Theory of Mind and Moral Cognition: Exploring the Connections[1]

Joshua Knobe
(Forthcoming in *Trends in Cognitive Sciences*)

It is widely recognized that people sometimes use theory-of-mind judgments in moral cognition. A series of recent studies show that the connection also goes in the opposite direction. It appears that moral judgments can sometimes be used in theory-of-mind cognition. Thus, there appear to be cases in which people's moral judgments actually serve as input to the process underlying their application of theory-of-mind concepts.

Over the past few decades, cognitive scientists have gained an improved understanding both of how people understand psychological phenomena ('theory of mind') and of how people make moral judgments ('moral cognition'). For the most part, research in these two domains has proceeded separately. In the past few years, however, a series of studies have explored the connections between the two.

The most obvious connection, of course, arises in cases where people's beliefs about an agent's mind serve as input to the process by which they arrive at moral judgments about that agent's behavior. So, for example, if we are wondering whether or not to blame an agent for her behavior (a moral judgment), we may need to know whether she performed that behavior intentionally (a theory-of-mind judgment). In cases like these, it seems that we first arrive at a judgment about the agent's mental states and then use that judgment as input to a process that ultimately yields a moral judgment about the agent's behavior.

One surprising result from recent studies has been that there are cases in which the process appears to be working *in reverse* — cases in which people's moral judgments seem to be serving as input to the process by which they arrive at theory-of-mind judgments. In cases of this latter type, it seems that people first arrive at a judgment as to whether the behavior itself was morally good or morally bad and then use this moral judgment as input to the process by which they arrive at judgments about the mind.

Experimental evidence

For a simple example, consider the ordinary distinction between behaviors that are performed 'intentionally' and those that are performed 'unintentionally.' This distinction appears at first to be a purely psychological one, based entirely on facts about the agent's mental states and their causal roles. Nonetheless, it can be shown that people's application of the distinction is sometimes sensitive to the *moral* status of the behavior performed.

The key data here come from studies in which subjects were given brief vignettes and then asked to determine whether particular behaviors within those vignettes were performed 'intentionally.' By systematically varying aspects of the vignettes, researchers can determine which factors are influencing people's intuitions. It can thereby be shown that these intuitions show a systematic sensitivity to moral considerations.

[1] I am grateful for helpful comments from Stephen Stich and from an anonymous reviewer for *Trends in Cognitive Sciences*.

Here is one of the vignettes:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
>
> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed.

Faced with this vignette, most subjects (85%) said that the chairman *intentionally* harmed the environment [1].

One might think that this judgment was based entirely on certain information about the agent's mental states (e.g., the fact that he specifically knew the policy would harm the environment). But it seems that there is more to the story. For suppose we leave all of the agent's mental states the same but change the moral status of the behavior by simply replacing the word 'harm' with 'help.' The vignette then becomes:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
>
> The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was helped.

Faced with this second vignette, only 23% of subjects said that the chairman *intentionally* helped the environment [1]. This result suggests that people's intuitions as to whether or not a behavior was performed intentionally (a theory-of-mind judgment) can be influenced by their beliefs as to whether the behavior itself was good or bad (a moral judgment).

A similar effect emerges when people are asked to make judgments using the concept of performing a behavior *in order to* accomplish a goal. This is the concept we use when we say, for example, that an agent went to the kitchen 'in order to' get something to eat. Here we have what seems to be an especially pure case of a theory-of-mind judgment, untainted by any admixture of moral praise or blame. And yet, it can be shown that people's application of this concept is sensitive to the perceived moral status of the agent's behavior. Indeed, this effect can be seen even using the very same vignettes that were used in the intentional action experiment. Faced with the first of these vignettes, most subjects respond that it sounds right to say: 'The chairman harmed the environment in order to increase profits.' But when subjects are given the second vignette, they have a quite different response. Most respond that it does *not* sound right to say: 'The chairman helped the environment in order to increase profits.' [2] Here again, it appears that people's moral judgments are in some way shaping their intuitions about how to apply terms that we would normally associate with the domain of 'theory of mind.'

A number of subsequent papers have replicated and extended these results [3-8] (A. Leslie *et al.* unpublished), and at this point, there can be little doubt of the basic proposition that moral considerations are affecting people's use of certain terms that one would ordinarily associate with the domain of 'theory of mind.'

Explaining the effects

The remaining debate concerns the question as to whether moral considerations are actually playing a role in the fundamental competence underlying people's theory-of-mind judgments or whether these considerations are somehow exerting a distorting influence on that competence.

This debate has been dominated by three major views:

- One view holds that the emotional reactions triggered by morally bad behaviors can distort people's theory-of-mind judgments. [9-10]
- A second view holds that moral considerations play a role in the pragmatics of people's use of certain terms but not in the semantics of their theory-of-mind concepts. [11-12]
- Finally, a third view holds that moral considerations truly do play a role in the fundamental competence underlying people's theory-of-mind capacities. [13-14]

In their attempts to adjudicate between these conflicting hypotheses, researchers have been increasingly concerned with the role of people's emotional responses. Initially, it was thought that the effects might be due entirely to emotional biases, and a number of theoretical models were proposed to explain precisely how people's emotions might bias their application of certain theory-of-mind concepts [11-12]. However, subsequent studies showed that the effects emerge even in subjects with deficits in emotional processing as a result of prefrontal cortex lesions (M. Hauser *et al.* unpublished) or when the stimulus materials are specifically designed to minimize emotional response [8]. These results provide at least tentative support for the view that the effects can emerge even in the absence of emotional responses, and some researchers have recently suggested that the effects might be due, not to an emotional bias, but rather to an innate, domain-specific 'moral grammar' [15].

Future research will undoubtedly shed new light on the nature of the psychological processes at work here. It would be especially useful to have results from studies on additional clinical populations (people with autism, psychopaths, etc.) and from studies using neuroimaging techniques.

References

1. Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*. 63, 190-193.
2. Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*. 64, 181-187.
3. Malle, B. F. (forthcoming). The moral dimension of people's intentionality judgments. *Journal of Culture and Cognition*.
4. McCann, H. (forthcoming). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*.
5. Sverdlik, S. (forthcoming). Intentionality and moral judgments in commonsense thought about action. *Journal of Theoretical and Philosophical Psychology*.
6. Nadelhoffer, T. (forthcoming). Skill, luck, control, and folk ascriptions of intentional action. *Philosophical Psychology*.
7. Knobe, J. and Burra, A. (forthcoming). Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition.*
8. Knobe, J. and Mendlow, G. (forthcoming). The good, the bad, and the blameworthy: Understanding the role of evaluative considerations in folk psychology. *Journal of Theoretical and Philosophical Psychology*.
9. Malle, B. F. and Nelson, S. E. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law*. 21, 563-580.
10. Nadelhoffer, T. (forthcoming). On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology*.
11. Adams, F. and Steadman, A. (2004) Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*. 64, 173-181.
12. Adams, F. and Steadman, A. (2004). Intentional action and moral considerations: Still pragmatic. *Analysis.* 64, 268-276.
13. Knobe, J. (forthcoming). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*.
14. Mele, A. (2003). Intentional action: Controversies, data, and core hypotheses. *Philosophical Psychology*. 16, 325-340.
15. Hauser, M. (forthcoming). *Moral Minds: The Unconscious Voice of Right and Wrong*. NY: Harper Collins.