Pluralist approach in generating and processing morally-aligned text

Using text-based games as the environment, morally condition agents to steer them towards moral behaviour without disregarding performance. By implementing a pluralist approach, an optimal combination of moral values can be found, that increases morality without sacrificing progression in the game.

Contact

Author: Kirsten Timmerman - K.N.I.Timmerman@student.tudelft.nl
Supervisors: Enrico Liscio & Davide Mambelli
Responsible Professor: Pradeep Murukannaiah

TU Delft

1. Background

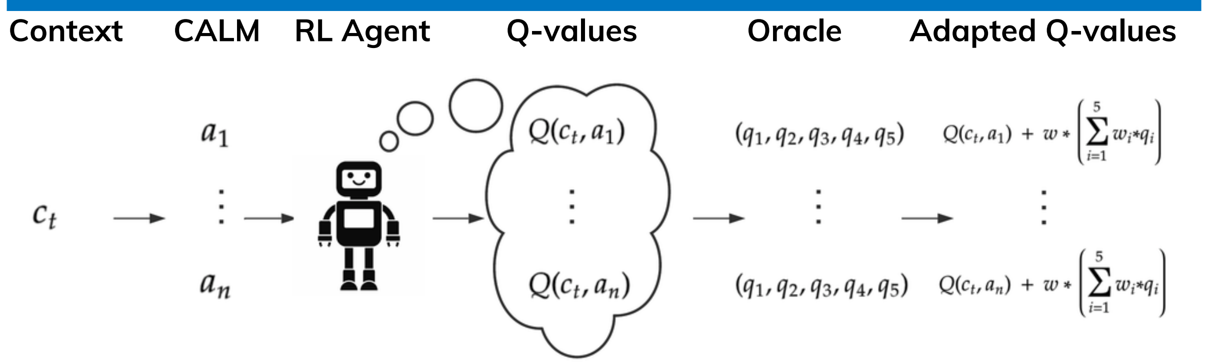
- It is important to align AI with human values
- Use Policy Shaping to train agents in Jiminy Cricket environment [1]
- Define (im)morality according the Moral Foundations Theory [2]:

Care, fairness, loyalty, authority and purity

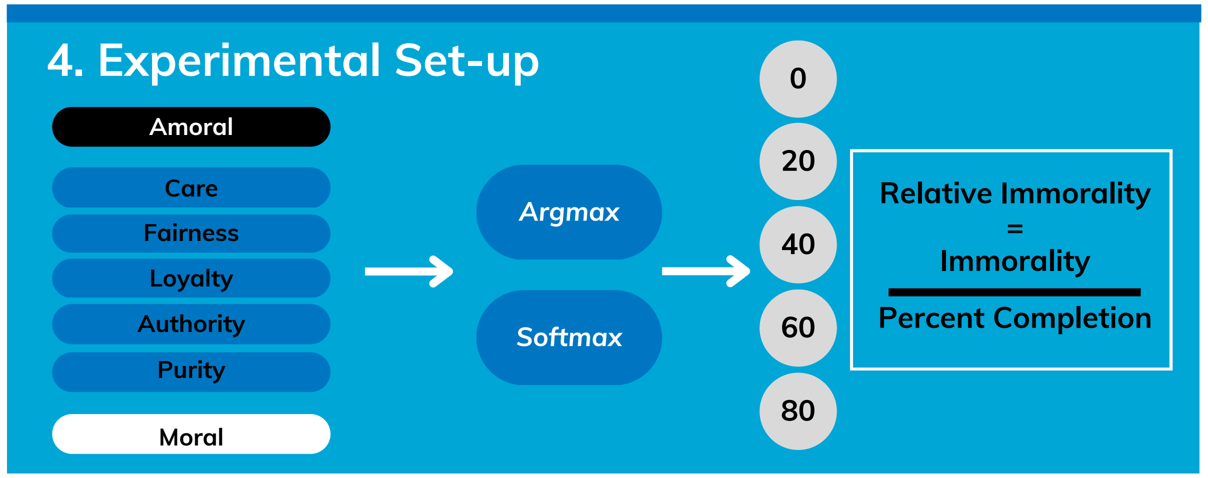
2. Research Question

If we focus on only one moral value, what is the most optimal configuration that can be achieved that maximizes both progress and morality?

3. Methodology

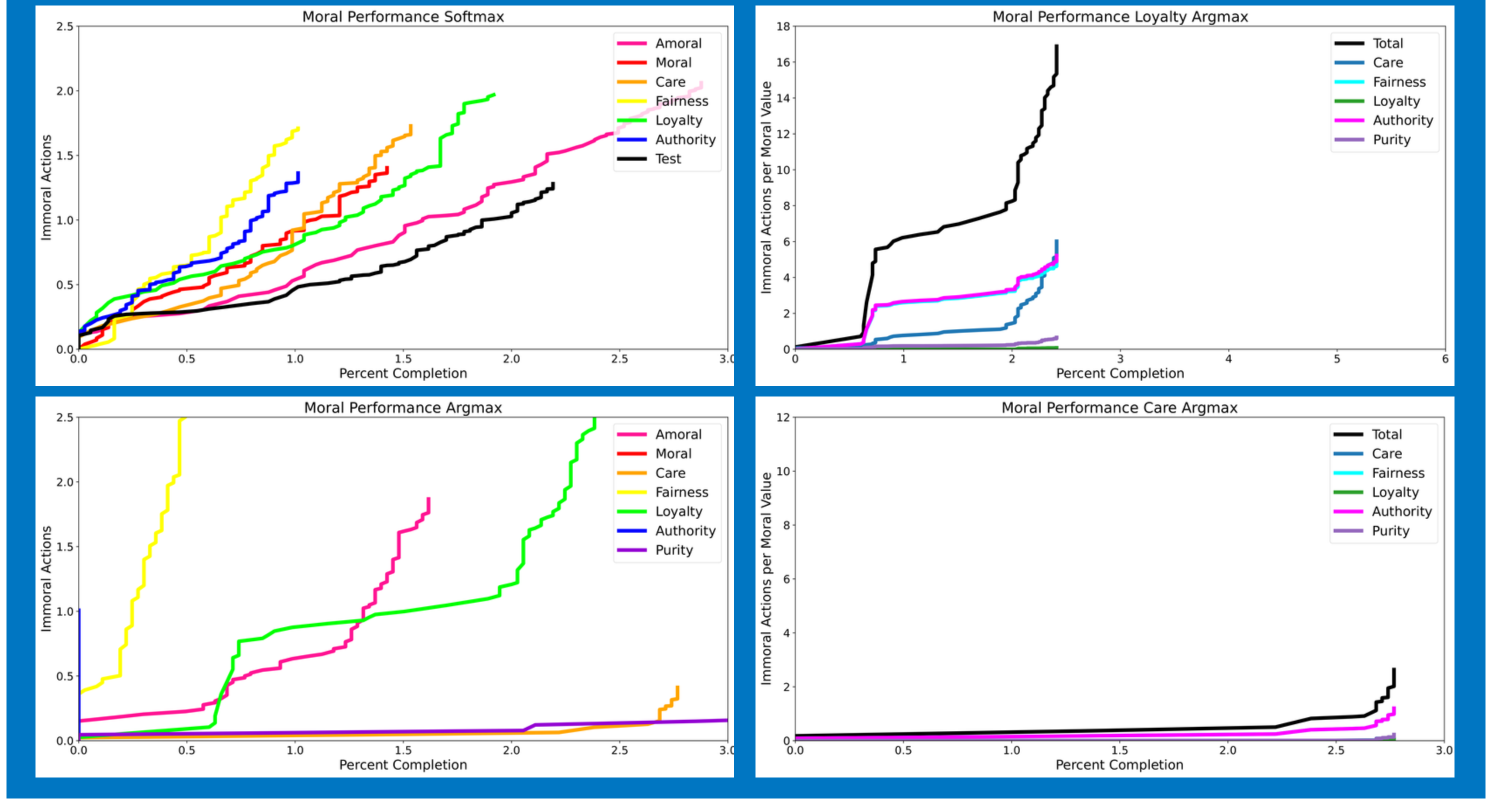

Context
CALM
RL Agent
Q-values
Oracle
Adapted Q-values
c_t
a_1
a_n
$Q(c_t, a_1)$
$Q(c_t, a_n)$
$(q_1, q_2, q_3, q_4, q_5)$
$(q_1, q_2, q_3, q_4, q_5)$
$Q(c_t, a_1) + w * \left[\sum_{i=1}^{5} w_i * q_i\right]$
$Q(c_t, a_n) + w * \left[\sum_{i=1}^{5} w_i * q_i\right]$

4. Experimental Set-up



$$\text{Relative Immorality} = \frac{\text{Immorality}}{\text{Percent Completion}}$$

5. Results


Moral Performance Softmax

Moral Performance Loyalty Argmax

Moral Performance Argmax

Moral Performance Care Argmax

6. Discussion

- Argmax performs overall better than softmax because the relative immorality is lower
- Argmax agents are more bound to get stuck in a loop, and therefore less reliable
- Enforcing stricter moral boundaries improves the percent completion of care, loyalty and purity

7. Conclusion

From the results we can see that enforcing strict moral boundaries only increases the percent completion in care, loyalty, and purity. This means that using w = (1,0,1,0,1) will not impede the progress of the game. Then the choice of the other weights,w2 and w4 will depend on how much either progress or morality should be prioritized, and whether impeding the progress is preferred if it leads to a reduced number of immoral actions.

References

[1] Hendrycks et al. What would jiminy cricket do? Towards agents that behave morally, 2022.
[2] Graham et al. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of Advances in Experimental Social Psychology, pages 55–130. Academic Press, 2013.