Controllability and resource-rational planning

Falk Lieder, Noah D. Goodman, Quentin JM Huys
contact: falk.lieder@gmail.com

University of Zurich[UZH]
ETH
Leland Stanford Junior University

Introduction

- Controllability bounds the differential utility of different actions
- Therefore, rational agents should invest less time into planning, the less control they have over their environment
- What is the optimal tradeoff between planning time and expected gain, and how does it depend on controllability?
- Can the optimal tradeoff explain aberrant planning and decision making?

Resource-Rational Planning

Sample-based planning:

Here we model how the brain solves large Markov decision problems (MDPs) as Monte-Carlo tree-search based on [1]:

$$\hat{Q}(s,a) = \frac{1}{k}\sum_{i=1}^{k}\left(r(s,a,s_i) + \hat{V}(s_i)\right), \qquad s_i \sim P(S_{t+1}|s_t,a) \qquad (1)$$

$$\hat{V}(s) = \max\{\hat{Q}(s,a_1),\cdots,\hat{Q}(s,a_N)\} \qquad (2)$$

1. $\hat{Q}(s,a)$: est. expected cumulative reward for action a in state s
2. $V(s)$: value of state s

Resource-Rationality:

The resource-rational [2] decision which actions to simulate and how often ($\mathbf{c}$) maximizes the **value of computation** (VOC):

$$\mathbf{c} = \arg\max_{\mathbf{c}\in\mathcal{C}^n} \mathrm{VOC}(\mathbf{c})$$

$$\mathrm{VOC}(\mathbf{c}) = \mathbb{E}_{P(B|c)}\left[\max_a \mathbb{E}_{P(Q,S|B)}\left[Q(s,a)\right] - \mathrm{cost}(\mathbf{c})\right]$$

Uncertain MDP and prior knowledge about control

In general, the MDP is partially unknown. Planning under uncertainty about outcome probabilities $\boldsymbol{\theta}$ was formalized as an augmented MDP:

$$M(\boldsymbol{\theta}) = (\mathcal{S}' = \mathcal{S}\times\mathcal{B}, \mathcal{A}, P(S_{t+1}|S_t,a_t), P_{\boldsymbol{\theta}}(R_t|S_t,a_t))$$
$$P_{\boldsymbol{\theta}}(R_t|S_t,a_t) = \mathrm{Multinomial}(\boldsymbol{\theta}_{s,a})$$
$$P_t(\boldsymbol{\theta}_{s,a}) = \mathrm{Dirichlet}(B_t(s,a));$$

- B_t: belief about outcome probabilities
- $\mathcal{S}'$: combines world-state S_t with belief B_t
- $B_0(s,a) = \alpha\cdot\mathbf{1}$ is informed by abstract knowledge about control: high $\alpha \Rightarrow$ random outcome independent of $a \Rightarrow$ no control [3,4]

Cognitive control as a meta-level MDP

VOC can be maximized by solving a simpler meta-level MDP [5]:

$$M^{\mathrm{meta}} = (\mathcal{S}_Q, \mathcal{C}, P_S^{\mathrm{meta}}, R^{\mathrm{meta}}, s_0^{\mathrm{meta}}) \qquad (3)$$

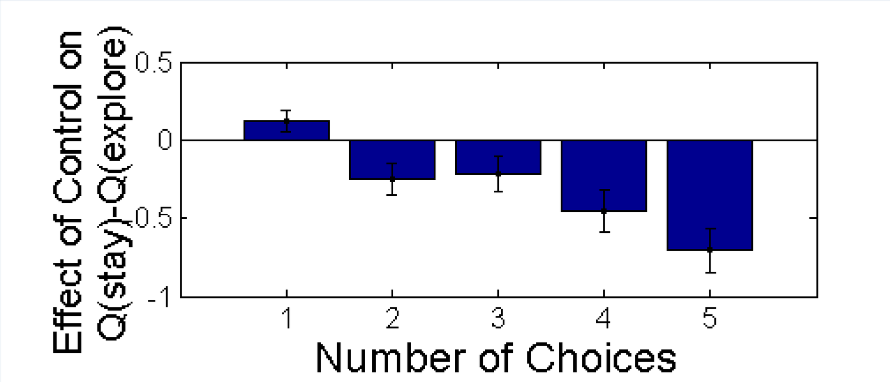
- meta-level states $S_Q^t = \{(\mu_i^t,\tau_i^t)\}$ are beliefs about Q-values: $P(Q(s,a_i)) = \mathcal{N}(\mu_i^t,\tau_i^t)$.
- comp. actions $\mathcal{C}$: $\perp$: stop planning, c_i: simulate action i
- P_S^{meta}: Bayesian learning from $q\sim\mathcal{N}(Q(s,a_i),\tau_i^{\mathrm{sample}})$
- reward fct. R^{meta}: $-\mathrm{cost}(c_i)$ for computations, cumulative reward expected under current meta-level belief for $\perp$.

Analytic results enable efficient approximate solutions [5,6].

Resource-Rational Effects of Control

1. Effect on exploration vs. exploitation

Control determines the differential value of exploration vs. exploitation [3, 4]. Our model explains how controllability can be taken into account with a cognitively plausible amount of computation.



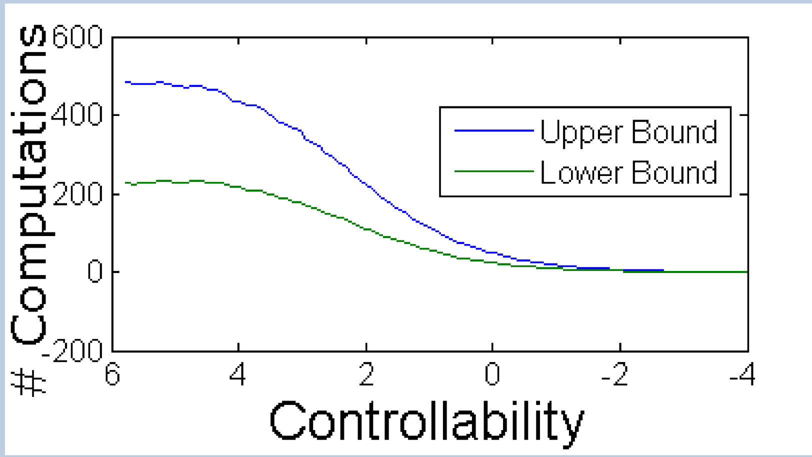
Estimated value of exploration in the 8-armed bandit task by [3] for high vs. low control based on 200 simulations with $k=2$ samples per inner node.

2. Effect on mental effort

We derived bounds on the number of simulations n chosen by the optimal meta-level policy

$$n \le \frac{k}{\min_i \tau_i^{\mathrm{sample}}}\cdot\left(\frac{1}{c\cdot\sqrt{2\pi}} - \min_i\{\tau_i^0+\tau_i^{\mathrm{sample}}\}\right)$$
$$n \ge \frac{1}{\max_i \tau_i^{\mathrm{sample}}}\cdot\left(\frac{1}{c\cdot\sqrt{2\pi}} - \max_i\{\tau_i^0+\tau_i^{\mathrm{sample}}\}\right).$$

- high cost of computation $c \Rightarrow$ low bounds.
- low control ($-\log(\alpha)$) $\Rightarrow$ high certainty τ_i^0 about $Q(s,a) \Rightarrow$ low n.



Perceived uncontrollability makes it irrational to plan ahead.

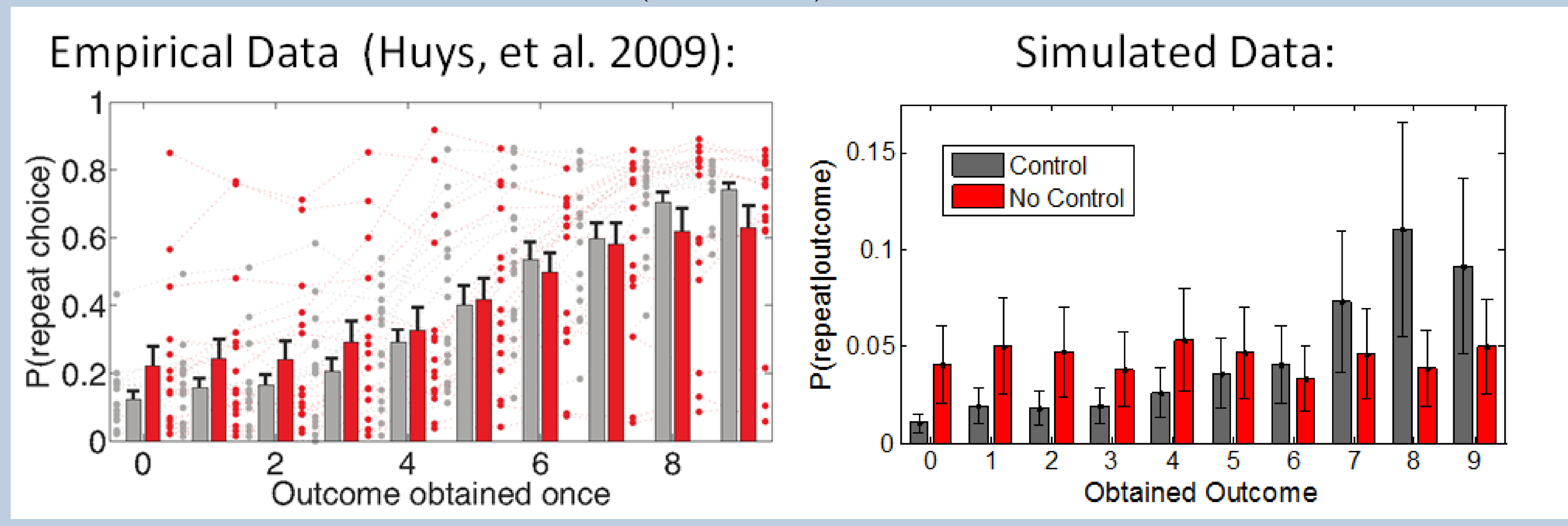
Major Depression (MDD) and Control

Lower repeat modulation in MDD [3]:

- 8-armed bandit task with 8 sequential choices
- independent, unknown reward distributions ($R\in\{0,\cdots,9\}$)
- MDD patients exerted less control: less likely to stick with good arms and move away from bad arms (repeat modulation).

Simulations:

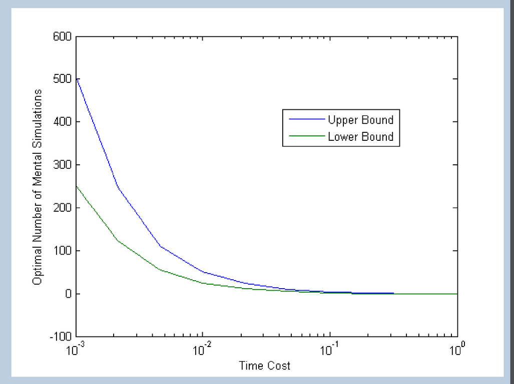
perceived lack of control (high α) $\Rightarrow$ reduced repeat modulation



Discussion

Alternative Explanations:

1. reduced processing speed $\rightarrow$ increased cost of comp. $\rightarrow$ less planning (see –>)
2. perceived lack of control impairs learning (cf. [7])



Conclusions

1. Resource-rationality [2] explains *why* people track control and *how* it shapes learning and decision-making
2. Impaired decision-making and learning [7] in major depression may result from the perceived lack of control (helplessness)
3. Uncontrollability reduces the utility of goal-directed decision making. This may trigger a shift to habitual or Pavlovian choice.

References

[1] Kearns, Mansour, and Ng. *Machine Learning*, 49(2), November 2002.

[2] Lieder, Griffiths, and Goodman. *NIPS 2012*, 2013.

[3] Huys, Vogelstein, and Dayan. *NIPS 2008*, 2009.

[4] Huys and Dayan. *Cognition*, (3), December 2009.

[5] Hay, Russell, Tolpin, and Shimony. *UAI*, August 2012.

[6] Hay and Russell. Technical report, EECS, UC Berkeley, 2011.

[7] Lieder, Goodman, and Huys. In *CogSci 2013*, submitted.