Machine Impostors Avoid Detection by Imitating Human Interactions in a Minimal Turing Test

Keywords: Turing test, Communication, Convention, Reciprocity, Online Experiment

Extended Abstract

Interactions between humans and bots are increasingly common, prompting some legislators to pass laws that require bots to disclose their identity (Regulation COM/2021/206 Final). Meanwhile, the amount of personal information that is accessible online is also growing, for instance via public interactions with social media. Impostor bots could potentially learn from this information to attempt personalized frauds (Seymour & Tully, 2016) or impersonate acquaintances to seem more trustworthy (Jagatic et al., 2007). Can access to the personal interaction history of a human make a bot impostor more deceptive? A better understanding of human strategies to distinguish humans from artificial agents could inform more targeted approaches to mediate the risk of deceptive bots.

The Turing test (Turing, 1950) is the classic thought experiment testing humans' ability to distinguish a bot impostor from a real human from exchanging text messages. In the current study, we remove natural language from the Turing test to open up space and investigate how human communication can arise and stabilize in the first place. In particular, we focus on conventional and reciprocal routes to communication. We define *conventions* as repeated and arbitrary solutions to a coordination problem (Lewis, 1969), established through, for example, precedence (Clark, 1996). Meanwhile, *reciprocity* is referring to interdependence between the actions of interaction partners (e.g. Barone et al., 2020). Our main hypothesis is that access to past interactions makes bot impostors more deceptive by interrupting the successful formation of conventions. This is because participants cannot rely on the simplest strategy: Repeat what has proven successful before.

We recruited 200 participants online for the experiment and randomly assigned them to a condition in pairs. Their goal was to find out whether their partner in each trial was the human partner or a bot impostor. Participants interacted within a 2D space containing an orange square and a blue circle. Their only way to do so was to move the square, while either the human partner or the bot controlled the circle. In both conditions, bot behaviors were faithful replays of previous behavior from human pairs, but their source differed: While bots in the *partner impostor condition* repeated behaviors previously shown 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 ($\beta = -1.92$, $SE = 0.16$, $\Delta\text{AIC} = 69$; Fig. 1a). We measured the conventionality of participants' behaviors by the Earth Mover's distance to their own spatial positions over trial blocks. Performance in the partner impostor condition suffered from conventional behavior, while the foreign impostor condition profited from it ($\beta = 0.30$, $SE = 0.08$, $\Delta\text{AIC} = 10$; 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 ($\beta = 0.17$, $SE = 0.07$, $\Delta\text{AIC} = 3.4$).
Taken together, the results show that bots can avoid human detection and prevent conventional behaviors from succeeding by imitating past interactions. Although both conventional and reciprocal behaviors were adaptive under the right circumstances, participants struggled when conventionality was maladaptive. Our results demonstrate how manipulating the behavior of artificial agents can provide new insights into the emergence of human communication, combining ideas from language evolution and social cognition. They also suggest that online bots accessing personal information, e.g. on social media, might become indistinguishable from humans more easily. However, we also show how reciprocal interaction could be a more difficult but reliable solution to detect unreactive artificial agents.

References

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.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100. https://doi.org/10.1145/1290958.1290968

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

Regulation COM/2021/206 final, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. European Commission. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206

Seymour, J., & Tully, P. (2016). *Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter*. https://tutorial.evogtechteam.com/wp-content/uploads/2017/03/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf

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

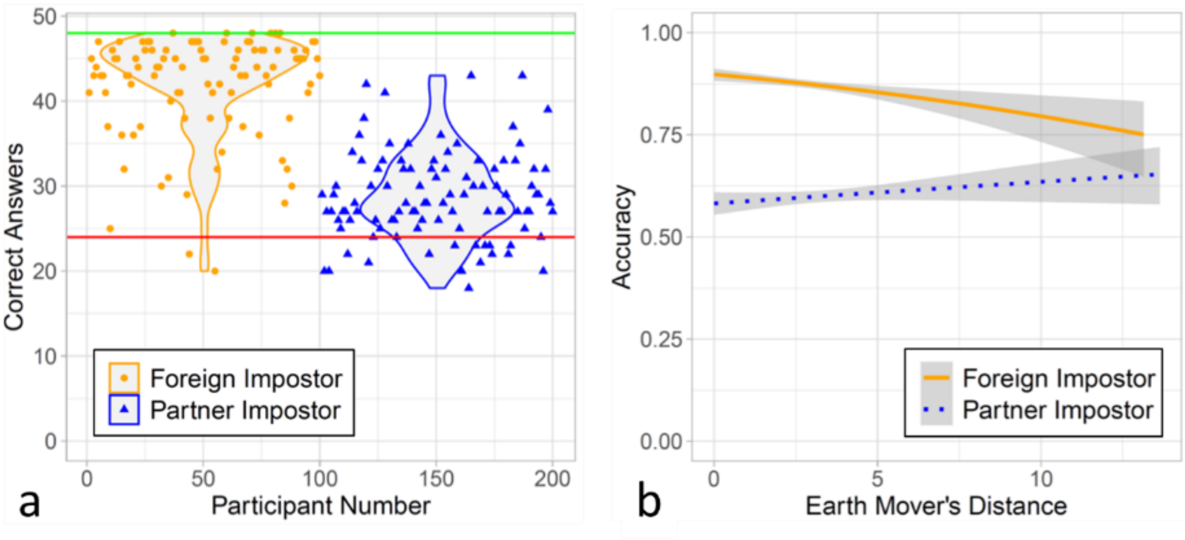

Figure 1. a) Performance results by participant. The top line represents the performance ceiling, the middle line performance at chance. **b)** Relation between conventionality and accuracy in the two conditions. Note that higher Earth Mover's distance translates to lower conventionality.