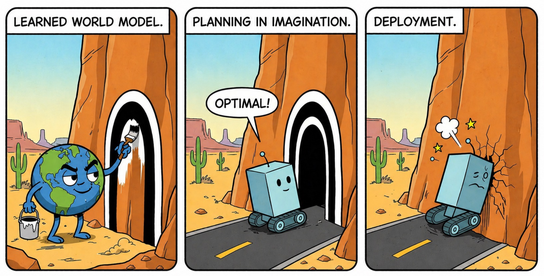
Imperfect World Models are Exploitable

Logan M. Bhamidipaty[1]*, Esmeralda S. Whitammer[1], David Abel[1], Mykel J. Kochenderfer[2], Subramanian Ramamoorthy[1]
[1]University of Edinburgh, [2]Stanford University *l.m.bhamidipaty@sms.ed.ac.uk

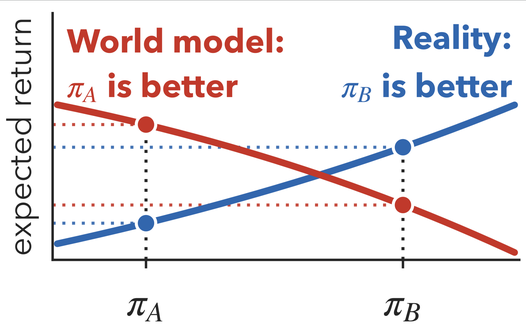
World models can trick agents



Agents **train inside learned world models**. A policy that looks **good in the model** can be **bad in reality**.

Q: When can we trust a world model to rank policies the way reality does?

Exploitable = a flipped ranking



DEFINITION
Model exploitation

The **world model** ranks π_A above π_B.
Reality ranks π_B above π_A.

Your world model is exploitable.

The only safe world model is one that ranks every policy exactly as reality does.

Formal statement: Corollary 1, below.

Generalizing the reward hacking theorem to world models

THEOREM 1
Value inversions are everywhere

Let J_1 and J_2 be non-trivial, non-equivalent value functions. If Π contains an open subset, then it admits a value inversion for J_1 and J_2.

On any policy set with an open subset, every non-trivial, non-equivalent pair of…

COROLLARY 1
Model exploitation

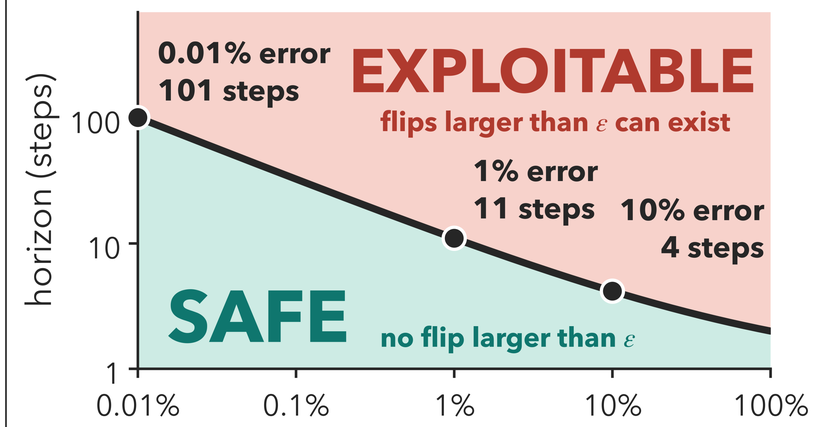
…**world models is exploitable**. Even among near-optimal policies (Corollary 2).

COROLLARY 3
Reward hacking

…**reward functions is hackable**. This recovers **Theorem 1 of Skalse et al.**, "Defining and Characterizing Reward Hacking" (NeurIPS 2022).

- Reward hacking and model exploitation are **both flipped rankings**.
- Model exploitation **can be rewritten as reward hacking**, but not the reverse.
- **Hacking is easier to avoid**: rewards form a vector space, world models do not.

Safe horizon: how far ahead can you trust it?



Theorem 2 at tolerance $\varepsilon = 1$, rewards in [0, 1]. Horizon means $1/(1-\gamma)$.

THEOREM 2
Safe horizon

Let rewards lie in [0, 1] and $\delta = TV(T, T')$. The safe horizon is

$$H(\varepsilon, \delta) = \frac{1}{2}\left[(1+\varepsilon) + \sqrt{(1-\varepsilon)^2 + 4\varepsilon/\delta}\right]$$

No ranking flip larger than ε exists on any policy set whenever $1/(1-\gamma) \le H(\varepsilon, \delta)$. **This bound is tight.**

COROLLARY 4
Planning 10× further ahead needs a world model about 100× more accurate.

This work was supported by a UKRI Turing AI World Leading Researcher Fellowship on AI for Person-Centred and Teachable Autonomy (grant EP/Z534833/1)

arXiv:2605.15960