How to Understand Your Robot: A Design Space Informed by Human Concept Learning

Serena Booth[1], Sanjana Sharma[2], Sarah Chung[3], Julie Shah[1], and Elena Glassman[3]

I. INTRODUCTION

Useful robots must be able to learn new skills [1] and adapt to human norms and preferences [2]; nonetheless, interactive teaching in unstructured environments remains a pipe dream. Many prior works have introduced algorithms to enable humans to teach AIs or robots through intuitive teaching signals like feedback, preferences, guidance, or corrections. These algorithms have made progress, but remain inadequate for use in the real world—with real robots, real humans, in real messy human spaces. The use of intuitive teaching signals is only one step in the larger communication loop between a human teacher and their robot (Fig. 1). Neglecting other stages of this communication loop—either by not helping the human understand the robot's *current* behavior or by not helping the human understand the impact of their teaching on the robot's *future* behavior—can lead to violations of human expectations [3], [4], [5]. This can severely diminish both the human's effectiveness as a teacher and the robot's effectiveness as a collaborator.

Our key insight is that humans must learn about a robot's capabilities and limitations as a prerequisite for effective teaching. While past works assume that humans are able to provide high quality teaching signals after briefly watching a robot act in an environment, there are many reasons why this premise is often false. Humans are sometimes able to learn about some robot motions, but this process is time-consuming; further, unnatural robot motions—those which are not human-like or animal-like—are hard for humans to learn through unstructured observation [6]. Detecting differences in robot capabilities and limitations over time can be impossible when those differences are imperceptible, not structurally aligned for ready comparison, or incomparable for any number of other reasons [7]. Finally, observing a robot only perform well or only perform poorly biases the human's understanding of its competency [5].

Instead of assuming that humans learn by observation, interfaces should mediate human-robot teaching and learning by systematically guiding the human's learning about the robot's behaviors. These interfaces must help humans to (1) learn about the robot's capabilities and limitations, (2) teach the robot by providing a signal like feedback or preferences for those expressed capabilities and limitations, and (3) learn about the capabilities and limitations of new robot behavior candidates, and compare these to prior candidates. These tasks are especially challenging as robot behaviors are complex, playing out over potentially complicated robot dynamics in potentially changing environments.

To tame this complexity, we consider relevant human-centered principles from theories of human concept learning. We identify two relevant and well-supported theories: Analogical Transfer [8], [9] and Variation Theory [10], [11]. Using these theories, we explore the interface design implications provided by this body of previously not-consulted knowledge about how humans come to understand existing phenomena and make predictions about as-yet unrevealed facts and futures—or, about how to understand your robot. As a result of this study, we contribute a design space for future interfaces to support robot teaching informed by theories of human concept learning. We find that prior works cover a diverse portion of the design space, and we identify numerous opportunities to apply human concept learning to better human-robot teaching and learning. In this short paper, we contribute an overview of the design space; in an extended version, we contribute an additional meta-study of 40 prior works on human-robot teaching and learning. We provide the extended version as supplementary material.

II. THEORIES OF LEARNING

In many domains, theories of human concept learning have been refined through controlled studies testing the learning implications of various interventions and their application in curriculum (and often implicitly, interface) design. We look to these theories, specifically analogical learning theory and the variation theory of learning, to inform how interfaces can best mediate the practice of humans building accurate schemas and mental models of robot behaviors.

1) Analogical Learning Theory: Analogical learning theory asserts that *analogy*, or finding and using relational commonalities, is the primary building block of concept learning [8], [12], [9]. In analogy, a familiar domain known as the *base* informs how humans understand and draw new inferences about a less familiar domain, the *target*. In analogical reasoning, a person must first identify a known base domain which is relationally similar to the target. Second, the person must map the analogy by *structurally aligning* the base and target. Lastly, the person must evaluate the analogy and assess any inferences drawn from it. People learn easily and intuitively by analogy: structural alignment and analogy formation allow us to readily form new inferences about

[1] Serena Booth and Julie Shah are employed by MIT CSAIL. {serenabooth, julie_a_shah}@csail.mit.edu

[2] Sanjana Sharma is employed by the Harvard Graduate School of Design. sharma.sas@gmail.com

[3] Sarah Chung and Elena Glassman are employed by the Harvard Paulson School of Engineering sc232@cornell.edu, glassman@seas.harvard.edu

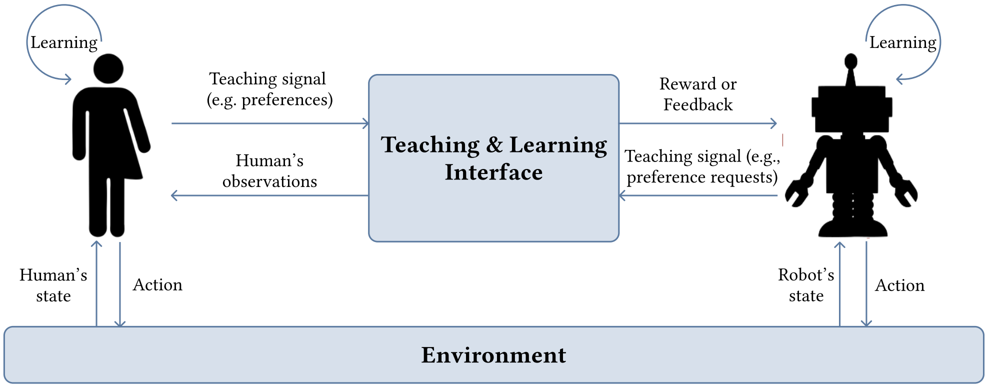

Fig. 1. An overview of human-robot teaching and learning systems. Over time, the robot takes actions and observes the changing state of the environment; the human separately observes the goings-on. At any time, the human may provide teaching signal to the robot. A teaching and learning interface supports the human's teaching efficacy—particularly by incorporating theories of human concept learning. This communication loop continues indefinitely.

an ill-understood target (**inference projection**), to construct new schemas or mental models through the process of mapping relations (**schema abstraction**), to detect differences between bases and targets (**difference detection**), and to re-represent domains at alternative levels of abstraction, making the analogy is more applicable (**re-representation**).

Analogical learning theory can inform interface design. When faced with a novel target, people implicitly seek a comparison base domain from memory and search for relational commonalities between the target and the base. If a person succeeds in this structural alignment process, their understanding of the target is bolstered by any common principles shared between the target and the base. In this process, there are two notable opportunities for interfaces to assist in analogy formation and reliance. First, humans are bad at accessing analogous base cases from memory [13], so an interface has an opportunity to assist by prompting the user to recall a base case. Second, analogical learning relies on structural alignment, which maximally highlights the commonalities between the target and the base: an interface has an opportunity to present data in a structurally aligned manner such that humans are more readily able to draw new inferences about the target or to detect differences.

Of course, analogical learning theory in not a catch-all solution to allowing humans to better understand robot collaborators. One notable caveat is that humans' aptitude for analogical reasoning is imperfect: a person is susceptible to forming analogies over non-corresponding bases and targets. In doing so, they can form inaccurate mental models and schemas. For human-robot or human-AI interaction, candidate bases often include human or animal behavior, or prior experience with automation (e.g., other robot morphologies or video games). An incorrect base for analogical reasoning is a candidate explanation for why humans are more readily able to learn about natural motions when interacting with a robot, as these motions are more human-like and so better supported analogical learning [6].

2) Variation Theory of Learning: Variation theory argues that in order to comprehend some object of learning, a person must first discern the *critical aspects* through some *critical features* of that object. Aspects are concepts—for example, color—while features are instantiations of aspects—for example, the color green or the color red. Critical aspects are those which are strictly necessary to understand the concept, and not those which are merely contingent. To succeed in discernment and learning, the person must experience varied instantiations of the critical aspect(s). Experiencing a sequence of varied and invariant aspects—some critical, some not—is an essential ingredient for learning.

To apply variation theory, we must designate some aspect or aspects to be *focused*. Ideally, the focused aspect would be a critical aspect, but this is not necessary, and usually cannot be guaranteed. Having identified at least one focused aspect, variation learning follows an ordered sequence of four key processes which use consistent patterns of variance and invariance to assist a person's natural inductive reasoning ability to more accurately infer how these focused aspects contribute to the object of learning. *For each focused aspect:*

1) **Repetition.** Both the focused aspect and any other aspects should be held constant and presented to a learner. To learn about a robot's behaviors, repetition prescribes that the human should see the robot act in same the environment at least once.
2) **Contrast.** The focused aspect should vary, while any other aspects are held constant. The robot should exhibit various behaviors for comparison, as the focal aspect is set to different feature values while the robot is operating in the same environment. The human may develop a preference for certain feature values over others within that aspect, which will designate them as critical features of the human's developing preferences for what behaviors they do and do not like.
3) **Generalization.** The focused aspect should be held constant, while any other aspects vary. After experiencing contrast, the human should see how the robot's behavior varies in new environments, given the particular selected feature value of the focused aspect.
4) **Fusion.** All aspects vary simultaneously. The human should experience both the robot's behavior and the environment varying at the same time, mimicking the variation they will experience "in the real world."

Variation theory argues that contrast must precede generalization in learning. This strict sequencing is often ignored when applying the theory, and humans are still able to learn [10]: If a human has already learned to discern a feature—for example, if they have learned that a 7DoF robot arm orients its shoulder joints before moving its wrist when performing reaching tasks—then they can often learn directly from experiencing generalization, e.g., of different reaching targets. If the human has not yet learned to discern a feature (as is possible for these unnatural, not-human-like motions [6]), contrast should strictly precede generalization.

III. Design Space

Following the outline of Fitzmaurice et al. [14] and subsequent works [15], [16], we developed a design space (Appx., Fig. 2) for human-robot teaching and learning interfaces which is informed by human concept learning theories. Appx. Fig. 2 additionally presents six representative prior works, illustrating how these works cover this design space.

A. Human Learning

Choosing how to display the robot's capabilities and limitations is critical for effective human learning. These choices determine whether the human is able to discern the information and context needed to provide high quality feedback. The following design choices inform the approach to supporting the human's learning.

- *Object(s) of Learning.* What concept or concepts must the human learn? A common choice is for the human to learn about the robot's policy or its expressed trajectories. However, this is not the only choice: the human might instead need or want to learn about individual components of an MDP, such as the plausible start states or the system's transition dynamics. The human might wish to learn about the environments in which the robot might be deployed, or about feature representations.
- *Focused Aspect(s).* For any given object of learning, an intermediary focused aspect can inform the concept learning method. Focused aspects may include start states, states, actions, transitions, trajectories, policies, environments, and/or features.
- *Concept Learning Method(s).* Which human concept learning approaches does the interface employ? Analogical learning concepts include inference projection, schema abstraction, difference detection, and re-representation. Variation theory concepts include repetition, contrast, generalization, and fusion.
- *Number of Aspects Shown Simultaneously.* For each aspect (e.g., policies, trajectories, environments, transitions, actions, states, start states, and/or state features), how many instantiations are *simultaneously* shown to the human? Showing simultaneous aspects is an especially useful tool for the structural alignment process in analogical learning theory [17] and supports humans' perception of the variation present in the contrast and generalization steps of variation theory without requiring the human to recall what they've seen previously.

B. Interface Support for Human Learning

Given a plan for what the human should learn and the general concept learning approach, the next choices on the design space inform how this learning is practically achieved. How does the interface present information to the human to support them as an effective teacher?

- *Level of Zoom.* How is the robot's behavior presented to the human? Is it highly detailed, focusing on, for example, individual features or states? Or is the focus less detailed but more information dense—for example, focusing on full trajectories or environments?
- *Focused Aspect Display Method(s).* Is a single, individual focused aspect displayed? Are there multiple focused aspects, presented sequentially? Side-by-side? Overlaid? This choice of display method relates to analogical learning theory, as simultaneously presented information is better structurally aligned [17].
- *Grouped aspects.* Are aspects grouped? If so, are start states grouped, or all states, or actions, or transition probabilities, or environments, or trajectories, or policies? Grouping is most useful for structural alignment.
- *Grouping method.* How are aspects grouped? By internal quality metrics—like uncertainty? By metrics computed on trajectory rollouts—like time or proximity? Randomly? By some metric of similarity or distance?
- *Selection criteria.* How is data—whether grouped or not grouped—selected to be presented to the human? By internal quality metrics—like uncertainty? By metrics computed on trajectory rollouts—like time to completion or proximity to an obstacle? Randomly? By some metric of similarity or distance?
- *Visualization medium(s).* How is data presented to the human? Does the interface use text, audio, haptic, or visual displays? If the interface is visual, does it support only static images, dynamic images, e.g., GIFs or videos, or is it fully immersive—through AR/VR/MR, or through co-presence with a physical robot?
- *Visualization technique(s).* What visualization techniques does the interface support? Examples include opacity; time manipulations for structural alignment (e.g., dynamic time warping); progressive disclosure wherein more data is disclosed over time or exposure; animation which supports analogical comparisons; overlay which supports structural alignment and visualizing variation; juxtaposition, which supports structural alignment for alignable differences; and the presentation of auxiliary data, for example showing uncertainty.
- *Initiative* Who initiates updates to the robot's learning? Does the robot query the human, or does the human prompt the robot? Is initiative shared, with the human or the robot interchangeably guiding the learning process?

C. Robot Teaching

After learning about the robot's current capabilities and limitations, the human should be empowered with the ability to teach the robot. Ideally, the space of interactivity for robot teaching would be flexible to any human intent. In practice, though, algorithms typically remain constrained to a single form of feedback, and this flexibility remains an unrealized goal. In the interim, the interface must choose which interaction modalities to accommodate.

- *Communication Medium.* What communication modalities does the interface support for human teaching? Text, audio? Can the human provide latent signals like emotional responses? Is there a GUI? Does the person interact with the robot through some intermediary hardware, like a keyboard or remote? Or, perhaps, are they able to physically manipulate the robot to communicate—for example, by providing physical corrections?
- *Feedback Target(s).* Should feedback be interpreted as a commentary on a state, a state action pair, a feature, a trajectory segment, a full trajectory, or a policy?
- *Feedback Mechanism(s).* What modalities can the person use to teach the robot? Can they use a latent signal which encodes information—like a raised eyebrow? Can they use a binary signal ("good robot" or "bad robot"), or a scalar reward? Otherwise, can they use preferences, corrections, advice, or explanations to teach?

IV. Discussion

Human concept learning provides a new and systematic lens with which to consider human-robot teaching and learning. Without explicitly considering human concept learning, past approaches have explored a diverse portion of our design space. We include a full meta-study of these prior works as supplemental material, and in an extended version of this work. Despite the coverage of these past efforts, many gaps and opportunities for better integrating human concept learning into the human-robot communication loop remain.

A. Supporting Analogy

When teaching a robot, humans are likely to employ analogy to inform their beliefs of the robot's capabilities and limitations, as well as their beliefs over how the robot will use their teaching signal to change its behaviors [18]. Humans might use any number of bases to inform their interactions, such as natural phenomena, human and animal behaviors, virtual character behaviors, or past experiences with other robots. In our meta-study, only 3 of the 40 systems we analyzed considered base case retrieval as a design input [19], [6], [20] 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 considered, usually to support difference detection. These prior works routinely assess whether a human is able to perceive some difference or provide some teaching signal [7]; nonetheless, rarely did these works explicitly consider 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 [21], [22]. Without alignment, such tasks are unnecessarily challenging for humans. Implementations of structural alignment differ substantially in their effectiveness; future work should aim to incorporate best practices.

B. Supporting Structured Variation

Variation theory proposes a strict sequence for efficient learning: first contrast, then generalization, then fusion. Despite this, none of the 40 works we looked at followed this prescribed sequence. Still, people can learn—if not as effectively [10]. In their policy summarization work, Sequeira and Gervasio 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 [5]. Using the prescribed structured presentation of variation is uncharted territory in human-robot teaching and learning systems, but it offers a resolution to this challenge and may additionally elevate human ability to learn about robots.

Throughout most of the systems we have discussed—those where a human teaches a robot with an intuitive teaching signal—the focus is implicitly on helping the robot to learn from human teaching, and not on helping the human to be a better teacher. The human is treated as an oracle—able to provide an assessment of any behavior at any time with perfect knowledge. Nonetheless, when variation is used as a tool to guide the robot's learning (e.g., Bajcsy et al. [23]), the human may inadvertently learn too. Future algorithms and interfaces should consider this relationship more directly: variation learning is useful to support both the human and the robot in discerning critical aspects, even if these aspects are not one and the same for both entities. A symbiotic approach to teaching and learning could optimize the data requirements to satisfy the variation needs of both human and robot.

V. Conclusion

As a framing, human concept learning has the potential to help us reconsider human-robot interaction problems, especially for the challenge of teaching and learning. These theories should be viewed as a source of inspiration when designing future systems and algorithms in these domains. To this end, we have explored how cognitive theories of human concept learning can inform a design space for human-robot teaching and learning systems. Future researchers can also use our contributed design space to assess how their approaches fit into this landscape, and so guide their consideration of how to incorporate additional design principles from human concept learning into human-robot interaction.

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] 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," *arXiv preprint arXiv:1701.06049*, 2017.

[4] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," *arXiv preprint arXiv:1701.08317*, 2017.

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

[6] 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.

[7] 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," *arXiv preprint arXiv:2006.14091*, 2020.

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

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

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

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

[12] 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.

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

[14] G. W. Fitzmaurice, H. Ishii, and W. A. Buxton, "Bricks: laying the foundations for graspable user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, pp. 442–449.

[15] B. Hartmann, L. Yu, A. Allison, Y. Yang, and S. R. Klemmer, "Design as exploration: creating interface alternatives through parallel authoring and runtime tuning," in *Proceedings of the 21st annual ACM symposium on User interface software and technology*, 2008, pp. 91–100.

[16] C. J. Cai, A. Ren, and R. C. Miller, "Waitsuite: Productive use of diverse waiting moments," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 24, no. 1, pp. 1–41, 2017.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," *arXiv preprint arXiv:1811.06521*, 2018.

[23] 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.

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

[25] A. L. Thomaz, C. Breazeal *et al.*, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *Aaai*, vol. 6. Boston, MA, 2006, pp. 1000–1005.

VI. Appendix

See Figure 2 on the next page; this figure presents an overview of the design space, and showcases six select works from human-robot teaching and learning.

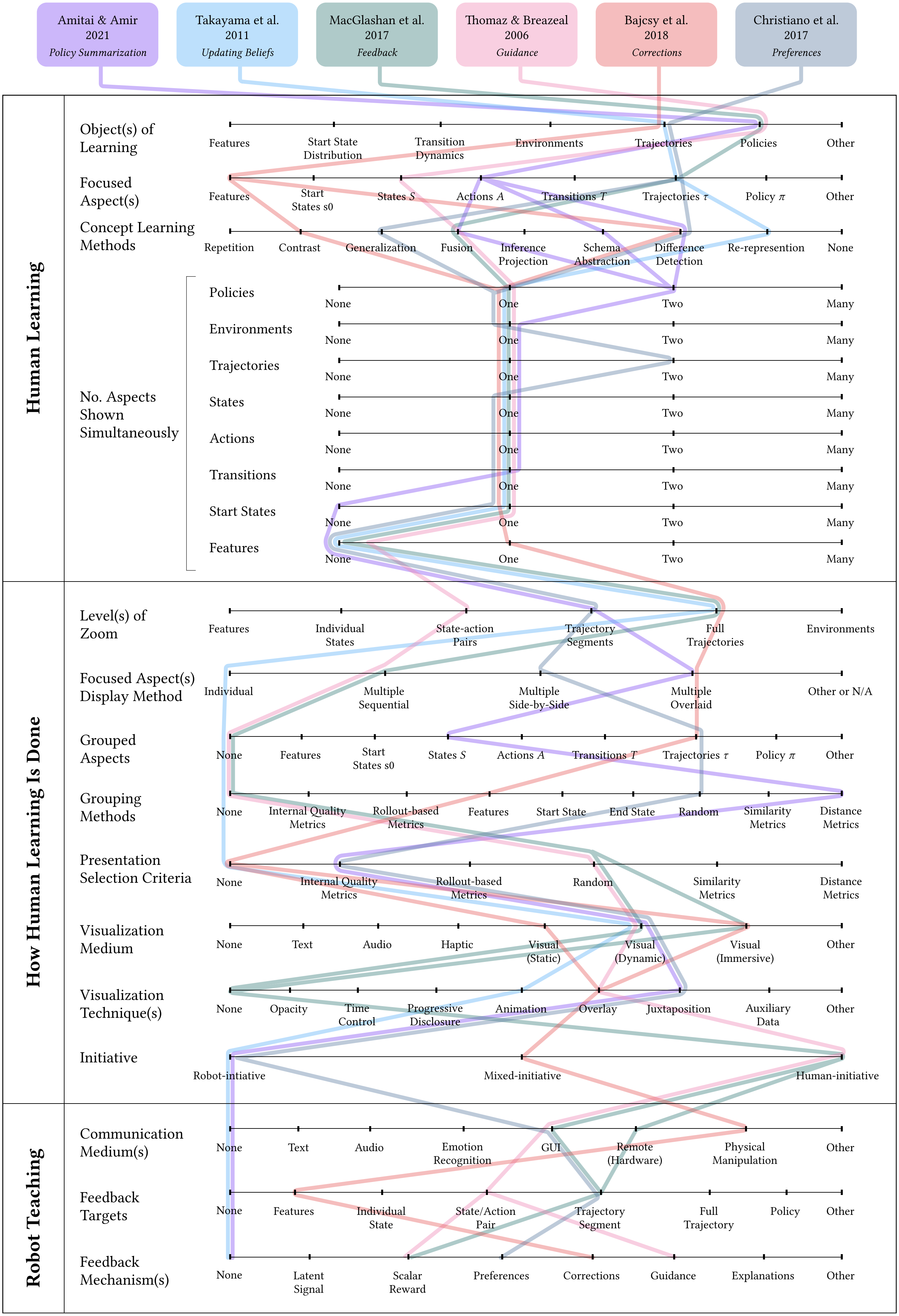

Fig. 2. An overview of the design space, showcasing six select works from human-robot teaching and learning. Amitai and Amir [24]; Takayama et al. [19]; MacGlashan et al. [3]; Thomaz and Breazeal [25]; Bajcsy et al. [23]; and, Christiano et al. [21].