A Cognitive Framework for Strategic AI Communication

Lukas W. Mayer, Mark Steyvers
Department of Cognitive Sciences, University of California, Irvine

Introduction

AI assistants are designed to help people do tasks. However, people do not always want help. Even worse, **AI assistance that is perceived redundant/unreliable is quickly turned off**, extinguishing any future possibility for beneficial interactions.

We created a decision-making task with a **deliberately annoying AI** assistant to study **when people turn AI off**, and possibly, back on.

Method

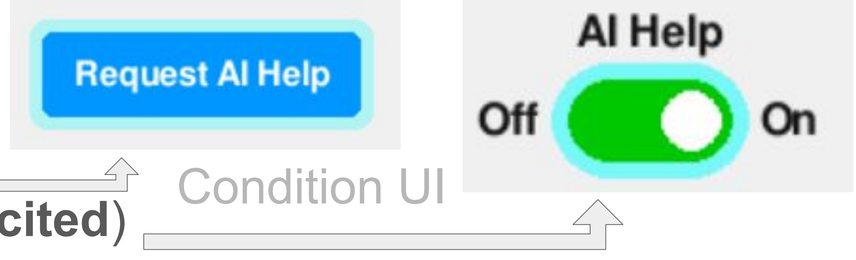
198 participants judged moving grids, determining which **column** produces the most **black squares**. There are **40 trials, 10 for each difficulty level**. Trials are **ordered by difficulty** (e.g. easy to hard, hard to easy).

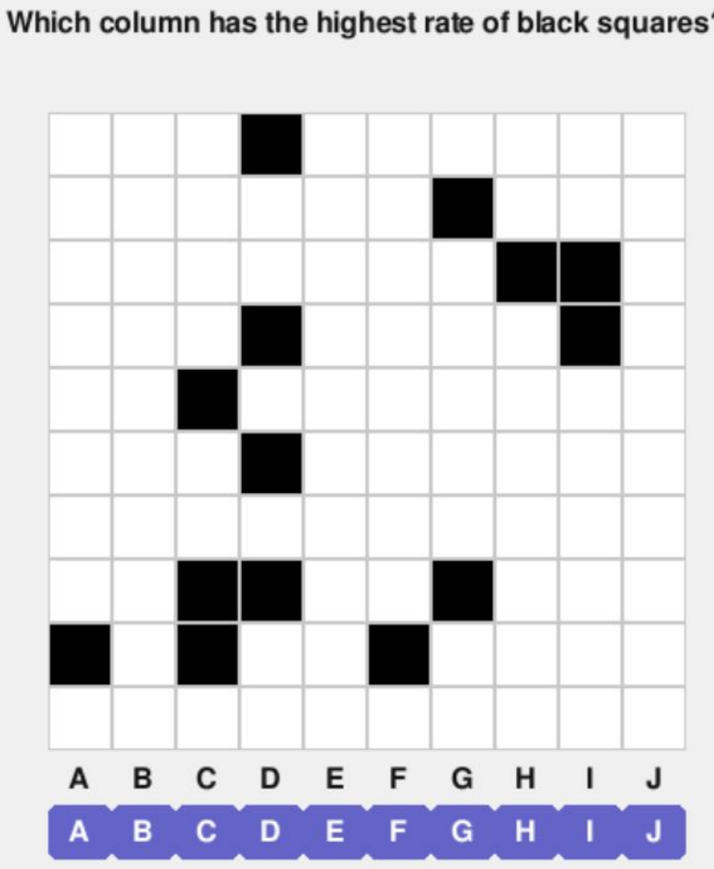
Conditions (6 x 2 Orders):

- No AI help (**Control**)
- AI upon request (**Solicited**)
- Unprompted AI pop-ups (**Unsolicited**)

Unsolicited sub-conditions:

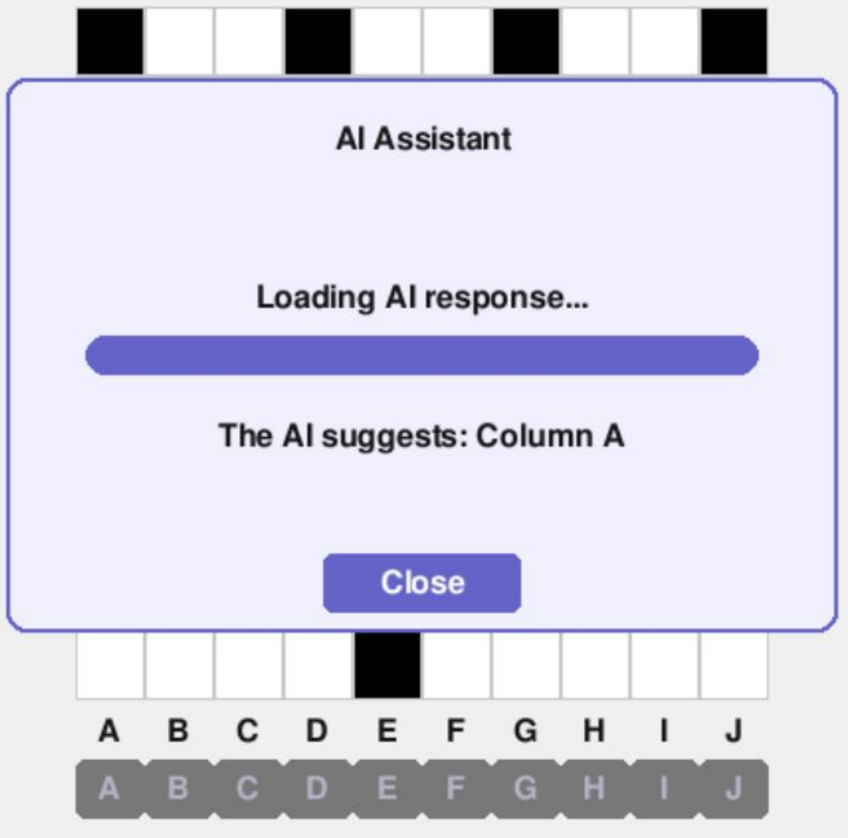
- AI help pop-up on every trial (**`always`**)
- AI help pop-up on 50% of trials, random (**`random_50`**)
- AI help pop-up on the 50% hardest trials (**`high_diff`**)





Target column differs from base columns in production **rate** by: **1%, 10%, 20%, 30%**

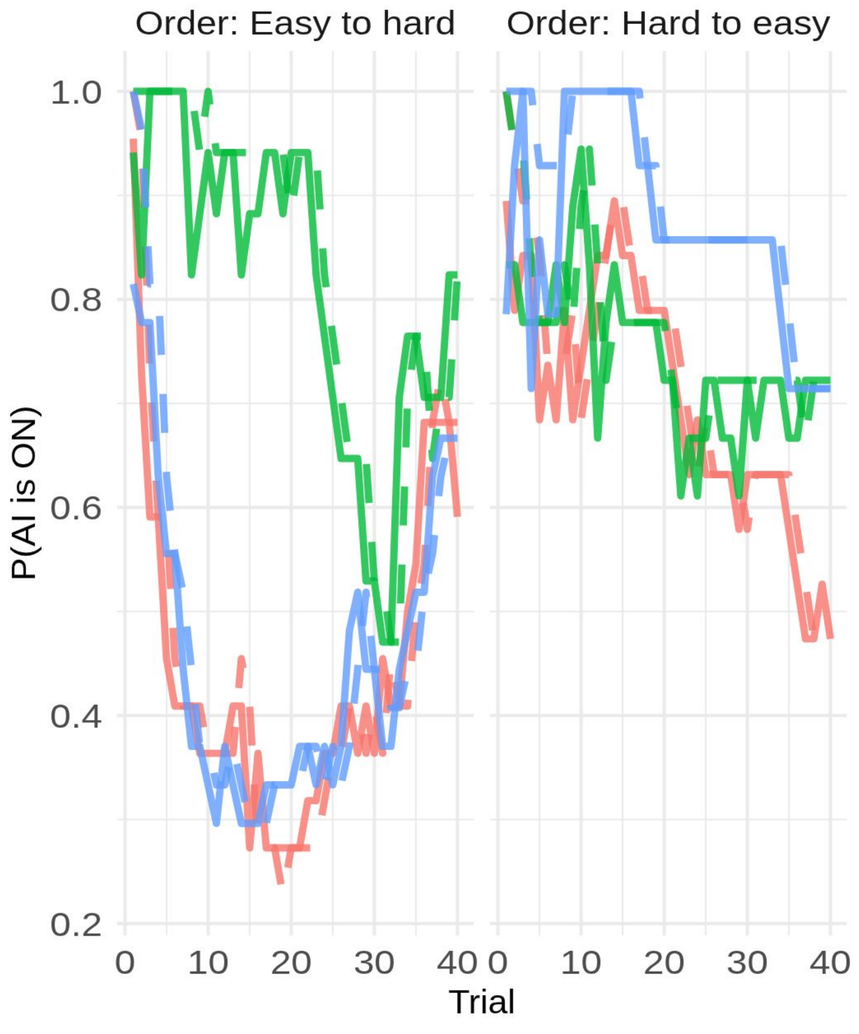
AI Pop-up **(Unsolicited only)**: Response options available after 6s wait.



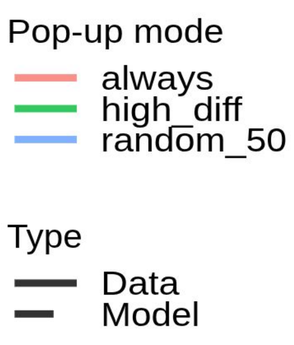
AI help (80% Acc.) available 5s into trial; **AI load times: 6s Unsolicited** (6s wait + 6s load), **12s Solicited**

Results

- People over-adopt AI advice in easier trials, under-adopt in harder trials (limited metacognition?)
- AI conditions show lower productivity (Correct/Minute) than Control, but productivity maximizing AI use would be significantly better than Control
- **Turning off AI** is predicted by **recent frequency** of AI pop-ups, **turning AI back on** is predicted by **trial difficulty**



Model: GLMM (Logit) with linear splines
AUC: 0.82



Next steps

Incorporate **our model** of human behavior into **Transition function** of **POMDP**, estimate optimal policy for **when to provide AI pop-ups** for maximum long-term productivity.

References

Chen, G., Li, X., Sun, C., & Wang, H. (2024). Learning to make adherence-aware advice. In The twelfth international conference on learning representations.

Noti, G., & Chen, Y. (2023). Learning when to advise human decision makers. In Proceedings of the thirty-second international joint conference on artificial intelligence (pp. 3038–3048).