Learning from Human Feedback: A Human-Compatible Perspective

Nguyen X. Khanh
UC Berkeley

Abstract

This paper takes steps toward developing a unified, rigorous framework that enables humans to effectively and efficiently adapt AI agents through natural communication. Drawing on foundational ideas from socio-cognitive science, we argue that both the diversity and efficiency of human teaching strategies are constrained by the design of the learning agent's cognitive system. This implies that, to learn effectively and efficiently from diverse forms of human feedback, an AI agent must adopt a **human-compatible** cognitive system—one that humans can readily theorize about and influence. To illustrate this principle, we examine a range of cognitive system designs for AI agents and analyze the types of learning each supports. We conclude with a proposal for future research directions.

1 Introduction

Learning from human feedback is a cornerstone of adaptive intelligence. For AI agents operating in complex, open-ended environments, fixed policies cannot anticipate every possible situation. Feedback enables these systems to improve performance, align with evolving goals, and adapt to novel conditions. In recent years, it has also become the primary mechanism for improving large language models (LLMs) and related multimodal systems. Techniques such as reinforcement learning from human feedback (RLHF) and instruction tuning have transformed LLMs from generic text predictors into capable assistants that can follow instructions, reason over multiple steps, and exhibit domain-specific expertise. This paradigm underscores the central role of human feedback—not only as a means of post-deployment correction, but as a driving force in the development of increasingly general and capable AI systems.

Among the many forms of feedback, natural language offers unique advantages. Its efficiency and naturalness for humans make it an ideal medium for transferring knowledge. Yet, despite decades of progress, the field of AI still lacks a unified and rigorous approach to learning from language feedback. Most notable milestones have been achieved through imitation learning and reinforcement learning, which provide strong theoretical guarantees but constrain humans to a limited communication channel. Existing approaches to language feedback either reduce the problem to imitation or reinforcement learning—thereby inheriting their limitations—or rely on the generalization capabilities of language models, which remain theoretically poorly understood.

We argue that the key to addressing this challenge is to precisely model the underlying mechanisms of human communication using the formalism of interactive learning theory. This paper demonstrates this approach by characterizing the essence of human communication and presenting principles for designing agents that are compatible with it.

2 Problem setting

We first formulate an iterative learning process between an AI agent and a human teacher. For simplicity, we instantiate this process in a deterministic environment, modeled by an MDP with with states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, start state s_0, horizon H, and transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. The goal is to learn a policy $\pi(y \mid s_0, x)$ that generates a distribution over *output solutions* $y \in \mathcal{A}^H$ given an *input task* $x \in \mathcal{X}$ and a start state s_0. Since the start state is always s_0, we omit it from the policy notation and simply write $\pi(y \mid x)$. As the environment is deterministic, a solution $y = (a_0, a_1, \cdots, a_{H-1})$ is essentially a sequence of actions. We refer to x as a "task" or

an "input", and y as a "solution" or an "output". The tasks x are drawn from a distribution $P_{\mathcal{X}}$. The quality of a solution is determined by a reward function $R(x, y)$.

We define the value of a policy as

$$V(\pi) \triangleq \mathbb{E}_{x\sim P_{\mathcal{X}},\, y\sim\pi(x)}\big[R(x,y)\big]. \tag{1}$$

The goal of learning is to find a policy with maximum value:

$$\max_{\pi} V(\pi)$$

A *learning process* consists of multiple rounds of interactions between the agent and the human. In each iteration, the agent receives a task x, generates a solution $y \sim \pi(x)$, and receives *feedback* f from the human. We impose no restrictions on the form of f: it may be numerical or linguistic, instructive or corrective, etc. After receiving feedback, the agent applies a *learning algorithm*, denoted by $\phi(\pi, f)$, to update its policy, generating a new policy.[1]

Learning from Human Feedback

For $t = 0, 1, 2 \ldots$

1. Agent receives task $x \sim P_0$;
2. Agent generates solution $y \sim \pi_t(x)$;
3. Human provides feedback f;
4. Agent updates policy: $\pi_{t+1} = \phi(\pi_t, f)$;

Cognitive system, architecture, and process. The *cognitive system* of an agent refers to all the mental representations and processes that enable it to think and make decisions. A *cognitive architecture* refers to the structure of a cognitive system, i.e., what its components are and how they are connected. A *cognitive process* is a mental activity that takes place within a cognitive system.

Learning framework. We use the term *learning framework* to denote the combination of a learning algorithm and a cognitive system that supports it. Viewing the cognitive system as a component of a learning framework is a key distinction of our perspective. As we will show, certain algorithms can only be executed when paired with a specific cognitive system.

Desiderata. We identify three desiderata for a learning framework:

1. **Effectiveness (near-optimal convergence).** The learning process induced by a learning framework should asymptotically result in a near-optimal policy: $\lim_{t\to\infty} V(\pi_t) \geq \max_\pi V(\pi) - \epsilon$ for a small ϵ.
2. **Efficiency.** Low "effort" is needed to reach a target policy value. Different notions of effort may be employed (e.g., number of interactions, time, or expense).
3. **Inclusiveness.** Humans use a variety of strategies to teach others. A learning framework should allow the human teacher to freely employ any of those strategies, rather than confining them to a restricted communication channel.

[1] Our formulation subsumes the special case where f is a pre-execution instruction. In this case, the agent generates a null output in the first iteration and obtains an instruction f guiding its actions in the next iteration.

3 Limitations of existing learning frameworks

In this section, We apply the formulation and desiderata introduced in the previous section to characterize and compare popular learning frameworks. Our aim is to gain a clearer understanding of their limitations and explain the motivation for moving beyond them.

First, we review two foundational frameworks: *imitation learning* and *reinforcement learning*. Imitation learning (IL) enables learning from *demonstrations*. There are two types of demonstrations. An instructive demonstration presents the solution desired by the human, which is independent of the agent's proposed solution. A corrective demonstration specifies the action that the agent should have taken in each state it visited while executing its solution; this type of feedback is therefore dependent on the agent's solution. Instructive demonstrations give rise to the behavior cloning framework [22], whereas corrective demonstrations are employed by adaptive approaches such as DAgger [24].

Reinforcement learning (RL) [27] enables learning from numbers, often called *rewards*. RL is generally considered less efficient than IL because a reward typically provides much less information about the desired behavior than a demonstration. However, in certain scenarios, rewards may require much less time and cost to obtain than demonstrations, so the verdict depends on how effort is quantified.

These days, IL and RL are mostly deployed in deep learning settings, where they are used to improve a neural network following a gradient-descent optimization approach. Thanks to the well-established theory of gradient descent, RL and IL offer strong convergence guarantees in convex optimization problems and empirically demonstrate reliable performance in non-convex problems. However, they are limited in inclusiveness, as each can only handle a single type of feedback. Moreover, the types of feedback they support are simple and low-bandwidth, constraining both the efficiency of communication and the diversity of strategies available to the human teacher.

Natural language feedback holds the promise of providing superior efficiency and inclusiveness. The argument for inclusiveness is straightforward: natural language is the predominant means of human communication. On the other hand, the promise of efficiency comes from two appealing properties of human language. First, the human language is referential: people can easily devise concise, abstract terms to substitute for lengthy descriptions, thereby reducing communication effort. Second, language-based communication is inferential: the signals that human speakers produce are only hints of what they want to convey. Listeners take these hints and use their reasoning capabilities to reconstruct the speakers' intentions. Given sufficient shared understanding of context, even a single word can convey a long story.

Recently, the advent of large language models (LLMs) has raised the question of whether the problem of learning from language feedback has already been solved. Several studies have demonstrated that many of these models can be improved simply by prompting them with free-form language feedback [17; 13; 30]. We refer to this framework as *context-based learning* (CBL). Does CBL meet the three desiderata we proposed? In terms of inclusiveness, CBL is a significant step-up from IL and RL: rather than having to communicate through actions or numbers, the framework allows the human teacher to speak free-form language and employ all the viable strategies in human-human communication. Nevertheless, the effectiveness of this approach remains questionable, as no satisfactory theory yet characterizes the generalizability of LLMs' language understanding. Empirically, Liang et al. [17] and Jin et al. [13] show that the improvement obtained from language feedback saturates after a few iterations, well before approaching optimality. With regard to efficiency, CBL can be considered efficient within the range of performance it is able to attain. However, for levels of performance beyond its reach, efficiency is non-existent.

The effectiveness of CBL can potentially be improved through additional IL and RL fine-tuning. Nevertheless, we contend that this approach can be unnecessarily costly, because the cognitive system of LLMs may be fundamentally incompatible with the intentions of human language feedback. We elaborate on this point in the next section, but put simply: human language feedback is intended to alter processes within a human-analogous cognitive system; yet it remains unclear whether LLMs possess such processes, without which the intentions of human language feedback cannot be realized.

4 Human-compatible learning

In this section, we characterize the intentions of human feedback and the type of cognitive system that can incorporate those intentions.

4.1 An overview of human communication

Giving feedback is a communicative act. Hence, to understand the nature of language feedback, we must first understand the nature of human communication. To provide a thorough description of human (cooperative) communication, we integrate three prominent lines of research.

The first line of research is theory of mind [23]. In a broad sense, theory of mind refers to the ability to mentally construct a hypothetical model of how an individual's mind directs their outward behavior. While most animals regard conspecifics as reactive systems—a straightforward mapping from input signals to output behaviors, humans postulate that there are non-trivial processes in between, operating on representations called "mental states".

Humans construct theory of mind not only to understand but also to influence others' behavior. This perspective is articulated in Michael Tomasello's seminal book Constructing a Language [21]. He distinguishes human communication from that of other animals as follows:

> " To oversimplify, **animals are aimed at the behavior and motivational states of others, whereas human symbols are aimed at the attentional and mental states of others**. It is this mental dimension that gives linguistic symbols unparalleled communicative power. "

In other words, animal communication primarily aims to regulate behavior, whereas human communication aims to shape the mind. The cognitive approach of humans is perhaps motivated by efficiency: changing a mental state can shape multiple behaviors at once. For instance, telling someone, "The floor is slippery!" prompts them to exercise greater caution in performing *all* tasks within the room. Compared to an approach that regulates each individual task performance, i.e., telling them to "do X more carefully" for every task X, the concise warning achieves a similar effect but with far less effort.

Whereas Tomasello's book portrays the intentions of human communication, pragmatics theory [11; 25] describes the mechanism by which these intentions are recognized. This theory proposes that human communication is an inferential process. Human communicative signals are merely hints of the actual intentions speakers wish to convey. Such communication is possible due to humans' exceptional reasoning capabilities. Specifically, the speaker produces a signal to convey an intention by reasoning about how the listener will interpret it. Similarly, the listener recognizes this intention by reasoning about how the speaker would convey it. This mutual recursion underlies humans' ability to infer meaning beyond the literal content of a message.

Combining these viewpoints, we propose a full characterization of the human communication process:

The human communication process

1. The speaker builds a hypothetical model of the cognitive system of the listener.
2. The speaker forms an intention to alter the mental states within that imaginary system, and reason to generate an appropriate linguistic expression to signal that intention.
3. The listener infers the speaker's intention hinted by the signal and adapts their mental states accordingly.

4.2 Human-compatible learner

Our characterization of human communication suggests that the ideal type of agent to learn from human language feedback are those whose cognitive systems are analogous to the human cognitive system. We refer to those agents as **human-compatible learners**. Moreover, we argue that human-compatibility is a requisite for achieving the three desiderata of learning laid out in section 2—namely effectiveness, efficiency, and inclusiveness. First, human-compatibility implies inclusiveness because it allows human teachers to flexibly select their teaching strategies. Second, human-compatibility is essential for effectiveness, as it is a prerequisite for incorporating human feedback precisely as intended. Finally, since human mind-shaping communication is naturally efficient, facilitating this form of communication offers improved efficiency compared to using only behavior-regulating strategies.

5 Cognitive architecture bounds learning ability

To better illustrate the influence of the cognitive system of an agent on its capability of learning from human feedback, we present three case studies featuring agents with various cognitive architectures. We demonstrate that the cognitive architecture of an agent places a constraint on the diversity and efficiency of teaching strategies. Specifically, more human-compatible cognitive architectures give rise to more inclusive and efficient learning frameworks.

5.1 Case 1: Opaque agent

This agent implements an inscrutable policy $\pi(y \mid x)$. A human teacher likely views it as a reactive system, ignoring its internal cognitive processes. With this mental model, the teacher can only directly shape the outward behavior of the agent. Two strategies are available: (1) present the desired behavior to the agent and (2) encourage or discourage a behavior of the agent. These strategies essentially translate to imitation learning (IL) and reinforcement learning (RL), respectively.

5.2 Case 2: Chain-of-thought agent

A chain-of-thought agent mimics how a human reasons by engaging in an inner monologue. This type of agent employs a policy $\pi(y, z \mid x) = \pi_0(z \mid x)\pi_1(y \mid x, z)$ that first produces a thought (chain) z before generating the final output y. The thought z is not always revealed to the human, but we assume it is expressed in natural language and that the human has observed enough examples to build a reasonable model of the thought-generation process. Such a model enables the human to predict the agent's thought even when it is hidden. Being able to make these predictions allows the human to plan communication signals that alter the thought, thereby influencing the agent's behavior.

Specifically, with a chain-of-thought agent, a human teacher can apply the following strategies.

IL or RL on thought and output. The human uses demonstrations or rewards to shape both z and y. They may employ different types of feedback for each. For example, y can be supervised with demonstrations and z with rewards.

IL or RL on output only. Alternatively, the human can view z as adaptable parameters of the policy and provide demonstrations or rewards on only y. The agent can then use gradient descent to search for the z that optimizes the learning objective. In practice, when z is a language utterance, techniques such as REINFORCE [28] are needed to enable backpropagation through discrete-token parameters.

Context-based learning (CBL). Instead of having the agent generate thoughts on its own, the human can do it for the agent. This strategy requires the agent to be pre-trained, either with IL or RL, to generate a good behavior conditioned on the teacher-provided thought (i.e., learning a good $\pi_1(y \mid x, z)$). While both use thought demonstration as feedback, CBL differs from performing IL on thought in that CBL does *not* alter the internal parameters of the trained policy. In addition, CBL typically requires thought demonstrations to come from the same distribution as the demonstrations used to learn π_1 (otherwise, this policy encounters an out-of-distribution input and may perform poorly). IL imposes no such restriction.

Interactive from activity description (ILIAD). This framework is proposed by [20]. The high-level idea is to ensure consistency between z and y. For example, if y is an undesirable behavior, then z should be an undesirable thought (e.g., $z =$ "make bad coffee" $\rightarrow y =$ actual bad coffee). In general, z must be an accurate description of y. If this consistency holds for all x, the human only needs to train the agent to produce desirable thoughts (i.e., learning $\pi_0(z \mid x)$). Due to the consistency guarantee, desirable thoughts implies that the following outputs must also be desirable.

Concretely, suppose the agent has generated (z, y) for an input x. The human feedback in ILIAD is a modified thought z' that is more consistent with y than z. Essentially, the human is telling the agent *"your thought should be z' instead of z if what you are about to do is y."* For example, if x is the command "make coffee," z is a

correct plan for making coffee (e.g., "First, I will get coffee..."), and y is a sequence of actions that produces tea, then the human feedback in this case is a correct plan for making tea—one that faithfully describes the actions y.

Because z is expressed in language, the feedback in CBL or ILIAD is essentially language feedback. Moreover, it carries a mind-shaping intention, as it aims to override z—the agent's internal thought. Hence, CBL and ILIAD exemplify learning frameworks that capture the nature of human communication, distinguishing itself from traditional approaches like IL and RL. However, both approaches must be combined with IL or RL pre-training. The main difference is that CBL requires IL or RL to learn the π_1, whereas ILIAD needs them to learn the π_0.

Both frameworks presuppose the chain-of-thought cognitive architecture. This highlights the necessity of a human-compatible cognitive system to effectively support human language feedback.

5.3 Case 3: Agent with a language-guided world model

Following [31], we consider an agent that, instead of reasoning through language, runs an actual optimization process to plan its actions. In addition to a policy $\pi(y \mid x)$, the agent possesses a *language-guided world model* $\tilde{T}(s' \mid s, a; z)$, which is an approximation of the environment transition function $T(s' \mid s, a)$. In particular, the behavior of the model is controlled by a textual context z.

For this agent, the more obvious teaching strategies are those that directly adapt the policy π, i.e., behavior-regulating approaches. IL or RL can be employed for this purpose. However, a more interesting approach is to adapt the world model—a mind-shaping approach. Before describing the approach, let us portray the cognitive process through which the world model influences the outward behavior. For simplicity, we assume that the agent has access to the reward function $R(x, y)$. This is a reasonable assumption in scenarios where this function can be expressed as a program (e.g., a video game, a programming problem graded by unit tests, a math problem graded by output only, a multiple-choice question).

We use π_x as a shorthand for $\pi(\cdot \mid x)$ and define the value of this policy under the world model as follows:

$$V(\pi_x; \tilde{T}, x) = \mathbb{E}_{y \sim P_{\pi_x, \tilde{T}}}[R(x, y)]$$

where $P_{\pi_x, \tilde{T}}$ denotes the distribution over action sequences y obtained by executing π_x and following the dynamics specified by $\tilde{T}$.

Given an input x, the agent generates an output as follows. First, it derives its policy π_x by approximately finding the maximizer of $V(\pi; \tilde{T}, x)$, potentially using optimization methods like gradient descent:

$$\pi_x \approx \arg\max_{\pi'} V(\pi'; \tilde{T}, x)$$

Next, it samples an output from this policy $y \sim \pi_x$. With this process, the world model effectively influences the distribution of the final output.

Teaching by altering the world model is remarkably efficient. In our formulation, the same world model is used to compute π_x for *every* x. Consequently, improving the accuracy of the world model enhances the solution quality for all tasks at once. This approach effectively enables simultaneous learning of multiple tasks using an amount of data that grows *independently* of the number of tasks. In other words, with this approach, adding more tasks does not necessarily require collecting more data. It is particularly advantageous for teaching a large number of tasks that share many subtasks.

A language-guided world model behaves similarly to a chain-of-thought policy: it is a function that depends on a language input. Hence, we can employ all the strategies described in the previous section to train this model.

6 Future directions

Going forward, we aim to develop frameworks that can incorporate a wider range of teaching intentions. Ultimately, our goal is to integrate fragmented frameworks into a unified framework that is provably effective and efficient. We believe the following directions hold strong potential for moving us toward this goal.

Build human-like cognitive system with language-guided, reusable modules. As human reasoning capabilities have been optimized for communication within the same species, the most human-compatible cognitive system would closely mirror the human cognitive system. While the human cognitive system has not yet been fully

understood, recent proposals such as [16] and [26], grounded in prominent cognitive science theories, offer promising starting points for further development.

Two additional principles are potentially helpful in this endeavor. First, incorporating modules that are shared across multiple cognitive processes can greatly improve learning efficiency, as exemplified by the use of a shared world model. Second, modeling each module as a language-conditioned function enables it to be adapted through language feedback—whether via CBL or ILIAD—ultimately enhancing inclusiveness. In building language-modulated models, besides work on language-conditioned world models [18; 19; 3; 4; 32], numerous studies on language-conditioned reward functions [8] offer useful insights and techniques that can be leveraged by future work.

Connect with the MDP framework to derive theoretical guarantees. A well-formulated learning framework should foster rigorous theoretical analysis, enabling formal proofs of its effectiveness and efficiency. The classical MDP framework and its derivatives provide a well-suited foundation for two main reasons. First, they have been extensively studied and adapted, offering a rich collection of theoretical tools and results. Second, the core components of this framework align naturally with the primary human mental states described by cognitive scientists: the transition function corresponds to beliefs about the environment, the reward function reflects desires, and the policy represents intentions.

Nevertheless, in most MDP-based formulations, the agent is treated as a black box, with no explicit specification of its cognitive system. The I-POMDP framework [9] and other works on belief inference depart from this trend by specifying a reasoning process for the agent. However, a true cognitive system, comprising multiple foundational mental states analogous to those of humans, remains absent. The main challenge for future work is to derive insightful theoretical models for agents equipped with a human-like cognitive system. Taxonomies such as ATOMS [1; 29; 10; 15] provide useful guidance on which types of mental states to include.

Exploit second-order influence. Directly shaping the agent's behavior can be regarded as a zero-order teaching intention, whereas altering its mental states to induce behavioral change is essentially first-order. Second-order teaching intentions involve influencing two intermediate mental states. Consider a process in which, if the policy for a particular task is modified so that it becomes suboptimal with respect to the world model, the world model can adjust itself to remain consistent with the updated policy. This, in turn, affects the policies of other tasks, which adapt to stay (nearly) optimal with respect to the revised world model. In this way, feedback provided to a single policy can efficiently adapt multiple policies. To realize this approach, we need optimization methods capable of propagating learning feedback through a complex cognitive process involving multiple intermediate mental states. Variational approaches (e.g., [5]) could provide the fundamental principles to tackle this problem.

Adopt a dual-system architecture to enhance speed and robustness. The second and third case studies presented in section 5 illustrate two different types of cognitive system: one reasoning based on language and the other reasoning through a mathematical optimization. These two systems are complementary: the language-based system is fast and intuitive but potentially fragile due to its reliance on pattern following, while the optimization-based can be more rigorous but also more computationally demanding. Humans combine two cognitive systems with similar properties to enjoy the best of both worlds, a design known as the *fast-and-slow* architecture [14]. We argue that future AI agents could benefit from a similar architecture. The central challenge lies in effectively coordinating the two systems: determining when each should make decisions and how they should interact to transfer knowledge between them.

Inferring human intentions. So far, we have only considered the problem of incorporating human teaching intentions. In practice, however, these intentions are rarely explicit; they must be inferred from language utterances, which are often indirect, ambiguous, and context-dependent. Therefore, equipping agents with strong reasoning capabilities is essential for developing robust communication skills. The key challenges of this problem have been discussed at length in various position and survey papers [2; 7; 12; 6]. Addressing these challenges will require integrating advances in language grounding, theory-of-mind modeling, and continual learning into a unified framework that can reliably infer and act upon human intentions in diverse real-world settings.

References

[1] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905, 2020.

[2] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

[3] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[4] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

[5] Benjamin Eysenbach, Alexander Khazatsky, Sergey Levine, and Russ R Salakhutdinov. Mismatched no more: Joint model-policy optimization for model-based rl. *Advances in Neural Information Processing Systems*, 35: 23230–23243, 2022.

[6] Jaime F Fisac, Monica A Gates, Jessica B Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S Shankar Sastry, Thomas L Griffiths, and Anca D Dragan. Pragmatic-pedagogic value alignment. In *Robotics research: the 18th international symposium Isrr*, pp. 49–57. Springer, 2019.

[7] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*, 2022.

[8] Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019.

[9] Piotr J Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[10] Prasoon Goyal, Scott Niekum, and Raymond J Mooney. Using natural language for reward shaping in reinforcement learning. *arXiv preprint arXiv:1903.02020*, 2019.

[11] H Paul Grice. 1975 logic and conversation. *The Philosophy of Language*, 1990.

[12] Guy Hoffman, Tapomayukh Bhattacharjee, and Stefanos Nikolaidis. Inferring human intent and predicting human action in human–robot collaboration. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2024.

[13] Di Jin, Shikib Mehri, Devamanyu Hazarika, Aishwarya Padmakumar, Sungjin Lee, Yang Liu, and Mahdi Namazifar. Data-efficient alignment of large language models with human feedback through natural language. *arXiv preprint arXiv:2311.14543*, 2023.

[14] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[15] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. *arXiv preprint arXiv:2303.00001*, 2023.

[16] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62 (1):1–62, 2022.

[17] Jacky Liang, Fei Xia, Wenhao Yu, Andy Zeng, Montserrat Gonzalez Arenas, Maria Attarian, Maria Bauza, Matthew Bennice, Alex Bewley, Adil Dostmohamed, et al. Learning to learn faster from human feedback with language model predictive control. *arXiv preprint arXiv:2402.11450*, 2024.

[18] Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language. *arXiv preprint arXiv:2308.01399*, 2023.

[19] Iman Nematollahi, Branton DeMoss, Akshay L Chandra, Nick Hawes, Wolfram Burgard, and Ingmar Posner. Lumos: Language-conditioned imitation learning with world models. *arXiv preprint arXiv:2503.10370*, 2025.

[20] Khanh X Nguyen, Dipendra Misra, Robert Schapire, Miroslav Dudík, and Patrick Shafto. Interactive learning from activity description. In *International Conference on Machine Learning*, pp. 8096–8108. PMLR, 2021.

[21] JULIAN M PINE. Tomasello, m., constructing a language: a usage-based theory of language acquisition. cambridge, ma: Harvard university press, 2003. pp. 388. hardback,£ 29.95. isbn 0-674-01030-2. *Journal of Child Language*, 32(3):697–702, 2005.

[22] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[23] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[24] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

[25] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

[26] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

[27] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[28] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[29] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. *arXiv preprint arXiv:2309.11489*, 2023.

[30] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[31] Alex Zhang, Khanh Nguyen, Jens Tuyls, Albert Lin, and Karthik Narasimhan. Language-guided world models: A model-based approach to ai control. *arXiv preprint arXiv:2402.01695*, 2024.

[32] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.