# Revisiting Human-Robot Teaching and Learning Through the Lens of Human Concept Learning

Serena Booth, *MIT CSAIL*, sbooth@mit.edu
Sanjana Sharma, *Harvard SEAS*, sharma.sas@gmail.com
Sarah Chung, *Harvard SEAS*, sc232@gmail.com
Julie Shah, *MIT CSAIL*, julie_a_shah@csail.mit.edu
Elena L. Glassman, *Harvard SEAS*, glassman@seas.harvard.edu

***Abstract*—When interacting with a robot, humans form conceptual models (of varying quality) which capture how the robot behaves. These conceptual models form just from watching or interacting with the robot, with or without conscious thought. Some methods select and present robot behaviors to improve human conceptual model formation; nonetheless, these methods and HRI more broadly have not yet consulted cognitive theories of human concept learning. These validated theories offer concrete design guidance to support humans in developing conceptual models more quickly, accurately, and flexibly. Specifically, Analogical Transfer Theory and the Variation Theory of Learning have been successfully deployed in other fields, and offer new insights for the HRI community about the selection and presentation of robot behaviors. Using these theories, we review and contextualize 35 prior works in human-robot teaching and learning, and we assess how these works incorporate or omit the design implications of these theories. From this review, we identify new opportunities for algorithms and interfaces to help humans more easily learn conceptual models of robot behaviors, which in turn can help humans become more effective robot teachers and collaborators.**

***Index Terms*—human-concept learning, mental models, HRI**

## I. Introduction

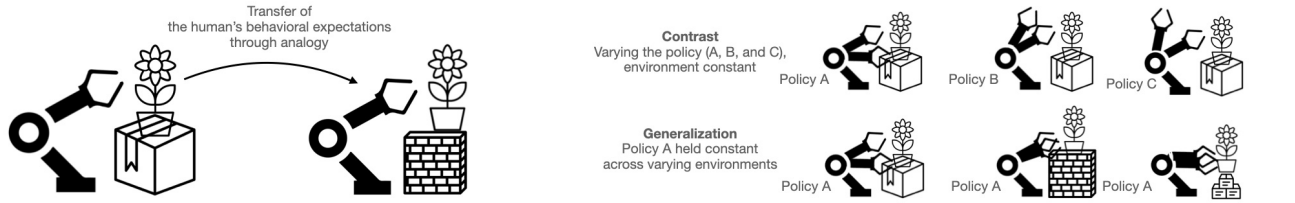
Humans should be able to teach robots new skills, norms, or preferences [1], [2], but challenges abound. Before teaching, the human can benefit from learning about the robot's current behaviors. Appropriately selecting robot behaviors to show to the human is challenging: observing a robot perform well or poorly biases the human's understanding of its competency [3]. Another challenge arises if the human cannot draw on preexisting mental models for robot behaviors. For example, the human may struggle to learn to predict robot motions if those motions are not human- or animal-like [4]. To assess the impact of their teaching, the human has to compare the robot's current behaviors to its past behaviors—a comparison which is not always straightforward. Bıyık et al. [5] found that humans are unable to provide preferences when robot behavior changes are imperceptible or of roughly equal utility, while Amitai and Amir found that independently selected behaviors are hard to compare [6]. In short, humans find it non-trivial to learn useful conceptual models of robot capabilities and limitations.

We review 35 papers from the human-robot teaching and learning literature, and we contextualize these works by analyzing whether and how they incorporate principles from cognitive theories of human concept learning. Applying these theories supports humans in developing conceptual models of robot capabilities and limitations faster, more accurately, and more flexibly, which in turn can help resolve the aforementioned interaction challenges. Specifically, we look to Analogical Transfer Theory [7], [8], which informs how humans use analogy to transfer prior knowledge to unfamiliar domains, and the Variation Theory of Learning [9], [10], which informs how humans learn to separate superficial details from core knowledge. Together, these complementary theories explain how humans come to understand complex, high-dimensional phenomena and make predictions about unrevealed facts and futures—as such, these theories can be applied to help humans understand robot behaviors. While the HRI community has not previously consulted these theories, each of the works we review inadvertently uses some of their guiding principles.

Going forward, humans need interfaces and algorithms that mediate human-robot teaching and learning by *systematically* guiding the human's learning about the robot's behaviors and how they change in response to human input. These interfaces can help humans to (1) learn about the robot's capabilities and limitations, (2) teach the robot by providing a response (e.g., feedback), and (3) learn about the capabilities and limitations of updated robot behavior candidates, and compare these to prior behaviors. Human concept learning theories provide design guidance, i.e., about the selection, sequence, and presentation of robot behaviors, for these interfaces and algorithms. Notably, Variation Theory prescribes an ordered sequence of variance and invariance to help humans distinguish core behaviors from superficial or incidental details, and Analogical Transfer Theory prescribes knowledge transfer by supporting the human's recall with a familiar entity or context.

## II. Human Concept Learning Theories

Cognitive theories of human concept learning have been refined by testing curriculum interventions. We look to two complementary theories, Analogical Transfer Theory and the Variation Theory of Learning, to inform how interfaces can best mediate the practice of humans learning about robot behaviors. Analogical Transfer Theory explains how humans transfer knowledge to new situations and domains, while the Variation Theory of Learning explains how particular patterns of variation and invariance can help humans discern the difference between superficial details and critical *features* and *aspects*. These processes are key to helping humans understand robots. Fig. 1 summarizes these learning theories graphically.




**Analogical Transfer Theory**
Humans can infer robot behavior by transferring observed robot behaviors from base situations to new target contexts

**Variation Theory**
Using patterns of contrast and invariance to help humans grasp concepts, e.g., robot reaching behaviors under different policies with a focus on Policy A

Fig. 1: Human concept learning prescribes curricula to help the human form useful conceptual models of robot behaviors.

*1) Analogical Transfer Theory:* Studies in cognitive psychology have shown that the parallel presentation of examples helps students attain knowledge gains. For example, simultaneous rather than sequential comparison has been shown to help mathematics students achieve greater gains in both procedural and conceptual knowledge [11]. When analogical encoding, or learning by drawing comparisons across examples, was incorporated in a negotiation strategy curriculum, Gentner et al. [12] found that comparing two parallel cases rather than studying the cases separately improved schema abstraction and transfer among novices, and that asking learners to describe the commonalities between these cases had the biggest positive impact. These controlled studies are examples of the body of work that collectively informs Analogical Transfer Theory.

Analogical Transfer Theory asserts that *analogy*, or finding and using relational commonalities, is a building block of human concept learning [7], [8], [13]. In analogy, a familiar base domain informs inferences about an unfamiliar target domain. First, a person identifies a candidate base. The person then maps the analogy by *structurally aligning* the base and target; this alignment should highlight relational similarities. Lastly, the person must evaluate the analogy by assessing any inferences drawn from it. Structural alignment and analogy allow people to form new inferences about novel targets (**inference projection**), construct new schemas or mental models by mapping relations (**schema abstraction**), detect differences between bases and targets (**difference detection**), and re-represent bases and targets at alternate levels of abstraction, making the analogy more applicable (**re-representation**).

Analogical Transfer Theory can inform HRI interface design. When faced with a novel domain, people implicitly seek a comparison base domain from memory and search for commonalities between the target and base. When forming an analogy, the person's understanding of the target is bolstered by these commonalities. There are two notable opportunities for interfaces to assist in analogy formation. First, humans are bad at recalling analogous base cases from memory [14], so an interface has an opportunity to prompt the human to recall a relationally-similar base. Second, analogies rely on structural alignment, which highlights the relational commonalities between the target and the base: an interface can present data in an aligned manner such that humans are more readily able to draw inferences about the target or to detect differences.

*2) Variation Theory of Learning:* Controlled studies in cognitive psychology have shown that presenting strategically varied examples improves learning outcomes. Students studying high-variability geometry problems required less mental effort than those studying low-variability examples, and their transfer performance was better and less effortful [15]. When tasked with solving statistical word problems and given either one or three examples with varying or constant superficial details, students with *multiple parallel examples that emphasized structural commonalities by varying superficial details* did best [16]. This variation positively impacted students' schema construction; interestingly, the impact was greatest for those with the least prior mathematical knowledge. Strategic variation illuminates otherwise difficult-to-discern latent structure of concepts [17]. These studies support the potential for variation to help end-users understand feasible robot behaviors.

Variation Theory argues that a person must first discern *critical aspects* and *features* to comprehend some object of learning. Aspects are parameters (e.g., color) while features in this context are instantiations of aspects (e.g., the color red). Aspects are critical when they are strictly necessary to understand the concept. To achieve robust discernment and learning, the person must experience variation across critical and non-critical (or superficial) aspects. To apply Variation Theory, we designate some aspect(s) as the focused object of conceptual learning. Having identified focused aspect(s), variation learning follows an ordered sequence of structured patterns of variance and invariance. These patterns support inductive reasoning to help humans more accurately infer how focused aspects contribute to the object of learning, e.g., a particular robot behavioral policy. For each focused aspect, Variation Theory prescribes the following sequence:

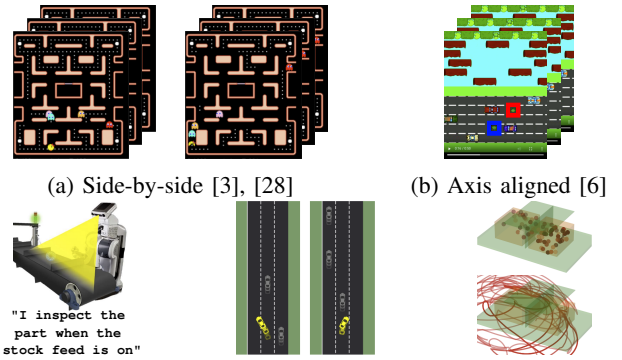
1) **Repetition.** All aspects are held constant. E.g., to learn about a robot's behaviors, the human sees the robot repeatedly act in the same environment.
2) **Contrast.** The focused aspect varies while other aspects are held constant. E.g., Fig. 1, the robot uses varied policies while operating in an otherwise fixed environment.
3) **Generalization.** The focused aspect is held constant, while other aspects vary. E.g., Fig. 1, the human sees how the robot's policy varies in new environments using the selected value of the focused aspect.
4) **Fusion.** All aspects vary to mimic "real world" variation.

Variation Theory has been used effectively in many domains, including story comprehension [17], learning vocabulary words [18], Chinese characters [19], the color of light [20], mathematics education [21], chemistry education [22], and computing education [23]. Books have discussed how Variation Theory can improve teaching and learning in schools [24], [25]. In HRI, Variation Theory has immediate application: many interfaces solicit human feedback as reflections on single executions of robot behaviors. But this fails to accommodate the backbone of Variation Theory: to provide high quality feedback, the person needs to understand the robot's behavior—which means they need to understand the underlying critical aspects of the robot's behaviors by first experiencing variation of the underlying critical features—*before* providing preferences or feedback over these behaviors.

*3) Concept Learning for Hypothetical Robot Applications:* In HRI, human concept learning occurs whenever the human must learn about robot behaviors. We consider two hypothetical robotics applications for exposition.

Consider collaborative assembly. Traditionally, robots halt whenever a human enters a shared work region. Modern approaches let robots predict human behavior and optimize their plans to increase system uptime [26]. For successful collaboration, the human should also learn about the robot's behaviors, both to increase their comfort in proximity and to help optimize robot uptime. For this, the human benefits from understanding both the robot's workspace and motion planner. A naive approach might show a human the limits of the robot's workspace or an example motion. In practice, however, the effective working patterns of the robot seldom reach these limits, and the human's learned conceptual model would be too conservative. By instead applying Variation Theory and experiencing variation in the robot's positioning and motions, the human can learn a conceptual model of the robot's effective workspace. With this, the human can feel safer (by predicting how the robot will move), and work to increase the robot's uptime (by avoiding interfering with the robot's plan).

In other robotics applications, the robot may have frequent, fleeting interactions with non-expert users—for example, a delivery robot needs to navigate alongside pedestrians and others. In such applications, particularly when humans have only brief encounters with robots, Analogical Transfer Theory can assist. One challenge with delivery robots is their potential use of omni-directional wheels: while these wheels provide flexibility to a large range of motions, humans experience discomfort when interacting with such a robot, since the motions these wheel structures exhibit are hard to predict [27]. One way to apply Analogical Transfer Theory is for the robot designer to leverage physical analogies: if the robot closely resembles a car (as is common in such applications), humans who interact with the system would anticipate car-like motions, which are dissimilar from many omni-directional wheel motions. Styling such robots after humans encompasses a larger range of acceptable omnidirectional motions, but challenges remain in aligning these motions to human expectations [27].




(a) Side-by-side [3], [28] (b) Axis aligned [6]

(c) Predicates [29] (d) Critical states [30], [31] (e) Overlaid [32]

Fig. 2: Policy summarization. (2a) uses contrast through side-by-side video summaries of varied policies [3], [28]. (2b) better aligns differences with videos of two varied policies—the red and blue agents—in an axis-aligned, shared state [6]. (2c) presents logical statements with states grouped by Boolean predicates [29]. (2d) presents varied critical states [30], [31]. (2e) overlays visuals to structurally-align both varied environment configurations (above) and trajectories (below) [32].

## III. Concept Learning in HRI

We review 35 papers from the literature on human-robot teaching and learning. We study how these works benefit from human concept learning principles, and how better curriculum design could support more effective interaction. Though these roles are inherently fluid, we focus on works in which the human is primarily the "teacher" and the robot the "learner"; i.e., we exclude robot tutoring. We selected works which have the following goals: policy summarization, updating human beliefs, or teaching with feedback, preferences, and/or corrections. The first two goals implicitly build curricula for informing conceptual models of robot behaviors; the latter three help humans teach robots with seemingly intuitive signals. We selected papers primarily from premier venues (e.g., HRI, NeurIPS, AAAI), which are highly topical or influential in these niches (e.g., #citations$\geq$50). See the supplementary material for further analysis and the appendix for a definition of "objects of learning" in these settings.

### A. Policy Summarization

Policy summarizations aim to help a human understand the robot's expected behaviors [33], allowing the human to determine an apt level of autonomy to afford the robot. Fig. 2 shows example policy summary interfaces.

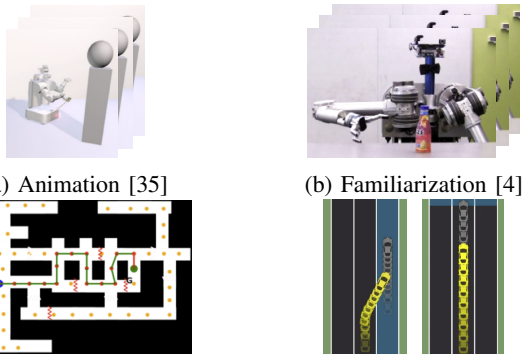
*1) Implementations:* Several policy summarization methods [3], [6], [28], [30] use $\mathcal{Q}$-values to select informative states; these quantify the benefit of taking action $a$ in state $s$. One approach selects states with the largest delta in $\mathcal{Q}$-values across actions [28], [30]. Huang et al. [30] call these *critical states* as they support learning about the robot's capabilities and limitations. In concept learning terms, critical states are *critical features*: a person cannot learn about the policy without understanding the robot's behaviors in these states.

Another method requires shared Boolean predicates between human and robot [29]. This method takes predicate-annotated trajectories, and solves a set cover problem over these traces to answer questions like, "When does the robot do $a$?" with answers like, "The robot does $a$ when p or q." Another approach presents counterfactuals by finding similar states which elicit different actions [31], and a final method applies Bayesian inference to find environments where the policy expresses a specific property, like maximal directness [32].

*2) Analogical Transfer Theory:* Policy summarization approaches make extensive use of structural alignment, the backbone of Analogical Transfer Theory. These works all facilitate schema abstraction, wherein the human assesses how the policy would apply in new target environments. Several approaches use structural alignment by visualizing shared states with different policies side-by-side [3], [28], [30] or overlaid [6]. Hayes and Shah [29] use shared predicates, and present textual summaries of these groupings to a user for both schema abstraction and inference projection, wherein a user learns how the agent will behave in a new state with shared properties. Countering this, Zhou et al. [32] instead structurally align environments and trajectories (but not individual states) by overlaying semitransparent visualizations to support schema abstraction for policy failure modes.

Several approaches also support difference detection, though to differing levels of success. Some compare multiple candidate policies [3], [6], [28], [30]; the goal is to assist a user in selecting the best policy. Most such methods find interesting states for two or three policies independently, and present these states or behavior samples for each policy side-by-side [3], [28], [30]. However, this independent search does not maximally find differences between these candidate policies. As such, these policy candidates are not well structurally aligned, and these side-by-side comparisons may be difficult or impossible for the user to assess. Amitai and Amir [6] propose a counter approach: instead of generating interesting states for each policy independently, they simulate both policies in parallel, and find policy disagreements, where the respective policies choose different actions in a shared state. They then show these two different policies in an axis-aligned manner, such that the user is more readily able to detect differences and select the more appropriate policy for their intended context.

*3) Variation Theory:* Using Variation Theory's contrast, Hayes and Shah [29] group states (through Boolean predicates) while maintaining a fixed policy and a fixed action. Zhou et al. [32] simultaneously present varied environments and trajectories, again for a fixed policy. Huang et al. [30] and Olson et al. [31] similarly present varied states as focused aspects, while maintaining a fixed policy. For fusion, Huang et al. [30] additionally compare policies; in this, every aspect is varied. Similarly, several other approaches [3], [6], [28] present a varied set of states alongside two [6], [28] or three [3] candidate policies. While none of these works use the language of Variation Theory, these incorporated patterns of variance and invariance help these works achieve their goals of supporting the human in learning about the robot's policy.



(a) Animation [35] (b) Familiarization [4]

(c) Annotations [36] (d) Examples [37], [38]

Fig. 3: Interfaces for updating human beliefs methods. 3a: [35] uses re-representation by using animation principles to induce legibility through anthropomorphized familiarity. 3b: [4] incorporates generalization by showing a curriculum of varied motions. 3c: [39] uses difference detection, wherein robots provide annotations to correct a human's model—here, as a map. 3d: [37] and [38] use generalization: they show varied trajectories while holding the policy constant.

*4) Takeaways and Frontiers:* Sequeira and Gervasio [3] found that users often anchor their perceptions of a robot's capabilities on their initial experiences [34]; as such, they found the importance of "appropriate" variation, and they employed Variation Theory's fusion as a learning strategy. Nonetheless, they do not define an "appropriate" amount of variation. Variation Theory can help: it can guide the design of strategic and sufficient variation to support the human's innate learning abilities. A more structured exposure with Variation Theory's prescribed sequence of contrast→generalization→fusion could improve the human's learning. *Fusion isn't all we need.* Learning to discern the limits of robot behaviors, through contrast and generalization, is needed, too.

### *B. Prompting Human Belief Updates*

Many methods explore how to prompt humans to update their beliefs. These include generating expressive motions [4], [35], [40] or state/action pairs [30], [38], or constructing "patches" to reconcile divergent models [36]. See Fig. 3.

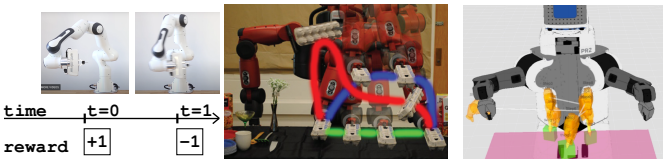
*1) Implementations:* Takayama et al. [35] observed that robots require time for planning, but the transition between "thinking" and acting often catches humans off guard. To address this, they used animation principles of anticipation and reaction for more expressive motion. Kwon et al. [40] observed that robot failures are not expressive—typically, the robot just stops. They generate expressive trajectories which mimic successful trajectories *in spite of failure*. Lastly, Dragan and Srinivasa [4] studied whether humans could learn a robot's motions through familiarization. They optimized motions with two cost functions: one which enables human-like motion; another which enables unnatural motion by prioritizing shoulder motion over wrist motion. They discovered users were more adept at predicting natural motions in new settings.

Huang et al. [37] and Lage et al. [38] seek expressive states which allow a person to update their beliefs about the robot's objective. Huang et al. [37] assumed the human uses inverse reinforcement learning (IRL) to model the robot's objective, and then used Bayesian inference to find environments which are maximally informative to the human's beliefs. Lage et al. [38] compared IRL with imitation learning. Both assess the human's learning by testing their knowledge in unfamiliar contexts. Finally, Chakraborti et al. [36] considered a search-and-rescue task, where the human has an outdated mental model of the environment while the robot acquires knowledge of how a disaster changed the environment. To update the human's beliefs, the robot explains model differences by providing a patch expressing why its new plan is acceptable.

*2) Analogical Transfer Theory:* Takayama et al.'s work [35] is unusual and interesting in its use of re-representation from Analogical Transfer Theory. In re-representation, a base and/or target is re-represented at a higher level of abstraction so a user is more readily able to perform analogical reasoning. They use animation to re-represent the robot as a more familiar entity to communicate robots switching from planning to acting. In using anticipation and reaction animations, they structurally align the robot's motions to those of anthropomorphic characters. This draws on humans' intuitive understanding and helps the human extrapolate their understanding to the robot by analogy.

The other human belief update works only lightly use Analogical Transfer Theory. Kwon et al. [40] used difference detection by structurally aligning—as much as possible—a failed trajectory to an imagined successful trajectory. From this, the human learns about the delta between these trajectories, which helps them comprehend the robot's failure. Chakraborti et al. [36] also drew on difference detection by assuming that the human and robot have divergent but partially-aligned models of the environment and aligning differences with model patches, which aim to reconcile the human's model to match the robot's. Lastly, several methods [4], [37], [38] employed both inference projection and schema abstraction, though with minimal structural alignment: they presented similar training environments before assessing the humans' abilities to project in a slightly unfamiliar environment.

*3) Variation Theory:* For generalization, Dragan and Srinivasa [4] discretized a robot's goal space and generate motions for each goal. The human learned through familiarization: they showed the human examples of the robot's motions with varied goals while fixing the underlying policy. They report that familiarization—using generalization—improves the human's accuracy in predicting the robot's motion, but not as substantially as anticipated. To test humans' accuracy, this study asks users to select a motion trajectory from three different choices, each generated by a different policy. This choice relates to the principle of contrast, though it is used only to *test* the human and not to *teach* the human about the robot's behaviors. Variation Theory shows contrast should precede generalization: to learn about the robot's policy, the human might benefit from seeing the results of *varied* policies as focused aspects before seeing the results of varied environments.




(a) Feedback (b) Preferences [42] (c) Corrections [43]

Fig. 4: Teaching with reward (4a), preferences (4b), and corrections (4c). With reward (4a), the person observes the robot and rewards it for past actions, e.g., with a $\pm 1$ signal. With preferences (4b), the person selects between two or more candidate trajectories; the example in 4b structurally aligns three trajectories, supporting difference detection. Finally, with corrections (4c), a person modifies past trajectories—either through a GUI [43], [44] or physical manipulation [45], [46].

Huang et al. [37] and Lage et al. [38] incorporated generalization when teaching the human about a robot's objective. Both methods hold the policy, the focused aspect, constant while varying the environment. Lage et al. [38] additionally incorporated fusion by showing multiple varied policies side-by-side. Countering this, Chakraborti et al. [36] primarily used contrast. In their approach, the object of learning is not the policy, as is typical; instead, their object of learning is the true environment. They assume the human and the robot have different models of the environment—effectively, varying the environment. They reconcile these differences by updating the human's model with patches which explain the robot's model.

*4) Takeaways and Frontiers:* Analogical reasoning is an especially useful tool for assisting in the task of guiding humans to update their beliefs about a robot, especially in new encounters. The goal in such settings is to push the human to establish correct assumptions; analogy can rapidly accomplish this. Takayama et al.'s [35] approach of using animation principles to change how the human understands the robot is compelling, as this uses re-representation through anthropomorphic characteristics. Other works have explored the related question of how robots can benefit (or suffer) from playing to stereotypes [41]. This presents a frontier for efforts to support humans in learning about robots: designing and structurally aligning robots to appropriate analogical bases—whether through animation, stereotype play, visual design, or other means—can support the humans' belief update process and human-robot teaching and learning interactions in general.

### C. Teaching with Feedback

Fig. 4 shows interfaces for intuitive teaching. When teaching with feedback, a human gives feedback as they watch an agent act. These works are motivated by the idea that "you know it when you see it." This assumption seemingly undermines the need for human concept learning: if a human *knows it* just by seeing, why should the human first need to learn about the robot's capabilities and limitations? For some tasks, a human is able to provide feedback with little learning. For example, if giving feedback to a simulated car, a human

might say "Good robot!" when on the road, or "Bad robot!" otherwise [47]. For other tasks, though, a human might lack intuition for the constraints of a robotic platform, and give poor feedback due to malformed expectations, e.g., they might believe the robot can learn a behavior which its morphology cannot accommodate [39], [40]. Or, the robot might be making progress toward a satisfactory-but-unexpected policy (as seen in, e.g., [48]). In these cases, the human must learn about the robot's capabilities and limitations *before* providing feedback.

*1) Implementations:* In TAMER, a human-supplied reward signal is used to train a supervised learning module, which approximates the human's reward [49]. TAMER was first tested in simulated Tetris and Mountain Car environments [49], and later evaluated in robotics settings [50]. A modification to TAMER biases the agent toward taking non-optimal actions for the sake of stimulating human engagement [51].

The Advise method instead interprets feedback as a commentary on actions: reward is determined by the environment, and feedback is used to guide action exploration. Griffith et al. [52] demonstrated Advise on simulated game environments with simulated teachers, while Cederborg et al. [48] demonstrated Advise in a user study. Curiously, the real users outperformed the simulated users as real users were able to adapt to different but equally good strategies, while simulated users provided negative feedback if the behavior did not match their pre-planned strategy. The real human might also have preferred a different policy; nonetheless, they learn about the agent's policy and adapt to its learning trajectory.

TAMER and Advise do not consider the *teacher's strategy*. Loftin et al. [53] conducted a user study showing that some users bias toward giving positive feedback while others are more balanced. They introduced SABL: a method which numerically manipulates human feedback based on the teacher's strategy; this strategy is learned during interaction. A similar approach, COACH, uses the insight that human feedback depends on the robot's current aptitude [54]. Accounting for this policy-dependency in feedback, COACH interprets feedback as an advantage function. COACH assumes that the human teacher is learning over time: to give feedback, the human must first learn about the current policy.

*2) Analogical Transfer Theory:* None of these works extensively use structural alignment—instead, these all require prolonged observation to learn about a policy. Nonetheless, COACH uses difference detection [54], though without the structural alignment recommended by Analogical Transfer Theory. COACH assumes that human feedback is policy-dependent, varying as the policy improves or degrades over time, and difference detection is needed to assess how the policy changes. This presents an opportunity for the application of Analogical Transfer Theory. Instead of tasking the human with watching the robot repeatedly attempt a task, and hope that the human identifies the differences or similarities over different trials, applying structural alignment would help. After policy updates, a supporting communication and intervention interface could provide highlighting or other means of identifying differences to support the human's assessment.

*3) Variation Theory:* In feedback user studies [48], [49], [53]–[56], variation is implicitly present, though not espoused as a core principle to support human learning. These works assume that a human is able to watch the agent act and give appropriate feedback in response. As the human observes, they are presented with varied environments, states, and actions in all cases. For several of these approaches [49], [53]–[56], the policy is updated in real-time in response to the human's feedback, and is therefore also varied simultaneously. All of these examples use fusion. Countering this, Cederborg et al. [48] hold the policy constant during feedback collection, and thus supports generalization, instead. They do so to support different experiment conditions, but this may coincidentally increase the human's aptitude for teaching, as they are better able to learn about the effects of the policy. To better support human concept learning, future assessments and algorithms should present this variation with a deliberate and managed approach. This is an opportunity for future research.

*4) Takeaways and Frontiers:* While none of these feedback-based approaches extensively incorporate principles from either theory of human concept learning, "you know it when you see it" isn't enough to support human-robot teaching and learning in the forseeable future. Both SABL [53] and COACH [54] observe that learning from feedback cannot be formulated as a person- or policy-independent algorithm; nonetheless, the tasks tested in all of these feedback works conform to the expectation that the human can give good feedback after short periods of unstructured observation. As these techniques accommodate increasingly complex tasks, and as the features used to complete a task further diverge between humans and robots, this assumption will ring hollow. Using human concept learning theories—particularly by using variation to communicate the behaviors of the current and future policies—can mitigate these challenges.

### *D. Teaching with Preferences*

Implicitly, preferences mandate that interfaces present multiple options: does the human prefer A or B? In this manner, preferences naturally rely on concepts of variation, and the requisite structural alignment supports the human in teaching.

*1) Implementations:* Sadigh et al. [57] take an active learning approach to selecting trajectory pairs for soliciting human preferences. They formulate this as volume removal over the distribution of potential reward functions, wherein each preference should maximize the volume removed from this hypothesis space. Their interface asks humans to compare trajectories generated by different policies in the same scenario. Bıyık et al. [5] observe that volume removal can fail to support human teaching as the robot can ask the human to compare two trajectories with imperceptible differences. They instead introduce an information gain approach for trajectory selection; this approach selects queries by maximizing both the robot's uncertainty over the human's response and the *human's uncertainty* in providing a preference.

Instead of generating trajectories from different policies, Christiano et al. [58] roll out the same policy numerous

times in slightly varied environments. Stochasticity in the policy, transition dynamics, and environment introduce variation. They similarly use active learning to select trajectory clips which are maximally uncertain under their reward model. Ibarz et al. [59] uses this same method, but, instead of learning from tabula rasa, they initialize their agent through imitation learning and then use preferences for policy refinement. Curiously, they found active querying did not increase the agent's learning performance, while slowing down sampling. They thus opted to adopt random sampling for trajectory clips instead.

Jain et al. [42] formulate learning from preferences as an iterative process. They observe that humans are typically unable to provide optimal demonstrations, but can re-rank trajectories iteratively. For this, they learn a model of a user's scores for trajectories. To generate trajectories for comparison, they fix the starting and ending states, and use a rapidly-exploring random tree planner with heuristics to encourage diversity. Lastly, Wilson et al. [60] approximate a policy distribution using a Bayesian likelihood function. Using this distribution, they sample two policies and generate two trajectories for comparison. They compare two active approaches for selecting policies for comparison. First, they consider policies which generate different behaviors when rolled out. Second, they consider the expected belief change in the hypothesis space.

*2) Analogical Transfer Theory:* Preferences naturally incorporate structural alignment. If two choices are not structurally aligned, it can be challenging or impossible to discern their differences—a prerequisite for providing preferences. While most of these works engage structural alignment and difference detection by presenting trajectories side-by-side [5], [57]–[59], they do not take full advantage of Analogical Transfer Theory. Christiano et al. [58] and Ibarz et al. [59] present side-by-side trajectory snippets with differing start and end states; these differences make the snippets are hard to compare. Most notably, Jain et al. [42] present figures which show structurally-aligned, overlaid trajectories with shared start and end states (Fig. 4b), but, in their experiments, humans watched a robot perform trajectories sequentially, and were then asked to rank them. These authors cite this as a limitation of their work: they note that making users memorize these trajectories and not aligning them hinders the efficacy of their approach.

*3) Variation Theory:* In all of these works, variation in trajectories is a prerequisite for *robot* learning. Through experiencing this variation, the human is also better equipped to understand the robot's policy. These works all incorporate contrast [5], [42], [57], [60] or generalization [58], [59]. They support contrast by varying the underlying policy [5], [42], [57], [60]—whether by varying their parameterizations (e.g., [57]) or by using an alternative for comparison (e.g., [42]). To support comparisons between policies, these works hold all other aspects invariant—e.g., the starting state, and sometimes the end state [42]. These works incorporate generalization by requesting preferences over multiple trajectory segments from the same policy [58], [59]. The same policy may also present different trajectories for comparison, as the policies and transition dynamics may be stochastic.

*4) Takeaways and Frontiers:* Teaching a robot with preferences naturally incorporates principles from both Variation and Analogical Transfer Theory. Variation is implicitly present, as the person is tasked with choosing between two or more options. Analogical Transfer Theory is also incorporated through the use of structural alignment, either presenting choices side-by-side or overlaid to support difference detection. In these settings, structural alignment is often still hard to comprehend, and could be improved through by *maximizing* this alignment—starting by presenting trajectories with shared start and/or end states. Despite this natural proclivity for engaging human concept learning, preferences approaches again commonly assume that the human either learns about the robot's policy from observation or is able to give feedback without any context. This is a missed opportunity.

### *E. Teaching with Corrections*

Lastly, we consider teaching with corrections. A robot starts with an initial policy, and the human is tasked with correcting it—e.g., by teaching the robot about preferred action choices.

*1) Implementations:* Alexandrova et al. [43] introduced a corrections interface where a user first provides a demonstration to a robot and subsequently corrects its policy. This interface is gnarly and complex: after providing demonstrations, users can modify past demonstrations by changing frames of reference or by deleting intermediary poses and/or landmarks—all from the robot's point of view. They found that visualizing the robot's learning was extremely useful, and deleting poses was also helpful. Using the same interface, Forbes et al. [44] sourced corrections from a crowd.

Bajcsy et al. [45] reframed corrections from the perspective of *physical* HRI. Humans often physically engage with robots—for example, by pushing it out of the way. These interactions are typically regarded as disturbances, but Bajcsy et al. noted some useful information. They introduced an optimization approach to update the robot's trajectory to align with the human's physically-corrected trajectory. In subsequent work, Bajcsy et al. [46] introduced a method where instead of using the full, corrected trajectory as the optimization target, they allow only one feature to vary at a time.

*2) Analogical Transfer Theory:* Correction-based systems naturally incorporate difference detection. In all of these approaches, the robot starts with some trajectory which needs to be corrected. To support difference detection, this trajectory is structurally-aligned with a corrected trajectory—either through a visualization to teach the human about what to teach [45], [46] or through an omnipresent interface used for supporting the user to correct the robot's behaviors and evaluate progress [43], [44]. In the former, the difference is only shown at the beginning of the interaction. While this assists the human in learning about the task expectations, it requires them to recall the behavior. A better interface supports and maintains this visualization throughout the interaction.

*3) Variation Theory:* Alexandrova et al. [43] found that repetition is remarkably beneficial for corrections-based teaching: in their system, a human trains a robot through demonstration

and subsequent corrections in a GUI. They find the mere presence of these visualizations and the ability to repeatedly observe actions to be the greatest benefit to humans' teaching. The repetition step of Variation Theory is often overlooked or skipped—but this result suggests it can be an effective supporting methodology. Bajcsy et al. [45] incorporated fusion in their first approach: they tasked a human with physically manipulating a robot to correct its expressed trajectories, where the robot learns from these corrections. Bajcsy et al. [46] instead use contrast, wherein the robot isolates updates to the single feature which changed most in the human's corrections. They compare their fusion and contrast implementations, and find the latter to be more effective for the robot's learning.

*4) Takeaways and Frontiers:* Corrections are a powerful teaching tool. Intuitively, we might expect humans would prefer to correct every aspect of a robot's behavior simultaneously; after all, that is efficient! In practice, Bajcsy et al. [46] demonstrate that this assumption is flawed—and their implementation reflects Variation Theory's insights. They find that humans are in fact more adept at correcting a trajectory by varying *one feature at a time*. Although this work is framed from the perspective of robot learning, Variation Theory suggests its implications will also hold for the inverse, the human's learning. By iteratively changing one feature at a time, both the human and robot learn from a varied critical aspect, while holding all else invariant. This approach supports the human in discerning the impact of that individual change.

## IV. Design Guidance & Future Directions

Human concept learning provides a new lens for human-robot teaching and learning. Without explicitly consulting these theories, past approaches have incorporated a number of their insights. Still, many gaps and opportunities remain.

*1) Supporting Analogical Transfer:* When teaching a robot, humans are likely to employ analogy to inform their beliefs about the robot, as well as their beliefs over how the robot will use their teaching signal to change its behaviors [61]. Humans might use any number of bases to inform their interactions: human or animal behaviors, virtual character behaviors, or past experiences with other robots. Only three systems we analyzed considered base case retrieval as a design input [4], [35], [40] by using exaggerated, anthropomorphic, and/or animated behaviors. Future efforts in human-robot teaching and learning should build on these ideas, and provide further support for base anchoring: instead of giving the person independence in selecting their own base, the presentation of the robot should guide the person to select an appropriate and desirable base.

Analogy's backbone is structural alignment. This is used throughout many of the human-robot teaching and learning systems we analyzed, usually to support difference detection. These prior works often assess whether a human is able to perceive some difference or provide some teaching signal [5]; nonetheless, these works rarely considered how to maximally-align information such that the human is best positioned to make these assessments. For example, in asking users to compare trajectory snippets, some works showed trajectories to users that both started and ended in different states, while also expressing variation in the interim [58], [59]. Such tasks ignore structural alignment, and are unduly challenging. Designing for maximal structural alignment is a promising path forward.

*2) Supporting Structured Variation:* Variation Theory informs how humans learn to discern the latent structure of new concepts, and to understand the bounds of their applicability. This theory does not, however, inform us of exactly how it should be applied in HRI settings. Specifically, Variation Theory requires the designation of an object of learning and a number of aspects related to that object of learning. These can be inferred, to some extent, from the task structure (e.g., see the appendix). Even so, identifying exactly how to group and present aspects to facilitate human concept learning is a design task, and requires substantial experimentation and prototyping.

Variation Theory then proposes a strict sequence for efficient concept learning: repetition, then contrast, then generalization, then fusion. Despite this, none of the 35 works we looked at followed this prescribed sequence. In policy summarization, Sequeira and Gervasio [3] noted that finding an appropriate amount of variation when using fusion was challenging: too much and users were confused about an agent's capabilities and limitations; too little and users believed agents to be either more competent or less competent than they really are. Using the prescribed structured presentation of variation is uncharted territory in human-robot teaching and learning systems, but it offers a potential resolution to this challenge and may additionally elevate human ability to learn about robots.

In this review, the focus is implicitly on helping the robot learn from human teaching, and not on helping the human be a better teacher. The human is treated as an oracle—able to provide a perfect assessment of behavior at any time. Nonetheless, when variation is used as a tool to guide the robot's learning, the human inadvertently learns too (e.g., [46]). Future algorithms and interfaces should consider this more directly: structured variation can support both the human and the robot in discerning critical aspects, even if these aspects are not the same for both entities. A symbiotic approach to human-robot teaching and learning could optimize the data requirements to satisfy the variation needs of *both* human and robot.

*3) Explainability:* In AI and HRI, explanations aim to support debugging, to calibrate end user trust, and to moderate model reliance; these goals make explanations promising for human-robot teaching and learning. Despite the introduction of many methods (of sometimes dubious quality [62], [63]), explanations often do not help people achieve these goals [64], [65]. Engaging human concept learning can help humans use generated explanations effectively. Onboarding for these methods is important [66] but often overlooked [67]. Onboarding can use Variation Theory to help users understand the bounds and limitations of explanations, and Analogical Transfer to help users bootstrap prior knowledge onto these new methods.

## V. Acknowledgements

This material is based upon work supported by the NSF under Grant No. 2107391. SB is supported by an NSF GRFP.

## REFERENCES


[1] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: lessons we have learned," *The International Journal of Robotics Research*, p. 0278364920987859, 2021.

[2] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[3] P. Sequeira and M. Gervasio, "Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations," *Artificial Intelligence*, vol. 288, p. 103367, 2020.

[4] A. Dragan and S. Srinivasa, "Familiarization to robot motion," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014, pp. 366–373.

[5] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences," *International Journal of Robotics Research*, 2020.

[6] Y. Amitai and O. Amir, ""i don't think so": Disagreement-based policy summaries for comparing agents," *arXiv preprint arXiv:2102.03064*, 2021.

[7] D. Gentner and L. A. Smith, "Analogical learning and reasoning," *The Oxford handbook of cognitive psychology*, pp. 668–681, 2013.

[8] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive science*, vol. 7, no. 2, pp. 155–170, 1983.

[9] F. Marton, *Necessary conditions of learning*. Routledge, 2014.

[10] F. Marton and S. A. Booth, *Learning and awareness*. psychology press, 1997.

[11] B. Rittle-Johnson and J. R. Star, "Does comparing solution methods facilitate conceptual and procedural knowledge? an experimental study on learning to solve equations." *Journal of Educational Psychology*, vol. 99, no. 3, p. 561, 2007.

[12] D. Gentner, J. Loewenstein, and L. Thompson, "Learning and transfer: A general role for analogical encoding," *Journal of Educational Psychology*, vol. 95, pp. 393–408, 2003.

[13] K. J. Kurtz, C.-H. Miao, and D. Gentner, "Learning by analogical bootstrapping," *The Journal of the Learning Sciences*, vol. 10, no. 4, pp. 417–446, 2001.

[14] M. L. Gick and K. J. Holyoak, "Schema induction and analogical transfer," *Cognitive psychology*, vol. 15, no. 1, pp. 1–38, 1983.

[15] F. G. Paas and J. J. Van Merriënboer, "Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach." *Journal of educational psychology*, vol. 86, no. 1, p. 122, 1994.

[16] J. L. Quilici and R. E. Mayer, "Role of examples in how students learn to categorize statistics word problems." *Journal of Educational Psychology*, vol. 88, no. 1, p. 144, 1996.

[17] S. Y. A. Tong, "Applying the theory of variation in teaching reading." *Australian Journal of Teacher Education*, vol. 37, no. 10, p. 1, 2012.

[18] A. Driver, K. Elliott, and A. Wilson, "Variation theory based approaches to teaching subject-specific vocabulary within differing practical subjects," *International Journal for Lesson and Learning Studies*, 2015.

[19] H. C. Lam, "Elaborating the concepts of part and whole in variation theory: The case of learning chinese characters," *Scandinavian Journal of Educational Research*, vol. 58, no. 3, pp. 337–360, 2014.

[20] L. M. Ling, P. Chik, and M. F. Pang, "Patterns of variation in teaching the colour of light to primary 3 students," *Instructional Science*, vol. 34, no. 1, pp. 1–19, 2006.

[21] M. Mhlolo, "The merits of teaching mathematics with variation," *Pythagoras*, vol. 34, no. 2, pp. 1–8, 2013.

[22] T. J. Bussey, M. Orgill, and K. J. Crippen, "Variation theory: A theory of learning and a useful theoretical framework for chemical education research," *Chemistry Education Research and Practice*, vol. 14, no. 1, pp. 9–22, 2013.

[23] J. Suhonen, E. Thompson, J. Davies, and G. Kinshuk, "Applications of variation theory in computing education," in *Seventh Baltic Sea Conference on Computing Education Research. Koli, Finland*, 2008.

[24] M. Ling Lo, *Variation theory and the improvement of teaching and learning*. Göteborg: Acta Universitatis Gothoburgensis, 2012.

[25] F. Marton, A. B. Tsui, P. P. Chik, P. Y. Ko, and M. L. Lo, *Classroom discourse and the space of learning*. Routledge, 2004.

[26] V. V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah, "Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2394–2401, 2018.

[27] R. Kitagawa, Y. Liu, and T. Kanda, "Human-inspired motion planning for omni-directional social robots," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 34–42.

[28] D. Amir and O. Amir, "Highlights: Summarizing agent behavior to people," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 1168–1176.

[29] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 2017, pp. 303–312.

[30] S. H. Huang, K. Bhatia, P. Abbeel, and A. D. Dragan, "Establishing appropriate trust via critical states," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3929–3936.

[31] M. L. Olson, R. Khanna, L. Neal, F. Li, and W.-K. Wong, "Counterfactual state explanations for reinforcement learning agents via generative deep learning," *Artificial Intelligence*, p. 103455, 2021.

[32] Y. Zhou, S. Booth, N. Figueroa, and J. Shah, "Rocus: Robot controller understanding via sampling," *Conference on Robot Learning*, 2021.

[33] O. Amir, F. Doshi-Velez, and D. Sarne, "Summarizing agent strategies," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 5, pp. 628–644, 2019.

[34] A. Furnham and H. C. Boo, "A literature review of the anchoring effect," *The journal of socio-economics*, vol. 40, no. 1, pp. 35–42, 2011.

[35] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: improving robot readability with animation principles," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 69–76.

[36] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "Plan explanations as model reconciliation–an empirical study," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 258–266.

[37] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Autonomous Robots*, vol. 43, no. 2, pp. 309–326, 2019.

[38] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir, "Exploring computational user models for agent policy summarization," in *IJCAI: proceedings of the conference*, vol. 28. NIH Public Access, 2019, p. 1401.

[39] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[40] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing robot incapability," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 87–95.

[41] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[42] A. Jain, S. Sharma, T. Joachims, and A. Saxena, "Learning preferences for manipulation tasks from online coactive feedback," *The International Journal of Robotics Research*, vol. 34, no. 10, pp. 1296–1313, 2015.

[43] S. Alexandrova, M. Cakmak, K. Hsiao, and L. Takayama, "Robot programming by demonstration with interactive action visualizations." in *Robotics: science and systems*, 2014.

[44] M. Forbes, M. Chung, M. Cakmak, and R. Rao, "Robot programming by demonstration with crowdsourced action fixes," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 2, no. 1, 2014.

[45] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning robot objectives from physical human interaction," in *Conference on Robot Learning*. PMLR, 2017, pp. 217–226.

[46] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning from physical human corrections, one feature at a time," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 141–149.

[47] S. Reddy, A. Dragan, S. Levine, S. Legg, and J. Leike, "Learning human objectives by evaluating hypothetical behavior," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8020–8029.

[48] T. Cederborg, I. Grover, C. L. Isbell Jr, and A. L. Thomaz, "Policy shaping with human teachers." in *IJCAI*, 2015, pp. 3366–3372.

[49] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proceedings of the fifth international conference on Knowledge capture*, 2009, pp. 9–16.

[50] W. B. Knox, P. Stone, and C. Breazeal, "Training a robot via human feedback: A case study," in *International Conference on Social Robotics*. Springer, 2013, pp. 460–470.

[51] W. B. Knox, B. D. Glass, B. C. Love, W. T. Maddox, and P. Stone, "How humans teach agents," *International Journal of Social Robotics*, vol. 4, no. 4, pp. 409–421, 2012.

[52] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," 2013.

[53] R. Loftin, J. MacGlashan, B. Peng, M. Taylor, M. Littman, J. Huang, and D. Roberts, "A strategy-aware technique for learning behaviors from discrete human feedback," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.

[54] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, D. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[55] Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. B. Knox, "The empathic framework for task learning from implicit human feedback," in *Conference on Robot Learning*. PMLR, 2020.

[56] J. Lin, Q. Zhang, R. Gomez, K. Nakamura, B. He, and G. Li, "Human social feedback for efficient interactive reinforcement agent learning," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 706–712.

[57] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions." in *Robotics: Science and Systems*, 2017.

[58] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in neural information processing systems*, 2017, pp. 4299–4307.

[59] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," *Advances in neural information processing systems*, 2018.

[60] A. Wilson, A. Fern, and P. Tadepalli, "A bayesian approach for policy learning from trajectory preference queries," *Advances in neural information processing systems*, vol. 25, pp. 1133–1141, 2012.

[61] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, "Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2429–2437.

[62] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *arXiv preprint arXiv:1810.03292*, 2018.

[63] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?" *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[64] H. Suresh, N. Lao, and I. Liccardi, "Misplaced trust: Measuring the interference of machine learning in human decision-making," in *12th ACM Conference on Web Science*, 2020, pp. 315–324.

[65] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[66] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, ""hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making," *Proceedings of the ACM on Human-computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019.

[67] A. Hopkins and S. Booth, "Machine learning practices outside big tech: How resource constraints challenge responsible development," *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2021.