Demonstration and Imitation of Novel Behaviors under Safety Aware Shared Control

Alexander Broad*, Todd Murphey†, and Brenna Argall*†
*Department of Electrical Engineering and Computer Science
†Department of Mechanical Engineering
Northwestern University, Evanston, IL 60208
Email: alex.broad@u.northwestern.edu

***Abstract*—We describe a shared control algorithm that improves both a human operator's ability to provide demonstrations, and a learning algorithm's ability to recreate the novel behavior. Our method introduces an autonomous agent that assists the human partner by enforcing safety and stability constraints. The autonomous agent has no *a priori* knowledge of the desired task and therefore only interferes when there is concern for the safety of the system. We evaluate the impact of our shared control algorithm on a user's ability to provide successful demonstrations in a variety of environments with a human subject study consisting of 20 participants. We then use the collected demonstration data to train a neural network policy through simple behavior cloning. A preliminary evaluation reveals that the continued application of the safety aware shared control algorithm is integral in producing autonomous policies that successfully mimic the desired behavior. We discuss limitations and future work in the conclusion.**

I. Introduction

To improve the acceptance and efficacy of robotic devices in human environments, we must design autonomous agents that are capable of learning novel behaviors without explicit programming or interference from a robotics expert. This is particularly important when we consider the class of robotic devices that are designed explicitly for interaction with a human partner. For example, mechanical devices in assistive and rehabilitation medicine can be used to help restore lost functionality to people suffering from motor control disorders or impairments due to physical injury. By designing robots that can grow in functionality alongside their human partner, we offer greater freedom to the person in need.

Learning from Demonstration (LfD) is a paradigm that solves this problem by enabling a robotic partner to reproduce novel behaviors demonstrated by a human operator [3, 5]. One challenge not currently addressed in the LfD literature is how to help users provide high-fidelity demonstrations when there are difficulties related to the complexity of the control problem, the complexity of the task, or limitations due to the skill of the human partner. We address this problem by allowing users to provide demonstrations under *shared control*.

In general, shared control is a paradigm that can be used to produce joint human-machine systems that are more capable than either the human or machine on their own [11]. In this work, we are interested in offloading challenging aspects of the control problem to an autonomous partner to *enable human operators to provide demonstrations when they would otherwise not be able to.* This is particularly important when the dynamic system is inherently challenging to control (e.g., unstable systems like exoskeletons) or when it is easy to provide demonstrations that may be dangerous for the human or environment (e.g., navigation through a dense crowd). We therefore define a shared control system that allows a human operator the freedom to demonstrate desired motions while the autonomous partner ensures safety constraints (see Section III). The provided demonstrations can then be used to train an autonomous control policy that is capable of recreating the desired motion [16]. Importantly, we demonstrate how the structural knowledge encoded in the safety aware autonomous controller can be used to improve the efficacy of simple behavior cloning techniques in recreating learned system behavior (see Section IV). Finally, we conclude with a discussion of limitations and future work in Section V.

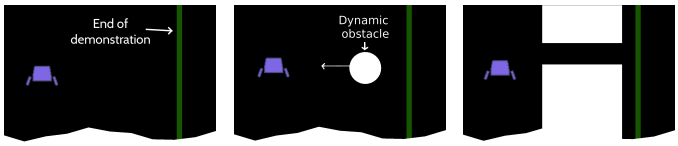

Fig. 1. Visualization of the experimental environments and task. A demonstration is considered complete when the lander moves outside the boundary defined by the green line. The lunar lander is enlarged for visualization.

II. Experimental System

We begin by describing the experimental system, a simulated "lunar lander" (see Figure 1) from OpenAI's Gym testbed [8]. The lunar lander is defined by a six dimensional vector, $\boldsymbol{x}$, in which the first three components (x_{1-3}) define the 2D position and heading, while the final three components (x_{4-6}) define their rates of change. The control input, $\boldsymbol{u}$, is a continuous two dimensional vector which represents the throttle of the main (u_1) and rotational thrusters (u_2). This system was chosen to demonstrate the impact that shared control can have on the safety of demonstrations when the control problem and environment are complex. The lunar lander exhibits nonlinear dynamics and can easily become unstable as it rotates away from its point of equilibrium.

III. Safety Aware Shared Control

We now describe our method for allowing human operators to provide demonstrations of novel behaviors *under a safety aware shared control paradigm*.

A. Model-based Shared Control

To implement a shared control paradigm we need to define a method for computing the policy of the autonomous agent and a method for dynamically allocating control between the two partners. In this work, we use model-based optimal control [4, 18] (MBOC) to compute autonomous policies. MBOC uses a model of the system dynamics learned directly from data [1, 12], which is then incorporated into an optimal control algorithm to produce autonomous policies [2]. To compute policies that are specifically concerned with the safety of the dynamic system we define a cost function based on the underlying structure of the system and task. In particular, we consider two notions of safety: stability around points of equilibrium and collision avoidance. Therefore we define a cost function that (1) penalizes states that are far from points of equilibrium using a quadratic cost, and (2) penalizes the system from entering dangerous portions of the state space using polynomial barrier functions [6].

To close the loop in our shared control system, we must define a dynamic allocation method that intelligently integrates the control provided by each partner. If the control allocation method is too permissive of the human operator, it may do a poor job enforcing the necessary safety requirements. However, if the control allocation method is too stringent, it can negatively impact the ability of the human operator to produce their desired motion. In this work, we use a modified version of Maxwell's Demon Algorithm (MDA) [17] to allocate control authority. At a high-level MDA uses the output of an optimal control algorithm as a guide by which to evaluate the input from a human partner [7, 9]. We define a modified, safety aware MDA in Algorithm 1.

Algorithm 1 Safety Aware Maxwell's Demon Algorithm

```
if δ(system, object) < ϵ then
    u = u_a;
else
    if ⟨u_h, u_a⟩ ≥ 0 then
        u = u_h;
    else
        u = 0;
    end if
end if
```

where δ is a function that computes the distance to the nearest obstacle, ϵ is a hand-selected distance threshold, $\langle \cdot, \cdot \rangle$ is the inner product, $\boldsymbol{u_h}$ is the input from the human operator, $\boldsymbol{u_a}$ is the input produced by the autonomy, and $\boldsymbol{u}$ is the control applied to the dynamic system. Under this paradigm, if the system is deemed to be in a dangerous state (i.e. too close to an obstacle), the autonomous partner's signal is used to control the system. If the system is outside of the distance threshold ϵ, and the user's input is close enough to the input computed by the autonomy (i.e. *safe enough*), the user's input is used to control the system. In all other cases, zero input is passed to the dynamic system. This algorithm is defined to balance the control authority given to the human and autonomous partners such that the human partner can provide demonstration data, while the autonomous partner improves safety and stability.

B. Experimental Evaluation

To evaluate the efficacy of our shared control paradigm, we performed a human subjects study consisting of 20 total participants (16 female, 4 male). Each subject was asked to provide 10 demonstrations of desired behaviors in three experimental environments (see Figure 1) under both a *user only control* paradigm and the defined *shared control* paradigm. There was no goal location specified to the participants, instead a demonstration was considered complete when the human operator navigated the lunar lander outside of a barrier defined by a green line in the environment. The first environment included only the lunar lander and the ground surface. This environment illuminates the challenges associated with maintaining the stability of a complex dynamic system, while simultaneously providing demonstrations of a new behavior. The second environment incorporated dynamic obstacles that obstructed the motion of the system. In this environment, a series of circular obstacles moved directly across the screen at the same height as the lunar lander (one at a time). The third environment included two static obstacles that forced the operator to navigate through a narrow passageway, thereby increasing the required control fidelity.

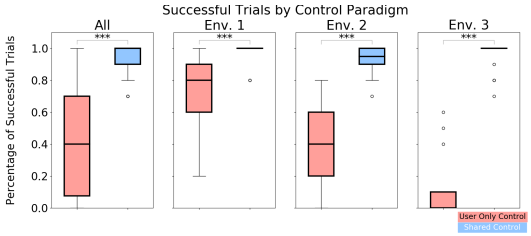

Fig. 2. Average number of successful trials under each control paradigm in each environment. We find that participants under the safety aware shared control paradigm provide successful demonstrations significantly more often then participants under the user only control paradigm ($p < 0.005$).

C. Average Success Rate of Demonstration

The primary metric we evaluate in this study is the ability of the human operator to navigate the lunar lander beyond the green border indicating the successful demonstration of a desired behavior. This metric can be used as a simple binary indicator of control competency. We therefore compare how often users are successful under each control paradigm. We use the non-parametric Wilcoxon signed-rank test to statistically analyze the results. We display pertinent metrics of the data and results of the described statistical tests in Figure 2. The results of the statistical tests reveal that our described shared control paradigm significantly improves the human partner's ability to provide safe demonstrations of a desired behavior.

IV. Learning from Demonstration

We now show that we can use the demonstration data collected under shared control to produce autonomous control policies that imitate behaviors exhibited by human partners [13, 15]. Our goal, then, is to learn an autonomous policy

$$\pi_a^*(s) = \arg\min \int_{s \in S} ||\pi_a(s) - \pi_h(s)||_2 \, ds \qquad (1)$$

where s is the state, $\pi(s) : s \rightarrow u$ defines a control policy, π_h represents the human's policy and π_a represents the autonomy's policy. To generate π_a, we define a neural network which we train on the successful demonstration data using a behavior cloning objective [13]. There are, however, well known problems with autonomous policies trained using vanilla behavior cloning techniques. One particularly common issue is that the data used to train the policy may come from a different distribution then the data observed at runtime [14]. We address this issue by combining the learned neural network model with the shared control paradigm described in Section III. That is, by incorporating the same shared control algorithm used during data collection, we encourage the system to operate in a similar distribution of the state space to what was observed during the demonstration phase. One can view this solution as a shared control paradigm in which the *control is shared between two autonomous agents* : the autonomy mimicking the human control and the autonomy enforcing safety constraints.

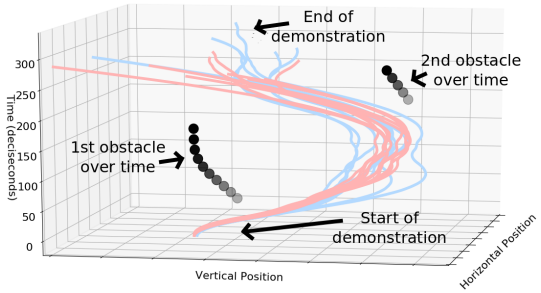

Fig. 3. A visualization of the user's demonstration data (blue) provided under our safety aware shared control algorithm, and trajectories generated autonomously (pink) using the learned control policy. This image displays the vertical and horizontal position of the system and obstacles (black) over time.

To evaluate the efficacy of our LfD approach, we examine whether the resulting policies are capable of safely reproducing the behavior demonstrated by the human operator under shared control. Figure 3 is a visualization of data collected from Environment 2. Figure 4 is a visualization of data collected from Environment 3. In both figures, we see that the learned control policy, *operating under our safety aware shared control algorithm*, is able to mimic the behavior demonstrated by the human operator under the same shared control paradigm. In fact, all trajectories produced autonomously under this paradigm safely avoid both the dynamic and static obstacles. Additionally, in Figure 4 we see that the user was unable to provide any successful demonstrations

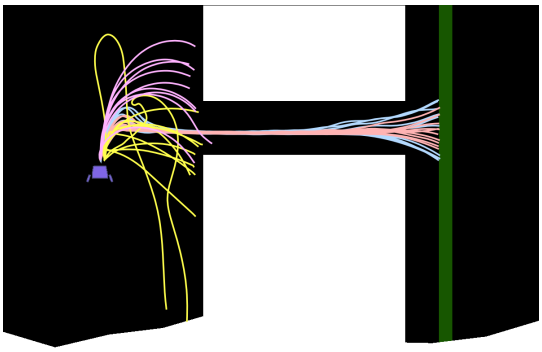

Fig. 4. A visualization of (1) user demonstration data provided under our safety aware shared control paradigm (blue), (2) user demonstration data provided under the user only control paradigm (yellow), (3) trajectories produced autonomously using the learned shared control policy (pink), and (4) trajectories produced autonomously using solely the behavior cloning policy (purple). All data is overlaid on Environment 3.

without safety assistance. Similarly, the learned control policy was unable to avoid obstacles in the environment without the safety assistance. *These two final points elucidate the need for our safety aware shared control system in both the demonstration and imitation phases.* Without assistance from an environment aware shared control algorithm, the human operator is unable to demonstrate desired behaviors, and the learned neural network policy fails to generalize.

V. Conclusion and Limitations

The results of our human subjects study show that our safety aware shared control paradigm is able to help human partners provide demonstrations of novel behaviors in situations in which they would otherwise not be able to (see Figure 2). An evaluation of our ability to use this data in a LfD paradigm demonstrates that the integration of structural knowledge based on the underlying system is *integral in the successful application of simple behavior cloning techniques*. Without this controller, the autonomous agent is unable to reproduce the motion trajectories demonstrated by the human operator.

We now discuss some of the limitations of this early work. The first limitation relates to the safety aware shared control algorithm. Ideally, our dynamic allocation algorithm would formally guarantee safety over the course of the entire interaction. However, with each additional constraint added to the system, we increase the complexity of the shared control algorithm and reduce the freedom afforded to the human operator. In future work we plan to explore notions of shared control that address this balance. A second limitation of this work relates to the learned autonomous policy. As we see in Figure 3, the successful application of the policy requires a continued integration with our safety aware shared control algorithm. In future work, we plan to explore methods that use the safety aware optimal controller as a supervisor in training more robust policies that can reproduce the behavior on the own. This idea is related to the use of an optimal control-based supervisor in Guided Policy Search [10].

References

[1] Pieter Abbeel, Morgan Quigley, and Andrew Y Ng. Using Inaccurate Models in Reinforcement Learning. In *International Conference on Machine Learning*, pages 1–8. ACM, 2006.

[2] Alexander R Ansari and Todd D Murphey. Sequential Action Control: Closed-form Optimal Control for Nonlinear and Nonsmooth Systems. *Transactions on Robotics*, 32(5):1196–1214, 2016.

[3] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A Survey of Robot Learning from Demonstration. *Robotics and Autonomous Systems*, 57 (5):469–483, 2009.

[4] Christopher G Atkeson and Juan Carlos Santamaria. A Comparison of Direct and Model-Based Reinforcement Learning. In *International Conference on Robotics and Automation*, volume 4, pages 3557–3564. IEEE, 1997.

[5] Christopher G Atkeson and Stefan Schaal. Robot Learning from Demonstration. In *International Conference on Machine Learning*, volume 97, pages 12–20, 1997.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] Alexander Broad, Todd Murphey, and Brenna Argall. Learning Models for Shared Control of Human-Machine Systems with Unknown Dynamics. In *Robotics: Science and Systems*, 2017.

[8] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv*, abs/1606.01540, 2016.

[9] Kathleen Fitzsimons, Emmanouil Tzorakoleftherakis, and Todd Murphey. Optimal Human-In-The-Loop Interfaces Based on Maxwell's Demon. In *American Control Conference*, pages 4397–4402, 2016.

[10] Sergey Levine and Vladlen Koltun. Guided Policy Search. In *International Conference on Machine Learning*, pages 1–9, 2013.

[11] Selma Musić and Sandra Hirche. Control Sharing in Human-Robot Team Interaction. *Annual Reviews in Control*, 2017.

[12] Duy Nguyen-Tuong and Jan Peters. Model Learning for Robot Control: A Survey. *Cognitive Processing*, 12(4): 319–340, 2011.

[13] Dean A Pomerleau. ALVINN: An Autonomous Land Vehicle in a Neural Network. In *Advances in Neural Information Processing Systems*, pages 305–313, 1989.

[14] Stéphane Ross, Geoffrey J Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 627–635, 2011.

[15] Claude Sammut, Scott Hurst, Dana Kedzier, and Donald Michie. Learning to Fly. In *Machine Learning Proceedings*, pages 385–393. Elsevier, 1992.

[16] Stefan Schaal and Christopher G Atkeson. Learning Control in Robotics. *Robotics and Automation Magazine*, 17(2):20–29, 2010.

[17] Emmanouil Tzorakoleftherakis and Todd D Murphey. Controllers as Filters: Noise-Driven Swing-Up Control Based on Maxwells Demon. In *Conference on Decision and Control*, 2015.

[18] Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information Theoretic MPC for Model-Based Reinforcement Learning. In *International Conference on Robotics and Automation*, 2017.