People use theory of mind to craft lies exploiting audience desires

Marlene Berke
Yale University, New Haven, Connecticut, United States

Ben Sterling
Yale University, New Haven, Connecticut, United States

Kartik Chandra
MIT, Cambridge, Massachusetts, United States

Julian Jara-Ettinger
Yale University, New Haven, Connecticut, United States

Abstract

Theory of Mind enables us to attribute mental states like beliefs and desires. We use it cooperatively, but we also use it adversarially, as when we lie. Prior work has shown people use Theory of Mind to craft lies to be believable to their audience, based on their audience's beliefs. But we usually also know something about our audience's desires. In this work, we ask a new question: Do people cater to their audience's desires by telling them what they want to hear? We propose that people expect others to be wishful thinkers—allowing their desires to color their beliefs—and exploit this by tailoring lies to audience desires. We implement this theory as a computational model and test it against human behavior in a novel task. This model quantitatively captures people's patterns of lying—both at the population and subject levels. This work advances our understanding of social cognition in adversarial interactions.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*