Received: 9 September 2026 | Revised: 9 September 2026 | Accepted: 6 October 2026

DOI: 10.1002/aaai.70114

SPECIAL TOPIC ARTICLE

From "how" to learn to "what" to learn in multiagent systems and robotics

Peter Stone

Department of Computer Science, The University of Texas at Austin, Austin, Texas, USA

Correspondence
Peter Stone, Department of Computer Science, The University of Texas at Austin, Austin, Texas, USA.
Email: pstone@cs.utexas.edu

Funding information
University of Texas at Austin

Abstract

There remains much more to Artificial Intelligence than understanding how to learn with neural networks. Inspired in part by how babies, and humans in general, tend to choose for themselves *what* (and *when*) to learn in addition to *how* to learn, this article discusses similar opportunities, in the areas of multiagent systems and robotics. Grounded in past research on the topic, it categorizes opportunities for choosing *what* to learn as coming from task decomposition or simplification for difficult problems and/or large environments; unsupervised pre-training of representations and priors; multiagent learning; and learning from human interaction. The article is based closely on a keynote talk with the same title, delivered at the AAAI-2026 conference in Singapore on January 22nd, 2026.

INTRODUCTION

The past decade and a half has featured immense progress in the field of Machine Learning, especially with regards to the power of neural networks. Much of the community's effort over this time has been devoted to understanding *how* to enable neural networks to learn as effectively as possible, and how they manage to generalize as well as they do.[1]

Roughly speaking, many of the research papers that have appeared can be categorized into four classes.

Neural network algorithms

Some of the most influential and impactful work has been around neural network algorithms and architectures, both for supervised learning (for example, transformers (Vaswani et al., 2017), state-space models (Gu & Dao, 2024), and diffusion (Ho et al., 2020)) and for deep reinforcement learning (RL) (for example, PPO (Schulman et al., 2017), SAC (Haarnoja et al., 2018), and GRPO (Shao et al., 2024)).

Neural network "psychology"

This class of research has focused on identifying under what conditions these neural network algorithms succeed or fail. For example, can they plan (Kambhampati, 2024)? Can they solve math problems? What can they do and what can't they do? This research studies the macro-level behavioral properties of trained neural networks.

Neural network "neuroscience"

This class of research studies the conditions under which various neurons in a neural network fire, and seeks con-

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2026 The Author(s). *AI Magazine* published by Wiley Periodicals LLC on behalf of the Association for the Advancement of Artificial Intelligence.

trol and/or understanding of their operation by examining them at the micro-level.

Neural network applications

Finally, there is a lot of important work on how to put these new technological advances to good use in targeted applications.

While this work is all very exciting and important, contrary to popular belief, there is much more to AI than how to learn with neural networks. Especially when speaking to students and young researchers looking to carve out a niche for themselves in the field, I tend to advise strongly as follows: Be different! I advise exploring different AI paradigms other than neural networks, including by leveraging and combining past paradigms such as symbolic reasoning and probabilistic models, and by exploring completely new paradigms, for example to address current shortcomings in causal reasoning. I also advise being different by shifting one's focus from *how* to learn to *what* and *when* to learn, which is the main focus of this article.

LESSONS FROM HUMAN BABIES

Babies are sometimes described as scientists in the crib (Gopnik et al., 1999). In contrast to when they will later be in grade school and teachers assign them curricula so that they can focus on *how* to learn, in their capacity as baby scientists they need to actively decide *what* to learn. This contrast has been emphasized prominently by the work of Adolph et al. on toddlers' acquisition of walking skills (for example, Adolph & Hoch, 2019). Traditionally, psychologists have enticed babies to walk straight on treadmills or instrumented mats as a way of focusing on how they learn to walk. On the other hand, Adolph and colleagues have eschewed such an approach in favor of studying babies in natural settings, where the babies choose for themselves *when* and where to walk, often moving along curved or erratic paths, and often stopping, sitting, falling, and so forth. Their premise is that the choice of what and when to learn is at least as important as how to learn.

This insight has also had direct relevance to robot learning, via a comparison between how babies learn to walk with how robots learn to walk. Specifically, in our efforts to train world champion robot soccer agents for the RoboCup 3D simulation league (MacAlpine et al., 2022), we also observed that it is important for robots to train on varied paths beyond just straight forward trajectories. In Ossmy et al. (2023), we trained simulated humanoid robots on paths that babies were observed to follow naturally, leading to much greater robustness than hand-coded, regimented trajectories.

OPPORTUNITIES TO IDENTIFY WHAT TO LEARN

Within Artificial Intelligence (AI), there are several opportunities for focusing on what to learn. While it has not been an explicit focus of my lab's research, over the years we have contributed to this endeavor along four separate dimensions, each of which provides many more opportunities for follow-up research. Specifically, it is useful for an agent to focus on what to learn under at least the following four conditions:

1. When the task is too difficult or the world is too big
2. When the agent's future tasks are unknown
3. When there are other agents in the environment
4. When humans react to the agent's actions

In the remainder of this section, I touch on each of these conditions in turn. While there has been relevant work in these areas from many research labs, I focus mainly on illustrative work from the Learning Agents Research Group (LARG) at the University of Texas at Austin.

Difficult tasks or big worlds

Many tasks are too difficult to learn in one shot, end to end. For example, when teaching a child to play chess, one often starts with simpler games, such as tic-tac-toe and checkers. When doing so, the assumption is that the child will be able to *transfer* knowledge from the simpler domain to the more complex setting. Transfer learning is the area of Machine Learning that focuses on imbuing autonomous agents with the similar ability to learn from simpler tasks en route to learning more difficult tasks (Taylor & Stone, 2009).

While transfer learning relies on task simplification, another approach to dealing with difficult problems is via task *decomposition*. For example, *layered learning* has been used to decompose complex robot soccer behaviors into more easily learnable skills (MacAlpine & Stone, 2018; Stone 2000). However both transfer learning and layered learning have typically relied on manual task simplification or decomposition, based on human intuition. That is, a person has generally identified *what* to learn, leaving the autonomous agent to focus on *how* to learn.

In contrast, the burgeoning area of automated curriculum learning (ACL) brings the decision of what to learn within the purview of the learning agent (Narvekar et al.,

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114. By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2020). Specifically, Narvekar and Stone (2019) identify the three main components of ACL as follows:

- **Task creation**: Identifying which simpler tasks might be helpful for eventually learning the target task. Returning to the example of teaching a child to play chess, the game QuickChess[2] provides simplified games with only pawns, then adding one new piece at a time. Task creation automates the process of creating such simplified games (or more generally, tasks).
- **Task sequencing**: Identifying a useful sequence for learning the simplified tasks. This step automates the process of deciding the order in which the simplified tasks should be learned. For example, should the child learn tic-tac-toe before checkers? Or vice versa?
- **Transfer learning**: Finally, transfer learning provides the ability to leverage learning in the sequence of simplified tasks toward more efficient learning of the eventual target task.

ACL can be seen as a sort of meta-learning problem, where the goal is to create a teacher agent that can create and sequence tasks for a student agent based on its current capabilities. This problem can be formalized as requiring a meta-teacher to generate the curriculum which will assist the student to train toward any given target task from a task distribution based on the similarity of these tasks to one another. To this end, Xu et al. (2023) introduced a model-based meta automatic curriculum learning algorithm (MM-ACL) that learns to predict the performance improvement on one task when the student is trained on another, given the current internal learning state of the student. This predictor can then be used to generate curricula for different target tasks. Empirical results demonstrate that MM-ACL outperforms the state-of-the-art curriculum learning algorithms in a grid-world domain and a more complex visual-based navigation domain in terms of sample efficiency. The key novelty of this approach lies in its ability to select a dynamic sequence of tasks depending on what concepts or skills the learning agent has mastered and which ones it is still struggling with. Thus different curricula may emerge for different students.

Another opportunity for identifying what to learn relates to the concept of *Event Tables* in RL (Kompella et al., 2023). Experience replay (ER) is a crucial component of many deep RL systems. However, uniform sampling from an ER buffer can lead to slow convergence and unstable asymptotic behaviors. Motivated by the challenge of creating superhuman car racing agents for the high-fidelity simulator/game Gran Turismo (Wurman et al., 2022), Stratified Sampling from Event Tables (SSET) partitions an ER buffer into Event Tables, each capturing important subsequences of optimal behavior. Empirical results in challenging MiniGrid domains, benchmark RL environments, and Gran Turismo demonstrate the advantages and versatility of SSET over existing ER buffer sampling approaches (Kompella et al., 2023). Automating the selection of which behavior subsequences qualify as important enough for capturing in their own Event Table is a ripe opportunity for identifying *what* to learn.

Finally, as an alternative to focusing on what *to* learn, the TEXPLORE algorithm (Hester & Stone, 2013) instead identifies what *not to* learn. TEXPLORE is designed for big worlds, defined as environments that have too many states for an agent to explore exhaustively. As a result, the agent must actively decide where *never* to explore. For example, as people, the world we live in is too large to explore fully. We cannot visit every city before deciding where to live. And many of us cannot even try every restaurant in our home city when choosing a favorite. Instead, we must actively choose where never to go. TEXPLORE is a model-based RL method that learns a random forest model of the domain which generalizes dynamics to unseen states. The agent explores states that are promising for the final policy, while ignoring states that do not appear promising.

In summary, ACL, Event Tables, and TEXPLORE exemplify opportunities to identify what to learn in difficult tasks and/or big worlds.

Unknown future tasks

Toward the ambition for robust, general-purpose service robots in the home, a perspective that often inspires me is that they will generally have two learning phases. The first phase can be seen as what happens in the factory, prior to deployment. The second phase can be seen as what happens in the home, after deployment.

The second phase is more or less the realm of standard RL and/or imitation learning. The robot will need to learn from interaction in the physical world and/or via observation. Algorithmic innovations tend to focus on *how* to learn efficiently in such a setting.

In contrast, during the first phase, there is a great opportunity for the robot to identify for itself *what* to learn. Without as much of a limit on data and computation, but also without full knowledge of the tasks it will be asked to perform once it is deployed, the robot ought to focus on learning priors and representations.

One promising area of enquiry toward this end is on causal representations for RL, such as Causal Bisimulation Modeling (CBM) (Wang et al., 2024b). Two desiderata of RL algorithms are the ability to learn from relatively little experience and the ability to learn policies that generalize to a range of problem specifications. In factored state spaces, one approach toward achieving both goals

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114. By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is to learn state abstractions, which only keep the necessary variables for learning the tasks at hand. CBM is a method that does so by learning the causal relationships in the dynamics and reward functions for each task to derive a minimal, task-specific abstraction. CBM leverages and improves implicit modeling to train a high-fidelity causal dynamics model that can be reused for all tasks in the same environment. Empirical validation on manipulation environments and DeepMind Control Suite reveals that CBM's learned implicit dynamics models identify the underlying causal relationships and state abstractions more accurately than explicit world models.

ELDEN, Exploration via Local Dependencies, is another approach that leverages discovery of causal dependencies among actions and effects for efficient RL (Wang et al., 2023). Tasks with large state spaces and sparse reward functions present a longstanding challenge to RL. In these tasks, an agent needs to explore the state space efficiently until it finds a reward. To deal with this problem, the community has proposed to augment the reward function with intrinsic reward, a bonus signal that encourages the agent to visit *interesting* states. ELDEN introduces a new way of defining interesting states for environments with factored state spaces and complex chained dependencies, where an agent's actions may change the value of one entity that, subsequently, may affect the value of another entity. Our insight is that, in these environments, interesting states for exploration are states where the agent is uncertain whether (as opposed to how) entities such as the agent or objects have some influence on each other. ELDEN introduces a novel intrinsic reward that encourages the discovery of new interactions between entities. ELDEN utilizes the partial derivative of the learned dynamics function to model the local dependencies between entities accurately and computationally efficiently. The uncertainty of the predicted dependencies is then used as an intrinsic reward to encourage exploration toward new interactions. We evaluated the performance of ELDEN on four different domains with complex dependencies, ranging from 2D grid worlds to 3D robotic tasks. In all domains, ELDEN correctly identifies local dependencies and learns successful policies, significantly outperforming previous state-of-the-art exploration methods.

Causal modeling can also help with skill discovery, during the pre-deployment learning phase. Unsupervised skill discovery carries the promise that an intelligent agent can learn reusable skills through autonomous, reward-free environment interaction. Existing unsupervised skill discovery methods learn skills by encouraging distinguishable behaviors that cover diverse states. However, in complex environments with many state factors (for example, household environments with many objects), learning skills that cover all possible states is impossible, and naively encouraging state diversity often leads to simple skills that are not ideal for solving downstream tasks. Skill Discovery from Local Dependencies (SkiLD) leverages state factorization as a natural inductive bias to guide the skill learning process (Wang et al., 2024a). The key intuition guiding SkiLD is that skills that induce diverse interactions between state factors are often valuable for solving downstream tasks. To this end, SkiLD develops a novel skill learning objective that explicitly encourages the mastering of skills that effectively induce different interactions within an environment. We evaluated SkiLD in several domains with challenging, long-horizon sparse reward tasks including a realistic simulated household robot domain, where SkiLD successfully learned skills with clear semantic meaning and showed superior performance compared to existing unsupervised RL methods that only maximize state coverage.

Another thing that can be learned during the pre-deployment phase is an efficient abstract action space for enabling RL on a real robot in the real world. Direct RL in the real world demands both safe exploration and high sample efficiency, which are difficult to achieve in practice. Sim-to-real RL, on the other hand, is often brittle due to the reality gap. SLAC is a recent method that renders real-world RL feasible for complex embodiments by leveraging a low-fidelity simulator to pretrain a task-agnostic latent action space (Hu et al., 2025). SLAC trains this latent action space via a customized unsupervised skill discovery method designed to promote temporal abstraction, disentanglement, and safety, thereby facilitating efficient downstream learning. Once a latent action space is learned, SLAC uses it as the action interface for a novel off-policy RL algorithm to autonomously learn downstream tasks through real-world interactions. Notably, SLAC learns contact-rich whole-body tasks in under an hour of real-world interactions, without relying on any demonstrations or hand-crafted behavior priors.

This notion of learning priors and representations during a pre-deployment phase has come to be known in the RL community as "Unsupervised RL" (Agarwal et al., 2025) to indicate that learning happens prior to having any supervisory reward function. In the extreme, after exploring an environment, an intelligent agent ought to be able to transfer its knowledge to most downstream tasks within that environment without additional interactions. Referred to as "zero-shot learning", this ability remains elusive for general-purpose RL algorithms. While recent work has attempted to produce zero-shot RL agents, it tends to make assumptions about the nature of the tasks or the structure of the Markov Decision Process (MDP). To mitigate this requirement, we introduced Proto Successor Measure (PSM): the basis set for all possible behaviors of an RL agent in a dynamical system (Agarwal et al.,

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114, By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2025). We prove that any possible behavior (represented using visitation distributions) can be represented using an affine combination of these policy-independent basis functions. Given a reward function at test time, we simply need to find the right set of linear weights to combine these bases corresponding to the optimal policy. We derive a practical algorithm to learn these basis functions using reward-free interaction data from the environment and show that PSM can produce the near-optimal policy at deployment time for any given reward function without additional environmental interactions.

PSM can then be leveraged to define post-deployment tasks entirely via natural language. The reward hypothesis (Silver et al., 2021) states that all goals and purposes can be understood as the maximization of a received scalar reward signal. However, in practice, defining such a reward signal is notoriously difficult, as humans are often unable to predict the optimal behavior corresponding to a reward function. Natural language offers an intuitive alternative for instructing RL agents, yet previous language-conditioned approaches either require costly supervision or test-time training given a language instruction. Leveraging PSM, we introduce a new approach that uses a pretrained RL agent trained using only unlabeled, offline interactions without task-specific supervision or labeled trajectories to get zero-shot test-time policy inference from arbitrary natural language instructions. We introduce a framework comprising three steps: imagine, project, and imitate (Sikchi et al., 2025). First, the agent imagines a sequence of observations corresponding to the provided language description using video generative models. Next, these imagined observations are projected into the target environment domain. Finally, an agent pretrained in the target environment with unsupervised RL (PSM) instantly imitates the projected observation sequence through a closed-form solution. To the best of our knowledge, our method, RLZero, is the first approach to show direct language-to-behavior generation abilities on a variety of tasks and environments without any in-domain supervision. We further show that components of RLZero can be used to generate policies zero-shot from cross-embodied videos, such as those available on YouTube, even for complex embodiments such as humanoids.

Up to this point in this section, I have considered the situation in which there is a clear division between the pre-deployment and post-deployment learning phase. But that division is not always so clear. In the related paradigm of Lifelong Learning, or Continual Learning, an agent keeps learning from a never-ending stream of experiences (Ring, 1994). Unlike traditional lifelong learning problems in image and text domains, which primarily involve the transfer of declarative knowledge of entities and concepts, lifelong learning in decision-making (LLDM) also necessitates the transfer of procedural knowledge, such as actions and behaviors. To advance research in LLDM, we introduced LIBERO, a novel benchmark of lifelong learning for robot manipulation (Liu et al., 2023). Specifically, LIBERO highlights five key research topics in LLDM: (1) how to efficiently transfer declarative knowledge, procedural knowledge, or the mixture of both; (2) how to design effective policy architectures and (3) effective algorithms for LLDM; (4) the robustness of a lifelong learner with respect to task ordering; and (5) the effect of model pretraining for LLDM. We developed an extensible procedural generation pipeline that can in principle generate infinitely many tasks. For benchmarking purposes, we created four task suites (130 tasks in total) that we use to investigate the above-mentioned research topics. To support sample-efficient learning, we provide high-quality human-teleoperated demonstration data for all tasks. Our extensive experiments uncovered several unexpected discoveries: sequential fine-tuning outperforms existing lifelong learning methods in forward transfer, no single visual encoder architecture excels at all types of knowledge transfer, and naive supervised pretraining can hinder agents' performance in the subsequent LLDM. LIBERO has since become a standard, widely used benchmark for LLDM.

Multiple (unknown) agents

As autonomous agents proliferate in the real world, both in software and robotic settings, they will increasingly need to band together for cooperative activities with previously unfamiliar teammates. In such ad hoc team settings, team strategies cannot be developed a priori. Rather, an agent must be prepared to cooperate with many types of teammates: it must collaborate without pre-coordination. An AAAI 2010 paper challenged the AI community to develop theory and to implement prototypes of ad hoc team agents (Stone et al., 2010). It defines the concept of ad hoc team agents, specifies an evaluation paradigm, and provides examples of possible theoretical and empirical approaches to the challenge.

Ad hoc teamwork (AHT) can be seen as a specialization of the Unsupervised RL problem to a multiagent setting. In the pre-deployment phase, an agent has the opportunity to learn from lots of different possible teammates in lots of possible environments, with the goal of building representations and priors pertaining to likely types of teammates it may encounter once deployed. Then, during the deployment phase, it must adapt to its current teammates with very little additional training.

A recent paper frames this challenge directly in terms of deciding *what* to learn, by actively choosing training part-

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114. By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ners who are likely to teach the agent new coordination skills. Most existing AHT approaches adopt a two-stage pipeline, where first, a fixed population of teammates is generated with the idea that they should be representative of the teammates that will be seen at deployment time, and second, an AHT agent is trained to collaborate well with agents in the population. To date, the research community has focused on designing separate algorithms for each stage. This separation has led to algorithms that generate teammates with limited coverage of possible behaviors, and that ignore whether the generated teammates are easy to learn from for the AHT agent. Furthermore, algorithms for training AHT agents typically treat the set of training teammates as static, thus attempting to generalize to previously unseen partner agents without assuming any control over the set of training teammates. In contrast, in order to bring *what* to learn within the purview of the learning agent, we introduced a unified framework for AHT by reformulating the problem as an open-ended learning process between an AHT agent and an adversarial teammate generator. Specifically, we introduced ROTATE, a regret-driven, open-ended training algorithm that alternates between improving the AHT agent and generating teammates that probe its deficiencies (Wang et al., 2025). Experiments across diverse two-player environments demonstrate that ROTATE significantly outperforms baselines at generalizing to an unseen set of evaluation teammates, thus establishing a new standard for robust and generalizable teamwork.

I hope and expect to see much more future research that combines the challenge of identifying *what* to learn with multiagent systems.

Human input

Finally, there are opportunities for identifying *what* to learn when a learning agent has the opportunity to interact with people in its environment. Should it copy their demonstrations? Provide opportunities to intervene with targeted corrections? Pay attention to explicit or implicit feedback regarding its actions? We have considered all of those modalities in our work.

Reinforcement Learning from Human Feedback (RLHF) has recently played an important role in the development of large language models, specifically providing the crucial post-training that turned GPT-3 into GPT-3.5, also known as ChatGPT. But the first example of RLHF was introduced more than a decade earlier in the guise of the TAMER framework (Knox & Stone, 2009), motivated by the ambition to reduce sample complexity of learning a task via knowledge transfer from humans to agents. Ideally, methods of transfer should be accessible to anyone with task knowledge, regardless of that person's expertise in programming and AI. The "Training an Agent Manually via Evaluative Reinforcement," or TAMER, framework allows a human trainer to interactively shape an agent's policy via reinforcement signals. Differing from previous approaches to interactive shaping, a TAMER agent models the human's reinforcement and exploits its model by choosing actions expected to be most highly reinforced.

Since empirical results showed that TAMER allows for much faster learning than traditional RL, but ultimately converges to worse final behavior, TAMER was later extended to be able to learn from both manual and environmental feedback (Knox & Stone, 2012). Several years later, it was extended to take visual input as DeepTAMER (Warnell et al., 2018).

While TAMER focuses only on learning from explicit human positive and negative feedback, Adaptive Planner Parameter Learning (APPL) (Xiao et al., 2022) also incorporates learning from demonstrations and interventions, specifically for autonomous robot navigation in constrained spaces. While current autonomous navigation systems allow robots to successfully drive themselves from one point to another in specific environments, they typically require extensive manual parameter re-tuning by human robotics experts in order to function in new environments. Furthermore, even for just one complex environment, a single set of fine-tuned parameters may not work well in different regions of that environment. These problems prohibit reliable mobile robot deployment by non-expert users. APPL leverages non-expert human interaction via several modalities — including teleoperated demonstrations, corrective interventions, and evaluative feedback — and also unsupervised RL to learn a parameter policy that can dynamically adjust the parameters of classical navigation systems in response to changes in the environment. APPL inherits safety and explainability from classical navigation systems while also enjoying the benefits of machine learning, that is, the ability to adapt and improve from experience.

Perhaps the most challenging opportunity for identifying what to learn from people comes from their implicit reactions, such as gestures, facial expressions, and vocalizations. These signals comprise an abundant, naturally occurring channel of information that humans provide during interactions. A robot or other agent could leverage an understanding of such implicit human feedback to improve its task performance at no cost to the human. This approach contrasts with common agent teaching methods based on demonstrations, critiques, or other guidance that need to be attentively and intentionally provided. In a 2020 paper, we first defined the general problem of learning from implicit human feedback and then addressed

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114, By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

this problem through a novel data-driven framework, EMPATHIC (Cui et al., 2020). This two-stage method consists of (1) mapping implicit human feedback to relevant task statistics such as reward, optimality, and advantage; and (2) using such a mapping to learn a task. We instantiated the first stage and three second-stage evaluations of the learned mapping. To do so, we collected a dataset of human facial reactions while participants observed an agent executing a sub-optimal policy for a prescribed training task. We trained a deep neural network on this data and demonstrated its ability to (1) infer relative reward ranking of events in the training task from prerecorded human facial reactions; (2) improve the policy of an agent in the training task using live human facial reactions; and (3) transfer to a novel domain in which it evaluates robot manipulation trajectories. In the long term, we envision personalized learning agents that can attend to all different manners of a person's implicit feedback and identify for itself what to learn from them.

SOCIETAL IMPLICATIONS

As mentioned in the Introduction, I tend to advise current students to "Be Different!" which I used as a jumping off point to motivate moving beyond focusing on how neural networks learn to also innovate with regards to *what* to learn across the full range of AI paradigms. As AI systems become more generally capable, including choosing for themselves what to learn, I find it also increasingly important to encourage students to pay attention to the societal implications of their research. Of course I am not at all the first to say it, but now that AI-based technologies are permeating society, it is essential that all AI researchers become well-versed in the potential cost-benefit tradeoffs of these powerful systems.

With regards to the theme of this article, considerations of AI's societal implications are often framed as shifting the focus from what AI technologies *can* do to what they *should* do. Eventually, this consideration may also become a part of the calculus when AI systems decide for themselves what to learn. It is not too difficult to imagine situations in which an AI system could identify a useful step in an automated curriculum (as described above) that would be widely considered by people to be unwise or unethical to actually execute. I include this section both because incorporating societal implications into the calculus of *what* to learn is an important open challenge for autonomous AI systems, and because it is becoming increasingly difficult to separate one's technical work in AI from societal implications in general.

By way of example, in the talk associated with this paper, I pointed the audience to cross-disciplinary efforts I have helped lead, including the One Hundred Year Study on AI (AI100) (Stone et al., 2022), as well as courses I have developed on *Ethical AI and Robotics*[3] for graduate students from multiple disciplines, and on AI literacy, aimed at the entire undergraduate population, and accessible to all (Xu et al., 2026). We have made these materials broadly available based on our belief that every college student, and indeed all adult members of society, no matter what their field of study or occupation, ought to take one course on the basic history of, and the current state of the art of, AI. Both the AI100 Study and these courses make a point of explicitly relating AI capabilities and opportunities (what one *can* do with AI) to AI responsibilities and governance (what one *should* do with AI).

I also pointed out two recent publications directly pertaining to societal implications, one on the potential harms of AI-based companions, and one on curating ethical datasets. First, amid the growing prevalence of human-AI interaction, large language models and other AI-based entities increasingly provide forms of companionship to human users. Such AI companionship—that is, bonded relationships between humans and AI systems that resemble the relationships people have with family members, friends, and romantic partners—might substantially benefit humans. Yet such relationships can also do profound harm. Knox et al. (2025) proposed a framework for analyzing potential negative impacts of AI companionship by identifying specific harmful traits of AI companions and speculatively mapping causal pathways back from these traits to possible causes and forward to potential harmful effects. We provide detailed, structured analysis of four potentially harmful traits—the absence of natural endpoints for relationships, vulnerability to product sunsetting, high attachment anxiety, and propensity to engender protectiveness—and briefly discuss fourteen others. For each trait, we propose hypotheses connecting causes—such as misaligned optimization objectives and the digital nature of AI companions—to fundamental harms—including reduced autonomy, diminished quality of human relationships, and deception. Each hypothesized causal connection identifies a target for potential empirical evaluation. Our analysis examines harms at three levels: to human partners directly, to their relationships with other humans, and to society broadly. We examine how existing law struggles to address these emerging harms, discuss potential benefits of AI companions, and conclude with design recommendations for mitigating risks. This analysis offers immediate suggestions for reducing risks while laying a foundation for deeper investigation of this critical but understudied topic.

Second, computer vision is central to many AI applications, from autonomous vehicles to consumer devices. However, the data behind such technical innovations are

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114, By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

often collected with insufficient consideration of ethical concerns. This oversight has led to a reliance on datasets that lack diversity, perpetuate biases and are collected without the consent of data rights holders. These datasets compromise the fairness and accuracy of AI models and disenfranchise stakeholders. Although awareness of the problems of bias in computer vision technologies, particularly facial recognition, has become widespread, the field lacks publicly available, consensually collected datasets for evaluating bias for most tasks. In response, we introduced the Fair Human-Centric Image Benchmark (FHIBE, pronounced "Feebee"), a publicly available human image dataset implementing best practices for consent, privacy, compensation, safety, diversity and utility (Xiang et al., 2025). FHIBE can be used responsibly as a fairness evaluation dataset for many human-centric computer vision tasks, including pose estimation, person segmentation, face detection and verification, and visual question answering. By leveraging comprehensive annotations capturing demographic and physical attributes, environmental factors, and instrument and pixel-level annotations, FHIBE can identify a wide variety of biases. The annotations also enable nuanced and granular bias diagnoses, enabling practitioners to understand sources of bias and mitigate potential downstream harms. FHIBE therefore represents an important step forward toward trustworthy AI, raising the bar for fairness benchmarks and providing a road map for responsible data curation in AI.

While all of the examples presented in this section assume that humans are actively in charge of assessing societal implications of AI technologies, it is an important research challenge to imbue AI systems themselves with the ability to make such judgements in a way that is aligned with the values of human stakeholders, especially when these systems are deciding for themselves *what* to learn.

CONCLUSION

This article, based closely on an AAAI-2026 keynote talk with the same title, focuses mainly on contributions from the UT Austin LARG pertaining to opportunities for identifying *what* to learn in an autonomous learning system. Categorizing these opportunities into four classes, it provides examples that come from task decomposition or simplification for difficult problems and/or large environments; unsupervised pre-training of representations and priors; multiagent learning; and learning from human interaction.

Although the provided examples come mainly from my own lab, many other colleagues have worked on related problems, as fully documented in the related work sections of the cited papers. And more importantly, it is not yet a solved problem. I hope the article (and the associated talk) help motivate the next generation of young researchers to look beyond the current prevalent research questions regarding how neural networks learn, toward the full range of past and future AI paradigms, and especially toward enabling autonomous agents to identify for themselves *what* (and *when*) to learn.

ACKNOWLEDGMENTS

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by the National Science Foundation (FAIN-2019844, NRT-2125858), the Office of Naval Research (N00014-24-1-2550), Army Research Office (W911NF-17-2-0181, W911NF-23-2-0004, W911NF-25-1-0065), DARPA (Cooperative Agreement HR00112520004 on Ad Hoc Teamwork), Lockheed Martin, and Good Systems, a research grand challenge at the University of Texas at Austin. The views and conclusions contained in this document are those of the authors alone. Peter Stone serves as the Chief Scientist of Sony AI and receives financial compensation for that role. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article.

ENDNOTES

[1] The talk is available online at https://underline.io/embed/146013-from-how-to-learn-to-what-to-learn-in-multiagent-systems-and-robotics

[2] https://quickchess.info/

[3] https://utexas.instructure.com/courses/1339675

REFERENCES

Adolph, K. E., and J. E. Hoch. 2019. "Motor Development: Embodied, Embedded, Enculturated, and Enabling." *Annual Review of Psychology* 70(2019): 141–64.

Agarwal, S., H. Sikchi, P. Stone, and A. Zhang. 2025. "Proto Successor Measure: Representing the Behavior Space of an RL Agent." In *ICML'25: Proceedings of the 42nd International Conference on Machine Learning*, 566–86. Vancouver, Canada.

Cui, Y., Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. B. Knox. 2020. "The EMPATHIC Framework for Task Learning From Implicit Human Feedback." In *Proceedings of the 4th Conference on Robot Learning (CoRL 2020)*.

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114, By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Gopnik, A., A. N. Meltzoff, and P. K. Kuhl. 1999. *The Scientist in the Crib: Minds, Brains, and How Children Learn*. New York: Harper Collins.

Gu, A., and T. Dao. 2024. "Mamba: Linear-Time Sequence Modeling With Selective State Spaces." In *First Conference on Language Modeling*.

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine. 2018. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning With a Stochastic Actor." In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 1861–70.

Hester, T., and P. Stone. 2013. "TEX-PLORE: Real-Time Sample-Efficient Reinforcement Learning for Robots." *Machine learning* 90(3): 385–429.

Ho, J., A. Jain, and P. Abbeel. 2020. "Denoising Diffusion Probabilistic Models." *Advances in Neural Information Processing Systems* 33: 6840–51.

Hu, J., P. Stone, and R. Mart´ın-Mart´ın. 2025. "SLAC: Simulation-Pretrained Latent Action Space for Whole-Body Real-World RL." In *Conference on Robot Learning (CoRL)*.

Kambhampati, S. 2024. "Can Large Language Models Reason and Plan?" *Annals of the New York Academy of Sciences* 1534(1): 15–18.

Knox, W. B., and P. Stone. 2009. "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework." In *The Fifth International Conference on Knowledge Capture*.

Knox, W. B., and P. Stone. 2012. "Reinforcement Learning From Simultaneous Human and MDP Reward." In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Knox, W. B., K. Bradford, S. V. Castro, D. C. Ong, S. Williams, J. Romanow, C. Nations, P. Stone, and S. Baker. 2025. Harmful Traits of AI Companions.

Kompella, V., T. Walsh, S. Barrett, P. Wurman, and P. Stone. 2023. "Event Tables for Efficient Experience Replay." *Transactions on Machine Learning Research (TMLR)*.

Liu, B., Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. 2023. "Libero: Benchmarking Knowledge Transfer in Lifelong Robot Learning." In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*.

MacAlpine, P., and P. Stone. 2018. "Overlapping Layered Learning." *Artificial Intelligence* 254: 21–43.

MacAlpine, P., B. Liu, W. Macke, C. Wang, and P. Stone. 2022. "UT Austin Villa: RoboCup 2021 3D Simulation League Competition Champions." In *RoboCup 2021: Robot World Cup XXIV*, edited by Alami, R., Biswas, J., Cakmak, M., and Obst, O., 314–26. Springer International Publishing.

Narvekar, S., and P. Stone. 2019. "Learning Curriculum Policies for Reinforcement Learning." In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Narvekar, S., B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone. 2020. "Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey." *Journal of Machine Learning Research* 21(181): 1–50.

Ossmy, O., D. Han, P. MacAlpine, J. Hoch, P. Stone, and K. E. Adolph. 2023. "Walking and Falling: Using Robot Simulations to Model the Role of Errors in Infant Walking." *Developmental Science* 27: e13449.

Ring, M. B. 1994. *Continual Learning in Reinforcement Environments*. PhD diss., The University of Texas at Austin, Austin, Texas: 78712.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. *Proximal Policy Optimization Algorithms*.

Shao, Z., P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, et al. 2024. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*.

Sikchi, H., S. Agarwal, P. Jajoo, S. Parajuli, C. Chuck, M. Rudolph, P. Stone, A. Zhang, and S. Niekum. 2025. "Rlzero: Direct Policy Inference From Language Without In-Domain Supervision." In *Conference on Neural Information Processing Systems (NeurIPS)*.

Silver, D., S. Singh, D. Precup, and R. S. Sutton. 2021. "Reward is Enough." *Artificial Intelligence* 299: 103535.

Stone, P., G. A. Kaminka, S. Kraus, and J. S. Rosenschein. 2010. "Ad Hoc Autonomous Agent Teams: Collaboration Without Pre-Coordination." In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence*.

Stone, P., R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, et al. 2022. *Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence*.

Stone, P. 2000. *Layered Learning in Multiagent Systems: A Winning Approach to Robotic Soccer*. MIT Press.

Taylor, M. E., and P. Stone. 2009. "Transfer Learning for Reinforcement Learning Domains: A Survey." *Journal of Machine Learning Research* 10(1): 1633–85.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polo-sukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30: 5998–6008.

Wang, Z., J. Hu, P. Stone, and R. Mart´ın-Mart´ın. 2023. "Elden: Exploration via Local Dependencies." In *Conference on Neural Information Processing Systems*.

Wang, Z., J. Hu, C. Chuck, S. Chen, R. Mart´ın-Mart´ın, A. Zhang, S. Niekum, and P. Stone. 2024a. "Skild: Unsupervised Skill Discovery Guided by Factor Interactions." In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, Z., C. Wang, X. Xiao, Y. Zhu, and P. Stone. 2024b. "Building Minimal and Reusable Causal State Abstractions for Reinforcement Learning." In *AAAI Conference on Artificial Intelligence*.

Wang, C., A. Rahman, J. Cui, Y. Sung, and P. Stone. 2025. *Rotate: Regret-Driven Open-Ended Training for Ad Hoc Teamwork*.

Warnell, G., N. Waytowich, V. Lawhern, and P. Stone. 2018. "Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces." In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Wurman, P. R., S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, et al. 2022. "Outracing Champion Gran Turismo Drivers With Deep Reinforcement Learning." *Nature* 602: 223–28.

Xiang, A., J. T. A. Andrews, R. L. Bourke, W. Thong, J. M. LaChance, T. Georgievski, A. Modas, et al. 2025. "Fair Human-Centric Image Dataset for Ethical AI Benchmarking." *Nature* 648: 97–108.

Xiao, X., Z. Wang, Z. Xu, B. Liu, G. Warnell G. Dhamankar, A. Nair, and P. Stone,. 2022. "APPL: Adaptive Planner Parameter Learning." *Robotics and Autonomous Systems* 154: 104132.

Xu, Z., Y. Zhang, S. S. Shperberg, R. Mirsky, Y. Jiang, B. Liu, and P. Stone. 2023. "Model-Based Meta Automatic Curriculum Learning." In *The Second Conference on Lifelong Learning Agents (CoLLAs)*.

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114. By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Xu, Z., K. Procko, M. Munje, K. Patterson, L. Sabatini, J. Biswas, and P. Stone. 2026. "The Essentials of AI for Life and Society: A Full-Scale AI Literacy Course Accessible to All." In *Symposium on Educational Advances in Artificial Intelligence (EAAI)*.

How to cite this article: Stone, P. 2026. "From "how" to learn to "what" to learn in multiagent systems and robotics." *AI Magazine* 47: e70114. https://doi.org/10.1002/aaai.70114

AUTHOR BIOGRAPHY

Peter Stone is the founder and director of the Learning Agents Research Group (LARG) at the University of Texas at Austin, where he also serves as Head of the School of Computing and Founding Director of Texas Robotics. His research interests include machine learning, particularly reinforcement learning, along with multiagent systems and robotics. He received his Ph.D. in Computer Science from Carnegie Mellon University in 1998 and was a Senior Technical Staff Member at AT&T Labs Research from 1999 to 2002. Stone is a Fellow of AAAI, IEEE, AAAS, and ACM, as well as an Alfred P. Sloan Research Fellow, Guggenheim Fellow, and Fulbright Scholar. His honors include the 2007 IJCAI Computers and Thought Award, the 2016 ACM/SIGAI Autonomous Agents Research Award, and the 2024 ACM/AAAI Allen Newell Award. He also serves as Chief Scientist of Sony Corporate Technology Center.

23719621, 2026, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aaai.70114. By University Of Texas at Austin, Wiley Online Library on [01/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License