Legibility-Aware Learning from Corrections

Anjiabei Wang
anjiabei.wang@yale.edu
Yale University
New Haven, CT, USA

Tesca Fitzgerald
tesca.fitzgerald@yale.edu
Yale University
New Haven, CT, USA

ABSTRACT

Corrections offer an easy way for end-users to teach and collaborate with a robot, while also offering rich information about task constraints. However, these corrections reflect more than just the optimality of the robot behavior, and are subject to additional influences such as task tolerance, physical effort, and the human's subjective expectation of whether the robot will succeed. We present a predictive model of corrections that accounts for the impact of these factors. We propose a user study to collect empirical data to study how robot's behavior influences when and how humans will modify it. Finally, we discuss how our predictive model can help robots learn more effectively from these corrections.

CCS CONCEPTS

• **Computer systems organization** → **Robotics**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

KEYWORDS

Corrections, Inverse Reinforcement Learning, Legibility

ACM Reference Format:
Anjiabei Wang and Tesca Fitzgerald. 2024. Legibility-Aware Learning from Corrections. In *Proceedings of Workshop on Human-Interactive Robot Learning (HRI '24).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

With the continued integration of machines into our everyday lives, the principle of non-technical human teachers being able to effectively communicate with and efficiently train robots becomes increasingly relevant. Current and prior research has looked into how people can train a robot to complete manipulation-based tasks using different modalities or interaction types, such as demonstrations [2] and ranked preferences [7]. Alternatively, a person can monitor a robot as it attempts to complete a task, interceding to provide a *correction* when they deem it necessary to modify the robot's behavior [3, 10]. For example, if a robot that is supposed to pick up a mug from the table is moving away from the table instead, a human teacher may offer assistance by correcting the robot's motion and pushing it in the right direction. This correction should inform how the robot behaves in future variations of the task, while also implying how the robot should *not* behave (i.e., the behavior that prompted the teacher to intercede in the first place).

HRI '24, March 11–15, 2024, Boulder, CO

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

From a physical human-robot interaction perspective, corrections are a natural method for collaboration and communication between humans and robots [15]. For non-technical users, it is easier to express their intent by directly interacting with the robots, as opposed to programming them. Furthermore, corrections put the human in the role of a supervisor rather than teacher, where they only step in as needed rather than try to teach the robot from scratch. This reduces the amount of data required to improve task performance and increases the efficiency of robot learning [17].

From a ML perspective, corrections have the potential to provide rich information for a robot to learn an optimal model of the task objective. Based on the initial behavior of the robot, a human's correction of that behavior can indicate what the robot did right or wrong, and how to avoid making similar mistakes in the future. However, corrections are complicated to interpret. Bayesian Inverse Reinforcement Learning (BIRL) provides a method for learning a reward function that maximizes the likelihood of the teacher's feedback [6, 15, 18]. Yet, this approach requires that we have a model of how the human feedback is influenced by the task objectives. When learning from corrections, there are other conflating influences such as the *abruptness* of the corrections (caused by the human's binary decision of whether or not to correct the robot's behavior) and the subjective *bias* that influences the teacher's belief over whether or not the robot will succeed at the task. This decision to interrupt and modify the robot's motion may occur after the robot has made a mistake or in anticipation of a future mistake that has not even occurred yet, and may be biased by the robot's previous behavior [16]. Prior work has focused on interpreting and learning from corrections [5, 6, 15], but has not looked into how the teacher decides to intervene and provide a correction in the first place.

The key challenge we address is how to isolate the influence of task objectives on corrections from other conflating influences. To do this, we need to first model the effects of these influences on corrections. In this abstract, we hypothesize the effects of three variables (legibility, task tolerance, and physical effort) on corrections and formalize them as predictive models. We then propose a user study to evaluate these models. Finally, we discuss how these models can be used to improve ML algorithms in future work.

2 RELATED WORK

To provide *corrections*, a human user monitors the robot as it attempts to complete a task. Throughout the robot's motion, the user may physically interact with the robot in order to help it to perform the task better (i.e., nudging the robot arm in the right direction). While the robot in in motion, corrections can be provided at regular intervals [10] or at any time the human chooses to intervene [6, 15]. These modifications are recorded as the torque that participants exert on the robot arm, the resulting changes in the robot's position and movement, and the timing of this interaction.

Using this data, corrections are typically interpreted as the teacher indicating a preference for the modified trajectory over the robot's originally-planned one, and that this preference is due to the optimality of the modified trajectory [6, 9, 15]. In [6, 15], the robot updates its reward model immediately in order to change its future behavior for the rest of the task. These prior works focus on how the robot should interpret the correction based on how they reflect the teacher's intentionality [15] and physical effort [6], but they do not model how or when the human teacher chooses to correct the robot's motion in the first place. Thus, for our study, we would like to define a probability model of corrections that captures how the robot's behavior within a task affects the mental model and expectations of the teacher for giving the corrections.

2.1 Bayesian Inverse Reinforcement Learning

Bayesian Inverse Reinforcement Learning (BIRL) involves inferring the most probable reward function based on its expected influence on the policy demonstrated by a teacher [19, 21]. For a collection of human feedback (trajectories) Ξ, we can formalize the optimal reward parameters ω for a reward function $R_\omega(\xi)$ as:

$$\omega^* = \arg\max_{\omega} \prod_{\xi \in \Xi} P(\omega|\xi) = \arg\max_{\omega} P(\omega) \prod_{\xi \in \Xi} \frac{P(\xi|\omega)}{P(\xi)} \quad (1)$$

Within this formulation, $P(\xi|\omega)$ is particularly important, as it models how a human provides feedback ξ based on their understanding of the task objectives (represented by ω). The Boltzmann distribution is often used to represent this probability by framing the human's feedback as a choice that a "noisily rational" human makes from a set of possible choices (represented by C) [19, 21]:

$$P(\xi|\omega) = \frac{e^{R_\omega(\xi)}}{\sum_{\xi \in C} e^{R_\omega(\xi)}} = \frac{e^{\beta \cdot \phi(\xi) \cdot \omega}}{\sum_{\xi \in C} e^{\beta \cdot \phi(\xi) \cdot \omega}} \quad (2)$$

where $\phi(\xi)$ returns the feature trace of the trajectory ξ. Depending on the interaction modality being used [9, 12], C may consist of other trajectories that the human could have demonstrated [9], other preferences that the human could have selected from a set of options [4], or other ways that the human could have corrected the robot's motion [3, 15]. Importantly, the reward function R_ω is scaled by β, which represents the expected optimality of the human feedback (i.e., how well it adheres to R_ω). When $\beta \to 0$, the human's feedback is independent of ω and thus will be chosen based on a uniform distribution. When $\beta \to \infty$, the teacher will only ever indicate the highest-reward choice according to R_ω.

This probability formulation assumes that the human's feedback is determined only by (1) the task objectives that the robot is trying to learn and (2) the optimality of the teacher. However, due to the complex and abrupt features of corrections, other factors may influence how people provide corrections as feedback. The question becomes: how can we more accurately model the probability of corrections by considering these additional influences?

3 ADDITIONAL INFLUENCES ON CORRECTIONS

We expect that determining *when* and *how* people correct robot behavior can provide valuable insights, in addition to the optimality of the resulting trajectory. In this section, we outline two factors that may influence how people provide corrections.

3.1 Task Tolerance

In the commonly-used Boltzmann distribution (Eq. 2), β represents the expected optimality of the teacher's feedback with respect to the reward function R_ω. Alternatively, it can be interpreted as the teacher's tolerance for non-optimality in their feedback. In either interpretation, β is a constant variable for the teacher, and does not reflect how different tasks involve different tolerance levels based on the task constraints. We propose that corrections are provided according to *task-specific* tolerances corresponding to each feature: $\omega = \langle < \omega_1, \beta_1 >, \cdots, < \omega_n, \beta_n > \rangle$. The various tolerance β_i for each constraint i indicates the distribution of each ω_i, and maps the trajectory ξ to a distribution of rewards: $R_\omega(\xi) = \phi(\xi) \cdot \omega^T$.

3.2 Human Expectation of Robot Behavior

In prior work on learning from corrections, the robot *always* requires a correction at some point in its motion. We are unaware of any studies of how people decide *whether* to intercede in the first place. We expect that people make this decision by observing the robot's behavior and constantly updating their belief over whether the robot will succeed or fail at the task. Prior studies on humans' trust in robots [13] and mental models of robot behavior and capabilities [8, 16] indicate that this belief can be influenced by the robot's reliability during its prior performance. We expect that the robot's competency in the task (i.e., the frequency at which it has previously succeeded at the task) will influence people's trust in it and thus the frequency with which they provide corrections.

Additionally, trustworthiness can be influenced by the human's confidence in inferring the goal of a robot as they observe the robot's motions [1, 11, 14, 20]. Performing highly-legible motions enables a robot to more clearly indicate its intentions, which increases the human's ability to predict the robot's future actions [8]. Based on this, we expect that legibility will also improve the human teacher's ability to determine whether they need to intervene and correct a robot's motions to prevent it from failing at the task. Overall, we propose that the robot's legibility and competency level influence the human teacher's expectation of the robot's success, and further influence how quickly they will intervene to provide a correction.

4 APPROACH

We now propose a model of how task tolerance and human expectations influence corrections. Our model is based on two hypotheses:

- **H1**: We can model corrections more accurately by separately representing *when* and *how* people provide them.
- **H2**: To model *when* people provide corrections, the legibility of the robot's motion has a positive correlation with how quickly people will correct the robot's behavior.

We start by representing the robot's total motion as a combination of its pre-correction and post-correction trajectories:

$$P(\xi_H|\omega, \xi_R) = P((\xi_{H1}, \xi_{H0})|\omega, \xi_R) \quad (3)$$

$$= P(\xi_{H1}|\xi_{H0}, \omega, \xi_R) P(\xi_{H0}|\omega, \xi_R) \quad (4)$$

where ω contains feature weights and corresponding tolerance constraints; ξ_R is the robot's original trajectory that it attempted to

execute; ξ_H is the actual trajectory that the robot executed (including the human-provided correction). We divide ξ_H into ξ_{H0} and ξ_{H1}, which correspond to the part of the trajectory before and after intervention, respectively. The correction probability now consists of the product of $P(\xi_{H0}|\omega, \xi_R)$ (representing the probability of *when* to intervene) and $P(\xi_{H1}|\xi_{H0}, \omega, \xi_R)$ (representing the probability of *how* to correct the robot's motion after intervening).

4.1 Modeling *when* people correct

We propose modelling *when* people correct the robot as their perceived probability of the robot completing the task successfully, based on the robot's actions prior to the current timestep. This depends on what it means for the robot to be successful (defined by the task tolerance) and the human's expectation of the robot's future behavior (defined by the legibility of the robot's motion):

$$P(\xi_{H0}|\omega, \xi_R) \propto \int_{s_g} P_H(s_g|\xi_{H0}) R_\omega(s_g) \tag{5}$$

$P_H(s_g|\xi_{H0})$ describes the probability of the human inferring the robot's goal state s_g given the trajectory ξ_{H0}, which should be directly related to the legibility of the robot motion, the competency level of the robot, and the task constraints. We will use the legibility score formulated by [8]:

$$legibility(\xi) = \frac{\int P(G^*|\xi_{S\to\xi(t)}) f(t) dt}{\int f(t) dt} \tag{6}$$

The score for legibility tracks the probability assigned to the actual goal G^* across the trajectory: the more legible a trajectory is the higher the probability is, weighted by a function $f(t)$ giving more weight to the earlier part of the trajectory. $S \to \xi(t)$ indicates the trajectory from the start S until the timestep t.

4.2 Modeling *how* people correct

We propose modelling *how* people correct the robot as a function of the task tolerance and the physical effort of applying that correction to the robot. We adapt the effort-based probability function posed by [6] as follows:

$$P(\xi_{H1}|\xi_{H0}, \omega, \xi_R) = P(u_H|\xi_R; \omega = \langle < \omega_1, \beta_1 >, \cdots, < \omega_n, \beta_n > \rangle) \tag{7}$$

$$= \frac{e^{-(\omega^T \phi(\xi_H) + \lambda ||u_H||^2)}}{\int e^{-(\omega^T \phi(\bar{\xi_H}) + \lambda ||\bar{u_H}||^2)} d\bar{u_H}}, \tag{8}$$

Here, ω is the vector that represents the task tolerance for each feature, u_H represents the torques that the teacher exerts on the robot in order to correct its motion, and λ reflects the degree to which physical effort compromises the optimality of the teacher's feedback. We expect λ to be consistent within each person from one task to the next.

4.3 User Study Design

We propose a user study in which we measure *when* and *how* people correct a robot's motion during a series of pick-and-place tasks. The task objective is for the robot to place a block of particular color and shape into its corresponding hole. The robot will then move slowly as it attempts to complete the task, and participants may choose to interrupt or modify the robot's motion as it moves in order to help it succeed.

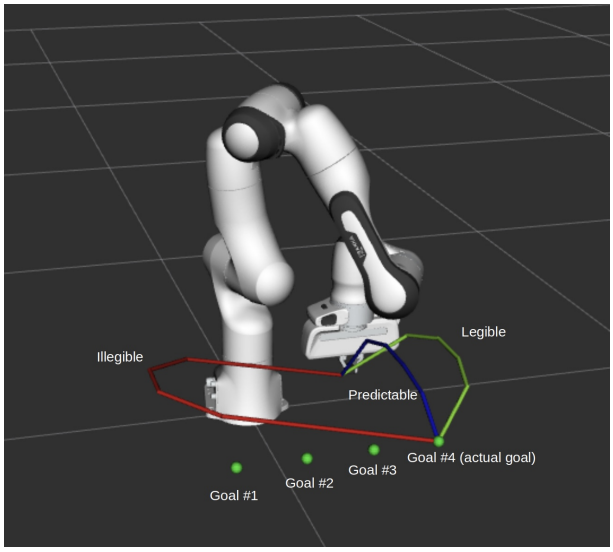

Figure 1: Examples of a legible trajectory (green), predictable trajectory (blue), and illegible trajectory (red) for a robot moving to goal #4.

We plan this as an across-participants study consisting of 12 conditions: 3 legibility conditions (legible, illegible, and predictable motion, as shown in Fig. 1) x 4 competency conditions (consistently poor performance, consistently high performance, improving performance, and worsening performance). Within each condition, the robot performs variations of the block-placing task that differ in their tolerance (i.e., how constrained the block's position and orientation is with respect to the target hole).

Our data analysis will involve fitting and evaluating Eqs. 5 and 8 according to these condition parameters. During the process we will record the timing of the interactions, the torque that participants exert on the robot arm, and the resulting changes in the robot's position and movement. In future work, we will evaluate the effect of using these models for Bayesian IRL in comparison to baselines reflecting traditional Boltzmann-rational models for correction data and train more effective Machine Learning algorithms that produce better future robot behavior.

5 CONCLUSION

Corrections have the potential to provide valuable information about how robots should and should not complete tasks. Yet, they must be carefully interpreted due to their abruptness (based on the teacher's decision to interrupt and correct the robot's motion) and bias from the robot's initial motion and previous performance. We have proposed two important factors that may influence when and how people correct robot behavior: (1) task tolerance and (2) the human's expectations of whether the robot will succeed. We expect that by empirically modelling the effects of these factors on corrections, we can develop more effective robot learners.

REFERENCES

[1] Rachid Alami, Aurélie Clodic, Vincent Montreuil, Emrah Akin Sisbot, and Raja Chatila. 2006. Toward Human-Aware Robot Task Planning.. In *AAAI spring symposium: to boldly go where no human-robot team has gone before*. 39–46.

[2] Christopher G Atkeson and Stefan Schaal. 1997. Robot learning from demonstration. In *ICML*, Vol. 97. Citeseer, 12–20.

[3] Andrea Bajcsy, Dylan P Losey, Marcia K O'malley, and Anca D Dragan. 2017. Learning robot objectives from physical human interaction. In *Conference on Robot Learning*. PMLR, 217–226.

[4] Erdem Bıyık, Malayandi Palan, Nicholas C Landolfi, Dylan P Losey, and Dorsa Sadigh. 2019. Asking easy questions: A user-friendly approach to active reward learning. *arXiv preprint arXiv:1910.04365* (2019).

[5] Andreea Bobu, Andrea Bajcsy, Jaime F Fisac, Sampada Deglurkar, and Anca D Dragan. 2020. Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *IEEE Transactions on Robotics* 36, 3 (2020), 835–854.

[6] Andreea Bobu, Andrea Bajcsy, Jaime F Fisac, and Anca D Dragan. 2018. Learning under misspecified objective spaces. In *Conference on Robot Learning*. PMLR, 796–805.

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

[8] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.

[9] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. 2023. INQUIRE: INteractive querying for user-aware informative REasoning. In *Conference on Robot Learning*. PMLR, 2241–2250.

[10] Tesca Fitzgerald, Elaine Short, Ashok Goel, and Andrea Thomaz. 2019. Human-guided trajectory adaptation for tool transfer. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1350–1358.

[11] Michael J Gielniak and Andrea L Thomaz. 2011. Generating anticipation in robot motion. In *2011 RO-MAN*. IEEE, 449–454.

[12] Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems* 33 (2020), 4415–4426.

[13] Zahra Rezaei Khavas, S Reza Ahmadzadeh, and Paul Robinette. 2020. Modeling trust in human-robot interaction: A survey. In *International conference on social robotics*. Springer, 529–541.

[14] Christina Lichtenthäler, Tamara Lorenz, and Alexandra Kirsch. 2011. Towards a legibility metric: How to measure the perceived value of a robot. In *International Conference on Social Robotics, ICSR 2011*.

[15] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. 2022. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research* 41, 1 (2022), 20–44.

[16] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*. PMLR, 2285–2294.

[17] Çetin Meriçli, Manuela Veloso, and H Levent Akın. 2011. Task refinement for autonomous robots using complementary corrective human feedback. *International Journal of Advanced Robotic Systems* 8, 2 (2011), 16.

[18] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. 2017. Should robots be obedient? *arXiv preprint arXiv:1705.09990* (2017).

[19] Deepak Ramachandran and Eyal Amir. 2007. Bayesian Inverse Reinforcement Learning.. In *IJCAI*, Vol. 7. 2586–2591.

[20] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*. 69–76.

[21] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.