Towards Mutual Understanding between Mortals and Machines in Motion for Safety

David McPherson

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-74
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-74.html

May 12, 2022

Towards Mutual Understanding between Mortals and Machines in Motion for Safety

by

David Livingston McPherson

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor S. Shankar Sastry, Chair
Professor Claire J. Tomlin
Assistant Professor Steven T. Piantadosi

Spring 2022

ABSTRACT

Every good collaboration is built on solid mutual understanding. Without understanding their machines' behavior, human operators cannot plan around them. Yet increasing automation is distancing them from active understanding. This dissertation will apply cognitive science to build automation that boosts human understanding.

The need for transparency is urgent for safety supervising tasks. Humans' environmental awareness and expansive understanding of safety can save robots from unforeseen edge cases. But only if those humans can also think through the robot's ongoing activity. Actions can be optimized to evidence safety or clearly anticipate faults, enabling supervisors to develop evidence-based appropriate trust. This work explores how observing action allows both humans and robots to construct better working models of the other.

In research on assured autonomy we focus on how machines can autonomously guarantee safety. Yet there will always remain a modeling gap that we require human collaborators to help fill: that's why after decades of autopilot experience and improvements, we still require two human pilots to validate ongoing safe operation. This thesis contends that safe robotics must work to inform these safety collaborators; that choices don't only function to complete objectives but are also *evidence* that other agents ultimately judge.

Characterizing *how agents judge* can empower our machines to choose actions to win correct judgements. First[1] we will exposit how to learn humans' safety concerns from data despite noisy dynamics and demonstrations. After learning humans' concerns, we typify how they perceive and forecast danger[2]. Building on cognitive science we present a model of human safety forecasting structured by reachability analysis. This structure induces data-efficient learning on small datasets so we can learn each su-

1 in Chapters 2 and 3
2 in Chapter 4

pervisor's idiosyncratic ways of thinking – enabling designers to conform their intelligent systems like a glove to a hand.

We build on these models of human safety judgement to support that judgement through machine choices. After learning each supervisor's unique alarms, respecting that safe set[3] lets robot teams decrease supervisory false positives. Extending this approach[4] to anticipate safety concerns ahead of the decision point, we optimize motion as evidence to reject the null hypothesis of danger.

The approaches in this dissertation contributes a mathematical lens for further inquiries into human risk-taking, safety negotiation, and technology learning[5]. By employing the formalisms of intelligent safety to sketch human safety behavior, we imbue machines with a "theory of mind" that is essential to fluent collaboration for our societal systems.

3 in Chapter 5
4 in Chapters 6 and 7
5 sketched in Chapter 8

This work is dedicated to the beloved memory of William Clyde Bryson, Jr. in the hopes that it opens a world of engineering to others like the one he opened for me.

1939 – 2016

Not enjoyment, and not sorrow,
Is our destined end or way;
But to act, that each to-morrow
Find us farther than to-day.

— Henry Wadsorth Longfellow [44]

ACKNOWLEDGMENTS

This book is the product of seven sweet years meeting real scientists: gushing about fast-twitch versus slow-twitch muscle fibers, giggling over duck papers and pronking, and just generally sharing ideas like sugar-rushed grannies giving Halloween candies. My growth in graduate school was formed from these souls. I have only gotten to see the world from a new perspective thanks to their showing me how to open my eyes. Words cannot capture my gratitude. But I'll be damned if I don't try and approach it asymptotically! So here goes. Never in my life can I repay the gift these people gave me:

- To all the Accountibilibuddies – Victoria Tuck, Kristen, Laura Hallock, and Jaimie Swartz: I hit my stride as a professional researcher under your encouragement and by your inspiration.
- Early collaboratos like Kene Akametalu and Shromona Ghosh, Vicenc Rubies Royo and Dexter Scobee imparted their sparks of research process that first kindled my inquiries and keeps me baking now. When I had nothing to contribute, they brought me in.
- To the whole "Autonomous Anonymous" community for solidarity through science's stresses and their volumes of experience. Especially to my fellow organizers Aaron Bestick, Katie Driggs-Campbell and Michael Laskey for making robotics a little more home-y. Beyond more time interacting with all the above friends, I also got to meet wonderful fellow wanderers like Meghan Clarke and Alan Dong who filled study with good humor.

- To all my students: for always asking the "dumbest" questions. There's no such thing as a stupid question, and you consistently proved to me how fresh eyes reveal the true foundations that I've lost sight of. Your creativity always refreshed my enthusiasm to see new ways of thinking about technology. Especial thanks to Peru for being a friend and for honoring my perspective throughout: I never thought of myself as wise until you asked for life advice.

- To all my labmates for always being the first test audience for crazy creative swings and gamely riffing on ideas.
  - Gluing fingers together like kindergarteners while origami folding cardboard robots or dreaming up power-ups for our lemur-robot with tails, wings, jets, rockets, fans, spring-legs, and more – comrades in the Biomimetic Millisystems Lab encouraged me to dream in paper and carbon fiber. You all launch as strongly as a lemur's elastic tendons. My appreciation for Anusha Nagabandi, Carlos Casarez, Ethan Schaler, Cameron Rose, Liyu Wang, Jane Esterline, Duncan Haldane, Austin Buchan, Justin Yim, Eric Wang, Jessica Lee, Mark Plecnik is as indestructible as laminated cockroach construction.
  - For inspiring animations that breakdown the dynamic bones of control and guts to always ask the definitional questions, the people of the Hybrid Systems Laboratory are role models to me: Sylvia Herbert, Mo Chen, Somil Bansal, Palak Bhushan, Andrea Bajcsy, David Fridovich-Keil, Jaime Fisac, Vicenc Rubies Royo, Donggun Lee, Margaret Chapman, Ellis Ratner, Forrest Laine, Michael Lim, Varun Tolani, Kaylene Stocking.
  - For stepping back and giving me a view of the big picture, my peers under Shankar Sastry taught me rigor and vision. From first throwing out a curiosity to Joe to each member gradually coming 'round the cube to pitch in their two cents to ramping up bombastic declarations and witty retorts as everyone's intrigued grins spread wider and wider – I'll always cherish our time together. These lively discursions profoundly shaped the ideas in this work and wouldn't be possible with-

out the unique wit of each of my lab members: Tyler Westenbroek, Joe Menke, Dexter Scobee, Jaime Fisac, Dapo Afolabi, Chih-Yuan Frank Chiu, Victoria Tuck, Kshama Dwarakanath, Mike Estrada, Josh Achiam, Kamil Nar, Eric Mazumdar, Roy Dong, Kshitij Kulkarni, Chinmay Maheshwari, Valmik Prabhu, Sally Hui, Michael Psenka, Ritika Shrivastava, Stella Seo, Josephine Koe. These rousing discussions pulled in ideas from neighboring labs too, so I will always be thankful to Laura Hallock, Katie Driggs-Campbell, Isabella Huang, Andreea Bobu, Aaron Bestick, Sara Fridovich-Keil, Mo Keshavarzi, Robert Matthew, Sarah Seko.

- For teaching me gentleness and awareness of my position to others, for teaching me that I will chase forms my entire life, for giving me many opportunities to laugh at my mistakes, I will always be grateful to the self-defense Yongmudo club at UC Berkeley and especially to masters Norman Link, Elaine Chao, Sally Ho, Lily Chou, Susan Link, David Commins, Raymond Cheung, Vera Chan, Stephanie Siu, Thomas Smart, Patrick Baur and fellow students Ashis Ghosh, Laura Hallock, Nikita Acharya, Han Feng, Mary Tan and Felix Pan.

- To my family at Christ Church, thank you for being my Shire and always reminding me to look up beyond the forest I'm lost in.

- To my friends of the Monday Bible Study – Emmeline Kao, Paul Riggins, Christie Forbes, Ida Bjorkgren, Linda Li, Sam and Fiona Meyer, Will and Jing Qi Smith: for sincerely welcoming me in both my manic joy and many fears. For teaching me how to proactively extend community to strangers. I praise God for you.

- To Carlos Casarez: when I first started researching and felt frustrated and lost, his warm wry smile reassured me that this is the way it goes. "Feeling lost is what happens when you're wandering uncharted territory."

- To Austin Buchan: for putting up with my never-ending questions on getting setup with robotics research.

- To Duncan Haldane: who, when I finished my first project and it seemed old and obvious, saw its simple novelty and taught me that what's old to the discoverer is still new to others. "If it dawns on you that your result is obvious, that means you've just gotten a good enough look at it to share it clearly"

- To Robert Matthew, who illustrates colorful inquiry and makes me excited to get hands-on. Especial thanks for my second-year crash, when you welcomed me when I was alone between labs and dejected. You taught me that loneliness is only for a day and that there's always more going on in rejections. "A rejection just means they don't know what they're missing out on".

- To Margaret Chapman, who taught me grace for the lows and highs of research work and justice for defying the oppression of exclusionary explanations. Learning to listen to my body for setting daily research tasks has been possibly *the* single enabling factor for completing radically open-ended self-guided work. Through her I witnessed the generosity of taking space.

- To Eugen Solowjow, Daniel Duecker, and all the good folks at the Meerestechnik Institut: for currywurst when worry-cursed, for nacken when knackered. You taught me how the same research content can be rejected or accepted based on how you translate it for different audiences. You gave me hope for a different way of doing academia.

- To Dexter Scobee, who believed in me when I didn't believe in myself and showed me the strength of pushing through. Also just generally for demonstrating how to be a professional.

- To Joseph Menke, who could always spot the promise inside others' research and blossomed it into discussion with our whole lab. This touch opened up a talk like a pop-up book, and did more to motivate further research than the most precise presentation feedback. We could see the future through your eyes.

- To Laura Hallock: for your camaraderie every day in the everyday labor of working science. You showed me what diligence is and what it means to be a Scientist. And you accepted me as a peer to join you in doing the same.

- To Victoria Tuck: for your rigor in taking research work with seriousness and the strength to question failings. I learned dignity and carefulness from you. Also just for macaron and cookie exchanges and neighboring desk decorations; you've made the isolated months of the pandemic brighter.

- To Kaylene Stocking: for always being down for a wild musing on learning. Your intellectual curiosity coupled with tactical discussions let me keep exploring even when endings were closing in. I am honored to have gotten to research with you, and I'm excited to follow your science.

- To Justin Yim, who showed me how to wave my hands fast enough to tell if a streetlamp is fluorescent or LED. From bouncing into elevators to detect their cantilever constants and spring supports or forming a two-man horse to demonstrate quadrupedal gaits like pronking past departmental dinner tables – after running with you, every mundane step is now brimming with wonder.

And of course to my many mentors amongst the faculty and especially the members of my thesis committee for reviewing my work and making research stronger.

- To Dr. Tomlin, for introducing me to the elegant structures of optimal control. I literally skipped for excitement after our weekly meetings exploring learning and safety, your connecting me to collaborate with your students started me on the research in Chapter 4, and your wisdom from just one year working with you continues to resound as I find new applications and build new collaborations.

- To Dr. Sastry, for supporting my research direction and nurturing an inquisitive and supportive environment in your lab that welcomed me. Your striving towards big questions and passion for proof has defined my ideals for science. Teaching and researching with you has made me strong.

- To Dr. Piantadosi, for inviting me to discuss research with you and your lab and for sharing an entryway into cognitive science that grounds this research. I am only sad the pandemic prevented me from learning with you more.
- To Dr. Allen Yang for entrusting me with students, for teaching me how to pitch an idea and ground research in demonstrations, and just for giving me the opportunity to research human-robot systems.
- To Dr. Ronald Fearing for introducing me to designing robots for collaboration.
- To Dr. Anca Dragan for early discussions around communicative motion.
- To Dr. Venkat Anantharam for teaching me tools of stochastic analysis and examining my prospectus.

This research exists due to the generous gifts of wonderful wonderers like these. I will pass this on and create curious communities like the ones you created for me. You have changed the world.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

IOC	Inverse Optimal Control
MPC	Model Predictive Control
LQR	Linear Quadratic Regulator
iLQR	Iterative Linear Quadratic Regulation

LFD	Learning from Demonstration
MDP	Markov Decision Process
PDE	partial differential equation
HJB	Hamilton Jacobi Bellman

1

MORTALS AND MACHINES IN MOTION

To date, every application of automation has a human backup. Industrial robots screech into sticking points and rely on line supervisors to halt the crash. Airlines' autopilots may clock-in decades of flight, yet they still rely on multiple pilots to check their work and intervene for edge-cases. Behind every successful robot, there is a capable human – waiting in the wings. They are tasked with continuously monitoring robot performance for evidence that it is operating within acceptable parameters. By integrating a human element into the system, the combined human-robot system benefits from humanity's adaptivity and critical thinking to spot the unquestioned.

Yet even this flexibility and capability has its limits, and joint human-machine systems report failures. When the system fails, it is the responsible caretaker on the scene who often bears the brunt of the blame. Rather than defer responsibility by crying "human error", we should demand better overall system designs since it is always a *joint system* failure.

Whether the fault "originates" with the human or the machine it is the overall human-machine system that failed together. After their career researching "human error", Norman concludes in their seminal textbook [55, p. 215]:

> What we call "human error" is often simply a human action that is inappropriate for the needs of technology. As a result, it flags a deficit in our technology. It should not be thought of as error. We should eliminate the concept of error: instead, we should realize that people can use assistance in translating their goals and plans into the appropriate form for technology.

> Given the mismatch between human competencies and technological requirements, errors are inevitable. Therefore, the best designs take that fact as given and seek to minimize the opportunities for errors while also mitigating the consequences.

When "pilot-induced oscillations" rack airliners with shuddering flight, it is fundamentally due to the mismatch between the machine's and human's frequencies.

Even when the failure originates with an automation fault, any lack of intervention from the human supervisor is better thought of as a failure at the *interface* between the human and machine. When self-driving cars misread a white truck as a bright open sky [8], some blame the owner for not supervising more closely. However, the handoff design failed in calibrating *appropriate trust* and transparently telegraphing that something was amiss in time for the human to react. Or when a new autopilot fails to report angle-of-attack sensor mismatch [13, 75], the pilot's inability to wrangle the elevators back is thwarted by the erroneous automation invisibly turning itself back on after a few moments. For systems to benefit from human safety input, the joint system must equip human decision making with the information they need.

For humans to use our systems effectively, designers must recognize that they are neither fools for foolproofing against nor idealized minds with infinite thinking time, attention, and perception. Instead, like any collaborators, they have strengths within limits. Mixed human-machine collaborations can tap humanity's strengths and bolster their limitations through assistance. The following chapters starts the work of harnessing humanity's safety expertise while clearly learning limitations through statistical models. They ground cognitive science into system theory to formalize human particularities. Only then can our machines ergonomically conform to their cognition.

If our robot designs are to successfully serve peoples' needs, we must prioritize enabling humans' agency and decision-making and build our machines around their strengths and fill in the areas humans are not suited for. This is human-centered design as defined in Norman's seminal textbook [55, p. 8]:

> The solution is human-centered design (HCD), an approach that puts human needs, capabilities, and be-

> havior first, then designs to accommodate those needs, capabilities, and ways of behaving. Good design starts with an understanding of psychology and technology. Good design requires good communication, especially from machine to person, indicating what actions are possible, what is happening, and what is about to happen.

Rather than displace or assimilate human intelligence to mechanistic processes, we work to extend human agency through human-centered automation. To support human intelligence, our machines must be designed considerately of cognition's needs. By sketching mathematical models of cognition, we can design emergent AIs to empower human thinking rather than replace it. My research supports humans' judgment on ongoing robot safety to ensure human concerns are prioritized. Our data-efficient machine learning algorithms can adapt to individuals' concerns and respect diverse understandings of safety through continual improvement. With models of how humans perceive and judge, our robots can optimize their actions to give humans the information they need. We can design transparent systems by studying how humans learn in action.

1.1 THESIS OVERVIEW AND CONTRIBUTIONS

Machines reason in the language of mathematical structures. Equipping robots to consider humanity requires sketching human quirks as mathematical formalisms. Even though human will is far more than the clockwork of ratios and algorithms, the work of trying to describe humanity is an opportunity to inspect parts up-close and personal. Testing these computational translations in application consistently reveals how humanity's complexity exceeds our expectations.

Incorporating more and more of this complexity into behavioral models must be tempered with what can be tractably solved. Even though physicists have incorporated a zoo of particle interactions into the Standard Model, only a fraction of the full model is ever used in practical solutions. Likewise, this work focuses on developing the equivalent of a Newtonian mechanics for behavioral science in engineering application. This tension between fidelity and

tractability defines the art of applied cognitive science and is the primary interest of my research. Instead of investigating root systems in neurology or cognitive science, this thesis will identify applicable distillations of these principles. The value of this work is in distilling as much complex behavioral detail down into a model and keeping it tractable through simplifying mathematical structure.

We start by expositing an application of the traditional rational actor assumption to model human safety behavior. By statistically fitting this model to expert demonstrations, our machines can learn human concerns for safety and risk. How a human navigates an unsafe environment can reveal the constraints they are avoiding. In Chapter 2, we advance the state of the art in constraint inference to rigorously work in even stochastic environments.

Modeling the constraint rules themselves turns out not to be enough, as humans diverge on how they enforce these rules. We can capture divergent safety forecasting behaviors by regressing to their binary labels of when to intervene. In Chapter 4 we show that we can incorporate dynamic safety structure into these regressions to form a data-efficient model of how human supervisors judge our robots' safety. Employing human experts to catch dangerous edge cases for automation helps make the system more efficient. Yet this dependable failsafe cannot work if their safety forecasting is distracted by too many false alarms, an especially common feature of employing one supervisor for a whole team of robots. These false alarms happen when the human's safety forecasting does not align with the robots. Though each supervisor may have distinct dynamic parameters for their mental simulation, learning from their alarm data reveals the contours of their concerns. By modeling these diverse safety forecasts as a reachable set, we can data efficiently learn supervisor's unique concerns and decrease false alarms across a team of robots.

With a mathematical approximation of human judgment, we can optimize choices to be judged correctly despite uncertainty. Chapter 6 optimizes control to evidence autonomous agents' underlying goals. We will see that this communicative optimization can be computed with the same computational complexity rates as the original non-communicative motion optimizations.

Part I

UNDERSTANDING HUMAN CONCERNS

2

MAXIMUM LIKELIHOOD CONSTRAINT INFERENCE

As a first step to supporting human safety judgement, our autonomous systems must first understand those humans' safety concerns. This chapter exposits constraint inference approaches and presents my work extending constraint inference to stochastic applications.

Uncertain transitions make up everyday situations. Under the steaming Georgia sun [80], tires slip as the gravel spits up a puff of caking chalk. Though the vehicle stumbles, innards floating for a beat until the wheels bite the skittering ground again, the driver recovers in time to hem the robot away from the tangled labyrinth of weeds scratching at the track's edge. Elsewhere along the west's brown stubbled hills [15, 16], the ground gulps down long-missed rainwater and the old lake struggles to reserve the surging runoff. Channeling new catchments and reservoirs to handle the thrashing swings of weather, from droughts that crack clay like chapped lips to floods , is essential to keep livelihoods flowing. From split-seconds to sunburnt seasons, we need to steer safely through stochastic surges.

When an expert operates a safety-critical dynamic system, constraint information is tacitly contained in their demonstrated trajectories and controls. These constraints can be inferred by modeling the system and operator as a constrained Markov Decision Process and finding which constraint is most likely to generate the demonstrated controls. Prior constraint inference work has focused mainly on deterministic dynamics. Stochastic dynamics,

This chapter is an adaptation of "Maximum Likelihood Constraint Inference from Stochastic Demonstrations" [50] written in collaboration with Kaylene C. Stocking and S. Shankar Sastry

however, can capture the uncertainty inherent to real applications and the risk tolerance that requires.

This chapter exposits prior art in constraint inference and a novel extension to stochastic applications by using maximum causal entropy likelihoods. This extension to more complicated systems does not cost any extra computational complexity. The derivations in this chapter will reveal the simpler structure underneath the forward-backward algorithms of prior art. We will derive an algorithm that computes constraint likelihood and risk tolerance in a unified Bellman backup, thereby keeping the same computational complexity as prior art.

2.1 INTRODUCTION

Optimization-based control (such as Model Predictive Control (MPC)) promises autonomous behavior [4] even in nonlinear [54] or stochastic dynamics [11, 80]. It has already impacted industrial practice [58] as "model-predictive control", and its recent incarnation as "reinforcement learning" [54, 69] pushes the paradigm further by leveraging large datasets and computing clusters.

Yet these optimizations only work if the clients' goals can be encoded as reward functions and their concerns encoded as safety constraint sets. One approach to this translation is to first solve the inverse of optimal control: given near-optimal demonstrations from the client, recover the reward function whose optimum would match the demonstrator's performance [39]. After fitting the task specification in this way, the objective can then be optimized to imitate the expert behavior [1] or used to predict human motion [85].

Often, inverse optimal control focuses on inferring the magnitude of the reward function. But as optimal control increasingly emphasizes working within constraints, inverse optimal control is interested in identifying those constraints [3, 18, 42, 56, 64]. Chou [17] inferred constraints along the paths that would be low cost but were never observed. This intuition was grounded into a probabilistic framework by Scobee [64] by translating maximum entropy inverse reinforcement learning [84] to work for hard constraints. Unfortunately, the maximum entropy used in Scobee's paper only works for deterministic systems.

Non-deterministic models capture the uncertain dynamics inherent in applications. That uncertainty is especially important to consider when designing for robust safety constraint satisfaction. Stochasticity can stand in for a variety of unpredictable dynamics in applications: from unpredictable power sources in renewable power systems [40] to hard-to-model turbulence in road conditions [80], from tumor cell growth in cancer treatment [63] to unforeseen changes in stormwater reservoirs [16].

The maximum entropy likelihoods can be extended to uncertain transition dynamics by conditioning the entropy at each time step only on the previously revealed state transitions [83]. This maximum *causal* entropy has been extended from running state-based rewards to learn signal temporal logic specifications [78]. Focusing this to just inclusion-for-all-time specifications that make up safety constraints allows for simpler algorithms as Scobee and Sastry [64] did for deterministic systems. This chapter similarly focuses on constraints, paralleling Scobee and Sastry[64], but goes further to model stochasticity by factoring in the causality of dynamics as in Ziebart's later work [83].

2.1.1 *Contributions and Guide*

This work advances prior art [64] in inferring state-action constraints:

- by respecting causality using the principle of maximum *causal* entropy for likelihood generative models
- and by streamlining the algorithm into one backwards pass, thereby maintaining the same computational complexity as the non-stochastic version [64]
- by extending the hypothesis family to include risk-tolerating *chance* constraints

2.2 BACKGROUND

Fitting models entails choosing the model out of some hypothesis class that is "best" along some metric. A natural metric is how likely the model would be to generate the observed demonstration data $\hat{x}_i$ for $i \in [0, 1, \cdots, N]$. Formally, assuming the space of

possible models (called the hypothesis family) is indexed by some vector of parameters θ, the best model is the one with the highest probability of generating the dataset:

$$\theta = \arg\max_{\theta \in \Theta} P_\theta(X_0 = \hat{x}_0, X_1 = \hat{x}_1, \cdots X_N = \hat{x}_T) \tag{2.1}$$

Which is the maximum likelihood estimate of the parameter θ of the probability distribution P_θ.

2.2.1 *Markov Decision Processes*

This probability distribution $P_\theta(X_0, X_1, \cdots X_T)$ can factor into simpler terms when the X_i are states sampled over time from a causal dynamical system. If the state X_i contains all the evolving information, then the Markov property means that datapoints only depend on the past through the most recent preceding state. In particular:

$$P_\theta(X_i | X_{i-1}, X_{i-2}, \cdots, X_0) = P_\theta(X_i | X_{i-1}) \tag{2.2}$$

When these probabilistic dynamics over state are controlled by some exogenous input a to optimize some reward, these Markov dynamics become a Markov Decision Process (MDP). In this work, we focus on discrete time and discrete state and action spaces, so the MDP can be written as a 4-tuple:

- state space $\mathcal{X} = \{x^0, x^1, \cdots, x^{N_X}\}$,
- set of actions $\mathcal{A} = \{a^0, a^1, \cdots, a^{N_A}\}$,
- transition probability function

  $$P(X_{t+1} = x_{t+1} | X_t = x_t, a_t) = S(x_t, a_t, x_{t+1})$$

  where $S : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$. Because of this strong coupling between actions and states across time, it will be useful to discuss their joint distribution as a combined variable $\xi = (a_{[0:T-1]}, x_{[0:T]})$ and the space of such trajectories as $\Xi = \mathcal{X}^T \times \mathcal{A}^{T-1}$

- and objective metric $R(\xi) : \Xi \to \mathbb{R}$ This work assumes the reward to have a form that is decomposable over timesteps:

$$R(\xi) = w(x_T) + \sum_{t=0}^{T-1} r(x_t, a_t) \tag{2.3}$$

where $w(x_T)$ is the final reward and $r(x, a)$ is the running reward.

2.2.2 *Estimating the Task's Rewards*

This probabilistic model can statistically fit to trajectory datasets; either by tuning a parameter θ on the dynamics $S_\theta(x_t, a_t)$ or by tuning a parameter θ on the rewards $R_\theta(\xi)$ to model the expert demonstrator themselves. This latter modeling is the Inverse Optimal Control (IOC) approach to the imitation learning problem: to replicate expert performance given a set of their demonstrated trajectories $\mathcal{D} = \{\xi^1, \xi^2, \cdots, \xi^M\}$.

Continuing with using the maximum likelihood framework to estimate parameters θ, all that is needed is a likelihood that a particular $R_\theta(\xi)$ will generate $a_{[0:T-1]}$ and thereby $x_{[1:T]}$. Ziebart [84] introduced a random distribution on ξ designed to be robust to possible reward phenomena outside the necessarily limited hypothesis class. Specifically, they assume that the parameter to be estimated θ will multiply candidate feature functions $\phi(x, a)$ that will form the spanning basis functions of the hypothesis class. Mathematically:

$$r_\theta(x, a) = \theta^T \phi(x, a) \tag{2.4}$$

The estimation, then, is able to choose the best distribution along the $\phi(x, a)$ basis set, but is incapable of describing distributions outside of this linear subspace of function space. With the goal of remaining maximally agnostic to (and therefore robust to) this non-capturable space, Ziebart [83] deploys the distribution that maximizes entropy outside the candidate $\phi(x, a)$:

$$P_\theta(\xi) = \frac{e^{w(x_T) + \sum_{t=0}^{T-1} \theta^T \phi(x_t, a_t)}}{Z_\theta} \tag{2.5}$$

where Z_θ is a normalizing constant.

This exponential family distribution makes state trajectories that are equally rewarding[1] equally likely and more optimal trajectories exponentially more likely. This likelihood can be used to estimate what rewards $r_\theta(x, a)$ would generate the demonstrated behavior [84] which can in turn be used to replicate that expert performance [85].

Ziebart [84] optimizes the parameters θ via gradient descent on the log likelihood[2]. Following the derivations to Equation 5 of Ziebart [85], let $f(\xi) = \sum_{t=0}^{T-1} \phi(x_t, a_t)$ denote the trajectory-wide equivalent of the feature vector $\phi(x, a)$ that'll define the trajectory's reward:

$$R_\theta(\xi) = w(x_T) + \sum_{t=0}^{T-1} \theta^T \phi(x_t, a_t) \tag{2.6}$$

$$= w(x_T) + \theta^T f(\xi) \tag{2.7}$$

This reward function is linear in the parameter θ which will make the gradient of the log likelihood in Equation 2.5 straightforward:

$$\nabla L(\theta) = \sum_{\xi \in \mathcal{D}} f(\xi) - \mathop{\mathbb{E}}_{\hat{\xi} \sim e^{R_\theta(\xi)}/Z_\theta} f(\hat{\xi}) \tag{2.8}$$

Therefore forward simulating the system dynamics under the θ-optimal control and adding up the features accrued by each trajectory is sufficient. The θ-optimal control can be simply computed for MDPs after a Bellman backup. Running the Bellman backup followed by the forward simulation creates a forward-backward algorithm similar to that in message passing that you can see recreated in Algorithm 1.

1 Note, however, that this distribution distributes based only on the *states* and not the *actions* that the agent actually has control over. This will result in a non-causality that we will explore later on in Section 2.2.4.

2 Since the logarithm is a monotonic function, the log likelihood is guaranteed to share optima with the likelihood

Algorithm 1: Ziebart's Algorithm to compute the gradient on θ

Data: Final reward $w(x)$ and running reward $r(x,a)$, Dynamics $S(x,a,x')$, Vector of features $\phi(x,a)$

Result: Computes for a given current reward model the expected accruals of each feature stacked into a vector F to compute the gradient on the reward model's parameters

```
// Bellman Backup to calculate the strategic transition
   probabilities
```

for $x \in \mathcal{X}$ **do**
  $Z(T,x) \leftarrow \exp(w(x))$
end
for $t \in [T-1, 0]$ **do**
  for $x \in \mathcal{X}$ **do**
    $Z(t,x) \leftarrow 0$
    for $a \in \mathcal{A}$ **do**
      $Za(t,x,a) \leftarrow 0$
      for $x' \in \mathcal{X}$ **do**
        $Za(t,x,a) += S(x,a,x')e^{r_\theta(x,a)}Z(t+1,x')$
      end
      $Z(t,x) += Za(t,x,a)$
    end
  end
end
$P_x(a,t) = Za(t,x,a)/Z(t,x)$

```
// Forward Simulate to calculate expected state visitation
   frequency (for calculating the expectation in gradient
   on parameters)
```

for $x \in \mathcal{X}$ **do**
  $D(0,x) \leftarrow p_0(x)$
  $F(x) \leftarrow 0$
end
for $t \in [0, T-1]$ **do**
  for $x \in \mathcal{X}$ **do**
    $D(t,x) \leftarrow 0$
    for $a \in \mathcal{A}$ **do**
      for $x' \in \mathcal{X}$ **do**
        $D(t+1,x') += S(x,a,x')P_x(a,t)D(t,x)$
      end
    end
  end
end
return $(\sum_{t\in[0,T-1]} \sum_{x\in\mathcal{X}} D(t,x)\phi(x,t))$

2.2.3 *Estimating the Task's Constraints*

Ziebart's algorithm works well for modeling preferences and ideas of optimality, but is insufficient to learn the expert's safety rules that must be satisfied to avoid dangers (e.g. how bicyclists avoid potholes). This question of finding the feasible *domain* of the reward function $R(\xi)$ is addressed by Scobee and Sastry [64] as an optimization with the pre-identified reward magnitudes as an input. The task then becomes to find the constraint set C that will form the support of the maximum entropy distribution as:

$$P_{C^{[0:T]}}(\xi) = \begin{cases} \frac{e^{w(x_T)+\sum_{t=0}^{T-1} r(x_t,a_t)}}{Z_C} & \text{, if } \xi \in C \\ 0 & \text{, if } \xi \notin C \end{cases} \tag{2.9}$$

where $\xi \in C^{[0:T]}$ means that all of trajectory ξ's states $x_{0:T}$ and actions $a_{0:T-1}$ are safe.

Similarly to how we decomposed the reward function over time into a sum of $r(x_t, a_t)$, we focus on the class of constraints that rule on individual timepoints' states $x_t \in C_X$ or actions $a_t \in C_A$ for all time.

$$x_t \in C_X \quad \forall t \tag{2.10}$$
$$a_t \in C_A \quad \forall t \tag{2.11}$$

These constraints can be rolled together into a state-indexed set of allowable actions $C_a(x)$. The indicator for this set, that taking an action a from state x is safe, is:

$$\begin{aligned} &\Phi_C(a,x) = \\ &= \mathbb{I}[a \in C_A \\ &\quad \cap x_t \in C_X)] \end{aligned} \tag{2.12}$$

Let the set of all such combined safety constraints C be $\mathscr{C}$.

Scobee's [64] central insight to identify these constraints is that narrowing the support C (after fixing the reward magnitude $r(x,a)$) will uniformly scale the distribution for all likelihoods still within the support C. As long as the demonstrations stay inside this C,

tightening the constrained safe regions will increase the likelihood of observing those demonstrations. To avoid over-fitting to ruling out all non-visited states and unused actions, states and actions were inferred as unsafe one-by-one until improvement rate tapered off. The best such x^i or a^j to cut out is the one that produces the best scaling factor corresponding to the new normalizing constant Z_{C^+} where C^+ denotes the constraint set C after the candidate x^i or a^j is ruled out.

$$Z_{C^+}^{0:T} = \sum_{\xi \in C^+} e^{w(x_T) + \sum_{t=0}^{T-1} \theta^T \phi(x_t, a_t)} \tag{2.13}$$

This sum over all trajectories can be computed for MDPs by forward simulating the probability distributions. In fact, this ranking metric can be computed for all candidate constraint sets $C^+ \in \mathscr{C}$ since this quantity Z_{C^+} is proportional to the summed probability of all trajectories satisfying C^+:

$$\begin{aligned} P_{C^0,\theta}(\xi \in C^+) &= \\ &= \frac{\sum_{\xi \in C^+} e^{w(x_T) + \sum_{t=0}^{T-1} \theta^T \phi(x_t, a_t)}}{Z_{C^0}^{0:T}} && (2.14) \\ &= \frac{Z_{C^+}^{0:T}}{Z_{C^0}^{0:T}} && (2.15) \end{aligned}$$

where $C^0 \in \mathscr{C}$ is the baseline constraint set with no additional x^i or a^j cut out, and so Z_{C^0} is a constant across all candidates' forward simulations. This means that the $P_{C^0,\theta}(\xi \in C^+)$ will form an equivalent ranking on C^+ as Z_{C^+} would, and can be used interchangeably to identify the likelihood maximizing C^+. Scobee and Sastry linked this probability to the features used in Ziebart's reward inference [84] by casting the probability as the expectation of an indicator $\mathbb{I}[\xi \in C^+]$. The indicator of constraint satisfaction is akin to the features tracked and counted by Ziebart's forward-backward algorithm, but requires an extra memory component: once a trajectory has violated a constraint, that trajectory is still unacceptable even after the state later re-enters the safe set. Adapting Algorithm 1 to include this memory results in

Scobee and Sastry's algorithm (shown[3] in Algorithm 2) for computing the probabilities $P_{C^0,\theta}(\xi \in C^+)$ for ranking hypothesized constraints.

2.2.4 *Issues with Stochastics in Prior Art*

Unfortunately the distribution in Equation 2.5 (and thereby the constrained distribution in Equation 2.9 derived from it) only work for non-stochastic dynamics as shown in [83]: the distribution does not factor in how the probabilistic dynamics reveal transition information over time and thereby makes the agents' selection of x non-causal. Ziebart's 2010 follow-up paper [83] rigorously handled stochastic dynamics for inferring rewards by incorporating the dynamics' information structure via a recursive definition between the actions and succeeding states:

$$P_\theta(a_t|x_t) = \frac{e^{Q^{soft}_{\theta,t}(a_t,x_t)}}{e^{V^{soft}_{\theta,t}(x_t)}} \tag{2.16}$$

$$Q^{soft}_{\theta,t}(a_t,x_t) = r(x_t,a_t) + \mathbb{E}_{X_{t+1}} V^{soft}_{\theta,t+1}(x_{t+1}) \tag{2.17}$$

$$\begin{aligned} V^{soft}_{\theta,t}(x_t) &= \log \sum_{a_t} e^{Q^{soft}_{\theta,t}(a_t,x_t)} \\ &= \operatorname*{softmax}_{a_t} Q^{soft}_{\theta,t}(a_t,x_t) \end{aligned} \tag{2.18}$$

where Q^{soft} can be interpreted as a state-action soft-optimal value-to-go and V^{soft} the state's soft-optimal value-to-go.

The present work will apply this improved causal distribution from [83] to the constraint inference approach designed in [64] in order to apply constraint inference to stochastic demonstrations. Towards this end, constraints can be added to Equations 2.16, 2.17, 2.18 as:

3 Note that the $\phi(x,a)$ in this algorithm is the complement of $\Phi_C(a,x)$ in Equation 2.12 and so F is calculated as the complement of H

Algorithm 2: Scobee and Sastry Feature Accrual History Calculation

Data: Final reward $w(x)$ and running reward $r(x,a)$, Dynamics $S(x,a,x')$, Vector of indicators of constraint *violation* $\phi(x,a) = 1 - \Phi_C(x,a)$ over all candidate constraints $C^+ \in \mathscr{C}^+$.

Result: Computes for a given reward model the expected accruals of each hypothesized constraint feature stacked into a vector F to compute the ranking on constraint hypotheses

// Repeat lines 1-16 of Algorithm 1 to obtain $P_x(a,t)$

// Forwards Simulation for constraint accrual calculation

for $x \in \mathscr{X}$ **do**
  $D(0,x) \leftarrow p_0(x)$
  $H(0,x) \leftarrow 0$
end
for $t \in [0, T-1]$ **do**
  for $x \in \mathscr{X}$ **do**
    for $a \in \mathscr{A}$ **do**
      $\Delta(t,x,a) = \phi(x,a) \odot \left(D(t,x)\mathbf{1}_{n_\phi \times 1} - H(t,x) \right)$
    end
  end
  for $x' \in \mathscr{X}$ **do**
    $D(t+1,x') \leftarrow 0$
    $H(t+1,x') \leftarrow 0$
    for $a \in \mathscr{A}$ **do**
      for $x \in \mathscr{X}$ **do**
        $D(t+1,x') += S(x,a,x') P_x(a,t) D(t,x)$
        $H(t+1,x') +=$
          $(H(t,x) + \Delta(t,x,a))\, P_x(a,t) S(x,a,x')$
      end
    end
  end
end
return $(1 - \sum_{x' \in \mathscr{X}} H(T,x'))$

$$P_C(a_t|x_t) = \frac{e^{Q^{soft}_{C,t}(a_t,x_t)}}{e^{V^{soft}_{C,t}(x_t)}}\Phi_C(a_t,x_t) \tag{2.19}$$

$$Q^{soft}_{C,t}(a_t,x_t) = r(x_t,a_t) + \mathbb{E}_{X_{t+1}} V^{soft}_{C,t+1}(x_{t+1}) \tag{2.20}$$

$$\begin{aligned} V^{soft}_{C,t}(x_t) &= \log \sum_{a_t} \Phi_C(a_t,x_t) e^{Q^{soft}_{C,t}(a_t,x_t)} \\ &= \underset{a_t}{\text{softmax}}\, \Phi_C(a_t,x_t) Q^{soft}_{C,t}(a_t,x_t) \end{aligned} \tag{2.21}$$

2.3 CONSTRAINT STATISTICS

The first step to applying the Scobee and Sastry's [64] key insight to the causal maximum entropy distribution defined in Equations 2.19-2.21 is writing the distribution joint over all timesteps in a horizon $[t:T]$ (with any starting $t \in [0,1,2,\cdots,T]$) as:

$$P_C(A_{[t:T]} = a_{[t:T]}|X_t = x_t) \tag{2.22}$$

$$= \begin{cases} \frac{e^{\mathbb{E}\left[R(X_{[t:T]},a_{[t:T]})\right]}}{e^{V^{soft}_{C,t}(x_t)}} & , \text{ if } a_{[t:T]} \in W_C^{[t:T]} \\ 0 & , \text{ if } a_{[t:T]} \notin W_C^{[t:T]} \end{cases} \tag{2.23}$$

where $W_C^{[t:T]}$ is the set of feedback-controller sequences that satisfy the condition in Equation 2.12 for all times $\tau \in [t:T]$. This distribution over *controls*, rather than *states* as in the non-causal Equation 2.9, means that the dynamics' probability distribution $S(x_t,a_t,x_{t+1})$ is truly incorporated.

With this causal correction in place, the insight from Scobee and Sastry [64] applied to Equation 2.9 can be applied to Equation 2.23. Changing the constraint set C only changes the normalizing constant $Z_{C,t}$:

$$Z_C^{t:T} = e^{V^{soft}_{C,t}(x_t)} \tag{2.24}$$

Therefore incrementing from a constraint set C^0 to any tighter constraint set C^+, as long as this C^+ still includes the demonstrations, will strictly increase the likelihood of the observed demonstrations. To clarify the connection to the quantity in Equation 2.15 that was tracked in [64], note that:

$$P_{C^0,\theta}(\xi^{t:T} \in C^+) = \frac{Z^{t:T}_{C^+}}{Z^{t:T}_{C^0}} = \frac{e^{V^{soft}_{C^+,t}(x_t)}}{e^{V^{soft}_{C^0,t}(x_t)}} \tag{2.25}$$

which is the ratio for converting between C^0's normalizing constant and C^+'s normalizing constant. For brevity, we will denote this probability by $F_{C^+,t}(x_t)$. Most crucially, the probability of a trajectory starting at x_t and staying inside C^+ and C^0 scales by $1/F_{C^+,t}(x_t)$. Therefore this F forms a ranking on possible tightened constraint sets C^+; whichever has the smaller $F_{C^+,t}(x_t)$ will have larger likelihoods of generating the measured demonstrations dataset.

The ratio $F_{C^+,t}(x_t)$ can be computed by modifying the soft Bellman backup defined in Equations (2.19) - (2.21). This modified backup procedure is described in the theorem below:

Theorem 2.3.1. *Let C^0 be a set of constraints and C^+ be an augmented version of C^0 with more states constrained. Then $F_{C^+,t}(x_t)$ can be computed as:*

$$F_{C^+,t}(x_t) = \mathbb{E}_{a_t \sim P_{C^0}}\left[\Phi_{C^+}(a_t, x_t) e^{\mathbb{E}_{x_{t+1}} \log(F_{C^+,t+1}(x_{t+1}))}\right]$$

Proof. It will be helpful to notate the set of legal actions from x under constraint set C as
$A_C(x) = \{a \mid \Phi_C(a,x) = 1\}$

$$\begin{aligned} F_{C^+,t}(x_t) &= \frac{e^{V^{soft}_{C^+,t}(x_t)}}{e^{V^{soft}_{C^0,t}(x_t)}} \\ &= \frac{\sum_{a_t \in A_{C^+}(x_t)} e^{Q^{soft}_{C^+,t}(a_t,x_t)}}{e^{V^{soft}_{C^0,t}(x_t)}} \\ &= \sum_{a_t \in A_{C^+}(x_t)} \frac{e^{r(x_t,a_t)+\mathbb{E}_{x_{t+1}} V^{soft}_{C^+,t+1}(x_{t+1})}}{e^{V^{soft}_{C^0,t}(x_t)}} \end{aligned}$$

It will be convenient to define the logarithm of our $F_{C,t}$. Let it be Δ^t_C:

$$\begin{aligned}\Delta^{t+1}_{C^+}(x_{t+1}) &= \log(F_{C^+,t+1}(x_{t+1})) \\ &= \log(\frac{e^{V^{soft}_{C^+,t+1}(x_{t+1})}}{e^{V^{soft}_{C^0,t+1}(x_{t+1})}}) \\ &= V^{soft}_{C^+,t+1}(x_{t+1}) - V^{soft}_{C^0,t+1}(x_{t+1})\end{aligned}$$

Then the ratio can be redefined in terms of previously calculated terms on C^0 and our iterating $F_{C,t}$

$$\begin{aligned}F_{C^+,t}(x_t) &= \sum_{a_t \in A_{C^+}(x_t)} \frac{e^{r(x_t,a_t)+\mathbb{E}_{x_{t+1}} V^{soft}_{C^0,t+1}(x_{t+1})}}{e^{V^{soft}_{C^0,t}(x_t)}} \\ &\quad \cdot e^{\mathbb{E}_{x_{t+1}} \Delta^{t+1}_{C^+}(x_{t+1})} \\ &= \sum_{a_t \in A_{C^+}(x_t)} \frac{e^{Q_{C^0}(x_t,a_t)+\mathbb{E}_{x_{t+1}} \Delta^{t+1}_{C^+}(x_{t+1})}}{e^{V^{soft}_{C^0,t}(x_t)}} \\ &= \sum_{a_t \in A_{C^+}(x_t)} \frac{e^{Q_{C^0}(x_t,a_t)}}{e^{V^{soft}_{C^0,t}(x_t)}} e^{\mathbb{E}_{x_{t+1}} \Delta^{t+1}_{C^+}(x_{t+1})} \\ &= \sum_{a_t \in A_{C^+}(x_t)} P_{C^0}(a_t|x_t) e^{\mathbb{E}_{x_{t+1}} \Delta^{t+1}_{C^+}(x_{t+1})} \\ &= \mathbb{E}_{a_t \sim P_{C^0}} \Phi_{C^+}(a_t, x_t) e^{\mathbb{E}_{x_{t+1}} \Delta^{t+1}_{C^+}(x_{t+1})} \qquad (2.26)\end{aligned}$$

■

2.4 ALGORITHM

Theorem 2.3.1 implies that an algorithm can compute the conversion ratios $F_C(x)$ for all candidate constraints (which will correspond to how much the hypothesis C shrinks the support thereby increasing the likelihoods) concurrent with the Bellman backup

for the baseline set of constraints C^0. The Greedy Iterative Constraint Inference procedure pioneered in [64] suggests this selection can be performed iteratively adding just one constraint at a time. This iterative approach can be shown to be bounded suboptimal compared to selecting all the constraints simultaneously [64]. In this iterative approach, the F_{C^+} optimizing C^+ will become the baseline set of constraints for the next iteration C^i.

2.4.1 *Comparing States and Actions Satisfaction Frequencies*

Let $\mathscr{C}^+ \subset \mathscr{C}$ be the subset of constraint sets that restrict only one more action or state than the nominal constraint set C_0. The most likely constraint $C \in \mathscr{C}^+$ is whichever still allows the observed demonstrations while having the smallest satisfaction frequency $F_{C^+,0}(x_0)$ from the starting state. This quantity can be computed via our proposed algorithm as described in Algorithm 3's pseudocode.

Analyzing the looped computations in Algorithm 3, we see that they scale with the MDP spaces as $O(|\mathscr{X}|^2|\mathscr{A}||\mathscr{C}|)$ – which is identical to the computational complexity of the Scobee [64] constraint inference for deterministic dynamics!

This Algorithm 3 was constructed to maintain the same hypothesis evaluation metric $F_C = P_{C^0,\theta}(\xi \in C^+) = \frac{Z_{C^+,t}}{Z_{C^0,t}}$ as prior art to highlight analogies. With it, we have combined prior art's forward and backward passes into a single backup pass. However, we can simplify further and only calculate the ranking $F_C = \frac{Z_{C^+,t}}{Z_{C^0,t}}$ *after* all passes have completed. Instead we only track the individual $Z_{C^+,t}$ through the pass, which strips the algorithm down back to just the soft Bellman backup update. This calculation would exactly replace the need to calculate the F_C at each time step and would take the same number of operations (both will scale as $O(|\mathscr{X}|^2|\mathscr{A}||\mathscr{C}|)$). Indeed, if the individual $Z_{C^+,t}$ are calculated we don't even need to normalize by $Z_{C^0,t}$ to rank hypotheses. We can see that the ranking purely on the normalization factors $Z_{C^+,t}$ corresponds to ranking hypotheses directly by the mass of their support – the smaller normalization factors will result in larger likelihoods on the demonstrations. Beyond Algorithm 3's comparative value in drawing analogies to state-of-the-art in maximum

Algorithm 3: Modified Soft Bellman Backup with Value Ratio

Data: Final reward $w(x)$ and running reward $r(x,a)$, Dynamics $S(x,a,x')$, Vector of indicators of constraint satisfaction $\Phi_C(x,a)$ for nominal constraint set C^0 and all candidate constraints $C^+ \in \mathscr{C}^+$.

Result: A column vector F where each entry corresponds to the $F_{C^+,0}$ for $C^+ \in \mathscr{C}^+$

for $x \in \mathscr{X}$ **do**
  $Z(T,x) \leftarrow \exp(w(x))$
  $F(T,x) \leftarrow \mathbf{1}_{|\mathscr{C}^+|\times 1}$
end
for $t \in [T-1, 0]$ **do**
  for $x \in \mathscr{X}$ **do**
    $Z(t,x) \leftarrow 0$
    $F(t,x) \leftarrow \mathbf{0}_{|\mathscr{C}^+|\times 1}$
    for $a \in \mathscr{A}$ **do**
      $Q(t,x,a) \leftarrow r(x,a)$
      $D(t,x,a) \leftarrow \mathbf{0}_{|\mathscr{C}^+|\times 1}$
      for $x' \in \mathscr{X}$ **do**
        $Q(t,x,a) += S(x,a,x') \log(Z(t+1,x'))$
        $D(t,x,a) += S(x,a,x') \log(F(t+1,x'))$
      end
      $Z(t,x) += \Phi_{C^i}(x,a) \exp(Q(t,x,a))$
      $F(t,x) += \Phi_{C \in \mathscr{C}^+}(x,a) \exp(Q(t,x,a))$ $\exp(D(t,x,a))$
    end
    $F(t,x) = F(t,x)/Z(t,x)$
  end
end
return$(F(0,x_0))$

entropy IOC, the simplest algorithm to implement will just be the pure Soft Bellman backup for each hypothesized constraint (see the listing in Algorithm 4).

Algorithm 4: Pure Soft Bellman Backup

Data: Final reward $w(x)$ and running reward $r(x,a)$, Dynamics $S(x,a,x')$, Vector of indicators of constraint satisfaction $\Phi_C(x,a)$ for nominal constraint set C^0 and all candidate constraints $C^+ \in \mathscr{C}^+$.

Result: Z as a column vector where each entry corresponds to the Z_{C^+} for $C^+ \in \mathscr{C}^+$

```
for x ∈ 𝒳 do
    Z(T,x) ← exp(w(x)) 1_{|𝒞⁺|×1}
end
for t ∈ [T − 1, 0] do
    for x ∈ 𝒳 do
        Z(t,x) ← 0_{|𝒞⁺|×1}
        for a ∈ 𝒜 do
            Q(t,x,a) ← r(x,a) 1_{|𝒞⁺|×1}
            for x′ ∈ 𝒳 do
                Q(t,x,a) += S(x,a,x′) log(Z(t+1,x′))
            end
            Z(t,x) += Φ_{C∈𝒞⁺}(x,a) exp(Q(t,x,a))
        end
    end
end
return(Z(0,x_0))
```

Whereas when working with a continuous hypothesis space of parameters as in Ziebart [85] makes calculating gradients from a current guess better than evaluating every hypothesis, when the hypothesis space is finite performing a Bellman backup to rank every hypothesis is not only feasible but equally efficient.

Although we extended the algorithm to handle stochastic dynamics, we did not have to increase the computational complexity at all. It remains at $O(|\mathscr{X}|^2|\mathscr{A}||\mathscr{X}+\mathscr{A}|)$ as in prior art. That is, control engineers can extract safety rules from expert demonstration data for the same cost in both stochastic and deterministic dynamics.

2.5 SUMMARY

This chapter progressed prior art in constraint inference past its roots in forward-backward algorithms into a unified Bellman backup. Consolidating to a backwards pass allowed calculations to marginalize over all futures instead of picking favorite futures anti-causally. Shifting the maximum entropy distribution's definition from realized states to chosen actions finally restored causality. With causality rigorously respected in the constraint likelihood calculation, we can now infer constraints for systems with uncertain transitions. We can now learn human's safety concerns despite noise in the dynamics as well as in their choices. Beyond learning what states or actions experts deem dangerous, human demonstrations through stochastic transitions can also teach us what risks humans are willing to take. In the next chapter we will extend the learnable hypothesis class to include constraints on *chances*.

3

CHANCE CONSTRAINTS AND LEARNING RISK THRESHOLDS

Chapter 2 enabled us to learn human understandings of safety even in stochastic dynamics. Now with notions of negotiating uncertainty placed on human experts, we can now investigate not just what states concern the human but also how much they're willing to risk entering that state. This chapter identifies a technique for identifying constraints on *chances* of transitions.

3.1 CHANCE CONSTRAINTS

We now extend the definition of constraints used in Section 2.2.3. Rather than prohibit all visitation to an undesirable state x, we can instead allow some tolerance $\psi \in [0,1]$ of risking entering x using chance constraints: For any state $x \in C_X$ we allow some small probability $\psi(x)$ of transitioning to that state:

$$P(X_{t+1} = \bar{x} | X_t = x_t, a_t) \leq \psi(\bar{x}), \quad \forall \bar{x} \notin C_X \tag{3.1}$$

To deterministically constrain out a state x set $\psi(x) = 0$. On the other hand, setting $\psi(x) = 1$ means the constraint is inactive and transitioning to x is freely allowed. Therefore the set of state constraints C_X can be encoded as a $\psi(x)$ over all states $x \in \mathcal{X}$. As before, these constraints can be bundled with the action constraints into an indicator that an action a from state x is safe as:

This chapter is an adaptation of "Maximum Likelihood Constraint Inference from Stochastic Demonstrations" [50] written in collaboration with Kaylene C. Stocking and S. Shankar Sastry

$$\begin{aligned}&\Phi_C(a,x) = \\ &= \mathbb{I}[a \in C_A \\ &\quad \cap (P(X_{t+1} = \bar{x} | X_t = x, a) \leq \psi(\bar{x}) \, \forall \bar{x} \notin C_X)] \end{aligned} \tag{3.2}$$

Now the set of all such safety constraints is $\mathscr{C}$. This indicator can replace the one in Equation 2.12 used in Algorithm 3 once the risk thresholds $\psi(x)$ are identified.

3.2 LEARNING CHANCE THRESHOLDS

For each possible unsafe state x there are a continuum of possible values for ψ. Rather than resorting to gridding up the interval $[0, 1]$ we can instead recognize a simple relation between the demonstrations and the acceptable risk for any state x:

Lemma 3.2.1. *When considering constraining out a single state x from C_x, it will maximize the likelihood of the demonstrations to choose the lowest possible $\psi(x)$ that doesn't rule out any demonstrations.*

Proof. Consider two candidate constraints C^+ and $C^\pm$ that differ only by $C^\pm$ having exactly one $\psi(x)$ lower than C^+ has. $C^\pm$ will always have $F_{C^\pm,0}(x_0) \leq F_{C^+,0}(x_0)$. Since a smaller $F_{C^\pm,0}(x_0)$ means that $C^\pm$ will have a larger likelihood of observing the demonstrated trajectories if and only if it doesn't rule out those trajectories as infeasible, the smallest possible $\psi(x)$ will be the maximum likelihood estimator. ∎

Therefore the risk thresholds $\psi(x)$ will maximize the likelihood of the demonstrations by lying exactly on the demonstrations' transition probabilities. We can quickly infer the risk thresholds as:

$$\psi^*(x) = \max_{\xi \in \mathscr{D}} \max_{t \in [0:T-1]} S(x_t, a_t, x) \tag{3.3}$$

3.3 RESULTS

We demonstrate the process of iteratively inferring constraints ranked by Algorithm 3 (shown in Figures 3.1, 3.2, and 3.3) on a

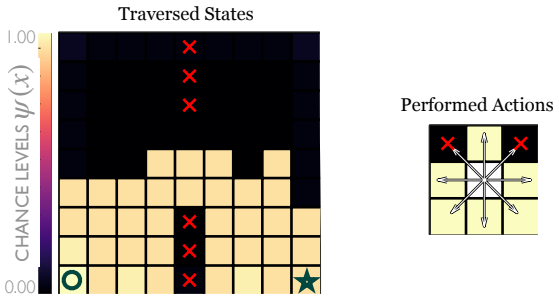

Figure 3.1: Plot of risk thresholds for each state and action hypothesis that are maximally allowed (following Lemma 3.2.1) by the demonstration dataset. The data-allowed risk thresholds for each hypothesis are plotted in their respective state and action spaces. The dark squares are states and actions the demonstrator never took. Conversely, the light state and actions squares were those chosen by the demonstrator. This relative shading on the states corresponds to the largest chance of transition to that state that was demonstrated in $\mathcal{D}$. That transition chance is as low as the risk threshold $\psi(x)$ on that state can be set for a constraint without rejecting the demonstrations as infeasible and making their likelihood 0. Therefore, inferred constraints will have $\psi(x)$ *equal* to that largest demonstrated transition chance.

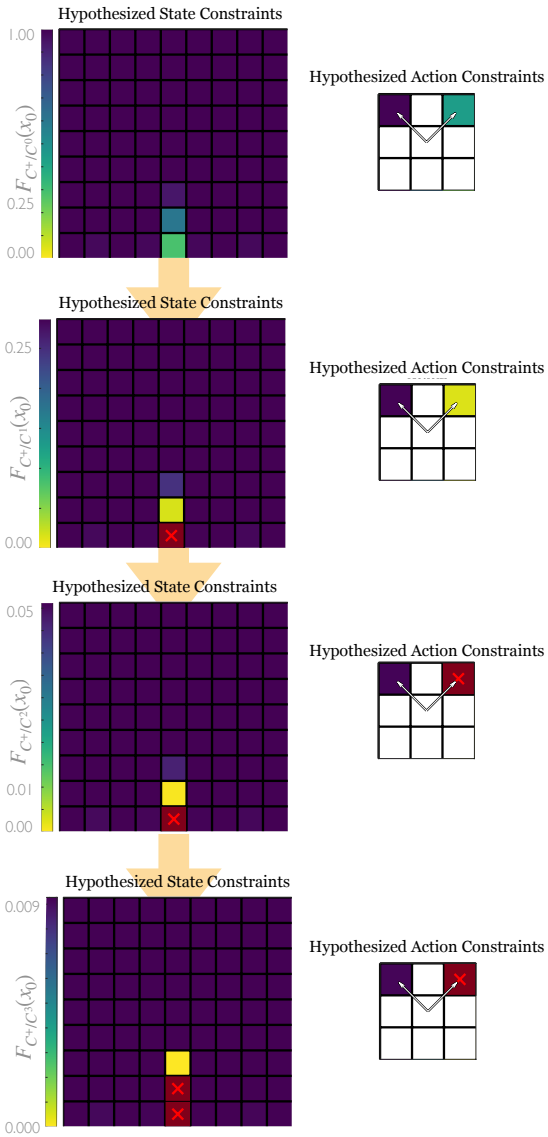

Figure 3.2: Executing our algorithm ranks how likely the states are to be constraints. Compared to the demonstrator, the first panel shows how the initially hypothesized agent (that is fully unconstrained with a vacuous constraint set C^0) would be unlikely to avoid the straightshot between its start and end like the demonstrator did. Indeed, the bottom middle gridcell should only be avoided by an unconstrained demonstrator $F_{C^+/C^0} = 29\%$ of the time. It is unlikely that an unconstrained agent would avoid this straightshot state – more likely there is a constraint there. After identifying a constraint at this bottom state (now marked with a red X), the following panels continue to identify avoidance behavior that is unlikely without an explanatory constraint.

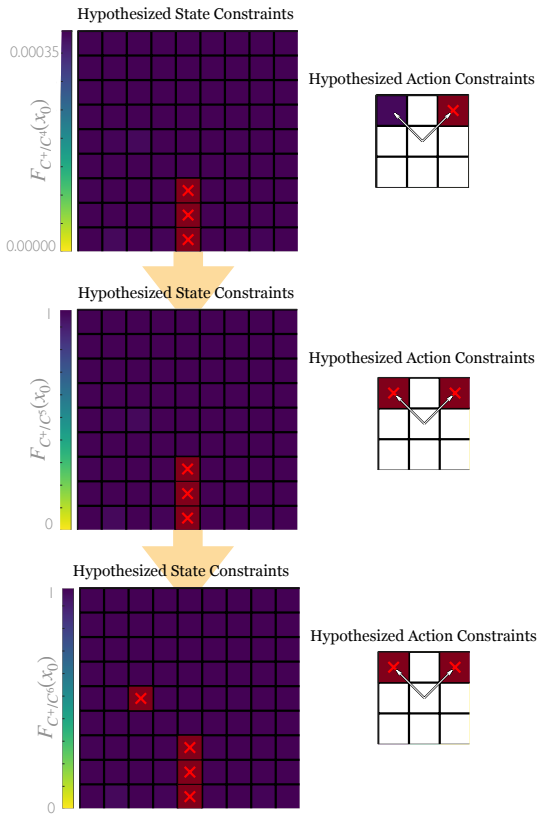

Figure 3.3: After the fourth constraint gets inferred, the continued scaling shrinks by an order of magnitude and then effectively halts as F retracts to $F = 0.96$. This corresponds to inferring an untrue constraint. The last three constraints in the upperhalf of the grid are difficult to infer since those states are already unlikely just from the unconstrained goal – adding those constraints won't change behavior much. Only constraints that are relevant to the task can be identified. If the task is changed those unidentified constraints in the blindspot may become more relevant. Though this is a problem for generalization to new tasks, these new tasks also solve the blindspot by making those states relevant in their new demonstrations.

synthetic dataset of one hundred trajectories from the bottom left to the bottom right avoiding the constrained states and actions marked off with a red X in the figures. This dataset was synthesized from simulated trajectories of a stochastically optimal agent minimizing distance traveled on a two-dimensional "Gridworld" MDP with movement in all eight compass directions. These eight directions made up the action space $\mathscr{A}$ along with a loitering terminal action for once the goal was reached. Each directional action was given a fixed "slippage" chance of 0.1 where a random direction out of the other seven was followed instead. All groundtruth state constraints were fixed at a constant chance threshold of $\psi = 0.25$.

The simulated demonstrator only noisily optimized the task, following a Boltzmann choice distribution as described in Equation (2.23). The constraint inference algorithm was evaluated on this dataset as shown in Figure 3.1. By the fifth iteration (shown in the first panel of Figure 3.3), the algorithm succeeded in recovering the groundtruth constraints (as shown in Figure 3.1).

3.4 LIMITATIONS AND FUTURE WORK

The algorithms discussed in this chapter focused on discretized state and action spaces. For controlling many systems on practical timescales, the state must be handled as a continuous parameter. Ongoing work is investigating how gridded state spaces like in Figure 3.2 can be refined to approximate continuous state spaces [67]. Furthermore, the result from this chapter showed how the constraint ranking passes can be reduced to a variant Bellman backup. This result, as seen in Theorem 2.3.1, suggests that continuous state and time constraints can also be ranked by a similar variant to the continuous analogue to the Bellman backup: the Hamilton Jacobi Bellman (HJB) partial differential equation (PDE). Solving these partial differential equations can pull on a rich literature of solutions, such as the fluid mechanical viscosity solution [53]. Through analogy we would expect the form of the HJB but with a softmax instead of its maximum as in:

$$-\frac{\partial V}{\partial t}(x,t) = \operatorname{soft}\max_a r(x,a) + \nabla_x V \cdot f(x,a)$$

yet rigorously deriving this differential equation from the maximum causal entropy distribution will take a solid effort into the theory.

Extending constraint inference to stochastic systems raises questions of whether human experts might be better modeled using a prospect-theoretic or risk-sensitive measure as in [47]. Future work should investigate how human heuristics for statistical prediction might impact the way demonstrations are generated. The algorithm should be designed to be robust to these biases or even leverage their structure.

3.5 DISCUSSION

Humans are still the current gold-standard for safe operation over machines. Machines must follow their lead to learn what safety concerns impact the workplaces they are deployed in. In this chapter, we saw how data from safe human operation can inform computational understandings of safety. We interpret expert demonstrations as chosen to perform well despite state and action constraints. This generative model lets us infer their underlying planning parameters using maximum likelihood estimation. Building on distributions robust to reward misspecification [83] and inheriting constraint approximation guarantees from successive coverage [64], we extended that prior art to work with noisy transitions. In this setting of precarious performances we introduce a notion of risk with transition chance constraints. Fitting the constraint's chance thresholds equips the machine with a learned approximation of how expert operators balance risk and reward.

This modeling opens the door to exploring human risk-taking behavior in a optimal control or reinforcement learning framework (unlike the more task planning framework of [78]). Prior art in inverse reinforcement learning [60] explored how demonstrators risk the prospect of winning or losing reward [38]. How do the decision-making models that are averse to reward risk extend to constraint violation risks? The model-based data analysis advanced in this work starts to investigate this question, future work can build from here to explore how prospect-theoretic distortions affect decision-making.

By learning risk-reward balancing norms from demonstrations, we sidestep the requirement for users to reason about statistics explicitly. Kahneman and Tversky made careers out of highlighting how humans are ill-equipped to explicitly reason about odds. Yet an expert's intuition is able to implicitly work over distributions through their aggregation of experience. By tapping into demonstrations rather than textual responses, we bypass erroneous statistical distortions from the conscious mind and interface directly with the subconscious mastery. It is important to interface with human capability along humanity's strengths rather than forcing them to match the computer's language. The work in learning from demonstration shows one way that machines can do this.

3.6 SUMMARY

This chapter extended techniques for learning safety from demonstrations. The inverse optimal control problem learns characteristics of planning that can be interpreted as an agent's intent. By carefully factoring in the causal structure of the stochastic dynamics problem, we enabled maximum likelihood constraint inference to be rigorously applied to stochastic systems.

With an understanding of the human's safety concerns in the state and action space, our autonomous agents can follow their lead on acting safely. Yet just behaving safely is not enough; for a team of human supervisor and autonomous robot to work, the robot must also be transparently safe to the human's perception. In the next chapter we will see how each human's safety perception differs, and solve this divergence by data-efficiently learning each supervisor's unique safety forecasting.

4

LEARNING HUMANS' SAFETY FORECASTING

Chapters 2 and 3 learned subsets of the state space that are unsafe from the human's understanding through their demonstrations. This chapter goes from safety *performance* by a human to safety *supervision* by a human. Going from active engagement to passive monitoring forces the human to rely on their ability to *forecast*. We will see that every person forecasts differently. Through their distinct records of interventions, we can learn each supervisor's forecasting concerns. We can learn data-efficiently by structuring the learning with mathematical structures from formal verification.

We propose a model of an idealized supervisor to describe human behavior. Such a supervisor employs an internal model of the robots' dynamics to judge safety. Yet this internal potentially diverges from the model used to forecast the machine's own safety predictions. This difference creates safety disagreements in the human-machine team. We represent these safety judgments by constructing a *safe set* from this internal model using reachability theory. When a robot leaves this safe set, the idealized supervisor will intervene to assist, regardless of whether or not the robot remains objectively safe. False positives, where a human supervisor incorrectly judges a robot to be in danger, needlessly consume supervisor attention. In this work, we propose a method that decreases false positives by learning the supervisor's safe set and using that information to govern robot behavior. We prove that robots behaving according to our approach will reduce the occurrence of false positives for our idealized supervisor model.

This chapter is an adaptation of "Modeling supervisor safe sets for improving collaboration in human-robot teams" [51] written in collaboration with Dexter R.R. Scobee, Joseph Menke, Allen Yang, and S. Shankar Sastry

Furthermore, we empirically validate our approach with a user study that demonstrates a significant ($p = 0.0328$) reduction in false positives for our method compared to a baseline safety controller.

4.1 BACKGROUND

Our deployments of automated systems always fall back on human flexibility as a failsafe for safety. The most common inputs these humans have onto these systems is a halt alert or takeover switch. We believe that learning from this data will equip machines with models of safety closer to human experts. Not only can this inject subject-area expertise into our automation designs, but more closely following human understandings of safety can help avoid confusing the supervisor when the humans' safety concerns don't align with the robots' safety models.

Learning from supervisor's corrections can be done interactively (or "coactively") while supervisors intervene on robots [35] and has been studied as a learning from demonstration (LfD) problem [5]. Learning from demonstration observes full performances from expert demonstrators and attempts to extract statistics sufficient for a robot to imitate the performance. The trajectory focus of foundational LfD means the dynamics focus on timed quantities ranging from dynamics models [19], to reference trajectories [1], to optimization objectives [84]. Prior art in learning from corrections [5, 66] inherited the goal of inferring optimization objectives from Inverse Optimal Control [39, 84]. This chapter focuses on the discrete quality of the instantaneous event of takeover. We will see that it corresponds to a discrete crossing from safe to unsafe. This binary flip will be captured by a binary-valued set inclusion function as we infer subsets of states for safety: constraints.

Constraints can also be learned from expert demonstration [50, 64, 67, 78]. Here too the algorithms inherit from Ziebart [84] a focus on full-horizon trajectories for demonstrator data. Compared to the data from takeover states, full-horizon trajectories contain many more datapoints. This wealth of datapoints both promises the possibility of more expertise information to be encoded in it but also detracts from efficiency as larger datasets need to be gathered and processed. This paper examines whether just the

starting state of a takeover (removing all the time-coupled following states of how the expert drives the system to recovery) can inform us of safety. Indeed, this distillation will reveal how the *timing* of takeovers can provide insights beyond just the safety task into the supervisor's unique perception of safety. Understanding each supervisor's unique way of perceiving safety will empower robots to avoid alarming supervisors unnecessarily, as we will demonstrate in a user study in Chapter 5.

Based on the success of cognitive dynamical models for explaining humans' understanding of physical systems, we posit that human operators may have some notion of reachable sets which they employ to predict collisions or avoid obstacles. We propose a noisy idealized model to describe the behavior of the human supervisor of a robotic team, and we develop a framework for estimating the human supervisor's mental model of a dynamical system based on observing their interactions with the team. We then propose a control framework that capitalizes on this learned information to improve collaboration in such human-robot teams.

4.2 DEFINING REACHABILITY FOR SAFETY

Consider a dynamical system with bounded input u and bounded disturbance d, given by

$$\begin{aligned} \dot{x} &= f(x,u,d), \\ x \in \mathbb{R}^n, \quad u \in \mathcal{U} \subset \mathbb{R}^{n_u}, \quad d \in \mathcal{D} \subset \mathbb{R}^{n_d}, \end{aligned} \tag{4.1}$$

where $\mathcal{U}$ and $\mathcal{D}$ are compact. We let $\mathscr{U}$ and $\mathscr{D}$ denote the sets of measurable functions $\mathbf{u} : [0,\infty) \to \mathcal{U}$ and $\mathbf{d} : [0,\infty) \to \mathcal{D}$, respectively, which represent possible time histories for the system input and disturbance. Given a choice of input and disturbance signals, there exists a unique continuous trajectory $\xi : [0,\infty) \to \mathbb{R}^n$ from any initial state x which solves

$$\begin{aligned} \dot{\xi}(t) &= f(\xi(t), \mathbf{u}(t), \mathbf{d}(t)), \text{ a.e. } t \geq 0, \\ \xi(0) &= x, \end{aligned} \tag{4.2}$$

where $\xi(\cdot)$ describes the evolution of the dynamical system [20].

Obstacles in the environment can be modeled as a "keep-out" set of states $\mathcal{K} \subset \mathbb{R}^n$ that the system must avoid. We define the

safety of the system with respect to this set, such that the system is considered to be safe at state $\xi(0) = x$ over time horizon T as long as we can choose $\mathbf{u}(\cdot)$ to guarantee that there exists no time $t \in [0, T]$ for which $\xi(t) \in \mathcal{K}$. The task of maintaining the system's safety over this interval can be modeled as a differential game between the control input and the disturbance. Consider an optimal control signal $\mathbf{u}(\cdot)$ which attempts to steer the system away from $\mathcal{K}$ and an optimal disturbance $\mathbf{d}(\cdot)$ which attempts to drive the system towards $\mathcal{K}$. By choosing any Lipschitz pay-off function $l : \mathbb{R}^n \to \mathbb{R}$ which is negative-valued for $x \in \mathcal{K}$ and positive for $x \notin \mathcal{K}$, we can encode the outcome of this game via a value function $V(x,t)$ characterized by the following Hamilton-Jacobi-Isaacs variational inequality [25]:

$$\begin{aligned} &\min \begin{cases} l(x) - V(x,t), \\ \frac{\partial V}{\partial t}(x,t) + \max_{u \in \mathcal{U}} \min_{d \in \mathcal{D}} \frac{\partial V}{\partial x}(x,t) \cdot f(x,u,d) \end{cases} = 0 \\ &V(x,T) = l(x). \end{aligned} \tag{4.3}$$

The value function $V(x,t)$ that satisfies the above conditions is equal to $\min_{\tau \in [t,T]} l(\xi^*(\tau))$ for the trajectory with $\xi^*(t) = x$ driven by an optimal control $\mathbf{u}(\cdot)$ and an optimal disturbance $\mathbf{d}(\cdot)$. We can therefore find the set of states

$$\mathcal{R}_T = \{x \in \mathbb{R}^n : V(x,0) < 0\}$$

from which we cannot guarantee the safety of the system on the interval $[0,T]$, also known as the backward-reachable set of $\mathcal{K}$ over this interval. That is, for all initial states $x \in \mathcal{R}_T$ and feedback control polices $\mathbf{u}(t) = g(\xi(t))$, there exists some disturbance $\mathbf{d}(\cdot) \in \mathcal{D}$ such that $\xi(t) \in \mathcal{K}$ for some $t \in [0,T]$.

If there exists a non-empty controlled-invariant set Ω that does not intersect $\mathcal{K}$, then we deem this set Ω a "safe set" because there exists a feedback policy that guarantees that the system remains in Ω, and thus out of $\mathcal{K}$, for all time. It follows from their properties that Ω is the complement of $\mathcal{R}_T$, and the relationship between $\mathcal{K}$, $\mathcal{R}_T$, and Ω is visualized in Fig. 4.1. Within a safe set Ω, the value function becomes independent of t as $T \to \infty$ [25]. Because we focus on the case where the system is initialized to some safe state $\xi(0) \in \Omega$ and we aim to maintain $\xi(t) \in \Omega$ for

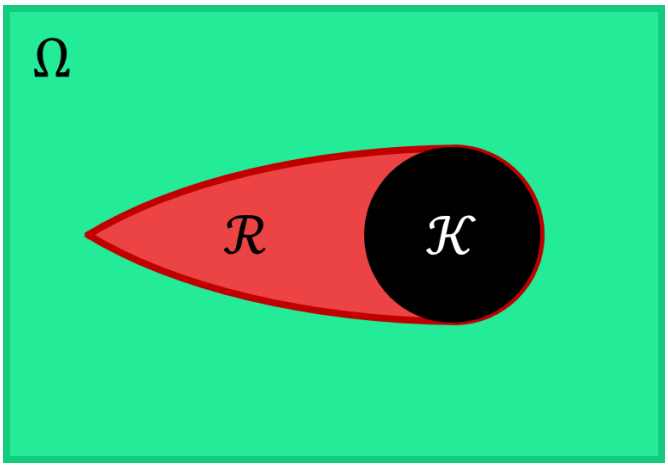

Figure 4.1: Illustration of the relationship between a keep-out set $\mathcal{K}$, the derived backward-reachable set $\mathcal{R}$, and the resulting safe set Ω. Note that $\mathcal{K} \subseteq \mathcal{R}$, and Ω is equal to the complement of $\mathcal{R}$. This illustration approximates the result obtained using the Dubins car dynamics given in (5.1).

all $t \in [0, \infty)$, we simplify notation by defining the terms $V(x) \triangleq \lim_{T \to \infty} V(x, \cdot)$ and $\mathcal{R} \triangleq \mathcal{R}_\infty$.

One approach to guaranteeing the safety of the system is to apply a "minimally invasive" controller which activates on the zero level set of $V(x)$ [30]. This approach allows complete flexibility of control as long as $\xi(t) \in \text{interior}(\Omega)$, and applies the optimal control to avoid $\mathcal{K}$ when $\xi(\cdot)$ reaches the boundary of Ω. We refer the interested reader to [25, 30] for a more thorough treatment of reachability and minimally invasive controllers.

4.3 THE NOISY IDEALIZED SUPERVISOR MODEL

We define an idealized model of the supervisor of a robotic team whose responsibility it is to ensure that no robots collide with obstacles represented by the keep-out set $\mathcal{K}$. The idealized supervisor behaves as a minimally invasive controller as described in Section 4.2. However, while the robotic team members' true dynamics are given by the function $f(x, u, d)$ as in (4.1), the supervisor possesses an internal model of the robots' dynamics given by $f_S(x, u, d)$, which is not necessarily equal to the true dynamics. Following the differential game characterized by (4.3), the super-

visor also possesses an internal value function $V_S(\cdot)$ and safe set Ω_S which they use to evaluate the safety of each state x in the environment. We allow for the possibility that the supervisor adds some amount of margin μ to their internal safe set, such that $\Omega_S = \{x \in \mathbb{R}^n : V_S(x) \geq \mu\}$. Therefore, the idealized supervisor will always intervene when a robotic team member reaches the μ level set of $V_S(\cdot)$, rather than the zero level set of the true $V(\cdot)$. We further specify that the idealized supervisor is *conservative*:

$$\forall x \in \mathbb{R}^n,\ V(x) \leq 0 \implies V_S(x) \leq \mu \tag{4.4}$$

This condition implies that the supervisor will never let a robot teammate leave the true safe set Ω since $\Omega_S \subseteq \Omega$. Additionally, we propose a noisy version of this idealized supervisor: the noisy idealized supervisor will intervene when they observe a robot reach the $\mu + w$ level set of $V_S(\cdot)$, where w is drawn from $\mathcal{N}(0, \sigma_S^2)$ whenever a supervisor makes a safety judgement.

4.4 LEARNING SAFE SETS FROM INTERVENTIONS

We choose to model the human supervisor of a robotic team as approximating the behavior of the idealized supervisor model presented in Section 4.3. That is, the human supervisor will allow the robots to perform their task however they choose, but intervene whenever they *perceive* that a robot is approaching an obstacle $\mathcal{K}$ in the state space. Given this model, we can interpret the points at which the human intervenes as corresponding to the unknown μ level set of some value function $V_H(\cdot) : \mathbb{R}^n \to \mathbb{R}$, which characterizes the human's mental safe set Ω_H. Our goal is to use observations of human interventions to derive an estimated value function $\hat{V}_H(\cdot)$ and $\hat{\mu}$ which describe the observed behavior and induce an estimated $\hat{\Omega}_H$. We approach this task by deriving a Maximum Likelihood Estimator (MLE) of the human's mental safe set. If we assume that a human supervisor always intends to intervene at the μ level set of $V_H(x)$, but their ability to precisely intervene at this level is subject to Gaussian noise, either from observation error or variability in reaction time, then we can consider the value at an intervention point x_i as being drawn from a normal distribution centered at μ (that is, $V_H(x_i) \sim \mathcal{N}(\mu, \sigma^2)$).

Given a proposed value function $\hat{V}_H(\cdot)$ and a set of intervention points $\{x_1, x_2, \cdots, x_p\}$ with corresponding values

$$\{\hat{V}_H(x_1), \hat{V}_H(x_2), \cdots, \hat{V}_H(x_p)\}$$

,

we wish to estimate the most likely μ and σ^2 to explain these interventions. Gaussian distributions induce the following probability density function for a single observation $\hat{V}_H(x_j)$

$$f\left(\hat{V}_H(x_j) \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\hat{V}_H(x_j) - \mu\right)^2}{2\sigma^2}\right) \tag{4.5}$$

which leads to the following probability density for a set of p independent observations

$$\begin{aligned} f\left(\hat{V}_H(x_1), \cdots, \hat{V}_H(x_p) \mid \mu, \sigma^2\right) &= \prod_{j=1}^{p} f\left(\hat{V}_H(x_j) \mid \mu, \sigma^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{p}{2}} \exp\left(-\frac{\sum_{j=1}^{p}\left(\hat{V}_H(x_j) - \mu\right)^2}{2\sigma^2}\right). \end{aligned} \tag{4.6}$$

The likelihood of any estimated parameter values $\hat{\mu}$ and $\hat{\sigma}^2$ being correct, given the observations and the proposed value function $\hat{V}_H(\cdot)$, is expressed as

$$\mathscr{L}\left(\hat{\mu}, \hat{\sigma}^2 \mid \hat{V}_H(\cdot)\right) = f\left(\hat{V}_H(x_1), \cdots, \hat{V}_H(x_p) \mid \hat{\mu}, \hat{\sigma}^2\right) \tag{4.7}$$

It can be shown that the values of the unknown parameters μ and σ^2 that maximize the likelihood function are given by

$$\hat{\mu}^* = \frac{1}{p}\sum_{j=1}^{p} \hat{V}_H(x_j) \quad \text{and} \quad \hat{\sigma}^{*2} = \frac{1}{p}\sum_{j=1}^{p}\left(\hat{V}_H(x_j) - \hat{\mu}^*\right)^2, \tag{4.8}$$

which are simply the mean and variance of the set of observations.

Notice that the estimates given by (4.8) are computed with respect to a given value function $\hat{V}_H(\cdot)$. If we were to assume that

the human supervisor has a perfect model of the system dynamics, then we could simply set $\hat{V}_H(\cdot)$ to equal the true $V(\cdot)$ of the system in (4.1), and $\hat{\mu}^*$ would be the maximum likelihood estimate for the level at which the supervisor will intervene. However, it is unlikely that a human supervisor's notion of the dynamics will correspond exactly to this model, and we would like to maintain the flexibility of estimating value functions that are not strictly derived from (4.1). To this end, we define the maximum likelihood of $\hat{V}_H(\cdot)$ being the $V_H(\cdot)$ that produced our observations as $\mathscr{L}^*(\hat{V}_H(\cdot)) = \max_{\hat{\mu},\hat{\sigma}^2} \mathscr{L}(\hat{\mu}, \hat{\sigma}^2 \mid \hat{V}_H(\cdot))$. The value of $\mathscr{L}^*(\hat{V}_H(\cdot))$ is obtained by substituting the estimates from (4.8) into the probability density function from (4.6). That is,

$$\mathscr{L}^*\left(\hat{V}_H(\cdot)\right) = f\left(\hat{V}_H(x_1), \cdots, \hat{V}_H(x_p) \mid \hat{\mu}^*, \hat{\sigma}^{*2}\right) \tag{4.9}$$

We seek the most likely value function to explain our observations, which will be the value function $\hat{V}^*(\cdot)$ with the greatest maximum likelihood $\mathscr{L}^*(\hat{V}^*(\cdot))$ (the maximum over maxima)

$$\hat{V}^*(\cdot) = \underset{V(\cdot)\in\mathscr{V}}{\arg\max}\, \mathscr{L}^*\left(V(\cdot)\right), \tag{4.10}$$

where $\mathscr{V}$ is the set of all possible value functions.

In order to make this optimization tractable, we can restrict ourselves to a set of value functions $\{V_\theta(\cdot)\}_{\theta\in\mathbb{R}^m}$ corresponding to a family of dynamics functions $\{f_\theta(\cdot,\cdot,\cdot)\}_{\theta\in\mathbb{R}^m}$ parameterized by $\theta \in \mathbb{R}^m$, making the optimization in question

$$\hat{V}^*(\cdot) = \underset{\theta\in\mathbb{R}^m}{\arg\max}\, \mathscr{L}^*\left(V_\theta(\cdot)\right). \tag{4.11}$$

In practice, we may not be able to find an expression for the gradient of $\mathscr{L}^*(V_\theta(\cdot))$ with respect to θ, since the value function is derived from the dynamics $f_\theta(\cdot,\cdot,\cdot)$ via the differential game given by (4.3). The lack of a gradient expression restricts the use of numerical methods to solve the problem as presented in (4.11). In these cases, we can compute a representative library of b value functions $\{V_i(\cdot)\}_{i=1}^b$ corresponding to a set of b representative parameter values $\{\theta_i\}_{i=1}^b$ (see Fig. 4.2 for an example library). The optimization then reduces to choosing the most likely value function from among this library

$$\hat{V}^*(\cdot) = \underset{i\in\{1,\cdots,b\}}{\arg\max}\, \mathscr{L}^*\left(V_i(\cdot)\right). \tag{4.12}$$

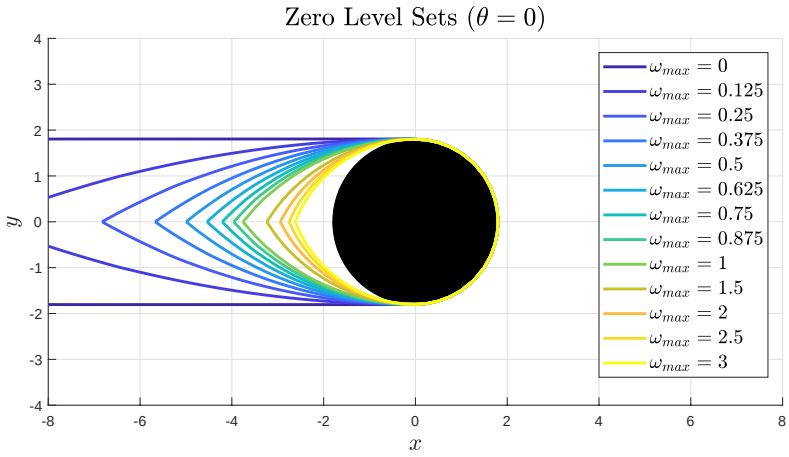

Figure 4.2: Two dimensional slices of the zero level sets of the value functions $V_i(\cdot)$ from the library used for the experiment described in Chapter 5. We used a family of Dubins car dynamics (see (5.1)) parametrized by ω_{max}. Notice that as ω_{max} decreases (the modeled control authority is decreased), the level sets extend farther away from the obstacle, indicating that a robot is expected to turn earlier to guarantee safety.

In order to ensure that the learned safe set is conservative, we can extend our MLE to a Maximum A Posteriori (MAP) estimator by incorporating our prior belief that, regardless of the safe set that the supervisor uses to generate interventions, they do not want the robots to be unsafe with respect to the true dynamics. In this case, we maintain a uniform prior $P(\theta)$ that assigns equal probability to all $V_\theta(\cdot)$ whose zero sublevel sets are supersets of the zero sublevel set of the true $V(\cdot)$, and zero probability to all other $V_\theta(\cdot)$. In other words, we assume that the supervisor does not overestimate the agility of the robots, and in practice we can enforce this condition by choosing the library in (4.12) to only contain appropriate value functions. Moreover, regardless of the choice of $\hat{V}_H(\cdot)$, we assume that the supervisor intends to intervene before reaching the zero level set of $\hat{V}_H(\cdot)$, which always includes the boundary of $\mathcal{K}$. If we choose a prior $P(\mu)$ that assigns

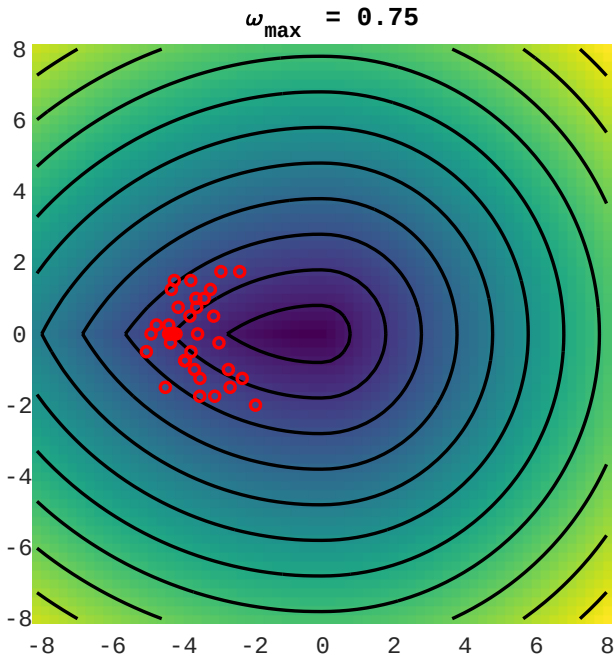

Figure 4.3: An example data set of how supervisors intervene for a simple car model. This data was gathered from the experiment described in Chapter 5 and illustrates how one supervisor's interventions spread around the contours of a reachable set. The red circles represent the location of supervisor interventions, and the colored background represents the learned value function $V(\cdot)$ with contour lines shown in black. In this case, the learning algorithm chose a dynamics model parametrized by $\omega_{max} = 0.75$.

zero probability to all non-positive μ and uniform probability elsewhere, it can be shown that the MAP estimates are obtained by letting $\hat{\mu}^*$ equal $\max\{\hat{\mu}^*, 0\}$ and otherwise proceeding as before. Fig. 4.3 provides an example of this algorithm estimating a safe set from human supervisor intervention data.

4.5 CONTROL ON LEARNED SAFE SETS

We propose that safe sets learned according to the approach in Section 4.4 can be used to create effective control laws for the robotic members of human-robot teams. Recall our model of the

human supervisor of a robotic team: the supervisor must rely on each robot's autonomy to complete the majority of their tasks unassisted, but the supervisor may intervene to correct a robot's behavior when necessary (such as by avoiding an imminent collision with the keep-out set $\mathcal{K}$). We put forth that in the scenario where the human intervenes to prevent a collision, they do so because they observe that a robot has violated the boundaries of their mental safe set Ω_H.

Now, consider a team of robots navigating an unknown environment, and which are able to avoid any obstacles that they detect. One approach to safely automating this team is to have each robot behave according to a minimally invasive control law: the robots are allowed to follow trajectories generated by any planning algorithm, so long as they remain within Ω, the reachable set computed using the baseline dynamics model (4.1) with associated value function $V(\cdot)$. Whenever these robots detect an obstacle, they add it to the keep-out set $\mathcal{K}$, thus modifying Ω and $V(\cdot)$. If a robot reaches the boundary of Ω, it applies the optimal control to avoid $\mathcal{K}$ until it has cleared the obstacle. However, it is possible that a robot does not detect an obstacle, and a human supervisor must intervene to ensure robot safety.

As stated above, the human supervisor will intervene when a robot reaches the boundary of Ω_H, not the boundary of Ω. This discrepancy leads to the possibility that the supervisor will intervene when the robot reaches some state x, even if the robot would have avoided the obstacle without intervention. These situations arise whenever $V_H(x) \leq \mu$ but $V(x) > 0$. These "false positive" interventions represent unnecessary work for the human supervisor, and we seek to eliminate them in order to improve the human's experience and the team's overall performance.

We propose using a safe set $\hat{\Omega}_H$ learned from previous observations of supervisor interventions, as outlined in Section 4.4, as a substitute for Ω in the robots' minimally invasive control law. By estimating the human's internal safe set, we take advantage of the following property:

Property. For an idealized supervisor collaborating with a team of robots as described in Section 4.5, if the robots avoid detected obstacles $\mathcal{K}$ by applying an optimally safe control at the boundary of safe set Ω_S, then if the supervisor plans to intervene because

they observe $\xi_i(t) \in R_S$ for robot i, the supervisor can infer that robot i has not detected an obstacle and any supervisor intervention will not be a false positive.

Proof. The proof of this property follows constructively from the definitions of safe set, idealized supervisor, and false positive. If robot i had correctly detected an obstacle and adjusted its representation of Ω_S accordingly, then it would have applied the optimal control to remain within the supervisor's safe set. Therefore, if the supervisor is able to observe that robot i has left Ω_S, it must be the case that the robot has not detected the obstacle. False positives are defined to be supervisor interventions that occur when a robot has detected an obstacle but the supervisor still intervenes. In this case, the supervisor can correctly infer that robot i has not detected an obstacle, so any intervention at this point cannot be a false positive. ■

For an idealized supervisor, as $\hat{\Omega}_S$ becomes an arbitrarily good approximation of Ω_S, the number of false positive interventions will approach zero. For a *noisy* idealized supervisor, the supervisor will intervene whenever $V_S(x) + w \leq \mu$ where $w \sim \mathcal{N}(0, \sigma_S^2)$. This noise will continue to produce false positives, even with a perfect fit $\hat{\Omega}_S = \Omega_S$, if the robots apply the optimally safe control at the μ-level set of Ω_S. Instead, the level set α where the optimally safe control is applied can be raised arbitrarily high to drive the false positive rate to zero. For example, $\alpha = \mu + 2\sigma_S$ is sufficient to avoid over 97% of intervention states used for learning, in expectation. We test the efficacy of our approach through the human-subjects experiment described in Section 5.2.

4.6 DISCUSSION

These simple avoidance controllers assumes the noisy Newton hypothesis for human cognition and analyzes forecasting through reachability calculations. By modeling which system states command supervisory attention, we can program autonomous systems to avoid those states when they do not require attention. We find that building from cognitive scientific principles into tractable mathematical structures borrowed from established verification produces predictive and data-efficient results. This work suggests

the potential of using other dynamic verification tools to model facets of human judgement. For example, the human interventions of this chapter could be modeled as responses to violated assume-guarantee contracts between the agents suggesting what actions need to be delegated and when. The rich toolsets of verification and their different concerns are fertile ground for rigorous development in human-robot collaboration. Formalizing concepts from psychology into mathematical structures equips machines with the computational language to learn human concerns.

By considering judgement mathematically, we witness just how much human cognition varies from the "rational" standard used to approximate mean human behavior. Our machines had to consider how each supervisor works with a distinct mental simulator for safety forecasting. This discrepancy between purely rational "econs" and humans in-the-wild is well noted in econometrics' prospect theory and broader behavioral economics. Behavior that diverges from mean models not only happens within markets – in our user study we see how every user has distinct concerns. When dealing with supervisory precision as measured by false positives, we see that this discrepancy has concrete impact on how we design our systems of machines and workers. We can make tractable headway in modeling this discrepancy through parametrized models like safe sets. Research considering cognition's idiosyncrasies through divergent intelligent systems models can tractably improve human-machine collaboration.

4.7 SUMMARY

Rather than work from long demonstrations of safe behavior, this chapter examined learning from the volumes of safety takeovers. In this supervisory context, new mental phenomena enter into the problem: how does the human *forecast* safety from the present state? This chapter applied reachability analysis to structure data-efficient regression to this phenomenon. In the next chapter we will examine an application of learning human forecasting: by being cognizant of human safety judgement, robots can mitigate false alarms due to mismatched dynamics models.

Part II

SUPPORTING HUMAN UNDERSTANDING

5

AVOIDING FALSE ALARMS THROUGH MODELING

When a human supervisor collaborates with a team of robots, the human's attention is divided, and cognitive resources are at a premium. We aim to optimize the distribution of these resources and the flow of attention. The ways that robot performance alarms the supervisor's attention can be predicted in part with the model derived in Chapter 4. We empirically validate the usefulness of the noisy idealized supervisor model with a user study. In partnering users with a team of robots that considered their perspective, we reduced unneeded interventions ($p = 0.0328$) over the default safety controller. This reduction in false positives also tracked the amounts predicted by the noisy idealized supervisor model.

5.1 INTRODUCTION AND BACKGROUND

As automation becomes more pervasive throughout society, humans will increasingly find themselves interacting with autonomous and semi-autonomous systems. These interactions have the potential to multiply the productivity of human workers, since it will become possible for a single human to supervise the behavior of multiple robotic agents.

While a human may be able to successfully exert direct control over a single robot, it becomes intractable for a human to directly

This chapter is an adaptation of "Modeling supervisor safe sets for improving collaboration in human-robot teams" [51] written in collaboration with Dexter R.R. Scobee, Joseph Menke, Allen Yang, and S. Shankar Sastry

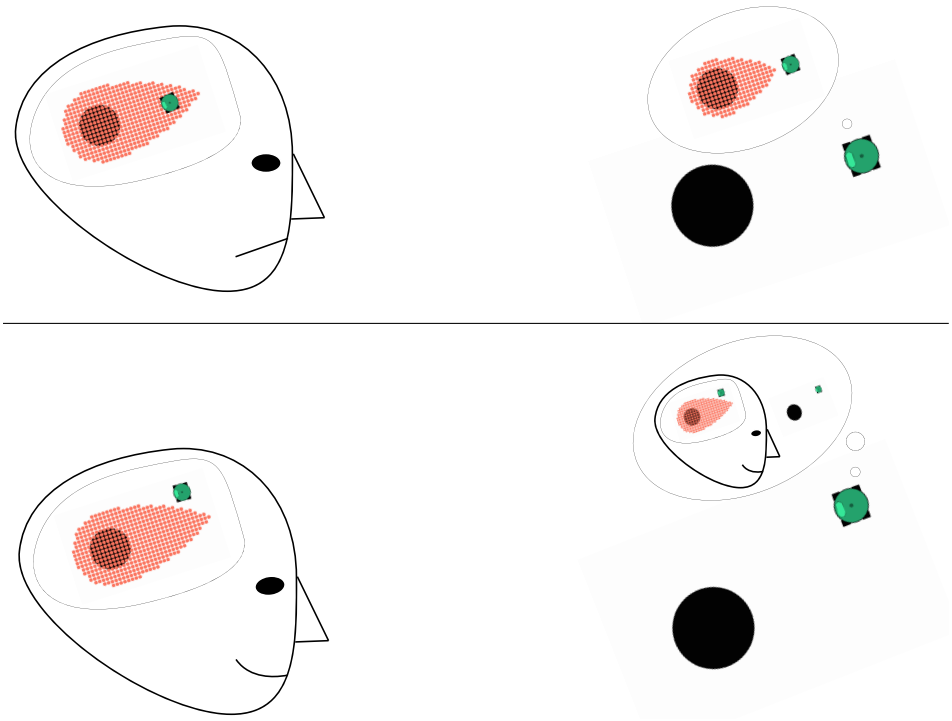

Figure 5.1: Top: if a robot's behavior does not take into account a human supervisor's notion of safety, the misaligned expectations can degrade team performance. Bottom: When a robot acts according to a human supervisor's expectations, the supervisor can more easily predict the robot's behavior.

control teams of robots [1]. In order to manage the increased complexity of multi-robot teams, the human must be able to rely on increased autonomy from the robots, freeing the human to focus their attention only on those areas where they are most needed. Our goal is to model what grabs the supervisor's attention in order to modify robot behavior to reduce the occurrence of distractions.

This project is inspired by work like Bajcsy et al [5] and Jain et al [35] that learn from supervisor interventions in a "coactive" learning framework. These works apply Learning from Demonstration techniques to the more challenging domain where the given data is just a correction from a trajectory rather than a full

1 in fact, even when working with a single robot, semi-autonomous assistance helps the human-machine collaboration as discussed in the literature on assistive teleoperation [23, 37]

trajectory. The authors of [5] posed this correction challenge in model-based framework that interprets the human's signals as resulting from an optimization problem. This inverse optimization framework has also been used in Inverse Reinforcement Learning [1, 84] which applies Inverse Optimal Control (as conceived of by Kalman [39]) to interpreting human trajectories. Our work applies the inverse optimization framework to learn from the supervisor's decisions to intervene.

Results in cognitive science suggest that humans observing physical scenes can be modeled as performing a noisy "mental simulation" to predict trajectories [9, 65]. We posit that human supervisors utilize this same cognitive dynamic simulation to predict robot safety and intervene accordingly. Specifically, we theorize that the intervention behavior is driven by an internal "safe set" which we can attempt to reconstruct by observing supervisor interventions.

5.2 EXPERIMENTAL DESIGN FOR USER VALIDATION

Our goal in understanding and modeling the supervisor's conception of safety is to improve team performance by decreasing cognitive overload. Although we have based our human modeling on the cognitive science literature, we do not intend to verify humans' exact cognitive processes. Instead, we aim to apply our inspiration from cognitive science toward building better human-robot teams. To this end, our hypotheses are:

- *H1*: Representing supervisor behavior as cognitive keep-out sets allows intervention signals to be distilled into an actionable rule which will decrease supervisory false positives and cognitive strain, thereby increasing team performance and trust.

- *H2*: Fitting danger-avoidance behavior to a supervisor's beliefs is preferable to generic conservative behavior.

In our experiment, we gather supervisor intervention data, fit our model to the data, and then run a human-robot teaming task that assesses performance.

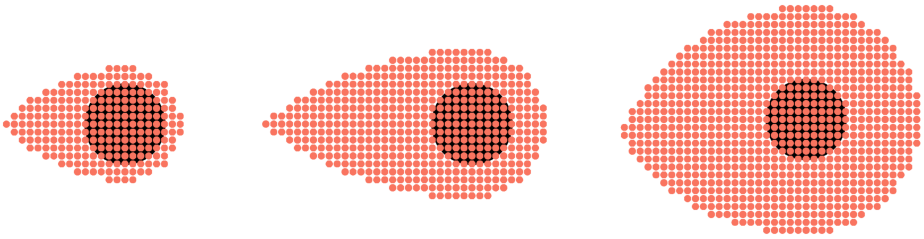

Figure 5.2: Safe sets tested in our experiment (illustrated by their complementary reachable set): (left) Standard safe set (calculated from true dynamics and obstacle size), (middle) example Learned safe set (calculated from fitted supervisory perception of dynamics and obstacle size), (right) Conservative safe set (calculated from true dynamics and inflated obstacle size)

5.2.1 *Procedure*

Our experiment applies the idealized supervisor theory and learning algorithm to supervising simulated robots. The robots moved according to the Dubins car model:

$$\begin{aligned} \dot{x} &= 3\cos(\theta) \\ \dot{y} &= 3\sin(\theta) \\ \dot{\theta} &= u \\ u \in \mathcal{U} = [-\omega_{max}, \omega_{max}],\ \omega_{max} &= 1 \end{aligned} \tag{5.1}$$

The experiment is divided into three phases. In Phase I, the subject is given an opportunity to familiarize themselves with the robotic system's dynamics. The user is allowed to directly apply the full range of controls through the computer keyboard for one minute. After ensuring the user has some experience from which to build an internal dynamics model, we then assess their emergent conception of safety. In Phase II, supervisory data is extracted from the subject by showing them scenes where the robot is driving towards an obstacle, and the supervisor decides where to intervene to avoid a crash. After building their conceptual model of the robot's dynamics, they share data on it through veering the robot out of crashes. By detecting when they flinch in the game of chicken we can learn what states alarm this supervisor uniquely through our algorithm described in Chapter 4. Our

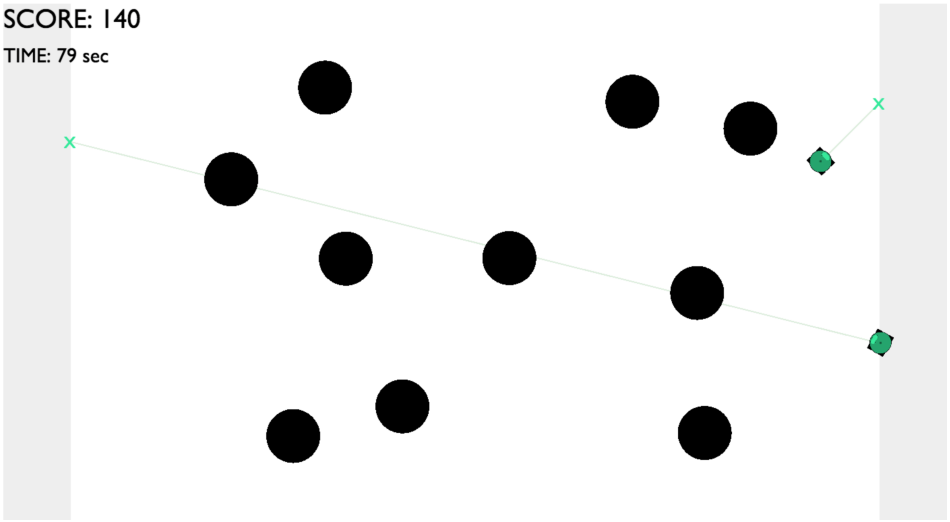

Figure 5.3: Screenshot of the task from Phase III of the experiment. Robotic vehicles make trips back and forth across the screen, detecting and avoiding each obstacle with 80% probability. The human supervisor must remove an obstacle in the event that it is undetected, but must infer this information from the robots' motion.

estimator used a library of candidate dynamics functions parameterized by values of ω_{max} between 0 and 3, as shown in Fig. 4.2. In this experiment, we enforced conservativeness by excluding subjects whose Learned sets were not supersets of the Standard safe set, rather than enforcing a prior directly on ω_{max}. The Learned safe set is assessed in Phase III against two fixed safe sets (see Fig. 5.2) pre-calculated from the true dynamic equations.

These safe sets were calculated using Hamilton-Jacobi reachability as described in Section 4.2 using the Level Set Toolbox [52] for MATLAB.

We compare the vehicles that have learned human safety concerns against vehicles that don't consider the supervisor but only avoid obstacles using the engineered concerns designed into the minimum safety guarantees (as per [30]). Vehicles rumble across the pixelated screen, delivering packages between depots that reward the team. Skittering towards dark holes strewn across their workspace, their autonomous understanding of safety swerves them away from a crash. Yet 20% of the time their simulated sensors "fail" to perceive the obstacles and they can only blindly

crash, costing a full delivery's worth of rewards. Therefore, the subject hovers above their mouse, watching for oncoming collisions with dark holes strewn across the vehicles' workspace. With a click they can displace an obstacle at the expense of half the rewards scored by one of their vehicles' successful deliveries. Still, it beats losing a full delivery's worth in a crash. Humans must balance relying on delegated autonomy and supporting team members' blindspots. In that decision, the supervisor will be considering their understanding of safety and forecasts of vehicle motion. In turn, team members will either obey only their understanding of safety using the minimum interventions from the baseline safe sets (as in [30]) or they can match the concerns their collaborator is working with. We contend that considering their collaborator's concerns will help the human discern the robot's (un)safety.

5.2.2 *Independent Variables*

To assess our hypotheses, we manipulate the safe set used between team supervision trials. We exposed the human subject to three teams, each driving using one of three safe sets. The Learned set is derived from Phase II supervisor intervention observations as described in Section 4.1, using $\alpha = \mu$. The two baseline kernels are calculated using Hamilton-Jacobi-Isaacs reachability on the true dynamic equations. The Standard set is calculated using the true obstacle size. The Conservative set adds a buffer that doubles the effective size of the obstacle, inducing trajectories that give obstacles a wide berth.

5.2.3 *Dependent Measures*

5.2.3.1 *Objective Measures*

The team was tasked with making trips across the screen to reach randomized goals. The robots' task was to travel across the screen, safely dodging obstacles along the way, while the human was tasked with supervising as a failsafe to remove an obstacle if the robots should fail to observe and avoid it.

Team performance was quantified using three objective metrics: number of trips completed, number of supervisory interven-

tions, and the number of obstacle collisions. These metrics were presented to the subject as an aggregated, arcade-style score. To incentivize participants to only intervene when necessary, obstacle-removal interventions reduced the score, but only by half as much as an obstacle collision.

The number of interventions taken by the supervisor can also serve as a proxy measurement to quantify the amount of cognitive strain they experience while working with the robotic team. Of particular note are the number of interventions that were not actually required, as the supervisor incorrectly judged that a robot had not detected an obstacle. These false positives needlessly drain supervisor attention and indicate a lack of trust in the system. We aim to increase the human's trust in the system, which we quantify by a decrease in these false positives.

5.2.3.2 *Subjective Measures*

After each round of pairwise comparison (completing the task with two different robotic teams), we gauged how the choice of safe set impacted the subject's subjective experience with a questionnaire. The questionnaires asked subjects how much they agreed with a series of simple statements on a 7-point Likert scale (1 - Strongly Disagree, 7 - Strongly Agree). These statements were designed to measure Trust, Perceived Performance, Interpretability, Confidence, Team Fluency, and overall Preference between the teams in the comparison.

5.2.4 *Subject Allocation*

The subject population consisted of 6 male, 5 female, and 1 non-binary participants between the ages of 18-29. We used a within-subjects design where each subject was asked to complete all three possible pairwise comparisons of our three treatments (the safe sets used). We used a balanced Latin Square design for the order of comparisons, with no treatment being first in a pair twice. Furthermore, we generated six randomized versions of the task so that subjects were presented with a different version of the task for each trial across the three pairwise comparisons. To avoid coupling the treatment results to a particular version of the task,

each treatment was paired with each task version an equal number of times across our subject population.

5.3 ANALYSIS AND DISCUSSION

5.3.1 *H1: False Positive Reduction over Standard*

Our first hypothesis is that a Learned safe set that reflects the supervisor's intervention behavior would decrease the number of false positives compared to the Standard safe set. To test this, we performed a one-way repeated measures ANOVA on the number of supervisory false positives from Phase III of the experiment with safe set as the manipulated factor. A false positive was any supervisor intervention where the removed obstacle was actually detected by all nearby robots, which would have avoided it successfully. The robot team's safe set had a significant effect on the number of supervisory false positives ($F(2, 20) = 8.72, p < 0.01$). An all-pairs post-hoc Tukey method found that the Learned safe set significantly decreased ($p = 0.0328 < 0.05$) false positives over the Standard safe set, but there was no significant difference between the Learned safe set and the Conservative safe set (which also significantly decreased false positives over the Standard safe set, with $p < 0.01$). These results support our main hypothesis that *representing supervisor behavior as cognitive keep-out sets allows intervention signals to be distilled into an actionable rule which will decrease supervisory false positives*.

The second half of that hypothesis, that *decreasing supervisory false positives will increase trust and team performance* was not shown conclusively from our data. We performed a one-way, repeated measures ANOVA on the pairwise comparison surveys between the teams using the Learned and the Standard safe sets. Measures of trust showed no significant improvement ($F(1, 9) = 1.86, p = 0.21$).

5.3.2 *H2: Preference over Conservative*

For 9 of 11 participants, the Learned safe set had shorter avoidance arcs than the Conservative set. We hypothesized that this greater efficiency would make the tailored conservativeness of

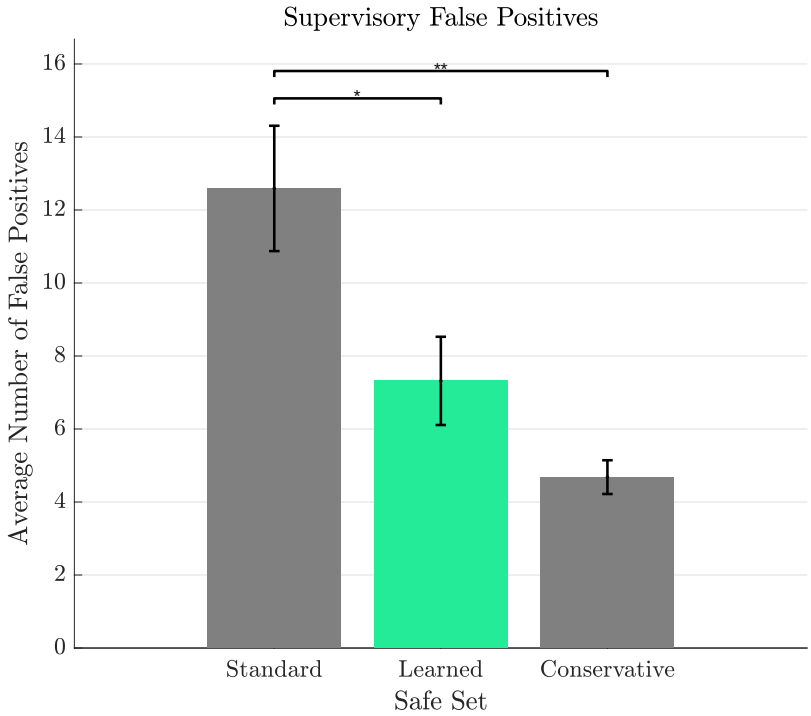

Figure 5.4: Average number of false positives per trial plotted against the three safe set types. There were significant differences between Standard and Learned ($p < .05$) and between Standard and Conservative ($p < .01$). There was no significant difference between Learned and Conservative.

the Learned set preferable to the baseline Conservative safe set. However, a t-test showed that the survey responses for preference were statistically indistinguishable ($p = 0.8$) from a neutral score: an inconclusive result for Hypothesis 2. We believe that this result stems from users judging preference more on intelligibility, the ease of avoiding false positives, than on efficiency, the shortness of paths. As discussed in Section 5.3.1, both the Learned and Conservative safe sets led to significant false positive reductions over the Standard set.

This indistinguishability is further compounded since a preference for intelligibility seems to be expressed by some subjects in their Phase II intervention data, resulting in their Learned safe sets having similar arcs as the Conservative safe set (see Fig. 5.5). Future work could investigate this efficiency-intelligibility

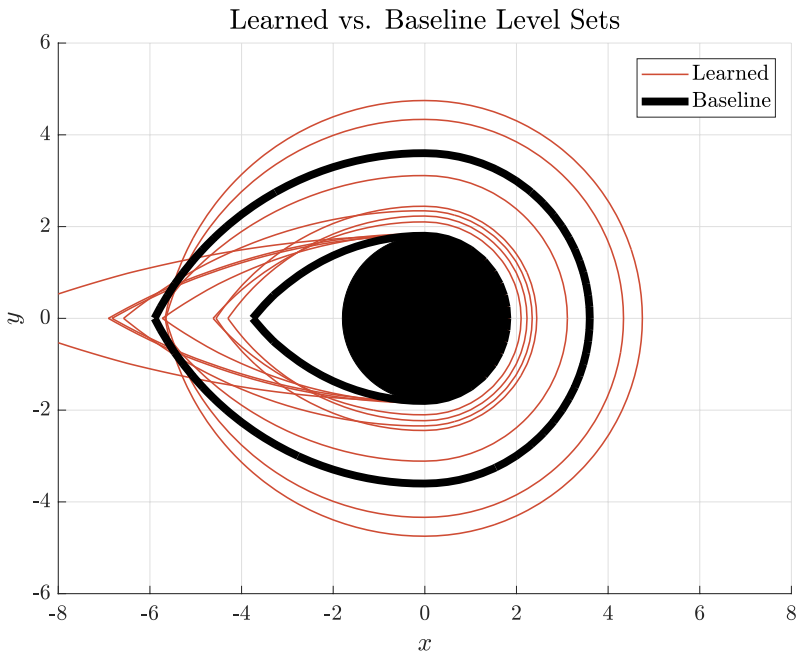

Figure 5.5: Regressed safe sets (viewed on the $\theta = 0$ slice) from supervisor intervention data overlaid on baselines. Three users' safe sets clustered to arcing like the Standard safe set. Three others clustered to arcing like the Conservative safe set. The final five safe sets exhibit a distinct behavior that reflects supervisors' preference for gradual, pre-emptive arcs.

trade-off further by using a conservative baseline that is distinguishably more conservative than user safe sets and by making efficiency more central to the team task.

5.3.3 *Model Validity*

The statistically significant decreases in false positives observed in Phase III agree with the decreases predicted by the supervisor model based on intervention data from Phase II. Our model posits that interventions occur at states noisily distributed about a safe set boundary. Therefore, it predicts that the empirical distribution of Phase II intervention states contained within a proposed safe set (see Fig. 5.6) will mirror the proportion of false positive interventions observed in Phase III: if states are deemed safe by the controller, they will not be avoided, even when the noisy supervi-

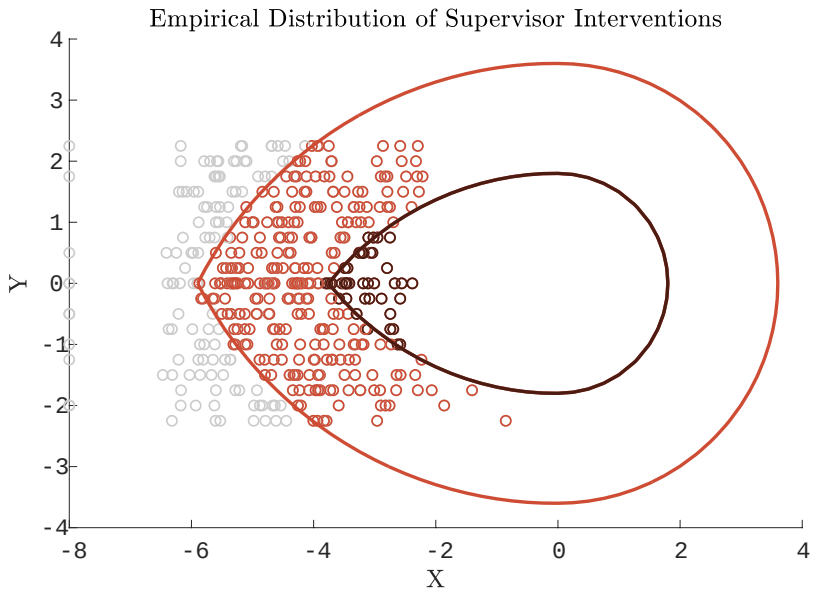

Figure 5.6: Empirical distribution of intervention states observed during data collection (Phase II of the experiment). The interventions within the Conservative reachable set are colored in red, leaving 115 interventions in the corresponding safe set. Similarly, the interventions within the Standard reachable set are colored darker, leaving 397 interventions in the corresponding safe set. Intervention states not contained within a reachable set would have generated a false positive during the human-robot teaming task.

sor would judge them to be unsafe. Since the Learned safe set controller intervenes at the $\hat{\mu}^*$ level set (see Section 4.4), exactly half the intervention states will be contained within the Learned safe set in expectation. The model's predictions are compared against observed false positives in Table 5.1.

5.4 SUMMARY

Our user study demonstrates a significant reduction in false positives over baseline behavior when learning from human safety forecasting. We can model human safety forecasting data-efficiently

	Interventions in Safe Set	Predicted F.P. vs Std.	Average F.P.	Observed F.P. vs Std.
Standard	397 / 440	100%	12.54	100%
Learned	220 / 440	55.4%	7.31	58.3%
Conservative	115 / 440	29%	4.68	37.3%

Table 5.1: Predicted and observed false positives. Left: Predicted false positives from Phase II data. Right: Observed false positives in Phase III.

using a statistical model built on top of reachability analysis for safe sets. We employ the noisy idealized supervisor as the generative model in a learning algorithm to predict supervisor safety judgements, and we present a safety controller for robotic agents that respects the supervisor's perception of safety. This safety controller is guaranteed to reduce false positives for idealized supervisors and real supervisors match those trends.

Automation with human supervisors relies on leveraging the human supervisor's cognitive resources for success. Respecting these resources is essential for creating well performing human-robot teams. It is especially important to avoid overtaxing the human as automated teams continue to scale up, and a single human worker both accomplishes more and bears more cognitive load than ever. To alleviate this burden, we can decrease the number of issues that command the supervisor's attention by reducing false positives.

Our results show that it is possible to reduce false positives, and thus cognitive load, by aligning robot behavior with humans' expectations. Yet since corrective action follows the "minimum intervention safety controller" framework, it also means the human is only informed of safety at the last possible moment. How can we incorporate anticipation into robot motion to inform supervisors in a timely manner? This is the inquiry for the next chapter.

6

COMMUNICATING THROUGH PRAGMATIC OPTIMIZATION

Previous chapters have equipped our machines to consider humans' safety concerns, both in constraints and dynamic forecasting. The next chapters will equip humans to understand the robots' concerns. We can equip human understanding this way through optimizing robot behavior to transparently evidence its concerns.

Prior art in transparent motion largely ignores the robot's dynamics, preferring pure motion planning instead. Yet without dynamic information included about the robot, the robot might not actually be able to execute the plan. Furthermore, this limited scope precludes the ability to express dynamic properties like motion constraints.

This high-level scoping helped make optimizations tractable. For example, the Bayesian perspective introduced by Dragan and Srinivasa [22] drew paths for robot arms that pointed towards where they were reaching. Extending the algorithm to controlling a car made the planning too slow to execute in real-time [12].

This chapter streamlines the legibility objective to have equivalent computational complexity as the original objective to be communicated. This enables any optimal control algorithm that can solve the original task to also solve the legible version of that task. This algorithm's simplicity will enable the next chapter to communicate robot intents beyond just end-states to include characteristics like safety.

This chapter will demonstrate how to replicate the properties introduced in prior art (like legibility, exaggeration, and antici-

This chapter is an adaptation of "An Efficient Understandability Objective for Dynamic Optimal Control" [49] written in collaboration with S. Shankar Sastry

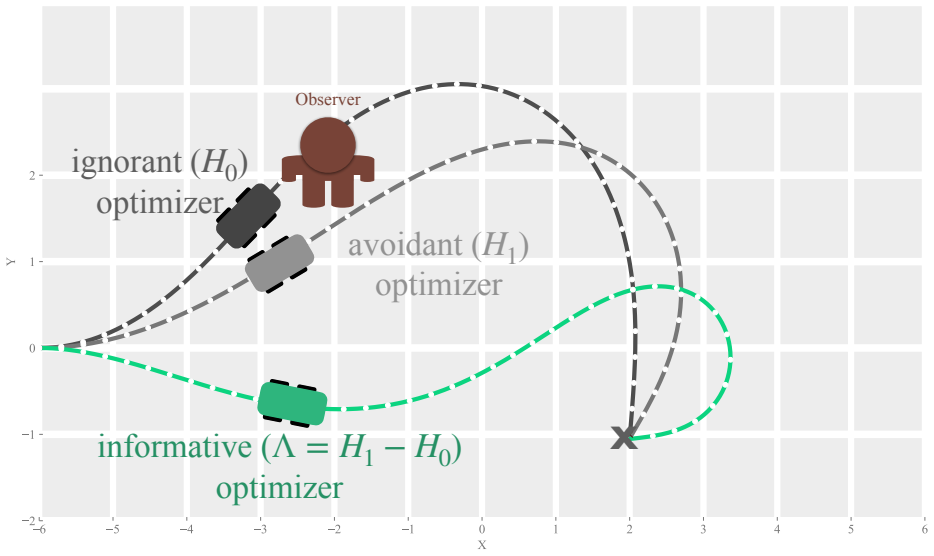

Figure 6.1: *Overview*: A human observes three possible robot motions with bicycle dynamics. The black path (top) is ignorant of needing to avoid the human and its control-minimizing path is a collision course. The gray path (middle) is optimized with a collision avoidance term in addition to minimizing control effort. It's path succeeds in avoiding the human, but to minimize control cost it comes concerningly close. The light green path (bottom) is optimized to evidence its awareness of the human's needs so the human is informed. This legibility optimization metric was historically too complex to apply to non-holonomic dynamics tractably (see conclusions of [12]). This work derives a tractable equivalent metric.

pation) more tractably using a streamlined algorithm that arises from a simpler observer model.

6.1 INTRODUCTION

Social and assistive robots need to communicate their intentions to collaborators. This is transparency. Indeed, studies in the psychology of human-human collaboration [74] emphasize the importance of theory of mind: the ability to reason about what your collaboration partner is thinking. Already roboticists have seen this need in building robots to work around people. Avoiding crashing with humans required robots to forecast their ways [76]. Follow-

ing their users' lead required robots to consider humans' ends and goals [57]. Working together means being mindful of the other.

Of course, humans need to consider the same information about their partner. Even when coordinating something as simple as handing over an object, task intent of "what", "when", and "where" must be synchronized [68]. And for those tasks, robots can help inform handovers by how they hold their hands [10]. Showing your hand this way generalizes to myriad modalities for coordination: through text, speech, graphics, blinking lights, and a multitude of other media [2]. Yet no amount of exclamation can convince users of intention if the robot is not acting that way.

The hard evidence of the robot's choices are the final judge of performance. Safety supervisors must judge critically, wary of bugs and mistakes in perception. For them, the robot's *actions* are ultimately what must be kept in line. Even for humans who are not explicitly in the loop as critical supervisors, they should engage the system with *appropriate* trust. Calibrating their trust to the appropriate level requires critically examining the robot's performance. Therefore the robot's actions will be the critical communication medium.

6.1.1 *Prior Art in Commmunicative Action*

Communication and action are so entwined, our language does not distinguish between acting (for agency) and acting (for expression). Not only does humanity act out their intent, but millennia of crafty puppeteers animate scraps of cloth and wood to act out intents [33]. Unfortunately we cannot afford to employ a puppeteer to run every robot (though we've tried before[1]). Instead we need to automate the animating principles [71]. Beyond learning functionality from human demonstrations as in main-

1 Though the history of artificial motion can be traced through medieval records [77] to antiquity, artificial intelligence has been more elusive. Eighteenth century tech hype marveled at Jacquet Droz's clockwork porcelain dolls that could replicate handwriting, yet they were limited to pre-programmed sentences and sketches [81]. Harnessing the runaway speculation at that time, one Wolfgang von Klempelen claimed to use similar clockwork to play chess (two full centuries before deep blue!). Despite the elaborate showcasing of its internal gears and cogs, its humanlike reactions and behavior turned out to be produced by a human chess master secreted away in the cabinet under the chessboard.

line Learning from Demonstration (LFD) (e.g. [1, 62, 83]), Gielniak introduced extracting specifically communicative keyframes to exaggerate the expression of motion [28, 29]. Whereas previous human-aware motion planning would optimize for human space (whether for following [57] or proxemics [46]), Gielniak emphasized the expressive elements by translating principles of animation like secondary motion [27] into optimizations. Actions not only accomplish goals but aspects of the motion communicate robot purposes [70, 82].

Following the motion optimization approach, Dragan [22] applied the linguistics' inference-based pragmatics models to make an explicitly informative objective. These inference models follow the "theory theory" of cognitive science that humans construct and test models from observed data [31, 32]. Assuming humans read actions from the "teleogical[2] stance" [21] , interpreting actions can be viewed as inferring hidden goals or preferences on actions [6, 7, 36, 45]. Designers of explainable AIs have turned this cognitive model of action understanding into an optimization objective for motion planning [24, 26, 34]. Dragan formalized the distinction between purely optimizing for expected intent versus acting to be interpreted as *predictability* versus *legibility*. Though generating distinct optima for motion, these two objectives mostly align [43] explaining the appeal of imitative approaches like LFD.

This inference-based principle promises to generalize to a variety of morphologies and applications, such as self-driving vehicles [12]. However, this generalization to dynamic control is rare amongst prior art (both amongst inference-based [24, 26, 34] and general understandable motion [14, 46]). The only prior work on applying legibility optimization to nonlinear dynamics [12] emphasized that the algorithm lacked the efficiency to run online. Mixing the belief-space dynamics of the inference objective with the non-holonomic dynamics of continuous control often stymies low-level communicative motion.

This chapter derives a simplified equivalent metric to legibility that is in the identical complexity class as the uncommunicative task. That is, optimizing the communicative motion problem requires no extra complexity over the barebones task.

2 Reading purpose into non-human objects seems to be an instinctual reflex for humans, as our inherited myths are teeming with personalities and motives for everything from trees to rivers to the sun itself.

6.1.2 *Contributions and Outline*

The main contribution of this work is the simplified legibility formulation that will be laid out in this chapter. Along the way, a more powerful hypothesis class for modeling "intent" is introduced in Section 6.3. This broader class opens the door to communicating more than just endpoints which we will explore in Chapter 7.

6.2 MATHEMATICAL BACKGROUND

The robotic motion problem is to select actions u from the set of available choices $U = \mathbb{R}^{n_u}$ to steer the evolving state x in the state space $X = \mathbb{R}^{n_x}$. The actions influence the state through the dynamic difference equation:

$$x(t+1) = \bar{f}(x(t), u(t)), \quad \forall t \in [0, 1, \cdots, T] \tag{6.1}$$

for discrete-time systems or through the dynamics differential equation:

$$\dot{x}(t) = f(x(t), u(t)), \quad \forall t \in [0, T) \tag{6.2}$$

for continuous time systems. For these differently timed dynamics, the time-indexing set ($[0, 1, \cdots, T] \subset \mathbb{Z}$ or $[0, T) \subset \mathbb{R}$) is called the time horizon $\mathcal{T}$ with T being the final time. Let $\mathcal{X}$ be the set of functions $x(\cdot) : \mathcal{T} \to X$ and $\mathcal{U}$ be the set of functions $u(\cdot) : \mathcal{T} \to U$.

Aside: For the rest of this chapter, we will focus on the continuous-time case but the mathematics will apply straightforwardly to discretely-indexed functions. Similarly we focus attention on continuous state and action spaces, but the contributions can be re-derived for discrete state or action spaces with some change in notation.

Actions can be chosen following a variety of paradigms. This chapter follows the optimization-based paradigm for reflecting goal-driven "intelligent" behavior. Here possible action choices, along with their resultant state trajectories starting from some

given initial state $x(0) = x_0$, are ranked by some objective function $J(x(\cdot), u(\cdot))$. Often these objectives are time-decomposable and so can be broken into a running cost $L(x(t), u(t))$ and terminal cost $\phi(x(T))$ as:

$$J(x(\cdot), u(\cdot)) = \phi(x(T)) + \int_0^T L(x(\tau), u(\tau)) d\tau \tag{6.3}$$

The goal is to choose the actions to achieve an extremum of this objective function (a minimum when J is considered a *cost* or a maximum when J is considered a *reward*). This is notated mathematically:

$$\begin{aligned} &\min_{u(\cdot)} J(x(\cdot), u(\cdot)) \\ &\text{subject to} \\ &\qquad \dot{x}(t) = f(x, u, t) \quad \forall t \in \mathcal{T} \\ &\qquad x(0) = x_0 \end{aligned} \tag{6.4}$$

Algorithms focused on planning the path abstract the problem by allowing the agent to directly choose states x. This configuration is equivalent to setting the action space equal to the state space $U = X$ and setting

$$f(x, u, t) = u \tag{6.5}$$

which is a linear, fully controllable dynamic system making it simple to optimize. Applied systems rarely have this luxury, but sacrificing dynamic feasibility is often made in order to tackle more complex objectives, like in joint task-and-motion planning or for non-convex human-factors objectives (like in [46]). This work will simplify the legibility metric enough that it can be optimized with respect to even nonlinear dynamics.

For notational concision and to focus on the actual decision variable $u(\cdot)$, we will combine the cost function $J(x(\cdot), u(\cdot))$ with the constraints $\dot{x}(t) = f(x, u, t) \quad \forall t \in \mathcal{T}$ and $x(0) = x_0$

$$\mathcal{H}_J(u(\cdot)) = J(\rho(u(\cdot), x_0), u(\cdot)), \tag{6.6}$$

using the unique solution map $\rho(u(\cdot), x_0)$ (that exists from the Picard-Lindelof theorem). We will narrow $\mathcal{U}$ to the set of piecewise continuous functions and require that the dynamics $f(x, u, t)$ are Lipschitz continuous in state x, continuous in u and piecewise continuous in t.

6.3 EXTERNAL OBSERVER MODELING

Traditionally, when animating a robot using optimal control, the robotic agent optimizes its behavior to concord with some agenda $J(x(\cdot), u(\cdot))$ (e.g. merge onto a highway without collisions). Yet optimizing just a task objective ignores the needs of external observers working around the robot. Robots do not operate in an informational vacuum. As robots start working in human-spaces, these other agents will observe and interpret the robot's motion to understand their behavior. In this work, we are particularly interested in how humans understand the robot's intended goals. These can be encoded as optimization parameters. So the human's interpretation is deciding between hypothesized optimization problems that might be generating the robot's behavior. These optimizations can be summarized by a metric $\mathcal{H}_{J_i}(u(\cdot))$ corresponding to hypothesis H_i. This chapter focuses on binary hypothesis testing: the human is deciding between a default case H_0 and an alternative H_1. Multiple hypothesis testing can be performed as iterated binary hypothesis testing with corrections (e.g. Bon Ferroni), but this problem is left for future work.

Following the hypothesis testing framework, we assume there is a null hypothesis that the observer will default to believing and an alternative hypothesis that describes the true characteristic of our robotic agent. These could be some essential binary (e.g. safe/unsafe) or a choice between two options (reaching for the left object or the right one). Applications like these are explored further in Chapter 7.

For any binary hypotheses, the uniformly most powerful test that an observer can use to judge the robot's performance is a Likelihood Ratio Test:

$$\begin{aligned} &\text{if } \Lambda(u(\cdot)) \leq \eta && \text{, rejects } H_0 \\ &\text{if } \Lambda(u(\cdot)) > \eta && \text{, fails to reject } H_0 \end{aligned}$$

where the deciding factor is the ratio between the probability of observing the robot's choices given the different hypotheses $P(u(\cdot)|H_i)$:

$$\Lambda(u(\cdot)) \coloneqq \frac{P(u(\cdot)|H_0)}{P(u(\cdot)|H_1)} \tag{6.7}$$

Following cognitive scientific models such as [45, 48], we formulate the hypothesized likelihood of observing a performed motion as an exponential distribution with cost as the sufficient statistic:

$$P(u(\cdot)|H_i) \propto \frac{e^{-\mathscr{H}_{J_i}(u(\cdot))}}{Z_i} \tag{6.8}$$

This distribution models the human observer as expecting optimal behavior (according to their hypothesized $\mathscr{H}_{J_i}$), but leaving sub-optimal behaviors as possible due to their inconfidence in $\mathscr{H}_{J_i}$ [48, 83]. This distribution is referred to as the "Boltzmann distribution" by analogy to the statistical mechanical equilibrium distribution where the negative cost function is interpreted as the energy.

This Boltzmann distribution was used in the prior state-of-the-art [24], but required calculating the normalization constant Z_i. This requires integrating over all possible time-varying controllers:

$$\begin{aligned} P(u(\cdot)|H_i) &= \frac{e^{-\mathscr{H}_{J_i}(u(\cdot))}}{Z_i} \\ &= \frac{e^{-\mathscr{H}_{J_i}(u(\cdot))}}{\int_{u(\cdot)} e^{-\mathscr{H}_{J_i}(u(\cdot))} du(\cdot)} \end{aligned} \tag{6.9}$$

Unfortunately, this is an infinite dimensional space. Dragan and Srinivasa sidestepped this Sisyphean task by approximating it instead by fully solving for the optimal cost-to-go in the control problem. For their focus on unconstrained planning problems

with quadratic costs, this could be done in closed form. Unfortunately, when working with nonlinear dynamics, the problem can no longer be solved analytically.

In the next section we will introduce an equivalent optimization problem that *does not require calculating the partition function Z_i at all*.

6.4 OPTIMIZING CONTROLS FOR INFORMATIVENESS

The model of human action interpretation in Section 6.3 classifies some actions as evidence for the null hypothesis and others as evidence for the alternative. Our robot can optimize its chosen actions to ensure it evidences the correct alternative:

Lemma 6.4.1. *For every likelihood ratio testing observer, the control that optimizes:*

$$\min_{u(\cdot)} \Lambda(u(\cdot)) = \min_{u(\cdot)} \frac{P(u(\cdot)|H_0)}{P(u(\cdot)|H_1)} \tag{6.10}$$

is guaranteed to be evidence to reject the null hypothesis for every obsever with non-empty rejection region:

$$R_{NP} = \left\{ x : \Lambda(u(\cdot)) = \frac{P(u(\cdot)|H_0)}{P(u(\cdot)|H_1)} \leq \eta \right\} \neq \emptyset \tag{6.11}$$

Proof of Lemma 6.4.1. From the assumed property (6.11) there exists an $u_r(\cdot) \in R_{NP}$ that, by definition, satisfies:

$$\Lambda(u_r(\cdot)) \leq \eta$$

Yet this $u_r(\cdot)$ cannot have smaller $\Lambda(u(\cdot))$ than $u^*(\cdot)$ since $u^*(\cdot)$ optimizes Λ. Indeed,

$$\Lambda(u^*(\cdot)) \leq \Lambda(u_r(\cdot)) \leq \eta$$

Therefore the optimizer is in the rejection region for any testing value of η.

■

This optimization turns out to have a simple form in terms of the hypotheses' costs:

Theorem 6.4.2 (Simplified Equivalent of Maximizing Self-Evidence)**.** *The problem of maximizing the observer's likelihood ratio in favor of the alternative hypothesis:*

$$\min_{u(\cdot)} \Lambda(u(\cdot)) = \min_{u(\cdot)} \frac{P(u(\cdot)|H_0)}{P(u(\cdot)|H_1)} \tag{6.12}$$

has the same optima as:

$$\min_{u(\cdot)} \mathscr{H}_{J_1}(u(\cdot)) - \mathscr{H}_{J_0}(u(\cdot)) \tag{6.13}$$

Call this equivalent objective:

$$L(u(\cdot)) \coloneqq \mathscr{H}_{J_1}(u(\cdot)) - \mathscr{H}_{J_0}(u(\cdot))$$

Proof of Theorem 6.4.2.

$$\begin{aligned}
\min_{u(\cdot)} \Lambda(u(\cdot)) &= \min_{u(\cdot)} \frac{P(u(\cdot)|H_0)}{P(u(\cdot)|H_1)} \\
&= \min_{u(\cdot)} \frac{\frac{e^{-\mathscr{H}_{J_0}(u(\cdot))}}{Z_0}}{\frac{e^{-\mathscr{H}_{J_1}(u(\cdot))}}{Z_1}} \\
&= \min_{u(\cdot)} \frac{Z_1}{Z_0} \frac{e^{-\mathscr{H}_{J_0}(u(\cdot))}}{e^{-\mathscr{H}_{J_1}(u(\cdot))}} \\
&= \min_{u(\cdot)} \frac{Z_1}{Z_0} e^{-\mathscr{H}_{J_0}(u(\cdot)) + \mathscr{H}_{J_1}(u(\cdot))}
\end{aligned}$$

An objective can be composed with any non-decreasing function and have equivalent optima. Since the partitions Z_i are non-negative and the logarithm is non-decreasing we can compose an equivalent objective as:

$$\begin{aligned}
L(u(\cdot)) &\coloneqq \log\left(\frac{Z_0}{Z_1} \Lambda(u(\cdot))\right) \\
&= \log\left(\frac{Z_0}{Z_1} \frac{Z_1}{Z_0} e^{-\mathscr{H}_{J_0}(u(\cdot)) + \mathscr{H}_{J_1}(u(\cdot))}\right) \\
&= \log\left(e^{-\mathscr{H}_{J_0}(u(\cdot)) + \mathscr{H}_{J_1}(u(\cdot))}\right) \\
&= -\mathscr{H}_{J_0}(u(\cdot)) + \mathscr{H}_{J_1}(u(\cdot)) \\
&= \mathscr{H}_{J_1}(u(\cdot)) - \mathscr{H}_{J_0}(u(\cdot))
\end{aligned} \tag{6.14}$$

■

Therefore the equivalent objective $L(u(\cdot))$ defined in Equation 6.14 can be used instead of the likelihood ratio, thereby avoiding calculating partition functions Z_i. This goal can be understood as aiming to improve the alternative hypothesis' cost while performing worse at the null hypothesis' cost. The simple linearity of Equation 6.14 makes informative control as tractable as the original optimization:

Corollary 6.4.2.1 (Time Complexity of Theorem 6.4.2's objective)**.** *For any gradient-based control optimization method, finding the optima of the likelihood ratio inherits the order of time-complexity from the original (non-communicative) optimizations in equation (6.4) (whichever has higher time-complexity)* .

Proof of Corollary 6.4.2.1. The optima of the likelihood ratio can be found as the optima of Equation 6.14. Because of the linear form of Equation 6.14, all queries to the objective function and its derivatives must only query the two hypotheses' respective lookups and combine them. Therefore the computational complexity of each iteration will be at most the sum of the two hypotheses' complexities. Thus they will share the same growth-rate/order of time-complexity.

■

The hypotheses' rewards are combined linearly with equal weights in Theorem 6.4.2. However, if the designer desires the robot to not only optimally communicate but also optimize the original reward, more weight on the H_1 term can be added in:

$$\begin{aligned} &\min_{u(\cdot)} L(u(\cdot)) + \alpha \mathcal{H}_{J_1}(u(\cdot)) \\ &= \min_{u(\cdot)} (1+\alpha)\mathcal{H}_{J_1}(u(\cdot)) - \mathcal{H}_{J_0}(u(\cdot)) \end{aligned} \tag{6.15}$$

This weight α could be interpreted formally as a Lagrange multiplier on a sub-optimality bound as suggested in Section VI of [24]. This relative weighting between the original optimization and the informativeness objective can be dynamically shifted to create another desirable property: anticipativeness.

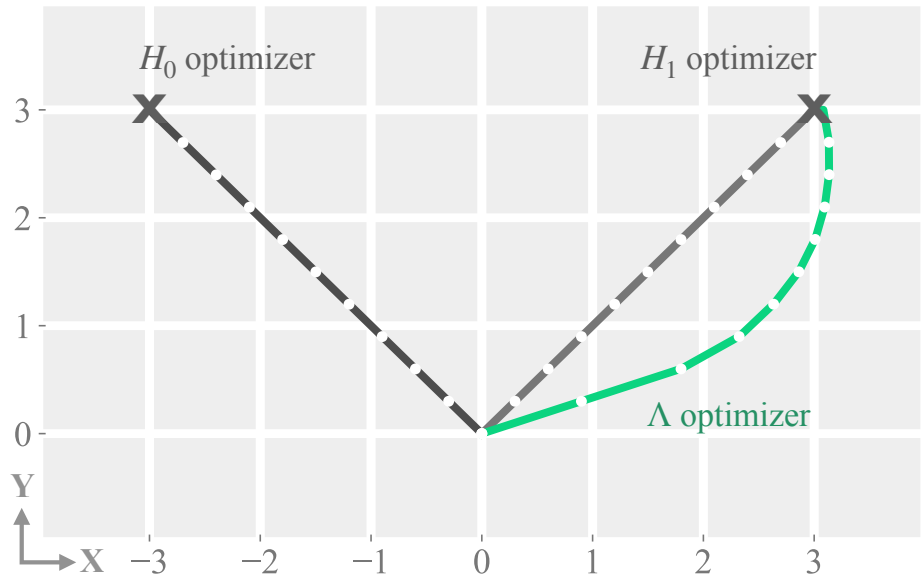

Figure 6.2: Optimized paths through x-y space reaching for either the leftwards destination (H_0) or the rightwards destination (H_1). The anticipative trajectory (in green) that optimizes Λ leads rightwards early; as opposed to the non-anticipative trajectories (in gray) which indicate much more slowly. The H_1 trajectory takes three times longer to move rightwards to $x = 2$

6.5 ANTICIPATION THROUGH RECEDING HORIZON CONTROL

Previous work emphasized the importance of communicating intent earlier in the motion. Gielniak and Thomaz [28] promoted the concept of "anticipativeness" and tweaked motions to express salient gestures earlier on in the time horizon.

Dragan and Srinivasa [24] incorporated this concept of anticipativeness by ensuring that unfinished viewings of the motion plan would also push the Bayesian observer towards the correct conclusion. They formulated this early expressiveness by optimizing for all incomplete viewings on time horizons $[0, t]$ for $t \in [0, T]$ simultaneously. They chose to combine these multiple sub-objectives through a weighted linear combination:

$$\int_{t=0}^{T} (T - t) P(H_1 | u(0 : t)) dt$$

This sum of probabilities again requires weighting by the normalizing partition constants. Instead, an equivalent prioritization of communication for earlier actions can be accomplished through receding horizon control. For earlier horizons more weight can be placed on informativeness through lower α in Equation 6.15. As we approach the final horizons, we can gradually shift more α-weight to prioritizing the original task.

Concretely, let the problem be replanned at times $\bar{t}_0, \bar{t}_1, \cdots \bar{t}_M$. Let $\alpha(t)$ be some increasing function of time that will prioritize efficiency over communicativeness in later replanning horizons. Let $\bar{u}_i(\cdot)$ be the optimal controls from one of M optimization problems on horizons $[\bar{t}_i, T)$

$$\bar{u}_i = \arg\min_{u(\cdot)} (\alpha(\bar{t}_i) + 1)\mathcal{H}_{J_1}(u(\cdot)) - \mathcal{H}_{J_0}(u(\cdot)) \tag{6.16}$$

The robot will follow the controls from $\bar{u}_i(\cdot)$ for all times $\bar{t}_i \leq t < \bar{t}_{i+1}$.

Whereas the original path planning problem would have been too expensive to replan, the streamlined objective in Theorem 6.4.2 is tractable enough to replan online. In fact, for the problem in Dragan and Srinivasa [24] it becomes a Linear Quadratic Regulator (LQR). This is because they use the path-planning approximation of Equation (6.5) for the dynamics and a quadratic penalty:

$$J_i(x(\cdot), u(\cdot)) = \|x(T) - g_i\|_2^2 + \int_0^T \|u(t)\|_2^2 dt \tag{6.17}$$

where g_i was one of two goals the robot could be reaching to grab.

This receding horizon anticipation optimization was used to recreate Dragan and Srinivasa's [24] exaggerated arcing path plans, as seen in Figure 6.2. Unlike that method, the solutions could be found analytically thanks to the streamlined objectives of Equation (6.13) leaving the LQR structure intact.

The above result focuses entirely on differentiating between the true concept and the distractor hypothesis. This can be useful in cases where there are two clear hypotheses (e.g. left or right, safe or unsafe, ignorant or corrected). If there are multiple distractor

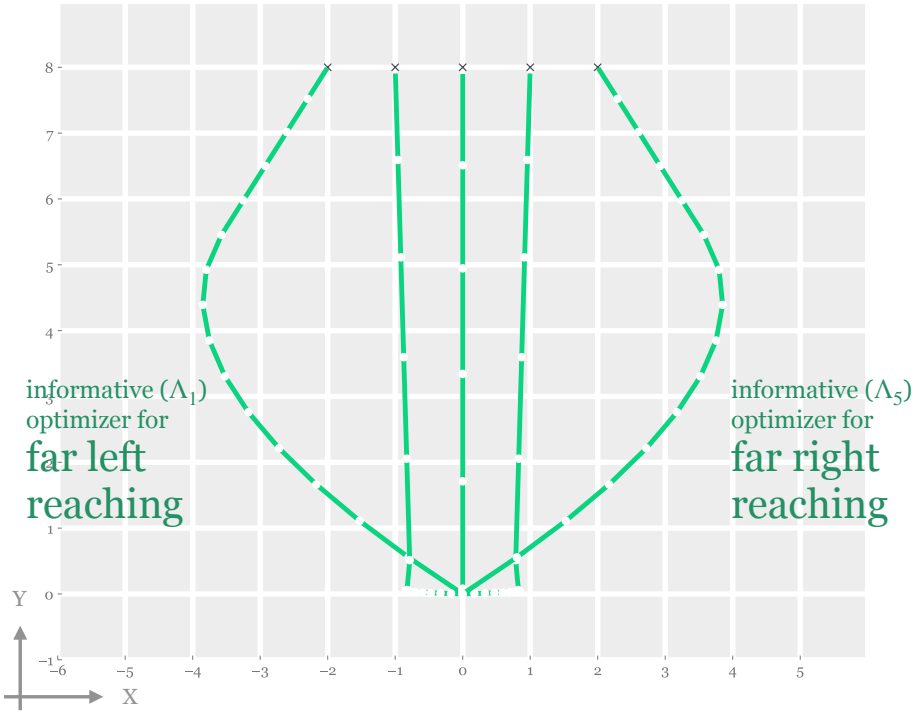

Figure 6.3: Paths through x-y space each optimized to evidence one of five reaching targets reaching. The anticipative trajectories for the far ends Λ_1 and Λ_5 each exaggerate motion out in their respective direction. Meanwhile the anticipative trajectories for the interior goals (Λ_2, Λ_3, and Λ_4) have their exaggeration hemmed in, instead anticipating via a sharp juke early and then straight shooting to the goal after alignment.

hypotheses $H_{0,i}$ for $i \in [1, N]$, informativeness could be encoded as constraints ensuring correct decisions along each pair:

$$\begin{aligned} &\min_u H_1(u) \\ &\text{subject to } H_1 - H_{0,1} \leq \eta_1 \\ &\qquad\qquad H_1 - H_{0,2} \leq \eta_2 \\ &\qquad\qquad\qquad\vdots \\ &\qquad\qquad H_1 - H_{0,N} \leq \eta_N \end{aligned}$$

This approach still optimizes the original objective while bounding the confusion on each pairwise hypothesis test between the truth and a distractor. Figure 6.3 demonstrates this optimization for five different true objectives H_1 corresponding to five different reach targets.

A Lagrangian analysis reveals that this multi-objective optimization will still reduce into a linear combination of the hypotheses' objectives as in the binary hypothesis case. Indeed, inspection of Dragan's gradient derivations in Equation 11 of [24] shows this underlying linear structure but required the coefficients be calculated explicitly via solving for the optimal cost-to-go over multiple horizons. Their fixation on optimizing for multiple horizons simultaneously obfuscated the intuitive structure from the linear combination: that evidencing the true objective from the distractors just requires performing well on the true objective and poorly on the distractors.

6.6 PREDICTABILITY VERSUS LEGIBILITY DISCUSSION

Section 6.4 proved the interchangeability of motion communicativeness with the streamlined objective $L(u(\cdot))$ along with time-complexity and outcome guarantees. The linear structure of $L(u(\cdot))$ also clarifies a disagreement in the literature on the relation of "predictability" and "legibility". Dragan and Srinivasa [24] emphasized the distinction between "predictability", defined as optimizing expectedness given the task $H_1(u(\cdot))$, and "legibility", defined as optimizing making the task clear which in our binary hypothesis setting is $\Lambda(u(\cdot))$. Crucially they state in [22]:

> "Predictability and legibility are fundamentally different and often contradictory properties of motion."

Lichtenthaler and Kirsch questioned this contradictoriness and concluded that "the two factors are coherent" [43].

Our work clarifies the exact relation between "predictability" $H_1(u(\cdot))$ and "legibility" $\Lambda(u(\cdot))$ through the simplified relation stated in Theorem 6.4.2. By reformulating the legibility problem from $\Lambda(u(\cdot))$ into the equivalent, simplified linear combination $L(u(\cdot))$ in Eq. 6.13, we can state the difference between predictability and legibility quite simply: legibility $L(u(\cdot))$ increases lin-

early with predictability $H_1(u(\cdot))$, meaning they are indeed coherent and typically correlated as [43] asserts, while also having an uncorrelated term (equal to adding in $-H_0(u(\cdot))$) that will cause the legibility optimizers to be fundamentally different from predictability optimizers.

6.7 SUMMARY

Prior art in communicative motion shifted motion planning away from the view that the robot is in an informational vacuum. Instead, the motion is observed and interpreted to infer latent intents. This view shifts the centering from machine performance to its impact on the human-machine collaboration.

And taking the human-centered perspective does not need to be costly. The derivations in this chapter showed how moving communicatively can be done for practically no extra computational cost. We were able to achieve this by removing the fixation on optimizing for multiple observation windows simultaneously that bloated Dragan's algorithm [24]. By focusing on a simple hypothesis testing framework we revealed the intuitive linear combination structure of evidential motion. From first principles of cognitive scientific models[3] we rigorously derived the simple intuition that communicating one task hypothesis over another simply requires performing well by the first and poorly on the other.

This chapter's approach to evidential motion was able to recreate the desirable properties identified by previous work [24, 28] through a more efficient objective. We will see in the next chapter how this simplicity will allow for easy extension to broader messages and optimization structures including true control.

3 such as Baker [7] or Jara-Ettinger [36]'s "naive utility calculus" for action understanding

7

EVIDENT SAFETY IN CONTROL

The previous chapter introduced the hypothesis testing model to streamline communicative motion optimization. This simplified framework makes it straightforward to extend the communicated payload from end-state goals to broader classes of "intent". We expand the definition of "intent" to include any parameter of the optimal control problem, thereby opening the door to extend communications to running preferences or even hard capabilities or safety constraints. Beyond replicating previous art with greater efficiency, the legible model predictive control (LMPC) algorithm can solve entirely new problems. We can now communicate back full understandings of corrected task specifications during value alignment interactions as in those proposed by Jain [35] or Bajcsy [5], closing the loop for robots to dialog with users about their preferences.

Of particular interest to our goal of *mutual understanding for safety*, is to share understandings of the constraints learned by methods like those in Part i. Choosing actions to provide *evidence* of safety well ahead of critical response times[1] assuages concerns and empowers humans to plan on the machine's safety compliance.

7.1 RESPONSIVENESS TO UPDATES

The previous chapter extended the expressible payload from just end-points (as in Dragan and Srinivasa [24]) to include the dy-

This chapter is an adaptation of "An Efficient Understandability Objective for Dynamic Optimal Control" [49] written in collaboration with S. Shankar Sastry

1 cf. Chapter 5

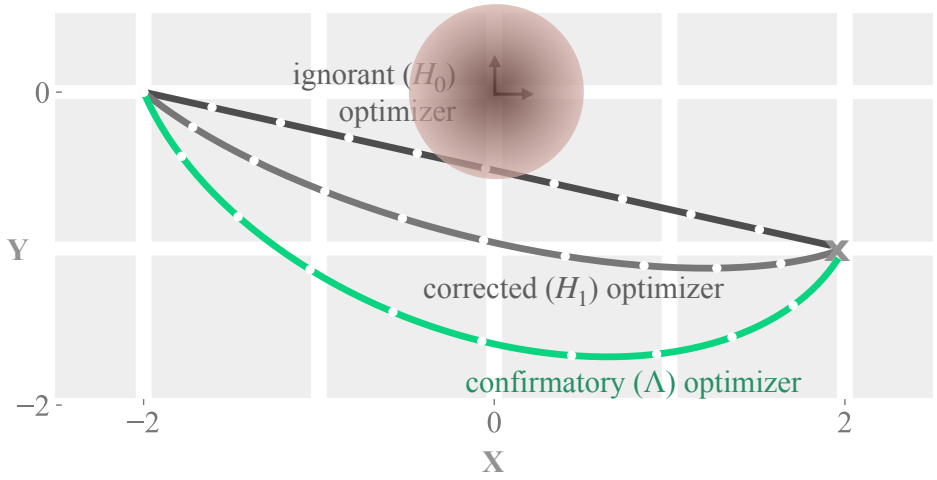

Figure 7.1: Optimized paths in x-y space: After the instructor corrects the robot to avoid the red region around the origin, the robot must demonstrate its new understanding. The dark path is the optimum pre-correction H_0 (from Equation 7.1), the gray path is the optimum post-correction H_1 (from Equation 7.2), and the green path is the optimum for informing the corrector of the successful correction Λ; all here with $g = [2, -2]^T$, $a_1 = 40$, $a_2 = 25$, $a_3 = 1$.

namic cost and constraints as well. This opens the door to communicate differences in running costs as in [34] but through communicative motion instead of scenario generation. This could help improve in-task teaching (like the physical in-task value alignment in [5]) by confirming whether the lesson was learned correctly, thereby completing the communication loop proposed in [2]'s roadmap. Figure 7.1 takes the use case of [5] where the user corrects the robot and adds a penalty on approaching the obstacle at the origin:

$$J_0(x(\cdot), u(\cdot)) = a_1 \|x(T) - g\|_2^2 + \int_0^T a_2 \|u(t)\|_2^2 dt \tag{7.1}$$

$$J_1(x(\cdot), u(\cdot)) = a_1 * \|x(T) - g\|_2^2 + \int_0^T a_2 \|u(t)\|_2^2 - a_3 \|x(t)\|_2^2 dt \tag{7.2}$$

By accentuating the alternative hypothesis H_1 over the null hypothesis H_0 the robot clarifies whether or not it's understood the correction. If the human likes the robot's understanding of J_1

they can rest easy. If they find the new behavior still unsatisfactory, they are now informed to know what else must be added.

7.2 EVIDENT SAFETY FOR NONLINEAR SYSTEMS

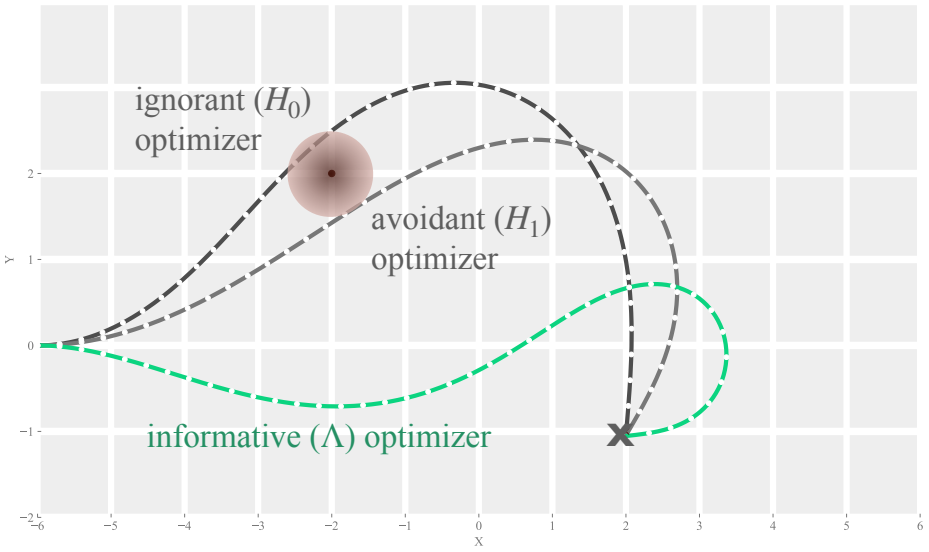

Figure 7.2: The informative control optimization can even apply to nonlinear dynamics. After adding a quadratic penalty to nearing state $[2,-2]^T$, the avoidant optimum to H_1 (in gray) indeed has a farther *integral* than the ignorant optimum to H_0 (in black), but the path still looks qualitatively the same. In contrast, the informative optimizer to Λ makes its avoidance obvious. Here $g = [2,-1]^T$, $h = [-2,2]^T$, $a_1 = 800$, $a_2 = 10$, $a_3 = 2$.

In communicating full optimizations rather than just endpoints we can now communicate running tasks such as obstacle avoidance. Following a potential field approach [59], we could encode the constraints identified in previous chapters with a repellent cost around the blocked states.

The null hypothesis minimizes the control effort added to the final distance from the goal $x-y$ point $g = [2,-1]$ (it is agnostic to angle), while the alternative hypothesis also quadratically penalizes proximity to an undesirable $x-y$ position at $a = [-2,2]$:

$$J_0(x(\cdot), u(\cdot)) = a_1 \|Px(T) - g\|_2^2 + \int_0^T a_2 \|u(t)\|_2^2 dt \tag{7.3}$$

$$J_1(x(\cdot), u(\cdot)) = a_1 \|Px(T) - g\|_2^2 + \int_0^T a_2 \|u(t)\|_2^2 - a_3 \|Px(t) - h\|_2^2 dt \tag{7.4}$$

where P is the projection from the three dimensional state of planar position and angle down only to planar position:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Not only does the reformulation in Chapter 6 allow communicating differences in running *costs*, it is also built to apply to systems with running dynamic *constraints*. And the new formulation in Equation 6.14 is lightweight enough to be tractable for the numerical methods often necessary for nonlinear optimal control.

This capacity is demonstrated in Figure 7.2 for the following example non-holonomic dynamical control problem. The streamlined objective in Equation 6.14 can be optimized for nonlinear dynamics using established nonlinear control frameworks, for example the iterative Linear Quadratic Regulator approach [73][72]. Consider the three dimensional Dubins vehicle with constant velocity $v = 3$:

$$\frac{d}{dt} x(t) = \begin{pmatrix} v \cos(x_3(t)) \\ v \sin(x_3(t)) \\ u(t) \end{pmatrix} \tag{7.5}$$

7.3 DISCUSSION

By viewing motion not only as needing to be safe but also doubling as *evidence* for an external validator to observe, we were able to generate motion that *anticipates* safety well ahead of time (as seen in Figure 7.2). This improves over the the avoidance algorithms in Chapter 5 that only demonstrated safety *exactly on* the human's critical decision boundary. This inability to assuage concerns ahead of the point when humans needed to make decisions caused the robot's to still trigger some false positives. Now

by considering the informativeness of their motions, robots can address supervisor's concerns ahead of time while still balancing efficiency.

We streamlined the "legible" control problem to be on the exact same order of complexity as the original (uncommunicative) control problems and demonstrated some communicative motions. Opening communicative motion to new optimization frameworks and new applications also opens potentials for future research in human modeling and communication.

7.3.1 *Human Modeling*

This work derived a simplified objective for choosing controls that will communicate the robot's task-intent. And every communication requires assuming how receivers will interpret the signals. We have laid out our assumed model based on the receiver testing optimally (i.e. uniformly most powerfully) with respect to binary Boltzmann hypotheses. Which in turn are optimal distributions around a known characteristic reward function with random hidden preferences [48, 83]. Yet this is not the only tenable receiver model.

There is a rich space of alternative decision models for human observers. In particular, humans rarely have infinite computation resources to judge and so will likely use heuristics (e.g. truncated cost-to-gos on their time horizon judgements, or not consider alternative outcomes fully and clip the infinite support of the Boltzmann distribution Equation 6.8). And even with infinite computational resources, decisions may be skewed by risk-averseness (like in [61]) if their decision has tangible outcomes (e.g. whether or not to trust an oncoming autonomous vehicle with their safety).

7.3.2 *Communication Extensions*

Casting the hypotheses as differentiating between full optimal control tasks allowed us to communicate about more than just endpoint states. For example, when users issue a new command to their robot, the robot's should express their understanding of the new goal. In Bajcsy and Losey's conception [5], physical cor-

rections should be interpreted as updates to the robot's running reward function. These requests for better value alignment can be transparently respected by contrasting the old task's optimization with the corrected task's formulation. Future work should explore what cognitive phenomena arise in how users perceive and plan around systems that communicate reception of their inputs.

Generalizing the hypotheses from destinations to full optimal control problems also opens the possibility of communicating what *constraints* the robot is bound to through motion. This framework can provide a new perspective on communicating capability as Kwon et al. prompted in [41]. This constraint communication could also be used to inform a supervisor whether or not the robot is obeying safety rules. The hard *evidence* of performed motion may be a uniquely suited communication channel to debug system failures; as the old adage goes "actions speak louder than words". For constraint communicating to work, future work must develop what it means to subtract reward-hypotheses that don't share the same support. After deriving, we can explore cognitive phenomena on safety-critical judgement (such as risk-sensitivity [61]).

7.4 SUMMARY

The more efficient objective derived in the last chapter let us communicate dynamic properties rapidly in control, unlocking more complex systems such as the illustrative Dubins car. Furthermore, by extending prior art in optimization-based communicative motion by broadening the subject to be conveyed from endpoint goals (as in [24, 26]) to full optimal control problems, we can express dynamic properties such as ongoing safety (illustrated with a obstacle avoidance cost function as promoted by the "elastic bands" [59] or potential field approaches [79]).

The more efficient hypothesis testing model for communicative motion was able to extend anticipation and exaggeration to [24, 28] across optimal control frameworks like LQR, Iterative Linear Quadratic Regulation (iLQR), and MPC. Though simplified models of intelligent behavior like these may sacrifice some detailing features (in this chapter's case, neglecting the Bayesian view decreases the ability to model observers' prior beliefs), the detail is

traded for tractability that enables more powerful algorithms to simply integrate the findings. Though not demonstrated here, we contend that control approaches ranging from dynamic programming to deep reinforcement learning approaches can all shift to choosing communicativeness with this simple difference objective.

Part III

NEXT STEPS

8

LOOKING FORWARD

Mutual understanding is the foundation of solid collaboration, yet our autonomous agents' inner workings are scarcely understood by their users. These users' understandings are constructed from observing machine behavior, so that autonomous behavior could be optimized to inform collaborators of key planning parameters. Yet knowing what information the human needs requires understanding how they perceive and reason. This thesis' research equips robots with cognitive models to contribute to group learning and activity. Looking forward, we foresee future research programs that can design machines to data-efficiently model human cognition across applications. With these cognition models, autonomous agents can adapt optimization choices to inform humans how to work with our machines.

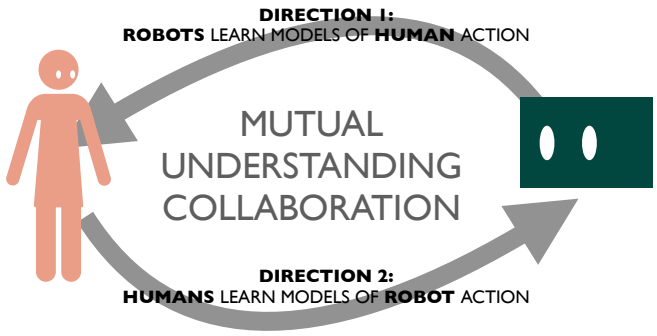

Figure 8.1: The two thrusts of mutual action understanding needed for human-AI-machine collaboration: machine learning models of human activity and human learning models of robot activity. Together these thrusts collaborate to create a virtuous cycle; if our robots understand human learning processes then we can optimize actions to support that learning.

Figure 8.2: Observing human driving behavior on one-tenth scale vehicle in motion-capture track for the Robot Autonomous Racing (ROAR) project for constraint inference research as in Chapter 2

8.1 MACHINE LEARNING MODELS OF HUMAN BEHAVIOR

Insofar as artificial intelligence algorithms are *intelligent* they can double as approximations[1] to sketch human intelligence. With these approximations in hand, machines can now consider how humans think, perceive, and make decisions. That is to say, machines can be considerate. For example, Chapter 4 used the mathematics of formal safety guarantees to model human supervisory behavior which reduced the number of false alarms our robots generated. Powerful tools from machine learning and AI optimization can become mathematical representations of cognition; formalisms as varied as random utility models from econometrics[2] to likelihood threshold testing from communications theory[3] all form the language for our machines to understand humanity. These parsimonious cognition models are data-efficient enough to adapt online to each unique users' needs[4], resulting in artificial intelligence behavior that ergonomically conforms to each individuals' cognition.

1 Importantly we contend that these models are *useful* for rough forecasts instead of underlying the true neural structure of human thought. It is unlikely humans solve computations to think. Yet letting computers respond to computational *models* of humans can still generate considerate behavior.
2 cf. Chapter 2
3 cf. Chapter 6
4 cf. Chapter 4

In Chapter 2 we saw how optimal choice models could be inverted to explain expert avoidance. As urban planners test novel designs for sharing the streets, driving behavior will adapt slowly due to unfamiliarity or resistance. Vehicle movement data reveals how drivers are responding to street design interventions like protected bike lanes or pedestrian yielding. We can quantify what priorities and constraints drive motorists' behavior through inverse optimal control [67]. Inferring these constraints on motorists' behavior can quantitatively describe how citizens understand new rules and how many follow them.

Modeling human needs can also augment clinical diagnostics with quantifiable interpretations. In collaboration with UCSF's Musculoskeletal Research Consortium, we are translating these models to examine neurological recovery through motion data. Current practice in clinical assessments of post-stroke movement recovery are confounded by changes in strength and patients' compensation strategies. Research developments are providing measures and insight on the movement smoothness but stops short of examining how the observed movements are linked to underlying changes in multi-joint motor control. Our statistical models can explain movement stereotypes dynamically and can be generated from smaller datasets enabling them to be integrated into clinical practice. If successful, this work can improve the underlying models of motor recovery as well as tracking and treatment in the hospital and clinic. Human modeling will support doctors' diagnostics with explicable machine learning and parsimonious models.

8.2 INFORMING HUMAN LEARNING OF MACHINE BEHAVIOR

With these learned models tailored to specific human behavior, our robots and AI systems can optimize their performance to support human thinking. Chapter 5 demonstrated how just replicating safety interventions learned from human supervisors can preempt concerns and decrease false alarms. Chapter 6 showed how optimizing motion against models of evidence-based validation can inform users of updated priorities or safety awareness. We can extend communicative motion to support human understanding in our public infrastructure. Equipping these applica-

tions with interpretations of human behavior provides some quantifiable approaches to explore the idiosyncrasies of human judgement.

Communicative motion research will increase fluent flow of traffic for self-driving cars. Contemporary self-driving vehicles cause accidents predominantly through being rear-ended or side-swiped. Far from being pure "human error", these accidents are prompted by uninterpretable autonomous motion and must be corrected by optimizing motion that considers observers' informational needs. Future research in this area can optimize driving motion to evidence braking to improve other motorists' reaction times and decrease rear-end rates. Evidencing plans through motion is also crucial for successful unprotected left turns: a major stall point for autonomous driving. With models of how human perceive ongoing safety, we can quantify the informational value of behavior like nosing into an intersection, enabling coordination needs to be incorporated into motion planning alongside efficiency optimization and safety guarantees. This traffic flow setting will allow us to research the structure of how concepts like *right-of-way* can be formally incorporated into our systems.

Beyond optimizing autonomous vehicles (AVs) to support safe *motorists*, our AVs must also yield city streets to residents' needs for walking, wheelchairing, and bicycling. These more exposed road users need dependable guarantees that our machines are mindful of their safety needs. It is critical that our vehicles not only yield to humans but that humans can plan on that priority. By modeling pedestrians' safety concerns and fitting to their inference idiosyncrasies, we can reliably help road users know they are safe. Supporting human understanding in our shared public places will guarantee there is always space for humans.

Studying this safety collaboration in the streets with evidence thresholds like those exposited in Chapter 6 can let us investigate the variance of wariness, anxiety, or trust across the public. Expanding the safety judgement models from Chapter 5 or the evidence gathering models from Chapter 6, we can explore cognitive phenomena like prospect theory or recency bias through curbside interactions.

By researching how humans come to understand machines' capabilities and constraints, we can also advance how we teach technology to students. Active interaction with machines' operations

is already a cornerstone of engineering education through both laboratories and design studios. With algorithms that generate evidence for understanding machines like those in Chapter 6, machines can automatically generate informative demonstrations for students' inquiries. By introducing quantifiable statistics like the evidence thresholds for hypothesis acceptance, we can investigate how students support hypotheses and how previous experience in related areas can transfer. Elaborating on these hypothesis testing models, we can explore how students might discover the need for new hypotheses and begin to inquire how these novel explanations are invented. We can improve technology and engineering education through AI that explicitly informs human understanding.

8.3 SUMMARY

Following from the mathematical models presented in this thesis, future research can nurture societal systems to support human understanding and agency across healthcare, transportation, and education.

- The struggles of patients recovering limb mobility will shed light onto their specific blockages; in contrast to label-based deep learning black-boxes, seeing patients as agents with a few impediments produces parsimonious interpretations as localized constraints. By using interpretable models, we can partner *with* healthcare providers' knowledge and increase understanding altogether.

- Interactions at dangerous intersections can be grounded in the perspectives and confusions of road-users, connecting city planners to users' needs through data. Entering into that same ecosystem, robotic vehicles can consider other road-users' perceptions and work to improve their understanding for smooth flow and assurance.

- Active learning for technology education can be amplified through intelligent machines that generate evidence for students' inquiries.

9

TOWARDS MUTUAL UNDERSTANDING

This thesis demonstrated how benefitting from human expertise requires conforming to each person's unique cognitive needs. From Chapter 5, we see how each operator's divergent thinking requires distinct actions. And in Chapter 4 we saw how we could data-efficiently conform to each individual thanks to the rich structure borrowed from formal control verification. We exhort you, reader, to continue leveraging the rich literature of intelligent control to approximate human behavior in mathematical terms.

This modeling can statistically make sense of datasets on human behavior to inform decisions. These behavioral sketches in computable terms will also equip algorithms with a "theory of mind". Once our machines understand the process, they can take actions to support it. By considering human thinking, our dynamic safety-critical systems can take action to augment human intelligence and decision-making. From Chapter 6 we saw how straightforward it is to take action to inform human judgement. Simply modeling how machine's actions inform human thinking in new applications can give improved joint decision-making for little computational complexity. Consider what judgements users in your system need to make, draw an analogy to how statistics would make that judgement, and furnish what the algorithm would need in their shoes: people need at least that much.

With the mechanisms of contemporary machine learning and intelligent controllers, we can formulate cognitive scientific principles into computable models. These empower robots to consider human needs for understanding and decision-making and optimize their own actions to support humanity's learning. By supporting understanding, our machines become usable and useful, empowering human activity rather than replacing it.

BIBLIOGRAPHY

[1] Pieter Abbeel and Andrew Y Ng. "Apprenticeship learning via inverse reinforcement learning." In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 1.

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. "Explainable agents and robots: Results from a systematic literature review." In: *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for AAMAS. 2019, pp. 1078–1088.

[3] Leopoldo Armesto, Jorren Bosga, Vladimir Ivan, and Sethu Vijayakumar. "Efficient learning of constraints and generic null space policies." In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 1520–1526.

[4] Anil Aswani, Humberto Gonzalez, S Shankar Sastry, and Claire Tomlin. "Provably safe and robust learning-based model predictive control." In: *Automatica* 49.5 (2013), pp. 1216–1226.

[5] Andrea Bajcsy, Dylan P Losey, Marcia K O'Malley, and Anca D Dragan. "Learning robot objectives from physical human interaction." In: *Proceedings of Machine Learning Research* 78 (2017), pp. 217–226.

[6] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. "Bayesian models of human action understanding." In: *Advances in neural information processing systems*. 2006, pp. 99–106.

[7] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. "Action understanding as inverse planning." In: *Cognition* 113.3 (2009), pp. 329–349.

[8] Victoria A Banks, Katherine L Plant, and Neville A Stanton. "Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016." In: *Safety science* 108 (2018), pp. 278–285.

[9] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. "Simulation as an engine of physical scene understanding." In: *Proceedings of the National Academy of Sciences* 110.45 (2013), pp. 18327–18332.

[10] Aaron Bestick, Ruzena Bajcsy, and Anca D Dragan. "Implicitly assisting humans to choose good grasps in robot to human handovers." In: *International Symposium on Experimental Robotics*. 2016, pp. 341–354.

[11] Bart van den Broek, Wim Wiegerinck, and Hilbert Kappen. "Risk sensitive path integral control." In: *arXiv preprint arXiv:1203.3523* (2012).

[12] Tim Brüdigam, Kenan Ahmic, Marion Leibold, and Dirk Wollherr. "Legible Model Predictive Control for Autonomous Driving on Highways." In: *IFAC-PapersOnLine* 51.20 (2018), pp. 215–221.

[13] Aircraft Accident Investigation Bureau. "Aircraft Accident Investigation Preliminary Report: Ethiopian Airlines group. Boeing 737-8 (MAX); registered ET-AVJ, 28 NM South East of Addis Ababa, Bole International Airport, March 10, 2019." In: *Addis Ababa, Ethiopia: Federal Democratic Republic of Ethiopia, Ministry of Transport, Aircraft Accident Investigation Bureau* (2019).

[14] Baptiste Busch, Jonathan Grizou, Manuel Lopes, and Freek Stulp. "Learning legible motion from human–robot interactions." In: *International Journal of Social Robotics* 9.5 (2017), pp. 765–779.

[15] Margaret P Chapman, Kevin M Smith, Victoria Cheng, David L Freyberg, and Claire J Tomlin. "Reachability analysis as a design tool for stormwater systems." In: *2018 IEEE Conference on Technologies for Sustainability (SusTech)*. IEEE. 2018, pp. 1–8.

[16] Margaret P Chapman et al. "A risk-sensitive finite-time reachability approach for safety of stochastic dynamic systems." In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 2958–2963.

[17] Glen Chou, Dmitry Berenson, and Necmiye Ozay. "Learning constraints from demonstrations." In: *International Workshop on the Algorithmic Foundations of Robotics*. Springer. 2018, pp. 228–245.

[18] Glen Chou, Necmiye Ozay, and Dmitry Berenson. "Learning parametric constraints in high dimensions from demonstrations." In: *Conference on Robot Learning*. PMLR. 2020, pp. 1211–1230.

[19] Adam Coates, Pieter Abbeel, and Andrew Y Ng. "Learning for control from multiple demonstrations." In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 144–151.

[20] Earl A Coddington and Norman Levinson. *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.

[21] Gergely Csibra and György Gergely. "'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans." In: *Acta psychologica* 124.1 (2007), pp. 60–78.

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

[23] Anca D Dragan and Siddhartha S Srinivasa. "A policy-blending formalism for shared control." In: *The International Journal of Robotics Research* 32.7 (2013), pp. 790–805.

[24] Anca Dragan and Siddhartha Srinivasa. "Generating Legible Motion." In: *Proceedings of Robotics: Science and Systems*. Berlin, Germany, 2013. DOI: 10.15607/RSS.2013.IX.024.

[25] Jaime F. Fisac, Anayo K. Akametalu, Melanie N. Zeilinger, Shahab Kaynama, Jeremy H. Gillula, and Claire J. Tomlin. "A General Safety Framework for Learning-Based Control in Uncertain Robotic Systems." In: *CoRR* abs/1705.01292 (2017). arXiv: 1705.01292. URL: http://arxiv.org/abs/1705.01292.

[26] Jaime F Fisac, Chang Liu, Jessica B Hamrick, Shankar Sastry, J Karl Hedrick, Thomas L Griffiths, and Anca D Dragan. "Generating plans that predict themselves." In: *Algorithmic Foundations of Robotics XII*. Springer, 2020, pp. 144–159.

[27] Michael J Gielniak, C Karen Liu, and Andrea L Thomaz. "Secondary action in robot motion." In: *19th International Symposium in Robot and Human Interactive Communication*. IEEE. 2010, pp. 310–315.

[28] Michael J Gielniak and Andrea L Thomaz. "Anticipation in robot motion." In: *RO-MAN*. IEEE. 2011, pp. 449–454.

[29] Michael J Gielniak and Andrea L Thomaz. "Enhancing interaction through exaggerated motion synthesis." In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 2012, pp. 375–382.

[30] Jeremy H. Gillula, Gabriel M. Hoffmann, Haomiao Huang, Michael P. Vitus, and Claire J. Tomlin. "Applications of hybrid reachability analysis to robotic aerial vehicles." In: *The International Journal of Robotics Research* 30.3 (2011), pp. 335–354. DOI: 10.1177/0278364910387173.

[31] Noah D Goodman, Joshua B Tenenbaum, and Tobias Gerstenberg. *Concepts in a probabilistic language of thought*. Tech. rep. Center for Brains, Minds and Machines (CBMM), 2014.

[32] Alison Gopnik and Henry M Wellman. "Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory." In: *Psychological bulletin* 138.6 (2012), p. 1085.

[33] Kenneth Gross. *Puppet: An essay on uncanny life*. University of Chicago Press, 2011.

[34] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. "Enabling robots to communicate their objectives." In: *Autonomous Robots* 43.2 (2019), pp. 309–326.

[35] Ashesh Jain, Shikhar Sharma, Thorsten Joachims, and A. Saxena. "Learning preferences for manipulation tasks from online coactive feedback." In: *The International Journal of Robotics Research* 34.10 (2015), pp. 1296–1313.

[36] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. "The naïve utility calculus: Computational principles underlying commonsense psychology." In: *Trends in cognitive sciences* 20.8 (2016), pp. 589–604.

[37] Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, S. S. Srinivasa, and J. Andrew Bagnell. "Shared autonomy via hindsight optimization for teleoperation and teaming." In: *The International Journal of Robotics Research* 37.7 (2018), pp. 717–742. DOI: 10.1177/0278364918776060.

[38] Daniel Kahneman and Amos Tversky. "Prospect theory: An analysis of decision under risk." In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127.

[39] Rudolf Emil Kalman. "When is a linear control system optimal?" In: *Journal of Basic Engineering* 86.1 (1964), pp. 51–60.

[40] Mohammad E Khodayar, Mohammad Shahidehpour, and Lei Wu. "Enhancing the dispatchability of variable wind generation by coordination with pumped-storage hydro units in stochastic power systems." In: *IEEE Transactions on Power Systems* 28.3 (2013), pp. 2808–2818.

[41] Minae Kwon, Sandy H Huang, and Anca D Dragan. "Expressing robot incapability." In: *Proceedings of the 2018ACM / IEEE International Conference on Human-Robot Interaction*. 2018, pp. 87–95.

[42] Changshuo Li and Dmitry Berenson. "Learning object orientation constraints and guiding constraints for narrow passages from one demonstration." In: *International symposium on experimental robotics*. Springer. 2016, pp. 197–210.

[43] C. Lichtenthäler and A. Kirsch. "Goal-predictability vs. trajectory-predictability: which legibility factor counts." In: *Proceedings of the 2014 acm/ieee international conference on human-robot interaction*. 2014, pp. 228–229.

[44] Henry Wadsworth Longfellow. *A Psalm of Life. What the Heart of the Young Man Said to the Psalmist*. 1891.

[45] Christopher G Lucas, Thomas L Griffiths, Fei Xu, Christine Fawcett, Alison Gopnik, Tamar Kushnir, Lori Markson, and Jane Hu. "The child as econometrician: A rational model of preference understanding in children." In: *PloS one* 9.3 (2014), e92160.

[46] Jim Mainprice, Emrah Akin Sisbot, Thierry Siméon, and Rachid Alami. "Planning safe and legible hand-over motions for human-robot interaction." In: *IARP / IEEE-RAS / EURON workshop on technical challenges for dependable robots in human environments*. 2010.

[47] Eric Mazumdar, Lillian J Ratliff, Tanner Fiez, and S Shankar Sastry. "Gradient-based inverse risk-sensitive reinforcement learning." In: *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE. 2017, pp. 5796–5801.

[48] Daniel McFadden. "Conditional logit analysis of qualitative choice behavior." In: *Frontiers of Econometrics*. Ed. by P. Zarembka. New York: Academic Press, 1973.

[49] D Livingston McPherson and S Shankar Sastry. "An Efficient Understandability Objective for Dynamic Optimal Control." In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2021).

[50] D Livingston McPherson, Kaylene C Stocking, and S Shankar Sastry. "Maximum Likelihood Constraint Inference from Stochastic Demonstrations." In: *2021 IEEE Conference on Control Technologies and Applications* (2021).

[51] D Livingston McPherson, Dexter RR Scobee, Joseph Menke, Allen Y Yang, and S Shankar Sastry. "Modeling supervisor safe sets for improving collaboration in human-robot teams." In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 861–868.

[52] Ian M Mitchell. "A toolbox of level set methods." In: *Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, http://www. cs. ubc. ca/~ mitchell/ToolboxLS/toolboxLS. pdf, Tech. Rep. TR-2004-09* (2004).

[53] Ian M Mitchell and J A Templeton. "A toolbox of Hamilton-Jacobi solvers for analysis of nondeterministic continuous and hybrid systems." In: *International workshop on hybrid systems: computation and control*. Springer. 2005, pp. 480–494.

[54] Andrew Y Ng, H Jin Kim, Michael I Jordan, Shankar Sastry, and Shiv Ballianda. "Autonomous helicopter flight via reinforcement learning." In: *NIPS*. Vol. 16. Citeseer. 2003.

[55] Donald A. Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[56] Claudia Pérez-D'Arpino and Julie A Shah. "C-learn: Learning geometric constraints from demonstrations for multi-step manipulation in shared autonomy." In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 4058–4065.

[57] Erwin Prassler, Dirk Bank, and Boris Kluge. "Key technologies in robot assistants: Motion coordination between a human and a mobile robot." In: *Transactions on Control, Automation and Systems Engineering* 4.1 (2002), pp. 56–61.

[58] S Joe Qin and Thomas A Badgwell. "A survey of industrial model predictive control technology." In: *Control engineering practice* 11.7 (2003), pp. 733–764.

[59] Sean Quinlan and Oussama Khatib. "Elastic bands: Connecting path planning and control." In: *[1993] Proceedings IEEE International Conference on Robotics and Automation*. IEEE. 1993, pp. 802–807.

[60] Lillian J Ratliff and Eric Mazumdar. "Inverse risk-sensitive reinforcement learning." In: *IEEE Transactions on Automatic Control* 65.3 (2019), pp. 1256–1263.

[61] Lillian J Ratliff and Eric Mazumdar. "Inverse risk-sensitive reinforcement learning." In: *IEEE Transactions on Automatic Control* 65.3 (2019), pp. 1256–1263.

[62] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. "Maximum margin planning." In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 729–736.

[63] Tyler Rispom et al. "Differentiation-state plasticity is a targetable resistance mechanism in basal-like breast cancer." In: *Nature communications* 9.1 (2018), pp. 1–17.

[64] Dexter RR Scobee and S Shankar Sastry. "Maximum likelihood constraint inference for inverse reinforcement learning." In: *arXiv preprint arXiv:1909.05477* (2019).

[65] Kevin A Smith and Edward Vul. "Sources of uncertainty in intuitive physics." In: *Topics in cognitive science* 5.1 (2013), pp. 185–199.

[66] Jonathan Spencer, Sanjiban Choudhury, Matthew Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Sidd Srinivasa. "Expert intervention learning." In: *Autonomous Robots* 46.1 (2022), pp. 99–113.

[67] Kaylene C Stocking, D Livingston McPherson, Robert P Matthew, and Claire J Tomlin. "Discretizing Dynamics for Maximum Likelihood Constraint Inference." In: *arXiv preprint arXiv:2109.04874* (2021).

[68] Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, Siddhartha S Srinivasa, Maya Cakmak, and Vincenzo Micelli. "Toward seamless human-robot handovers." In: *Journal of Human-Robot Interaction* 2.1 (2013), pp. 112–132.

[69] Niko Sünderhauf et al. "The limits and potentials of deep learning for robotics." In: *The International Journal of Robotics Research* 37.4-5 (2018), pp. 405–420.

[70] Daniel Szafir, Bilge Mutlu, and Terrence Fong. "Communication of intent in assistive free flyers." In: *Proceedings of the ACM/IEEE international conference on Human-Robot Interaction*. 2014, pp. 358–365.

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

[72] Yuval Tassa, Nicolas Mansard, and Emo Todorov. "Control-limited differential dynamic programming." In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, pp. 1168–1175.

[73] Emanuel Todorov and Weiwei Li. "A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems." In: *Proceedings of the 2005, American Control Conference, 2005.* IEEE. 2005, pp. 300–306.

[74] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. "Understanding and sharing intentions: The origins of cultural cognition." In: *Behavioral and brain sciences* 28.5 (2005), pp. 675–691.

[75] Komite Nasional Keselamatan Transportasi. "Aircraft Accident Investigation Report: PT Lion Mentari Airlines. B737-8 (MAX); registered PK-LQP, Tanjung Karawang, West Java, Republic of Indonesia, 29 October 2018." In: *Jakarta, Indoneisa: Komite Nasional Keselamatan Transportasi* (2018).

[76] Peter Trautman, Jeremy Ma, Richard M Murray, and Andreas Krause. "Robot navigation in dense human crowds: the case for cooperation." In: *2013 IEEE international conference on robotics and automation*. IEEE. 2013, pp. 2153–2160.

[77] Elly Rachel Truitt. *Medieval robots*. University of Pennsylvania Press, 2015.

[78] Marcell Vazquez-Chanlatte, Susmit Jha, Ashish Tiwari, Mark K Ho, and Sanjit Seshia. "Learning task specifications from demonstrations." In: *Advances in Neural Information Processing Systems*. 2018, pp. 5367–5377.

[79] Charles W Warren. "Global path planning using artificial potential fields." In: *1989 IEEE International Conference on Robotics and Automation*. IEEE Computer Society. 1989, pp. 316–317.

[80] Grady Williams, Paul Drews, Brian Goldfain, James M Rehg, and Evangelos A Theodorou. "Aggressive driving with model predictive path integral control." In: *2016 IEEE Interna-*

tional Conference on Robotics and Automation (ICRA). IEEE. 2016, pp. 1433–1440.

[81] Gaby Wood. "Edison's eve." In: *A magical history of the quest for mechanical life. 1st American ed. New York: AA Knopf* (2002).

[82] Min Zhao, Rahul Shome, Isaac Yochelson, Kostas Bekris, and Eileen Kowler. "An experimental study for identifying features of legible manipulator paths." In: *Experimental robotics*. Springer. 2016, pp. 639–653.

[83] Brian D Ziebart. "Modeling purposeful adaptive behavior with the principle of maximum causal entropy." PhD thesis. Carnegie Mellon University, 2010.

[84] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. "Maximum entropy inverse reinforcement learning." In: *AAAI*. Vol. 8. Chicago, IL, USA. 2008, pp. 1433–1438.

[85] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. "Human Behavior Modeling with Maximum Entropy Inverse Optimal Control." In: *AAAI Spring Symposium: Human Behavior Modeling*. Vol. 92. 2009.

COLOPHON

= The LaTeX template `classicthesis` laid out this document with the typographical stylings developed by André Miede and Ivo Pletikosić. They were inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*", and provided this template free to use under the GNU General Public License. Many thanks to them.

Final Version as of May 12, 2022 (classicthesis v4.6).