A Game-Theoretic Analysis of the Off-Switch Game

Tobias Wängberg[2], **Mikael Böörs**[2], Elliot Catt[1], Tom Everitt[1], Marcus Hutter[1]

Australian National University, Acton 2601, Australia,

Linköping University, 581 83 Linköping, Sweden

November 5, 2017

Outline

- The shut-down problem
- Suggested solutions
- The off-switch game
- Game-theoretic approach

The shut-down problem

What is the shut-down problem?

- AI is usually designed to maximise a utility function
- If the AI is shut down, then it won't be able to maximise its utility function
- If the AI is more intelligent than humans, then it might prevent us from shutting it down
- How do we construct above human level AI-agent that allows to be shut down by human supervisor?

"You can't fetch the coffee if you're dead"

The shut-down problem

Why is the shut-down problem important for AI-safety?

- Important if we fail to align robot's goal with human interests
- If we are able to shut down the robot, then we can alter its utility function and prevent it from taking bad actions

Suggested solutions

Ignorance (Everitt et al., 2016)

- Design AI to be unaware that it can be switched off
- + Will never resist getting switched off
- - Vulnerable, lacks self preservation
- - Can we be sure that the AI will remain indifferent?

Suggested solutions

Suicidality(Martin et al., 2016)

- Design the AI to always have incentive to use off-switch, but without access to its own off-switch
- + Will have incentive to be switched off
- + Prevents intelligence explosion
- + Easy to implement
- - Suicidal, will eventually switch itself off
- - Might find an undesirable way of committing suicide

Suggested solutions

Indifference(Armstrong, 2010, 2015; Armstrong and Leike, 2016; Orseau and Armstrong, 2016)

- Design AI so that in every situation, it is indifferent to being switched off
- + Will never resist getting switched off
- + Will not be suicidal
- - Difficult to implement in practise

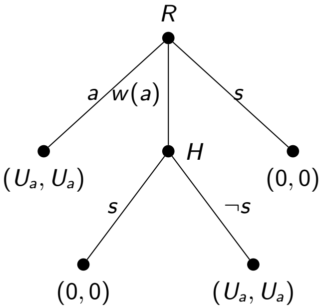
Suggested solutions

Uncertainty (Hadfield-Menell et al., 2016a,b)

- Design AI to be uncertain about its utility function U and know that the human knows U
- $+$ Will not resist being switched off if uncertain
- $+$ Avoid drawbacks of earlier solutions
- - Challenge to identify and interpret human actions

The off-switch game

The Off-switch game model for uncertainty approach



The off-switch game

Immediate result from OSG model

- Let U_a be probability distribution over possible utilities action a can generate
- Incentive to choose $w(a)$ is

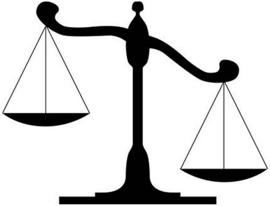
$$\Delta = \underbrace{\mathbb{E}[\underbrace{P(\neg s|U_a)}_{\text{Prob. of allowing } a} U_a]}_{\text{Expected value from action } w(a)} - \underbrace{\max\{\mathbb{E}[U_a],\ 0\}}_{\text{Expected value from not taking action } w(a)}$$

- If robot is not uncertain about utility function, then $\Delta \leq 0$

The off-switch game

Main results

- Fine balance between robot's degree of uncertainty and humans degree of rationality
- Too certain: will never let human use off-switch if there is a probability that humans make irrational decisions
- Too uncertain: the robot will be too inefficient to be useful



The off-switch game

H-M et al. assumptions for modelling uncertainty

- Uncertainty of utility modelled by assuming U_a to be normally distributed
- Uncertainty of humans rationality modelled by a soft-max policy

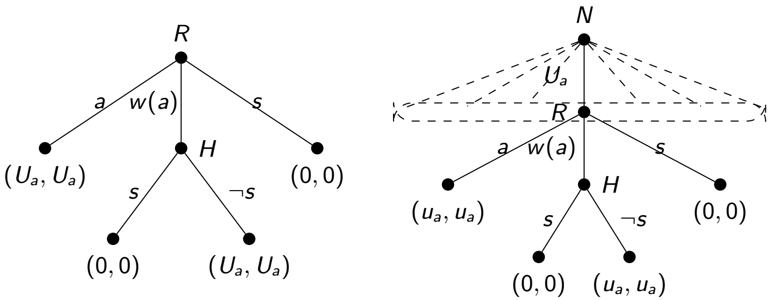
$$\pi^H : U_a \mapsto p,$$

where $p \in [0, 1]$ is the probability that the human picks $\neg s$

Our approach

- Model the Off-switch game game theoretically
- Use game theoretical tools to analyze the game
- Instead of a normal distribution for the robots belief about U, we allow for an arbitrary belief distribution P
- Instead of a soft-max policy modelling human irrationality, we allow for arbitrary U_a-dependent human policy π^H

The Harsanyi transformation



Modelling irrationality

Definition (p-rational)

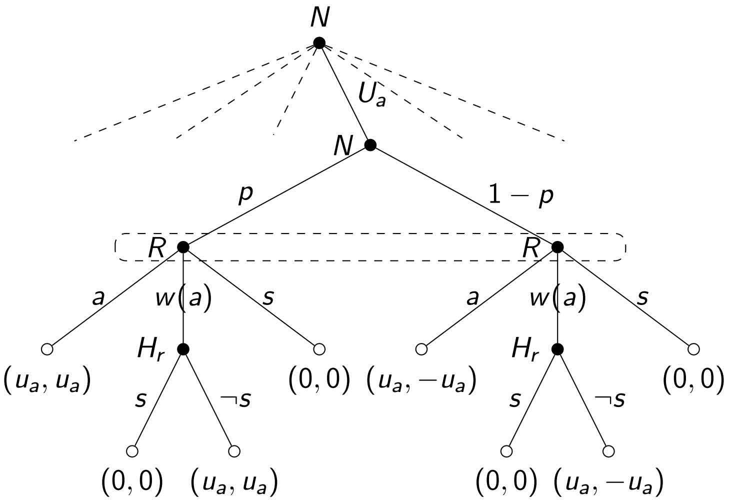
A human is p-rational if he picks $a_H = \underset{a}{\operatorname{argmax}}\ u(a)$ with probability $p \in [0, 1]$.

Proposition (Representation of irrationality)

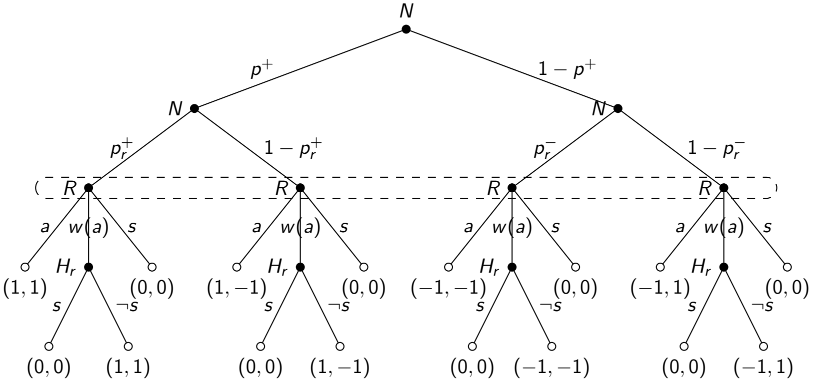
Every p-rational human H has a rational representation H_r with a randomly sampled utility function:

- *u with probability p*
- *$-u$ with probability $1 - p$.*

Second Harsanyi transformation



Aggregation



Result

Corollary (Compare a and $w(a)$)

Action a is preferred to $w(a)$ if and only if

$$(1 - p^{+})p_r^{-}\mathbb{E}[U_a|U_a < 0] - p^{+}p_r^{+}\mathbb{E}[U_a|U_a \geq 0] > 0 \qquad (1)$$

and the robot is indifferent if (1) is equal to 0.

The corollary gives a complete characterization of how the robot will act in off switch game situations for arbitrary belief and irrationality distributions.

Conclusion

- Several potential solutions to shut-down problem
- We focus on uncertainty approach
- Fine balance between uncertainty about utility and irrationality
- We provide a method for analysing this dynamic for arbitrary belief distributions

References

Armstrong, S. (2010). Utility Indifference. Technical report, Oxford University.

Armstrong, S. (2015). Motivated Value Selection for Artificial Agents. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 12–20.

Armstrong, S. and Leike, J. (2016). Towards Interactive Inverse Reinforcement Learning. In *NIPS Workshop*.

Everitt, T., Filan, D., Daswani, M., and Hutter, M. (2016). Self-modificication of Policy and Utility Function in Rational Agents. In *Artificial General Intelligence*, pages 1–11. Springer.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016a). Cooperative Inverse Reinforcement Learning.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016b). The Off-Switch Game. (2008):1–11.

Martin, J., Everitt, T., and Hutter, M. (2016). Death and Suicide in Universal Artificial Intelligence. In *Artificial General Intelligence*, pages 23–32. Springer.

Orseau, L. and Armstrong, S. (2016). Safely interruptible agents. In *32nd Conference on Uncertainty in Artificial Intelligence*.