PSYCHOLOGY

Flexible updating of beliefs in order to forgive

We rapidly make inferences about the moral character of others. Observing a single immoral behaviour is often sufficient to make us think of them as morally 'unworthy'. But our beliefs about others' 'badness' (as opposed to 'goodness') are more uncertain. That is, we allow ourselves more space to re-assess and, if needed, rectify these beliefs.

Alexander Todorov

We automatically form impressions of others and act on these impressions[1]. The impression that we most heavily weigh in decisions is about moral character. We are eager to jump to conclusions about the moral character of others from surprisingly little information[2]. But we are also eager to change our minds in light of new information. While there is extensive research on the processes underlying first impressions, there is much less research on the processes underlying the updating of these impressions. Writing in *Nature Human Behaviour*, Siegel, Mathys, Rutledge and Crockett[3] present an ingenious paradigm and computational modelling to study the dynamics of impression updating. Given many previous findings that 'bad' agents — those who commit immoral acts — command attention and that extreme negative behaviours outweigh extreme positive behaviours in impressions[4–7], one would predict that people would be less willing to update their beliefs about morally bad than morally good agents. In other words, once we decide that a person is bad, they remain bad forever. However, Siegel and colleagues[3] find very different dynamics of impression updating. Beliefs about 'bad' agents are more volatile or uncertain than beliefs about 'good' agents. That is, we allow ourselves the flexibility to reconsider the 'badness' of others in light of new information.

In the studies, participants repeatedly predicted the choices of two agents choosing between a less and a more moral action. The choices were unpalatable and involved earning money while inflicting pain (electric shocks) on another person. The less morally palatable choice was to maximize one's earning while inflicting more pain. While one of the agents' choices followed the less palatable course of action, the other followed the more palatable course of minimizing both earning and pain. Participants were given feedback about the accuracy of their predictions and rapidly learned to accurately predict the choices of the agents. They also reported their impressions of the agents ('nasty' versus 'nice') and the certainty of their impressions. Not surprisingly, the impressions of the 'good' and 'bad' agents rapidly diverged and stabilized in the expected direction, but perhaps surprisingly, the impressions of 'bad' agents were more uncertain than the impressions of 'good' agents.

Credit: erhui1979/DigitalVision Vectors/Getty

Siegel and colleagues[3] used a Bayesian learning model to study the dynamics of impression updating. A critical parameter was the volatility of beliefs — monotonically related to their uncertainty — and this parameter reliably distinguished between beliefs about 'good' and 'bad' agents. The latter beliefs were more volatile, indicating more uncertainty and faster change of beliefs about 'bad' agents. Importantly, this phenomenon was specific to learning about the moral character of others. In a different study, where participants made predictions based on the skill level of agents (for example, making a number of basketball shots within a limited time frame), there was no difference between the volatility of beliefs about low-skilled and high-skilled agents, although the probabilities of high- versus low-skill actions were matched to the probabilities of more- versus less-moral choices.

The findings that beliefs about 'bad' agents are more volatile, giving us the flexibility to change these beliefs in light of new information, are fascinating and to some extent optimistic. They suggest that we are willing and ready to forgive moral infractions. But the infractions in the studies were not truly self-relevant, were relatively minor and unambiguous. What if you were a participant and the 'bad' agent was inflicting physical pain on you or somebody you love to maximize their earnings rather than on a stranger? What if the transgressions are major? Most social psychology studies in the past have used rich, narrative descriptions that often describe extreme immoral acts like stealing from an orphanage. In these cases, a single immoral act may trump multiple moral acts, suggesting that beliefs may be firmly set after a single immoral act. And what if the assessment of the morality or immorality of the act depends on our prior beliefs and, more importantly, on our values? We are particularly likely to judge harshly people who commit acts inconsistent with our values. In all of these cases, it is unclear whether our beliefs about 'bad' agents would be volatile or uncertain. Of course, a single set of studies can only answer a handful of questions.

Siegel and colleagues[3] have introduced an elegant paradigm and a simple computational model that captures the dynamics of how we change our beliefs about morally good and bad agents. The model is based on witnessing minor moral infractions, but these are the kind of infractions that most of us witness in everyday life. And perhaps, this model is a faithful representation of how we change our minds about others. Whether the model explains more extreme (and rare) moral infractions remains to be seen. ❐

Alexander Todorov
Princeton University, Princeton, NJ, USA.
e-mail: atodorov@princeton.edu

Published online: 17 September 2018
https://doi.org/10.1038/s41562-018-0442-0

References

1. Todorov, A. *Face Value: The Irresistible Influence of First Impressions* (Princeton Univ. Press, Princeton, 2017).
2. Uhlmann, E. L., Pizarro, D. A. & Diermeier, D. *Perspect. Psychol. Sci.* **10**, 72–81 (2015).
3. Siegel, J. Z., Mathys, C., Rutledge, R. B. & Crockett, M. J. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-018-0425-1 (2018).
4. Rozin, P. & Royzman, E. *Pers. Soc. Psychol. Rev.* **5**, 296–320 (2001).
5. Skowronski, J. J. & Carlston, D. E. *Psychol. Bull.* **105**, 131–142 (1989).
6. Fiske, S. T. *J. Pers. Soc. Psychol.* **38**, 889–906 (1980).
7. Reeder, G. D. *Pers. Soc. Psychol. Bull.* **19**, 586–593 (1993).

Competing interests

The author declares no competing interests.