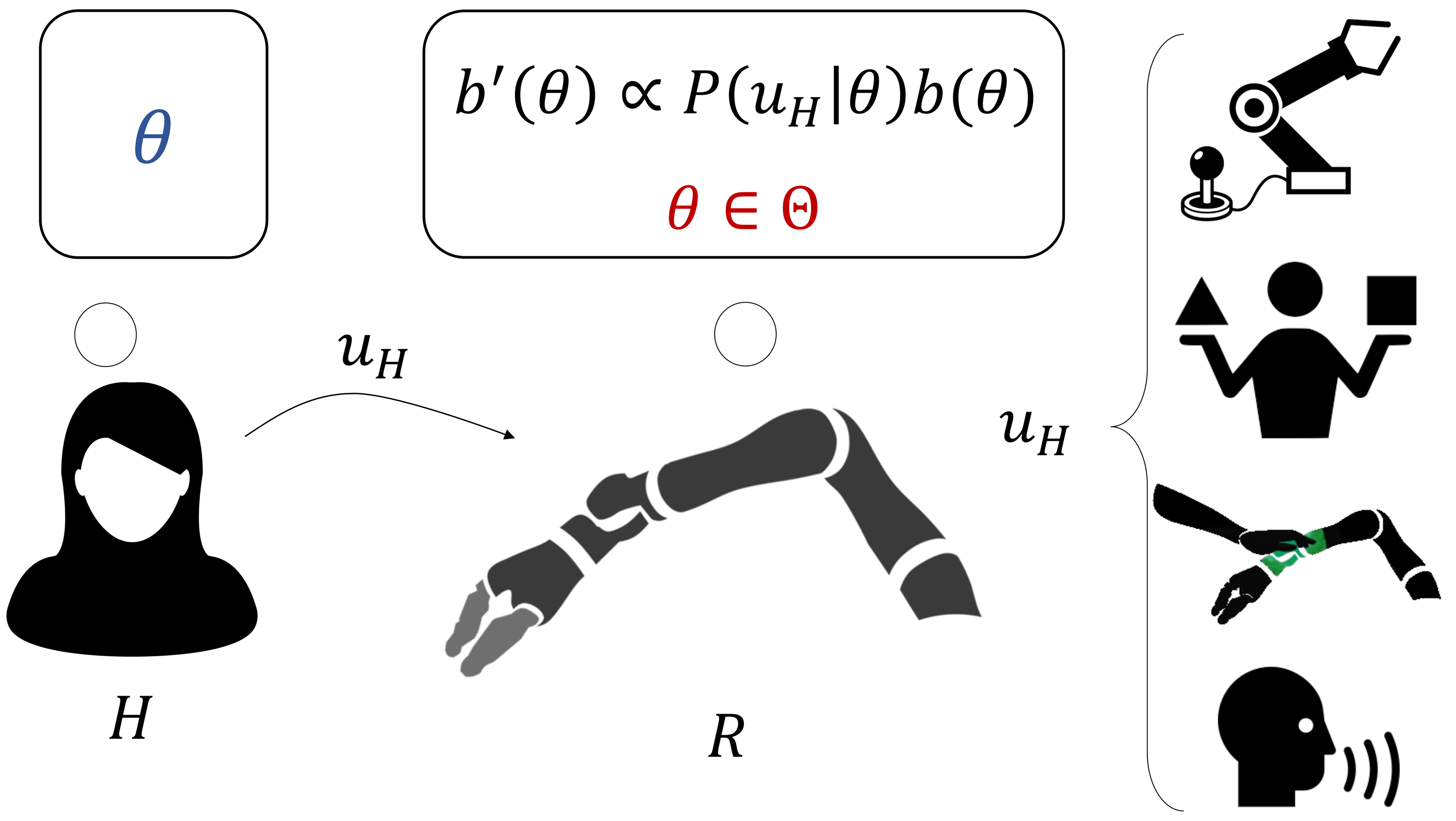
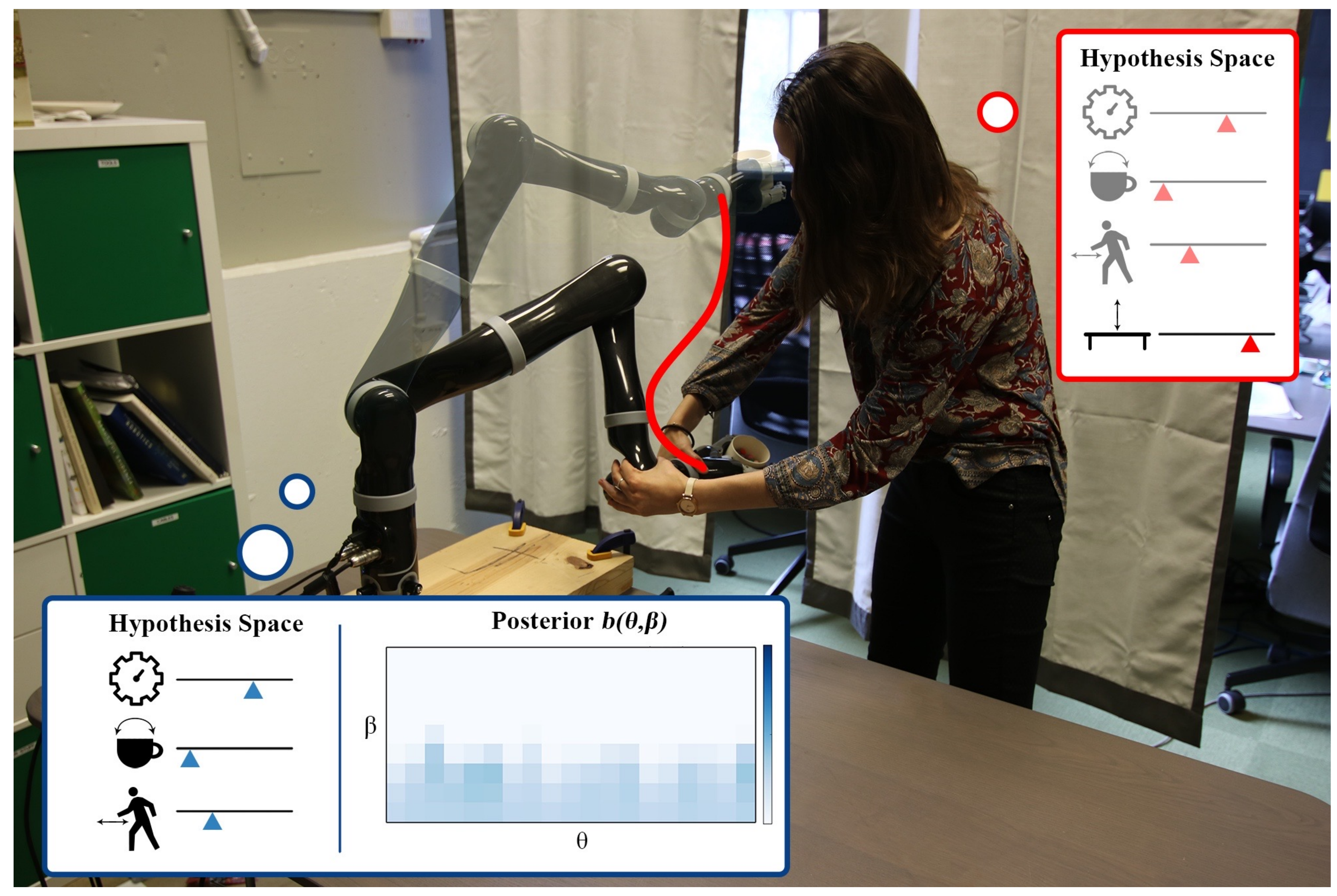
Detecting Hypothesis Space Misspecification in Robot Learning from Human Input

Andreea Bobu



Ultimate Goal

optimize what H wants: $\min_{\xi} C(\xi;\theta)$



Challenge

*What if what H wants is outside R's hypothesis space **Θ** ?*

Insight: If the human seems suboptimal for all hypotheses, chances are we don't have the right hypothesis space.

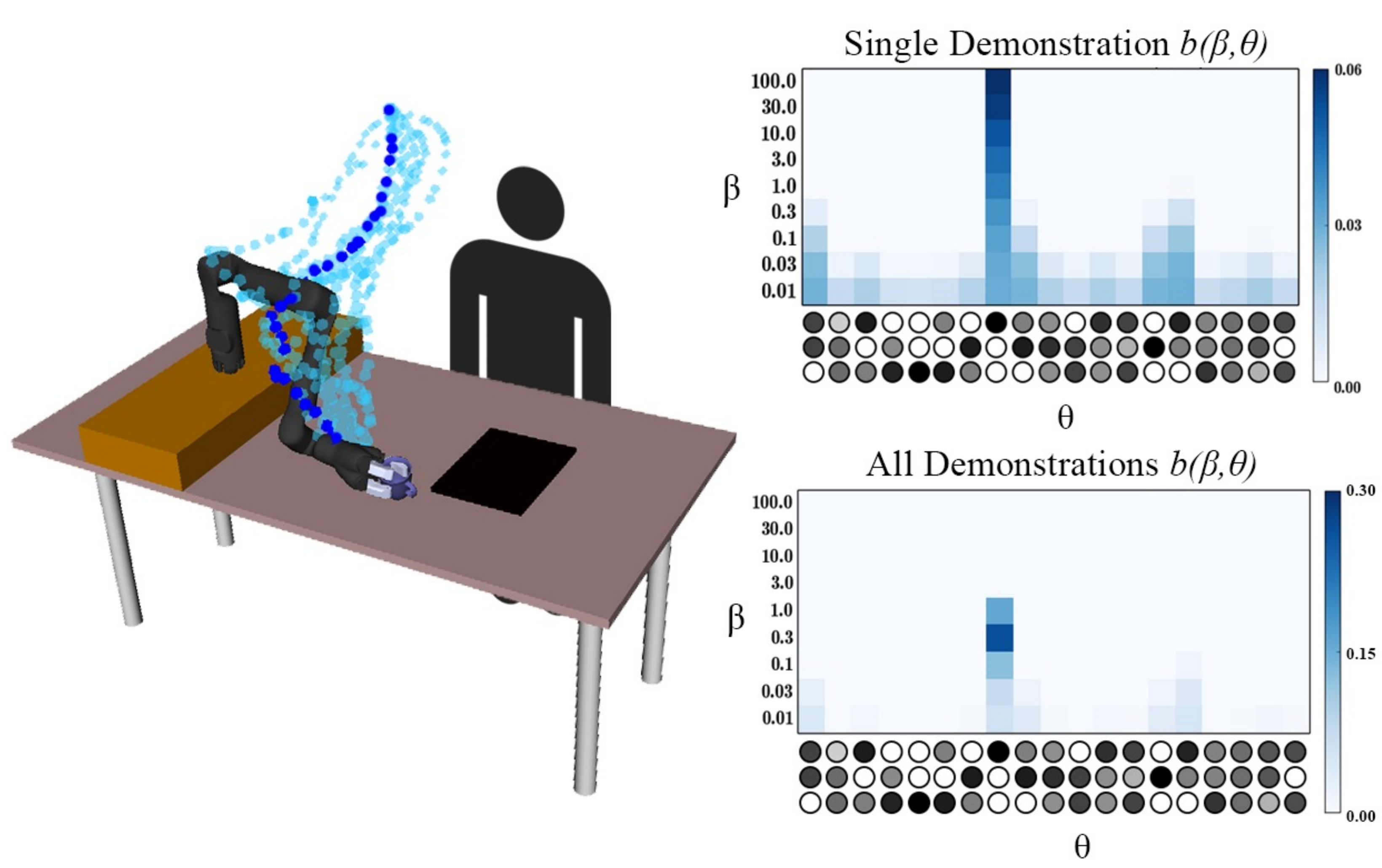
Demonstrations: Joint inference on discretized space

Demonstration Weight Confidence

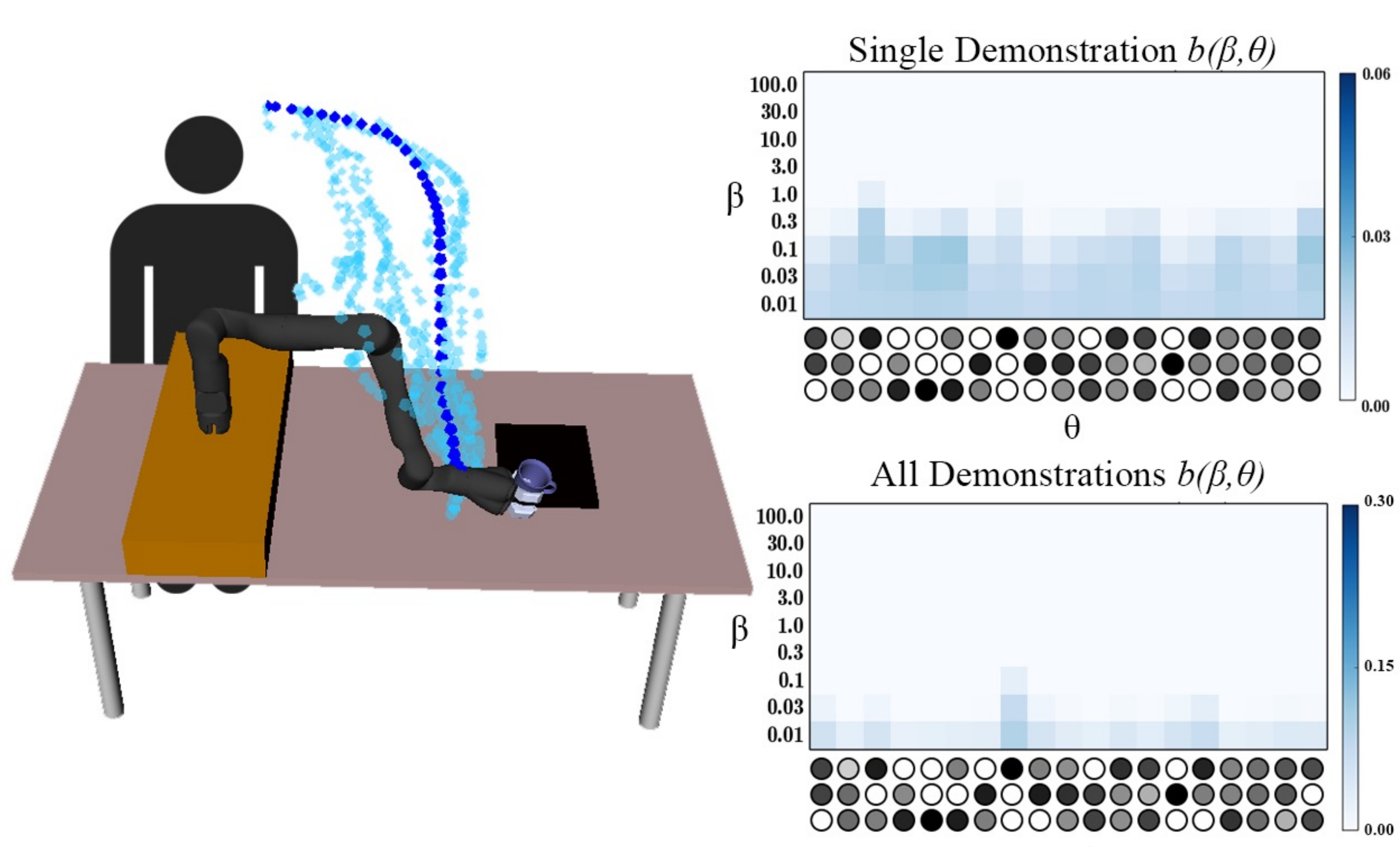
$$P(\xi_H \mid \beta, \theta) = \frac{e^{-\beta C_\theta(\xi_H)}}{\int e^{-\beta C_\theta(\bar{\xi}_H)} d\bar{\xi}_H}$$

$$b'(\beta, \theta) = \frac{P(\xi_H \mid \beta, \theta) b(\beta, \theta)}{\int P(\xi_H \mid \bar{\beta}, \bar{\theta}) b(\bar{\beta}, \bar{\theta}) d\bar{\theta} d\bar{\beta}}$$

a) Well-specified hypothesis space



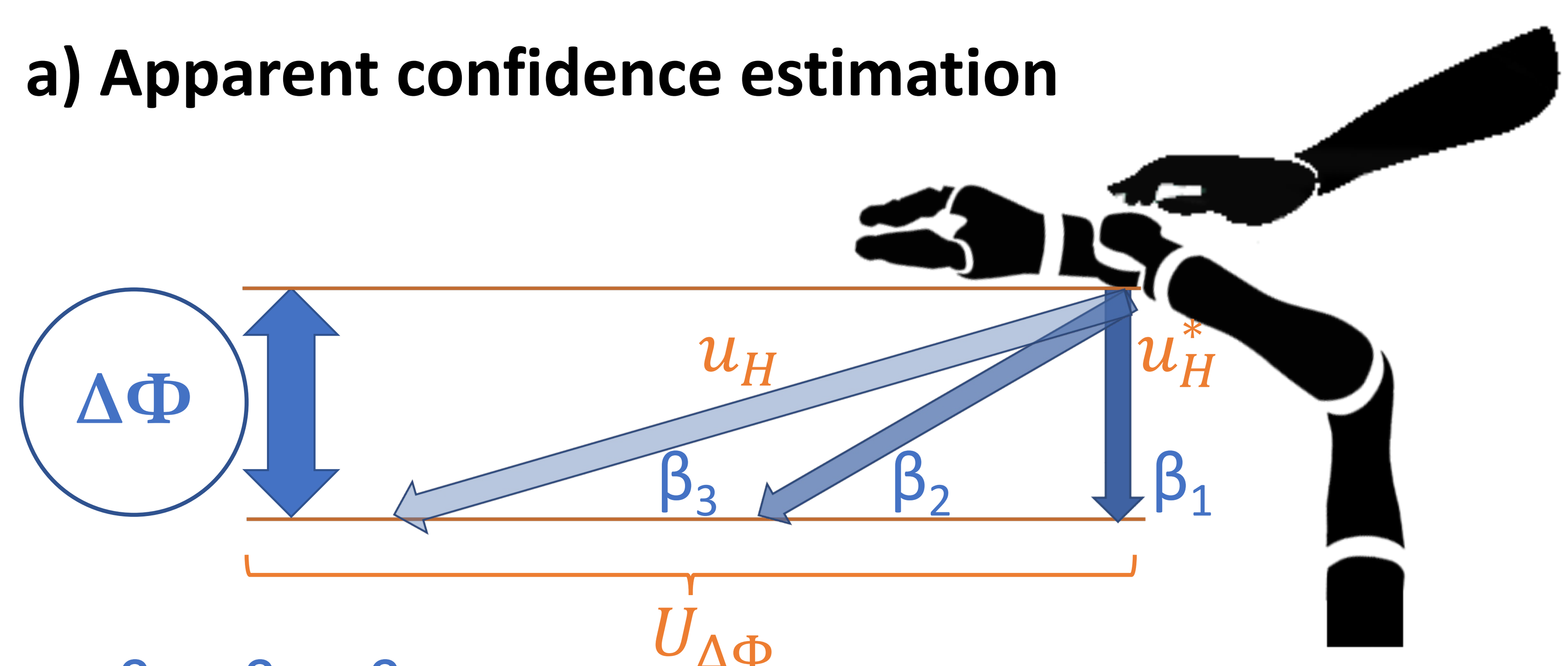
b) Misspecified hypothesis space



Physical Corrections: Real-time approximation

$$P(u_H \mid \xi_R; \beta, \theta) = \frac{e^{-\beta(\theta^T \Phi(\xi_H) + \lambda ||u_H||^2)}}{\int e^{-\beta(\theta^T \Phi(\bar{\xi}_H) + \lambda ||\bar{u}_H||^2)} d\bar{u}_H}$$

a) Apparent confidence estimation



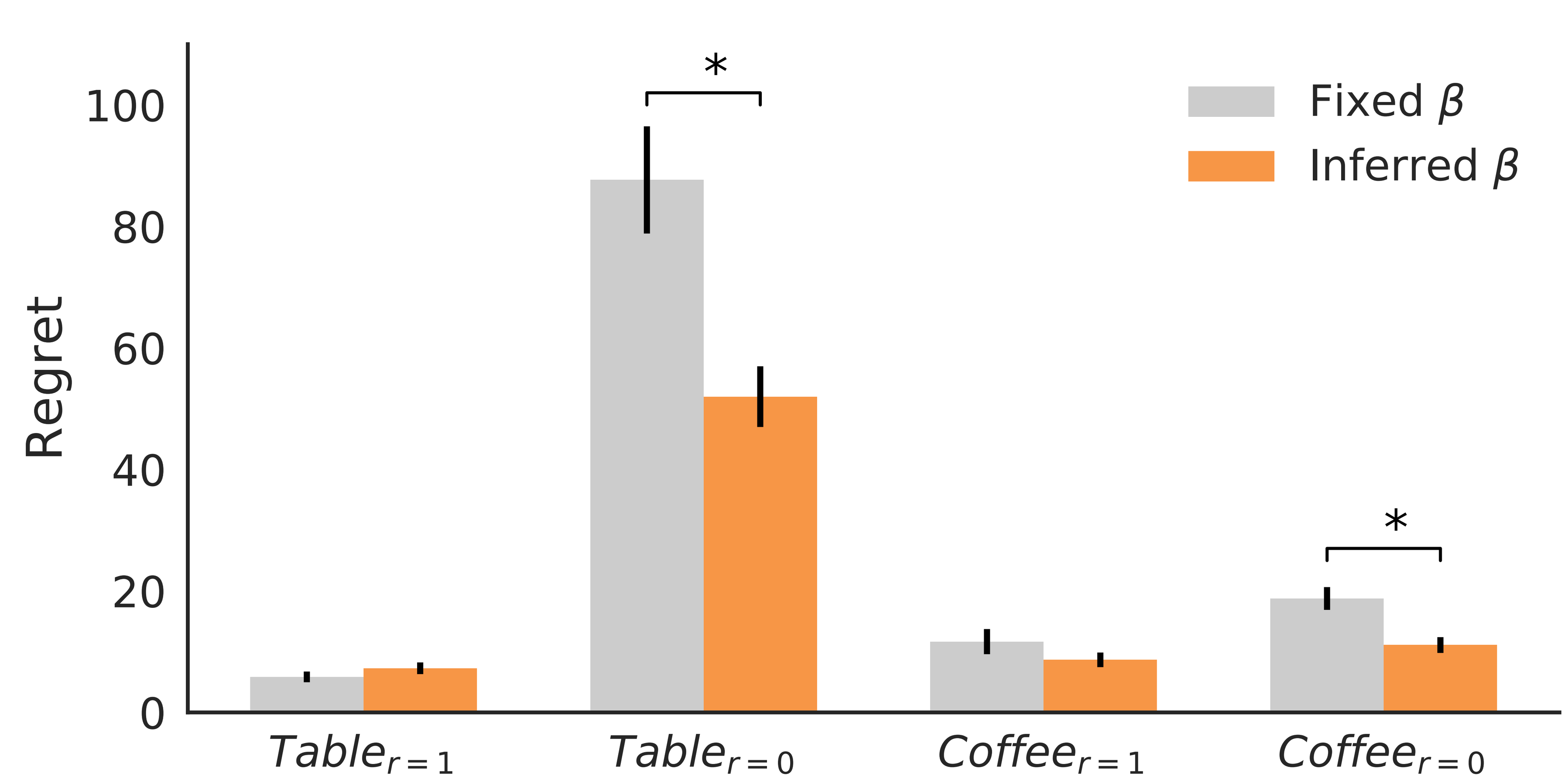
$\beta_1 > \beta_2 > \beta_3$

$||u_H^*||^2 = ||u_H^1||^2 < ||u_H^2||^2 < ||u_H^3||^2 = ||u_H||^2$

$$\hat{\beta} = \operatorname{argmax} P(u_H \mid \beta, \Phi(\xi_H), \xi_R) \approx \frac{k}{2(||u_H||^2 - ||u_H^*||^2)}$$

b) Confidence-aware approximate MAP estimate:

$$\hat{\theta}' = \hat{\theta} - \alpha f(\hat{\beta}, \hat{\theta}') (\Phi(\xi_H) - \Phi(\xi_R))$$



When misspecified (2&4), confidence-aware reduces unintended learning, while maintaining good accuracy when the hypothesis space is well-specified (1&3).