Many modern digital systems—from automotive vehicles to social media platforms—have an increasing ability to measure, store, and process data. Optimism about the potential to benefit from this data is driven by parallel progress in machine learning, where large datasets and computational power have led to advances on complex tasks like image recognition and machine translation. However, applying techniques developed for static datasets to real world problems requires grappling with the effects of dynamic feedback and systems that change over time. In these settings, classic statistical and algorithmic guarantees do not always hold. Even rigorously evaluating performance can be difficult. I work on analyzing and developing principled techniques for these dynamical settings, towards a vision of reliable machine learning.

My research draws on tools and concepts from **control theory**, which has a long history of formulating guarantees about the behavior of dynamical systems, **optimization**, which provides a language to articulate goals and tradeoffs, and **machine learning**, which uses data to understand and act on the world. In pursuing principled ways to use data from systems that change over time, I have studied interconnected problems, which can be broadly categorized into:

- **Guaranteeing safety in feedback control.** By combining machine learning and robust control, I have developed a framework for data-driven optimal control design with non-asymptotic performance guarantees. Considering settings where dynamics are unknown, I have used this framework to study the sample complexity of the linear quadratic regulator (LQR) [1], with extensions to safety constraints [2] and adaptive control [3]. Considering instead settings where challenges arise from complex sensing modalities, I have used this framework to develop guarantees for perception-based control of linear [4, 6] and nonlinear [5] systems.
- **Ensuring values in social-digital systems.** Consequences matter when learning algorithms interact with people. Predictive models alone do not ensure desirable outcomes, so I develop ways to incorporate and verify notions like fairness, well-being, and user agency. Motivated by consequential decision-making, I have rigorously characterized the impact of fairness constraints [8] and their connections to social welfare [9]. In the setting of content recommendation, I introduced a novel metric to characterize user agency in interactive systems [10].

While I have a strong theoretical focus, my research is enriched through collaborations across disciplines and in applications including robotics, computational imaging, and recommendation systems. I am both grounded and inspired by conversations with aerospace engineers, roboticists, economists, app developers, and social scientists. My activities as a founder of a transdisciplinary student group on technology and society have shaped my views on formulating technical research questions around important human values. I also appreciate the practical perspectives of industry: my work has been deployed at Canopy, a personalization startup, and in the Google ML Fairness Gym. Looking forward, I plan to bring the same combination of cross-disciplinary collaboration and industry interaction to my work as a professor.

Vision. I will pursue a research agenda aimed at realizing the promise of machine learning in modern digital systems, contending with challenges posed by dynamics and feedback. I look forward to leading a research group with a broad range of overlapping interests, blending strong theoretical foundations with a commitment to building computational and mathematical tools with meaningful impact in domains ranging from digital infrastructure to robotics.

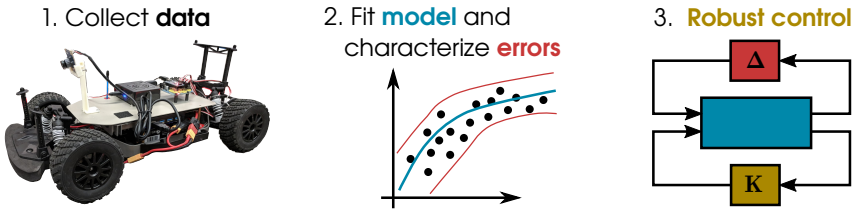

Figure 1: The data-driven optimal control framework has three steps.

ML for Safety in Feedback Control

Modern autonomous systems face challenges well-suited to a combination of feedback control and machine learning. To explore the risks and trade-offs that arise in such settings, I developed a data-driven optimal control framework, consisting of three steps: first, collect data from the system; second, fit a model of the uncertain component and characterize its potential errors; third, use the model to design a controller robust to these errors (Figure 1).

Unknown Dynamics. My work on developing theoretical foundations focuses on LQR, a classical optimal control problem that forms the backbone of several modern control techniques. Using the three step framework, I bounded the sample complexity of a data-driven approach to LQR [1], illustrating how the relationship between the number of datapoints and resulting performance depends on properties of the system to be controlled. While the procedure of noise injection and identification through least-squares is easy to specify, its analysis required modern tools from random matrix theory and recently developed techniques in robust control synthesis.

Specifying a learning procedure is more complicated when a system is required to satisfy safety constraints, as care must be taken to prevent constraint violation. To address this problem, I developed a novel method for synthesizing robustly safe linear controllers [2]. The synthesis guarantees state and input constraint satisfaction on an infinite time horizon in the presence of bounded disturbances and uncertain dynamics. While existing methods for constrained control use finite horizon problems strung together in a receding horizon strategy, my approach does not require online recomputation and therefore sidesteps issues of recursive feasibility. Using such a robustly safe controller during the first and third steps, I showed that the three step framework guarantees identification, safety, and bounded sub-optimality.

Online and adaptive algorithms follow naturally by repeating the three steps illustrated in Figure 1 during system operation. My coauthors and I presented the first computationally tractable adaptive control algorithm for LQR with guaranteed sublinear regret [3], a problem which has since received renewed interest in the theoretical reinforcement learning community.

Perception-Based Control. Synthesizing information from complex sensors is a highly practical problem, even for engineered systems with well-understood dynamics. Using data to learn a map from complex observations to physical quantities (e.g. from images to lane position) simplifies the planning problem. This approach shifts the challenge to ensuring that the learned perception map remains valid. Proving this validity requires generalization guarantees stronger than are typical in machine learning.

For continuous perception maps and linear systems, I showed that robust control can be used to guarantee generalization [4]: ensuring that the closed-loop sensitivity is bounded proportionally to the continuity of the perception errors certifies stability. This insight applies both to classic computer vision strategies like visual odometry and modern machine learning approaches. Extending to nonlinear control, I developed measurement-robust control barrier functions which

can guarantee safety in the presence of perception errors [5], and showed how safety constraints induce requirements on the density of training data. Taking a more statistical perspective, I showed that even in the presence of noise, a class of nonparametric perception maps has uniformly bounded errors under a dynamically feasible dense sampling procedure [6]. This means that with enough training data, it is no longer necessary to incorporate robustness into control design; i.e. that the certainty equivalent controller has bounded sub-optimality.

My work on perception-based control is directly relevant to robotic applications. In [5], I demonstrate safe control-from-pixels for a realistic robotic Segway simulation, with a software stack directly translatable to the physical robot and full implementation in progress. I have also worked with the Berkeley Autonomous Race Car platform to lead a team of graduate and undergraduate students in building a proof of concept system. Using open source computer vision libraries, the autonomous car races around arbitrary tracks with only a mounted camera to measure position. Previously, I developed an automatic imaging system for high-throughput fluorescence microscopy based on optimized illumination sequences and motion deblurring [7].

Future Directions. There are many opportunities for further investigation at the intersection of learning and control, and I am most excited about problems that grapple explicitly with nonlinearities and are motivated by concrete applications in robotics and sensing.

- **Nonlinear theory.** I plan to extend the data-driven optimal control framework for sample complexity guarantees in nonlinear control. Possible approaches for identification have emerged in recent work on active learning and uniform convergence analyses. However, avoiding exponential dependence on dimension will require the clever use of structure in classes of dynamics or control tasks. On the control side, barrier functions and receding horizon strategies can guarantee robust safety and stability, but defining a notion of suboptimality requires carefully analyzing nonlinear closed-loop behavior.
- **Robotics.** Beyond theoretical guarantees, I hope to use nonlinear data-driven control on robotic platforms to perform realistic and challenging tasks. I am currently in the early stages of a collaboration to demonstrate ice- and roller-skating on a bipedal robot. This problem setting highlights challenges in stability, contact dynamics, and sensing that I plan to address with a combination of modern nonlinear control and data-driven adaptation. There is a rich problem space of learning tasks to explore, from adapting to environmental uncertainties (rough or sloped ice) to executing complex trajectories (advanced figure skating moves).
- **Computational sensing.** To reduce the data requirements for perception-based control, I hope to leverage known properties of sensors and dynamics (e.g. projective camera geometry and rigid body mechanics). This blending of models and data-driven perception falls on the spectrum between classic computer vision, which is specialized for specific image sensors, and deep reinforcement learning, which is in principle general but also data-hungry and brittle. These methods are especially relevant for domains with complex sensing modalities not traditionally considered in computer vision.

ML with Values in Social-Digital Systems

From credit scores to video recommendations, many machine learning systems that interact with people have a temporal feedback component. However, it can be difficult to explicitly model or plan around these dynamical interactions due to limited predictability, difficulties of measurement, and the indeterminacy of translating human values into mathematical objectives. I address

these challenges by developing methods for incorporating considerations of impact into the design of learning algorithms that interact with people.

Fairness constraints are a popular method for mitigating bias. They are especially relevant for cases of consequential decision-making, where notions of equality between groups are legally protected. I worked with my coauthors to contextualize common group fairness criteria from a dynamical perspective by analyzing the delayed impact of the "fair" decisions [8]. Framing the problem in terms of a temporal measure of wellbeing, I demonstrated that static criteria alone cannot ensure favorable outcomes. In a followup collaboration with economists, we proposed an alternate framework: dual optimization of institutional (e.g. profit) and individual (e.g. welfare) objectives directly [9]. In this work, I showed that decisions constrained to obey fairness criteria can be equivalently viewed through the dual objective lens by defining welfare in a particular group-dependent way. This insight, arising from the equivalence between constrained and regularized optimization, sheds light on the connections between the two approaches.

Digital content recommendation offers a distinct set of challenges. Despite its influence over much of the internet, crisp definitions of the values at stake are lacking. During an internship at Canopy, a personalization startup which combines differential privacy with on-device machine learning, I sought to address the extent to which individuals can control their recommendations. I introduced a novel metric of *reachability* to characterize user agency in interactive systems [10]. Using this metric, it is computationally tractable to audit dynamical properties of a recommender system prior to deployment [11]. To understand how such properties might change during deployment, I built an open source recommendation system simulation and evaluation package,[1] which I also used to investigate the reliability of recommender evaluation practices [12].

My work in these areas attempts to re-imagine the goals of predictive models ubiquitous in machine learning, moving towards new design principles that prioritize human values. An important component of this work is articulating tractable criteria for values like safety, fairness, and human agency. This goal cannot be addressed on a technical basis alone [13], which is why I co-founded GEESE, a transdisciplinary student group that aims to give graduate students a constructive place to reflect on issues of society and technology. We have received grants to study sociotechnical pedagogy among graduate students in AI disciplines [14] and to convene an interdisciplinary workshop around questions of reinforcement learning and public policy.

Future Directions There are many open problems, both in developing crisp technical criteria that capture important human values and in incorporating these criteria into algorithm design.

- **Algorithmic reachability.** I am eager to further develop my work on recommendation systems by pursuing learning algorithms which ensure reachability by design. The reachability audit depends on key geometric properties of model parameters, which suggests an approach based on regularization or adjusting models post-hoc. Reachability is also a potential framework for understanding gaming phenomena, allowing for investigations into the tension between protecting against adversarial behavior and providing user agency.
- **Networked systems.** There is much promise for incorporating data-driven methods into networked systems like the power grid and traffic networks. Doing so reliably requires reasoning about both physical dynamics and social concerns. I am interested in extending my work on data-driven control, fairness, social welfare, and agency into these new application areas.

[1] https://github.com/berkeley-reclab/RecLab

References

[1] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. "On the Sample Complexity of the Linear Quadratic Regulator." *Foundations of Computational Mathematics* (2019): 1-47.

[2] Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. "Safely Learning to Control the Constrained Linear Quadratic Regulator". *2019 American Control Conference (ACC)*. IEEE, 2019.

[3] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. "Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator." *Advances in Neural Information Processing Systems*. 2018.

[4] Sarah Dean, Nikolai Matni, Benjamin Recht, and Vickie Ye. "Robust Guarantees for Perception-Based Control." *Learning for Dynamics and Control*. 2020.

[5] Sarah Dean, Andrew Taylor, Ryan Cosner, Benjamin Recht, and Aaron Ames. "Guaranteeing Safety of Learned Perception Modules via Measurement-Robust Control Barrier Functions." *Conference on Robot Learning*. 2020. **Best Paper Finalist.**

[6] Sarah Dean and Benjamin Recht. "Certainty Equivalent Perception-Based Control." Preprint at arXiv:2008.12332.

[7] Zachary Phillips, Sarah Dean, Benjamin Recht, and Laura Waller. "High-throughput fluorescence microscopy using multi-frame motion deblurring." *Biomedical Optics Express* 11.1 (2020): 281-300. Extended abstract awarded **Best Student Paper in Imaging Systems** at OSA Congress 2018.

[8] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed Impact of Fair Machine Learning." *Proceedings of the 35th International Conference on Machine Learning*. 2018. **Best Paper Award.**

[9] Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T. Liu, Danial Bjorkegren, Moritz Hardt, and Josh Blumenstock. "Balancing Competing Objectives with Noisy Data: Score-Based Classifiers for Welfare-Aware Machine Learning." *Proceedings of the 37th International Conference on Machine Learning*. 2020. Short version awarded **Best Paper** at NeurIPS Joint Workshop on AI for Social Good 2019.

[10] Sarah Dean, Sarah Rich, and Benjamin Recht. "Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020.

[11] Sarah Dean, Mihaela Curmei, and Benjamin Recht. "Quantifying Reachability in Collaborative Filtering Recommender Systems." In preparation.

[12] Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I. Jordan. "Do Offline Metrics Predict Online Performance in Recommender Systems?" Submitted to The Web Conference. Preprint at arXiv:2011.07931.

[13] Roel Dobbe, Sarah Dean, Thomas Krendl Gilbert, and Nitin Kohli. "A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics." *5th Workshop on fairness, accountability, and transparency in machine learning*. 2018.

[14] McKane Andrus, Sarah Dean, Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. "AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks." *IEEE International Symposium on Technology and Society (ISTAS)*. 2020.