What you see is what you guess: Explanations for how a choice was made are paradoxically driven by the choice (regardless of accuracy) in a human vs. AI sorting task

David A. Bosch
New York University, New York, New York, United States

Hetvi Parekh
New York University, New York, New York, United States

Yuhri Ishizaki
New York University, New York, New York, United States

Yayun Zheng
New York University, New York, New York, United States

Edward S. Kim
New York University, New York, New York, United States

Neerali Patel
New York University, New York, New York, United States

Abstract

We present evidence that explicitly explaining holistic classification choices may promote inverted mental models. That is, people believe their choices are based on observed properties but, paradoxically, their choices drive illusory observations. Participants viewed images, some generated by a human artist and others by Google Gemini. They answered questions about each, sorted them into human- or AI-generated categories, then described how they made their choices. The vast majority provided explicit explanations for how they sorted the images. Critically, many gave similar explanations (e.g., perfection indicates AI), even when incorrectly identifying human-created work as AI and vice-versa. In addition, dimensions offered as diagnostic tended to be more abstract when associated with AI (e.g., "artificialness") and more concrete (e.g., "attention to detail") for human-generated attributions. We propose a psychological mechanism for how these mental models may become inverted. Implications for inductive reasoning, naïve theories, human-computer interactions, and (mis)information effects are discussed.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*