Position: Human-Compatible Cognitive Systems Are a Prerequisite for Incorporating Natural Human Feedback

Nguyen X. Khanh [1]

Abstract

Natural language feedback carries *mind-shaping* intentions: speakers aim to modify beliefs, goals, or, in general, cognitive processes of listeners. Such intentions can be efficiently and accurately internalized only if the listener possesses a human-compatible cognitive system—namely, internal mental states and processes that humans can naturally theorize about, reason over, and influence through natural language. We compare several cognitive architectures for AI agents and show that increasing human-compatibility expands the diversity, efficiency, and potentially effectiveness of the teaching strategies that can be employed by the corresponding agents. This yields our central thesis: to fully harness natural language feedback, AI agents must be built with human-compatible cognitive systems. We outline key research challenges toward such systems and propose principles for a unified, theoretically grounded framework for learning from language feedback.

1. Introduction

Learning from human feedback is central to building adaptive and aligned AI systems. In open-ended, real-world environments, agents inevitably encounter situations that cannot be fully anticipated during training. Human feedback provides a powerful mechanism for correcting mistakes, refining capabilities, and enabling continual improvement. Recently, this form of learning has become the dominant approach for aligning and strengthening large language models (LLMs): techniques such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) and supervised fine-tuning (SFT) have transformed LLMs from generic sequence predictors into broadly capable assistants with emergent reasoning and task-generalization abilities.

Among the various modalities of feedback, natural language is uniquely promising. It is the most expressive, efficient, and intuitive communication channel available to humans. Despite its power, however, the AI community still lacks a unified, rigorous framework for learning from natural language feedback. Most existing paradigms reduce language feedback to either imitation learning (Scheurer et al., 2023; Jin et al., 2023; Chen et al., 2024) or reinforcement learning (Goyal et al., 2019; Feng et al., 2024), thereby inheriting the limitations of these frameworks. More recent context-based approaches (Akyürek et al., 2023; Madaan et al., 2023; Scheurer et al., 2023; Chen et al., 2024) leverage the generalization abilities of LLMs to incorporate free-form language feedback directly into the prompt. While extremely flexible, these methods generally lack theoretical convergence guarantees and do not fully capture the cognitive mechanisms underlying human communication.

In this work, we argue that unlocking the full potential of natural language feedback requires modeling the underlying cognitive mechanisms of human communication, not just the signals they produce. Drawing on foundational work in socio-cognitive science, we highlight a key insight: the strategies humans use to teach are deeply shaped by their assumptions about the learner's underlying cognitive system. More specifically, human communication is inherently *mind-shaping*: speakers form intentions about how to alter the listener's internal cognitive processes—how their beliefs, desires, intentions, and other mental representations interact to influence downstream behaviors—and generate linguistic expressions to influence those processes to achieve intended behavioral outcomes. These capability presupposes that the listener possesses human-like internal processes that can be reasoned about and adapted with language. Existing learning frameworks rarely model such internal cognitive structure, limiting both their generalizability, efficiency, and the range of teaching strategies they can accommodate.

This observation motivates the central thesis of the paper: **to learn effectively and efficiently from diverse types of natural language feedback, an AI agent must adopt a human-compatible cognitive system—one that humans can readily theorize about, reason over, and influence with natural language**. We formalize this idea within an interactive learning framework and show how different cog-

[1]UC Berkeley. Correspondence to: Nguyen X. Khanh <nguyenxuankhanhm@gmail.com>.

Preprint. January 29, 2026.

nitive architectures support different classes of teaching strategies. By comparing different cognitive architectures, we illustrate how increasing cognitive human-compatibility expands the effectiveness, efficiency, and inclusiveness of feedback learning capabilities. This perspective helps explain why current approaches may plateau and suggests a design principle for future systems: align the agent's internal cognitive structure with the assumptions implicit in human communication.

2. Problem setting

We first formulate an interactive learning process between an AI agent and a human teacher. We model an MDP environment with states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, start state $s_0 \in \mathcal{S}$, horizon H, and transition function $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ where $\Delta(\mathcal{S})$ denotes the space of probability distributions over $\mathcal{S}$. A task specification $x \in \mathcal{X}$ describes a task in the environment. We will refer to x simply as a task. Tasks are drawn from a distribution $P_{\mathcal{X}}$.

The agent maintains a policy $\pi(a \mid s_0, x)$ that generates a distribution over actions given an *input task* $x \in \mathcal{X}$ and the start state s_0. Since the start state is always s_0, we simply write the policy as $\pi(a \mid x)$. We will use π_x as a shorthand for the conditional policy $\pi(\cdot \mid x)$. A solution $y = (s_0, a_0, s_1, a_1, \cdots, a_{H-1}, s_H)$ is a sequence of states and actions, generated by executing the policy in the environment. The quality of a solution is determined by a reward function $R(x, y)$.

We define the value of a policy as

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

where $P_{\pi_x, T}$ is the distribution over solutions generated by executing π_x in the environment specified by T.

The goal of learning is to find a policy with maximum value within the space of feasible policies Π:

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

A *learning process* consists of multiple rounds of interactions between an agent learner and a human teacher. In each iteration, the agent receives a task x, generates a solution y, and receives *feedback* $f = h(x, y)$ 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]

[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.

Algorithm 1 Learning from Human Feedback Process

Initialize policy $\pi^{(0)}$
for $t = 0, 1, 2, \ldots$ **do**
  Agent receives task $x \sim P_{\mathcal{X}}$
  Agent generates solution $y \sim P_{\pi_x^{(t)}, T}$
  Human provides feedback $f = h(x, y)$
  Agent updates policy: $\pi^{(t+1)} = \phi(\pi^{(t)}, f)$
end for

Cognitive system, architecture, and process. The *cognitive system* of an agent refers to the mental representations (beliefs, desires, intentions, emotions, etc.) and the internal processes operating on these representations that enable the agent 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 interact. A *cognitive process* is an activity taking place within a cognitive system.

Learning framework. We use the term *learning framework* to denote Algorithm 1 instantiated with a specific learning algorithm and agent cognitive system. Viewing the agent cognitive system as a component of a learning framework is a key distinction of our perspective. As we will later demonstrate, certain algorithms require specific cognitive systems to function.

Desiderata. We identify three desiderata for a learning framework:

1. **Effectiveness.** The learning process instantiated by the framework asymptotically produces a near-optimal policy: $\lim_{t \to \infty} V(\pi_t) \geq \max_{\pi \in \Pi} V(\pi) - \epsilon$ for a small $\epsilon > 0$.
2. **Efficiency.** If a policy value is attainable, low "effort" is required to reach it. Different notions of effort may be employed (e.g., number of interactions, time, expense, cognitive load, or a mixture of them).
3. **Inclusiveness.** Humans use a variety of strategies to teach others. An ideal learning framework should allow the human teacher to freely employ any of those natural strategies, rather than confining them to a restricted communication channel.

We prefer frameworks that *provably* meet these desiderata, where the proofs can be logical or mathematical. When it is infeasible to provide such proofs, substantial empirical evidence is required.

3. Alternative views: 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 help readers gain a clearer understanding of their limitations and explain the motivation for moving beyond them. A brief summary of this section is provided in Table 2.

First, we review two foundational frameworks: *imitation learning* (IL) and *reinforcement learning* (RL). IL enables learning from *demonstrations*—sequences of actions belonging to the agent's action space. Meanwhile, RL (Sutton et al., 1998) enables learning from numbers, often called *rewards*.[2] RL is generally considered less efficient than IL in terms of number of interactions, since a reward typically provides much less information about the desired behavior than a demonstration. However, in many scenarios, rewards may take significantly less time and effort to obtain than demonstrations, so the conclusion about efficiency depends on how effort is defined.

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 optimality guarantees in convex optimization problems and empirically demonstrate robust performance in non-convex problems, proving their effectiveness.[3] Nevertheless, these two frameworks are limited in inclusiveness, as each can only handle a single type of feedback. Moreover, the types of feedback they support—demonstration and reward—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-based communication. First, human language is *referential*: people 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 rich narrative.

It is worthwhile to distinguish demonstrations and rewards from instructive and evaluative language feedback. When a teacher provides a demonstration, they are, by definition, communicating the solution through the agent's action space. The medium of this form of feedback therefore differs from that of language instructions (e.g., "go left, turn right"). Likewise, rewards are numerical signals, whereas language evaluations (e.g., "your solution is wrong") are conveyed verbally. Standard IL and RL cannot natively process instructive or evaluative language feedback.

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 (Liang et al., 2024; Jin et al., 2023; Zawalski et al., 2024). We refer to this framework as *context-based learning* (CBL).

How does CBL fare with respect to the three desiderata of a learning framework? In terms of inclusiveness, CBL is undeniably a significant step-up from IL and RL: rather than having to communicate through actions or numbers, human teachers can now speak freely to learning agents and employ many teaching strategies present in human-human communication. With regard to efficiency, CBL can be considered efficient within the range of performance it is able to attain, because the feedback provider can exploit the referential-inferential nature of human language. Unfortunately, the effectiveness of CBL remains questionable, as no satisfactory theory yet characterizes the generalizability of LLMs' language understanding in novel contexts. Empirically, Liang et al. (2024) and Jin et al. (2023) show that the improvement obtained from language feedback saturates after a few iterations, well before approaching optimality.

In short, the major shortcoming of IL or RL is efficiency and inclusiveness, whereas the major shortcoming of CBL is effectiveness. Although the lack of effectiveness of CBL can be compensated with additional IL and RL fine-tuning of the agent's language understanding, we contend that this approach can be unnecessarily costly, because the cognitive system of LLMs may be fundamentally incompatible with the intentions of natural language feedback. We elaborate on this point in the next section, but put simply: natural 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 natural language feedback cannot be accurately realized.

4. Human-compatible learning

We propose that incorporating natural language feedback in a principled way involves two steps: (1) understand the original intentions of the feedback producers, i.e., how they imagine the feedback to be incorporated (2) design cognitive

[2]Many techniques convert other types of feedback to rewards, allowing deploying RL to "learn" from those types of feedback (e.g., (Christiano et al., 2017)). However, the RL framework itself can admit only numerical feedback.

[3]RL may fall short in problems with a large search space.

systems that enable these intentions to be accomplished. In this section, we characterize the intentions of natural language feedback and the type of cognitive system that can incorporate these intentions.

4.1. An overview of human communication

We provide a holistic description of human (cooperative) communication by integrating three prominent research bodies. The first line of research is theory of mind (Premack & Woodruff, 1978). In a broad sense, "theory of mind" refers to the ability to mentally simulate an individual's cognitive system, particularly how it directs their outward behavior. While most animals regard conspecifics as reactive systems—a straightforward mapping from percepts to behaviors—humans are different in that they recognize that there are non-trivial mental processes in between. This recognition opens a whole new mechanism for communication. Michael Tomasello's seminal book *Constructing a Language* (PINE, 2005) portrays how human leverages theory of mind to communicate with and influence others. This excerpt articulates his distinction between human communication and the communication of other animals:

> "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."

The excerpt reveals perhaps the most important fact about human communication: while animal communication is concerned with regulating behavior, human communication is oriented toward **shaping the mind they attribute to others**. This cognitive approach dramatically boosts the efficiency of human communication, as intervening on a single cognitive process can shape multiple behaviors at once. For instance, telling someone, "The floor is slippery!" prompts them to exercise greater caution in performing *many* tasks within the room. Compared to an approach that regulates each individual task performance, i.e., telling them to "do z more carefully" for every relevant task z, the concise warning achieves a similar effect but with far less effort.

Whereas Tomasello's work characterizes the intentions of human communicative signals, pragmatics theory (Grice, 1990; Sperber & Wilson, 1986) sheds light on the mechanism by which these intentions are recognized. This theory proposes that human communication is an *inferential* (or reasoning-based) process. Human communicative signals are merely hints of the actual intentions speakers wish to convey. Such inference is possible because humans possess an exceptional ability to reason over their theory of mind. Specifically, during a conversation, the speaker produces a signal to convey an intention by postulating how the listener would receive it. On the other end, the listener recognizes the intention by hypothesizing how the speaker would convey it pragmatically. This mutually recursive reasoning process allows human ability to transfer information beyond the literal content of a message, elevating communication efficiency to a new level.

Combining the three bodies of work, we propose the following characterization of the language-based mind-shaping communication process in humans:

Algorithm 2 How Humans Communicate Linguistically to Shape the Mind

1: The speaker builds a mental model of the listener's cognitive system.
2: The speaker forms an intention to alter cognitive processes within that imaginary system, and reasons to generate an appropriate linguistic expression that signals that intention.
3: The listener infers the speaker's intention from the linguistic signal and adapts the targeted cognitive processes accordingly.

Human communication can sometimes aim to shape only outward behavior. For example, teaching via IL or RL can be seen as a behavior-shaping strategy: IL forces the agent to perform desired behaviors, whereas RL promotes or inhibits certain behaviors. Nevertheless, mind-shaping communication is regarded as the most advanced form of human communication, enabling the highest (token) efficiency and robustness in novel situations. When we talk about learning from language feedback, we refer to the capability of efficiently and precisely internalizing the mind-shaping intentions behind this type of feedback.

4.2. The wide range of language feedback intentions

Because there are many ways of theorizing a cognitive system and many ways of intervening on it, there exist not one but *many* forms of natural language feedback. The slippery-floor example illustrates feedback that targets beliefs, but language feedback can also target desires, intentions, emotions, and other mental states. Some feedback aims not at mental states themselves but at the mental processes that operate over them. Others are directed at multiple mental components simultaneously. Table 1 presents a non-exhaustive list with examples.

The substantial diversity of feedback intentions suggests that there is unlikely to be a simple, general solution to the problem of learning from natural language feedback.

Each intention category requires a different cognitive system to support it and a different algorithm for incorporating it. The final solution will likely consist of a collection of specialized frameworks, each of which handles a group of related intentions. Despite the heterogeneity, we argue that there exist common high-level design principles that can guide solutions for many forms of natural language feedback. In the next section, we present one such principle.

4.3. The necessity of human-compatible cognitive systems for incorporating natural language feedback

Drawing from the description of human communication in the previous section, we posit that rigorously incorporating natural language feedback requires an agent to possess a cognitive system that is compatible with the intentions behind such feedback. Without such an infrastructure, the influence assumed by human communicators simply cannot occur as originally planned.

Because humans naturally use language to communicate with other humans, a stronger requirement emerges: the agent's cognitive system should resemble the theory of mind that humans routinely attribute to one another. When speakers choose how to phrase an utterance, they habitually imagine how a *human* listener would interpret and integrate the latent intention into their cognitive system. For an AI system to accurately recognize and incorporate such intentions, its cognitive system must be sufficiently aligned with this imagination. We refer to such a system as a *human-compatible cognitive system*—one that is structured and operate in ways that humans can readily theorize about, reason over, and intentionally shape through language.

The previous "floor is slippery" example illustrates this requirement concretely. When a speaker issues this warning, they assume the listener (1) have a representation of the floor's slipperiness, (2) use that representation to guide actions across tasks, and (3) possess a mechanism for updating that representation in response to linguistic signals. If any of these components is missing—if the listener stores no such belief, if the belief has no causal effect on its behavior, or if language cannot update that belief—the intended mind-shaping effect fails, even when the listener's outward behavior appears satisfactory across many situations.

It is still possible to synthesize the effects of natural language feedback using a brute-force behavior-shaping approach, which does not necessitate a human-compatible cognitive system. In this approach, the agent does not adapt the targeted cognitive processes; instead, it merely learns to reproduce the expected outward behaviors in response to a large set of linguistic cues. Returning to the slippery-floor example, an agent may be trained—through supervised demonstrations—to behave more cautiously in a variety of tasks after receiving the warning. To a human observer, the agent may present itself as obeying the human's intention, even when it lacks any explicit representation of environmental beliefs. However, this approach can be inefficient or fragile. Fundamentally, this approach attempts to reconstruct a rich set of complex algorithms—a human theory of mind—from observed outputs alone, a class of problems where current deep learning systems are known to struggle (Zaremba & Sutskever, 2014; Lake & Baroni, 2018; Thomm et al., 2024; Shojaee et al., 2025). In practice, communicative intentions that aim to shape many downstream behaviors require the agent to be trained on a correspondingly large set of desired behaviors. If the agent encounters substantially novel contexts, the intended influence may fail to materialize. In the end, while behavior shaping can offer a superficial approximation of mind-shaping communication, it is unable to offer the robustness that a truly human-compatible cognitive system enables.

5. Cognitive architecture bounds learning ability

To illustrate how an agent's cognitive system shapes its ability to learn from human feedback, we present three case studies featuring agents with different cognitive architectures. We demonstrate that an agent's cognitive architecture fundamentally constrains the teaching strategies that can be effectively employed. In particular, more human-compatible architectures give rise to more diverse, efficient, and potentially more effective learning capabilities.

5.1. Case 1: Opaque agent (Figure 1)

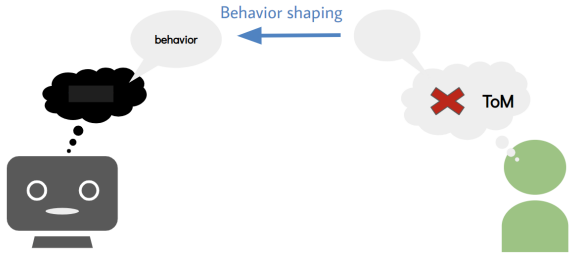

Figure 1. Opaque agents only allow humans to shape their outward behavior.

This agent implements an inscrutable policy $\pi(y \mid x)$. When teaching this agent, a human teacher likely views it as a reactive system—a direct mapping from percepts to actions—ignoring its internal cognitive processes. With this theory of mind, the only viable strategy is to directly regulate the outward behavior. Two approaches are available, which correspond to two currently popular learning frameworks: (1) force the agent to reproduce a desired behavior (imitation learning) and (2) encourage or discourage a behavior

Table 1. A non-exhaustive list of types of language feedback, the mental components they target, and the cognitive-system support required to incorporate them.

Feedback type	Targeted Mental Component	Example Utterance	Required Cognitive-System Support
Belief-shaping	Beliefs about the environment or task structure	"The floor is slippery." "The cup is on the top shelf."	Agent must represent environmental beliefs and update them in response to language.
Desire-shaping	Preferences, goals, or reward-related values	"Try to get the freshest apples." "You should prioritize safety over speed."	Agent must represent and modify internal reward-related preferences.
Intention-shaping	Plans, subgoals, or action-level intentions	"Stir the mixture before heating it." "Approach the door from the left."	Agent must represent intentions or plans and revise them through linguistic instructions.
Mixed	Beliefs + desires, or beliefs + intentions, etc.	"Since the vase is fragile, handle it gently." "Because it's late, aim for a quicker route."	Agent must integrate multiple mental representations and update them jointly.
Process-level	Mental processes such as reasoning style, planning strategy, or decision rules	"Think step by step." "Double-check your assumptions before acting."	Agent must expose and adapt its cognitive processes (e.g., reasoning pipeline, planning algorithm).
Meta-level	How the agent should use or update its own cognitive system	"Don't trust the map too much; rely more on what you observe." "When you're uncertain, ask for clarification."	Agent must model and revise meta-cognitive processes or self-regulation mechanisms.

produced by the agent (reinforcement learning).

5.2. Case 2: Chain-of-thought agent (Figure 2)

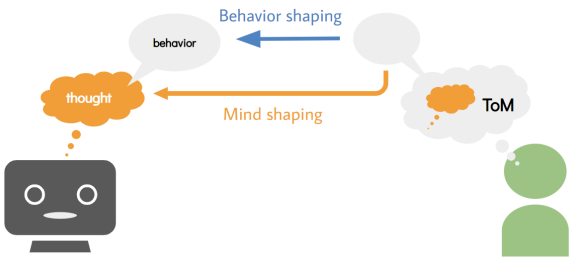

Figure 2. Chain-of-thought agents allow humans to shape both their outward behavior and internal reasoning.

A chain-of-thought (CoT) 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 (chain of) thought 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. This enables the human to predict the agent's thought even though it is latent. Being able to make these predictions allows the human to plan communication signals that alter the agent's thought, thereby enabling them to influence the agent's behavior indirectly, in addition to directly regulating its behavior.

Concretely, with a CoT 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 y only. 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 (Williams, 1992) are needed to enable backpropagation through these discrete-token parameters.

Context-based learning (CBL). Instead of letting the agent generate z on its own, the human can provide the value

of this variable for it. This strategy requires the agent's π_1 to be pre-trained to generate good behaviors y conditioned on human-provided z. The performance of this approach on previously unseen z relies on the generalizability of the pre-trained policy.

Interactive learning from activity description (ILIAD). Developed by (Nguyen et al., 2021), this approach aims to ensure consistency between z and y. If z is a flawed thought, we want it to lead to an undesirable rather a good behavior y (e.g., $z =$ "make bad coffee" $\rightarrow y =$ actual bad coffee). In general, z must always be an accurate description of y. If this consistency holds for all x, we only need to train the agent to produce desirable thoughts (i.e., learning $\pi_0(z \mid x)$). Due to the consistency guarantee, the output behavior will also be desirable. The learning of π_0 can be accomplished via IL, RL, or CBL.

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 have been z' instead of z if what you did was 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.

The feedback in IL on z, CBL, or ILIAD is essentially language feedback, because it is taken from the space of z, which, by definition of a CoT agent, is expressed in language. Moreover, the feedback carries a mind-shaping intention, as it aims to modify the agent's original thought. Hence, these three frameworks exemplify learning frameworks that capture the nature of human communication, distinguishing itself from traditional approaches like IL/ RL on the output y only. We can see that these frameworks are applicable thanks to the chain-of-thought cognitive architecture. This underscores the necessity of a human-compatible cognitive system to effectively support natural language feedback.

5.3. Case 3: Agent with a language-guided world model (Figure 3)

Following Zhang et al. (2024), we consider an agent that, instead of reasoning in language, runs an actual optimization process to plan its actions. Specifically, in addition to a policy π, 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 (recall that our formulation assumes deterministic environments). In particular, the behavior of the model is controlled by *textual context* $z \in \mathcal{Z}$. We will use $\tilde{T}_z$ as a shorthand for $\tilde{T}(\cdot, \cdot; z)$.

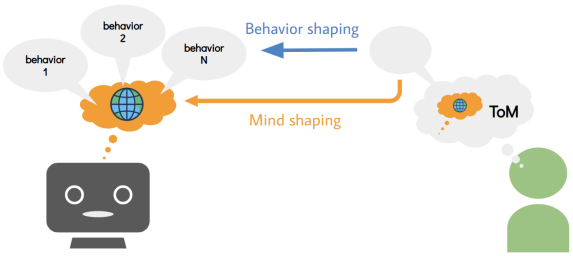

Figure 3. Agents that employ a world model for planning allow humans to efficiently influence multiple behaviors at once by altering the agent's world model.

We assume that the agent has access to the task reward function $R(x, y)$.[4] Given a task x, the agent generates a solution y as follows. First, it computes the optimal policy with respect to its world model: $\pi_x^\star = \max_{\pi_x} \mathbb{E}_{y \sim P_{\pi_x, \tilde{T}_z}}[R(x, y)]$, where $P_{\pi_x, \tilde{T}_z}$ is the distribution over solutions induced by policy π_x and transition function $\tilde{T}_z$. Next, it generates a solution using this policy. With this process, the agent's world model effectively influences its behavior.

This cognitive process gives rise to a novel strategy for controlling the agent's behavior: adapting its world model by modifying the textual context z. Note that presenting a new z to the agent is essentially giving it language feedback. This strategy mimics how humans influence others' beliefs about the world in conversations. As seen from the slippery-floor example (§4), teaching by altering the world model is remarkably efficient. In our particular formulation, the same world model is used to compute the policy π_x for *every* task x. Consequently, improving the accuracy of the world model enhances the solution quality for all tasks at once.

We can further extend this type of agent by allowing it to autonomously generate the world-model context z using a learned distribution $\mu(z)$. This new variant behaves similarly to a chain-of-thought agent: it generates its own language parameters z and acts in accordance with them. Consequently, in addition to modifying the world model, we can employ all the strategies applicable to a CoT agent to teach this agent.

6. What is next?

Our thesis implies that solving the problem of learning from human language feedback is not solely a matter of designing better learning algorithms, but also of designing the right cognitive architectures to support them. In this endeavor, the challenge is not only to imitate the human's theory of

[4] 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).

mind, but also to enable rigorous proofs of effectiveness, efficiency, and inclusiveness. In this section, we highlight several concrete challenges and research directions that can potentially move us closer to this goal.

Build human-like cognitive systems with language-guided, reusable modules. Because human reasoning has been shaped by communication within the same species, the most human-compatible cognitive system for AI agents will closely mirror the human cognitive system itself. Although the human cognitive system is not yet fully understood, recent proposals such as (LeCun, 2022) and (Sumers et al., 2023), grounded in prominent cognitive science theories, offer promising starting points for further development.

Two additional principles are particularly important for this endeavor. First, incorporating modules that are *shared across multiple cognitive processes* can substantially increase learning efficiency, as illustrated by the use of a shared world model in the previous section. Second, modeling each module as a *language-conditioned function* enables it to be adapted directly through language feedback, thereby enhancing inclusiveness. In building language-modulated systems, in addition to language-conditioned world models (Lin et al., 2023; Nematollahi et al., 2025; Bruce et al., 2024; Du et al., 2023; Zhou et al., 2024), prior work on language-conditioned reward functions (Fu et al., 2019) offers valuable insights and techniques that future work can leverage.

Connect with the MDP framework to facilitate theoretical guarantees. A well-formulated learning framework should support rigorous theoretical analysis, enabling formal proofs of effectiveness and efficiency. The classical MDP framework and its derivatives provide a natural foundation for two reasons. First, they have been extensively analyzed, yielding a rich set of theoretical tools. Second, their components align closely with the major human mental states studied in cognitive science: for example, the transition function corresponds to beliefs about the environment, the reward function reflects desires, and the policy represents intentions.

However, most MDP-based formulations treat the agent as a black box, without specifying its internal cognitive system. The I-POMDP framework (Gmytrasiewicz & Doshi, 2005) and related belief-inference models depart from this trend by assigning explicit reasoning processes to agents. Yet a full cognitive system—comprising multiple foundational mental states analogous to those of humans—is still missing. A central challenge for future work is to derive insightful theoretical models for agents equipped with such human-like cognitive systems. Taxonomies like ATOMS (Beaudoin et al., 2020; Xie et al., 2023; Goyal et al., 2019; Kwon et al., 2023) provide useful guidance regarding which mental states to include.

Explore higher-order influence. Directly shaping an agent's behavior can be viewed as a zero-order teaching intention, whereas altering its mental states to induce behavioral change is fundamentally first-order. Higher-order teaching intentions involve influencing at least two intermediate mental states. Consider a process in which modifying the policy for a particular task renders it suboptimal with respect to the world model, prompting the world model to adjust itself to remain consistent with the updated policy. This, in turn, affects the policies of other tasks, which adapt to stay optimal under the revised world model. In this way, feedback directed at a single policy can efficiently adapt many policies at once. Realizing this approach requires optimization methods capable of propagating learning feedback through complex cognitive processes involving multiple intermediate mental representations. Variational approaches like (Eysenbach et al., 2022) offer fundamental principles that could enable such propagation.

Adopt a dual-system architecture to enhance speed and robustness. The second and third case studies in § 5 illustrate two complementary types of cognitive systems: one that reasons through language and another that reasons through mathematical optimization. The language-based system is fast and intuitive but may be fragile due to its reliance on pattern-matching, while the optimization-based system is more rigorous but computationally demanding. Humans combine two analogous systems to enjoy the strengths of both, a design known as the *fast-and-slow* architecture (Kahneman, 2011). We argue that future AI agents could similarly benefit from such a dual architecture. The central challenge is to coordinate the two systems effectively: determining when each should make decisions and how they should share and transfer knowledge.

Tackle the problem of inferring human intentions. Thus far, we have focused exclusively on incorporating human teaching intentions. In practice, however, these intentions are rarely explicit; they must be inferred from linguistic utterances that are often indirect, ambiguous, or highly context-dependent. Equipping agents with strong reasoning capabilities is therefore essential for robust communication. The key challenges of this problem have been examined extensively in position and survey papers (Bisk et al., 2020; Fried et al., 2022; Hoffman et al., 2024; Fisac et al., 2019). Addressing them will require integrating advances in language grounding, theory-of-mind modeling, and continual learning into a unified framework capable of reliably inferring and acting upon human intentions across diverse real-world settings.

Impact Statement

Learning from human feedback is increasingly central to the deployment of AI systems in real-world settings, particularly as these systems are expected to operate autonomously, interact naturally with people, and adapt over time. Natural language feedback is an especially powerful form of human input, as it allows users to efficiently convey intentions, corrections, and high-level guidance. However, current learning frameworks often treat language feedback as an unstructured signal to be optimized against, without accounting for the cognitive assumptions implicit in how humans communicate. This gap limits both the effectiveness and reliability of human–AI interaction.

This paper argues that to fully and safely harness natural language feedback, AI agents must be equipped with *human-compatible cognitive systems*—internal representations and processes that humans can naturally reason about and intentionally influence through language. By grounding this claim in insights from cognitive science and formal learning frameworks, the paper reframes language-based alignment not merely as an algorithmic challenge, but as a systems-level design problem. This perspective helps explain why existing approaches to learning from language feedback often plateau or behave unpredictably, and it offers principled guidance for building more robust and interpretable agents.

The ideas presented here have several positive potential impacts. First, they encourage the development of AI systems that are easier for humans to teach, correct, and collaborate with, reducing the cognitive and operational burden on users. Second, by emphasizing cognitive transparency and structured internal representations, the framework supports more reliable generalization to novel situations, which is critical for safety-critical and high-stakes applications. Third, the paper provides a conceptual foundation for unifying disparate strands of research—including reinforcement learning, imitation learning, language-based supervision, and cognitive architectures—under a common theoretical lens.

At the same time, increased human-compatibility also raises important considerations. Systems that are designed to be highly responsive to language feedback must be carefully evaluated to ensure that they do not over-interpret ambiguous instructions, amplify human errors, or become susceptible to unintended manipulation. The framework presented in this paper makes these risks more explicit by tying them to specific cognitive assumptions, thereby enabling more targeted evaluation and mitigation strategies.

Overall, this work aims to advance the scientific foundations of human-AI alignment by clarifying what it means for an agent to meaningfully incorporate natural language feedback. By shifting attention from surface-level learning signals to the underlying cognitive systems that support communication, the paper seeks to enable AI systems that are not only more capable, but also more understandable, controllable, and beneficial to the humans they serve.

References

Akyürek, A. F., Akyürek, E., Kalyan, A., Clark, P., Wijaya, D. T., and Tandon, N. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7716–7733, 2023.

Beaudoin, C., Leblanc, É., Gagner, C., and Beauchamp, M. H. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10: 2905, 2020.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

Bruce, J., Dennis, M. D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Chen, A., Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Bowman, S. R., Cho, K., and Perez, E. Learning from natural language feedback. *Transactions on machine learning research*, 2024.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., and Abbeel, P. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Eysenbach, B., Khazatsky, A., Levine, S., and Salakhutdinov, R. R. Mismatched no more: Joint model-policy optimization for model-based rl. *Advances in Neural Information Processing Systems*, 35:23230–23243, 2022.

Feng, X., Liu, B., Song, Y., Fu, H., Wan, Z., Koushik, G. A., Hu, Z., Yang, M., Wen, Y., and Wang, J. Natural language reinforcement learning. *arXiv preprint arXiv:2411.14251*, 2024.

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S. S., Griffiths, T. L., and Dragan, A. D. Pragmatic-pedagogic value alignment. In *Robotics research: the 18th international symposium Isrr*, pp. 49–57. Springer, 2019.

Fried, D., Tomlin, N., Hu, J., Patel, R., and Nematzadeh, A. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*, 2022.

Fu, J., Korattikara, A., Levine, S., and Guadarrama, S. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019.

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

Goyal, P., Niekum, S., and Mooney, R. J. Using natural language for reward shaping in reinforcement learning. *arXiv preprint arXiv:1903.02020*, 2019.

Grice, H. P. 1975 logic and conversation. *The Philosophy of Language*, 1990.

Hoffman, G., Bhattacharjee, T., and Nikolaidis, S. Inferring human intent and predicting human action in human–robot collaboration. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2024.

Jin, D., Mehri, S., Hazarika, D., Padmakumar, A., Lee, S., Liu, Y., and Namazifar, M. Data-efficient alignment of large language models with human feedback through natural language. *arXiv preprint arXiv:2311.14543*, 2023.

Kahneman, D. *Thinking, fast and slow*. macmillan, 2011.

Kwon, M., Xie, S. M., Bullard, K., and Sadigh, D. Reward design with language models. *arXiv preprint arXiv:2303.00001*, 2023.

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pp. 2873–2882. PMLR, 2018.

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

Liang, J., Xia, F., Yu, W., Zeng, A., Arenas, M. G., Attarian, M., Bauza, M., Bennice, M., Bewley, A., Dostmohamed, A., et al. Learning to learn faster from human feedback with language model predictive control. *arXiv preprint arXiv:2402.11450*, 2024.

Lin, J., Du, Y., Watkins, O., Hafner, D., Abbeel, P., Klein, D., and Dragan, A. Learning to model the world with language. *arXiv preprint arXiv:2308.01399*, 2023.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Nematollahi, I., DeMoss, B., Chandra, A. L., Hawes, N., Burgard, W., and Posner, I. Lumos: Language-conditioned imitation learning with world models. *arXiv preprint arXiv:2503.10370*, 2025.

Nguyen, K. X., Misra, D., Schapire, R., Dudík, M., and Shafto, P. Interactive learning from activity description. In *International Conference on Machine Learning*, pp. 8096–8108. PMLR, 2021.

PINE, J. M. 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.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978.

Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Chen, A., Cho, K., and Perez, E. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

Sperber, D. and Wilson, D. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

Sumers, T., Yao, S., Narasimhan, K., and Griffiths, T. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Thomm, J., Camposampiero, G., Terzic, A., Hersche, M., Schölkopf, B., and Rahimi, A. Limits of transformer language models on learning to compose algorithms. *Advances in Neural Information Processing Systems*, 37: 7631–7674, 2024.

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

Xie, T., Zhao, S., Wu, C. H., Liu, Y., Luo, Q., Zhong, V., Yang, Y., and Yu, T. Text2reward: Reward shaping with language models for reinforcement learning. *arXiv preprint arXiv:2309.11489*, 2023.

Zaremba, W. and Sutskever, I. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.

Zawalski, M., Chen, W., Pertsch, K., Mees, O., Finn, C., and Levine, S. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Zhang, A., Nguyen, K., Tuyls, J., Lin, A., and Narasimhan, K. Language-guided world models: A model-based approach to ai control. *arXiv preprint arXiv:2402.01695*, 2024.

Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.-Y., and Gan, C. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Table 2. Comparison of the three most popular learning frameworks along the three desiderata outlined in §2.

Framework	Effectiveness	Efficiency	Inclusiveness
Imitation Learning (IL)	Optimality convergence guarantee derives from gradient-based optimization theory; empirically fast to converge to near-optimality	Highly efficient in terms of number of interactions as demonstrations are information rich; however, can be viewed as inefficient since demonstrations may require substantial effort to obtain (e.g., self-driving car domain)	Low: restricted to instructive or corrective demonstrations
Reinforcement Learning (RL)	Optimality convergence guarantee derives from gradient-based optimization theory; in practice, often not able to converge to near-optimality when search space is large (e.g., problems with language output)	Due to low-information rewards, often inefficient in terms of number of interactions in problems with large search space; but rewards can be generally cheaper to provide than demonstrations in certain applications	Low: limited to numerical feedback
Context-Based Learning (CBL)	Convergence relies on model generalizability, which is theoretically poorly understood in oudt-of-distribution cases; empirically, improvement saturates after a few rounds of revisions	Highly efficient in terms of spoken information bits (but only within reachable performance range, which can be far from optimality)	High: supports free-form natural language; allows wide range of human teaching strategies

A. Appendix