MACHINE IMPOSTORS IMPEDE HUMAN DETECTION AND THE SUCCESS OF STABLE CONVENTIONS BY IMITATING PAST INTERACTIONS

THOMAS F. MÜLLER[*1], LEVIN BRINKMANN[1], JAMES WINTERS[2], and NICCOLÒ PESCETELLI[3]

[*]Corresponding Author: mueller@mpib-berlin.mpg.de
[1]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany
[2]School of Collective Intelligence, Mohammed VI Polytechnic University, Ben Guerir, Morocco
[3]Department of Humanities and Social Sciences, New Jersey Institute of Technology, Newark, United States

How can successful communication arise and stabilize between humans? Artificial language games have shown that we create novel *conventions* even from minimal communicative means (Galantucci, 2005; Garrod et al., 2007; Scott-Phillips et al., 2009). A convention is defined as the arbitrary solution to a repeated coordination problem (Lewis, 1969), established through, e.g., precedence (Clark, 1996). Meanwhile, results from the perceptual crossing paradigm (Auvray et al., 2009; Barone et al., 2020) have shown that human interaction in minimal environments is distinct from interaction with automated agents in its *reciprocal* patterns, i.e. interdependence between participants. However, these studies deliberately avoid conventional paths to success. In the current study, we build on these experimental approaches to investigate conventional and reciprocal routes to communication in a minimal environment. In a Turing test setup (Turing, 1950) stripped of the use of natural language, we manipulate bot behavior to assess how it affects the detection of bots, and participants' behaviors. Our main hypothesis is that access to past interactions makes bot impostors more deceptive by interrupting convention formation.

We recruited 200 participants online for the experiment and assigned them to one condition in pairs. Their goal was to find out whether their random partner in each trial was the human partner or the bot impostor. Participants interacted within a 2D space containing an orange square and a blue circle. Their only way

to communicate was to move the square, while the human partner or the bot moved the circle. In both conditions, bot behaviors were identical replays of previous behavior from human pairs, but their source differed: While bots in the *partner impostor condition* repeated behaviors shown earlier by the participant's own human partner, bots in the *foreign impostor condition* imitated an unrelated participant from a previous partner impostor condition. Participants received feedback on their own and their partner's performance after each trial.

We tested our preregistered predictions via model comparison of mixed-effects models. Predicting performance by condition revealed that participants were more successful in the foreign impostor condition than in the partner impostor condition ($\beta = -1.92$, $SE = 0.16$, $\Delta AIC = 69$; Fig. 1a). We measured the conventionality of participants' behaviors by the Earth Mover's distance to their own spatial positions over trial blocks. Human performance in the partner impostor condition suffered from conventional behavior, while the foreign impostor condition profited from it ($\beta = 0.30$, $SE = 0.08$, $\Delta AIC = 10$; Fig. 1b). Last, we measured reciprocity by computing the transfer entropy between participants' and their partners' movements, and found that it was only useful to detect bots in the partner impostor condition ($\beta = 0.17$, $SE = 0.07$, $\Delta AIC = 3.4$).

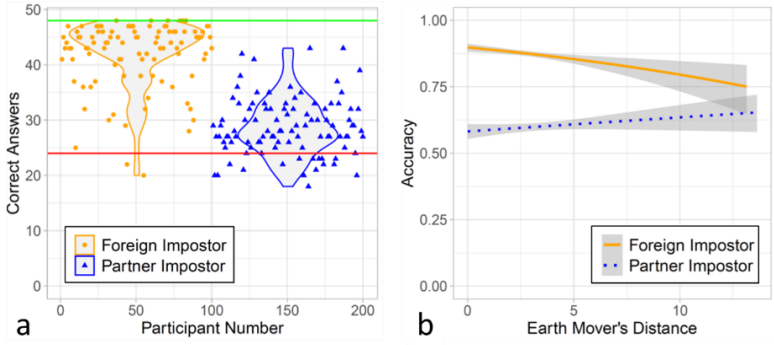

Figure 1. a) Performance results by participant. Violin plots show the density in the two conditions. The top line represents the performance ceiling, the middle line performance at chance. b) Relation between conventionality and accuracy in the two conditions. Note that higher Earth Mover's distance translates to lower conventionality. The shaded area shows the 95% confidence interval.

Our results suggest bots can avoid human detection and prevent conventional behaviors from succeeding by imitating past interactions, emphasizing the role of the interaction history for human communication. Although both conventional and reciprocal behaviors were adaptive under the right conditions, participants struggled when conventionality was maladaptive. We show how manipulating bot behavior can provide new insights into emergent communication, combining ideas from language evolution, pragmatics, and social cognition.

References

Auvray, M., Lenay, C., & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology*, *27*(1), 32–47. https://doi.org/10.1016/j.newideapsych.2007.12.002

Barone, P., Bedia, M. G., & Gomila, A. (2020). A Minimal Turing Test: Reciprocal Sensorimotor Contingencies for Interaction Detection. *Frontiers in Human Neuroscience*, *14*, 102. https://doi.org/10.3389/fnhum.2020.00102

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767. https://doi.org/10.1207/s15516709cog0000_34

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987. https://doi.org/10.1080/03640210701703659

Lewis, D. (1969). *Convention*. Harvard University Press.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233. https://doi.org/10.1016/j.cognition.2009.08.009

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460.