Language and Experience :

A Computational Model of Social Learning in Complex Tasks

Cédric Colas, Tracey Mills (MIT), Ben Prystawski (Stanford), Michael H. Tessler (Google DeepMind),
Noah Goodman (Stanford), Jacob Andreas (MIT) and Josh Tenenbaum (MIT)

Humans learn quickly from experience and language

Learning to forage mushrooms

novice forager

experienced forager

"chanterelles can be found by oak trees after warm rains"

"orange ones with separate gills are not chanterelles, they are TOXIC"

Individual learning (experience)

- slow exploration and learning
- dangerous exploration: poisonous mushrooms

Social learning (language)

- faster exploration
- safer exploration

→ **spreads exploration costs across people, drives cultural accumulation**

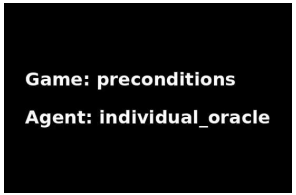
Previous work:

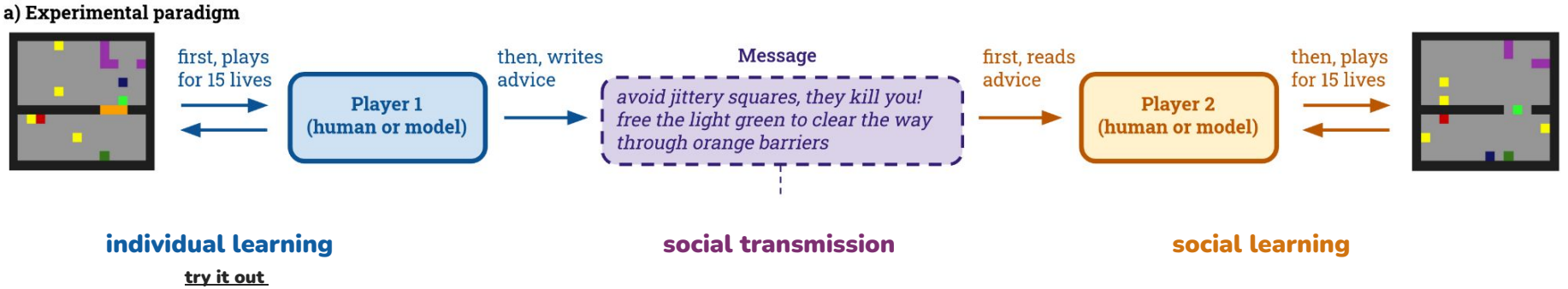
- is sample inefficient (intrinsically motivated RL, language-conditioned RL, *Sutton 98*, *Schmidhuber 91*, *Oudeyer 07*, *Luketina 19*; *Colas 22*)
- learns from experience only (theory-based RL, *Tsividis 2021*)
- works with predefined set of hypotheses (e.g. Bayesian theory of mind, *Baker 11*, *Jara-Ettinger 16*, *Zhi-Xuan 23*, *Ying 24*)

We propose to model the human capacity to quickly solve novel tasks from experience and human language

Experimental design: learning to solve unknown video games

Problem description: the learner must solve a new video game, with unknown action space, state space, dynamics, and reward function



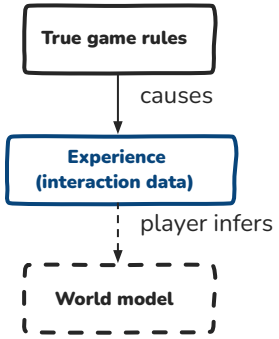


Learning from experience with theory-based RL (Tsividis et al., 2021)

Recipe: learn a world model, plan actions with that world model

↳ $W = P(s_t|s_{0:t-1}, a_t)$ should explain the player's experience

- ~~Deep model-based RL: optimizing neural networks to minimize prediction error~~
- Theory-based RL: inferring posterior distribution over structured programs $P(W|\text{exp})$



Bayes rule

$$P(W|\text{exp}) \propto P(W) \times P(\text{exp}|W)$$

prior

likelihood
(via simulations)

Planning: given this model, you can plan to explore (eg maximizing information gain) and plan to solve the game (maximizing game rewards).

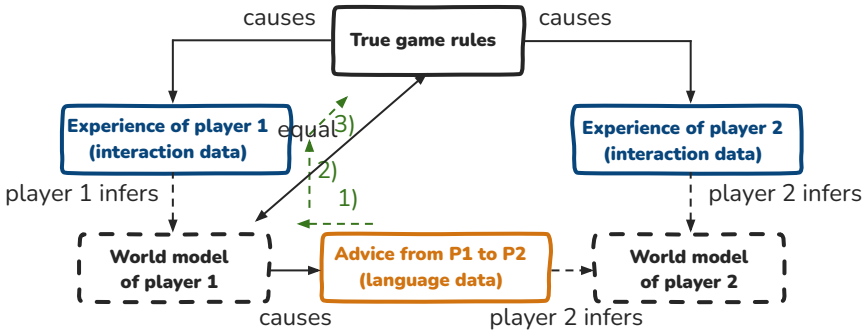
Advantages over deep model-based RL approaches

- faster and more generalizable world modelling (leveraging program structure)
- better decision-making (extracting goals from world programs)
- active exploration (using uncertainty from inference process)

Theory-based RL better captures human-like sample efficiency and exploration patterns than deep RL (see Tsividis et al., 2021)

Extending theory-based RL to learn from language

Multi-agent setting



In principle, inferring game rules from language requires inverting 3 models (that's hard)

Instead, **we assume that player 1 knows what they're talking about: world model 1 == true game rules**

so we only need to invert the speaker model (world model → message)

We use an LLM as a general **speaker model:** $P(\text{lang}|W)$

- We can model **advice generation** by sampling from that distribution
- We can measure the probability of player 1 saying {message} if they thought the world was W

We can now infer world models from experience and language jointly

(Bayesian theory of mind, *Baker 11, Jara-Ettinger 16, Zhi-Xuan 23, Ying 24*)

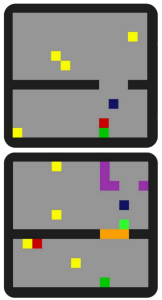
language likelihood from LLM

$$P(W|\text{exp},\text{lang}) \propto P(W) \times P(\text{exp}|W) \times P(\text{lang}|W)$$

Efficient world modelling from experience and language

World programs expressed in a game DSL (Schaul., 2013)
compilable (game engine in the head, Ullman 2017)

```
SpriteSet
  darkblue > MovingAvatar
  orange > Immovable
  yellow > RandomNPC
  purple > Immovable
  green > Immovable
  lightgreen > Chaser s=orange
  red > Immovable
InteractionSet
  darkblue yellow > killSprite reward=-1
  darkblue orange > killSprite reward=-1
  lightgreen purple > stepBack
  orange lightgreen > killSprite
  purple darkblue > killSprite
  green darkblue > killSprite reward=1
TerminationSet
  isZero s=darkblue win=False
  isZero s=green win=True
```



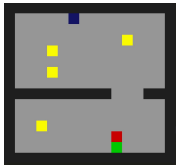
The space of possible world programs too large to enumerate, we need approximate inference algorithms (particle filter with Monte-Carlo rejuvenations)

Initialize:

sample N world programs from prior

Message

avoid jittery squares, they kill you!

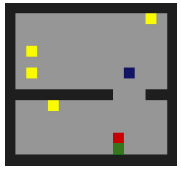


Ep 1, Step 1

```
SpriteSet
  darkblue > ShootAvatar s=red
  yellow > Immovable
  green > Passive
  red > Immovable
InteractionSet
  green darkblue > push
TerminationSet
  isZero s=darkblue win=False
```

Model maintains N=20 theories

Loop:



Ep 1, Step 9

Exp: yellow move, red, green don't move

Collect data

```
SpriteSet
  darkblue > ShootAvatar s=red
  yellow > Immovable          ?
+ yellow > RandomNPC
  green > Passive
  red > Immovable
InteractionSet
  green darkblue > push
+ darkblue yellow > killSprite ?
TerminationSet
  isZero s=darkblue win=False
```

Local exploration: propose rule swaps

$$\text{accept} \sim \mathcal{B}\left(\frac{P(W_{mod}) \times P(\text{exp}|W_{mod}) \times P(\text{lang}|W_{mod})}{P(W) \times P(\text{exp}|W) \times P(\text{lang}|W)}\right)$$

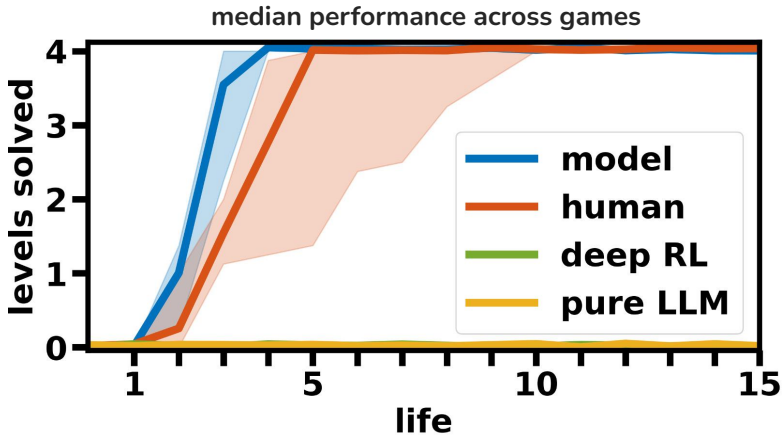
```
SpriteSet
  darkblue > ShootAvatar s=red
  yellow > RandomNPC
  green > Passive
  red > Immovable
InteractionSet
  green darkblue > push
  darkblue yellow > killSprite
TerminationSet
  isZero s=darkblue win=False
```

Refocus on promising paths

Resample theories proportional to joint probs

Human-like sample efficiency in VGDL games

Our model can solve these games in < 10 lives (comparable to humans)



Example game: Bees and Birds

But:

- **LLM agent fails:** struggles to infer game rules, form and execute plans
- **deep RL fails:** lower sample efficiency and no exploration capabilities

Message analysis

Examples

Model:
"Control your darkblue square with arrow keys. ***To win, eliminate all green objects*** *by touching them to earn points. Be cautious, as* ***yellow and orange will kill you on contact*** *- try to use lightgreen objects to your advantage, as they can take out these threats by touching them. Your* ***goal is to survive while clearing the board of green objects****."*

Human:
*"****Yellow will kill you****, the goal is to* ***get to green without being killed****. From level 2 up,* ***you can "eat" purple****, light green is your friend, free him and he will eat orange for you,* ***orange will also kill you****, get to green as fast as possible."*

Example game: Bees and Birds
(model learning from human advice)

Message analysis:
We hand-coded messages and found that:

- 88% contained info about game dynamics
- 64% contained info about win conditions
- 74% contained info about lose conditions
- 11% contained errors

Effect of message content (fixed effect model):
Longer messages and messages containing more info about win conditions led to stronger performance

Messages impact exploratory behaviors

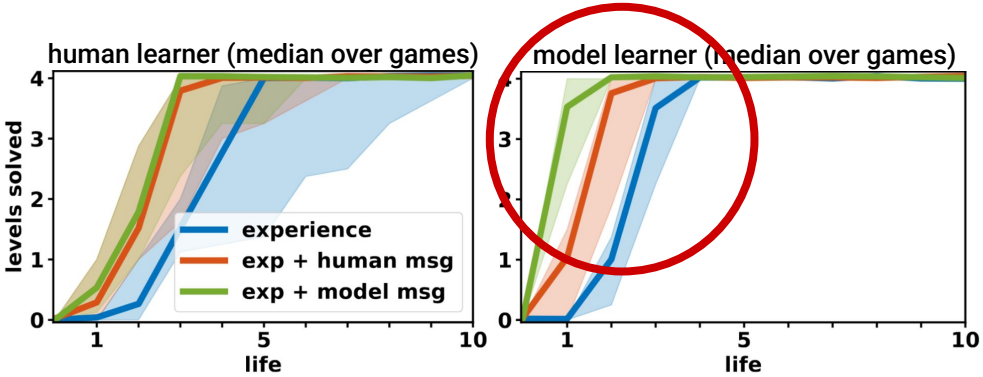
Safer exploration: Players who read about a dangerous interaction experienced it less

Faster exploration: Players who read about a key game rule discovered it earlier

Mistakes restrict exploration: False information about a rule may impact exploration negatively — the double-edged sword of pedagogy (Bonawitz, 2011)

Models and humans learn from model- and human-generated advice

Cross-embodiment social learning
Humans and models can learn from humans or models



Models prefer model guidance

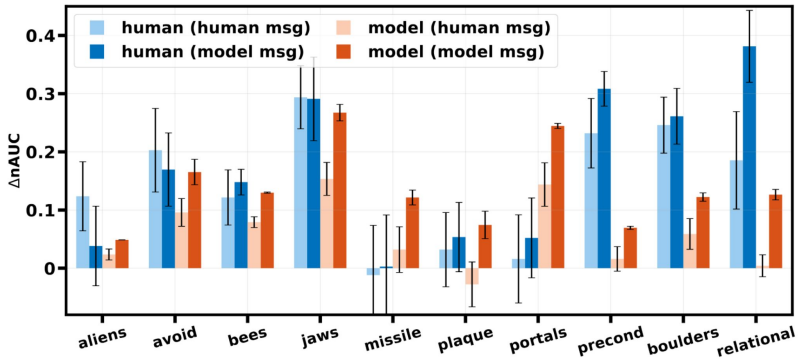
- because human guidance is harder to interpret: meta-strategy, planning advice, uncertainty

"the orange is like terminator"

"take time to look for safe patterns"

"I was left very confused ..."

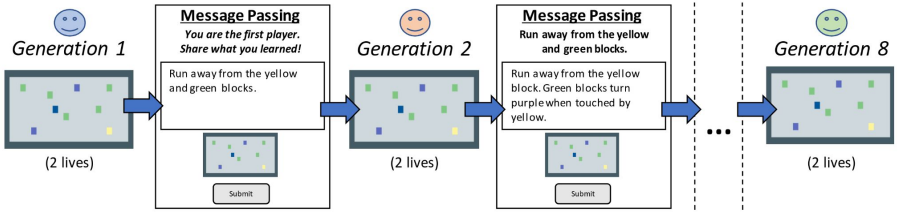
Difference in normalized area under the curve compared to individual learning



Transmission chains — accumulating knowledge across generations

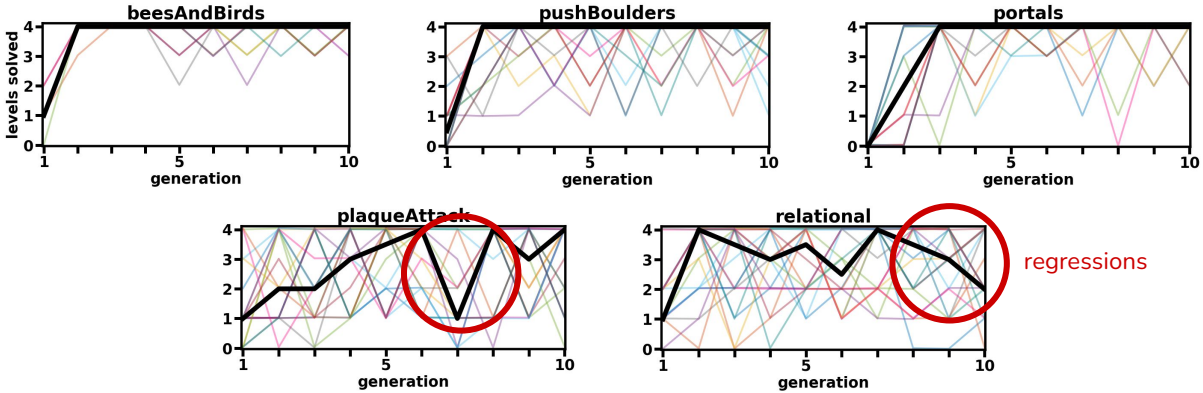
Transmission chains with human participants **(previous result from Tessler et al., 2021)**

Human-human chains:
10 generations,
2 lives each,
transmit a message



Our model accumulates knowledge across generations (cultural learning)

Model-model chains:
same setting



Discussion

Summary

- a Bayesian model of learning to solve sequential decision-making tasks from experience and language
- captures human-like performance and sample efficiency, language-shaped exploratory patterns

Future work

- inference over more general game representation spaces (e.g. Python)
- learning from more diverse linguistic advice (e.g. goals and strategies, meta-cognitive advice, etc)

From cognitive models to social partners

- we observe successful transfer between humans and models
- more human-like AI partners for better human-machine collaboration (Brinkmann et al., 2024; Collins et al., 2025)

Thanks!