Aligning Robot Representations with Humans

Andreea Bobu
University of California, Berkeley
Berkeley, CA, USA
abobu@berkeley.edu

Andi Peng
Massachusetts Institute of Technology
Cambridge, MA, USA
andipeng@mit.edu

***Abstract*—As robots are increasingly deployed in real-world scenarios, a key question is how to best transfer knowledge learned in one environment to another, where shifting constraints and human preferences render adaptation challenging. A central challenge remains that often, it is difficult (perhaps even impossible) to capture the full complexity of the deployment environment, and therefore the desired tasks, at training time. Consequently, the *representation*, or abstraction, of the tasks the human hopes for the robot to perform in one environment may be *misaligned* with the representation of the tasks that the robot has learned in another. We postulate that because humans will be the ultimate evaluator of system success in the world, they are best suited to communicating the aspects of the tasks that matter to the robot. Our key insight is that effective learning from human input requires first explicitly learning good intermediate representations and then using those representations for solving downstream tasks. We highlight three areas where we can use this approach to build interactive systems and offer future directions of work to better create advanced collaborative robots.**

I. Introduction

Imagine a world where you wake up in the morning, arise from bed, and your home robot assistant makes your bed. After getting ready, you head downstairs where your robot has placed a steaming mug of fresh coffee on the table exactly where it knows you will sit. After drinking the coffee, your robot picks up the empty mug and places it in the dishwasher as you leave the house and set off for work. The entire morning, your robot is incorporated seamlessly into your daily life and home. This scene of domestic bliss captures the essence of what we hope for from our advanced collaborative assistants – the ability to effectively complete desired tasks while integrating into our environments and adapting to our individual preferences, akin to human-like collaboration.

Today, autonomous systems are increasingly able to learn advanced behaviors like those mentioned above (Abbeel et al. [2], Kolter et al. [31], Wulfmeier et al. [51]). However, designing learning algorithms that match the adaptability and generalizability of human reasoning remains challenging: while these systems may perform their tasks successfully in the environment(s) and under the conditions they were trained on, their learned behaviors may not necessarily work well in novel deployment environments. This problem can rear its head in a variety of instances: when physical constraints change (while it's okay for the robot to break mugs when trying out new grip poses in the lab, we may wish for them to be more careful in a home), when environment conditions, layouts, or compositions change (we may wish for the robot to grasp an octopus-shaped mug that it's never before seen), or when the task preferences of the human that the robot interacts with change (one human may prefer that the coffee is prepared as quickly as possible irrespective of mess, while another may prefer that the robot prioritizes not spilling the coffee while navigating the kitchen).

The key issue in all these cases is that, while the designer can anticipate some of the possible task specifications when training the robot, these specifications do not necessarily reflect the desires of the other humans the robot will interact with in its lifetime (Ng et al. [38], Levine et al. [35]). In other words, the *representation*, or abstraction, of the tasks the human hopes for the robot to perform in one environment may be *misaligned* with the representation of the tasks that the robot has learned in another. Our observation is that because humans have adapted their environments to capture the full idiosyncrasies of completing tasks that they desire, they are best equipped to help insert knowledge specifically describing aspects of the environment that are useful to the robot for learning. Specifically, human input can best help solve the *representation alignment* problem of transferring task aspects that matter to the human when adapting to a new environment.

Traditional methods of robot learning from human input instantiate representations as a set of hand-engineered *features*—specific aspects of the task that a human may care about (Ziebart et al. [53], Hadfield-Menell et al. [27], Abbeel and Ng [1], Bajcsy et al. [4], Osa et al. [39]). These features are pre-specified by a system designer and function as state-space abstractions that insert structure for learning the task efficiently. However, they can be difficult to construct and impossible to exhaustively specify. Meanwhile, state-of-the-art deep learning methods (Wulfmeier et al. [51], Finn et al. [22], Christiano et al. [19], Fu et al. [25, 26], Brown et al. [15], Abbeel and Ng [1], Torabi et al. [48]) bypass feature specification by operating directly on high-dimensional state spaces, thereby automatically constructing an *implicit* representation from the person's task-specific input (e.g. demonstrations). Unfortunately, because these methods are optimized to learn the task while bypassing the explicit need to learn the representation, there is difficulty in disentangling the high-level representation from the specific task provided Fu et al. [25], Reddy et al. [43], Bobu et al. [9]. Consequently, effective task learning requires massive amounts of training data and renders generalization to new tasks difficult. In summary, one paradigm inserts useful structure to solve the robot learning problem efficiently but that structure is difficult to define; the

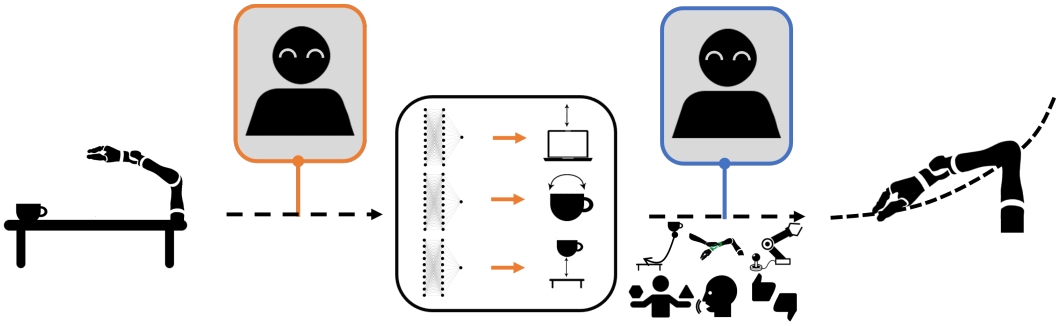

Fig. 1. Under our framework, the robot first learns *human-guided representations* by asking the human for representation-specific input to capture specific aspects of the task that they care about (e.g. distance to laptop, cup orientation, cup near table). The robot then uses the representation to learn how to perform the task from task-specific input like demonstrations, corrections, etc.

other avoids explicitly specifying the structure but requires too much human data to extract it implicitly and thus struggles to generalize across different domains.

We postulate that effective learning from human input requires methodologies that combine the best of both traditional feature engineering and highly-expressive deep learning worlds. Our core idea is to **divide and conquer** the learning problem: *explicitly* focus human input on teaching robots good intermediate representations before using those representations for downstream tasks. We call these *human-guided representations*: abstractions that, if learned well, can enable robots to better solve tasks when deployed into the real world. We discuss several directions for learning human-guided representations as well as strategies for identifying misalignment and improving effective downstream task learning.

II. Learning Human-Guided Representations

The representation learning literature has accrued a vast body of work on learning disentangled latent spaces in an unsupervised manner (Chen et al. [18], Higgins et al. [28], Chen et al. [17]). However, because these methods are purposefully designed to bypass direct human supervision, the disentangled factors in the learned embedding do not necessarily correspond to concepts in the human's representation. In other words, the robot's learned representation does not necessarily align with the human's, therefore adapting to how they want the task to be done is difficult. Self-supervised learning inserts some human guidance by allowing for the designer to specify proxy tasks useful for feature learning (Doersch et al. [21], Pathak et al. [40], Aytar et al. [3], Brown et al. [15], Laskin et al. [33]) (for example, predicting forward dynamics to capture what constrains movement). In this process, the human designer hopes to instill good representations into the robot by using their intuition to construct tasks which illustrate specific features. However, devising proxy tasks is an exercise that requires nontrivial effort and expertise: human effort to manually specify features is instead traded for human effort to specify objective functions for extracting those features.

A more direct way to guide representation alignment is to learn directly from human input. In standard imitation learning, the robot learns a policy that copies—or clones—human demonstrations (Osa et al. [39], Abbeel and Ng [1]). However, it cannot learn to imitate what it has not seen before, thus rendering human input non-generalizable to new tasks (Levine et al. [35], Torabi et al. [48]). Moreover, BC suffers from the problem of covariate shift, where once a learned policy drifts away from the demonstrations, errors compound more and more over time. Inverse reinforcement learning (IRL) attempts to extract a reward function from demonstrations that is intended to capture *why* a specific behaviour is desirable (Abbeel and Ng [1]), but unfortunately requires massive amounts of data to truly learn a fully-specified reward (Fu et al. [25], Reddy et al. [43]). IRL also requires expert or close to expert demonstrations (Ziebart et al. [53]). Meta-learning reduces this sample complexity by reusing demonstrations from an array of different tasks in the training distribution (Finn et al. [23], Xu et al. [52]), but still requires the human to know the test time task distribution *a priori*, which brings us back to the specification problem: we now trade hand-crafting features for hand-crafting tasks.

Because demonstrations are intended for teaching the robot *how* to do the tasks, not *what matters* for doing the tasks, they can only contribute to aligning representations implicitly. This might not result in learning algorithms extracting salient features that matter to the human for performing the desired tasks (Bobu et al. [9]). As shown in Fig. 1, we propose that the robot should explicitly ask for *representation-specific* human input to teach it the intermediate representation before using it to learn more generalizable downstream tasks from task-specific input. Importantly, because of this separation, these representations are not specific to any one particular task the human may want the robot to do; instead, they capture aspects causal for the potential task *distribution* in the environment.

Designing human input for representation learning. One option for learning intermediate human-guided representations is to instantiate them as feature sets like those in traditional methods, and let the human teach individual, novel features themselves (Bobu et al. [12, 9]). A natural way to represent any specific new feature is via a neural network which is trained by asking the human for supervision labels representing the feature values at different states. Unfortunately, querying the human for labels to train this neural network requires a burdensome amount of human interaction. Even worse, humans are notoriously imprecise at giving these types of numerical inputs, rendering learned representations likely

erroneous (Braziunas and Boutilier [13]). We propose that a key direction for future work is considering new types of representation-specific input that are highly informative about the feature without requiring too much effort from the human. For example, a new type of structured human input called a *feature trace* (Bobu et al. [12]), where a human guides the robot from states where the feature is highly expressed to states where it is not, has been found to recover more robust and generalizable rewards with far less human effort. Moving forward, we can study additional forms of human input such as language or gaze and pose, that can also be targeted for feature learning. Moreover, we can also consider types of human input that recover the feature representation as a whole (rather than one by one) via representation-specific proxy tasks – *calibration* tasks where the robot's goal is to specifically align itself with the demonstrating human.

Transforming the representation for human input. Instead of designing the type of input the human can teach the representation with, we can directly design the type of representation itself. 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 specify the necessary dimensions of the representation, just not the mapping to the representation itself. This could happen, for example, if the designer has prior knowledge that the class of features the robot needs to express for its tasks has a well-studied representation. For instance, recent work defines a model to relate emotions expressed in natural language, such as 'happy' or 'sad', into the Valence-Arousal-Dominance spectrum inspired by social psychology (Sripathy et al. [46]). The human can teach the representation efficiently with natural language by having the robot map their utterances to their emotive latent VAD equivalent. This way, all user feedback for this representation contributes to learning about all emotions, and the robot can model new emotions that interpolate those seen during training. Moving forward, we should leverage existing methods that transform human-comprehensible concepts, such as language or images, into robot-comprehensible representations for downstream learning (Shridhar et al. [45], Radford et al. [41]).

Designing the human-robot interface for learning. In order to truly deploy collaborative robots in the world, we must eventually develop usable interactive interfaces that allow for effective information exchange of representations understood by both the human and robot. Existing work has highlighted the importance of the interface when a human and robot collectively share the same workspace, with key considerations being ease of use, specificity of communication, and reliability of feedback (Wright et al. [50], Bansal et al. [6]). Current methods suggest using visual displays, hand or face gestures, physical interaction and haptics, and verbal language can all be viable solutions towards effective human communication (Berg and Lu [7]). However, less work has been done in interfaces for how the robot can effectively communicate the representation of what it has learned with the human. For example, it would be desirable to have an interface by which the robot can effectively demonstrate or show the human what it *thinks* is the correct desired task prior to actually deploying it in the real-world. This could be done in the form of mapping the proposed robot policy to simulated demonstrations or even natural language to communicate the intended behaviour. We propose that effective human-robot interaction which leads to learning human-guided representations will require the development of both streams of information flow in order to fully achieve its potential.

III. Identifying Misalignment

Along with learning transferable human-guided representations, it is also important to detect when misalignment exists in the first place. Misaligned representations may cause the robot to misinterpret the human's guidance for how to complete the task, execute unexpected or undesired behaviors, or degrade in overall performance (Bobu et al. [8]). Ergo, we wish for the robot to *know when it does not know* the aspects that matter to the human *before* it starts incorrectly learning how to perform the task. If misalignment is correctly detected, then a process which begins with expanding or re-learning the representation will better help ultimately learn the downstream task.

Several methods suggest an introspective approach where the robot can maintain uncertainty in its representation's ability to explain the human's input. By modeling humans as noisily rational agents choosing inputs in proportion to their exponentiated rewards (Baker et al. [5], Jaynes [30], Von Neumann and Morgenstern [49]), Bayesian approaches can jointly infer both the reward parameter and a *confidence* in whether the desired reward function can be captured by the current representation (Fridovich-Keil et al. [24], Bobu et al. [10], Losey and O'Malley [36], Bobu et al. [8], Zurek et al. [54]). When the human input refers to a reward that the robot's representation cannot support, the inferred confidence is low, signaling misalignment. Meanwhile, deep learning methods often study this uncertainty through an *ensemble* of neural networks (Lakshminarayanan et al. [32], Sun et al. [47]). The intuition here is that if multiple (identically trained) networks disagree on their predictions, this suggests that the input is out of distribution and therefore the representation is misaligned.

In both cases, once the robot detects misalignment there are a few options for how to proceed: discard the human input entirely, continue learning in proportion to its assessed confidence, or halt execution and ask the human to undergo the process of representation alignment from the previous section (Bobu et al. [8]). Assuming the robot identified misalignment correctly, any of these options are viable alternatives to re-learning from the original human feedback. Unfortunately, robustly detecting misalignment remains difficult in many real-world scenarios. We highlight three key areas where identifying misalignment is particularly challenging and offer brief suggestions for future work.

Disambiguating between misalignment and noise. When a robot's representation cannot explain the human input, it may be difficult to disambiguate whether this is due to representation misalignment or human noise (Bobu et al. [8]). 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, from just a few observations. We suggest that a more sustainable alternative is to investigate better human modeling for separating out these two errors (Ramakrishnan et al. [42]).

Poor feature learning. Misalignment can additionally occur due to two reasons: either the robot's representation does not fully capture an aspect that the human cares about or it does, but *poorly*. The latter can occur if some of the features were not learned well enough; for example, a feature might have required more data from the human to cover the state space and generalize to new areas. We propose that it is crucial for the robot to distinguish between misalignment due to an incomplete representation or to incorrectly learned dimensions of the representation so that instead of attempting to re-learn a new feature, the robot knows to query for more data on the existing one. Future work is needed to understand whether the robot needs to repair an existing learned feature, detecting which feature that may be, and developing interactive methods to elicit informative data to improve existing features.

Feature confusion. An even more fundamental issue exists when the human's input refers to something not captured by the robot's learned representation, but the representation nonetheless can explain their input. In this case, we have confused misalignment for human noise (Sun et al. [47], Bobu et al. [8]). This problem will especially occur if the representation is highly expressive and can only be solved by intaking additional human input: each input might be explainable by some hypothesis, but eventually no hypothesis can explain all input. More work is needed to study how to query for a broad and diverse set of input, how the robot would best demonstrate the features it has learned to the human, and how to best balance between querying vs. learning with existing data.

IV. Learning the Downstream Task

Once we have learned a human-guided representation, it is easy to then apply that representation towards learning a downstream task by using standard policy (Ng et al. [38], Levine et al. [35, 34]) or reward learning techniques (Jain et al. [29], Bajcsy et al. [4], Brown et al. [14], Fu et al. [26]). However, human-guided representations have important implications for how they impact the downstream learning pipeline. We subsequently discuss three considerations that future work should consider to fully close the learning loop.

Using the right features at the right time. We have advocated for learning a human-guided representation that is sufficiently decoupled from any specific task and focuses instead on capturing causal aspects for the potential downstream task distribution. When the robot specializes on a task, the representation by construction will contain features that are irrelevant. If all feature dimensions in the representation were orthogonal to one another, this would not cause any issue. However, in the real world, many relevant features may be related and, thus, *spurious correlation* between features could affect task learning (de Haan et al. [20]). Future directions should enable the robot to *focus on the right features at the right time*. One idea for accomplishing this is to employ feature selection strategies to activate the subset of the representation that matters for the specific task at hand. This strategy could be heuristic-based, like choosing the minimum set that maximizes coverage (Sax et al. [44]). Alternatively, since we would hope for learned representations to be interpretable, we could also consider building interfaces where the person themselves can quickly indicate to the robot which features are important for the desired task (Cakmak and Thomaz [16]).

Using representations to better understand humans. Human-guided representations also enable us to learn something about how the person generates the task input in the first place. In particular, the previously mentioned human decision-making models (Baker et al. [5], Von Neumann and Morgenstern [49], Luce [37]) assumed that, out of a set of choices, the person selects their input in proportion to these choices' exponentiated rewards. However, we suggest that human-guided representations inform the robot how it should interpret the person's task input, thus we should *reinterpret the available choices from the perspective of the learned representation* (Bobu et al. [11]). We suggest future research must revisit how robot learning methods are affected by reinterpreting human input through the lens of their representation.

Grounding representations to real-world tasks. Much of HRI has historically assumed that the robot already has access to all the aspects in the environment that the interacting human might care about. This assumption has enabled researchers to make progress on human-robot collaborative algorithms without needing to worry about how to formally ground the robot's behaviour to complex environments and tasks that we would see in the deployment scenarios. Human-guided representations can help bridge the gap towards learning from high-dimensional state spaces as we know the real-world to be, opening the door to HRI applications more challenging and tractable than ever before.

V. Conclusion

Ultimately, the true evaluators of any system deployed in the world will be the humans that it interacts with, and thus soliciting input from them to effectively learn downstream tasks is critical. Learning effective methods to learn from humans holds the promise of enabling more advanced, aligned robotic systems. We proposed several methods for learning more generalizable representations from humans and suggested directions for moving towards a continual and interactive learning framework. It is through understanding and utilizing this bi-directional communication flow that truly effective human-robot collaboration can exist.

References

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

[2] Pieter Abbeel, Adam Coates, and Andrew Y Ng. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research*, 29(13):1608–1639, 2010.

[3] 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.

[4] 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. URL http://proceedings.mlr.press/v78/bajcsy17a.html.

[5] 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.

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437, 2019.

[7] Julia Berg and Shuang Lu. Review of interfaces for industrial human-robot interaction. *Current Robotics Reports*, 1(2):27–34, 2020.

[8] 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.

[9] 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. doi: 10.1177/02783649221078031. URL https://doi.org/10.1177/02783649221078031.

[10] 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. URL http://proceedings.mlr.press/v87/bobu18a.html.

[11] Andreea Bobu, Dexter R. R. Scobee, Jaime F. Fisac, S. Shankar Sastry, and Anca D. Dragan. *LESS is More: Rethinking Probabilistic Models of Human Behavior*, page 429–437. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450367462. URL https://doi.org/10.1145/3319502.3374811.

[12] 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. ISBN 9781450382892. doi: 10.1145/3434073.3444667. URL https://doi.org/10.1145/3434073.3444667.

[13] Darius Braziunas and Craig Boutilier. Elicitation of factored utilities. *AI Magazine*, 29(4):79, Dec. 2008. doi: 10.1609/aimag.v29i4.2203. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2203.

[14] Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International Conference on Machine Learning*, pages 783–792. PMLR, 2019.

[15] 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. URL http://proceedings.mlr.press/v119/brown20a.html.

[16] Maya Cakmak and Andrea L. Thomaz. Designing robot learners that ask good questions. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 17–24, 2012. doi: 10.1145/2157689.2157693.

[17] 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.

[18] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2180–2188, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[19] 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.

[20] 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.

[21] Carl Doersch, Abhinav Kumar Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[22] 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.

[23] 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.

[24] 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.

[25] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkHywl-A-.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] E. T. Jaynes. Information theory and statistical mechanics. volume 106, pages 620–630. American Physical Society, May 1957. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

[31] J Zico Kolter, Christian Plagemann, David T Jackson, Andrew Y Ng, and Sebastian Thrun. A probabilistic approach to mixed open-loop and closed-loop control, with application to extreme autonomous driving. In *2010 IEEE International Conference on Robotics and Automation*, pages 839–845. IEEE, 2010.

[32] 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. ISBN 9781510860964.

[33] 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. URL https://proceedings.mlr.press/v119/laskin20a.html.

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

[35] 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.

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

[37] R. Duncan Luce. *Individual choice behavior.* John Wiley, Oxford, England, 1959.

[38] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287, 1999.

[39] 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.

[40] 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. doi: 10.1109/CVPRW.2018.00278.

[41] 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 lan-

guage supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[42] Ramya Ramakrishnan, Vaibhav Unhelkar, Ece Kamar, and Julie Shah. A bayesian approach to identifying representational errors, 2021. URL https://arxiv.org/abs/2103.15171.

[43] 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. URL https://openreview.net/forum?id=S1xKd24twB.

[44] Alexander Sax, Bradley Emi, Amir R. Zamir, Leonidas J. Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. In *Conference on Robot Learning*, 2018.

[45] 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.

[46] 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. URL https://arxiv.org/abs/2203.02091.

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

[48] 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. ISBN 9780999241127.

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

[50] Julia L Wright, Jessie YC Chen, and Shan G Lakhmani. Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems*, 50(3):254–263, 2019.

[51] 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.

[52] 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. URL https://proceedings.mlr.press/v97/xu19d.html.

[53] 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. ISBN 978-1-57735-368-3. URL http://dl.acm.org/citation.cfm?id=1620270.1620297.

[54] 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. doi: 10.1109/ICRA48506.2021.9561839.