Topics in Intelligence: World Models and Social Reasoning TCS Problem Formulation Whitepaper

Avrim Blum Sanjoy Dasgupta Madhur Tulsiani
Scribe: Shiry Ginosar

July 24, 2026

1 Caregiving

Caregiving is fundamental in the development of many different species including humans. Experiments show that children explore more when a caregiver is present [Gopnik, 2019, Tottenham et al., 2019], and more broadly, caregiving allows children to remain children for longer periods [Gopnik, 2020]. Caregiving, and being taken care of, are a basic part of the human experience, and may even be a model for "raising" ethical and well-behaved AI systems [Gopnik, 2025]. Yet caregiving remains an understudied phenomenon. This section proposes several possible theoretical directions in modeling and studying caregiving from an algorithmic, game-theoretic, and learning-theoretic perspective.

Exploration under caregiving and caregiving for exploration. It is intuitively natural that children would explore more when a caregiver is present, since the caregiver can protect the child from danger. Can we develop a mathematical model to study this phenomenon more quantitatively, e.g., to understand optimal exploration under caregiving and optimal caregiving for exploration? Below is one possible starting point, which takes an RL/MDP perspective, but leads to behavior that looks much more like pure curiosity-driven exploration rather than activity focused on immediate rewards. In this way, it aligns to an extent with empirically-observed behavior of children, who explore in ways that are focused more on world-model building, and the idea of a long childhood itself as an evolutionary solution to explore-exploit tradeoffs [Gopnik, 2020].

Specifically, as a starting point, consider a deterministic Markov Decision Process (MDP) which is unknown to the child, and where the child's eventual goal (one could view this as evolution's goal for the child) is to maximize average reward per time step. Let's imagine that rewards are in the range $[-\infty, B]$ for some finite bound B, where we can replace $-\infty$ with a quantity that goes to negative infinity in the limit if we assume all actions having this reward lead to a dead state. The child (or evolution) either (a) has a prior distribution P over MDPs, or (b) has a family H of MDPs and is interested in maximizing their worst-case performance over the MDPs in H.

In this scenario, what should the child do? Starting from its initial state s_0, all it can

do is pick some action to get to a next state, say s_1. If $s_1 s_0$, then again all it can do is pick some action getting to another state s_2. However, once the child reaches a state it's been in before (it's found a cycle), then for any interesting prior P (in which any untaken edge has nonzero probability of having reward $-\infty$) or any interesting set H (in which any untaken edge could have reward $-\infty$), the child now should just loop around the cycle it has found, because this cycle has a finite average reward whereas a new action choice could give $-\infty$. Therefore, optimal child behavior, from the perspective of evolution's goal of the child maximizing average reward, would be very conservative, no longer exploring.

However, suppose a caregiver can stop the child from going over the cliff, burning their hand, etc. More generally, assume that if the child is in some state s and chooses some action a, the caregiver can intervene and prevent the child from taking that action. What is the minimal amount of intervention that ensures a child will find an optimal policy (i.e., will find the cycle of maximum mean reward)? The caregiver needs to prevent the negative infinities, and it should prevent any edge that takes the child to a new strongly-connected component of the MDP with a lower optimum value (like poking out an eye or getting addicted to drugs). Then the child has no reason not to try everything until it has found the optimal cycle. In fact, unless the child has somehow discovered a self-loop with reward B, the child should try everything, because any edge it has not yet taken could possibly be part of an optimal cycle. So, it is optimal behavior for the child to perform pure exploration until it has tried every edge.

Notice that if the caregiver just stops the negative infinities but does not prevent the actions that cause "permanent damage" by taking the child to a lower-valued strongly-connected component, then an evolutionarily-optimal child might explore less (depending on the prior P or set H) and not find the reward-maximizing cycle. On the other hand, there is no need for the caregiver to prevent the child from taking actions it can recover from.

The above represents just some initial thoughts. It would be interesting to consider optimal behaviors under specific families H, or under different assumptions on discount factors and reward values. Below are a few additional directions that would be interesting to investigate.

Models of caregiver knowledge and abilities. One might ask why the caregiver in the above scenario doesn't just direct the child to the optimal cycle. One answer could be that it would require too much effort to do so; however, another reason might be that the caregiver only knows an approximate version of the MDP. For example, if the caregiver knows only a multiplicative approximation to the child's reward function, then it could still stop the negative infinity actions, but the child now might want to disobey the caregiver when the caregiver simply believes the component being transitioned to is somewhat worse than the current component, depending on the child's priors. More broadly, it could be interesting to consider different models of caregiver knowledge and abilities.

The caregiving game. From an evolutionary (or mechanism design) perspective, what reasonable utilities for the caregiver might lead to overall good outcomes of the caregiver-child system? This could be interesting to model and analyze in simpler repeated-game settings.

Learning to learn and data-driven algorithm design. Another reason that a caregiver may wish to allow a child to explore rather than to directly lead the child to an optimal policy is that the MDP only represents the current environment. The next day, the child may be faced with a different, related MDP. Exploration today may help the child learn how to better explore a new environment tomorrow, including in the future when the caregiver is no longer around. This setting fits nicely into the framework of data-driven algorithm design, in which past instances are used to learn a good algorithm for future instances. This is a direction with substantial recent TCS work developing analytical and algorithmic techniques (see, e.g., [Balcan, 2020, Balcan et al., 2024]), so it would be interesting to explore this problem from that perspective.

More generally, we believe there should be many more interesting things one can say theoretically about caregiving strategies and exploration under caregiving.

2 Convergence of Representations

Vision, language and multimodal models encode data (e.g. embed images or sentences into a vector space) and then use this encoding for a variety of tasks (e.g. translating the sentence to another language). The encoding can be thought of as the model's internal representation of a data instance. There is growing empirical evidence that the representations used by different models are converging, in the sense of becoming more similar to one another.

This phenomenon has been noted repeatedly over the past decade and was called the "Platonic representation hypothesis" in a recent paper Huh et al. [2024] that conducted a thorough empirical study. The paper considered a large number of different vision and language models, with different training sets and training criteria, and assessed the similarity of the representations found by them. Cross-modal representations were also compared, by relating the embedding of an image by a vision model to the embedding of the image's caption by a language model.

To discuss this work a bit more precisely, let's focus on two representations (corresponding to two different models) of a single data space X: call these $f : X \to R^p$ and $g : X \to R^q$. An "alignment" between the embeddings $f(x)$ and $g(x)$ is typically taken to mean one of three things:

1. $f(x)$ is well approximated by an affine function of g(x).
2. Pairwise dot products $< f(x), f(x') >$ are strongly correlated with $< g(x), g(x') >$.
3. The points x' whose embeddings $f(x')$ lie in the neighborhood of $f(x)$ are roughly the same as those in the neighborhood of $g(x)$.

Among these, (1) is the strongest form of equivalence and (3) the weakest. The experiments in the earlier-mentioned paper Huh et al. [2024] use both (2) and (3) but find the clearest evidence of convergence in the sense of (3).

There are many applications of convergent representations, for instance in mechanistic interpretability and steering of large models, or in using one modality to interpret another (e.g. brain signals) with which it is aligned. But the basic question remains of why such convergence is plausible. Here are some cases to consider, in order of

increasing generality.

1. Two models with the same architecture are trained on the same data. The difference between them stems from different initialization and from random sequencing of data while training.
2. The two data sets are not identical but are both sampled from the same underlying distribution.
3. The architectures differ.
4. The modalities differ: e.g. one is based on language while the other is based on images.

Some recent theoretical results tackle these questions under various assumptions. Ziyin and Chuang [2025] focus on deep linear architectures. Bangachev et al. [2026] study data distributions that satisfy the (related) "linear representation hypothesis" and treat representation formation using a dictionary learning approach. Other relevant work includes recent papers by Ciernik et al. [2024] and Insulla et al. [2025]. It would be interesting to more generally provide theoretical understanding of why convergence occurs in such a broad range of settings.

3 World Models

World models are widely considered as a next frontier in the development of agentic AI systems, going beyond "just" the modeling of language or visual data, and allowing an agent to reason about actions in an environment. The environment is modeled as "state" (which can be a partially written text for an LLM), and the reasoning involves predicting and evaluating new states reached as outcomes of "actions". The workshop included an extensive discussion on how to formally define a world model in a way which allows for evaluation and comparison of different models:

What properties should such a definition capture? How can we say one world model is better than another? What does it mean to learn and improve a world model? Is the abstraction of a world model really necessary? What other abstractions are used to model navigation and reasoning in the cognitive science literature?

We summarize below some of the ideas and theoretical directions emerging from the above discussions.

3.1 What is a good world model?

Early definitions such as one by Craik [1967] describe organisms as possessing a "small scale model of external reality and its own actions" allowing for simulation (predicting and evaluating results of actions) and generalization (going beyond past experience). A commonly used notion is define a world model as a function f capable of predicting the new state resulting from the outcome of a given action in a given state: $s_{t+1} = f(s_t, a_t)$. However, there are implicit or explicit assumptions about how rich are the state and action spaces, since they might depend on input/sensor fidelity and internal memory. Thus, a definition needs to compare models representing the same environment with possibly different state spaces.

Another related issue is regarding the tasks that can be used to evaluate a world model. While the models are meant to be agnostic to the tasks they are later used for, any nontrivial definition needs some limits on possible tasks since no nontrivial model can perform tasks of arbitrary complexity (such as predict the position of every atom in the universe, or solve the halting problem). In most discussions, one implicitly works with a large family of tasks, such as tasks which can be done by humans, when comparing world models. Motivated by notions in cryptography, pseudorandomness, and learning theory, one approach could be then to ask: what is a good world model for a given family of tasks $\mathcal{F}$?

To formalize the above, we then consider the following objects:

- A "ground truth" universe (world) $\mathcal{U} = (S_0, A_0, f_0)$, where S_0 is the set of states, A_0 are allowed actions, and $f_0 : S_0 \times A_0 \to S_0$ predicts the results of actions in the real world.
- A world model $\mathcal{M} = (S, A, \phi, f)$ with an internal state set S, action set $A \subseteq A_0$, a projection $\phi : S_0 \to S$ and a prediction function $f : S \times A \to S$
- A family of tasks $\mathcal{F}$

We say two different world models $\mathcal{M}$ and $\mathcal{M}'$ are (almost) equally good for a family of tasks $\mathcal{F}$ if for every task in the family, both models perform (almost) equally well. To measure performance, we also need to formally define tasks. One simple definition could be defining a task as a pair of sets (U_1, U_2), where $U_1, U_2 \subseteq \mathcal{U}$ are sets of source and goal states respectively, and the task is to go from states in U_1 to ones in U_2 (here, a set of states can be equivalently thought of as a property of the states.) A more general definition can be $T = (U_1, U_2, c)$ where $c : \mathcal{U} \to \mathbb{R}$ is a cost/reward function, to track the cost incurred by the model in going from U_1 to U_2. Since the world itself can be thought of as a "model" (with the projection as identity), we can say that is a model is ε good for a family $\mathcal{F}$ if for every task in the family the cost achieved by the model is at most $(1 + \varepsilon)$ times that of the "optimal" model $\mathcal{U}$.

As a toy example, one can consider the case where the "world" is a graph, with the states corresponding to the vertices. The relevant family of tasks is to go from a starting vertex u to target vertex v with the cost function being 1 on all vertices (so that the cost is the length of the shortest path). Then one possible "world model" is a spanner [Peleg and Schäffer, 1989], which is a potentially much sparser graph allowing for approximately the same shortest path lengths.

We note that the above definition is also related to the notion of "Omnipredictors" from learning theory [Gopalan et al., 2022] which can compete with the best predictor from a hypothesis class $\mathcal{H}$ over a large class of loss functions $\mathcal{L}$ (and do not need to be trained with a specific loss function). These are in-turn related to older notions of "dense models" indistinguishable by a large class of test functions [Reingold et al., 2008] used in pseudorandomness. Task-based metrics have also been used in recent works on evaluating world models implicit in LLMs [Vafa et al., 2024].

3.2 Learning world models

Note that the above definition is purely functional or test-based in nature: a good world model is defined as one which performs well on a family of tasks. However, it does not

address some other desirable properties for models such as "compression" or simplicity, and the complexity or learning or updating the model.

For the first property, we can note that the above definition still allows for a comparison based on aspects such as compression (the size of the state space) between two models with similar performance on tasks. While compression is not necessarily included as a property of a good model, one can ask if performance can be retained with significant compression. However, the functional definition does not address the complexity of learning a model, and what actions/state sequences can be used to update an existing model.

With respect to available actions and observations, one distinction is between egocentric models, exocentric models and "models of nature". The first concerns models of an individual's own behavior, where there is rich observable data of actions and resulting states. The second kind of models concern behavior of other individuals (who also form a dynamic and updating system), where only partial observations resulting from actions may be available. The third kind of models (models of nature) are thought to model the physical world as opposed to agents in a system. The differences in observable properties of states in each of these cases can be incorporated in the projection $\phi : S_0 \to S$ which can capture more fine-grained egocentric state information, but only partially incorporates exocentric or physical states.

The above still leaves several directions which need formalization: What algorithms can be used to train a good model? Does the order of training examples (curriculum) matter for efficient training? How should one choose what information to remember i.e., how should a world model be updated based on observed action/state sequences? While it can be helpful for abstraction to separate evaluation from training, the connections to learning theory above may provide some starting points for algorithmic approaches.

3.3 Biological realizations: world models vs. other abstractions

While world models are very active area of research, it is not clear that such models are indeed necessary for intelligent systems or agents. Brooks [1991] argues that intelligence may not necessarily arise from a central representation (model) but rather from a collection of units interacting with the environment in parallel. This was also reflected in the research on nest construction by weaver ants Trocki et al. [2026], which construct nests out of leaves many times larger than any single ant, effectively adapting to different leaf geometries and other environmental constraints.

Other works in evolutionary biology, such as ones by Bennett [1996] and van Woerkum [2023] also consider whether cognitive maps or other mental representations are indeed required by animals for navigation and spatial reasoning, or whether this is a primarily anthropomorphic view. We note that some simple alternatives such as (randomized or deterministic) policies $\pi : S \to A$ mapping states to actions are actually captured by the definition above, by simply having the predictor function $f : S \times A \to S$ return a null state for actions not allowed by the policy. However, it is interesting to consider if there are other abstractions, particularly in case of distributed parallel action, which are not captured by the definitions considered here.

Even for the case of policies or other abstraction captured above, one can consider the question of tradeoffs between efficiency (size) and complexity of the models. For

example, in case of the models for a graph with the tasks as navigation, a single policy may not be able to simultaneously navigate between all possible pairs of start and end points (all tasks in the family). On the other hand, it is possible have a list of policies dependent on the end point (using the shortest path tree from the current node) but the total storage required for all the policies can be as large as $\Omega(n^2)$ for a graph on n vertices, which is larger than the storage required for a "world model" (spanner).

Another relevant consideration from the point of view of biological realizations is that of energy efficiency. While this can potentially be considered as a part of the cost function, this is not currently included in the discussion above. This may also require some care in modeling, as one may need to consider not just the energy costs of transitions or simulations performed by the model, but also those of storage and retrieval for state information.

References

Maria-Florina Balcan. Data-Driven Algorithm Design. In *Beyond the Worst-Case Analysis of Algorithms (Chapter 29)*. Cambridge University Press, 2020. URL https://arxiv.org/pdf/2011.07177.

Maria-Florina Balcan, Dan Deblasio, Travis Dick, Carl Kingsford, Tuomas Sandholm, and Ellen Vitercik. How much data is sufficient to learn high-performing algorithms? *Journal of the ACM*, 71(5):1–58, 2024.

Kiril Bangachev, Guy Bresler, and Yury Polyanskiy. Representation alignment rests on linear structure. *arXiv preprint arXiv:2605.28870*, 2026.

Andrew TD Bennett. Do animals have cognitive maps? *Journal of Experimental Biology*, 199(1):219–224, 1996.

Rodney A Brooks. Intelligence without representation. *Artificial intelligence*, 47(1-3): 139–159, 1991.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. *arXiv preprint arXiv:2411.05561*, 2024.

K.J.W. Craik. *The Nature of Explanation*. Philosophy: science CAM. Cambridge University Press, 1967. ISBN 9780521094450. URL https://books.google.com/books?id=wT04AAAAIAAJ.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215. Schloss Dagstuhl–Leibniz-Zentrum f {\" u} r Informatik, 2022.

Alison Gopnik. Life history, love and learning. *Nature Human Behaviour*, 3(10):1041–1042, 2019.

Alison Gopnik. Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, 375(1803):20190502, 2020.

Alison Gopnik. Three ages and three intelligences: Exploit, explore, empower. University Lecture, 2025. Available online at: https://gradlectures.berkeley.edu/lecture/three-ages-and-three-intelligences-exploit-explore-empower/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

Francesco Maria Gabriele Insulla, Shuo Huang, and Lorenzo Rosasco. Towards a learning theory of representation alignment. In *International Conference on Learning Representations*, volume 2025, pages 37594–37608, 2025.

David Peleg and Alejandro A Schäffer. Graph spanners. *Journal of graph theory*, 13(1): 99–116, 1989.

Omer Reingold, Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. Dense subsets of pseudorandom sets. pages 76–85. IEEE, 2008. URL https://ieeexplore.ieee.org/document/4690942.

Nim Tottenham, Mor Shapiro, Jessica Flannery, Christina Caldera, and Regina M Sullivan. Parental presence switches avoidance to attraction learning in children. *Nature human behaviour*, 3(10):1070–1077, 2019.

Gadi Trocki, Michal Roitman, Ehud Fonio, and Ofer Feinerman. Local rules and geometric constraints enable robust leaf-nest construction in weaver ants. *Current Biology*, 36(1):188–198, 2026.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Bas van Woerkum. Animal navigation without mental representation. *Phenomenology and the cognitive sciences*, pages 1–18, 2023.

Liu Ziyin and Isaac Chuang. Proof of a perfect platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025.