whether a belief is true or false, justified or not, morally positive or malicious,' says Saxe. It is in the later integration stage that these factors come into play, with the RTPJ playing a particularly important role. 'During integration, the RTPJ is sensitive to the valence of the belief (whether the agent has a good or a bad intention) and the valence of the outcome (whether harm occurs or not), and we see an interaction of these factors in the RTPJ response,' explains Young. 'RTPJ response is selectively enhanced for failed attempts to harm, when subjects must rely solely on the mental state information.' Recent results show that individual RTPJ responses can be even used to predict subsequent moral judgements (Young & Saxe, 2009).

This is an unfolding story, and other researchers inevitably have different takes on the results. As Young notes, 'We're trying to figure out both the cognitive and neural divisions in a very complicated domain.'

Young, L., Cushman, F., Adolphs, R. et al. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? *Journal of Cognition and Culture, 6*, 265–278.

Young, L., Cushman, F., Hauser, M. & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences USA, 104*(20), 8235–8240.

Young, L. & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*, 1912–1920.

Young, L. & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*, 2065–2072.

autism, who have difficulties with conversational pragmatics yet also show the Knobe effect. Knobe has also used the CEO examples to explore whether it sounds natural or right to say 'The CEO did X *in order* to help/harm…', the idea being that it only feels natural to say someone did something *in order* to achieve something else when we think they did it intentionally. Mirroring the previous studies, a majority of people said it felt right to say that 'the CEO harmed the environment in order to make a profit', but not 'the CEO helped the environment in order to make a profit', suggesting a difference in how the concept of intentionality is being deployed (Knobe, 2004).

Subsequent studies have explored other supposedly non-moral judgements about peoples' actions. Take another key distinction in moral philosophy between actively 'doing' and more passively 'allowing'. This is more than a matter of mere semantics or framing of issues: think of the legal and moral difference between killing and letting die, which comes to light in particularly stark terms in debates about euthanasia. Reflecting this distinction, the American Medical Association stipulates that is sometimes morally acceptable for doctors to allow patients to die, but never acceptable for a doctor to kill a patient.

In a recent study, Knobe teamed up with Fiery Cushman, a psychologist at Harvard University, and Walter Sinnott-Armstrong, a philosopher at Dartmouth College, to explore whether moral appraisals affect doing/allowing judgements (Cushman et al., 2008). In one experiment, the trio presented participants with the case of a doctor removing life support from a patient, resulting in their death. In a 'morally bad' version, the doctor did this simply as a matter of convenience (to free up resources), while in a morally ambiguous vignette the doctor acted out of concern for the dignity of the patient (this example was deemed ambiguous because while some people will think the doctor's actions were permissible or even commendable, others will view the same act as wrong regardless of the doctor's motives).

As before, moral appraisals exerted their backward causation on doing and allowing judgements, which might ordinarily be expected to drive moral judgements. The morally bad scenario was more likely to elicit a judgment that the doctor actively did something – killing the patient, in this case – while the same actions in the ambiguous version tended to be judged as allowing the patient to die. In addition, those with stronger attitudes against euthanasia were more likely to say that the doctor killed the patient in the morally ambiguous scenario.

The finding was replicated in a second experiment, which also points to the role of pre-existing moral attitudes in determining doing/allowing judgements. Participants read a story in which a pregnant woman discovers that her baby has a dangerous vitamin B6 deficiency, and then deliberately avoids foods that would provide the vitamin B6 required to bring the baby to term. Pro-life participants were more likely to describe the woman's actions as making the fetus die than allowing it to die, and accorded greater causal responsibility to the woman in this death. 'When people construe a particular action as morally bad, they are more likely to describe the individual performing that action as actively bringing about the outcome,' says Cushman.

Pre-existing moral attitudes play a role in determining judgements

Crime and punishment

New insights into the ways moral appraisals affect judgements about the intentionality of actions also shed light on other aspects of legal debates. Ordinarily, we expect punishment to fit the crime, and further that crimes committed intentionally deserve greater punishment. In principle, it might be argued that what someone has done in the past is irrelevant to how we judge what they have done now. Yet we intuitively feel that repeat offenders should receive stiffer sentences, an intuition that is reflected in much legal practice.

In another recent study, Dorit Kliemann, Liane Young, Jonathan Scholz and Rebecca Saxe of MIT explored these intuitions by having participants judge the behaviour of people with a prior record of acting either fairly or unfairly. While positive prior record had little effect on moral judgements about actions and outcomes, a negative history led to greater moral blame, and increased ascriptions of intentionality, for the same actions when they led to bad consequences (Kliemann et al., 2008).

Brain-imaging scans in the same study also revealed that being presented with a negative prior record led to greater neural activity in a region of the brain that has been associated with representing and reasoning about the mental states of others (the right temporo-parietal junction, or RTPJ) (Saxe & Kanwisher, 2003; see also the box 'Intentional action, ToM and the brain'). The researchers suggest that a negative prior record redirects attention towards the mental states of others, and the intentional status of their actions, leading to increased activation in RTPJ (Kliemann et al., 2008).

Further complexity has been added to this picture in new studies by Fiery Cushman, who has looked at how intentions and consequences factor into judgements about the moral permissibility of certain actions, and the punishment that should be meted out for moral transgressions (Cushman, 2008). 'We need to distinguish between two categories of moral judgement,' suggests Cushman. 'One deals with punishment, and the other the permissibility and wrongness of actions.'

Take the following case. Jake and Jeff leave a party drunk, and get into their cars to drive home. Jake loses control on a corner and crashes into a tree. Jeff swerves off the road and mows down a family, killing them all. Intuitively, we feel that while the actions of Jake and Jeff – getting into a car while stocious – are equally condemnable and morally impermissible, but that they deserve different punishments. In a real legal setting, Jake is likely to get fined a few hundred pounds, and Jeff a spell in jail. For most people, it would seem unfair and unjustified to sentence Jake to a lengthy jail term, and morally obscene to

let Jeff off with a comparatively trivial fine.

Cushman's studies, in which he used a series of stories to manipulate the causal and intentional properties of actions, and their relationship with certain outcomes, suggest the following picture. 'Judgements about punishment depend relatively more on accidental outcomes than judgements about whether an action was morally permissible, or right or wrong,' says Cushman. 'When judging whether an action was permissible, we ask ourselves "Did the person do it intentionally?", and that's the end of the story. There is an effect of outcomes, but it's minuscule compared with the role of intentions. In contrast, when we make judgements about punishment, intentions matter a lot, perhaps more than outcomes, but there's a large effect size for outcomes as well.'

The two systems of moral judgement – one assessing causal responsibility for harm, the other assessing intent to harm – not only operate separately but can also come into conflict, as demonstrated by the following experiment (Cushman, 2008). Participants read about Smith, an athlete who wants to kill his competitor Brown: believing that Brown has a fatal allergy to poppy seeds, Smith sprinkles some into his salad.

In a 'no harm' version of the story, Brown eats the salad but survives (he is not, in fact, allergic to poppy seeds). Nonetheless, Smith's actions were deemed to be deserving of punishment – this is attempted murder, after all. In the 'harm' version, Smith laces the salad with poppy seeds (again causing no harm), but the chef preparing the salad adds hazelnuts (as usual for this dish), which Smith is fatally allergic to; as a result, he dies. In the harm case, where Brown is killed by a route independent of Smith's actions, people said that Smith deserved less punishment than in the 'no harm' condition, when Smith intended to kill Brown but failed.

Why should this be so? After all, Smith's murderous intentions are the same in both cases. Cushman's interpretation is that in the no-harm condition, attention is focused on Smith's immoral intentions, and he is naturally judged harshly. In the harm case, however, Brown's death initiates an unconscious search for the causal path leading to this outcome. This, in turn, refocuses attention on the chef's innocent intentions and actions, and away from Smith's malicious beliefs and desires, a process Cushman calls 'blame blocking'.

Despite all these intriguing and at times puzzling findings, the picture of the ways that moral judgements relate to ToM and other aspects of our understanding of other people – our folk psychology, in short – remains incomplete, and controversial. New studies continue to emerge, bringing to light new ways in which moral considerations affect other aspects of cognition. And some big questions remain unsolved.

One of the biggest is whether all of these strange effects of moral or evaluative considerations on other aspects of cognition can be explained by a single, umbrella theory, or whether each effect – on judgements of intentionality, causality, responsibility and so on – has its own idiosyncratic explanation. 'I think this is a relatively open question,' says Cushman, 'though I'm more inclined to think that there is some over-arching explanation that generalises across all of these cases.'

Knobe, meanwhile, leans the other way: 'Although Cushman's view may turn out in the end to be correct, my best guess at the moment is that there are really a whole mess of different processes at work in producing these different effects'. These debates in turn touch on questions about the role and purpose of folk psychology in everyday life (see box 'The function of folk').

New studies and alternative explanations continue to emerge, all of which will help build a more complex, nuanced picture. In the meantime, however, this growing body of research should give us pause for thought when we make judgement calls about whether someone did something on purpose (and consequently deserves more severe punishment), particularly when the actions and its consequences are deemed morally bad, or violate our prior moral convictions. The righteous mind must beware of becoming a biased and self-serving mind that fudges the facts to fit our convictions.

Dan Jones *is a freelance science writer*
dan.jones@multipledrafts.com

The function of folk

Humans come equipped with an intuitive understanding of the world in many domains. Our folk physics enables us to navigate a three-dimensional world at a 'mid-level' scale somewhere between the atomic and the cosmological. As such, it has much in common with scientific physics, allowing us to make predictions about how the world of objects will behave.

But what of folk psychology – our intuitive understanding of other minds, and how we reason about people's actions? Is it like scientific psychology, which aims to provide us with objective explanations of human psychology that enable us to make sense of, and at times predict, human behaviour?

The Knobe effect and more recent studies in moral cognition suggest otherwise. If our folk psychology is built to fulfil the same goals, then it might be expected to take the 'facts on the ground' as inputs into making sense of other people. For example, whether or not someone had a belief or desire to achieve some end is a matter of fact (though often hard to discern), and so should be an input into social and moral cognition.

Ascriptions of intentions, however, are often not taken as hard facts on which to build an objective assessment of the thoughts and actions of others. Rather, the 'facts' of other people's mental lives are frequently viewed through a lens coloured by the very issue that the facts are supposed to help us settle – a moral judgement in this case.

Explanations for why scientific and folk psychology diverge sit along a spectrum. At one end, favoured by Knobe, is the proposal that folk psychology has the features it does because of the way it is cobbled together, evolutionarily and developmentally, from other systems. 'The study of moral cognition is increasingly moving to the view that it's driven by a hodgepodge of different mechanisms,' says Knobe. 'There's not one single process underlying everything that's important about moral cognition – we need to invoke a diverse array of mechanisms.' No one oversaw the process of linking together these different systems, some for representing intentional actions or beliefs, others for assessing causation or assigning blame or praise. As such, the overall system has a few glitches, the Knobe effect being an example.

An alternative view sees folk psychology serving a more coherent functional or adaptive role, even if different from scientific psychology. Indeed, the task of social living may not be well served by a scientific-style psychology. 'As long as it's adaptive for us to detect harmful agents in the world, whatever helps us do so will be adaptive too,' says Liane Young of MIT. 'The Knobe effect may not look adaptive if our ultimate goal is to reason "scientifically" about people's knowledge – if that's how we want to think about intentional action, then the Knobe effect shows we're doing a pretty poor job of it. However, if our ultimate goal is moral reasoning – making sure that people aren't given credit when it's not deserved, and that potentially harmful people don't get away with their actions – then the Knobe effect makes us look pretty reasonable,' argues Young.