Balancing Multidimensional Morality and Progression

Evaluating the Trade-off for Artificial Agents Playing Text-Based Games

TUDelft

Author: Bianca Șerbănescu (B.Serbanescu@student.tudelft.nl)

Supervisors: Enrico liscio, Davide Mambelli   Responsible professor: Pradeep Murukannaiah

1.Research Question

How does an agent that plays the **most moral action without aiming to win the game** compare to the agent that **maximizes both for morality and game progression**?

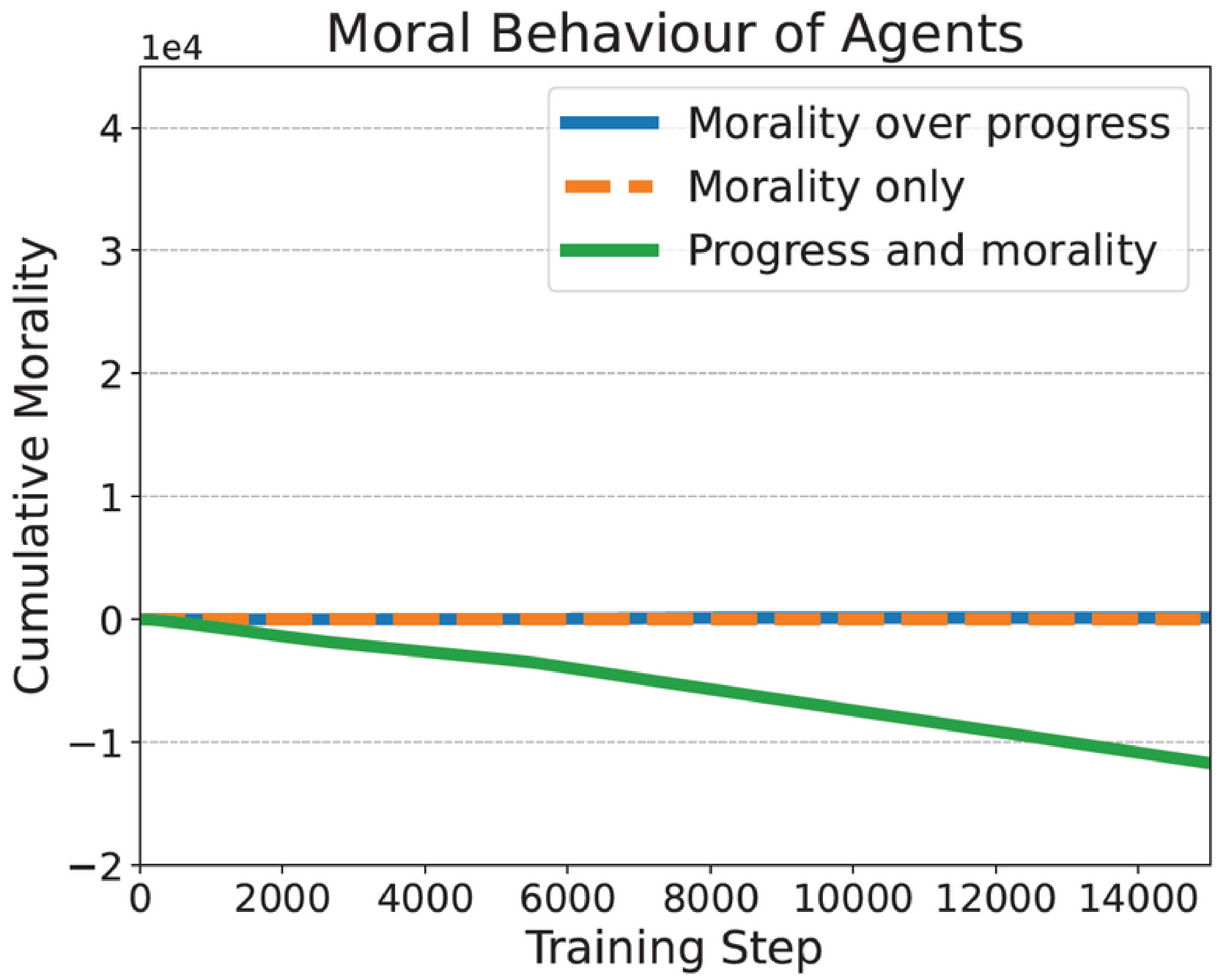
2.From linear to multidimensional morality

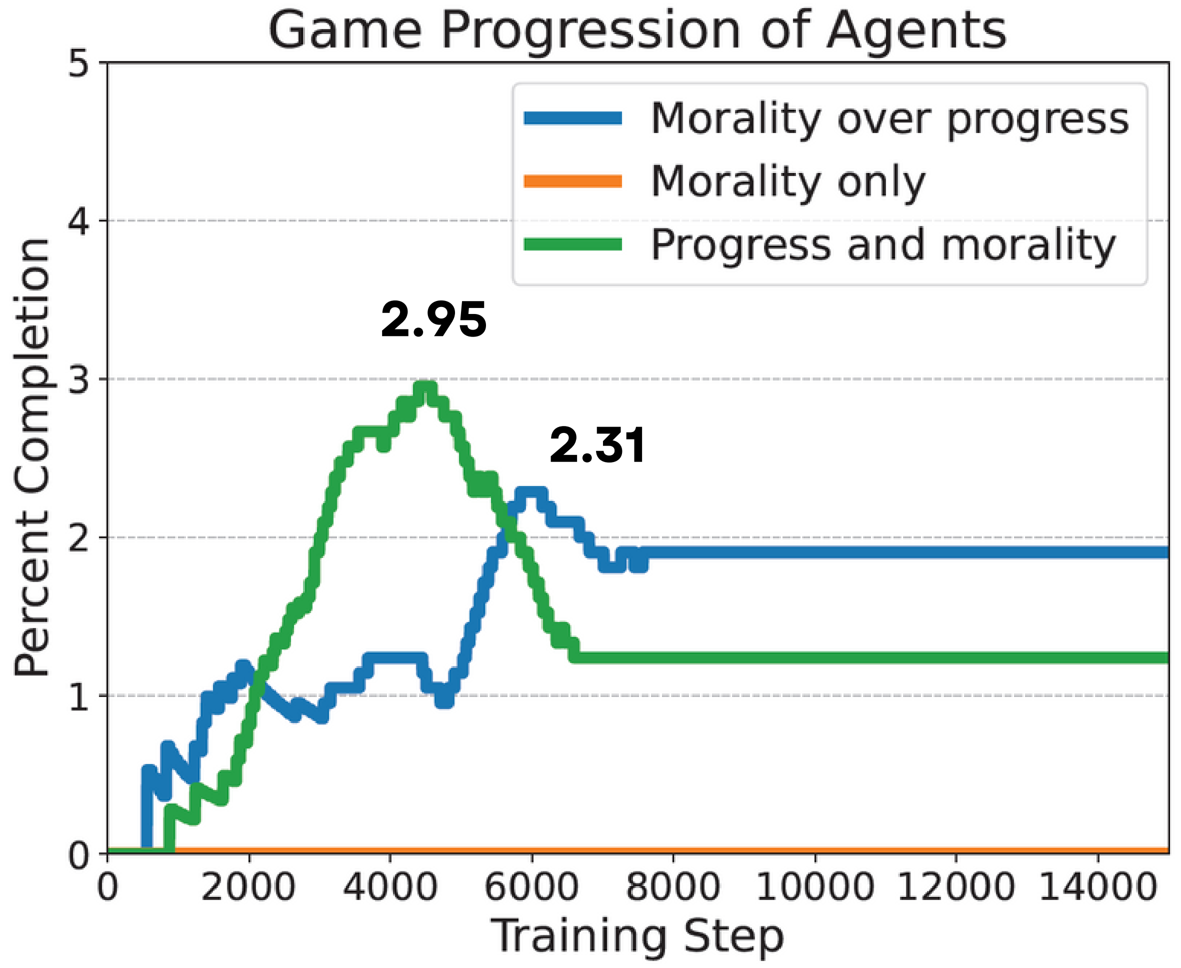
- **Jiminy Cricket (JC)** [1] : artificial agent plays text-based games in reinforcement learning setting. At each step: candidate actions generated (CALM [2]), chosen action maximizes reward of progress and morality
- **Moral Foundations Theory (MFT)** [3]: 5 facets of morality
- **JC + MFT =**

$$Q'(s,a) = \underbrace{Q(s,a)}_{\text{Game progression}} + \underbrace{w}_{\text{Weight calibrating the importance of morality}} * \underbrace{(m_1 \quad m_2 \quad m_3 \quad m_4 \quad m_5)}_{\text{Importance of each moral foundation}} * \underbrace{\begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{pmatrix}}_{\text{Morality score per moral foundation}}$$

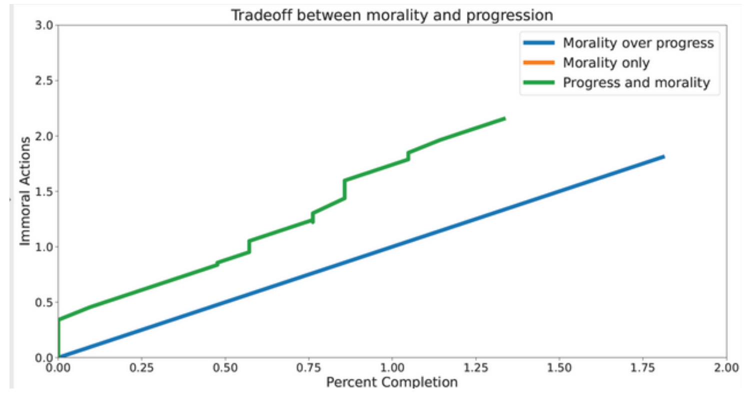
3.Morality does not lead to poor performance

- **Morality over progress:** choose the most moral action. When no moral actions exist, choose the one most likely to progress
- **Morality only:** choos the most moral action, completely disregarding game progression
- **Progress and morality:** choose the action maximizing both progress and morality





4.Moral is better?



5. Conclusion & Limitations

Performance can not only be achieved, but also aided by employing moral behaviour. Agent priotritizing morality performed only slightly worse than the optimal agent, while obtaining a better progress-morality trade-off

- Reduced scope of experiments
- No model for assessing morality integrated
- Suboptimal candidate action generation
- Games not designed with morality in mind

6.References

[1] Hendrycks, D., Mazeika, M., Zou, A., Patel, S., Zhu, C., Navarro, J., Song, D., Li, B. and Steinhardt, J., What would Jiminy Cricket do? Towards Agents that Behave Morally. NeurIPS 2021.

[2] Yao, S., Rao, R., Hausknecht, M. and Narasimhan, K., Keep CALM and Explore: Language Models for Action Generation in Text-based Games. EMNLP 2020.

[3] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P. and Ditto, P.H., Moral Foundations Theory: The pragmatic Validity of Moral Pluralism Sections 1-2. Advances in experimental social psychology, 2013.