Robot Action *for* and *around* People

Most robotics problems look something like this: 1) an engineer writes down an objective, in the form of reward, cost, loss, constraints, goal, acceptability threshold, etc. 2) the robot is in charge of finding behavior that meets the objective. But when robots need to interact with people, the way we formulate these problems needs to fundamentally change, and, with them, the algorithms we use to generate the robot's behavior.

First, robots will not act in isolation. From autonomous cars to quadrotors to mobile manipulators in the home, robots will work with and around us. This makes robot action chosen in isolation far from sufficient – robots will need to choose actions that **coordinate** well with ours. Second, robots **assisting** us will need to do what *we* want them to do. Only the passenger in a self-driving car knows what it means for the ride to be comfortable, and how much they want to prioritize comfort over efficiency. As for us, the robot engineers, we might know what behavior we want to see, but not necessarily what magic set of numbers will make the robot produce it reliably across any new situation it might face. Thus, figuring out how to generate the behavior is only half the battle. The other half is figuring out the objective itself. And the key to that lies with us, people – what *we* want internally, be it as end-users or as engineers, should be the robot's objective, even if we can't always explicate it.

My group's research agenda is to formalize and algorithmically solve the problem of robot action not in isolation, but for assistance of and in coordination with people – this is what I call robot action *for* people, and *around* people.

Approach. We formalize the problem as a human-robot system. The key to our approach is modeling people as *intentional* agents, whose behavior is driven (albeit imperfectly) by what they want. We formalize intent – someone's goals, preferences about the task, and preferences about their interaction with the robot – generally via a reward function whose parameters are unknown to the robot, and thus latent state. Robot action *for* people is then defined by optimality with respect to this reward function. The same function drives human behavior, so robot action *around* people becomes a response to this behavior. This framework gives us a systematic way to break down interaction into its core components: identifying sources of *evidence* about the latent state (1), *modeling* their relation to the latent state to perform estimation (2), generating the robot's actions with respect to the latent state, in light of *the human also acting* simultaneously (3), and accounting for the engineer/designer as a human that is part of the human-robot system (4).

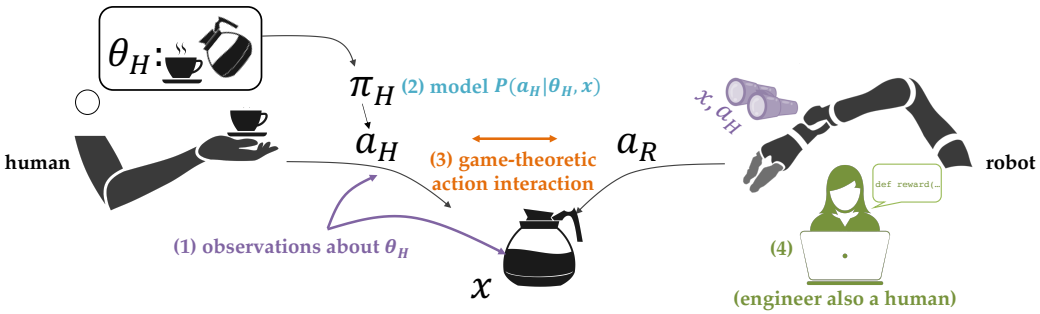


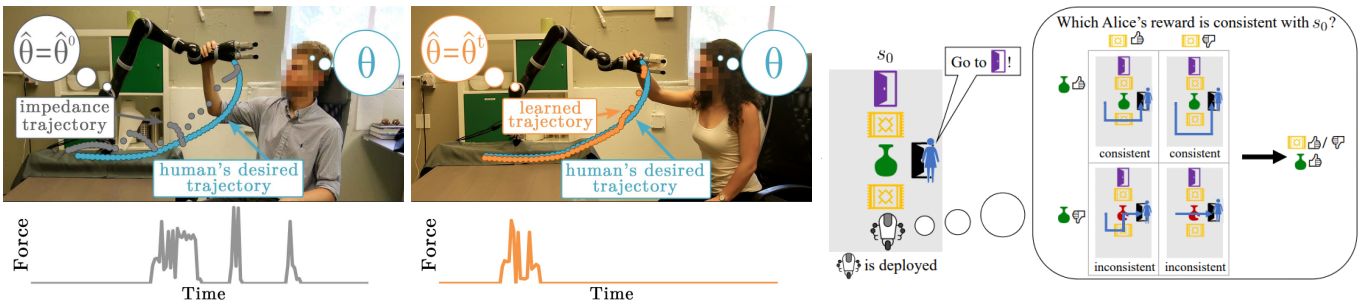
Figure 1: The human wants coffee, which the robot does not directly observe (latent state θ_H). The robot can treat human actions a_H and even the state of the world x as observations about θ_H (1), via a model of how the human acts based on θ_H (2). The human knows the robot is trying to help, and the human and the robot act simultaneously (3). Here, the fact that the pot has already brewed coffee (x) leaks information that the human values coffee enough to brew it. The human is moving to set the cup down next to the pot (a_H), expecting the robot will fill it. The robot uses x and a_H to figure out what the human wants, and starts reaching for the pot.

[1. Observations] Learning from explicit and leaked information under a unified formalism

To estimate human intent, the robot needs some source of evidence about it – it needs observations (or measurements). Traditionally in robotics, these come from the human demonstrating the task on their own, which is both tedious and often insufficient. Instead, we've shown that when you step outside of an artificial lab setting, observations beyond demonstrations become readily available:

> *People leak information about what they want when they physically react to what the robot does (e.g. push it away from them [1, 2]), when they tell it to stop what it's doing or switch it off [3], etc. And it gets even better: even the someone's environment itself leaks information about their preferences [4], because they have been acting in it according to their preferences!*

Imagine someone tells you to clean up a room, and as you walk in you see in the corner an elaborate house of cards. Even though the person didn't say it, you implicitly know you shouldn't clean up the house of cards. This is an example of the environment itself implicitly communicating about what the robot should optimize. We've



(a) Preferences implicit in physical corrections

(b) Preferences implicit in the state of the environment

Figure 2: Observations about what people want. (a) Physical corrections. Left: the robot is compliant, but does not learn from the human's intervention. Right: the robot uses the user's external torque to update its estimate of how the user wants the robot to move. (b) The state of the environment: the user tells the robot to go to the door, but the robot figures out, from the fact that the vase in the middle of the room is still intact, that it should probably go around it.

contributed algorithms for learning from these sources, as well as from explicit human feedback such as answers to comparison queries [5] or feature queries [6].

We've also enabled robots to capitalize on all these sources together by contributing a unifying formalism for reward learning from human feedback [7]. The idea started by looking at explicit sources, and generalize from there. For instance, when the person tells the robot they prefer a behavior over another, we know how to interpret that – we see it as a choice that the person is making, with respect to the reward. When the person gives a demonstration to the robot, they are also making a choice – it's just that this one is *implicit*: they are choosing the demonstrated behavior over all other behaviors they could have demonstrated, but chose not to. Again, this choice is relative to the reward. When the person, say, turns the robot off, we argue they are also making a choice: they could have done nothing and let the robot continue, but, implicitly, they chose not to. Thus, to make sense of such leaked information, as well explicit feedback like scalar rewards, comparisons, or credit assignment, all in one algorithm,

> *We proposed that all human feedback, despite its diversity, can be seen as a choice the human makes implicitly – a choice that is based on the reward function, even though the reward of a choice can often not be directly evaluated.*

The trick is that when choices are not robot behaviors, as in turning the robot off, correcting it, or even the state of the environment, they can still be *grounded* in behavior. Find the grounding, and we know how to link the choice to the reward. For instance, switching the robot off grounds to the trajectory the robot was pursuing, followed by staying stopped for the rest of the time horizon. This lens helped provide conceptual clarity on reward learning, enabled robots to combine different types of feedback together and actively select informative feedback types, and also gave us a recipe for formalizing yet-to-be-invented sources of information.

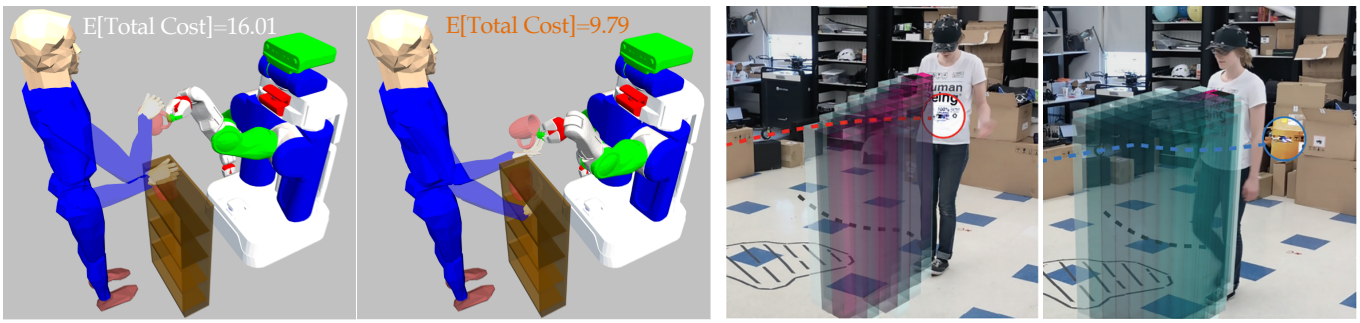
[2. Human Model] Modeling irrational human behavior and being robust to misspecification

How does human behavior, which is observable, relate to the human underlying intent? While thinking of human behavior as noisily rational has taken us far, behavioral economics has long warned us that people make systematic deviations from rationality. We've shown that such deviations can make intent inference go completely wrong, so the robot really needs to account for them. Unfortunately, while behavioral economics has identified a plethora of domain-specific heuristics, attempting to somehow encode these for the robot is not scalable. Instead, we noticed that much of the behavior that appears irrational might actually be derivable from first principles:

> *Our idea was to treat people as actually rational, but under* different *assumptions than those of the robot's.*

We can then leverage data to learn what these assumptions are [8]. For instance, while users operating a complex robot might seem really suboptimal, their actions make perfect sense under *their own, internal, dynamics model* – people have a notion of intuitive physics that fails to capture all intricacies of the real system; learn what this internal model is, and that gives us the key to figure out what they want and assist them in spite of the suboptimality. We've therefore used this generalized rationality framework to improve robot assistance by modeling people as assuming a different dynamics model [9], a shorter time horizon [10], or as still learning about their own preferences [11]. Further, noisy-rationality came from econometrics and discrete spaces, and we showed better performance by re-deriving it for continuous robotics spaces [12].

Nonetheless, no model is ever perfect, which leads robots to infer the wrong reward, or get unsafely close to the person due to wrong predictions. We've proposed that the robot should estimate the person's apparent rationality online:



(a) R assumes global H; H chooses grasp myopically and incurs high cost at the goal

(b) R accounts for myopic H; H chooses grasp myopically but still incurs low cost at the goal

(c) R does not know H is avoiding coffee spill, makes confidence predictions, collides

(d) R detects its assumptions are wrong and automatically becomes more conservative

Figure 3: a-b: The robot accounts for the human as rational under a shorter time horizon, and can better assist by compensating for the human's myopic behavior. c-d: The robot detects misspecification by estimating the human's apparent irrationality – as the human is avoiding the coffee spill which the robot does not know about, the human appears irrational, the robot starts making higher variance predictions, and its plans automatically give the person more room.

> *If the person appears irrational under the robot's assumptions, that simply means the robot has the wrong assumptions about the person.*

We have applied this idea to assistance, where the robot detects that the human's demonstrations or corrections cannot be explained by the set of features it currently has access to [14]. We also applied it to coordination, where this estimation naturally leads to higher variance predictions when the model is wrong [13]. What happens next is particularly exciting: rather than having to somehow intervene and heuristically make the robot more conservative, these higher variance predictions – modeling a person who appears less rational to the robot – *automatically* lead to plans where the robot gives the human more space, at least until their behavior starts making sense to the robot's model again.

[3. Human and robot simultaneous actions] The game-theoretic approach to interaction

The robot's actions and even its very existence influence human behavior, because people make different decisions when they interact than when they act in isolation. It is tempting to address coordination by predicting what people would do in isolation, and having robots stay out of the way. But that leads, for instance, to cars failing to merge on a highway because they can't get into their target lane, unable to realize that their actions can actually influence what people do.

> *Robots can turn seemingly infeasible plans into feasible ones if they account for the mutual influence between their actions and the humans'.*

We realized this mutual influence can is best characterized by a general sum dynamic game, and developed several approximations (static Stackleberg [15], hierachical decomposition [16], a game-theoretic equivalent of iLQR [17]). These enabled robots to figure out beautiful coordination strategies, like how to negotatiate merges, or that backing up at an intersection makes it more likely for the other driver to proceed first.

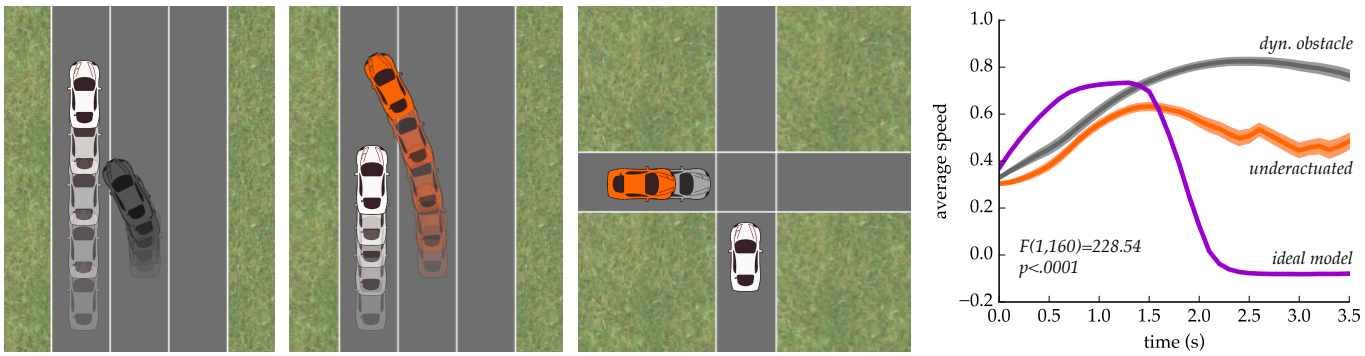
Similarly for assistance, people don't behave as if they are alone:

> *People are aware that the robot is observing them, trying to learn what they want – so* they try to be informative*! In turn, robots should not be stuck as passive observers, but can* leverage their physical actions to speed up learning.

Our idea was to formalize these phenomena via a common payoff game between the human and the robot, in which only the human can access the parameters of the shared utility – we call this the "assistance game" [18]. We've shown how to approximate it with an exponential reduction in complexity via a new Bellman update [19, 20], and we've also shown how robots can explicitly seek information gain via their physical actions, i.e. make "embodied queries": we've gotten cars to invent strategies like inching forward at intersections or nudging into a lane to probe someone's driving style [21], which otherwise would be handcrafted; and we've gotten manipulators to hold objects in a way such that the person reveals their ergonomic preferences when reaching for them [22].

[4. The Engineer as the Human] AI-assisted reward design:

Finally, there is a question of who is the human in this framework. Even for robots that are not meant to interact with end-users, a human – the engineer – still needs to specify the robot's objective (reward, cost, goal, constraints,



(a) Treating the Human as a Dynamic Obstacle

(b) Planning in the Underactuated System

(c) The Robot Inches Backwards to Get the Human to Cross First

(d) User Study: The Robot Does Affect the Human's Speed as Predicted

Figure 4: By treating interaction from a game-theoretic lens and finding approximations that work in real-time, cars leverage their influence on human actions: they merge in front of someone in heavy traffic knowing that they can slow down to let them in, and decide to back up at the intersection to get the human to proceed faster. These strategies emerge automatically out of the robot's optimization.

loss, etc.). Through my experience in both academia and industry on a variety of applications, I've come to realize that for any interesting enough problem we have no idea how to specify the right objective. First, it's always an iterative process. Second, even once we've iterated, we still get it wrong – we can never anticipate every single environment the robot will face and make sure the objective incentivizes the right behavior everywhere. With this realization,

> *We proposed that the specified objective should merely be evidence about, rather than the definition of, the true objective, to be interpreted only within the context it was specified for [23].*

Our algorithm enabled robots to learn from the specified objective, but maintain uncertainty about what they should optimize for in new environments and implicitly know what they don't know. This has lead to better test-time performance in arm motion planning [24] and autonomous driving. We have also closed the loop with the designer, leveraging the uncertainty to make queries about what the reward should be in hypothetical synthesized environments to narrow in on what they actually want.

References

[1] A. Bajcsy, D. Losey, M. O'Malley, and A.D. Dragan. Learning robot objectives from physical human interaction. In *Conference on Robot Learning (CoRL)*, 2017. **(oral), acceptance rate 10%**.

[2] A. Bajcsy, D. Losey, M. O'Malley, and A.D. Dragan. Learning from physical human corrections, one feature at a time. In *International Conference on Human-Robot Interaction (HRI)*, 2018.

[3] D. Hadfield-Menell, A.D. Dragan, P. Abbeell, and S. Russell. The off-switch game. In *International Joint Confernece on Artificial Intelligence (IJCAI)*, 2017.

[4] R. Shah and. Krasheninnikov, J. Alexander, P. Abbeel, and A.D. Dragan. Preferences implicit in the state of the world. In *International Conference on Learning Representations (ICLR)*, 2019.

[5] D. Sadigh, A.D. Dragan, S. Sastry, and S. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems (RSS)*, 2017.

[6] C. Basu, Singhal M, and A.D. Dragan. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *International Conference on Human-Robot Interaction (HRI)*, 2018.

[7] H.J. Jeon, S. Milli, and A.D. Dragan. Reward-rational (implicit) choice: a unifying formalism for reward learning. In *in review*, 2020.

[8] R. Shah, N. Gundotra, P. Abbeel, and A.D. Dragan. Inferring reward functions from demonstrators with unknown biases. In *International Conference on Machine Learning (ICML)*, 2019.

[9] S. Reddy, A.D. Dragan, and S. Levine. Where do you think you're going? inferring beliefs about dynamics from behavior. In *Neural Information Processing Systems (NeurIPS)*, 2018.

[10] A. Bestick, R. Bajcsy, and A.D. Dragan. Implicitly assisting humans to choose good grasps in robot to human handovers. In *International Symposium on Experimental Robotics (ISER)*, 2016.

[11] L. Chan, D. Hadfield-Menell, S. Srinivasa, and A.D. Dragan. The assistive multi-armed bandit. In *International Conference on Human-Robot Interaction (HRI)*, 2019.

[12] A. Bobu, D. Scobee, S. Satry, and A.D. Dragan. Less is more: Rethinking probabilistic models of human behavior. In *International Conference on Human-Robot Interaction (HRI)*, 2020. **(best paper award)**.

[13] J. Fisac, A. Bajcsy, D. Fridovich, S. Herbert, S. Wang, C. Tomlin, and A.D. Dragan. Probabilistically safe robot planning with confidence-based human predictions. In *Robotics: Science and Systems (RSS)*, 2018. **(invited to special issue)**.

[14] A. Bobu, A. Bajcsy, J. Fisac, and A.D.Dragan. Learning under misspecified objective spaces. In *Conference on Robot Learning (CoRL)*, 2018. **(invited to special issue)**.

[15] D. Sadigh, S. Sastry, S. Seshia, and A.D. Dragan. Planning for autonomous cars that leverages effects on human drivers. In *Robotics: Science and Systems (R:SS)*, 2016. **(invited to special issue)**.

[16] J. Fisac, E. Bronstein, E. Stefansson andD. Sadigh, S. Sastry, and A.D. Dragan. Hierarchical game-theoretic planning for autonomous vehicles. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[17] D. Fridovich-Keil, E. Ratner, A.D. Dragan, and C. Tomlin. Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum games. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[18] D. Hadfield-Menell, A.D. Dragan, P. Abbeell, and S. Russell. Collaborative inverse reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2016.

[19] D. Malik, M. Palaniappan, J. Fisac, D. Hadfield-Menell, S. Russell, and A. D. Dragan. An efficient, generalized bellman update for cooperative inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018. **(oral)**.

[20] J. Fisac, M. Gates, J. Hammrick, C. Liu, D. Hadfield-Menell, S. Sastry, T. Griffiths, and A.D. Dragan. Pragmatic-pedagogic value alignment. In *International Symposium on Robotics Research (ISRR)*, 2017. **(best bluesky paper award finalist)**.

[21] D. Sadigh, S. Sastry, S. Seshia, and A.D. Dragan. Information gathering actions over human internal state. In *International Conference on Intelligent Robots and Systems (IROS)*, 2016. **(best cognitive robotics paper award finalist)**.

[22] A. Bestick, R. Panya, R. Bajcsy, and A.D. Dragan. Learning human ergonomic preferences for handovers. In *International Conference on Robotics and Automation (ICRA)*, 2018.

[23] D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell, and A.D. Dragan. Inverse reward design. In *Neural Information Processing Systems (NIPS)*, 2017. **(oral, acceptance rate 1.2%)**.

[24] E. Ratner, D. Hadfield-Menell, and A.D. Dragan. Simplifying reward design through divide-and-conquer. In *Robotics: Science and Systems (RSS)*, 2018.

[25] M. Kwon, S. Huang, and A.D. Dragan. Expressing robot incapability. In *International Conference on Human-Robot Interaction (HRI)*, 2018. **(best paper award finalist)**.

[26] J. Fisac, C. Liu, J. Harick, K. Hedrick, S. Sastry, T. Griffiths, and A.D. Dragan. Generating plans that predict themselves. In *Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2016.

[27] S. Huang, P. Abbeel, and A.D. Dragan. Enabling robots to communicate their objectives. In *Robotics: Science and Systems (RSS)*, 2017. **(invited to special issue)**.

[28] A. Zhou, D. Hadfield-Menell andA. Nagabaudi, and A.D. Dragan. Expressive robot motion timing. In *International Conference on Human-Robot Interaction (HRI)*, 2017.