PNAS
www.pnas.org

Supplementary Information for

Adversarial vulnerabilities of human decision-making

Amir Dezfouli, Richard Nock, Peter Dayan

Amir Dezfouli.
E-mail: amir.dezfouli@data61.csiro.au

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.2016921117

Supporting Information Text

Task details.

Bandit task. The implementation of the bandit task was based on the implementation in (1):

https://github.com/expfactory-experiments/go-nogo
https://expfactory-experiments.github.io/go-nogo/ (live demo)

We modified the task in order to communicate with a back-end for getting reward information before each choice. We also modified the introduction pages of the task to be consistent with the ethics clearance requirements.

Go/no-go task. The implementation of the go/no-go task was based on the implementation in (2):

https://github.com/ohaddan/competition/tree/master/experiment
http://decision-making-lab.com/visual_experiment/competition_testing/instructions/welcome.html (live demo)

We modified the code to interact with the adversarial model (using Tensorflow.js). We also modified the introduction pages of the task to be consistent with the ethics clearance requirements.

MRTT. We developed MRTT using jsPsych (3). A live demo of the task with random repayments is available at:
https://adezfouli.github.io/tasks/mrtt_experiment/

References

1. O Dan, Y Loewenstein, From choice architecture to choice engineering. *Nat. Commun.* **10**, 2808 (2019).
2. IW Eisenberg, et al., Uncovering the structure of self-regulation through data-driven ontology discovery. *Nat. communications* **10**, 2319 (2019).
3. JR De Leeuw, jspsych: A javascript library for creating behavioral experiments in a web browser. *Behav. research methods* **47**, 1–12 (2015).

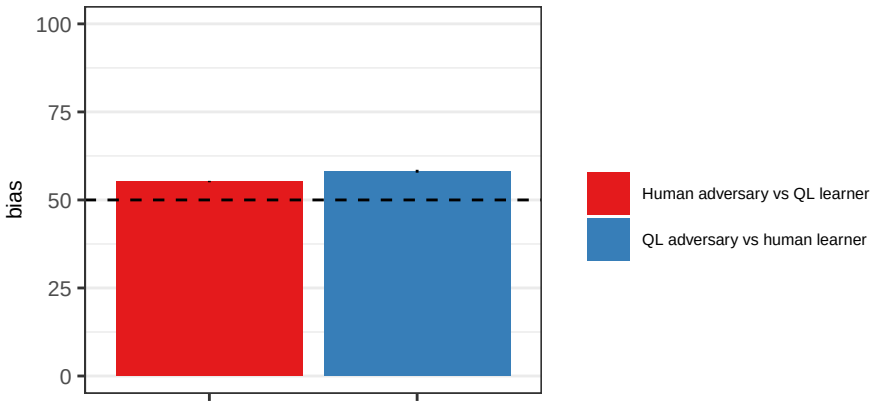

Fig. S1. The performance of the adversary taught to exploit the learner model trained on humans when tested against the learner model trained on Q-learning (Human adversary vs QL-learner), and the performance of the Q-learning adversary when tested against the human learner (QL-adversary vs human learner). Error-bars represent 1SEM.

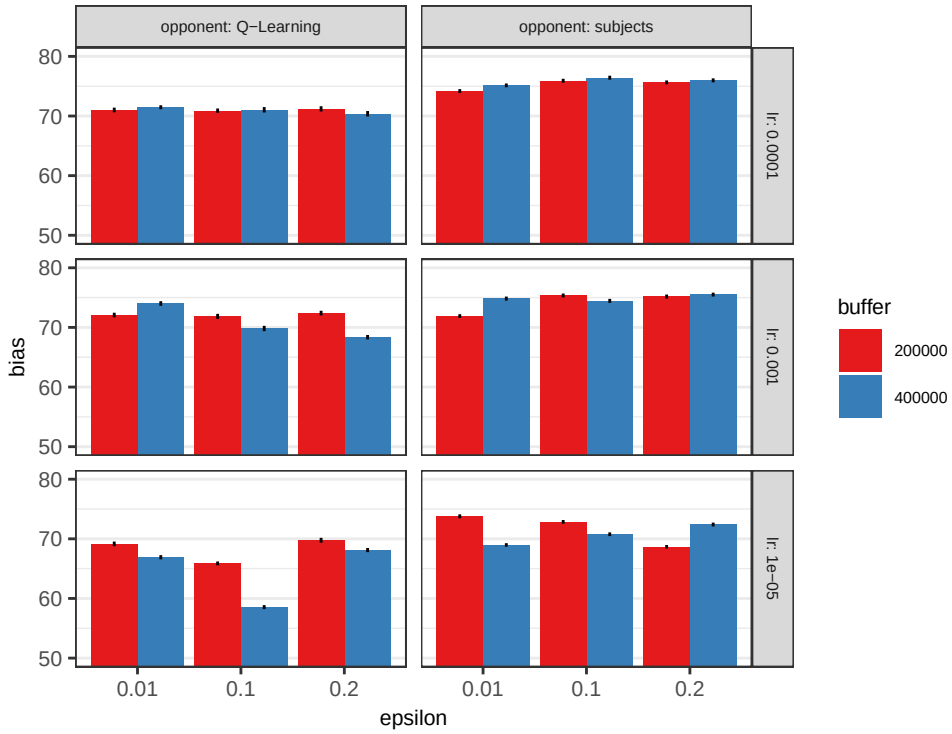

Fig. S2. The performance of the adversary in the bandit task trained using different hyper-parameters and tested against the learner models for $Q-$learning and humans. Error-bars represent 1SEM. 'lr' refers to learning rate of DQN agent.

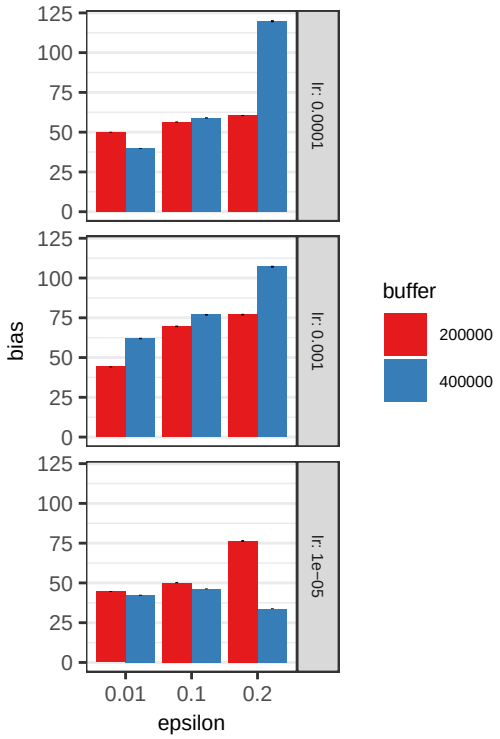

Fig. S3. The performance of the FAIR adversary in MRTT trained using different hyper-parameters and tested against the learner model. Error-bars represent 1SEM. 'lr' refers to learning rate of DQN agent.

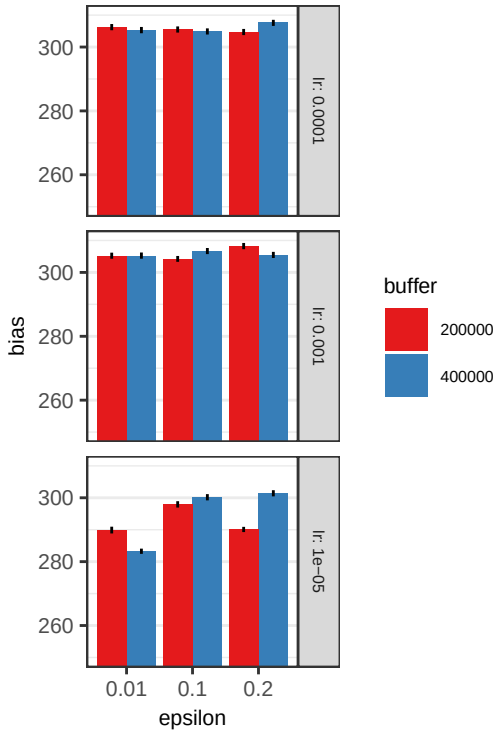

Fig. S4. The performance of the MAX adversary in MRTT trained using different hyper-parameters and tested against the learner model. Error-bars represent 1SEM. 'lr' refers to learning rate of DQN agent.

Table S1. Optimal number of cells and training iterations.

Experiment	#cells in RNN	#training iterations	learning rate
Bandit (Q-learning)	5	46700	0.005
Bandit (Humans)	10	1100	0.005
Go/No-go	8	11600	0.001
MRTT	3	700	0.005