Aligning Robot Representations with Humans

Andreea Bobu[*]
UC Berkeley
abobu@berkeley.edu

Andi Peng[*]
MIT CSAIL
andipeng@mit.edu

Pulkit Agrawal
MIT CSAIL
pulkitag@mit.edu

Julie Shah
MIT CSAIL
julie_a_shah@mit.edu

Anca Dragan
UC Berkeley
anca@berkeley.edu

Abstract

As robots are increasingly deployed in real-world environments, a key question becomes how to best teach them to accomplish tasks that humans want. In this work, we argue that current robot learning approaches suffer from *representation misalignment*, where the robot's learned task representation does not capture the human's true representation. We propose that because humans will be the ultimate evaluator of task performance in the world, it is crucial that we *explicitly* focus our efforts on aligning robot representations with humans, in addition to learning the downstream task. We advocate that current representation learning approaches in robotics can be studied under a single unifying formalism: *the representation alignment problem*. We mathematically operationalize this problem, define its key desiderata, and situate current robot learning methods within this formalism.

1 Introduction

In the robot learning community, we aspire to build robots that learn to accomplish tasks that human users want. We do this via two main paradigms: in imitation learning, we generate demonstrations of us completing the task for the robot to learn the human's *policy* [91, 132]; in reward learning, we instead supply task data for learning the human's *reward function* for the robot to optimize [61, 148]. Unfortunately, both methods often fail to accurately capture the human's *task representation*, or abstraction of what matters for solving the task, as learning from demonstrations is prone to capturing *spurious correlations* from a budgeted, non-representative sample of human task data [91]. This problem can be alleviated in two ways: we can increase the diversity of human task data to *implicitly* provide information about how to disentangle these different aspects, or we can *explicitly* build structure into our representations such that we extract all the relevant aspects from existing task data.

Unlike robots, human are remarkably adept at constructing representations that include only the relevant aspects for solving the task [71, 2, 72]. The failures above – where the robot's learned representation latches onto irrelevant aspects of the task – are only failures *because there is a mismatch between the human's representation and the robot's*. In other words, the human's representation is *misaligned* with the one learned by the robot. If we want to learn the human's true policy or reward for a task – if we want robots to know what *matters* to humans – we must *explicitly* align learned robot representations with humans'. We pose this as the *representation alignment problem*.

We propose a unifying mathematical formalism for tackling this problem and arrive at the following key takeaway: a better structured representation affords better alignment and therefore better task performance, but always with the unavoidable tradeoff of more human effort. Depending on the

[*]Equal contribution

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

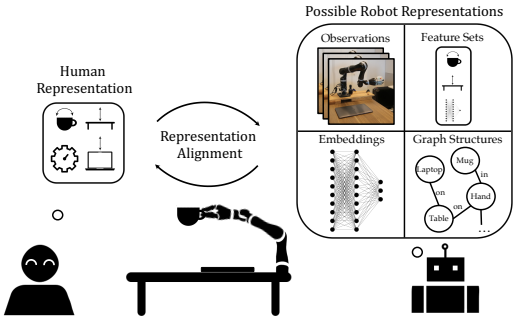

Figure 1: We instantiate the representation alignment problem as the search for a robot task representation that is easily able to capture the true human task representation.

representation, this human effort can be directed in the following three ways: 1) representations that operate directly on the observation space, e.g. end-to-end neural networks, direct human effort at increasing task data to avoid spurious correlations; 2) representations that build explicit task structure, e.g. graphs or feature sets, direct human effort at constructing and expanding the representation; and 3) representations that learn directly from implicit human representations, e.g. self-supervised models, direct human effort at creating good proxy tasks. Situated within our formalism, this provides a unifying lens to think about design tradeoffs that current and future robot learning approaches must make when faced with the challenge of aligning robot representations with humans.

2 Problem Formulation

Setup. We consider scenarios where a robot R seeks to learn how to perform a task desired by a human H. The human and the robot live in the same state $s \in \mathcal{S}$, where they can perform actions $a_H \in \mathcal{A}_H$ and $a_R \in \mathcal{A}_R$, respectively. The robot's goal is to learn a task expressed via a *reward function* $r^* : \mathcal{S} \mapsto \mathbb{R}$ capturing the human's preference ordering over states. The human knows the task they want, and, thus, implicitly know this reward function and how to act accordingly via a *policy* $\pi^*(a_H \mid s) \in [0, 1]$, but the robot does not.

We will consider two robot learning approaches of interest: *imitation learning*, where we seek to learn the human's policy for solving the task, and *reward learning*, where we seek to learn the reward function describing the task. Both approaches have received extensive interest, and have different benefits and downsides: imitation learning does not require modeling the human and simply replicates their actions [117, 1], but in doing so it also replicates their suboptimality and can't generalize well to changing dynamics or state distributions [91, 149]; meanwhile, reward learning attempts to capture *why* a specific behaviour is desirable and, thus, can generalize better to novel scenarios [1] but requires assuming a human model and large amounts of data [49, 129].

Partial Observability and Task Representations. Before diving into the details of imitation and reward learning, it's important to think about how to represent the state s that both the policy and the reward rely on. In theory, we could imagine the state to comprehensively capture the "true" components of the world down to their atomic elements, but in practice such a hypothetical state is neither fully observable nor usable by either actor. Hence, in this paper we assume that neither agent has full information of the state and, instead, they each *observe* it via observations $o_H \in \mathcal{O}_H$ and $o_R \in \mathcal{O}_R$. In the case of the robot, the observations o_R come from its (possibly noisy and non-deterministic) sensors $P(o_R \mid s)$ and may take the form of robot joint angles, RGB-D images from a camera, object poses and bounding boxes, etc. The human also senses observations o_H via their own "sensors" (e.g. the retinal visual inputs, audio signals, etc.) which we could model according to $P(o_H \mid s)$. Because of partial observability, both the robot and the human use the *history* of t observations $\mathbf{o}_R = (o_R^1, ..., o_R^t) \in \mathcal{O}_R^t$ and $\mathbf{o}_H = (o_H^1, ..., o_H^t) \in \mathcal{O}_H^t$, respectively, as a proxy for the state – or sequence of states – they observe $\mathbf{s} = (s^1, ..., s^t) \in \mathcal{S}^t$. We assume throughout the paper that $\mathbf{o}_R$ and $\mathbf{o}_H$ correspond to the same $\mathbf{s}$.

Neuroscience and cognitive psychology literature suggest that humans don't estimate the state directly from the complete $\mathbf{o}_H$ [16]. Instead, people focus on what's important for the task at hand, often

ignoring task-irrelevant attributes [26], and build a task-relevant *representation* to help them solve the task [22]. We, thus, assume that when humans think about how to complete or evaluate a task, they operate on a representation $\phi_H(\mathbf{o}_H)$ given by the transformation $\phi_H : \mathcal{O}_H^t \mapsto \Phi_H$, which determines which information in $\mathbf{o}_H$ to focus on or filter out and how to combine it into something useful for the task. For example, to determine if two novel objects have the same shape, a human might look around both of them (gather a sequence of visual information $\mathbf{o}_H$) to build an approximate 3D model (representation $\phi_H(\mathbf{o}_H)$) in order to answer the query. Intuitively, we can think of such a representation as an estimate of the task-relevant components of the state, in lieu of the true unknown state. We can, thus, model the human as approximating their preference ordering r^* with a reward function $r_H : \Phi_H \mapsto \mathbb{R}$, and their policy mapping π^* with $\pi_H(a_H \mid \phi_H(\mathbf{o}_H)) \in [0, 1]$.

The robot could similarly hold representations $\phi_R(\mathbf{o}_R)$ given by a transformation $\phi_R : \mathcal{O}_R^t \mapsto \Phi_R$. The most general ϕ_R is the identity function, where the robot uses the raw observations directly, but as we will see in Sec. A.4, other representations that are more structured are possible. For example, representations can be instantiated as handcrafted feature sets, where the designer distills their prior knowledge by manually pre-defining a set of representative aspects of the task [12, 110, 61], or as neural network embeddings, where the network attempts to implicitly extract such prior knowledge from data demonstrating how to do the task [46, 138, 160].

Imitation Learning. In imitation learning, the robot's goal is to learn a policy π_R that maps from its task representation (of possibly raw observations) to a distribution over actions $\pi_R(a_R \mid \phi_R(\mathbf{o}_R))$ telling it how to complete the task. To do so, the robot receives task *demonstrations* from the human and learns to imitate the exact actions that they take at every state [117, 149]. Let the demonstration that the human produces be a state trajectory $\xi = (s^0, \ldots, s^T)$ of length T. Importantly, both the human and the robot perceive this trajectory differently: the human observes $\xi_H = (o_H^0, \ldots, o_H^T)$ and the robot $\xi_R = (o_R^0, \ldots, o_R^T)$. Because the demonstrator is assumed to produce trajectories with high cumulative reward $r_H(\phi_H(\xi_H))$, i.e. be an expert at accomplishing the task, the intuition is that directly imitating their actions should result in good behaviour without the need to know the reward.

The issue with this approach is that the human's policy $\pi_H(a_H \mid \phi_H(\mathbf{o}_H))$ produces actions based on $\phi_H(\mathbf{o}_H)$, whereas the robot's actions are based on $\phi_R(\mathbf{o}_R)$. Thus, by directly imitating the human, the method implicitly assumes that $\phi_H(\mathbf{o}_H)$ is accurately captured by – or easily recoverable from – whatever $\phi_R(\mathbf{o}_R)$ was chosen to be. In other words, it assumes the robot and human's representations of what matters for the task are *aligned*. If this assumption does not hold, the robot might not be able to recover the right policy, and, thus, execute the right actions at the right state.

Reward Learning. Meanwhile, in our second problem of interest, the robot's goal is to recover a parameterized estimate of the human's reward function $r_\theta : \Phi_R \mapsto \mathbb{R}$, whether that's from demonstrations [170, 46, 50], corrections [12], teleoperation [81], comparisons [31], trajectory rankings [24], coactive feedback [80] etc. The intuition here is that the human's input can be interpreted as evidence for their internal reward function r_H, and the robot can use this evidence to find its own approximation of their reward r_θ. Given a learned r_θ, the robot can find an optimal policy π_R by maximizing the expected total reward $\mathbb{E}_{\pi_R}[\sum_{t=0}^{\infty} r_\theta(\phi_R(\mathbf{o}_R))]$.

Similar to imitation, because the human internally evaluates the reward function r_H based on $\phi_H(\mathbf{o}_H)$, their input is also based on $\phi_H(\mathbf{o}_H)$, whereas the robot interprets it as if it were based on $\phi_R(\mathbf{o}_R)$. Hence, if the two representations $\phi_R(\mathbf{o}_R)$ and $\phi_H(\mathbf{o}_H)$ are *misaligned*, the robot may recover the wrong reward function, and, thus, produce the wrong behaviour when optimizing it [17, 48].

The Problem of Misaligned Representations. In this paper, we argue that:

> *In real-world scenarios, it's infeasible to assume that robot and human representations will naturally align.*

We can see this in our earlier examples of robot representations $\phi_R(\mathbf{o}_R)$. First, the identity "representation" which maps $\mathbf{o}_R$ onto itself should, in theory, be able to capture everything in $\phi_H(\mathbf{o}_H)$ as long as $\mathbf{o}_R$ has enough information, but the high-dimensionality of the space $\mathcal{O}_R^t$ makes this representation challenging to apply in the real world: directly learning a downstream reward or policy that is robust across the state space and generalizes across environments would require a massive amount of diverse data – an expensive ask when working with human data [129, 49]. Meanwhile, while the example of sets of feature functions is lower dimensional, specifying all features that may be important to the human a priori is unrealistic, inevitably leading to incomplete representations $\phi_R(\mathbf{o}_R)$ that fail to

capture aspects in $\phi_H(\mathbf{o}_H)$ [17]. Lastly, in the neural network embedding example, learning a full mapping from the history $\mathbf{o}_R$ to the representation $\phi_R(\mathbf{o}_R)$ that robustly and generalizably covers all observations $\mathbf{o}_R$ (and, thus, $\mathbf{o}_H$) requires an extremely high amount of highly diverse data, similar to how reward and policy learning from the identity representation would.

To summarize, whether it's insufficient prior knowledge of what matters for the task or insufficient resources for exhaustively demonstrating the task, the robot's representation will more often than not be misaligned with the human's.

3 A Formalism for the Representation Alignment Problem in Robotics

How can we mathematically operationalize representation alignment[2]? While it is impossible for the robot and the human to perceive the world in the same way via observations $\mathbf{o}_R$ and $\mathbf{o}_H$, we would like for them to *make sense of their observations in a similar way*. We formalize the *representation alignment problem* as the search for a robot representation that is similar to the human's representation:

$$\phi_R^* = \arg\max_{\phi_R} \psi(\phi_R, \phi_H), \tag{1}$$

where ψ is a function that measures the similarity – or alignment – between two representation functions. The key question is: how exactly should we measure representation alignment, i.e. what should ψ be? We find the following ψ for measuring alignment:

$$\psi(\phi_R, \phi_H) = -\min_F \sum_{\mathbf{s}\in\mathcal{S}^t} \|F^T\phi_R(\mathbf{o}_R) - \phi_H(\mathbf{o}_H)\|_2^2 - \lambda \cdot dim(\Phi_R) \ , \tag{2}$$

D1: Recover the Human's Representation. For the robot's representation to capture *all* the relevant task aspects, we intuitively want alignment to be high when the human's representation $\phi_H(\mathbf{o}_H)$ can be *recovered* from the robot's $\phi_R(\mathbf{o}_R)$, no matter the state(s) $\mathbf{s}$ being observed by $\mathbf{o}_R$ and $\mathbf{o}_H$. Mathematically, we define "recovering" the human's representation from the robot's as a function $f : \Phi_R \mapsto \Phi_H$ mapping from the robot's representation to the human's, where $\phi_H(\mathbf{o}_H)$ is recoverable from $\phi_R(\mathbf{o}_R)$ if $f(\phi_R(\mathbf{o}_R)) \approx \phi_H(\mathbf{o}_H), \forall \mathbf{s}$. In other words, we can express the recovery error via an L_2 distance summed across all state sequences $\mathbf{s}$ with corresponding $\mathbf{o}_R$ and $\mathbf{o}_H$: $\sum_{\mathbf{s}\in\mathcal{S}^t} \|f(\phi_R(\mathbf{o}_R)) - \phi_H(\mathbf{o}_H)\|_2^2$. In equation 2, we want functions ϕ_R that have high alignment ψ with ϕ_H to have low error, hence we use the negative best distance as a measure of similarity.

D2: Avoid Spurious Correlations. We don't just want $\phi_R(\mathbf{o}_R)$ to recover $\phi_H(\mathbf{o}_H)$, i.e. be sufficient, but we want it to also be *minimal* to avoid spurious correlations that reflect irrelevant aspects of the task. We formalize this with a penalty on the size of the robot representation function's co-domain Φ_R. Together, in Equation 2 **D1** and **D2** describe a measure of representation alignment that rewards small representations that can be mapped close to $\phi_H(\mathbf{o}_H)$, where λ trades-off the two conditions and has to be designer-specified.

D3: Easily Recover the Human's Representation. Good representations $\phi_R(\mathbf{o}_R)$ enable *easy* recovery of human representations $\phi_H(\mathbf{o}_H)$. Looking at Equation 2, it's clear that finding an optimal solution via typical optimization is intractable given the arbitrarily large space of functions f. In theory, if we assume that the human's ϕ_H can be queried, the most straightforward way to optimize Equation 2 is to collect feedback $\langle \mathbf{o}_R, \phi_H(\mathbf{o}_H)\rangle$, where robot observation histories are labeled with the corresponding human representation mapping, and fit an approximation $\hat{f}$. Unfortunately, even if $\phi_R(\mathbf{o}_R)$ is low-dimensional, fitting an arbitrarily complex $\hat{f}$ that reliably results in high alignment for all states could require a large amount of diverse data from the human, i.e. it would not be *easy* to recover the human's representation. For this reason, we say that "easily" recoverable means that the transformation f is of relatively simple complexity. In practice recent work argues that linear transformations are a good proxy for small complexity [33, 85, 131, 5]. As such, we assume f to be a linear transformation given by a matrix F, but we note that we only need f to be low complexity.

D4: Explain the Robot's Representation. Lastly, human-aligned representations should be amenable to interpretability and explainability tools. If the human representation is easily recoverable, i.e. an estimate $\hat{f}$ can be easily learned from human feedback, then we get this condition almost for

[2]We assume the single human, single task case for ease of exposition. For extensions to the multiple humans or tasks, see App. A.3.

free. Since the robot has a good $\hat{f}$ estimate, it can communicate its representation to the human by showing them examples $\langle \mathbf{o}_H, \hat{f}(\phi_R(\mathbf{o}_R)) \rangle$ where observation sequences are labeled with the robot's current "translation" of its representation. The last piece we need for explainability, thus, is ensuring that $\hat{f}$ is understandable by the human, by, for example, having additional tools that can convert $\hat{f}$ into more human-interpretable interfaces, like language or visualizations.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Machine Learning (ICML), International Conference on*. ACM, 2004.

[2] David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 10–19. PMLR, 2018.

[3] David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34:7799–7812, 2021.

[4] Pulkit Agrawal. The task specification problem. In *Conference on Robot Learning*, pages 1745–1751. PMLR, 2022.

[5] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[6] Alnour Alharin, Thanh-Nam Doan, and Mina Sartipi. Reinforcement learning interpretation methods: A survey. *IEEE Access*, 8:171058–171077, 2020.

[7] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[8] Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[9] Sanjeev Arora, Simon S. Du, Sham Kakade, Yuping Luo, and Nikunj Saunshi. Provable representation learning for imitation learning via bi-level optimization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[10] Tsz-Chiu Au, Okhtay Ilghami, Ugur Kuter, J. William Murdock, Dana S. Nau, Dan Wu, and Fusun Yaman. SHOP2: an HTN planning system. *CoRR*, abs/1106.4869, 2011.

[11] Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2935–2945, Red Hook, NY, USA, 2018. Curran Associates Inc.

[12] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning robot objectives from physical human interaction. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 217–226. PMLR, 13–15 Nov 2017.

[13] Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (VPT): learning to act by watching unlabeled online videos. *CoRR*, abs/2206.11795, 2022.

[14] Chris Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 01 2007.

[15] Tom Bewley and Jonathan Lawry. Tripletree: A versatile interpretable representation of black box agents and their environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11415–11422, 2021.

[16] Daniel Birman and Justin L. Gardner. A flexible readout mechanism of human sensory representations. *Nature Communications*, 10(1):3500, 2019.

[17] A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan. Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *IEEE Transactions on Robotics*, pages 1–20, 2020.

[18] Andreea Bobu, Andrea Bajcsy, Jaime F. Fisac, and Anca D. Dragan. Learning under misspecified objective spaces. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 796–805. PMLR, 29–31 Oct 2018.

[19] Andreea Bobu, Chris Paxton, Wei Yang, Balakumar Sundaralingam, Yu-Wei Chao, Maya Cakmak, and Dieter Fox. Learning perceptual concepts by bootstrapping from human queries, 2021.

[20] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D. Dragan. Inducing structure in reward learning by learning features. *The International Journal of Robotics Research*, 0(0):02783649221078031, 0.

[21] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D. Dragan. Feature expansive reward learning: Rethinking human input. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, page 216–224, New York, NY, USA, 2021. Association for Computing Machinery.

[22] Tyler Bonnen, Daniel L.K. Yamins, and Anthony D. Wagner. When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, 109(17):2755–2766.e6, 2021.

[23] Cynthia Breazeal, Jesse Gray, and Matt Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *The International Journal of Robotics Research*, 28(5):656–680, 2009.

[24] Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast Bayesian reward inference from preferences. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1165–1177. PMLR, 13–18 Jul 2020.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[26] Frederick Callaway, Antonio Rangel, and Thomas L. Griffiths. Fixation patterns in simple choice reflect optimal information sampling. *PLOS Computational Biology*, 17(3):1–29, 03 2021.

[27] Kevin Chen, Nithin Shrivatsav Srikanth, David Kent, Harish Ravichandar, and Sonia Chernova. Learning hierarchical task networks with preferences from unannotated demonstrations. In Jens Kober, Fabio Ramos, and Claire J. Tomlin, editors, *4th Conference on Robot Learning, CoRL 2020, 16-18 November 2020, Virtual Event / Cambridge, MA, USA*, volume 155 of *Proceedings of Machine Learning Research*, pages 1572–1581. PMLR, 2020.

[28] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[29] Xin Chen, Sam Toyer, Cody Wild, Scott Emmons, Ian Fischer, Kuang-Huei Lee, Neel Alex, Steven H. Wang, Ping Luo, Stuart Russell, Pieter Abbeel, and Rohin Shah. An empirical investigation of representation learning for imitation. *CoRR*, abs/2205.07886, 2022.

[30] Jaedeug Choi and Kee-Eung Kim. Bayesian nonparametric feature construction for inverse reinforcement learning. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[31] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[32] Michael Jae-Yoon Chung, Abram Friesen, Dieter Fox, Andrew Meltzoff, and Rajesh Rao. A bayesian developmental approach to robotic goal-based imitation learning. *PloS one*, 10:e0141965, 11 2015.

[33] Adam Coates and A. Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade*, 2012.

[34] Anthony C. Constantinou. Learning bayesian networks with the saiyan algorithm. *ACM Trans. Knowl. Discov. Data*, 14(4):44:1–44:21, 2020.

[35] Angel Daruna, Mehul Gupta, Mohan Sridharan, and Sonia Chernova. Continual learning of knowledge graph embeddings. *IEEE Robotics and Automation Letters*, 6(2):1128–1135, 2021.

[36] Angel Andres Daruna, Devleena Das, and Sonia Chernova. Explainable knowledge graph embedding: Inference reconciliation for knowledge inferences supporting robot actions. *CoRR*, abs/2205.01836, 2022.

[37] Angel Andres Daruna, Lakshmi Nair, Weiyu Liu, and Sonia Chernova. Towards robust one-shot task execution using knowledge graph embeddings. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pages 11118–11124. IEEE, 2021.

[38] Devleena Das and Sonia Chernova. Semantic-based explainable ai: Leveraging semantic scene graphs and pairwise ranking to explain robot failures. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3034–3041. IEEE, 2021.

[39] Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[40] Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11693–11704, 2019.

[41] Anthony M. Dearden and Yiannis Demiris. Learning forward models for robots. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, pages 1440–1445. Professional Book Center, 2005.

[42] Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. *CoRR*, abs/2202.13903, 2022.

[43] Simon S. Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably, 2020.

[44] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[45] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1126–1135. JMLR.org, 2017.

[46] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 49–58. JMLR.org, 2016.

[47] Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Learning visual feature spaces for robotic manipulation with deep spatial autoencoders. *CoRR*, abs/1509.06113, 2015.

[48] David Fridovich-Keil, Andrea Bajcsy, Jaime F. Fisac, Sylvia L. Herbert, Steven Wang, Anca D. Dragan, and Claire J. Tomlin. Confidence-aware motion prediction for real-time collision avoidance. *International Journal of Robotics Research*, 2019.

[49] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[50] Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational inverse control with events: A general framework for data-driven reward definition. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8547–8556, Red Hook, NY, USA, 2018. Curran Associates Inc.

[51] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[52] Ilche Georgievski and Marco Aiello. HTN planning: Overview, comparison, and beyond. *Artif. Intell.*, 222:124–156, 2015.

[53] Dibya Ghosh, Abhishek Gupta, and Sergey Levine. Learning actionable representations with goal conditioned policies. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[54] Claire Glanois, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu. A survey on interpretable reinforcement learning. *arXiv preprint arXiv:2112.13112*, 2021.

[55] Adam Gleave and Oliver Habryka. Multi-task maximum entropy inverse reinforcement learning. *arXiv preprint arXiv:1805.08882*, 2018.

[56] Anna Goldenberg and Andrew W. Moore. Tractable learning of large bayes net structures from sparse data. In Carla E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

[57] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *International conference on machine learning*, pages 1792–1801. PMLR, 2018.

[58] Martin Günther, J. R. Ruiz-Sarmiento, Cipriano Galindo, Javier González Jiménez, and Joachim Hertzberg. Context-aware 3d object anchoring for mobile robots. *Robotics Auton. Syst.*, 110:12–32, 2018.

[59] Piyush Gupta, Nikaash Puri, Sukriti Verma, Sameer Singh, Dhruv Kayastha, Shripad Deshmukh, and Balaji Krishnamurthy. Explain your move: Understanding agent actions using focused feature saliency. *arXiv preprint arXiv:1912.12191*, 2019.

[60] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[61] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[62] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in neural information processing systems*, 30, 2017.

[63] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[64] Bradley Hayes and Brian Scassellati. Autonomously constructing hierarchical task networks for planning and human-robot collaboration. In Danica Kragic, Antonio Bicchi, and Alessandro De Luca, editors, *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 5469–5476. IEEE, 2016.

[65] Bradley Hayes and Julie A Shah. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 303–312. IEEE, 2017.

[66] Laura Hiatt, Cody Narber, Esube Bekele, Sangeet Khemlani, and J Trafton. Human modeling for human–robot collaboration. *The International Journal of Robotics Research*, 36:027836491769059, 02 2017.

[67] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[68] Irina Higgins, Arka Pal, Andrei A. Rusu, Loïc Matthey, Christopher P. Burgess, Alexander Pritzel, Matthew M. Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *ICML*, 2017.

[69] Sophie Hilgard, Nir Rosenfeld, Mahzarin R. Banaji, Jack Cao, and David C. Parkes. Learning representations by humans, for humans. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4227–4238. PMLR, 2021.

[70] Mark K Ho. The value of abstraction. *Current opinion in behavioral sciences*, 29, 2019.

[71] Mark K Ho, David Abel, Carlos G Correa, Michael L Littman, Jonathan D Cohen, and Thomas L Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[72] Mark K Ho and Thomas L Griffiths. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:33–53, 2022.

[73] Yordan Hristov, Daniel Angelov, Michael Burke, Alex Lascarides, and Subramanian Ramamoorthy. Disentangled relational representations for explaining and learning from demonstration. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pages 870–884. PMLR, 2019.

[74] Chao Huang, Wenhao Luo, and Rui Liu. Meta preference learning for fast user adaptation in human-supervisory multi-robot deployments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5851–5856. IEEE, 2021.

[75] Marcus Hutter. Feature dynamic bayesian networks. *CoRR*, abs/0812.4581, 2008.

[76] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8011–8023. Curran Associates, Inc., 2018.

[77] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[78] Tetsunari Inamura, Masayuki Inaba, and Hirochika Inoue. User adaptation of human-robot interaction model based on bayesian network and introspection of interaction experience. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2000, October 30 - Novemver 5, 2000, Takamatsu, Japan*, pages 2139–2144. IEEE, 2000.

[79] Guillaume Infantes, Malik Ghallab, and Félix Ingrand. Learning the behavior model of a robot. *Auton. Robots*, 30(2):157–177, 2011.

[80] Ashesh Jain, Shikhar Sharma, Thorsten Joachims, and Ashutosh Saxena. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research*, 34(10):1296–1313, 2015.

[81] Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research*, 37(7):717–742, 2018.

[82] E. T. Jaynes. Information theory and statistical mechanics. volume 106, pages 620–630. American Physical Society, May 1957.

[83] Nathan P. Koenig and Maja J. Mataric. Robot life-long task learning from human demonstrations: a bayesian approach. *Auton. Robots*, 41(5):1173–1188, 2017.

[84] Minae Kwon, Sandy H Huang, and Anca D Dragan. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95, 2018.

[85] Cheng-I Lai. Contrastive predictive coding based feature for automatic speaker verification. *arXiv preprint arXiv:1904.01575*, 2019.

[86] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc.

[87] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 13–18 Jul 2020.

[88] Elena Lazkano, Basilio Sierra, Aitzol Astigarraga, and José María Martínez-Otzeta. On the use of bayesian networks to develop behaviours for mobile robots. *Robotics Auton. Syst.*, 55(3):253–265, 2007.

[89] Kimin Lee, Laura M. Smith, and Pieter Abbeel. PEBBLE: feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6152–6163. PMLR, 2021.

[90] Timothée Lesort, Natalia Díaz-Rodríguez, Jean-Franois Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018.

[91] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[92] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Feature construction for inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2010.

[93] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 19–27, 2011.

[94] Nan Li, William Cushing, Subbarao Kambhampati, and Sung Wook Yoon. Learning probabilistic hierarchical task networks as probabilistic context-free grammars to capture user preferences. *ACM Trans. Intell. Syst. Technol.*, 5(2):29:1–29:32, 2014.

[95] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[96] Long-Ji Lin. Hierarchical learning of robot skills by reinforcement. In *Proceedings of International Conference on Neural Networks (ICNN'88), San Francisco, CA, USA, March 28 - April 1, 1993*, pages 181–186. IEEE, 1993.

[97] Weiyu Liu. A survey of semantic reasoning frameworks for robotic systems. 2022.

[98] Manuel Lopes, Francisco S. Melo, and Luis Montesano. Affordance-based imitation learning in robots. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 29 - November 2, 2007, Sheraton Hotel and Marina, San Diego, California, USA*, pages 1015–1021. IEEE, 2007.

[99] Dylan P. Losey and Marcia Kilchenman O'Malley. Including uncertainty when learning from human corrections. In *CoRL*, 2018.

[100] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2016.

[101] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pages 1113–1132. PMLR, 2019.

[102] Ashique Rupam Mahmood. Structure learning of causal bayesian networks: A survey. 2011.

[103] Zhao Mandi, Pieter Abbeel, and Stephen James. On the effectiveness of fine-tuning versus meta-reinforcement learning. *arXiv preprint arXiv:2206.03271*, 2022.

[104] Neville Mehta, Soumya Ray, Prasad Tadepalli, and Thomas G. Dietterich. Automatic discovery and transfer of MAXQ hierarchies. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 648–655. ACM, 2008.

[105] Ishai Menache, Shie Mannor, and Nahum Shimkin. Q-cut - dynamic discovery of sub-goals in reinforcement learning. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002, 13th European Conference on Machine Learning, Helsinki, Finland, August 19-23, 2002, Proceedings*, volume 2430 of *Lecture Notes in Computer Science*, pages 295–306. Springer, 2002.

[106] Anahita Mohseni-Kabir, Changshuo Li, Victoria Wu, Daniel Miller, Benjamin Hylak, Sonia Chernova, Dmitry Berenson, Candace Sidner, and Charles Rich. Simultaneous learning of hierarchy and primitives for complex robot tasks. *Autonomous Robots*, 43(4):859–874, 2019.

[107] Anahita Mohseni-Kabir, Charles Rich, Sonia Chernova, Candace L. Sidner, and Daniel Miller. Interactive hierarchical task learning from a single demonstration. In Julie A. Adams, William D. Smart, Bilge Mutlu, and Leila Takayama, editors, *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2015, Portland, OR, USA, March 2-5, 2015*, pages 205–212. ACM, 2015.

[108] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Modeling affordances using bayesian networks. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 29 - November 2, 2007, Sheraton Hotel and Marina, San Diego, California, USA*, pages 4102–4107. IEEE, 2007.

[109] Negin Nejati, Pat Langley, and Tolga Könik. Learning hierarchical task networks by observation. In William W. Cohen and Andrew W. Moore, editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 665–672. ACM, 2006.

[110] Andrew Ng and Stuart Russell. Algorithms for inverse reinforcement learning. *International Conference on Machine Learning (ICML)*, 0:663–670, 2000.

[111] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

[112] Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, page 33–40. IEEE Press, 2013.

[113] Kentaro Nishi and Masamichi Shimosaka. Fine-grained driving behavior prediction via context-aware multi-task inverse reinforcement learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2281–2287. IEEE, 2020.

[114] Guanglin Niu, Bo Li, Yongfei Zhang, and Shiliang Pu. Enginekgi: Closed-loop knowledge graph inference. *arXiv preprint arXiv:2112.01040*, 2021.

[115] Daniel Nyga, Subhro Roy, Rohan Paul, Daehyung Park, Mihai Pomarlan, Michael Beetz, and Nicholas Roy. Grounding robot plans from natural language instructions with incomplete world knowledge. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, pages 714–723. PMLR, 2018.

[116] Oliver Obst. Using a planner for coordination of multiagent team behavior. In Rafael H. Bordini, Mehdi Dastani, Jürgen Dix, and Amal El Fallah Seghrouchni, editors, *Programming Multi-Agent Systems, Third International Workshop, ProMAS 2005, Utrecht, The Netherlands, July 26, 2005, Revised and Invited Papers*, volume 3862 of *Lecture Notes in Computer Science*, pages 90–100. Springer, 2005.

[117] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179, 2018.

[118] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Fred Shentu, Evan Shelhamer, Jitendra Malik, Alexei A. Efros, and Trevor Darrell. Zero-shot visual imitation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2131–21313, 2018.

[119] Abhishek Paudel. Learning for robot decision making under distribution shift: A survey. *arXiv preprint arXiv:2203.07558*, 2022.

[120] Chris Paxton, Chris Xie, Tucker Hermans, and Dieter Fox. Predicting stable configurations for semantic placement of novel objects. In *Conference on Robot Learning (CoRL)*, 2021. to appear.

[121] Judea Pearl. Causal inference. In Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf, editors, *Causality: Objectives and Assessment (NIPS 2008 Workshop), Whistler, Canada, December 12, 2008*, volume 6 of *JMLR Proceedings*, pages 39–58. JMLR.org, 2010.

[122] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[123] Rouhollah Rahmatizadeh, Pooya Abolghasemi, Ladislau Bölöni, and Sergey Levine. Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 3758–3765. IEEE, 2018.

[124] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov, editors, *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018.

[125] Deepak Ramachandran and Rakesh Gupta. Smoothed sarsa: Reinforcement learning for robot delivery tasks. In *2009 IEEE International Conference on Robotics and Automation, ICRA 2009, Kobe, Japan, May 12-17, 2009*, pages 2125–2132. IEEE, 2009.

[126] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[127] Nathan Ratliff, David M Bradley, Joel Chestnutt, and J A Bagnell. Boosting structured prediction for imitation learning. In *Advances in Neural Information Processing Systems*, pages 1153–1160, 2007.

[128] Sid Reddy, Anca D. Dragan, and Sergey Levine. Pragmatic image compression for human-in-the-loop decision-making. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 26499–26510, 2021.

[129] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. SQIL: imitation learning via reinforcement learning with sparse rewards. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[130] Siddharth Reddy, Anca D. Dragan, Sergey Levine, Shane Legg, and Jan Leike. Learning human objectives by evaluating hypothetical behavior. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8020–8029. PMLR, 2020.

[131] C. J. Reed, X. Yue, A. Nrusimha, S. Ebrahimi, V. Vijaykumar, R. Mao, B. Li, S. Zhang, D. Guillory, S. Metzger, K. Keutzer, and T. Darrell. Self-supervised pretraining improves self-supervised pretraining. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1050–1060, Los Alamitos, CA, USA, jan 2022. IEEE Computer Society.

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

[133] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[134] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[135] Scott Sanner. Simultaneous learning of structure and value in relational reinforcement learning. In *Workshop on Rich Representations for Reinforcement Learning*, page 57. Citeseer, 2005.

[136] Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra Kumar Misra, and Hema Swetha Koppula. Robobrain: Large-scale knowledge engine for robots. *CoRR*, abs/1412.0691, 2014.

[137] Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R. Devon Hjelm, Philip Bachman, and Aaron C. Courville. Pretraining representations for data-efficient reinforcement learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12686–12699, 2021.

[138] Seyed Kamyar Seyed Ghasemipour, Shixiang Shane Gu, and Richard Zemel. Smile: Scalable meta inverse reinforcement learning through context-conditional policies. *Advances in Neural Information Processing Systems*, 32, 2019.

[139] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. The implicit preference information in an initial state. In *International Conference on Learning Representations*, 2019.

[140] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[141] Avi Singh, Eric Jang, Alexander Irpan, Daniel Kappler, Murtaza Dalal, Sergey Levine, Mohi Khansari, and Chelsea Finn. Scalable multi-task imitation learning with autonomous improvement. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 2167–2173. IEEE, 2020.

[142] Avi Singh, Larry Yang, Chelsea Finn, and Sergey Levine. End-to-end robotic reinforcement learning without reward engineering. In Antonio Bicchi, Hadas Kress-Gazit, and Seth Hutchinson, editors, *Robotics: Science and Systems XV, University of Freiburg, Freiburg im Breisgau, Germany, June 22-26, 2019*, 2019.

[143] Dan Song, Carl Henrik Ek, Kai Huebner, and Danica Kragic. Multivariate discretization for bayesian network structure learning in robot grasping. In *IEEE International Conference on Robotics and Automation, ICRA 2011, Shanghai, China, 9-13 May 2011*, pages 1944–1950. IEEE, 2011.

[144] Dan Song, Kai Huebner, Ville Kyrki, and Danica Kragic. Learning task constraints for robot grasping using graphical models. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan*, pages 1579–1585. IEEE, 2010.

[145] Arjun Sripathy, Andreea Bobu, Zhongyu Li, Koushil Sreenath, Daniel S. Brown, and Anca D. Dragan. Teaching robots to span the space of functional expressive motion, 2022.

[146] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 2021.

[147] Liting Sun, Xiaogang Jia, and Anca D. Dragan. On complementing end-to-end human behavior predictors with planning. *Robotics: Science and Systems XVII*, 2021.

[148] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[149] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4950–4957. AAAI Press, 2018.

[150] Mycal Tucker, Yilun Zhou, and Julie Shah. Latent space alignment using adversarially guided self-play. *International Journal of Human–Computer Interaction*, 0(0):1–19, 2022.

[151] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[152] Paul Vernaza and Drew Bagnell. Efficient high dimensional maximum entropy modeling via symmetric partition functions. In *Advances in Neural Information Processing Systems*, pages 575–583, 2012.

[153] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press Princeton, NJ, 1945.

[154] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

[155] Garrett Warnell, Nicholas R. Waytowich, Vernon J. Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. *ArXiv*, abs/1709.10163, 2018.

[156] Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[157] M. Wulfmeier, D. Z. Wang, and I. Posner. Watch this: Scalable cost-function learning for path planning in urban environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2089–2095, 2016.

[158] Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 285–294, New York, NY, USA, 2019. Association for Computing Machinery.

[159] Annie Xie, Avi Singh, Sergey Levine, and Chelsea Finn. Few-shot goal inference for visuomotor learning and planning. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, pages 40–52. PMLR, 2018.

[160] Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, and Chelsea Finn. Learning a prior over intent via meta-inverse reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6952–6962. PMLR, 09–15 Jun 2019.

[161] Jun Yamada, Karl Pertsch, Anisha Gunjal, and Joseph J. Lim. Task-induced representation learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[162] Mengjiao Yang and Ofir Nachum. Representation matters: Offline pretraining for sequential decision making. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11784–11794. PMLR, 2021.

[163] John Seon Keun Yi, Yoonwoo Kim, and Sonia Chernova. Incremental object grounding using scene graphs. *CoRR*, abs/2201.01901, 2022.

[164] Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. *Advances in Neural Information Processing Systems*, 32, 2019.

[165] Wentao Yuan, Chris Paxton, Karthik Desingh, and Dieter Fox. Sornet: Spatial object-centric representations for sequential manipulation. In *5th Annual Conference on Robot Learning*, pages 148–157. PMLR, 2021.

[166] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, volume 12368 of *Lecture Notes in Computer Science*, pages 606–623. Springer, 2020.

[167] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[168] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1–8. IEEE, 2018.

[169] Yichuan Zhang, Yixing Lan, Qiang Fang, Xin Xu, Junxiang Li, and Yujun Zeng. Efficient reinforcement learning from demonstration via bayesian network-based knowledge extraction. *Comput. Intell. Neurosci.*, 2021:7588221:1–7588221:16, 2021.

[170] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pages 1433–1438. AAAI Press, 2008.

[171] Matthew Zurek, Andreea Bobu, Daniel S. Brown, and Anca D. Dragan. Situational confidence assistance for lifelong shared autonomy. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2783–2789, 2021.

A Appendix

A.1 The Desired Representation

Value Alignment. First, learning human-aligned representations can help tackle the challenge of *value alignment* [7], enabling robots to perform well under the human's true desired objective, as opposed to optimizing mispecified objectives that lead to unintended side-effects. In "reward hacking" scenarios, the reward function may capture an ill-defined and sometimes misleading representation of the human's true intent [7]. Consider the canonical example of a robot that is tasked with sweeping dust off the floor [134]. An optimal policy for "maximize dust collected off the floor" leads to a robot that dumps dust onto the floor just to immediately sweep it up again. In this case, the reward function is defined on top of a representation that is *under-specified*, e.g.only the amount of dust that is collected, and fails to capture other important features, e.g. covering the whole house, not increasing the amount of dust on the floor, etc. Explicitly learning a representation that is aligned to the human's therefore ensures that all causal features for accomplishing the objective are captured.

Generalizable Task Learning. A human-aligned representation may afford more generalizable task learning [43, 9]. A central problem in robot learning remains our ability to learn a diverse set of behaviours across different environments and user preferences [91, 119]. While domains like language or vision have elicited impressive performance across many tasks by leveraging large-scale data collection and pre-training [126, 25, 122], robot learning remains bottlenecked by our ability to collect diverse data that captures the full complexity of the world. Hence, learning representations that fully describe all aspects relevant for completing the task, yet exclude potentially arbitrary perturbations from high-dimensional observation data is difficult. This is in large part due to neural networks learning non-causal correlates from task data, where many relevant features may be related [77, 39]. Thus, robot learning objectives that operate directly on high-dimensional observation spaces suffer from the problem of *spurious correlations*, where the implicit representations learned by the policy may contain irrelevant features that are not informative to accomplishing the task [4]. Consequently, the learned network may be based on these irrelevant features that appear causal in-distribution, but fail under distribution shift. Thus, explicitly aligning robot representations with those used by humans could afford more generalizable task learning that maintains robustness under distribution shift.

Reducing Human Burden. Learning methods that operate on human-aligned representations can reduce the teaching burden required from human users. Consider our above two scenarios, where human guidance is in the form of either task demonstrations or specified rewards. In both cases, if we possessed unlimited human time and effort, we would be able to provide a perfect task representation, i.e. a demonstration of the task in every environment for every user [44] or a reward function that comprehensively specifies every intended feature any user may find relevant for performing the task in any environment [62], and then fit the data with an arbitrarily complex function such as a neural network. In practice, both scenarios are intractable with low sample complexity, and therefore motivate the explicit need for representations that align on the task abstraction level [3, 70, 2].

Explainability. For eventually achieving robots that can be deployed in human environments, we desire representations that better enable system transparency and explainability. Ensuring that deep learning systems are interpretable to various concerns encompassing ethical, legal, safety, or usability viewpoints is a key focus of real-world deployments and a robust area of active research [54, 6]. Methods range from training generating post-hoc explanations via decision trees [54, 15], text-based descriptions of relational MDPs [65, 135], or Jacobian saliency maps [57] for explaining behaviour. However, it is important to consider system explainability from not only the perspective of usability and safety during deployment, but also as critically embedded within the design process itself [51, 59]. Moreover, not only are explainable representations necessary for safe real-world deployment, they additionally afford good human-AI teamwork [112, 23, 66]. Explicitly aligning representations with humans' *prior* to learning can create a more streamlined process for ensuring that the underlying features subsequently learned by the robot are primed for human understanding [133].

Desideratum 1: The representation should capture *all* the relevant aspects of the desired task, i.e. the human's true objective should be *realizable* when using the representation for task learning. This directly informs our ability to learn all the aspects that matter for producing generalizable task behaviour in novel scenarios, as well as ensures accurate optimization of the intended user objective.

Desideratum 2: The representation should not capture *irrelevant* aspects of the desired task, i.e. avoid spurious correlations. This can help alleviate learning feature correlates that can harm robust task performance under distribution shift, as well as avoid potentially unintended consequences.

Desideratum 3: Human guidance for learning the representation should require *minimal* time and effort, i.e. the human's representation should be *easily recoverable* from data. This ensures that budgeted user input can be fully leveraged for task learning.

Desideratum 4: The representation should enable *human interpretable* and *explainable* behaviour, affording safe, transparent systems that can fully integrate with human users in the real-world.

A.2 Examples of Robot Representations

Since solving Equation 1 is clearly intractable for an arbitrarily large set of functions ϕ_R, different ways of defining the robot's representation $\phi_R(\mathbf{o}_R)$ implicitly make different simplifying assumptions. When ϕ_R is the identity function, the underlying assumption is that there exists some $f : \mathcal{O}_R^t \mapsto \Phi_H$ that satisfies Equation 2 so long as $\mathbf{o}_R$ has enough information to capture $\phi_H(\mathbf{o}_H)$. Unfortunately, because f operates on an extremely large space of robot observation histories $\mathcal{O}_R^t$, it would have to be complex enough to reliably cover the space, violating **D3**. This, together with the large dimensionality of the representation space, result in a small alignment value in Equation 2. Meanwhile, methods that assume that $\phi_R(\mathbf{o}_R)$ has some more low-dimensional structure, like the feature sets or embeddings from earlier, could also have small alignment values: feature sets might be non-comprehensive, while learned feature embeddings might have not extracted what's truly important to the human, making it, thus, impossible to find an f that recovers $\phi_H(\mathbf{o}_H)$. As we will see in Section A.4, no representation is naturally human-aligned and every representation type comes with its trade-offs.

A.3 Extensions to Formalism

Extension to Multiple Tasks. Above, we considered the single task setting, where the robot's goal is to successfully perform one desired task. However, our formalism can be extended to account for multiple tasks. First, when the person wants to train the robot to correctly perform multiple tasks, the observation space $\mathcal{O}_R$ may be different for each task. In practice, these observation spaces are oftentimes the same or similar (e.g. multiple robot manipulation tasks can all still use images of the same tabletop as observations, although the observation distribution may differ if different objects are used). We can account for differing spaces by choosing the overall observation space $\mathcal{O}_R$ to be the union of all individual N task observation spaces $\mathcal{O}_{R_i}$: $\mathcal{O}_R = \mathcal{O}_{R_1} \bigcup ... \bigcup \mathcal{O}_{R_N}$. Additionally, in multi-task settings, the human representation $\phi_H(\mathbf{o}_H)$ will reflect aspects of the task *distribution* that matter to them, rather than of a single task. As a result, the robot's representation learning strategy should reflect this, as we will see in the survey portion of the paper.

Extension to Multiple Humans. Aligning the robot's representation to multiple humans requires acknowledging that each human may operate under a different observation space $\mathcal{O}_H$ or representation $\phi_H(\mathbf{o}_H)$. First, we could modify our formalism for differing spaces similarly to how we did in the multi-task setting, by choosing the overall observation space $\mathcal{O}_H$ to be the union of all individual M human observation spaces $\mathcal{O}_{H_i}$: $\mathcal{O}_H = \mathcal{O}_{H_1} \bigcup ... \bigcup \mathcal{O}_{H_M}$. Second, in such multi-agent settings, the robot could attempt to align its representation to a unified $\phi_H(\mathbf{o}_H) = \phi_{H_1}(\mathbf{o}_H) \bigcup ... \bigcup \phi_{H_M}(\mathbf{o}_H)$, individually to each $\phi_{H_i}(\mathbf{o}_H)$, or a combination of the two strategies where the unified representation is then specialized to each individual human's representation.

A.4 Survey of Robot Representations

We now present our survey of common representations in the robot learning literature. We situate them within our formalism and highlight their key tradeoffs. We primarily focus on overviewing *learned* representations and discuss four main categories: identity, learned feature sets, feature embeddings, and graph structures.

A.5 Identity Representation

As we alluded to in Sec. 3, an identity representation maps a history of observations onto itself, i.e. $\phi_R(\mathbf{o}_R) = \mathbf{o}_R$. In this case, the co-domain of the representation function is the space of histories

of observations: $\phi_R : \mathcal{O}^t \mapsto \mathcal{O}^t$. The methods we review here, thus, don't focus on learning some explicit intermediate representation that captures what matters for the task(s) and instead hope to implicitly extract what's important directly from human task data.

Because the inputs for reward or policy learning consist of raw (histories of) observations, which can be very high-dimensional, e.g. images, most of the approaches we cover here are based on high-capacity deep learning models. There are now numerous end-to-end methods for learning policies [117, 149, 129, 91] or reward functions [46, 49, 157] from demonstrations. While these methods perform very well by overparameterizing a high complexity function to overfit to the training tasks, they suffer from generalization failures due to *distribution shift* from the training to test distribution [132], resulting in behaviour that can be arbitrarily erroneous during deployment. Achieving good end-to-end performance to cover a large test distribution can require hundreds or even thousands of demonstrations for each desired task [168, 123, 124], which can be difficult and expensive to obtain in practice. In reward learning, this has been alleviated by introducing other simpler types of human reward feedback, like comparisons [31], numeric feedback [155], examples of what constitutes a goal [50], or a combination [76]. These are not only more user friendly alternatives to demonstrations, but they are also amenable to active learning techniques [130, 142], which can further reduce the human burden.

Another popular approach for reducing this sample complexity is meta-learning [45], which seeks to learn a representation that enables fast fine-tuning at test time [160, 164, 141, 74, 138, 159]. The intuition is that at training time, we can reuse human data from an array of different tasks, and if the training distribution is representative enough of the test task(s), this "warm-started" model can adapt to a new task with little data. Unfortunately, the human would have to know the test time task distribution *a priori*, which brings us back to the specification problem: we now trade hand-crafting representations for hand-crafting task distributions. Moreover, because deep learning methods do not explicitly learn a representation prior to task learning, they are inherently *uninterpretable* and difficult to debug in the event of a failure [133].

Takeaway. Despite the recent advances in higher performance from end-to-end learning systems, the identity representation, while easy to specify, is difficult to use in practice for robust and generalizable robot learning with minimal human input.

A.6 Feature Sets

We now discuss methods where the robot's representation $\phi_R(\mathbf{o}_R)$ is instantiated as a set $\{\phi_R^1(\mathbf{o}_R), ..., \phi_R^d(\mathbf{o}_R)\}$, where each member $\phi_R^i(\mathbf{o}_R)$ constitutes a different individual dimension of the robot's representation, with d much smaller than the dimensionality of $\mathcal{O}^t$. These dimensions are meant to represent concrete aspects of the task – or features, e.g. how far the robot's end effector is from the table, – which is why we call each ϕ_R^i a feature function and the output $\phi_R^i(\mathbf{o}_R)$ a feature value. In general, the feature function maps observation histories to a real number indicating how much that feature is expressed in the observations, $\phi_R^i : \mathcal{O}_R^t \mapsto \mathbb{R}$. Hence, under this instantiation, the robot's representation maps from observation histories onto a d-dimensional space of real values: $\phi_R : \mathcal{O}_R^t \mapsto \mathbb{R}^d$.

Handcrafted feature sets have been used to great success across robot policy and reward learning [1, 61, 80, 81, 139], but exhaustively specifying all aspects of the task the human will care about ahead of time is extremely difficult [18]. To address this, early reward and policy learning methods have looked at inferring a set of relevant feature functions directly from task demonstrations. **(author?)** [152] constrain the robot's representation to have a low-dimensional structure by projecting the observations via PCA and defining the feature functions to be the resulting lower dimensions. Other approaches specify a set of base feature components for constructing the feature functions, and then either jointly [30] or iteratively [92, 127] infer the task and feature parameters from demonstrations. **(author?)** [92] iterate between adding a new feature function as a logical conjunction of base integer components and learning a non-linear reward on top of the current feature set, while **(author?)** [127] follow a similar iterative procedure for imitation learning but train feature functions as regression trees.

Unfortunately, these early methods still rely on engineering a relevant set of base features, which can be very tedious and result in incomplete specification. Moreover, because they use low-capacity learning methods for the feature functions, they are limited to discrete or low-dimensional observation

spaces, and underperform the previously discussed deep learning methods which use more expressive architectures [93, 157]. Therefore, instead of learning features as logical conjunctions, more recent methods propose representing individual feature functions with neural networks [21, 20, 19, 120, 165], with the most straight-forward way being to provide labels for different diverse observations and train the feature function as a classifier [120, 165]. While **(author?)** [120] have been able to learn complex spatial relations mapping from high-dimensional point cloud observations, they require large amounts of data, making the approach unsuitable for learning many different feature functions individually from a human. A different approach reduces the data complexity for training feature functions by introducing a new type of structured input, a feature trace, which results in large amount of feature value comparisons to training the network with relatively little effort from the human [21, 20]. Meanwhile, **(author?)** [19] reduce the human burden by using a small amount of human labels to learn feature functions defined on a lower dimensional transformation of the observation space (object geometries) and using that to label data in a simulator (object point clouds).

Takeaway. While feature sets are advantageous for inserting structure in the reward or policy learning pipeline, making downstream learning more data efficient, robust, and generalizable [20], that added structure is only useful if it is complete. Luckily, the issue of under-specified feature sets can be alleviated by learning new feature functions over time, but we need ways of reducing the human burden for teaching individual features, like introducing new types of structured input [21] or bootstrapping the learning process [19].

A.7 Feature Embeddings

We review a vast body of work on representations learned as feature embeddings in a neural network. Here, the robot's representation $\phi_R(\mathbf{o}_R)$ is instantiated as a low-dimensional embedding, or vector, $\vec{\phi}_R(\mathbf{o}_R)$, where each dimension is a different neuron in the embedding. The representation function is, thus, $\phi_R : \mathcal{O}^t \mapsto \mathbb{R}^d$, with d much smaller than the dimensionality of the observation space. While feature set functions also map to $\mathbb{R}^d$, each dimension is learned individually (and is representative of some aspect of the task), whereas here the embedding vector is learned jointly all at once (and hopes to capture important task aspects more implicitly). We identify two broad areas in this space: unsupervised methods (also called self-supervised), which primarily use unlabeled data and proxy tasks to learn representations, and supervised methods, which use human supervision at the representation level. We also cover some in-between semi- or weakly-supervised methods.

Unsupervised methods. At the most human data-efficient extreme, unsupervised methods attempt to learn disentangled latent spaces from data collected while exploring the environment without any human supervision. Instead of explicitly giving feedback, the human designer hopes to instill their intuition for what is causal for the task by specifying useful *proxy tasks* [90, 29, 67, 28, 89, 162]. Many self-supervised proxy tasks have been explored in the robot learning community, from reconstructing the observation (to ignore visual aspects irrelevant to the observation) [156, 47, 68, 60, 101], to predicting forward dynamics (to capture what constrains movement) [156, 60] or inverse dynamics (to recover actions from observations) [118], to enforcing behavioural similarity between observations [167, 53, 11], to contrastive losses [151, 87, 8, 146], and many more, or some combination of these [63, 137]. The proxy task result itself does not matter; rather, these methods are interested in the intermediate representation extracted from training on the proxy tasks. However, because these methods are purposefully designed to bypass direct human supervision, the representations do not necessarily correspond to concepts in the human's representation, thus rendering explicit alignment challenging. In fact, the cases where the disentangled factors match human concepts appear to be primarily due to spurious correlations rather than theoretical guarantees [100]. Moreover, like all learned latent representations, they remain difficult to interpret and explain in the event of a failure.

Supervised Methods. On the other end of the spectrum, we have methods that do use human supervision. Some methods combine the human's reward or policy supervision with self-supervised proxy tasks to pre-train a useful low-dimensional feature embedding [24, 150] while other methods reduce the level of human supervision required by learning a simpler model that, when trained well, can automatically label large swaths of commonly available unlabeled videos of people performing tasks [13]. Multi-task methods also attempt to pre-train representations from human input coming from multiple tasks, and then fine-tune the reward or policy on top of the learned embedding at test time [55, 113, 161]. Similar to meta-learning, the motivation here is that across its life time the robot collects data from many different but related tasks, which it can then leverage for jointly

training a shared representation. The hope is that this representation holds meaningful structure for what's important for all tasks, thus helping the robot reuse the representation to efficiently learn new but related tasks. This has been shown to be more stable and scalable than meta-learning approaches [103], but still needs curating a large set of training tasks to robustly cover the test distribution.

Lastly, there is a growing body of work seeking to learn human-aligned representations by using direct supervision targeted at the robot representation level. *Implicit* methods make use of a proxy task for the human to solve and a visualization interface to change the displayed outputs based on the robot's current representation [128, 69]. The intuition is that if the human can still solve the proxy task well, the representation producing the visualization must contain causal behavioural aspects. If the representation dimensions are interpretable enough, *explicit* learning of human-aligned representations is also possible by directly labeling examples with the embedding vector values [73, 145]. What both these directions have in common is that the representation *is* or *can be* converted into a form that is interpretable to the human, thus opening the possibility of the human providing targeted feedback that is explicitly intended to teach the robot the desired task representation.

Takeaway. There appears to be a trade-off between the amount of human supervision at the representation level and how human-aligned the learned representations are. "Supervising" by coming up with proxy tasks certainly reduces the end user's potential labeling effort, but may also result in misaligned representations. On the other hand, direct supervision more explicitly aligns the robot's representation with the human's, but is also more effortful.

A.8 Graphical Structures

Representations in the form of graphical structures map observation histories onto a graph G, i.e. $\phi_R(\mathbf{o}_R) = G$ with $\phi_R : \mathcal{O}^t \mapsto \mathcal{G}$. Many graphical structure instantiations have been used for robot learning and planning, from knowledge graphs [37], to directed graphs [136], Markov random fields [58], Bayesian belief networks [83], hierarchical task networks [106], etc. Here, we briefly cover Knowledge Graphs (KG), Hierarchical Task Networks (HTN), Bayesian Networks (BN) and discuss their trade-offs.

KGs are comprised of world entities, e.g. "mug" or "table", and relationships or properties between them, e.g. "on top of". Because KGs are meant to be an explicit repository of complete world knowledge, they have been especially useful in situations where robust robot behaviour relies on strong priors for the task context, like interpreting ambiguous user commands [166, 163] or handling partially observable environments [115, 37]. Since their relational structure directly allows for probing the causal effect of a certain part of the representation on the robot's behaviour [158, 36, 37], they are also often leveraged in the interpretablity literature. The challenge is that building comprehensive KGs takes a considerable amount of human effort, as the entities and relations must be defined, stored, managed, extended, and made by the human [154, 97]. Furthermore, KGs can amass a large amount of entities and relationships, resulting in semantic redundancies (like synonyms) over a large search space that makes inference harder and slower [111]. Hence, recent methods have tackled this problem by learning KG entities and relationships as *embeddings*, which afford more efficient search and generalization [154, 114]. The downside is that we now lose the direct benefit of having more interpretable symbolic rules as in standard KGs [114], posing a clear tradeoff between the performance gains associated with more efficient data management vis-a-vis preserving the natural interpretable structure of graphs.

HTNs are tree-based representations that organize domain knowledge as hierarchies of primitive or compound tasks. Finding an appropriate robot policy on such a representation involves expanding all compound tasks until a legal chaining of primitive tasks is found [52]. This technique is very advantageous for fast and robust planning [10, 116, 96], but requires well-conceived, well-structured, and comprehensive domain knowledge (primitive tasks and hierarchy) to be successful: if one of the primitives on the optimal plan fails (due to, say, distribution shift), the representation may not contain enough information to recover [109, 107]. Various approaches have tried to alleviate this problem by learning the primitives themselves [105], the hierarchy given the primitives [104], or both [64, 27, 106, 94]. Unfortunately, most of these methods rely in turn on a set of hand-specified "base" primitives to construct useful primitives. To address missing or erroneous primitives, recent work has combined HTNs with knowledge graphs for extracting the necessary additional information

to solve the task [115, 37]. However, the hand-specified requirements associated with building base primitives still takes considerable human time and effort.

BNs are probabilistic models represented as directed acyclic graphs where the nodes are task variables (e.g. the observation history) and the edges between them symbolize a probabilistic conditional dependency. Many works in robotics hand-define a task-specific BN structure and focus on learning the corresponding task probabilities [144, 79, 125, 41, 32]. We are instead interested in the much harder problem of *learning* the BN structure itself before applying it for a task. One way is to define the nodes as an exhaustive set of atomic components (for example, histories of binary observations [78, 75, 88] or hand-crafted features of the observation histories [108, 98]) and find the appropriate graph edge structure via heuristic search methods. Such methods do not scale well to real-world settings, so alternative approaches attempt to build more compact representations of the node space by adaptively discretizing it [169, 56] or by using dimensionality reduction like the sparse Gaussian Process [143] or generative modeling [42]. Methods in causal structure learning have looked into constructing the graph structure based on the causal effect that each variable has [34, 121] on the others, being able to leverage neural networks to learn causal graphs from data [95, 40, 102].

Takeaway. While graphical structures are more interpretable to human users, they require significant human effort to construct and maintain relative to their neural network counterparts. Much like trying to specify rewards by hand, it is hard to specify all the nodes that matter, potentially resulting in under-specification.

A.9 Open Challenges

A.9.1 Learning Human-Aligned Representations

Designing human input for representation learning. One direction for learning human-aligned task representations is to provide the requisite tools for human users to directly give input informing the robot of the representation itself, rather than task inputs [21, 73, 19, 145]. The aspiration here is that the dimensionality of the robot task(s) representation is both smaller than that of the tasks themselves and also shareable between tasks, meaning that explicitly targeting human input for learning robot representations *prior* to learning the downstream task distribution should require less overall supervision. We propose that a key direction of future work is considering new types of *representation-specific* human input that are highly informative (and intuitive to understand) about desired task representations without being too laborious for a human to provide, such as natural language or gaze and pose. Moreover, we can also explore methods that recover the entire representation via representation-specific proxy tasks – *calibration* tasks where the robot's goal is to specifically align its full task representation with that of the demonstrating human. For this to be an actionable direction, we also encourage development of new interactive interfaces that allow for effective communication of desired human representation labels, such that even inexperienced users are able to provide useful input.

Transforming the representation for human input. A second complementary approach is to directly design robot task representations to resemble those naturally understood by humans. Previously, when we instantiated the representation as a set of learnable features, we gave the human freedom to decide what feature each dimension of the representation was and provide feedback for teaching it to the robot. This enabled the human to add desirable task aspects to the representation even if the system designer did not originally think of them. In some cases, though, it may be possible for the system designer to transform the full task representation into a form that is more aligned with how humans perceives of the task. This can happen if the designer has prior knowledge that the class of features the robot needs to learn for the desired task has a well-studied human representation. For instance, it is well studied in the cognitive science and neuroscience literatures that human planning operates on hierarchical abstractions conditioned on the desired task, such as visual masks that filter out irrelevant scene features for navigation [26], rather than on low-dimensional features of the raw observed state [71]. Knowing this, we can turn our attention to instantiating learnable robot representations that are well equipped for soliciting human input of the same form, such as masked image states for visual navigation. Moving forward, we should explore other avenues of leveraging human-comprehensible concepts, such as natural language, for instantiating robot representations (**(author?)** [140, 122]). This will be beneficial for not only downstream task learning, but also for forming a shared language by which the robot can effectively communicate to the human what it

thinks is the correct representation prior to actual deployment. We propose that effective human-robot interaction which leads to learning human-aligned representations will require approaches on both fronts to fully leverage information flow between humans and robots.

A.9.2 Detecting Misalignment

Robot Detecting Its Own Misalignment. In addition to learning human-aligned task representations, it is also important to build robots that detect when their own learned representations are not aligned with those of the humans', else they may misinterpret the humans' guidance for how to complete the task, execute unexpected or undesired behaviour, or degrade in overall performance [17]. Ergo, we wish for the robot to *know when it does not know* the human's representation *before* it starts incorrectly learning how to perform the task. If misalignment is detected, then the robot can begin a process for re-learning or expanding its existing representation rather than wastefully optimizing an incorrect representation.

Currently, there are two approaches for robots to detect misalignment: one Bayesian-based and one deep learning-based. In the Bayesian approach, the robot models the human as a noisily rational agent choosing input in proportion to their exponentiated reward [14, 82, 153], and can therefore jointly infer both the reward parameter and a confidence measure of whether the desired reward function can be accurately captured by the current representation [48, 18, 99, 17, 171]. If the human input refers to an input that the robot's learned representation cannot support, the inferred confidence is low, signaling misalignment. In the deep learning approach, representation uncertainty is measured through an *ensemble* of neural networks [86, 147], where if multiple (identically trained) networks disagree on their predictions, input is flagged as out of distribution and the representation potentially misaligned.

Once misalignment is detected, the robot can either discard the human input entirely, continue learning in proportion to its assessed confidence, or halt execution and begin representation alignment ([17]). Unfortunately, building in autonomous strategies for robots to detect their own misalignment remains difficult in many real-world scenarios, especially when there is inherent difficulty in disambiguating between representation misalignment and human noise [17]. This issue often arises from inexperienced users and is inherent to the types of data designers must work with in human-robot interaction scenarios. A proposed, albeit expensive, method of addressing this challenge is to collect more data to balance out noise, but this solution would not fare well in online learning scenarios where the robot must detect misalignment in real time. We suggest that developing methods for fast, online misalignment detection remains critical for real-world deployment scenarios.

Human Detecting Robot Misalignment. A second alternative direction of research is to instead build in methods that allow for human users to detect when a robot's learned representation is misaligned with their own. The advantage of this approach is simple: while the previous section identified a central challenge in robots needing to disambiguate between human input vs. noise, this would not be the case if the tools for identifying a correctly learned representation were instead given to the human themselves, i.e. *a human should know what they want the robot to do*.

In the simplest of cases, the human would be able to detect misalignment by observing bad behaviour produced by the robot, but such behaviours are rarely informative of the underlying reason of *why* the robot failed [84]. Because of this, the field of robot explainability has prioritized the development of tools that are informative of the causal factors behind an underlying system failure [38, 35, 37, 36]. Consequently, many methods focus on generating post-hoc explanations for explaining behaviour [54, 15, 65, 135, 57]. Unfortunately, in real-world deployments, especially those with the added risk of potential safety hazards, e.g. self-driving cars, users may not have the luxury of being able to observe the consequences of a robot's failed representation *after* the fact. Therefore, a growing body of work has started to build tools for allowing humans to interpret and correct robot representations *prior* to deployment [128]. We remain hopeful that this is a promising direction of inquiry, and suggest that building in mechanisms for humans to explicitly correct representations should be an integral part of the learning process.

A.9.3 Evolving a Shared Representation

Until now, we have assumed that the goal of the robot is to adapt its representation to that of the human's, i.e. the human representation $\phi_H(\mathbf{o}_H)$ is fixed and fully captures the desired task. However,

it is worth re-visiting this assumption. In collaborative scenarios, it is possible that the robot holds a more complete task representation that it wishes to communicate to the human, i.e. teach the human new aspects of the task that they were not aware of before. This may occur in situations of partial observability, where the robot's $\mathbf{o}_R$ contains information valuable to solving the task that are not captured by the human's $\mathbf{o}_H$ (say, the robot can see a useful tool that the human cannot), or incomplete knowledge, where the robot possesses knowledge of how to leverage an aspect shared by $\mathbf{o}_H$ and $\mathbf{o}_R$ that the human does not (say, the robot knows how to use a tool in a way that the human does not). One way for the robot to communicate this information is to show the human examples $\langle \mathbf{o}_H, \hat{f}(\phi_R(\mathbf{o}_R)) \rangle$ where observations are labeled with the robot's estimate of the representation transformation function. We can also envision a situation where neither the robot nor the human individually hold a complete representation, and must jointly communicate missing aspects of the desired representation. By alternating between the direct (robot learning about the human's representation) and the reverse (robot teaching the human about its representation) channels of communication, we can enable reaching a mutual representation that is most informative to completing the task.