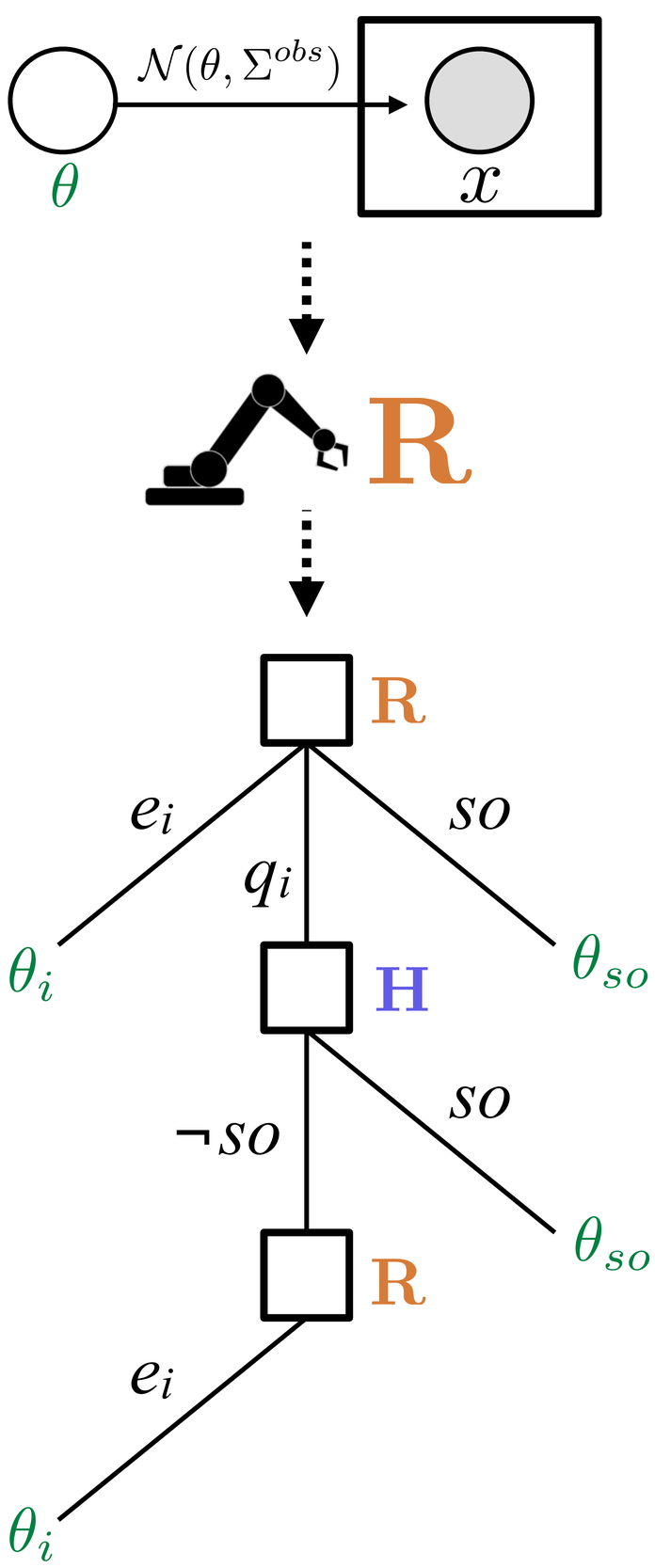
Should Robots Have Off Switches?

Smitha Milli (smilli@berkeley.edu), Dylan Hadfield-Menell, Stuart Russell

Introduction

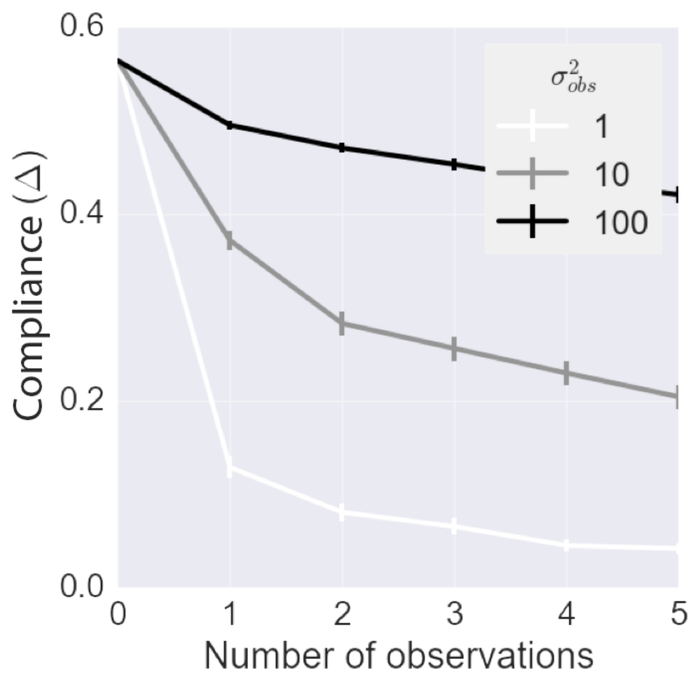
- How should a robot act after receiving orders from a human?
- Robot should be compliant when human knows more than it, previous approaches rely on robot's uncertainty to incentivize compliance
- But what if robot can learn about right thing to do, what are the effects of artificially forcing incentives for compliance?

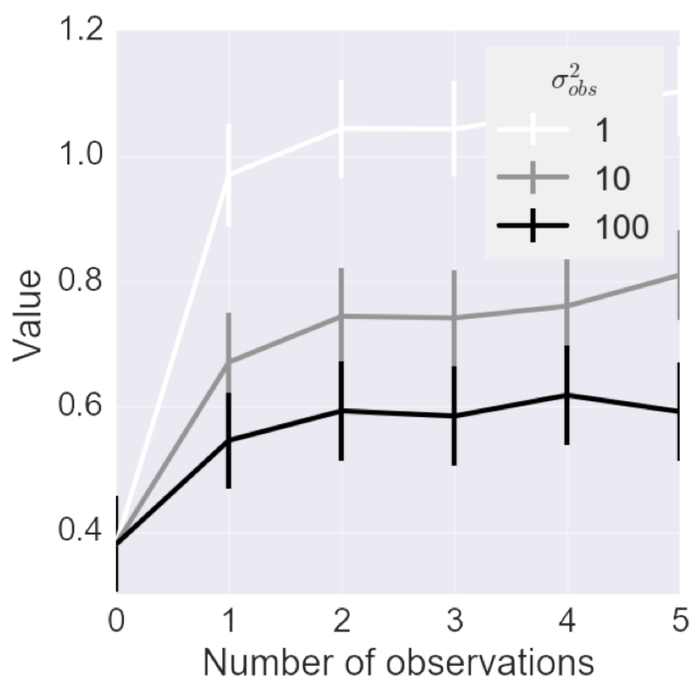
Learning in the off-switch game

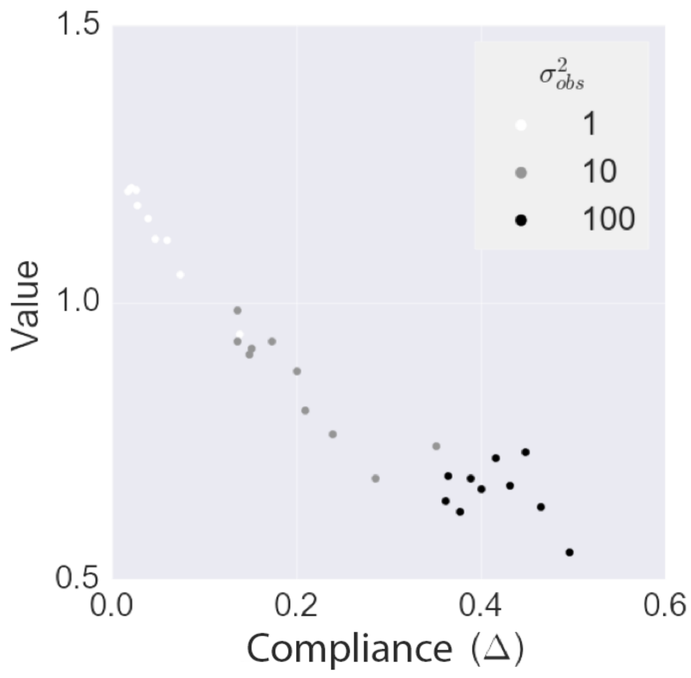


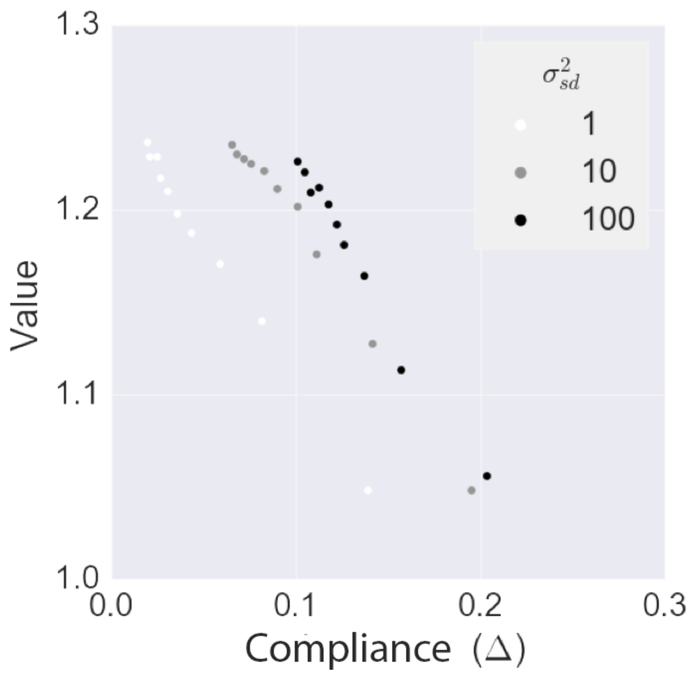
Value-Compliance Tradeoff

Compliance: $\Delta = \max_i Q^R(s_0, B^R, q_i) - \max_i Q^R(s_0, B^R, e_i)$









Conclusion

- Artificially incentivizing an agent to be compliant results in a loss in value
- Question is not "Should robots have off-switches?"
- Instead: How can we ***guarantee*** robots are compliant/not compliant when they should be?

References

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. "Cooperative Inverse Reinforcement Learning." NIPS 2016.
Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. "The Off-Switch." In Preparation, 2016 .

Berkeley
Artificial Intelligence Research Laboratory