Adaptive interventions for both accuracy and time in AI-assisted human decision making

Siddharth Swaroop, Zana Buçinca, Finale Doshi-Velez
Harvard University, USA; contact: siddharth@seas.harvard.edu

Summary

- Users can be both time-pressured and need high accuracy
- For example, doctors working in Emergency Rooms
- AI assistances trade off response time and accuracy differently
- We want to **adapt** AI assistance depending on question and user
- We introduce a task, run pilot studies, and find evidence we should:
  - On easy questions, show AI assistance **before** initial decision
  - On hard questions, show AI assistance **after** initial decision
- Future: (i) run full study, (ii) adapt using reinforcement learning

Alien prescription task

- Participants prescribe medicines to sick aliens, inspired from [1]
- Treat as many sick aliens as possible in medical shift (15-20 mins)
- Alien treatment plan: set of decision rules (easy for human to parse)
- Two correct medicines: suboptimal uses fewer observed symptoms

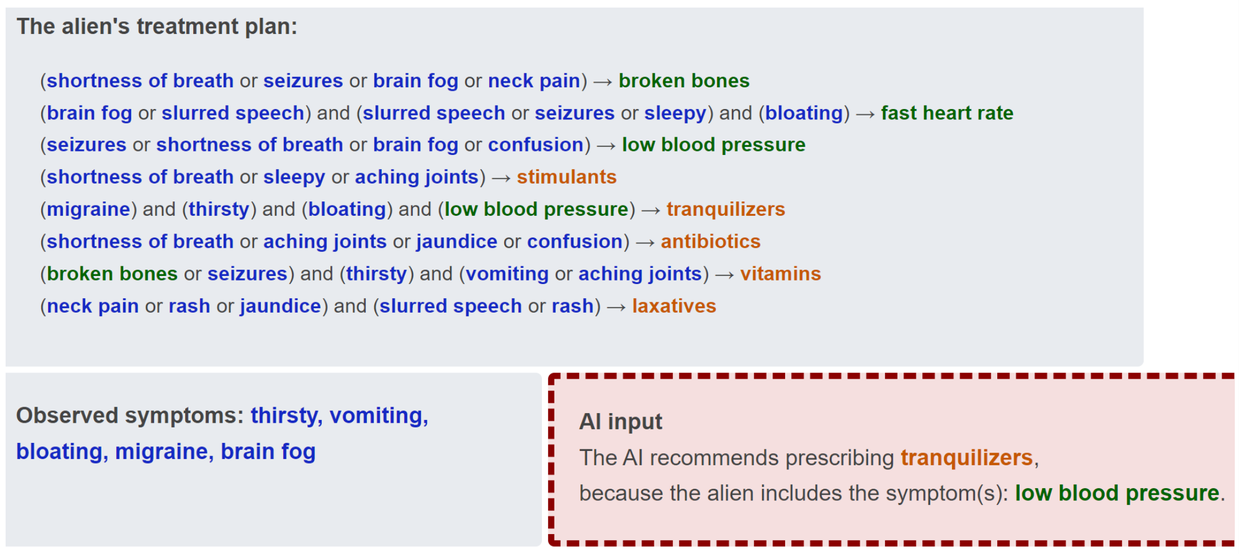

Figure 1: Alien prescriptions task.
- Participants must use observed symptoms and rules to prescribe a single medicine. Only the observed symptoms and intermediate (green) symptoms can be used, no other symptoms.
- AI assistance is shown in a red box
- Here, the AI recommendation is the best (tranquilizers uses the most observed symptoms)
- Vitamins is also correct, but is suboptimal: it uses fewer observed symptoms
- All other medicines are wrong/incorrect

- Three AI assistance types:
  - **No-AI**
  - **AI before**: AI recommendation + explanation before initial decision
  - **AI after**: AI recommendation + explanation after initial decision
- Also tested: timer shown on screen vs no timer
- Half of questions were 'easy', half were 'hard' (see paper)
- Six pilot studies on Prolific, 20 participants (remove 4-7 each study)
  - $12/hr (30-35 mins); $3 bonus to best-performing in each study

- Metrics (avg within participant; mean & std error across participants):
  - **Accuracy**: 1 for optimal medicine, 0.5 for suboptimal, 0 for wrong
  - **Response time**
  - **Overreliance**: proportion of times same as AI when AI non-optimal
  - **Underreliance**: proportion of times non-optimal when AI optimal

Results

Difficulty	Avg acc	Avg time (s)
All	0.80±0.05	66±7
Easy	0.91±0.02	60±7
Hard	0.69±0.08	76±8

Table 1: Mean and standard error on the No-AI condition (n=14 participants).
- Participants achieve 80% accuracy on average, taking 66 seconds per question
- Higher accuracy (91%) and quicker on easy questions

Question difficulty	AI condition	Change in avg acc	Change in avg time (s)
All	AI before	-0.005±0.03	-13±3
	AI after	0.09±0.03	9±1
Easy	AI before	-0.04±0.04	-11±4
	AI after	0.02±0.03	8±1
Hard	AI before	0.007±0.06	-12±6
	AI after	0.17±0.07	9±2

Table 2: The effect of AI assistance type (AI before and AI after), measured as within-participant differences to the No-AI condition (mean and standard error).
- AI before (n=17) saves time without impacting accuracy, can use on easy questions
- AI before overreliance high (48±9%), underreliance low (8±3%): people follow AI
- AI after (n=14) increases response time, but increases accuracy on hard questions
- AI after overreliance low (14±8%), underreliance low (12±4%): people perform well

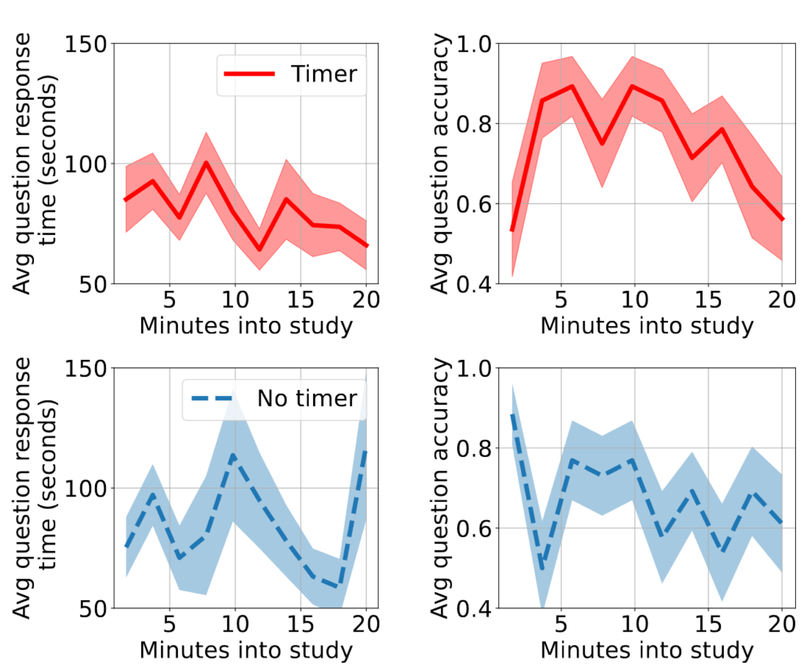

Figure 2: Plots of response time and accuracy during the course of the study.
- Top row: when a timer is shown to participants (n=14), they maintain a fast pace of answering questions (top left), but accuracy reduces later in the study (top right)
- Bottom row: when no timer is shown, they maintain a constant accuracy (bottom right), but average response time is high (bottom left)

Discussion / Future work

- Adapting AI assistance type leads to good accuracy/time tradeoffs
  - On easy questions, can show AI before
  - On hard questions, can show AI after
  - When time-pressured (as time runs out), slow down users (AI after)
- Benefit of adapting to users? Intrinsic motivation to think (NFC), reaction under time-pressure, overall skill, trust in AI.
- Will run larger study based on these results
- Future work: can use reinforcement learning to adapt quickly
- Effect of low-stakes environment?

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2107391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[1] Lage et al., 2019: An evaluation of the human-interpretability of explanation.