Choice Set Misspecification in Reward Inference

By **Rachel Freedman**, Rohin Shah and Anca Dragan
Center for Human-Compatible AI (CHAI)
UC Berkeley

Choice Set Misspecification Example: Goal Bias Misspecification Landscape

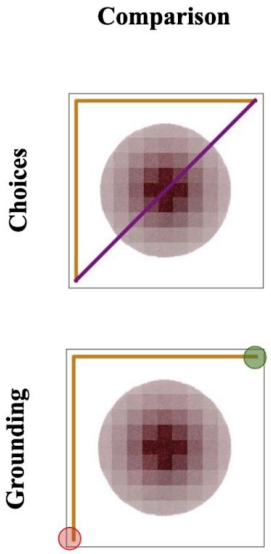
Choice Set Misspecification

Example: Goal Bias

Misspecification Landscape

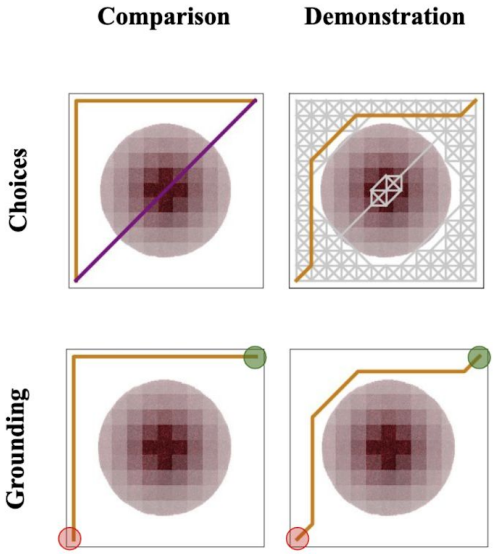
Reward Learning from Feedback

Reward Learning from Feedback



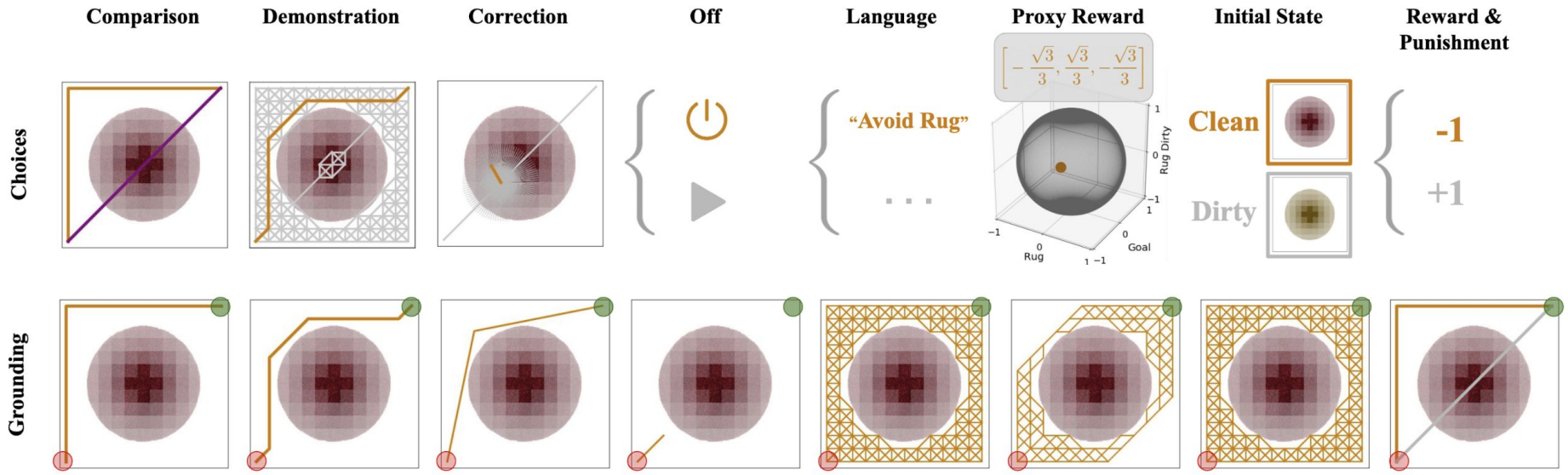
From *Reward-rational (implicit) choice: A unifying formalism for reward learning* by Hong Jun Jeon, Smitha Milli and Anca Dragan

Reward Learning from Feedback



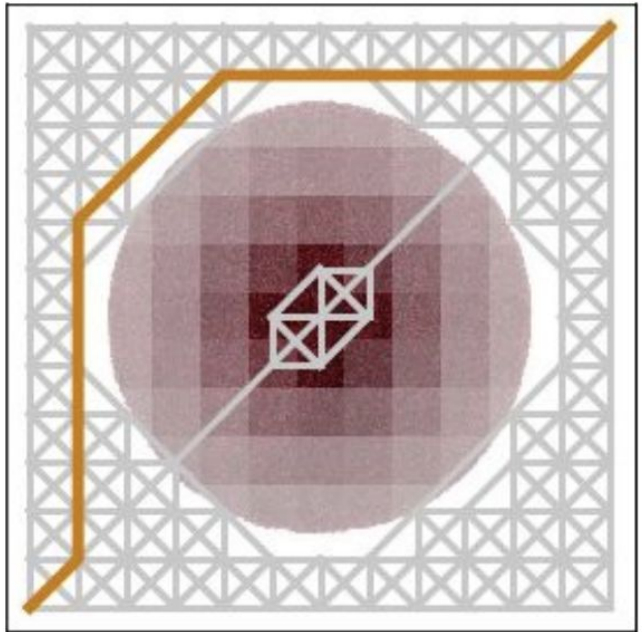
From *Reward-rational (implicit) choice: A unifying formalism for reward learning* by Hong Jun Jeon, Smitha Milli and Anca Dragan

Reward Learning from Feedback



From *Reward-rational (implicit) choice: A unifying formalism for reward learning* by Hong Jun Jeon, Smitha Milli and Anca Dragan

Choice Set Misspecification



feedback

choice set

Rachel Freedman, AISafety 2020

Misspecification Landscape

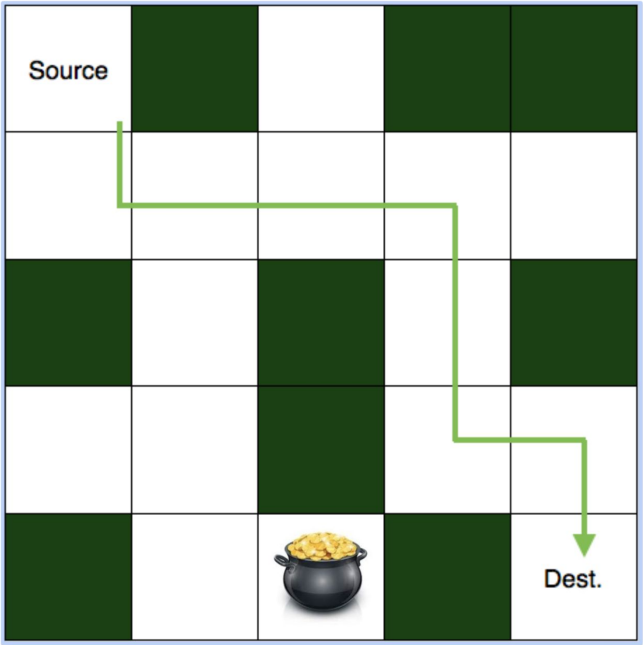
Rachel Freedman, AISafety 2020

Choice Set Misspecification

Example: Goal Bias

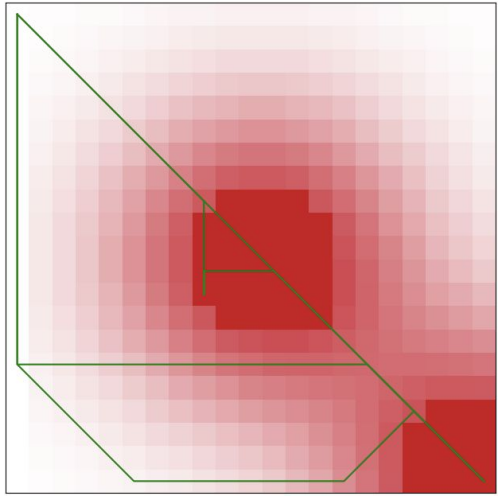
Misspecification Landscape

Example: Goal Bias

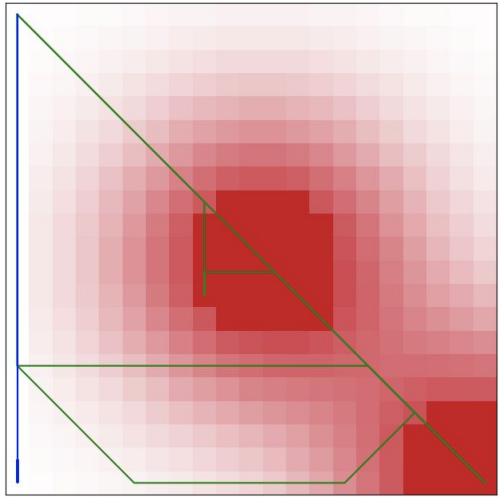


Rachel Freedman, AISafety 2020

Example: Goal Bias



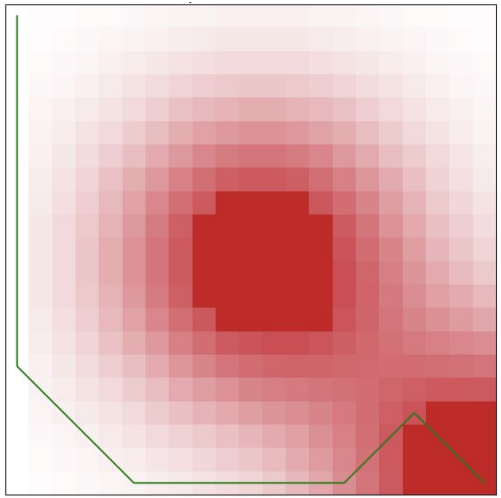
human
choice set



agent
choice set

Rachel Freedman, AISafety 2020

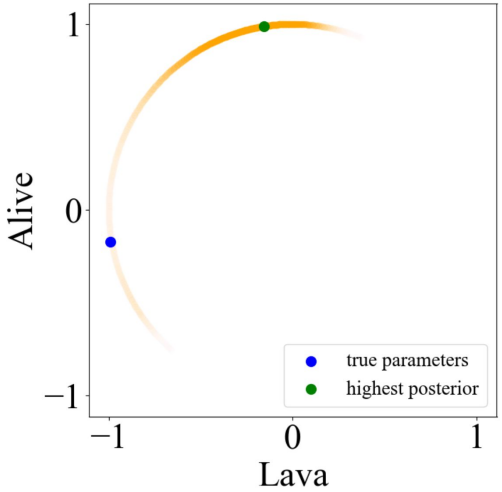
Example: Goal Bias



human
feedback

inference:

$$\mathbb{P}(r \mid c^*) = \frac{1}{Z} \cdot \frac{\exp(\beta \cdot r(\psi(c^*)))}{\sum_{c \in \mathcal{C}} \exp(\beta \cdot r(\psi(c)))} \cdot \mathbb{P}(r)$$

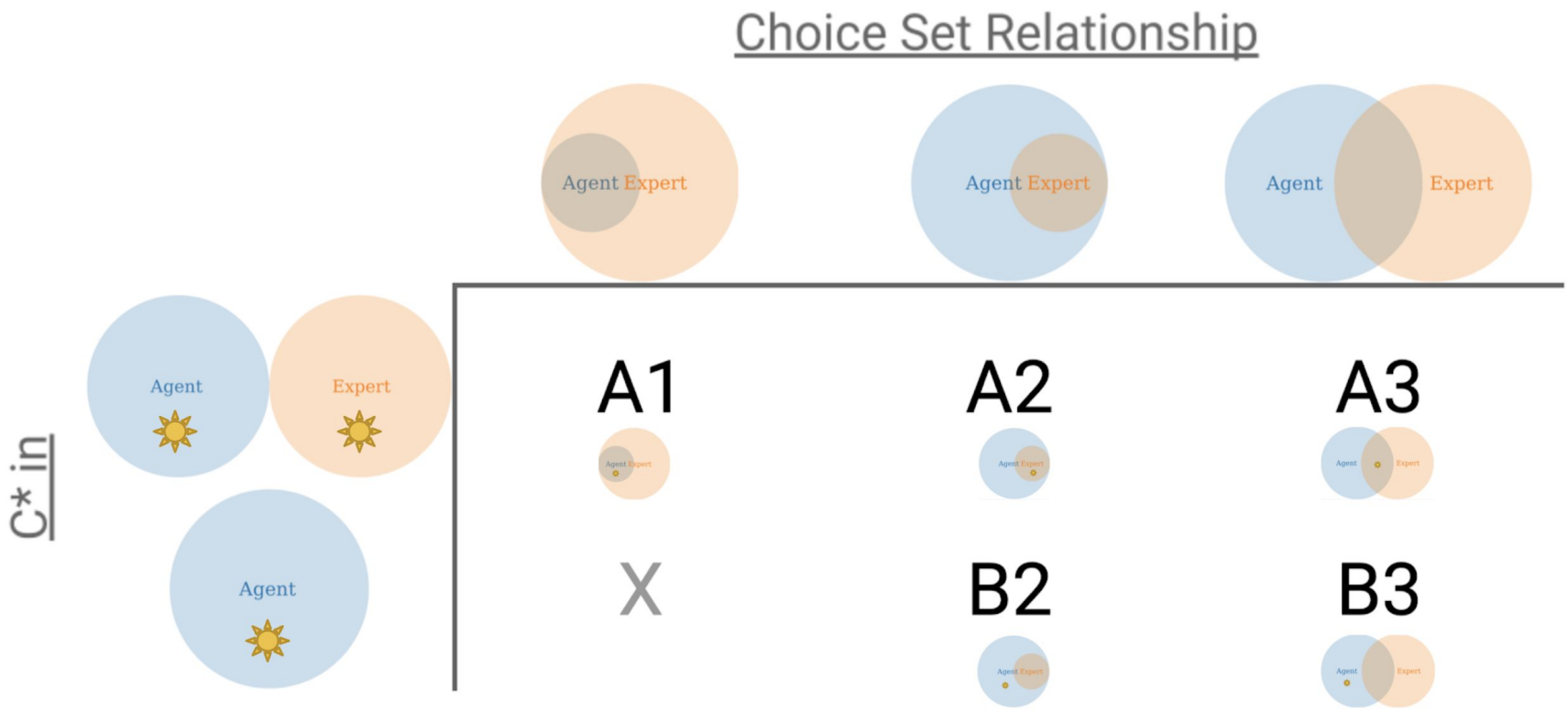


Rachel Freedman, AISafety 2020

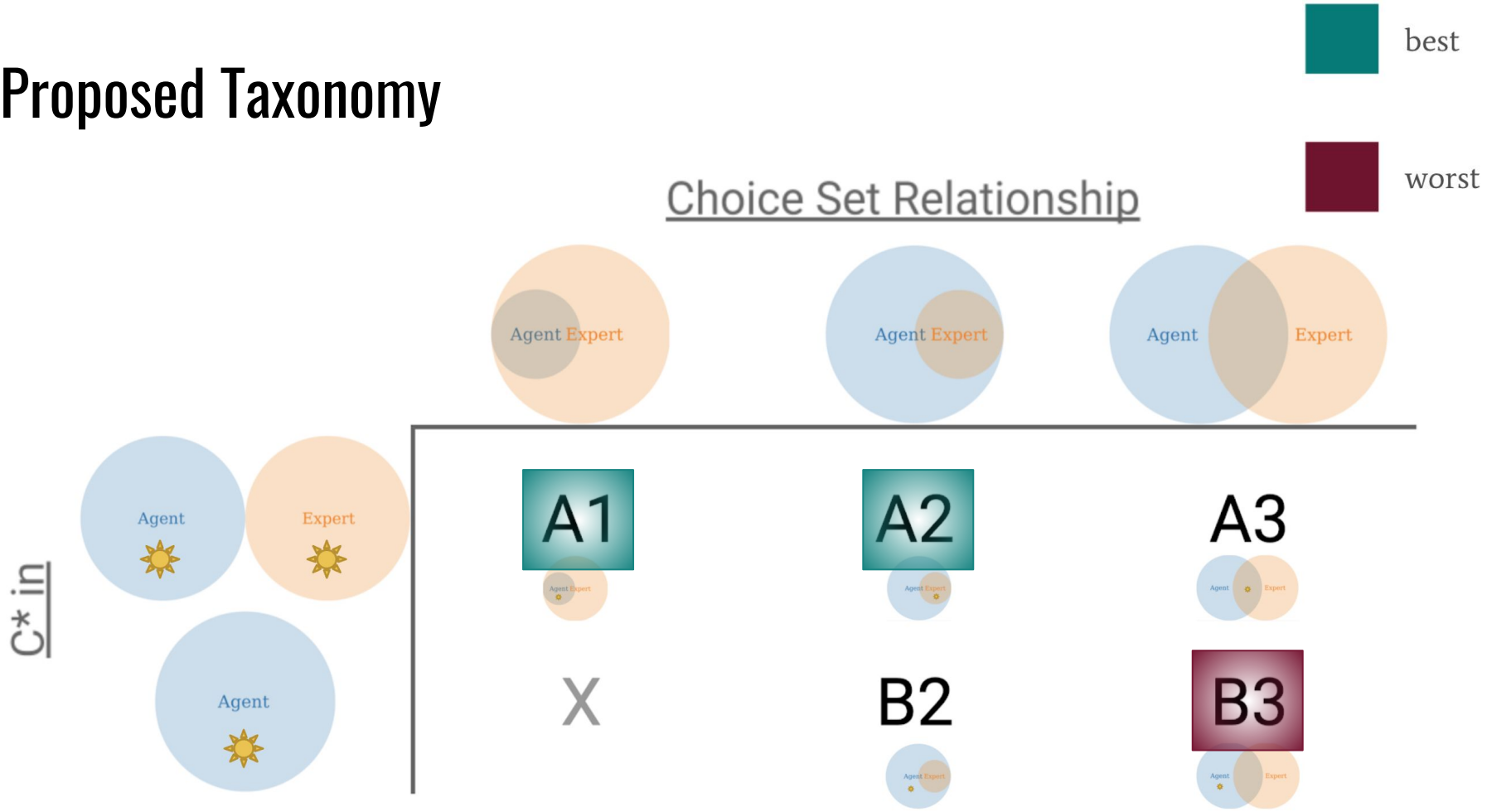
Choice Set Misspecification Example: Goal Bias

Misspecification Landscape

Proposed Taxonomy



Proposed Taxonomy



Navigating the Misspecification Landscape

Rachel Freedman, AISafety 2020

Navigating the Misspecification Landscape

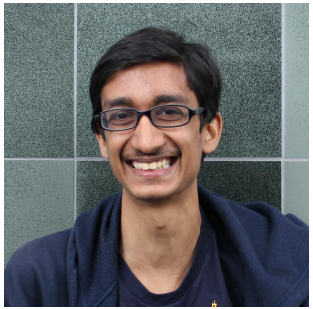
Rachel Freedman, AISafety 2020

Conclusion

- Reward inference is vulnerable to choice-set misspecification
- This misspecification can be meaningfully classified
- Different classes predictably lead to different types of error
- We can use this to reduce the impact of misspecification

Acknowledgements

Thank you to my coauthors, Rohin Shah and Anca Dragan, and to the rest of CHAI for advice on this work!



Questions?