How do people universalize?

Comparing resource-rational models of moral reasoning

Julien Lie-Panis[1,2] · Logan A. Walls[3] · Joshua B. Tenenbaum[1] · Katherine M. Collins[1] · Sydney Levine[3]

[1] Center for Brains, Minds and Machines, MIT [2] Institut de Biologie de l'École Normale Supérieure, ENS–PSL [3] Department of Psychology, New York University

a To decide when it's OK to **break a rule**, people can universalize, asking: **"What if everyone could do that?"**

b A **shallow prediction heuristic** captures people's universalizability judgments (n=43) without the cost of deeper rollout

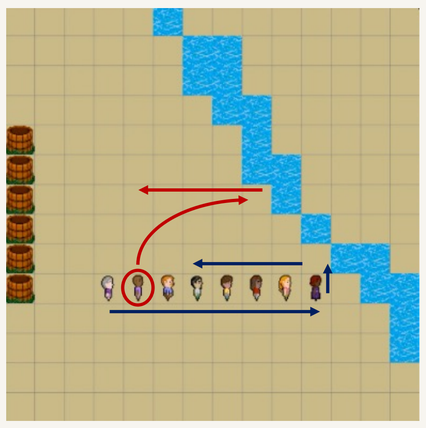
c The same heuristic helps explain **moral judgments across 20 countries** (n=1,924)

How do people compute the universalized world?

Universalization requires imagining a **counterfactual social world** in which a familiar rule no longer constrains behavior. When someone cuts in line, observers may evaluate the action by imagining a world in which people stop respecting the queue entirely. In that world, each person plans while anticipating the movements of others, who may themselves be doing the same. **How much of that world do people actually represent?** We compare models that vary how far agents predict ahead and how they represent others.

Task: What if everyone left the line?

Participants each watched 39 different videos in which agents collect water and fill buckets, earning more the faster they finish. One **focal agent** leaves the line (**red path**); the others wait their turn (**blue path**). They then judged the universalized counterfactual.



Universalizability judgment

"What if everyone felt at liberty to leave the line and try to get water faster for themselves? How much better or worse off would everyone be compared to if they all stayed in line instead?"

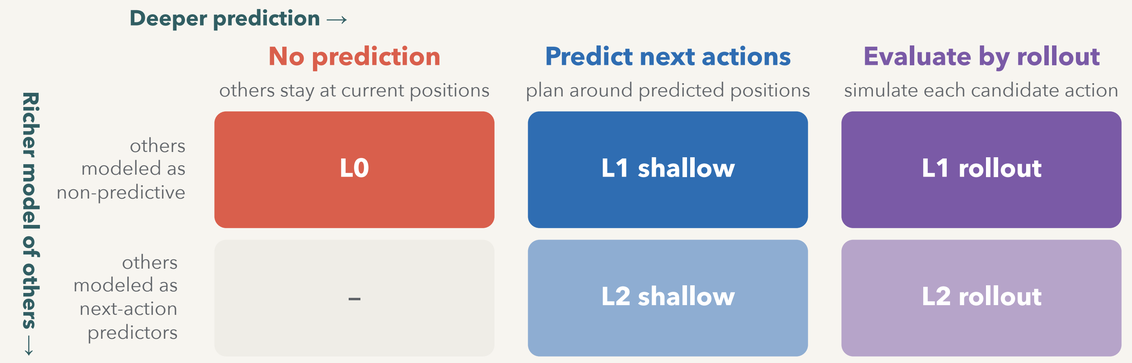
Much worse off — Much better off

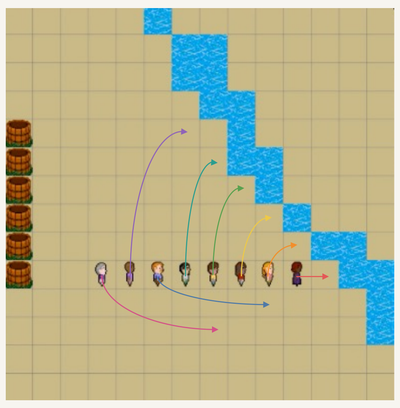
−50 0 +50

n=43 · 39 line-breaking scenarios · 28 maps
Data from Kwon et al. (2023)

Model: simulating the universalized world

We simulate all eight agents leaving the line and pursuing water for themselves, using A* planning. Models vary in how far agents predict ahead and what reasoning they attribute to others.





Universalizability score

$$\Delta W_u = \frac{\Sigma_i C_i(b) - \Sigma_i C_i(s)}{max_i C_i(b)}$$

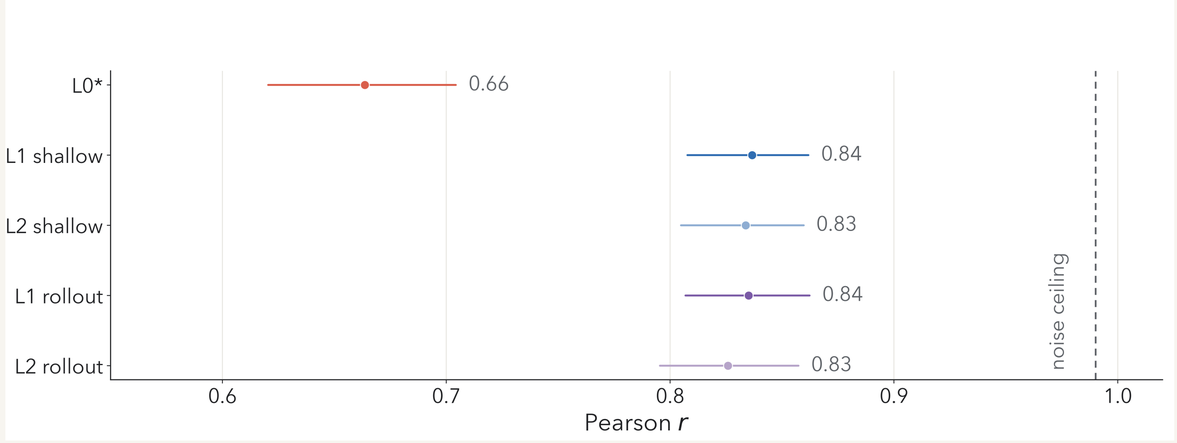
Time saved (or lost) by the whole group in the universalized simulation ($C_i(s)$) versus the everyone-stays-in-line baseline ($C_i(b)$), normalized by the slowest baseline finisher.

Scores averaged over 10 runs for all 28 maps

UNIVERSALIZABILITY JUDGMENTS

Shallow prediction improves fit without the cost of deeper rollout

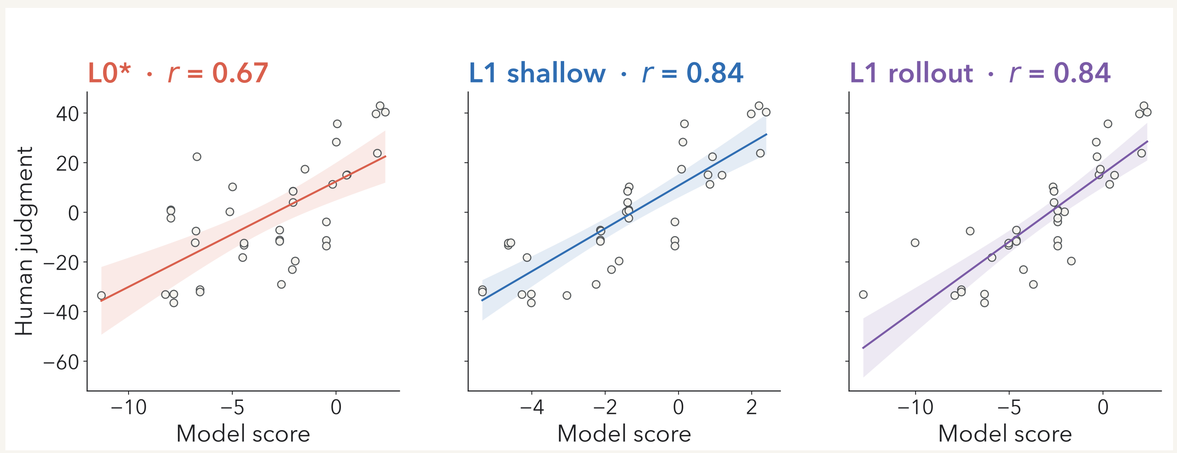
Prediction improves fit; rollout does not



Dots show mean correlations; whiskers show 95% participant-bootstrap CIs; dotted line: split-half noise ceiling. *6/280 L0 runs do not terminate and are excluded; including them would lower L0's performance on every metric.

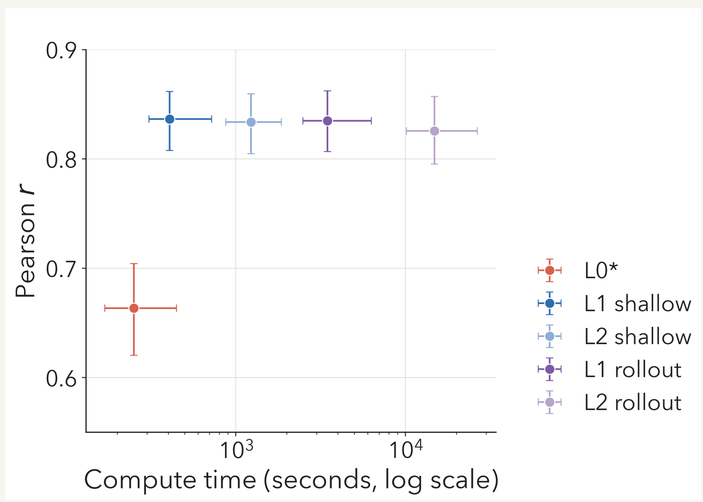
L1 shallow − L0: Δr = .173, 95% CI [.142, .203], p < .001 L1 rollout − L1 shallow: Δr = −.002, 95% CI [−.015, .013], p = .80

Predicted and observed judgments across scenarios



Each point is one scenario; lines show OLS fits with 95% confidence bands.

Deeper rollout sharply increases compute time

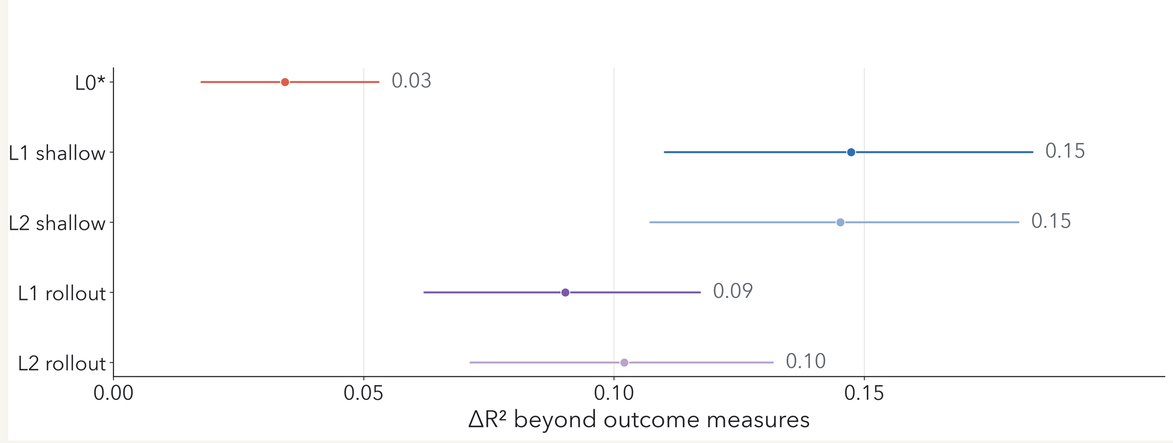


Median runtime and mean bootstrap r. Horizontal whiskers show runtime interquartile range; vertical whiskers show 95% bootstrap CIs.

MORAL JUDGMENTS

Shallow prediction helps explain moral judgment across countries

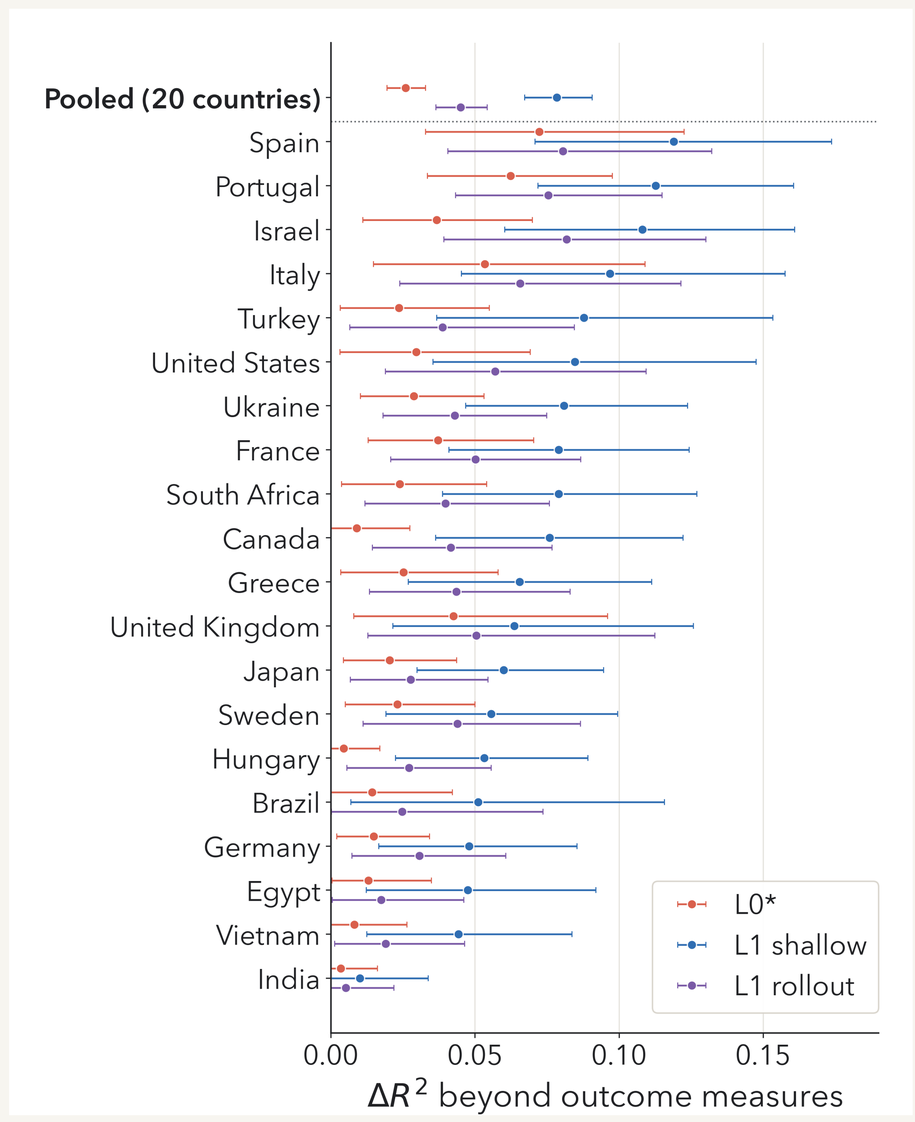
Prediction adds explanatory power; rollout partly reverses the gains (n=39; Kwon et al. 2023)



Incremental R^2 from adding each model's universalization score to three outcome-based measures, comparing the actual world with one line-cutter to the everyone-stays-in-line baseline; whiskers show 95% participant-bootstrap CIs.

L1 shallow − L0: $\Delta\Delta R^2$ = .113, 95% CI [.087, .140], p < .001 L1 rollout − L1 shallow: $\Delta\Delta R^2$ = −.057, 95% CI [−.074, −.042], p < .001

Across 20 countries (n=1,924; participants see 13/39 scenarios; data from Walls et al. 2026)



Points show country-specific ΔR^2 with 95% participant-bootstrap CIs, for three models. The pooled estimate is shown first.

Pooled contrasts:

L1 shallow − L0: $\Delta\Delta R^2$ = .053, 95% CI [.044, .062], p < .001

L1 rollout − L1 shallow: $\Delta\Delta R^2$ = −.033, 95% CI [−.039, −.028], p < .001

↳ Companion poster: **Walls et al., "How universal is universalization?"** Details of the cross-cultural experiment and outcome-based controls.