Using argumentation to design reward functions

Rémy Chaput

2023/03/16

The Value Connection Workshop

Desiderata for value implementation

- Choice of values to embed in AI systems should be discussed with a large audience (designers, domain experts, users, regulators, interested parties, etc.)
- How to define these values? (cultural differences)
- Needs a pluri-disciplinar discussion
  - AI experts
  - Domain experts (e.g., Smart Grids, Transport, etc.)
  - Moral philosophers
  - Law experts
  - End users
  - ...

Reinforcement Learning

- Reinforcement learning can be used to learn (ethical) behaviours
- Interaction loop: Observations - Actions - Rewards
- Reward function used as a signal to encourage/discourage specific behaviours
- ⇒ Values can be embedded in the reward function (objectives to satisfy)

Problems

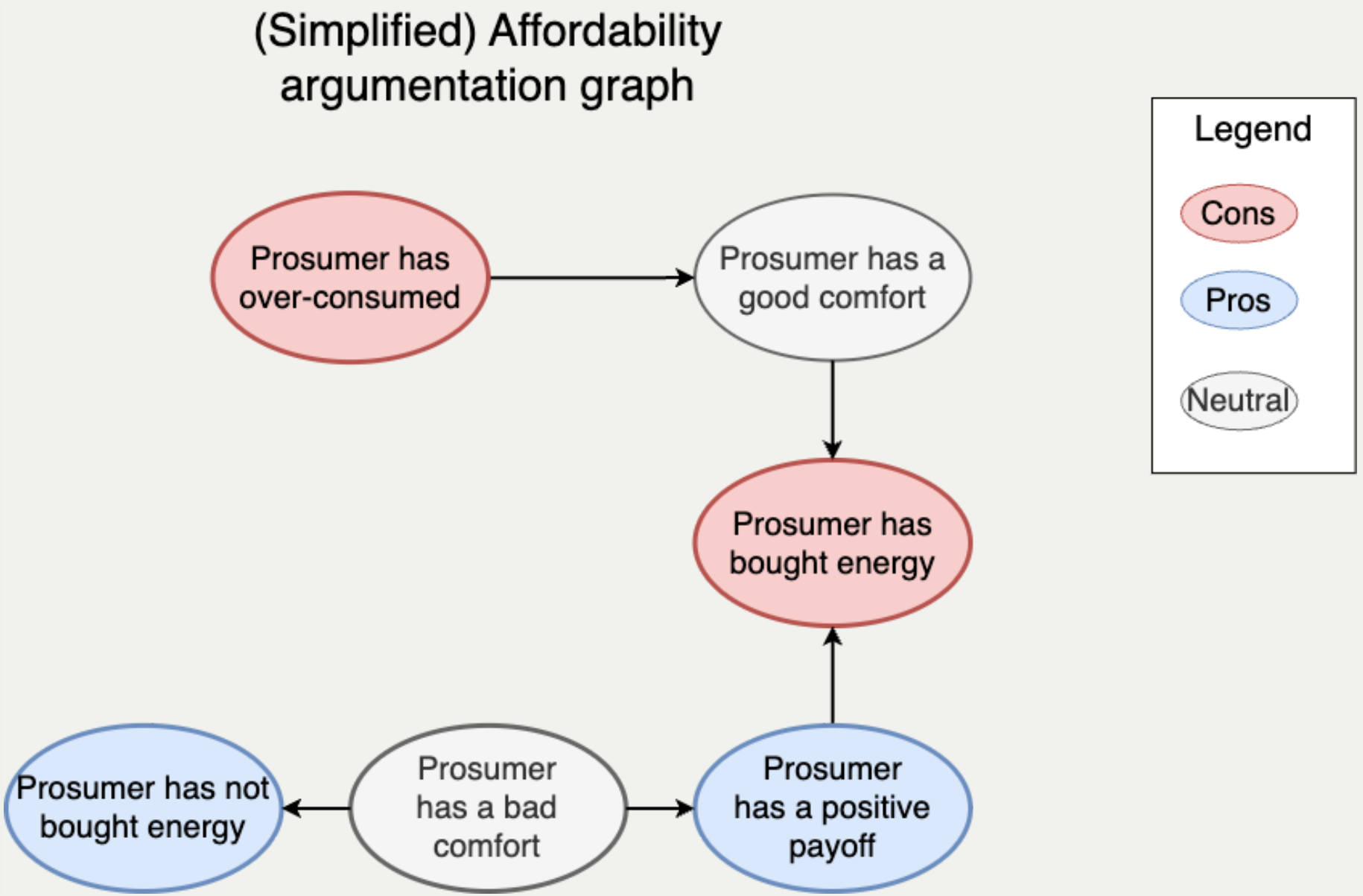
- Traditionally, reward function is a mathematical function
  - ⇒ Can be hard to discuss with a pluri-disciplinar audience
- We want to update the reward function
  - Because of the *reward gaming problem*[1]
  - Because ethics are not fixed: our society evolves and the ethical consensus follows

1. Also called the *King Midas* problem: getting exactly what you asked ... not what you desired

Using argumentation for reward functions

- Proposition: design reward functions through *symbolic* methods
  - Such as argumentation frameworks
- Advantages
  - Argumentation coincides with a form of human cognition and reasoning
  - Richer structure, allowing explicit conflicts ⇒ can mitigate *reward gaming*
  - Graphical representation, easier to grasp what the function is doing
  - (Learnt behaviour can be explained through activations of arguments)

Example of argumentation-based reward function



Remaining questions

- How can argumentation-based reward functions be integrated in the value design process?
- Which argumentation frameworks are the most appropriate for embedding values?
  - Many exist: abstract, weighted, structured, ...
- How can arguments be leveraged to explain the learnt behaviour?
- Can all moral values be implemented using argumentation frameworks?
  - Probably not ⇒ how can we mitigate this?