HUMAN MACHINE CO-LEARNING

Xiaojin Zhu
Bryan Gibson
KwangSung Jun
University of Wisconsin-Madison, USA

NIPS 2009 Workshop on Interactive Machine Learning

Outline

Guide human learning by machine learning

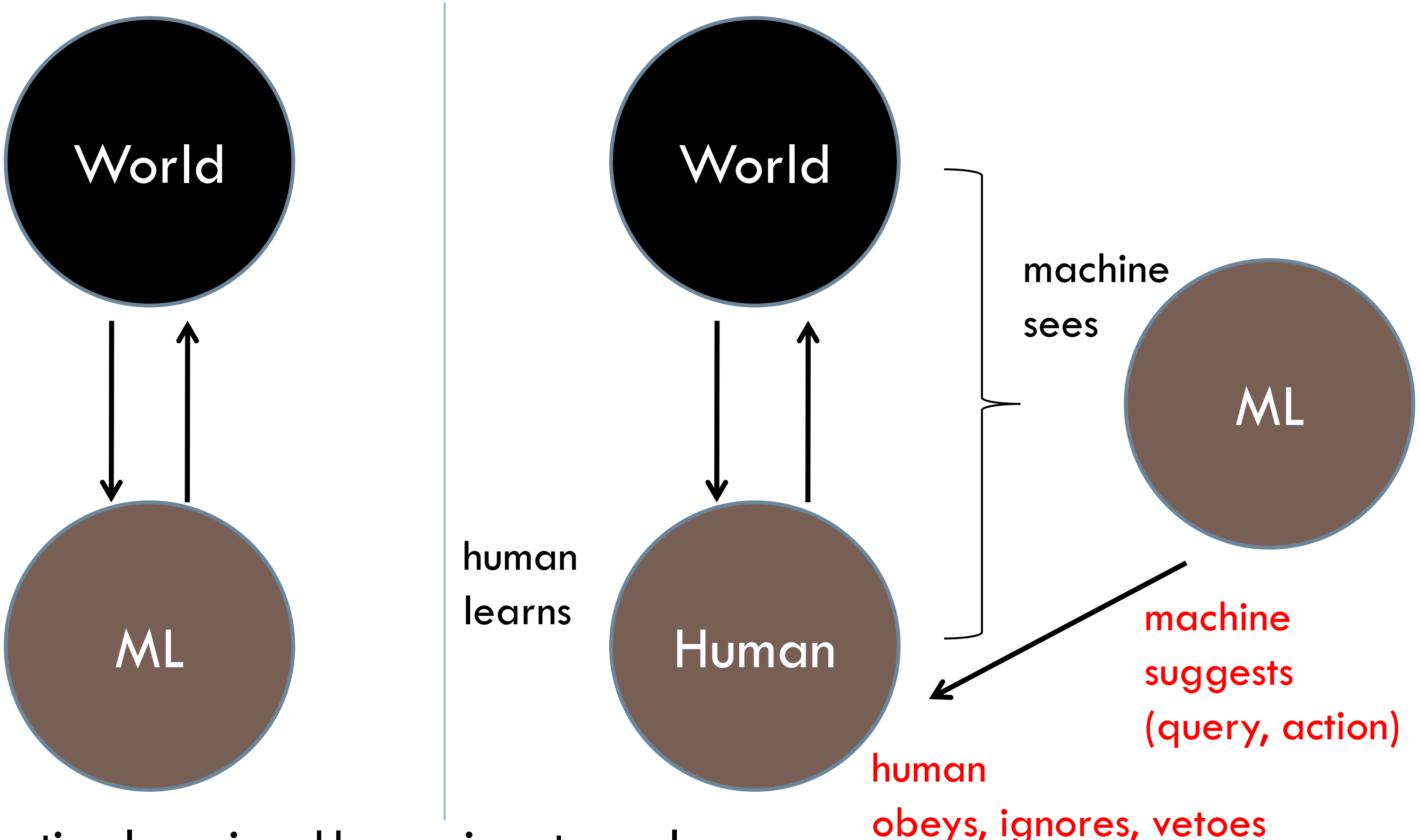
- Human is the boss
- Multi-Armed Bandit testbed
- Suggestions, more suggestions, and reverse psychology
- Speculations

Terminator 3

- General Brewster (PI): “Mr. Chairman, I need to make myself very clear. If we uplink now, Skynet will be in control of your military. ”
- “But you'll be in control of Skynet, right? ”
- “(pause) That is correct, sir. ”
- “Then do it.”

Human’s desire to control machine learning

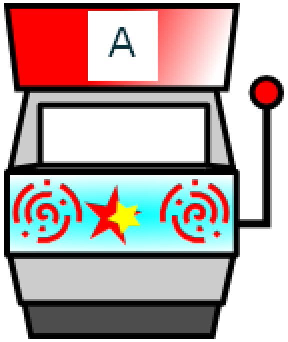
Human-Machine Co-Learning: Learning when human is the boss



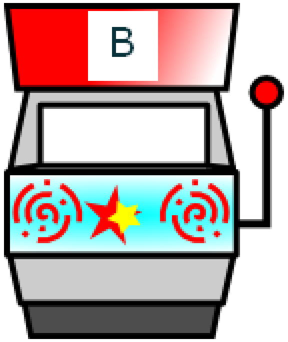
- Not active learning: Human is not oracle
- Not computer tutoring: Machine does not know the world either
- Two learning systems interact. Goal: maximally help the human learner

To make things concrete…

- Example:
  - World = Multi-Armed Bandit (Whistler Restaurant Problem)



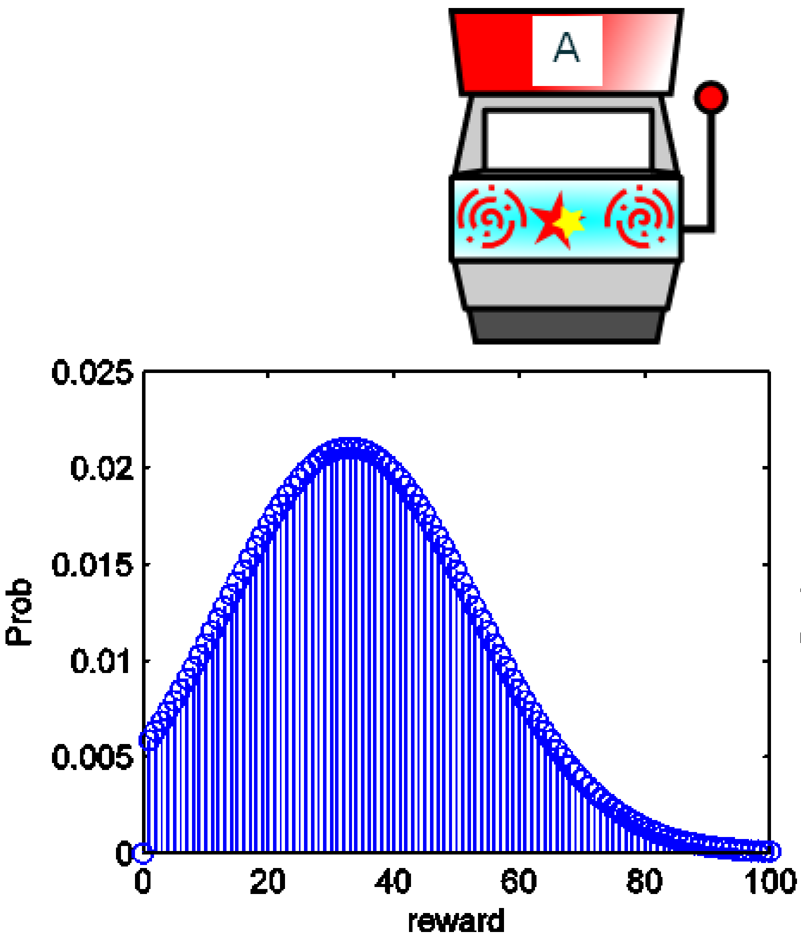
reward ~ P_A



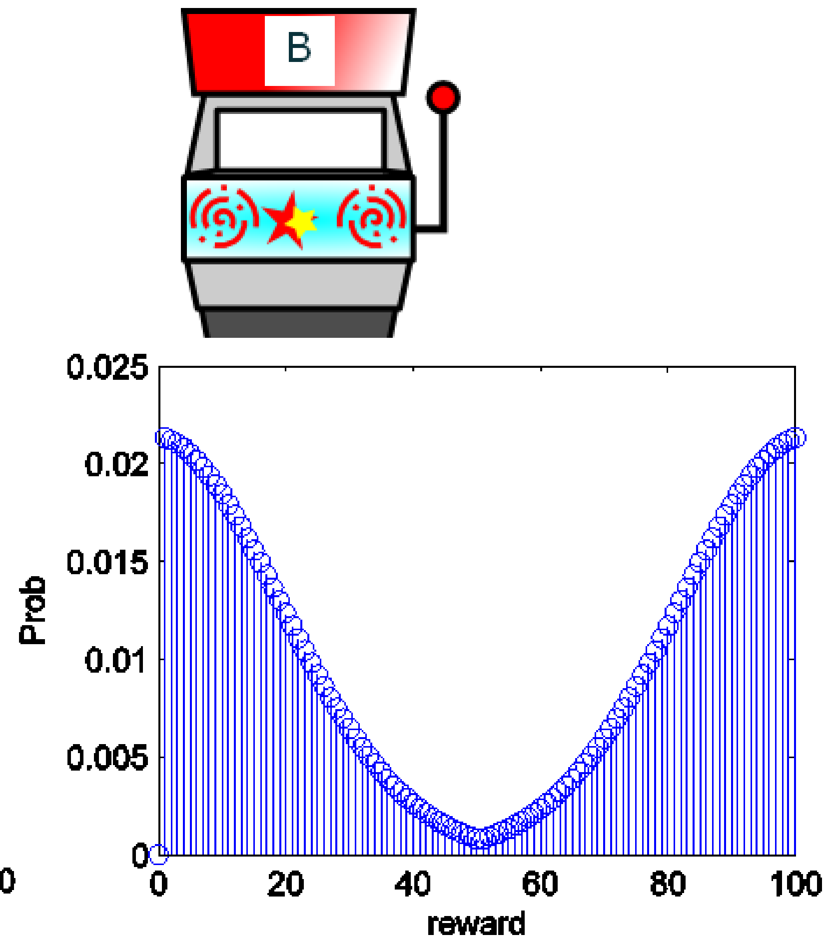
reward ~ P_B

  - Human = user
  - Machine learning = smartphone
- Demo

The truth



Mean μ_A=35.2



Mean μ_B=50.5

Machine good

- Let $x_1, \ldots, x_n$ be the rewards received in *n* trials
- Regret $n\mu^* - \sum_{i=1}^{n} x_i$ where $\mu^* = \max(\mu_A, \mu_B)$
- Per-trial regret $\mu^* - \frac{1}{n}\sum_{i=1}^{n} x_i$
- There is a rich literature in machine learning on optimal MAB strategies
  - e.g., UCB1

UCB1 [Auer, Cesa-Bianchi, Fischer]

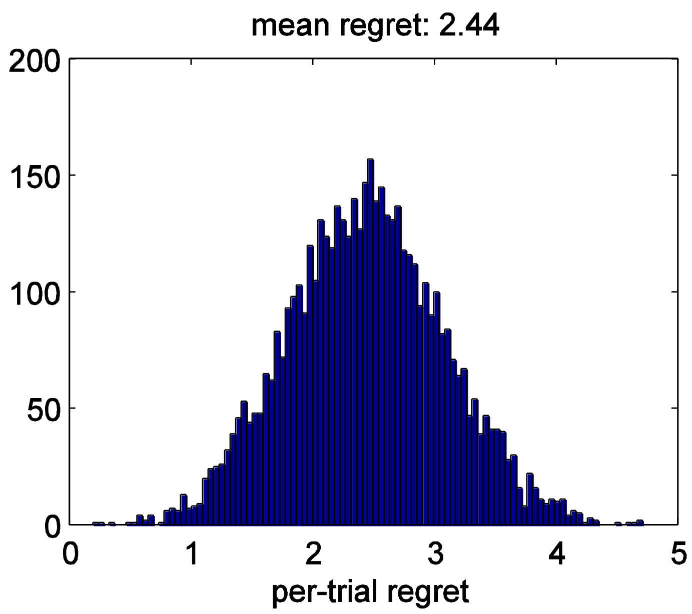
- Initialization: play each arm once
- Repeat:
  - Play arm $\arg\max_j \overline{x_j} + \sqrt{\frac{2\ln n}{n_j}}$
  - $\overline{x_j}$ is the average reward from arm *j* so far
  - n_j is the number of times arm *j* has been played
  - n is the overall number of plays
- Regret $O(\ln n)$

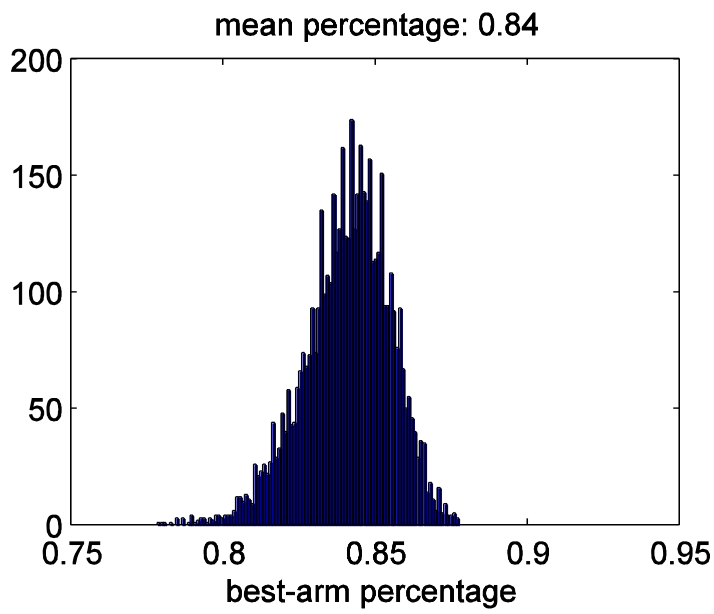
UCB1-tuned

- Empirical enhancement
- Play arm
$$\arg\max_j \overline{x}_j + \sqrt{\frac{\ln n}{n_j} \min(\frac{1}{4}, V_j(n_j))}$$
- Upper variance bound for arm *j* which is played *s* times in *t* trials:
$$V_j(s) = \left(\frac{1}{s}\sum_{r=1}^{s} x_{jr}^2\right) - \overline{x}_{js}^2 + \sqrt{\frac{2\ln t}{s}}$$

Machine good

- UCB1-tuned performance, averaged over 5000 sessions. Each session has 29 trials. Each trial has length 150.

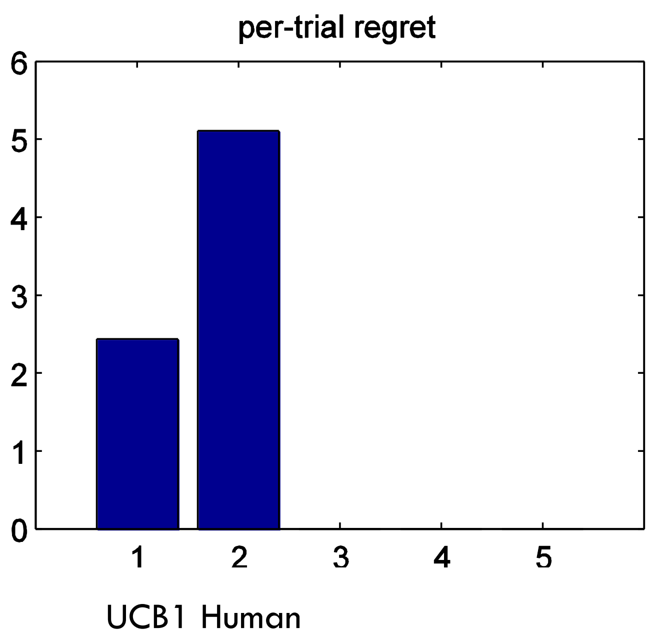


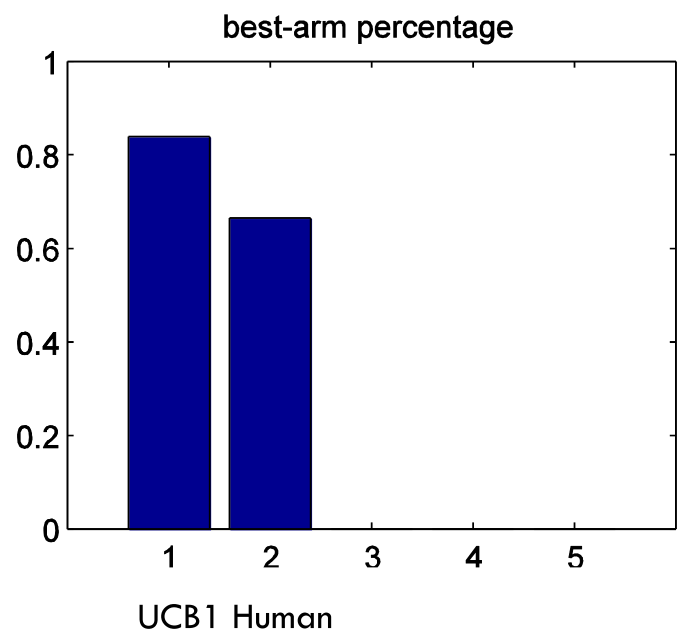


Human bad

- There is also a rich psychology literature on human sub-optimal performance on MAB [e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan 06; Lee, Zhang, Munro, & Steyvers 09; Acuna & Schrater 08]
- Psychology experiment
  - 28 undergrads
  - 150 pulls each

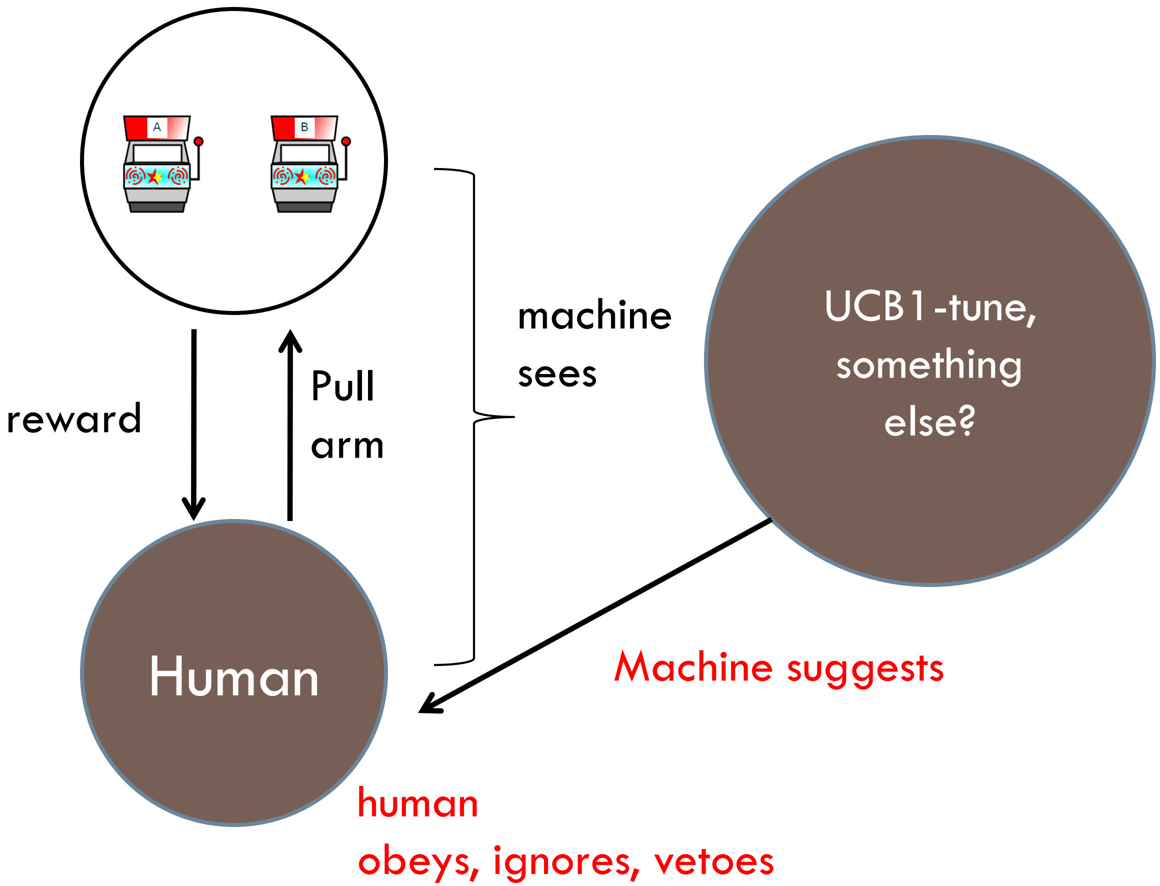
Human bad


per-trial regret
6
5
4
3
2
1
0
1
2
3
4
5
UCB1 Human


best-arm percentage
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
UCB1 Human

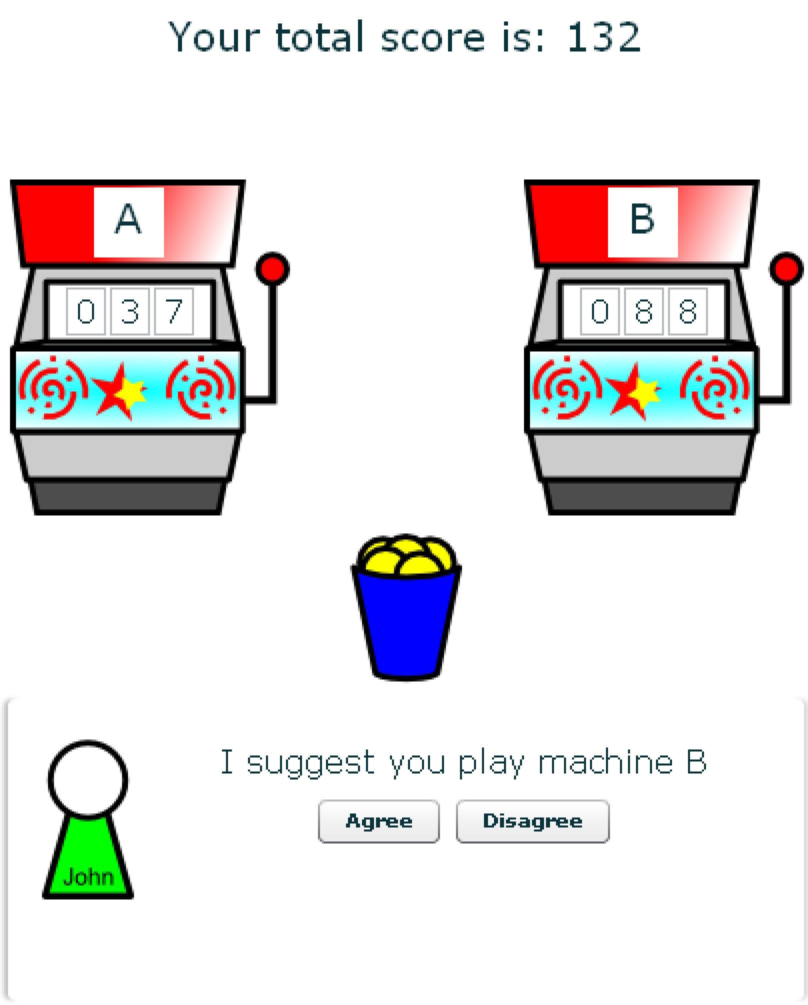
Co-Learning in MAB

- Q: how can machine help?

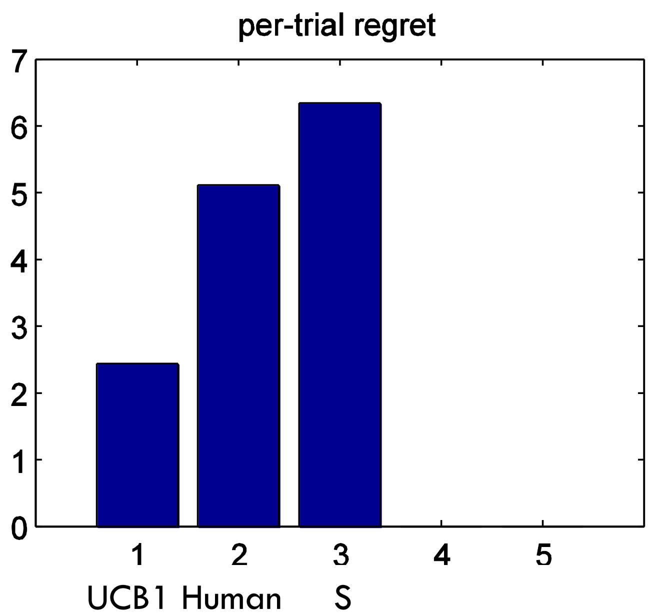


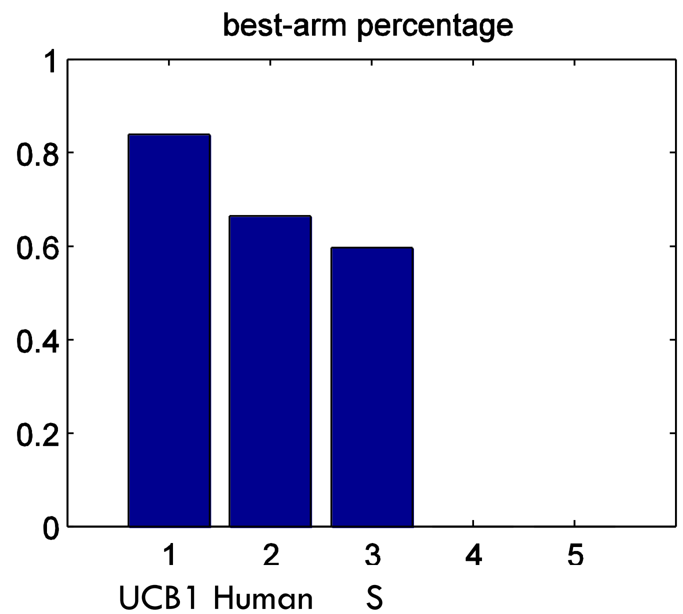
Idea 1: Giving suggestions

- Demo



Idea 1: Giving suggestions

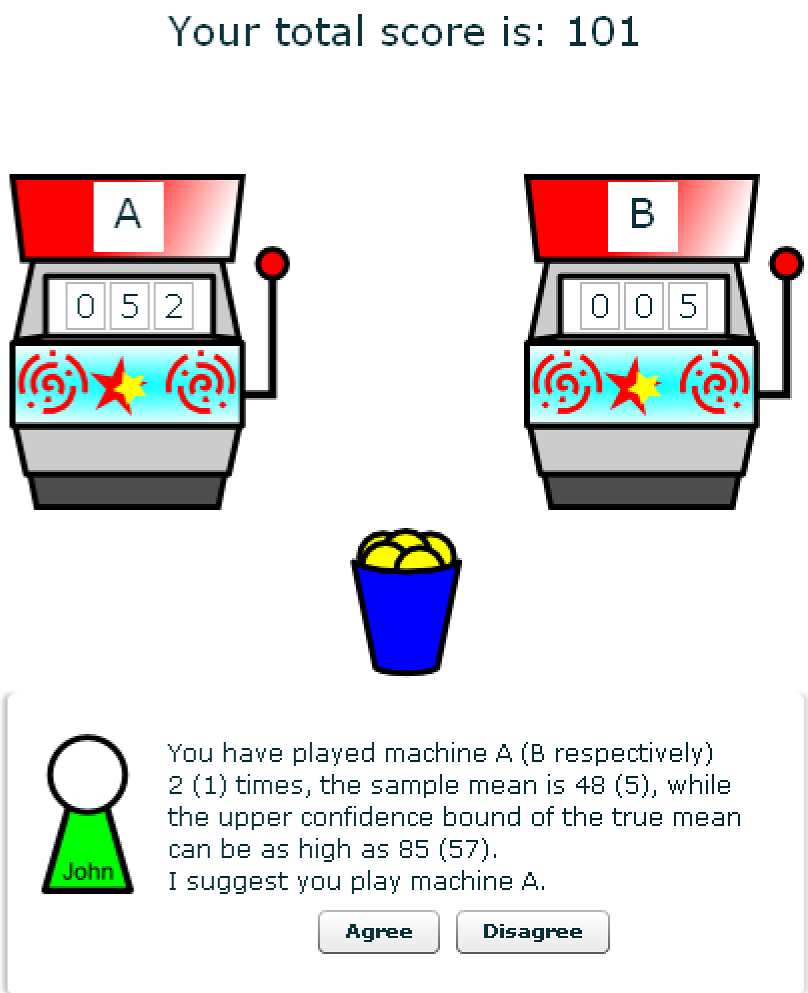




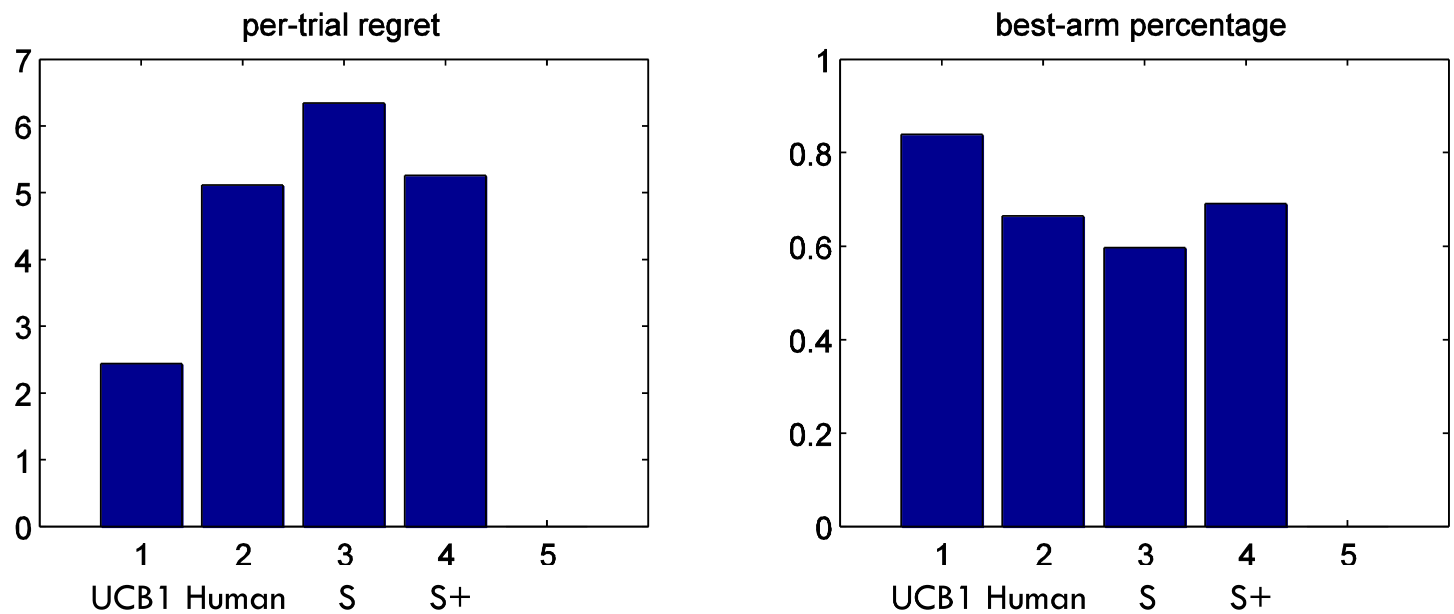
- “Human”: 28 subjects, “S”: 27 subjects

Idea 2: Giving detailed suggestions

- Demo



Idea 2: Giving detailed suggestions



- "Human": 28 subjects, "S": 27 subjects, "S+": 28 subjects

Idea 3: Reverse psychology

- Let's model humans
  - A_i: "agree" or "disagree" at iteration i
  - x_i: reward at iteration i
  - S_i: machine suggestion at iteration i

$$P(A_i \mid A_{1:i-1}, x_{1:i-1}, S_{1:i-1})$$
$$\approx P(A_i \mid A_{i-1})$$

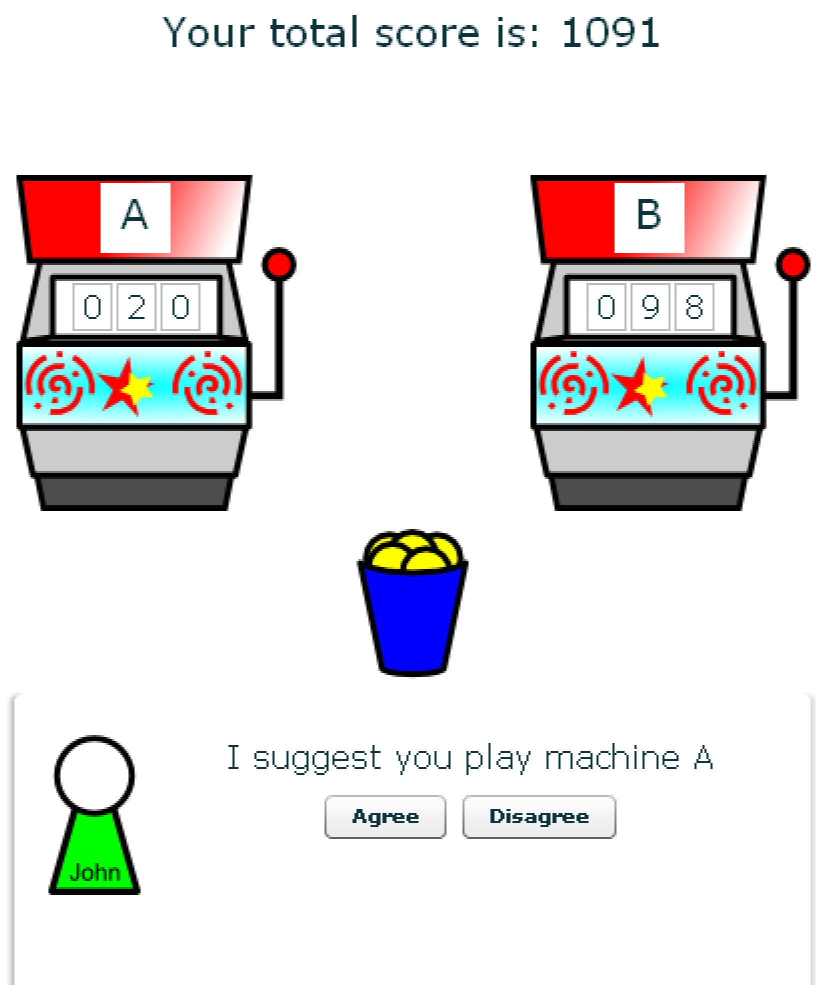
- Let M_i be the true intention of UCB1

$$S_i = \begin{cases} M_i & \text{if } P(A_i \mid A_{i-1}) \geq 1/2 \\ \neg M_i & \text{otherwise} \end{cases}$$

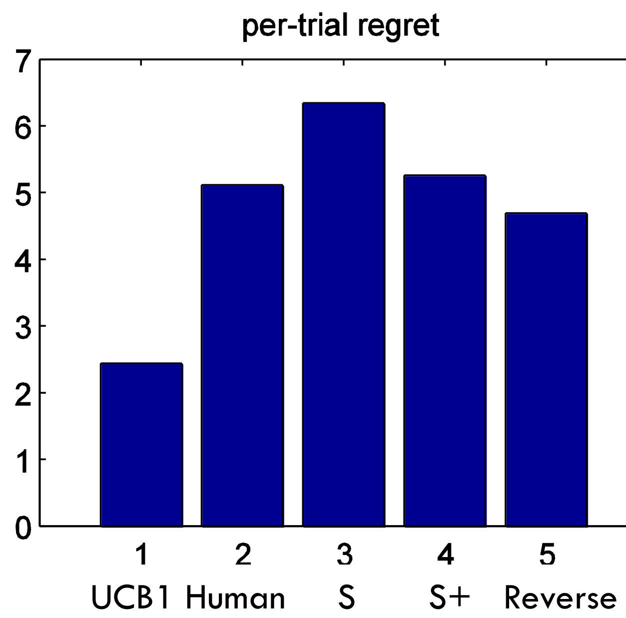
Reverse psychology

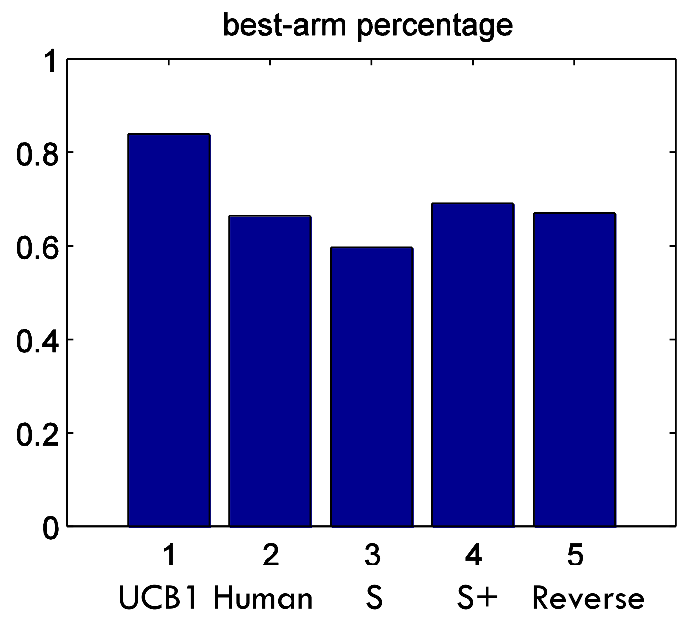
Idea 3: Reverse psychology

- Demo (always disagree)



Idea 3: Reverse psychology





- "Human": 28 subjects, "S": 27 subjects, "S+": 28 subjects; "Reverse": 29 subjects

Speculations

- Multi-Armed Bandit with trembling hands?
- RL?
- Ethics
- What if humans do better than machines?
- Synergy?