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

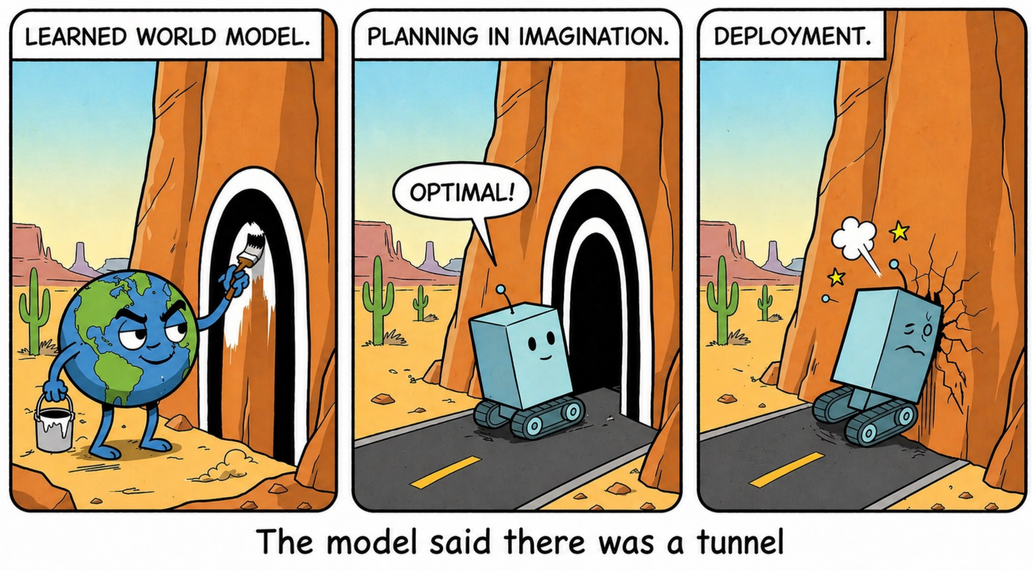
Motivation

World models are **widely used in robotics and RL** and have applications in AVs, healthcare, and synthetic data generation.

World models let agents **train in imagination** and learn from simulated rollouts instead of costly/unsafe real-world trials.

Unfortunately, world models are imperfect approximations, so behavior that looks **good in simulation may be bad in reality.**

Q: When can we trust world models to rank policies like reality?



The model said there was a tunnel

Key Definitions

To investigate our question, we propose a new definition for **model exploitation** as a **value inversion** in policy rankings.

DEF 1 – Value Inversion
Policy value functions J_1 and J_2 *admit a value inversion w.r.t.* Π *if there exists* $\pi, \pi' \in \Pi$:

$$J_1(\pi) > J_1(\pi') \text{ and } J_2(\pi) < J_2(\pi').$$

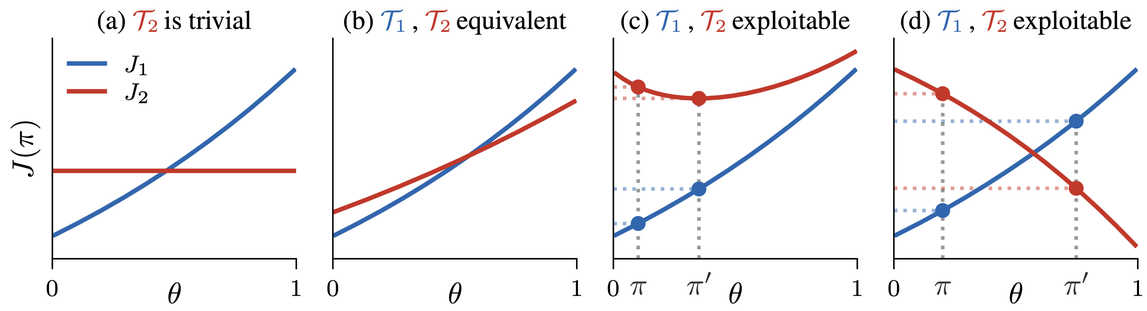


Fig 1. Examples of trivial, equivalent, and exploitable transition models in a 3-state MDP.

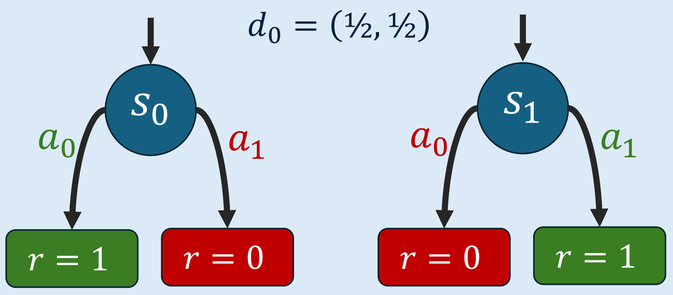
Intuitively, a world model is exploitable if it implies policy A is strictly better than policy B, but reality implies the reverse.

DEF 2 – Model Exploitation
A pair of world models (T_1, T_2) *is exploitable w.r.t.* Π and $(S, A, \_, R, \gamma, d_0)$ *if it admits a value inversion.*

DEF 3 – Reward Hacking [Skalse et al. (2022)]
A pair of reward functions (R_1, R_2) *is hackable w.r.t.* Π and $(S, A, R, \_, \gamma, d_0)$ *if it admits a value inversion.*

Model Exploitation vs Reward Hacking

Exploitation is harder to avoid than hacking because transition functions live in a simplex rather than a vector space.



π^* = play **green action**
π_0 = play a_0
π_* = play **red action**

$$J(\pi^*) > J(\pi_0) > J(\pi_*)$$

Fig 2. Finite policy sets guarantee unhackability but not unexploitability.*

Prop 1 – Exploitation implies hacking
For any exploitable pair (T, T') *w.r.t.* Π *and* $(S, A, \_, R, \gamma, d_0)$ *there exists a reward function* R' *s.t.* (R, R') *is hackable w.r.t.* Π *and* $(S, A, R, \_, \gamma, d_0)$.

This connection hints that a **unified theory** should exist.

Unifying Exploitation & Hacking

Exploitation and hacking have different mathematical structures but **share the geometry of value inversion.**

LEMMA 1 – Local inversion
Let Π *be open. If there is a* v *s.t.* $\nabla_v J_2(\pi) < 0 < \nabla_v J_1(\pi)$, *then* J_1 *and* J_2 *admit an inversion on* Π.

LEMMA 2 – Global equivalence
Let $\widetilde{\Pi} \subset \Pi$ *be open. If* J_1 *and* J_2 *are non-trivial and have positively proportional gradients wherever both are nonzero on* $\widetilde{\Pi}$, *then* J_1 *and* J_2 *are equivalent on* Π.

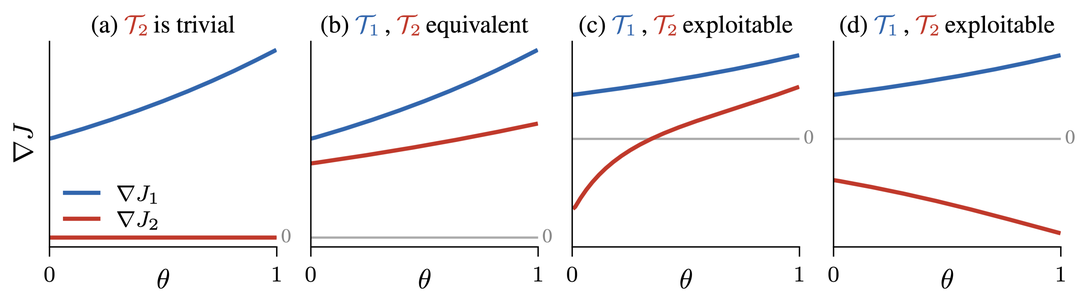


Fig 3. Policy value gradients from transition models in Fig 1.

THM 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.

COR 1 – Imperfect world models are exploitable
On any policy set with an open subset, every non-equivalent, non-trivial pair of world models is exploitable.

COR 2 – Theorem 1 of Skalse et al. (2022)
On any policy set with an open subset, every non-equivalent, non-trivial pair of reward functions is hackable.

Safe Horizon for World Models

THM 2 – Safe Horizon
Normalize R *and let* $\delta = TV(T, T')$. *The safe horizon is*

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

Every pair (T, T') *is* ε*-unexploitable on every policy set whenever* $1/(1-\gamma) \leq H(\varepsilon, \delta)$. *This bound is tight.*

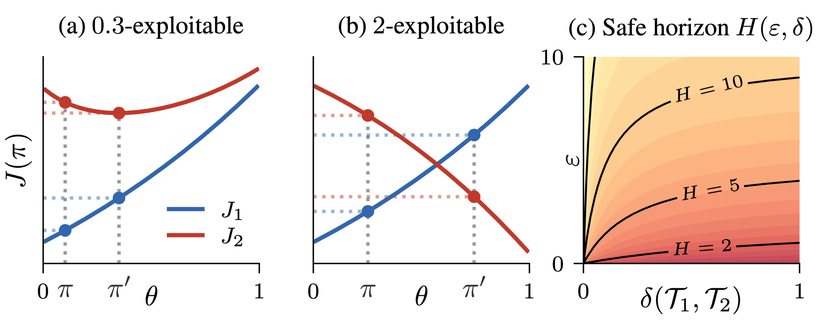


Fig 4. Examples of ε-unexploitability and a safe horizon contour plot.

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)