UNIVERSITY OF CALIFORNIA
Los Angeles

For All Intents, a Purpose:
A Causal Framework for Empirical Counterfactuals

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Andrew Francis Forney

2017

ABSTRACT OF THE DISSERTATION

For All Intents, a Purpose:
A Causal Framework for Empirical Counterfactuals

by

Andrew Francis Forney
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2017
Professor Judea Pearl, Chair

Unobserved confounders (UCs) are factors in a system that affect a treatment and its outcome, but whose states are unknown. When left uncontrolled, UCs present a major obstacle to inferring causal relations from statistical data, which can impede policy making and machine learning. Control of UCs has traditionally been accomplished by randomizing treatments, thus severing any causal influence of the UCs to the treatment assignment, and averaging their effects on the outcome in each randomized group. Although such interventional data can be used to appropriately inform population-level decisions, unit-level decisions are best informed by counterfactual quantities that provide information about the UCs relating to each unit. That said, arbitrary counterfactual computation can be performed in only certain scenarios, or in possession of a fully-specified causal model that requires knowledge of the distribution over UC states.

This work describes how additional information from a deciding agent can be utilized to empirically estimate certain counterfactuals, even in the presence of UCs and the absence of a fully-specified model of reality. The resulting technique yields strictly more information than standard randomization, and is specialized to personal decision-making. We first formalize this new strategy, called Intent-specific Decision-making (ISDM), in the context of the tools provided by causal inference. We then demonstrate its utility in online, reinforcement learning tasks with UCs, and support the efficacy of our technique in both human-subject

and simulation experiments. We demonstrate how ISDM accommodates a fusion of observational, experimental, and counterfactual data, which can be used to accelerate policy learning. Finally, we extend ISDM to the offline experimental design domain, detailing its application toward improving the established randomized clinical trial.

The dissertation of Andrew Francis Forney is approved.

Yingnian Wu

Michael Dyer

Adnan Darwiche

Judea Pearl, Committee Chair

University of California, Los Angeles

2017

To Chéla, my heart's intent

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It is no small endeavor to thank all of those who have positively impacted my work and education in pursuit of this thesis, since (over the past 6 years) I have had the pleasure of interacting with so many brilliant, kind, and helpful individuals. Though the list is long, and my memory poor, allow me to at least express my gratitude to the following:

To my advisor, Judea Pearl, thank you for your patient guidance, support, and invaluable insight. I always felt the glow of enlightenment after each of our meetings, and walked away with a memorable piece of advice, like to always "Make two epsilons into a delta." I thank you for opening my eyes unto the world of causality, and hope to spread its messages far and wide.

To my instructors and committee members, Adnan Darwiche, Michael Dyer, and Ying-nian Wu, I thank you not only for your feedback on this dissertation, and for serving on my committee, but also for giving me the quality education that has paved its path. I am likewise grateful to David Smallberg and Richard Korf for their tutelage and shared experience that shaped my approach to education. Apropos, I am lucky to have had such wonderful students throughout my twelve quarters as a TA; thank you all for your kind words, interest, and encouragement.

To my lab-mates, Karthika Mohan, Bryant Chen, and Ang Li, I am lucky to have shared the graduate experience with such supportive and brilliant colleagues. I am particularly indebted to Elias Bareinboim, whose sage advice and experience were invaluable throughout my graduate career; I look forward to our continued work together, and am genuinely grateful for your friendship and motivation. To Kaoru Mulvihill, I am extremely grateful for all of your support; I am unsure that I would have ever submitted a timesheet *on time* had it not been for your diligent intervention. There were so many tasks that you helped me with throughout the years that it felt like you were my administrative guardian angel.

To my friends, Matt Akiyama, Eric Debelak, Adam Diab, Robert Huizar, Greg Scott, Joe Shugt, and Brett Toyama, your friendship has has been an incredible source of support during the trials of graduate life. Although high school feels like it was ages ago, I'm thrilled

that we've all managed to stay close. To my best man, Jimmy Bresnahan, I doubt I would have survived the strain and stresses of these past six years were it not for our "deep" conversations, shared interests, and frequent contact; your friendship has meant more than just a bit to me, and I look forward to the many more inspiring chats to come.

To my parents, Mark and Bridget Forney, I can only begin to express my gratitude for your love and support through all of my scholarly endeavors and otherwise. Thank you for teaching me the value of science and curiosity from an early age, and enabling me to pursue my interests. I am likewise thankful to my many other relatives, in particular my grandparents Gene and Lee Forney and Gertrude Wirth.

Last, but certainly not least, I am grateful to my wonderful wife, Chéla Willey. Having your love and devotion throughout our graduate careers has meant an immeasurable amount to me, and I am happy that we have had the opportunity to attend UCLA together. I cannot wait to build our lives together, especially with a teammate as brilliant as you.

VITA

2012	B.S., Computer Science & Psychology (Maj.); Pure Mathematics (Min.). Loyola Marymount University.
2012–2017	Teaching Assistant, Department of Computer Science. University of California, Los Angeles.
2015	M.S., Computer Science. University of California, Los Angeles.
2015–2017	Department of Computer Science, Head TA. University of California, Los Angeles.
2016–2017	Research Assistant, Cognitive Systems Lab. University of California, Los Angeles.
2017–Present	Assistant Professor, Computer Science. Loyola Marymount University.

PUBLICATIONS

*E., Bareinboim, *A., Forney, and J. Pearl (2015). Bandits with Unobserved Confounders: A Causal Approach. UCLA Cognitive Systems Laboratory, Technical Report (R-460). *In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS).*

A., Forney, J., Pearl, and E., Bareinboim (2017). "Counterfactual Data-Fusion for Online Reinforcement Learners." UCLA Cognitive Systems Laboratory, Technical Report (R-471).

* These authors contributed equally.

In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of Proceedings of Machine Learning Research.

A., Forney, Willey., C., Bareinboim, E., and J., Pearl (in-prep.). "Regret as a Counterfactual Learning Mechanism in Human Decision-Making."

A., Forney, Bareinboim, E., and J., Pearl (in-prep.). "Counterfactual Randomization for Clinical Trials."

LIST OF ABBREVIATIONS

CDT	Causal Decision Theory
EDT	Evidential Decision Theory
ETT	Effect of the Treatment on the Treated
HI	Heterogeneous Intent
IEC	Intent Equivalence Class
ISDM	Intent-specific Decision-making
MAB	Multi-Armed Bandit
MABUC	Multi-Armed Bandit problem with Unobserved Confounders
RCT	Randomized Clinical Trial
RDT	Regret Decision Theory
SCM	Structural Causal Model
SDM	Structural Decision Model
TS	Thompson Sampling
UC	Unobserved Confounder

CHAPTER 1

Introduction

"My colleagues, they study artificial intelligence; me, I study natural stupidity."

—Amos Tversky

For better or for worse, humans remain the sole decision-makers in a variety of daily, personalized, and important situations: doctors make prescriptions of drugs and treatments, judges give rulings of bail and incarceration, and (perhaps with less gravity) shoppers decide whether to purchase an expensive item or not. If humans were perfectly rational decision-makers who could detail every contributing influence of their choices, then psychologists, data-scientists, and a host of other professions would have little trouble predicting human behavior; again, for better or for worse, this is not the case. The fact is that humans rely on a number of simplifying heuristics to help them make expedient decisions, but as the word implies, heuristics can be at times misleading and at others exploited. Cognitive scientists have studied such *cognitive biases* at length, namely, factors that influence human decision-makers but are unknown to the individual to be influential. Apropos, Kahneman and Tversky describe a two-system model of human decision-making: (System 1) the primitive, impulsive, and fast system that operates with almost no conscious effort, and (System 2) the effortful, rational, and slower system that most people identify with [Kah11].

Despite how we may wish that System 2 is always in control of our decisions, the influences of System 1 have been repeatedly documented. Even cognitive biases as seemingly innocuous as *order effects*[1] have been implicated in influencing important decisions like

[1]The recency and primacy effects are well documented cognitive biases categorized as "order effects" in which items in a serial presentation are remembered best, and are perceived most important, when they appear first (primacy) or last (recency) [Ebb13].

hiring practices, in which the order of presented candidates can matter more than their qualifications [HG97, FY75, Far73]. Additionally, recent investigations into recidivism have concluded that judges' sentencing decisions are not always based on objective metrics of likelihood that a criminal will re-offend; rather, judges may possess personal heuristics for sentencing that relate to the type of crime committed (e.g., violent vs. non-violent), how the defendant conducted themselves in the courtroom, or even desires to combat racial inequities [KLL17]. From the medical domain, physician diagnostics, triage, and treatment have been shown to be influenced by implicit biases, such as perceptions of patient race and socio-economic status, despite self-reported claims that these factors do not influence their judgments [Els99, GCP07]. These biases and heuristics, conscious or otherwise, are generally not recorded, and so their influences on outcomes of interest could be either helpful or detrimental and vary from actor to actor.

Consequently, the purpose of this dissertation is to provide a general-purpose approach for controlling for the effects of personal cognitive biases in autonomous agent decision-making (human or computerized) that is more precise than the traditional approach of randomization. In general, we wish to not only control for factors that are influential to both an individual's actions (e.g., judge's sentencing) and their outcomes (e.g., recidivism), but also exploit them to benefit personalized decision-making. Towards this goal, we will draw on tools from causal analysis, cognitive psychology, and machine learning. The remainder of this introduction will provide a roadmap for the more technical aspects of this endeavor, followed by an outline of the problems we address, and our approaches for addressing them. We begin by formulating the problem of cognitive biases as confounding factors, and discuss the modeling implications moving forward.

Cognitive Biases as Confounding

Scientific endeavor has fought a historically rich battle against the influence of *confounders*, or factors that mutually influence a treatment and its measured outcome. Because, at its core, the scientific method seeks to uncover relationships of cause and effect, confounders

can contaminate causal claims when left uncontrolled. For instance, in one of a number of legal trials debating the causal effect of smoking on lung cancer, the attorney for Philip Morris proposed that genetic factors may be to blame for the interaction between smoking and cancer. The attorney went on to claim that the lack of experimental evidence, in concert with other explanations like such a genetic predilection to both smoking and cancer, could create extenuating circumstances that leave cigarettes themselves to be only indirectly at fault [MDD06]. The implication of this argument, and danger of confounding in the general sense, is that in the presence of uncontrolled confounding (in this case, a genetic factor), there exist multiple explanations that can be used to interpret the same data; we are thus left with the question of what is to blame for the lung cancer in a smoker: their smoking, or the genetic disposition that caused them to both smoke and attain the cancer? To settle debates such as this, and in general, to determine true relationships of cause and effect, the identification and control of confounding factors has been a significant focus of the empirical sciences [Pea98].

In the present work, we focus on the challenges that arise due to *unobserved confounders* (UCs), namely, unmeasured variables that influence the treatment (or action) as well as the response (or reward) to that treatment. In the medical context, for example, variables such as age and sex qualify as observed confounders – they affect doctors' decisions to prescribe certain drugs as well as each patients' response to that treatment, but are also known to be causal influences and are recorded. That said, such factors are particularly subtle when left uncontrolled, especially when they are invisible to decision-makers and present the potential to introduce *confounding bias* [Pea00, Ch. 6]. To reference the findings mentioned earlier that order effects can influence hiring, if recency and primacy biases are left uncontrolled, then a company may hire a suboptimally qualified candidate simply because they happened to interview first or last.

Controlling for confounding bias is not a new problem, and was presented to Fisher in the context of agricultural experiments. Farmers needed to estimate the effect of soil fumigants on oat crop yields, but UCs such as the populations of eel worms and birds systematically affected the agricultural lots based on their locations, which were used by the farmers to base

their decisions. This specific bias led Fisher to the idea of assigning treatment (fumigants) to the agricultural lots at random, which controlled for the systematic bias introduced by the farmers. This development is a central component in the theory of experimental design (see [Fis51, Wai89, Pea00] for more detailed discussions).

Furthermore, advances in causal analysis have given detailed prescriptions for identifying causal effects in both observational and experimental settings containing both observed and unobserved confounders (a formal treatment of confounding will be detailed in Chapter 2) [Pea00, Ch. 3]. However, the vast majority of these advances (not least, the treatment of confounding) have taken place in the domain of offline data-analysis. In these settings, it is assumed that the data of interest has been collected from many participants prior to analysis; data sources scrutinized by these traditional causal analysis approaches include surveys (an example of observationally collected data) and randomized clinical trials (an example of experimentally collected data). The distinction between, and concrete examples of, observational and experimental data will be crystallized in Chapters 2 & 3.

From these offline datasets, causal analysis provides tools for estimating the effects of both population-level and individual-level interventions. Population-level decisions generally consult causal interventional calculus; for these policy decisions, practitioners are interested in determining the best treatment for the general population. For example, enactment of an after school program may improve the average grade point average of high school students, even though certain students in that population would better benefit from another treatment like a personal tutor. To address the need to make personalized decisions, in which "what is good for the goose may *not* be good for the gander," individuals can consult the calculus of structural counterfactuals. The benefit of reasoning counterfactually is that individuals can make the best decision under their *particular* circumstance, rather than consulting population-level data (which will only reveal the best decision *on average*). We will explore the semantics of structural counterfactuals in Chapter 2 and demonstrate their superiority for personalized decision-making throughout the rest of the work.

Structural counterfactuals are useful tools for reasoners who are in possession fully spec-

ified causal models[2] of their environment, but in many applications of interest, this requirement is neither available nor feasible. Importantly, in the absence of a fully specified model, confounding factors can prevent estimation of counterfactual quantities that, as earlier intimated, are important for personalized decision-making. Although previous work has identified several scenarios in which these counterfactual quantities can be estimated with minimal modeling assumptions, our framework provides a bias-free method for estimating them in *any* setting in which the agent is an active experimenter (the details of which we will explore throughout the work).

Before detailing the background of technical concepts to be used throughout this work, we will first describe the specific problems that it addresses and outline our approach to solving them.

Questions Addressed in this Work

1. Can we estimate arbitrary counterfactual decision-making quantities in the absence of a fully-specified model and in the presence of UCs?

2. In the presence of UCs, in which observational, experimental, and counterfactual data are considered heterogeneous and incompatible, can an autonomous agent employ these disparate datasets in pursuit of learning counterfactual quantities of interest?

3. Does a strategy exist that can control for the influences of cognitive biases in human decision-making? If so, can humans wield it to successfully improve their performance in a decision-making task with cognitive biases?

4. Can autonomous agents who are differently affected by unobserved confounders (i.e., suffer from different cognitive biases in the same environment) coordinate to form a more detailed model of their situation?

[2]Modelling assumptions are discussed in Chapter 2, but a fully-specified model implies that relationships between variables in the system, observed or otherwise, are known.

Summary of Chapters & Contributions

As we shall demonstrate, the answer to all of the above questions is "Yes." The following chapters will illustrate their answers in detail.

Chapter 2: Background, in which we describe the necessary technical tools from causal analysis that will be employed in the remainder of the work. Specifically, this chapter will introduce:

- Structural Causal Models (SCMs), their specification, and semantics.
- A formal definition of confounding and how to model it in a SCM.
- An overview of the calculus of interventions and structural counterfactuals, with examples.

Chapter 3: Intent-specific Decision-making, in which we introduce our approach for empirically estimating counterfactual quantities in scenarios where a reasoning agent does not possess the fully-specified model of reality, is subject to UCs, and must learn an optimal choice policy. The specific contributions of this chapter are:

- The notion of an agent's *intent*, which, briefly, is their intended action before execution. Intent serves as a proxy for the state of any unobserved confounders, which allows an agent to make unbiased, counterfactual, and "intent-specific" estimates of an action's outcome. Termed *intent-specific decision-making (ISDM)*, this approach is the first to estimate certain counterfactual quantities in the absence of a fully-specified model.
- The specification of an SCM used to model intents and learning in a decision-making task, called Structural Decision Models (SDMs). Using SDMs, we prove that ISDM represents an empirical means of estimating certain counterfactuals, which was previously not possible for the classes of models we discuss.
- Application of ISDM to new, more general, versions of classical reinforcement learning problems: the Multi-Armed Bandit Problem with Unobserved Confounders (MABUC).

We demonstrate that traditional approaches fail to converge to the optimal policy in MABUC scenarios, define a new ISDM-based optimization metric called Regret Decision Theory (RDT), and prove its superiority over the previous state-of-the-art MAB algorithms.

- A demonstration of RDT used to retrofit a traditional bandit-learning algorithm, Thompson Sampling. Simulation results support the efficacy of ISDM in MABUC problems.

Chapter 4: Human-subjects Intent-specific Decision-making, in which we determine if humans can use intent as a reasoning tool to control for cognitive bias in a MABUC reinforcement learning task. The results of this study assert that:

- Humans can indeed isolate the signal of their intent, and employ it to improve learning and performance on a confounded decision-making task.

- Intent is reactive to environmental influences, not experiential history; this suggests that intent does not respond differently to the same stimuli from learned experience.

- Although humans only rarely discover ISDM when left to their own devices, suggesting that it is not a natural or common reasoning tool, those that were instructed to use it saw significant improvements in performance compared to their peers who were not.

Chapter 5: Counterfactually Enabled Data-Fusion, in which we illustrate that observational and experimental datasets, although not the counterfactual estimation objectives of RDT, can accelerate learning of optimal policies in MABUC scenarios. Specifically, we demonstrate that:

- Observational and experimental data is heterogeneous in scenarios with unobserved confounders, but can be used to accelerate empirical learning of counterfactual quantities of interest.

- This data-fusion, which was previously thought only possible for binary treatments and certain classes of restricted models, can be generalized to scenarios with arbitrary action-choice dimension (i.e., non-binary action choices) when using ISDM.

- Simulation results in MABUC settings support the efficacy of the data-fusion approach, and demonstrate that agents with access to observational and experimental data will converge to an optimal policy more quickly than those without.

Chapter 6: Heterogeneous Intent-specific Decision-making, in which we generalize our assumption that all agents in the reasoning environment possess the same intent function. Specifically, we illustrate that:

- In certain confounded decision-making scenarios, a combination of different actors' intents can be more informative than any one individually.

- We can create a new type of randomized clinical trial that is more informative about confounders by grouping individual actors into intent equivalence classes. We then discuss this approach in applications to drug trials and general experimental design.

- Simulation results in MABUC settings involving agents with heterogeneous intents support this new approach.

Chapter 7: Discussion, in which we detail the implications of ISDM and its prospects for future directions of study in the disciplines of artificial intelligence, cognitive science, and experimental design.

Roadmap. In Figure 1.1, we give a brief pictorial roadmap of the dissertation by chapter, category of relevant application, and means of empirically supporting the theories within. We lay the groundwork for ISDM in Chapter 3 with motivating examples, theoretical underpinnings, the foundational algorithm for online, empirical, counterfactual learning, and support its efficacy with simulation results. Chapter 4 corroborates the validity of this approach in human subject samples, demonstrating its plausibility to work in real-world scenarios. Chapter 5 demonstrates that data collected in offline settings can be used to improve the

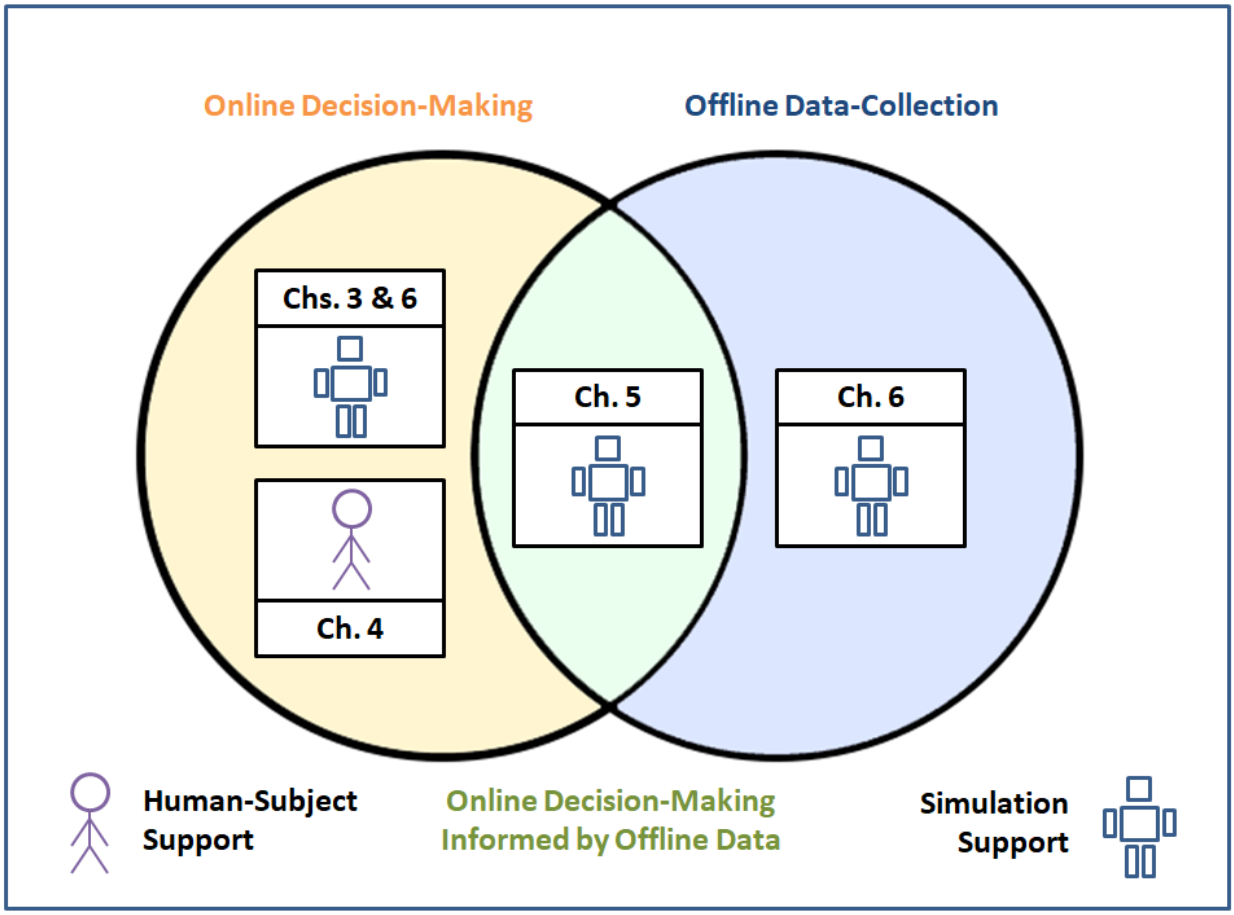

Figure 1.1: Roadmap of dissertation topics by chapter, category of relevance, and empirical support.

online counterfactual learning applied in Chapters 3 and 4, supporting its findings with simulations. Finally, Chapter 6 describes a means of applying ISDM to the offline experimental design domain and demonstrates how agents with different intent functions can actually benefit policy formation in a confounded decision-making scenario measured from both an online and offline perspective, once again supporting its theories with simulations. In briefest summary, this dissertation connects the primary avenues of scientific inquiry (online, active experimentation / learning and offline randomized clinical trials) with counterfactual reasoning under a traditionally difficult premise: that a decision and that decision's outcome are mutually affected by unknown factors.

CHAPTER 2

Background

In this chapter, we will review the technical tools from causal analysis that will be used throughout the remainder of the work. In particular, we will formalize: Structural Causal Models, d-separation, the back-door criterion, the front-door criterion, structural counterfactuals, the counterfactual back-door criterion, and a counterfactual quantity known as the Effect of the Treatment on the Treated.

2.1 Structural Causal Models (SCM)

We will employ the logic of Structural Causal Models to model our agents' decision-making, which will allow us to articulate the notions of observational, experimental, and counterfactual distributions as well as formalize the problem of confounding due to the influence of unobserved confounders (UCs).

Definition 2.1.1. (Structural Causal Model) [Pea00, pp. 204] A Structural Causal Model is a 4-tuple, $M = \langle U, V, F, P(u) \rangle$ where:

1. U is a set of *background* variables (also called exogenous), that are determined by factors outside the model.

2. V is a set $\{V_1, V_2, ..., V_n\}$ of *endogenous* variables that are determined by variables in the model, viz. variables in $U \cup V$.

3. F is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each f_i is a mapping from (the respective domains of) $u_i \cup PA_i$ to V_i where $U_i \subseteq U$ and $PA_i \subseteq V \setminus V_i$ and the entire set F forms a mapping from U to V. In other words, each f_i in $v_i = f_i(pa_i, u_i), i = 1, ..., n$ assigns

a value to V_i that depends on (the values of) a select set of variables.

4. $P(u)$ is a probability function defined over the domain of U.

The functional relationships between variables in a causal model can be graphically depicted in a causal diagram. Graphical models are useful for visualizing which variables influence one another in the system, and formalize modeling assumptions of independence that can be read directly from the structure of the graph (a procedure that we will detail shortly).

Definition 2.1.2. (Causal Diagram) Each SCM M is associated with a causal diagram G such that G encodes:

1. The set of endogenous variables V, represented as solid nodes (vertices).

2. The set of exogenous variables U, represented as hollow nodes.

3. The set of functional relationships F, represented as directed edges between nodes. Specifically, for each function $v_i = f_i(pa_i, u_i)$, a directed edge will point from variables on the function's right-hand side (i.e., from pa_i and u_i) to that on its left (i.e., to v_i). By convention, we will represent causal influences from endogenous variables as solid arrows, and influences from exogenous variables as dashed arrows.

Example 2.1.1. Suppose we wish to model a medical setting in which physicians treat a condition by administering one of two drugs X based on the patient's sex Z. The patient's chances of recovery Y depend on both the treatment and sex of the patient. These causal assumptions can be represented as the following system of structural equations in model M_1, with error terms ϵ_i that are generally assigned as parents to all endogenous variables to model background factors that have been omitted (but are conventionally not represented graphically due to their tendency to clutter the diagram):

$$\begin{aligned} Z &= f_Z(\epsilon_Z) \\ X &= f_X(z, \epsilon_X) \\ Y &= f_Y(x, z, \epsilon_Y) \end{aligned} \tag{2.1}$$

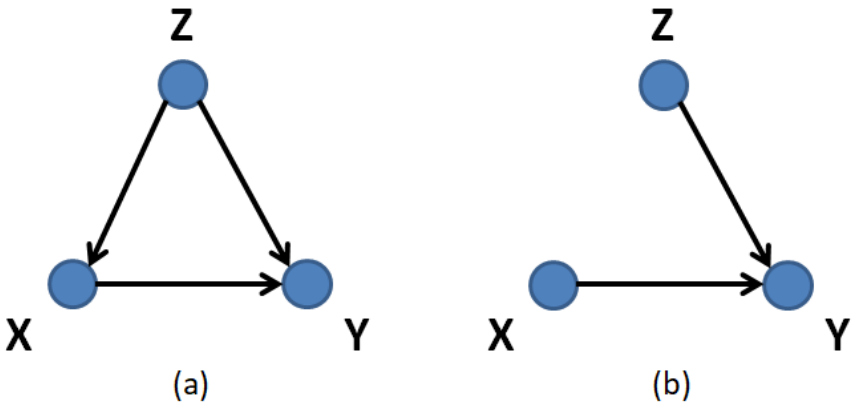

Figure 2.1: Causal diagrams of SCMs with error terms omitted for clarity: (a) Causal diagram $G(M_1)$ of observational study in Example 2.1.1 wherein X represents assigned treatment, Z patient sex, and Y recovery. (b) Causal diagram $G(M_{1x})$ of an experimental study in Example 2.1.1, in which the Natural influences on drug assignment $Z \to X$ are severed.

We can graphically depict the model expressed in Eqs. 2.1, M_1, as in Figure 2.1(a).

Causal models, and the graphical diagrams that depict them, encode the modeller's *causal assumptions* about the system under scrutiny; namely, if a variable X appears in the right-hand side of a equation for some other variable Y (e.g., in Eqs. 2.1, X and Z appear in the equation for Y, $y = f_Y(x, z, \epsilon_Y)$, then X is assumed to be a direct causal influence on Y, and is graphically referred to as its *parent*. It is important to distinguish the causal assumptions implicit in the functional model from the statistical data that is assumed to have been generated by that model, but collected empirically; as we shall soon exemplify, many causal queries cannot be answered from data alone, and require the aid of a modeling assumption to clarify. Note that structural causal models represent non-parametric versions of structural equation models (see [Pea00, Ch. 5]), in which the functions relating variables are known or estimated. Rather, SCMs admit that we rarely possess the true F and $P(u)$ (as in what is called *fully-specified model*), but can refer to the model to interpret data under certain causal assumptions (as in what is called a *partially-specified model* when we do not know the true F and $P(u)$).

The modeling assumptions made in each SCM will vary their classification and the conclusions that we can draw from each model class. For instance, the most basic classes of causal models, eliciting diagrams that are both acyclic (meaning no directed path in the graph visits a variable more than once) and contain no dependent error terms (to be discussed in relation to confounding, shortly), are called *Markovian*. In Markovian models, we see that for any given instantiation of exogenous variables $U = u$, there will be a unique instantiation of endogenous variables $V = v$. This implies that the joint distribution $P(v)$ is uniquely determined by the distribution of error terms $P(u)$, and is said to satisfy the Causal Markov Condition.

Theorem 2.1.1. (Causal Markov Condition) [PV91] Every Markovian causal model M induces a distribution over endogenous variables $P(v) = P(x_1, ..., x_n)$ that satisfies the parental Markov condition relative the causal diagram G associated with M; that is, each variable X_i is independent of all its non-descendants, given its parents PA_i in G, and allows for the *Markovian factorization* of the joint distribution:

$$P(x_1, ..., x_n) = \prod_i^n P(x_i|pa_i) \tag{2.2}$$

If the distribution $P(v)$ follows the Causal Markov Condition relative to $G(M)$, then P is said to be *Markov relative* to G. Put differently, in Markovian models, if a variable W does *not* appear on the right-hand side of the equation for Y, then it is assumed that Y is unaffected by any perturbations of W as long as the parents of Y (i.e., its causal influences) are held constant. Intuitively, the Causal Markov Condition stipulates that if we control for all immediate causes of Y, then Y should be unaffected by changes in any non-descendants in the system.

During model creation, investigators must determine which variables to model as parents of which others. Models that are faithful representations of the data that they are used to explain will follow two primary guidelines: (1) every variable that is a cause of two or more other variables is explicitly modeled, and (2) Reichenbach's [Rei56] common-cause assumption, stating that if two variables are dependent, then either one causes the other or there is another (sometimes unobserved) variable that causes both [Pea00, Ch. 1]. Broadly

speaking, when both of these guidelines are followed, the resulting model is Markovian (we will discuss exceptions shortly). In general, for any SCM, we have a procedure for determining if two sets of variables X and Y are independent given a third set $Z \subseteq V \backslash \{X, Y\}$. Moreover, these independence relationships can be read directly from the structure of the graphical model, $G(M)$, via the *d-separation* criterion.

Definition 2.1.3. (*d*-separation) [Pea00, Def. 1.2.3] A path p in $G(M)$ is said to be d-separated (or blocked) by a set of nodes Z if and only if:

1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or

2. p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is *not* in Z and such that no descendant of m is in Z.

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

The d-separation criterion also gives us a set of *testable implications* of the model such that if X and Y are d-separated by a set of variables Z in $G(M)$, then we should also find that $X \perp\!\!\!\perp Y|Z$ in P as well. Now that we have the semantics of SCMs specified, we will examine the various causal queries that they can be used to answer as well as several important mechanisms of data collection.

2.2 Causal Queries

Recalling Example 2.1.1, and assuming that P is Markov relative to G, there are a variety of *causal queries* that an investigator could employ M to answer. For instance, physicians may be interested in the *average* effect of one drug over the other across sexes. Traditionally, scientists answer causal queries by conducting studies that attempt to measure their exposure's effect on some outcome, which can vary in their data collection procedures; in this work, we will be contrasting two popular classes of study procedures, the models they entail,

and the queries that they can answer. The first type we will examine are *observational / non-experimental* studies.

Definition 2.2.1. (Observational Study) An observational study collects data over endogenous variables V (eliciting a distribution over observables $P(v)$) in some model M, and attempts to measure the effect of some exposure X on some outcome Y without any experimenter-exerted external intervention on the exposure assignment.

Suppose that Example 2.1.1 represents an observational study in which we are interested in evaluating the effect of a drug (X) on patient recovery (Y), but acknowledge that the effect can be different within sub-populations of the different sexes (Z). Likewise, we also acknowledge that doctors may modify their assigned treatment based on the patient's sex. In our model M_1 depicted in Figure 2.1(a), we have encoded this assumption with edges from $Z \to X$ and $Z \to Y$, indicating that Z is what is known as a *confounder*.

Definition 2.2.2. (Confounder) When evaluating the causal effect of some factor X on another factor Y in model M, a *confounding factor* (or *confounder*) Z is a common cause of X and Y such that in $G(M)$, $X \leftarrow Z \to Y$.

Controlling for confounders is integral when attempting to assess the causal effect of one factor X on another Y in observational studies.[1]. Intuitively, we are interested in the effect of X on Y within *homogeneous* populations, even if we are averaging the effect across multiple sub-populations. To see why this is important, consider that the data we have collected in Example 2.1.1 is observational, such as a survey of medical records in which each record contains the sex of the patient Z, the drug they were prescribed by their physician X, and whether or not they recovered from the condition being treated Y. Note that in observational studies such as this, the assignment of a drug to each participant is *not* randomized but instead is a function of some other selection criteria – in this case, the

[1]Simpson's Paradox is a demonstration that, from statistical data alone, the effect of some treatment X on outcome Y can be reversed by sequential conditioning on covariates; causal inference solves this problem by providing rules for which covariates should be conditioned upon in pursuit of causal queries (see [Pea14] [Pea00, Ch. 6])

patient's sex (and any unmodeled disturbance terms). Thus, if we measure the recovery rates of patients in the observational setting, ignoring that X is affected by Z (which also influences the outcome), we do not strictly measure the influence of X on Y (as intended), but contaminate this result with the physicians' selection mechanisms. To use an extreme example, consider that there is an even distribution of sexes in the observational sample (i.e., $P(z) = P(z') \; \forall \; z, z' \in Z$), but, by physician selection, drug A is given almost exclusively to men (i.e., $P(x|z) \neq P(x|z') \; forall \; z, z' \in Z$). If we wish to estimate the average effect of drug A across sexes, the sample will not properly scale the sex-specific effect of drug A on recovery by proportion of sexes[2].

Thus, the goal in measuring the effect of X on Y is to measure the exposure's influence on the outcome *independent from* any selection mechanism. This is precisely the procedure of our other study type of interest: an *experimental / interventional* study.

Definition 2.2.3. (Experimental Study) An experimental study collects data over endogenous variables V by randomly assigning treatments X (independent of the treatment's natural causes, i.e., f_X) and measures the outcome Y within each randomly assigned treatment condition. We call this random assignment an *intervention*, given that external forces (typically, the experimenter) have forced it to attain some value that it might normally not.

In experimental studies called Randomized Clinical Trials (RCTs), random assignment *fixes* a treatment to its intervened value for each participant, regardless of what treatment that individual would be assigned in the pre-intervention model. Interventions have an elegant interpretation for SCMs, both in terms of their effect on the functional model as well as the graphical representation.

Definition 2.2.4. (Intervention) An intervention represents an external force that fixes a variable to a constant value, and is denoted $do(X = x)$, meaning that X is fixed to the value x with $P(do(X = x)) = 1$. This amounts to replacing the equation for the intervened variable with its fixed constant such that, in the post-intervention model, $f_X = x$, and all

[2]A numerical example of this confounding bias will be presented in the following chapter.

mentions of X in other equations are likewise replaced with its fixed value, x. If the pre-intervention model is M, then we annotate the post-intervention model as M_x. Graphically, an intervention severs all inbound edges to X, indicating that all pre-interventional causal influences are no longer so.

Example 2.2.1. Consider now that we have performed an experimental study investigating the effects of the same two drugs X (from Example 2.1.1) on patient recovery Y, knowing that recovery is a function of both the drug choice and patient sex. However, instead of observing medical records containing physicians' drug assignment policies, we randomly assign our study's participants to be given either drug A or drug B (as by a coin flip). The causal assumptions implicit in M_1 (Figure 2.1(a)) are now slightly amended, such that, for each participant, the drug assigned is fixed by intervention $do(X = x)$. This gives us new causal assumptions as represented by the functions of the SCM M_{1x} below, and are echoed in the graphical representation, $G(M_{1x})$, in which the pre-interventional influences on X are severed (Figure 2.1(b)).

$$\begin{aligned} Z &= f_Z(\epsilon_Z) \\ X &= x \\ Y &= f_Y(x, z, \epsilon_Y) \end{aligned} \tag{2.3}$$

Let us now return to our causal query of interest, the effect of X on Y across genders. We now possess the vocabulary to express this query notationally: we are interested in measuring $P(Y|do(X))$. Given that our modeling assumption stipulates that recovery Y is a function of sex Z (and drug assignment), if we wished to assess the average causal effect of a drug across sexes in M_{1x}, we simply examine the effect of the drug in each assigned drug condition, given that the effects of sex on recovery will be distributed between conditions:

$$P(Y|do(X = x)) = P_{M_{1x}}(y) \tag{2.4}$$

Given that our causal query $P(Y|do(X))$ is measurable within the experimental study in Example 2.2.1, as represented by M_{1x}, we should address whether this same effect can be recovered from the observational study in Example 2.1.1. Recall that the distinguishing

assumption between the two studies is that in the observational case, the drug assignment mechanism is not randomized, but rather, is a function of sex – an influence that is indicated by the arrow from $Z \to X$ in M_1 that is severed by random assignment in M_{1x}. In Markovian models, causal queries like $P(Y|do(X))$ can always been identified, even from observational studies that have confounding factors.

[Pea95] determined that causal effects can be measured from non-experimental data if we are able to phrase all probabilistic expressions containing a *do* operator in terms of the observational distribution's parameters, $P(v)$. Put differently, a causal effect is identifiable if we are able to simulate an intervention in the scenario where no such intervention took place. The mechanisms for performing this algebraically are expressed in the rules of *do-calculus*, which use the graphical model $G(M)$, in concert with the rules of *d-separation*, to determine if a reduction from queries containing *do* operators to expressions that are *do-free* is possible. The rules of *do-calculus* are expressed as follows:

Theorem 2.2.1. (Rules of *do* Calculus) [Pea00, Theorem 3.4.1] Let G be the directed acyclic graph associated with a SCM M, let $G_{\bar{X}}$ be the subgraph in which all arrows into X are severed, let $G_{\underline{X}}$ be the subgraph in which all arrows emanating from X are severed, let $G_{Z(\bar{W})}$ be the subgraph of all arrows severed into Z nodes that are not ancestors of any W node in $G_{\bar{X}}$, and let P stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W, we have the following rules:

1. **Rule 1** (insertion / deletion of observations):

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\bar{X}}} \tag{2.5}$$

2. **Rule 2** (action / observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\bar{X}\underline{Z}}} \tag{2.6}$$

3. **Rule 3** (insertion / deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\bar{X}\overline{Z(W)}}} \tag{2.7}$$

Using these rules, we can show that our query of $P(Y|do(X))$ is indeed identifiable from the observational study, and is given by the expression:

$$P(y|do(x)) = \sum_z P(y|do(x), z)P(z|do(x)) \tag{2.8}$$

$$= \sum_z P(y|x, z)P(z) \tag{2.9}$$

Eq. 2.8 follows from conditioning on z, and the reduction of $P(y|do(x), z) = P(y|x, z)$ in Eq. 2.9 follows from Rule 2 of do-calculus, in which $(Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}}$. This derivation is summarized by what is known as the *back-door criterion*, which details the conditions under which confounders can be controlled in observational settings.

Definition 2.2.5. (Back-Door) [Pea00, Def. 3.3.1] A set of variables Z satisfies the *back-door criterion* relative to an ordered pair of variables (X, Y) in a directed, acyclic graph (DAG) G if:

1. No node in Z is a descendant of X; and
2. Z blocks every path between X and Y that contains an arrow into X.

When a *back-door admissible* set Z is available in G (meaning, a set of variables that satisfies Def. 2.2.5), then the causal effect $P(Y|do(X))$ can be computed using the *back-door adjustment formula*:

Theorem 2.2.2. (Back-Door Adjustment) [Pea00, Theorem 3.3.2] If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula

$$P(y|do(x)) = \sum_z P(y|x, z)P(z) \tag{2.10}$$

The back-door adjustment formula has an intuitive interpretation that removes the influences of a confounder's effect on treatment selection:

$$\text{[average effect of x on y]} = \sum_{z \in \text{sub-populations}} \text{[effect of x on y for z]} \times \text{[prevalence of z]}$$

Thus far, we have been considering only Markovian models wherein all common causes of variables have been explicitly defined as endogenous variables. If, however, error terms between two or more variables are found to be correlated, then it suggests that an unmodeled (latent or unobserved) common cause is to blame. In these *non-Markovian* systems, the influence of unobserved, exogenous variables can introduce *confounding bias*, which can compromise causal claims from observationally collected data. In the next section, we will look at how to formalize unobserved confounding influences and how to answer causal queries in their presence.

2.3 Challenges of Unobserved Confounding

In Example 2.1.1, we modeled sex Z as an *observed* confounder of drug assignment X and recovery Y. Adjusting for observed confounders is straightforward, as governed by the rules of *do*-calculus and the back-door criterion, just as we did in Eq. 2.8. However, in other scenarios, investigators may be unable to observe some confounding influences. This can be the case when the confounders are discovered after data collection (in which their values were not recorded during the study), when the confounders are known, but their values cannot be collected for each datum, or more insidiously, when we find that the errors of two or more variables are correlated (thus violating our "no correlation without causation" guideline for model creation). In either case, we would be naïve to simply assume that these common causes do not exist, lest we make our analysis susceptible to confounding bias.

Rather, we can explicitly model these latent or "unobserved" confounders by including an exogenous variable to represent the confounding.

Definition 2.3.1. (Unobserved Confounder) An *unobserved confounder (UC)*, C, is a common cause of two sets of endogenous variables X and Y (meaning $X \leftarrow C \rightarrow Y$), but C exists in the latent space of the model, such that $C \in U$. When there is at least one UC present in the system, and all UCs are modeled explicitly as exogenous variables, we call the model *Semi-Markovian*. Semi-Markovian models behave the same as Markovian models, except that UCs cannot be conditioned upon nor be subject to external intervention (as by

the *do*-operator).

The consequences of UCs for identifying causal effects from observational studies are immediate. If, in a Semi-Markovian model, we attempt to estimate $P(Y|do(X))$ but cannot find a back-door admissible set of variables that blocks all confounding paths between X and Y, then we cannot obtain an unbiased estimate for this query. In general, a question of whether or not a causal effect can be estimated from observational, or pre-interventional, data is a question of *identifiability*.

Definition 2.3.2. (Identifiability) [Pea00, p. 77] A causal query $Q(M)$ is identifiable, given a set of assumptions A (e.g., those in the diagram $G(M)$), if for any two (fully specified) models M_i and M_j that satisfy A, we have

$$P(M_i) = P(M_j) \Rightarrow Q(M_i) = Q(M_j) \tag{2.11}$$

In other words, if two models satisfy the same assumptions about the system, then the equality of the probability distributions in those models implies the equality of the query quantity, meaning that the query can be expressed in terms of the distributions alone.[3] Let us revisit our example from the introduction to demonstrate a causal effect that is *unidentifiable* from an observational study.

Example 2.3.1. Consider the case of a cigarette manufacturer arguing that their cigarettes are not to blame for incidences of lung cancer, but rather, that there is an unobserved confounder in the form of a genetic craving for nicotine that likewise increases susceptibility to lung cancer. They propose a Semi-Markovian model to interpret observational data in which each datum only records whether or not an individual smoked X and whether or not they attained lung cancer Y. Conspicuously absent is any record of whether or not the individual possesses the genetic appetite for nicotine, C, that also predisposes them to lung cancer. Suppose now that we consider C an unobserved confounder of X and Y, we can

[3] See [Pea12] for discussions on identifiability for a variety of causal queries and models

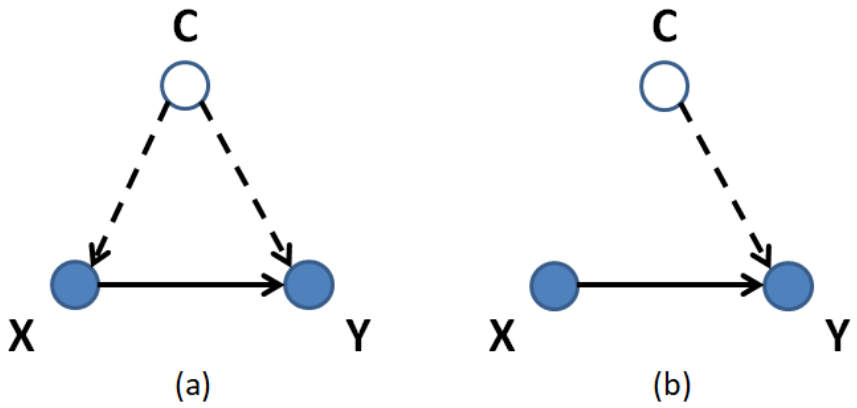

Figure 2.2: Causal diagrams of SCMs with error terms omitted for clarity: (a) Causal diagram $G(M_2)$ of observational study in Example 2.3.1 wherein X represents whether or not the patient smokes, Y whether or not they attained lung cancer, and C an unobserved genetic trait that causes a craving for nicotine and susceptibility to cancer. (b) Causal diagram $G(M_{2x})$ of an experimental version of 2.3.1, in which the Natural influences of the genetic craving $C \to X$ are severed.

model the system as follows (see also graph $G(M)$ in Figure 2.3(a)):

$$\begin{aligned} X &= f_X(c, \epsilon_X) \\ Y &= f_Y(x, c, \epsilon_Y) \end{aligned} \tag{2.12}$$

Now, suppose we wish to answer the causal query of whether smoking causes lung cancer, as we would formalize by the query $P(Y|do(X))$. To compute an unbiased estimate of this query in the observational setting, we would (by the back-door criterion, Def. 2.2.5) need to adjust for C, which is not possible because C is unobserved. As such, the confounding arc introduced by C cannot be controlled, and so the causal query $P(Y|do(X))$ is *unidentifiable* in M_2. Intuitively, this effect is not identifiable in the observational setting because any dependence of X on Y could instead be attributed to the UC. In litigation, the effect of smoking on lung cancer being unidentifiable from observational data could provide extenuating circumstances under which smoking is not at fault for incidences of lung cancer.

As mentioned earlier, to control for the influence of any UCs, the empirical sciences traditionally run an experimental study like an RCT. Ideally, this would leave us with a

post-interventional model like in Figure 2.3(b), whereby participants are randomly assigned to "smoke" vs. "do not smoke" conditions, and so the effect of smoking on lung cancer can be identified. Such a study would, of course, be ethically reprehensible, and is a prime example of why causal analysis has strived to provide formalisms for identifying causal effects in observational settings. Unfortunately, for many interesting and plausible models (like in Example 2.3.1), UCs can contaminate the causal claims of observational results. We will revisit a model similar to $G(M_2)$ in the following chapter, and demonstrate how UCs that may haunt offline data analysis can be used to the benefit of reasoners in the online domain.

Thus far we have been considering *population-level* causal effects, in which we are interested in the effect of some treatment averaged over sub-populations in which that effect may vary. In Examples 2.1.1 and 2.2.1, we discussed how the query $P(Y|do(X))$ encodes the average recovery rates Y of some drug X across sexes; after Example 2.3.1, we discussed how an unethical experimental study could measure the query $P(Y|do(X))$, which would average the effect of smoking on lung cancer across those with and those without the genetic UC. While these queries are interesting from a policy-analysis perspective (e.g., "Should we stock drug A or drug B for our pharmacy? Which will be the greatest good for the average patient?"), they may not be useful for personal decision-making in which each individual's characteristics specify the efficacy of a treatment. For instance, if we accept the existence of a genetic craving for nicotine that also causes a susceptibility to lung cancer, then it is possible that there exist individuals without the trait who would be less likely to get cancer if they started smoking. That said, if the trait was common, then both observational data (in which the trait is an uncontrolled UC between smoking and lung cancer) and hypothetical experimental data (in which the effect of smoking on cancer is averaged over those with the trait and those without) could over-exaggerate the risk for someone without it.

In summary, the problem with using population-level data to inform individual-level decisions is that the prior summarizes effects within heterogeneous populations (e.g., those with and without the genetic trait), only some of which apply to the latter (e.g., an individual without the trait). For individual decision-making, the best treatment or action (as measured by some response variable) is the one that is optimal for that individual's particular

characteristics and situation, $U = u$. However, in the offline domain, since an individual's situation is characterized by both endogenous and exogenous factors, the individual's behaviors serve as the only evidence for the state of the exogenous variables. Thus, the goal of attaining a desirable outcome (like not getting lung cancer) in a given situation (like not having the genetic disposition for smoking) depends not only on the action one ultimately takes (like smoking) but also any indicators of an individual's situation (like their desire to smoke).

An optimal individual-level decision can therefore be characterized by one in which, for the same situation $U = u$, no other decision would lead (or be more likely to lead) to a more desirable outcome. This sentiment is a known experience to most humans in the form of *regret*, in which we envision a hypothetical world where, had we chosen *differently* than we did, we would have attained a better outcome. An optimal decision, therefore, is one bereft of regret, which can be considered a *counterfactual* quantity. If measuring causal effects answers the population-level question of "What happens if I do x?" then counterfactuals answer individual-specific questions of "What *would have* happened had I done x' given that I did some other x?" Causal analysis provides another tool for examining structural counterfactuals in the context of SCMs, which we will review in the following section.

2.4 Structural Counterfactuals

Humans employ counterfactual reasoning effortlessly for a wide variety of applications, which include providing explanations of outcomes in the past, modifying policies for future action, and associating blame [Byr16]. Questions of "what if" or "if only" are laden with causal implications because they betray a cognitive model whereby, had some cause been different, a different outcome would have occurred. In general, counterfactuals are useful insofar as they compare the outcomes of different treatments undertaken in the *exact* same situation. However, counterfactuals have been traditionally difficult to quantify because we witness only a single outcome to one action under a precise situation $U = u$, and cannot simultaneously witness a parallel outcome to a different action in that same situation at the same time.

For instance, when selecting routes to a destination while driving, we may at time t need to decide between taking route A vs. route B. As soon as we decide to take route A, we are not physically able to return to time t to compare the driving time to our destination *had we taken* route B. We could of course return at time $t + \delta$ to the exact location we were at at time t and compare the time to our destination taking route B, but differences in driving conditions, traffic light states, road construction, and other perturbations in the situation between time t and $t+\delta$ would contribute differences that do not allow for an exact comparison of the outcomes from our choices at time t.

Yet, humans make these counterfactual comparisons regularly, regretting trips on the freeway that we believe could have been sped by taking side-streets. As such, these mental comparisons of hypothetical outcomes can be modelled by our conceptions of cause and effect that are captured in structural causal models. We can thus formally define a counterfactual in a SCM as:

Definition 2.4.1. (Counterfactual) [Pea00, pp. 204] In a SCM M, Let X and Y be two subsets of endogenous variables such that $\{X, Y\} \in V$. The counterfactual sentence "Y would be y (in situation $U = u$), had X been x" is interpreted as the equality with $Y_x(u) = y$, where $Y_x(u)$ encodes the solution for Y in a structural system where for every $V_i \in X$, the equation f_i is replaced with the constant x. Alternatively, we can write:

$$Y_x(u) = Y_{M_x}(u) \tag{2.13}$$

At first look, a counterfactual appears similar to our definition of an intervention, Def. 2.2.4. However, whereas the *do*-operator expresses an intervention across *all* possible situations $u \in U \ \forall \ u$, a counterfactual computes an intervention for a *particular* $U = u$. Just as we had causal queries for population-level interventions, so do we have counterfactual queries of the (probabilistic) format:

$$P(B_A \mid e) = P(\text{consequence}_{\ \text{antecedent}} \mid \text{evidence}) \tag{2.14}$$

Intuitively, counterfactual queries assess the probability of some outcome (or consequence) under the hypothetical intervention (or antecedent) given the evidence that was

witnessed in reality. Importantly, counterfactuals allow for the antecedent and evidence to contain contradictory claims. For instance, borrowing M_2 from Example 2.3.1, the counterfactual query of "What would be my chances of getting lung cancer had I smoked, given that I had not" can be posed as (for $X = 1$ representing "smoking" and $X = 0$ representing "not smoking") $P(Y_{X=1}|X = 0)$. Though it may appear odd to have evidence and hypothetical antecedent in contrast to one another, the interpretation of this counterfactual quantity is intuitive: we use the observed evidence to update our belief about the state of U at the time of observation and then assess the outcome under the model in which the antecedent is fixed to its hypothesized value. In other words, we adjust our belief about the pre-interventional state of the world from the observed evidence, and then compute the effect of the antecedent (fixed by intervention) on the outcome in this modified model of the world. Formally, this procedure can be accomplished in a fully-specified model using the following steps:

Theorem 2.4.1. (Counterfactual computation) [Pea00, Theorem 7.1.7] Given a fully-specified SCM $M = \langle U, V, F, P(u)\rangle$, the conditional probability $P(B_A|e)$ of a counterfactual sentence "If it were A, then B," given evidence e, can be evaluated using the following three steps.

1. **Abduction**: Update $P(u)$ by the evidence e to obtain $P(u|e)$.
2. **Action**: Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A.
3. **Prediction**: Use the modified model $M'_A = \langle U, V, \{F \setminus f_A\} \cup \{f_A = a\}, P(u|e)\rangle$ to compute the probability of B, the consequence of the counterfactual.

However, as we mentioned earlier, investigators are rarely in possession of fully-specified models. Yet, counterfactual quantities can still be useful for informed decision-making, and so this work will focus on their estimation in scenarios where we have only a partially-specified model. To this end, we will begin by reviewing a counterfactual quantity that has been studied in model-free domains and demonstrate how it applies to our purpose: *the effect of treatment on the treated (ETT).*

While average causal effects measure the influence of some treatment applied uniformly within a population (including sub-populations), the ETT considers the influence of a treatment counter to the one that they have already been assigned; it has traditionally been applied for policy makers interested in predicting, for example, the change in some outcome should they cease a program or treatment that is already in effect [Hec92]. Formally, we define the ETT as:

Definition 2.4.2. (Effect of the Treatment on the Treated (ETT)) [Pea00] The counterfactual ETT of some treatment $X = x$ on some outcome Y, written $ETT(X \to Y)$, is expressed as the difference between the outcome in the treated population and the outcome in that same population (i.e., for characteristics $U = u$) had they not been treated x'.

$$ETT(X \to Y) = P(Y_x|x) - P(Y_{x'}|x) \tag{2.15}$$

$$= \sum_u [P(Y|x,u) - P(Y|x',u)]P(u|x) \tag{2.16}$$

Consider that $Y = 1$ is a desired outcome like recovery, then intuitively, when $P(Y_x = 1|x) > P(Y_{x'} = 1|x)$, the treatment $X = x$ is more desirable than a lack of treatment x' – in other words, those treated were better off than had they not been. We will revisit the relevance of this facet of the ETT for personalized decision-making in the following chapter, but must first analyze its component quantities and methods of computation. Firstly, note that Eq. 2.16 expresses the difference in outcomes from each treatment x vs. x' within each homogeneous sub-population $U = u$ as indicated by their treated status, and weighted by likelihood of each unit belonging to u such that we scale by $P(u|x)$. While this is indeed the quantity sought after for the ETT, we again must note that in lieu of a fully-specified model, we would be unable to perform the calculation in Eq. 2.16. Rather, we can attempt to estimate the quantities on the right-hand side of Eq. 2.15.

To do so, consider the first term indicating the effect of the antecedent $X = x$ in the pre-intervention world in which $X = x$, denoted $P(Y_x|x)$. Note that here the evidence and the antecedent are in agreement – the intervention we are considering assessing was in reality the one administered to this population. As such, this term is reducible to an observational quantity $P(Y|x)$ by the *consistency axiom*:

Definition 2.4.3. (Consistency) [Pea00, Corollary 7.3.2] As a corollary to the definition of a null action, in which:

$$Y_{\emptyset}(u) \triangleq Y(u), \tag{2.17}$$

the consistency axiom states that for any set of variables Y and X in a causal model, we have:

$$X(u) = x \Rightarrow Y(u) = Y_x(u) \tag{2.18}$$

As such, $P(Y_x|x) = P(Y|x)$, a quantity that can be measured from the pre-interventional distribution, as from data one might gather from an observational study (Def. 2.2.1). To compute the remaining quantity, $P(Y_{x'}|x)$, we cannot use the consistency axiom as the antecedent and evidence disagree. However, consider the following for binary $X = x \in \{x_0, x_1\}$:

$$\begin{aligned} P(Y_{x_0}) &= P(Y_{x_0}|x_0)P(x_0) + P(Y_{x_0}|x_1)P(x_1) \\ &= P(Y|x_0)P(x_0) + P(Y_{x_0}|x_1)P(x_1) \\ P(Y_{x_0}|x_1) &= \frac{P(Y_{x_0}) - P(Y|x_0)P(x_0)}{P(x_1)} \end{aligned} \tag{2.19}$$

Note that in Eq. 2.19, $P(Y_{x_0}|x_1)$ is thus expressible in terms of observational quantities and interventional quantities, since $P(Y_{x_0}) = P(Y|do(x_0))$. This allows us to compute the counterfactual ETT for binary X when we are in possession of both observational and experimental data measuring the effect of X on Y. Furthermore, we are able to do so without a fully-specified model, and can instead measure the quantities of interest from the probability distributions alone. However, as soon as the treatment options become non-binary, (e.g., for $X = x \in \{x_0, x_1, x_2\}$) this data-driven approach fails in the offline domain. Without loss of generality, suppose we have a ternary X and wish to estimate the counterfactual $P(Y_{x_0}|x_1)$:

$$\begin{aligned} P(Y_{x_0}) &= P(Y_{x_0}|x_0)P(x_0) + P(Y_{x_0}|x_1)P(x_1) + P(Y_{x_0}|x_2)P(x_2) \\ &= P(Y|x_0)P(x_0) + P(Y_{x_0}|x_1)P(x_1) + P(Y_{x_0}|x_2)P(x_2) \\ P(Y_{x_0}|x_1) &= \frac{P(Y_{x_0}) - P(Y|x_0)P(x_0) - P(Y_{x_0}|x_2)P(x_2)}{P(x_1)} \end{aligned} \tag{2.20}$$

Note that to solve for the counterfactual quantity $P(Y_{x_0}|x_1)$, we would require observational data to estimate $P(Y|x_0)$, experimental data to estimate $P(Y_{x_0}) = P(Y|do(x_0))$, and the ability to compute a separate, unknown counterfactual quantity $P(Y_{x_0}|x_2)$. As such, because we have 1 equation and 2 unknowns ($P(Y_{x_0}|x_1)$ and $P(Y_{x_0}|x_2)$), our query is not identifiable in the absence of a fully-specified model, even when we are in possession of both observational and experimental data. This impediment to counterfactual estimation has been addressed for certain model formats (e.g., when we are licensed to use the *front-door adjustment*, see [Pea00, Def. 3.3.3]), but has thus far remained unsolved for arbitrary Semi-Markovian models. In the following chapter, we will demonstrate that the shortcomings of counterfactual estimation in the offline domain (wherein treatments are observed or fixed by intervention) are not inherited in the online domain (wherein extra information from an agent that may dynamically choose a treatment can aid counterfactual estimation).

Naturally, from the previous paragraph, we might wonder if there is at least some possibility to estimate counterfactuals from empirical data with a partially-specified, Semi-Markovian model. We can use a final tool from causal analysis to answer this in the affirmative, assuming that we possess a back-door admissible set of covariates Z which controls for any confounding influences that might otherwise bias the relationship between X and Y in a counterfactual query. This result is formalized in the following theorem:

Theorem 2.4.2. (Counterfactual Interpretation of Back-door) [PGJ16, Theorem 4.3.1] If a set Z of variables satisfies the back-door criterion relative to (X, Y), then, for all x, the counterfactual Y_x is conditionally independent of X given Z such that

$$P(Y_x|X, Z) = P(Y_x|Z) \tag{2.21}$$

Intuitively, Theorem 2.4.2 asserts that when we can control for all confounding influences between our treatment X and outcome Y, the pre-treatment observation of X tells us nothing more about the hypothetical Y_x that we did not already know from having observed Z. Of course, in some Semi-Markovian models like Figure 2.3(a), there exists no back-door admissible set Z (in the offline domain) to satisfy this theorem. In such a scenario, we would remain unable to compute counterfactual queries from empirical data in the absence of a

fully-specified model.

That said, the following chapter will make use of the counterfactual interpretation of the back-door by showing that, for online decision-making agents, there will always exist some set Z that can control for confounding influences (observed or otherwise) of treatments under the agent's control. We will begin this effort by providing the causal underpinnings of a decision-making task, formalize cognitive biases that may be present in such a task as unobserved confounders, and detail a novel approach for estimating counterfactual quantities empirically in a dynamic, online domain.

CHAPTER 3

Intent-Specific Decision-Making

As we detailed in the previous chapter, traditional approaches to counterfactual estimation take place in *offline* model and data analysis, wherein the investigator can identify counterfactual quantities of interest from:

1. Fully-specified Structural Causal Models (Def. 2.1.1, Theorem 2.4.1)
2. Partially-specified models where the treatment is binary and the investigator possesses observational and experimental data [Pea00, Ch. 9]
3. Partially-specified models wherein some set of covariates Z satisfy either the back-door or front-door criteria relative to counterfactual antecedent X and consequence Y [Pea00, Ch. 3]

However, given the rarity of possessing a fully-specified model in practice (which discards traditional approach #1), accounting for the prevalence of non-binary treatments of interest (which discards traditional approach #2), and, in the domain of decision-making, considering the pervasiveness of cognitive biases that introduce uncontrolled confounding bias (which discards traditional approach #3), there remain scenarios and counterfactual queries of interest that traditional, offline causal analysis cannot identify.

In this chapter we will detail a novel method for empirically estimating counterfactuals in arbitrary, partially-specified models where the treatment(s) are decisions that an agent can choose to enact online in real-time. However, in offline domains, data is generally collected

Chapter 3 is an extended version of [BFP15].

by the same mechanism within a population, like through an observational study (Def. 2.2.1, where the treatments are observed) or an experimental study (Def. 2.2.3, where the treatments are fixed, e.g. by random assignment). As we will see, the ability of individual decision-makers to fuse elements of observational and experimental strategies will allow them to estimate counterfactuals to create a more robust policy, even in the face of unobserved confounders.

We will begin this chapter with a motivating example that provides a concrete demonstration of the influence of unobserved confounders (UCs) in decision-making tasks, illustrates the difference between observational, experimental, and counterfactual quantities in such a scenario, and emphasizes the superiority of counterfactual data for individualized decision-making over the traditionally employed interventional data.

3.1 Motivating Example: The Greedy Casino

Example 3.1.1. Consider a scenario in which a greedy casino decides to demo two new models of slot machines, say $X = 0$ and $X = 1$ for simplicity, and wishes to make them as lucrative as possible. As such, they perform a battery of observational studies (using random sampling) to compare various traits of the casino's gamblers to their typical slot machine choices. From these studies, the casino learns that two factors well predict the gambling habits of players (unbeknownst to the players themselves): player inebriation and machine conspicuousness (say, whether or not the machines are blinking). Coding both of these traits as binary variables, we let $B \in \{0, 1\}$ denote whether or not the new machines are blinking, and $D \in \{0, 1\}$ denote whether or not the gambler is drunk. As it turns out, a gambler's *natural*[1] choice of machine, $X \in \{0, 1\}$, can be modelled by the structural equation indicating the index of their chosen machine:

$$X \leftarrow f_X(B, D) = (D \wedge \neg B) \vee (\neg D \wedge B) = D \oplus B \tag{3.1}$$

[1]By "natural" choice, we mean the choice that is reactive to environmental factors as a consequence of the individual's state, surroundings, and preferences; in cognitive science terms, these choices are those suggested by System 1, which is implicated in impulsive, non-deliberative, and heuristic decision-making [KK09], and in causal modeling terms, these are choices made in the unperturbed system without intervention.

Moreover, the casino learns that every gambler has an equal chance of being intoxicated and configure the machines to have an equal chance of blinking their lights at a given time, namely, $P(D = 0) = P(D = 1) = 0.5$ and $P(B = 0) = P(B = 1) = 0.5$.

The casino's executives decide to take advantage of their gamblers' propensities by programming the new slot machines to have reactive payouts that will tailor win rates to whether or not each believes (via sensor input, assumed to be perfect for this problem) a gambler is intoxicated, and predicated on whether or not its lights are presently blinking. The one catch: a new gambling law requires that casinos maintain a minimum attainable payout rate for slots of 30%. Cognizant of this new law, while still wanting to maximize profits by exploiting gamblers' natural arm choices, the casino executives modify their new slots with the payout rates depicted in Table 3.1a.

(a)	$D = 0$		$D = 1$	
$P(y\|X, D, B)$	$B = 0$	$B = 1$	$B = 0$	$B = 1$
$X = 0$	$^i0.10$	0.50	0.40	$^i0.20$
$X = 1$	0.50	$^i0.10$	$^i0.20$	0.40

(b)	$P(y\|X)$	$P(y\|do(X))$
$X = 0$	0.15	0.30
$X = 1$	0.15	0.30

Table 3.1: Greedy Casino: (a) Payout rates decided by reactive slot machines as a function of arm choice, sobriety, and machine conspicuousness. Players' natural arm choices under D, B are indicated by the superscript i, to indicate "intent" (to be formalized later). (b) Payout rates according to the observational, $P(Y = 1|X)$, and experimental $P(Y = 1|do(X))$, distributions, where $Y = 1$ represents winning (shown in the table), and 0 otherwise.

The state, blind to the casino's payout strategy, decides to perform a randomized study to verify whether the win rates meet the 30% payout requisite. Wary that the casino might try to inflate payout rates for the inspectors, the state recruits random players from the casino floor, pays them to play a random slot, and then observes the outcome. Their randomized experiment yields a favorable outcome for the casino, with payouts meeting precisely the 30% cutoff. The data from their investigation looks like Table 3.1b (third column), assuming binary payout $Y \in \{0, 1\}$, where 0 represents losing, and 1 winning.

As students of causal inference who are suspicious of the casino's ethical standards, we

decide to go to the casino's floor and observe the win rates of players based on their natural arm choices (through random sampling). We encounter a distribution close to Table 3.1b (second column), which shows that the casino is actually paying ordinary gamblers only 15% of the time. From our perspective, which is blind to the casino's reactive payout policies, the fact that $P(Y|X) \neq P(Y|do(X))$ implies that UCs are present in the system, though we do not know the states of these influencing factors ([Pea98]).

In summary, the casino is at the same time (1) exploiting the natural predilections of the gamblers' arm choices as a function of their intoxication and the machine's blinking behavior (based on Eq. 3.1), (2) paying, on average, less than the legally required (15% instead of 30%), and (3) fooling the state's inspectors since the randomized trial payout meets the 30% legal requirement. Unknown to the gamblers in this casino, their natural choice of slot machine represents a case of *inevitable regret* [Pea13] since, by examination of Table 3.1a, the payouts associated with every naturally chosen machine (indicated by asterisks) are inferior to the payout rates of the machine choice that is counter to it. We can express this as a comparison of counterfactual quantities equivalent to the ETT (Def. 2.4.2):

$$P(Y_x|x) < P(Y_{x'}|x) \; \forall \; x, x' \in X \tag{3.2}$$

Note that this comparison corresponds to a common, counterfactual human experience of *regret*: the choice that we did *not* make (x') appears to lead to a better outcome (a higher probability of payout) than the choice that we made (x). However, unlike the counterfactual comparison from Eq. 3.2 wherein one machine appears conditionally better than the other, the observational and experimental payouts from Table 3.1(b) give us no information about which machine is the optimal choice; in fact, these results seem to suggest that there is *no* optimal machine choice, and that a gambler will experience the same payouts regardless of their choice. Regret plays an important role in policy modification, since actions under a particular context that we regretted in the past should be chosen less often than their alternatives in the future. However, regret is only useful for decision-making insofar as it informs future decisions – in other words, regret is not simply a retrospective cognitive act, but also a predictive one. To formalize and then employ regret as a counterfactual,

personalized learning and reasoning tool, this chapter will accomplish the following:

1. Connects the existing machine learning literature to a new, more general, learning problem in which the assumption of unconfounded decision-making is removed.

2. Formalizes personal learning and decision-making scenarios in the language of structural causal models (a variant of which we will define as structural *decision* models).

3. Demonstrates how the quantification of retrospective regret is captured by traditional counterfactual computations from fully-specified structural causal models (and why this quantity is to be sought-after for decision-making).

4. Details a new strategy relating how target quantities of retrospective regret can be measured in absence of a fully-specified model, and then translated into active decision-making criteria that reduce the agent's regret.

5. Provides simulation results to support the efficacy of this new approach.

3.2 Regret as a Learning Problem

Throughout this work, we will treat regret not as a known quantity (as one could compute as a counterfactual from a fully-specified structural causal model), but rather, as one that must be learned. Indeed, many facets of intelligent decision-making are wrought from trial and error (including evidence for a neuronal basis, see [HC02]). The goal of learning which actions or treatments produce regret is to avoid those actions in the future, and so develop a choice policy that optimizes some desired quantity (like monetary payouts in the Greedy Casino example). Scenarios in which an agent must learn which action among a set of choices is optimal (in terms of some reward function) have long been studied in the Multi-Armed Bandit literature. We will thus consider a new variant of the MAB framework to model our system, but will first provide some context for the existing literature.

The Multi-Armed Bandit (MAB) setting represents a canonical sequential decision-making problem that spans many disciplines and applications from medicine to robotics, economics

to online advertising [Rob52, LR85, EMM06, Sco10, BC12]. In a traditional bandit instance, an agent is faced with $K \in \mathbb{N}, K \geq 2$ discrete action choices (often called "arms"), each with its own, independent, and initially unknown *reward*[2] distribution. The agent's task is to maximize cumulative rewards over a number of rounds T, which requires learning (over independent trials) about the underlying reward distributions associated with each arm. Though this goal may seem straightforward, its execution involves an "exploration vs. exploitation" challenge – agents must *explore* arms sufficiently to determine which has the best payout, but should be wary of exploring too much lest they delay *exploiting* the one they suspect is best; if the agent explores too little, however, it risks settling for a sub-optimal arm. It is usually understood that, if given enough time to sample arms and observe their payouts, the agent will eventually (i.e., asymptotically) converge to the optimal arm.

Example 3.2.1. Personalized online advertisement selection is a popularly cited application for MAB algorithms [LCL10]. Consider a simplified version of Google's advertising engine that, in an effort to maximize the number of times a user is interested in a shown advertisement, must select one of several ads (say, for home-improvement, clothing, and video games) to display to a website's visitor. Without any knowledge about which ads tend to convert the most "click-throughs" (i.e., times when a user clicks on an ad), the engine must first randomly experiment (i.e., *explore*) to determine which ads tend to be the most attractive to each type of user (based on any collected user information like demographics). Only after sampling the merit of each ad (as measured in click-throughs) to each user group can the system consistently display the most group-appropriate ad (i.e., *exploit*). The engine thus operates user by user (MAB trial by trial) selecting an ad to display by some policy that leverages exploration and exploitation.

The exploration-exploitation trade-off was studied extensively in the canonical setting (e.g., [AMS09]) and has been extended to accommodate a wide range of scenarios that appear in practice. For instance, the *contextual bandit* problem models the agent's reward as

[2]The term *reward* is used in the MAB literature to denote the outcome of the agent's chosen treatment or action. Rewards might be patient recovery (in medical settings), click-throughs (in online advertisement selection), or money (in examples like the Greedy Casino).

a function of some observed, environmental factors, and so the best arm choice in one context may not be the best in another [LZ08, DHK11, Sli14]. In the *adversarial bandit* problem, the agent must contend with an omnipotent adversary that may manipulate reward distributions to counter the agent's strategy [BK10, ACF95, BS12]. We refer readers to [Sze10, BC12] for a comprehensive overview.

The standard metric of success for bandit algorithms is a quantification of *regret*, representing the difference between the reward that the agent received using its choice policy and the reward that the agent would have received choosing optimally. In simulations, which generally test bandit algorithms across some finite $T = t$ number of trials, the traditional metric of comparison is by the *cumulative regret* experienced by each. In these traditional MAB formalizations, unconfoundedness is assumed, and so regret is defined over *experimental / interventional* quantities. We will thus define, and refer to, the traditional definition of MAB regret as "Experimental" Regret (or e-Regret for short), and will discuss this choice shortly:

Definition 3.2.1. (Experimental Regret (e-Regret)) For a MAB problem with time horizon T, action choice $X \in \{x_1, ..., x_k\}$ (where $K = |X| \in \mathbb{N}, K \geq 2$ represents the number of choices), and reward Y, the *optimal experimental action* x^* is considered the one that maximizes the interventional expected reward, defined as:

$$x^* = \underset{x \in X}{\operatorname{argmax}} P(y_x) \tag{3.3}$$

The *e-regret* experienced by an agent using choice policy π at trial $0 < t < T$ is defined as:

$$r_t = P(y_{x^*}) - y_{x_t^\pi} \tag{3.4}$$

The *cumulative e-regret* experienced by an agent across all T trials is thus:

$$R_T = \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} P(y_{x^*}) - y_{x_t^\pi} \tag{3.5}$$

In early trials, learning agents in MAB settings will typically experience high amounts of regret when little is known about each arm's reward distributions, which later attenuates as

learning occurs and $\pi \rightarrow \pi^*$ where π^* is the optimal policy. Thus, an agent that attempts to maximize its reward will equivalently attempt to minimize its regret.

Importantly, note that Def. 3.2.1 defines regret in terms of the interventional distributions $P(Y_x) = P(Y|do(x))$. The traditional MAB literature thus makes the assumption that the optimal arm is the one discovered by experimental methods, either assuming the inexistence of UCs between X and Y or, that if they do exist, their influence will be disrupted by randomized action exploration. In Example 3.2.1, the agent explores advertisement choices (i.e., its choice of actions) for each user group before continuously exploiting the one with the best click-through rate. As such, the agent in this example is minimizing the traditional definition of regret, as given in Def. 3.4. While this definition may be appropriate for some decision-making scenarios, the assumption of no-confounding between choice and reward is certainly less general than definitions that account for the potential presence of UCs. This begs the question of whether the standard definition of regret should be used in scenarios with confounded decision-making.

Consider again the Greedy Casino Example 3.1.1, in which the decision-makers are gamblers on the casino-floor. We know from the example's description that the unobservant gambler's slot-machine choice is confounded by the casino's manipulative payout policy, which is a function of the machines' conspicuousness and the gamblers' intoxication. These gamblers are simply choosing machines "by whim," as suggested by System 1 cognitive processes that make heuristic decisions (as described in Chapter 1; see [TK75]). However, if we, as observant gamblers, decide to take a more principled approach to maximizing our winnings (as by the dictates of the more methodical System 2 cognitive processes), we might consider applying MAB learning algorithms to find the optimal machine choice (if any). Acting as rational agents, we may be tempted to surmise that, since we cannot know the states of the system's UCs, the standard definition of regret should apply, and thus, we should experience no regret for any machine choice because $P(Y_{x_0}) = P(Y_{x_1})$ (see Table 3.1(b)). Yet, this stance ignores the difference between the observational and experimental payout rates ($P(Y|X) \neq P(Y|do(X))$, implicating UCs that must mutually affect gamblers' decisions and rewards; this inequality also implies that if we knew the state of the UCs at

the time of decision-making, we could potentially experience regret for some of our choices (see true payout distribution in Table 3.1(a)).

We will soon demonstrate that, in the presence of UCs, agents can indeed perform better than simply maximizing the experimental distribution $P(Y_x)$; however, to discuss improvements over experimental maximization, we require a new, more general definition of regret. As such, to distinguish confounded decision-making scenarios like the Greedy Casino from traditional MAB problems, we will refer to the former as a Multi-Armed Bandit problem with Unobserved Confounders (MABUC). The new definition of regret in MABUCs considers the regret an agent would experience had they known the state of the UCs at the time of their decision, which we define, and refer to, as "Unobserved Confounder" Regret (or u-Regret for short):

Definition 3.2.2. (Unobserved Confounder Regret (u-Regret)) For a MABUC problem with time horizon T, action choice $X \in \{x_1, ..., x_k\}$ (where $K = |X| \in \mathbb{N}, K \geq 2$ represents the number of choices), reward Y, and unobserved confounders U (where U is an unobserved common cause of X and Y), the *optimal action* $x^*(u)$ is considered the one that maximizes expected reward under confounder state $U = u$, defined as:

$$x^*(u) = \underset{x \in X}{\operatorname{argmax}} P(y_x|u) \tag{3.6}$$

The *u-regret* experienced by an agent using choice policy π at trial $0 < t < T$ is defined as:

$$r_t^u = P(y_{x^*(u_t)}|u_t) - y_{x_t^\pi} \tag{3.7}$$

The *cumulative u-regret* experienced by an agent across all T trials is thus:

$$R_T^u = \sum_{t=1}^{T} r_t^u = \sum_{t=1}^{T} P(y_{x^*(u_t)}|u_t) - y_{x_t^\pi} \tag{3.8}$$

Def. 3.2.2 is identical to Def. 3.2.1, except that the optimal action (and its associated reward) are predicated on the state of the UCs $U_t = u_t$ at each round, noting that we make no restriction that the state of the UCs need stay the same at every round (e.g., the Greedy Casino's machine lights may stop blinking between rounds, or the gamblers may sober up, etc.). Plainly, to compute this version of regret, we would require a fully-specified model in

which we know not only the UC states at each round, but also the rewards associated with each action under each $u_t \in U_t$. Still, for discussing theoretical algorithmic performance on various MABUC instances, it will prove useful to have a metric that compares the "true" regret of an agent's chosen action (as by full knowledge of the system) despite the fact that in realistic settings the agent will not possess the same knowledge.

Apropos, consider that we are not simply habitués of the Greedy Casino, but also machine learning researchers who decide to run a battery of experiments using standard bandit algorithms to test the new slot machines on the casino floor, including: ϵ-greedy [SB98], Thompson Sampling [Sco10], UCB1 [ACF02], and EXP3 [ACF03]. If we use the standard definition of regret (i.e., e-regret from Def. 3.2.1) to assess our algorithms' performances, then (since there is no optimal experimental arm choice) we would experience no e-regret. However, if the casino owners observed our experiments and instead used the more general MABUC definition of regret (i.e., u-regret from Def. 3.2.2), they would observe that none of our algorithms ever learn the optimal arm under each configuration of U. In general, agents that never learn an optimal policy experience *linear regret*, in which $\lim_{t\to\infty} r_t \neq 0$ and so $R_t = O(t)$. In Figure 3.1, we depict the linear u-regret experienced by traditional bandit algorithms in the Greedy Casino example by plotting the probability that the agent chooses optimally at trial t and the cumulative u-regret (R_t^u) experienced by trial t. We also plot the u-regret of a bandit player who decides which machine to play by a coin flip [Exp.] (akin to the investigator's experimental study) and that of the typical gambler playing by whim [Obs.] (i.e., the observational play strategy whereby gamblers playing by whim obey $f_x = D \oplus B$).

Figure 3.1 illustrates several important points about the traditional decision-making policies in our MABUC problem: (1) each experiences linear u-regret, (2) each performs no better than randomly choosing a machine to play (as compared to the [Exp.] graph), (3) gamblers playing by whim or heuristic [Obs.] (blue line) not only experience linear regret, but also *never* choose the optimal arm. Our goal throughout the rest of the chapter will thus be to devise a strategy that minimizes u-regret, and so maximizes reward in bandit learning scenarios where the decision and reward are confounded by unobserved factors. We begin

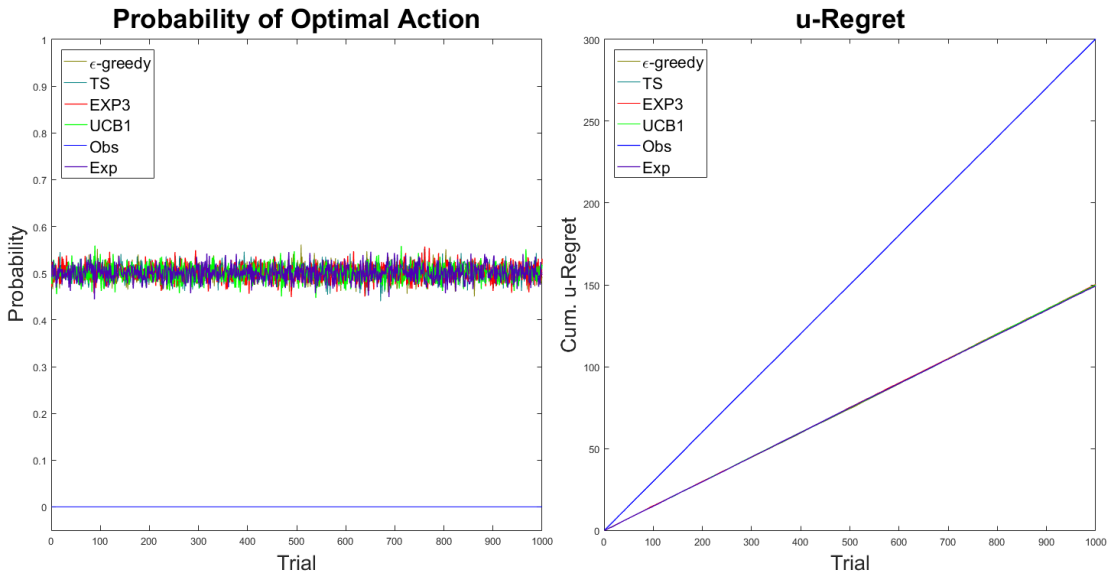

Figure 3.1: Plots of traditional MAB algorithm performance in Greedy Casino MABUC scenario. (Left) The probability that an algorithm chooses the optimal arm as a function of time. (Right) The cumulative u-regret experienced by each algorithm as a function of time.

this effort by formalizing bandit scenarios as causal inference problems, and use the tools developed in Chapter 2 (along with several new insights afforded by active learning agents) to create a novel MABUC algorithm.

3.3 Bandits as Causal Inference Problems

Towards developing a strategy for mitigating u-regret in scenarios with confounded decision-making, we must first formalize bandit problems in a language akin to that of Structural Causal Models (SCMs, Def. 2.1.1), which will allow us to model decision-making, confounding factors, and the means by which they distinguish observational, experimental, and counterfactual reward quantities.

We will begin by modeling two important facets of the Greedy Casino Example 3.1.1, which employs SCMs to relate the UCs B, D (whether or not the machines are blinking and whether or not the gambler is drunk, respectively), the gambler's machine choice $X \in x_0, x_1$,

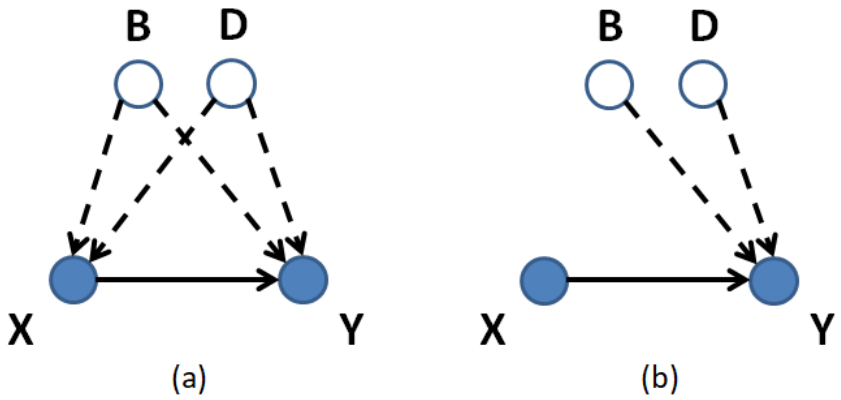

Figure 3.2: Graphical models of scenarios in the Greedy Casino Example 3.1.1. (a) Graph of observational model $G(M_3)$ wherein unobserved confounders B, D influence both the agents' decisions and their associated rewards. (b) Graph of the experimental / interventional model $G(M_{3x})$ wherein the decision-making influence of the UCs is severed by random assignment, though their influence on the reward Y remains.

and the received reward $Y \in 0, 1$:

1. *The observational setting*, M_3 in Figure 3.3(a), models the machine choices of gamblers playing "by whim" such that $f_x = D \oplus B$ (i.e., M_3 is the intervention-free model of the environment as described by Def. 2.2.1). The expected reward for agents in this model, as indicated in Table 3.1(b), can be computed:

$$\begin{aligned} P(Y = 1|x) &= \sum_{b,d} P(y_1|x, b, d)P(d, b|x) \\ &= \sum_{b,d} P(y_1|x, b, d)\frac{P(x|d, b)P(d)P(b)}{P(x)} \\ &= 0.15 \end{aligned} \tag{3.9}$$

2. *The experimental setting*, M_{3x} in Figure 3.3(b), models the machine choices of gamblers that were randomly assigned to play at a machine, as by intervention $do(X = x)$ (thus severing the influence of the UCs on each gamblers' decision) in the state investigator's experiment. The expected reward for agents in this model, as indicated in Table 3.1b,

can be computed:

$$\begin{aligned} P(Y=1|do(x)) &= \sum_{b,d} P(y_1|do(x),b,d)P(b,d|do(x)) \\ &= \sum_{b,d} P(y_1|x,b,d)P(b)P(d) \\ &= 0.30 \end{aligned} \tag{3.10}$$

Having detailed the observational and experimental models of the Greedy Casino example, note that we have yet to consider any counterfactual quantities implied by M_3. To see why counterfactuals may be instrumental to our decision-making in the Greedy Casino, consider what a gambler's natural machine choice tells us about the state of the UCs influencing their decision. In particular, with the fully-specified model, we know that $f_x = D \oplus B$, $P(x) = P(b) = P(d) = 0.5 \ \forall \ x, b, d$ and $P(d, b|x) = \frac{P(x|d,b)P(d)P(b)}{P(x)}$.

	$D=0$		$D=1$	
$P(D,B\|X)$	$B=0$	$B=1$	$B=0$	$B=1$
$X=0$	0.50	0.00	0.00	0.50
$X=1$	0.00	0.50	0.50	0.00

Table 3.2: Probability of each UC state $\{B=b, D=d\}$ given the observationally chosen arm $X=x$ in the Greedy Casino example as modeled by SCM M_3.

In words, if we observe a gambler choosing $X=0$, we know that either $\{B=0, D=0\}$ or $\{B=1, D=1\}$, and that all other configurations of B, D are impossible (and vice versa for the case when we observe $X=1$). Plainly, Table 3.2 illustrates that observing a gambler's observationally chosen arm provides more information about the state of the UCs than lacking such an observation (since $P(b,d) = 0.25 \neq P(b,d|x) \ \forall \ b, d, x$ in absence of observations about X). So, whether or not we know the states of the UCs, we *can* observe the choices that they have influenced. We are thus left with an important, counterfactual question: "Given that a gambler chose machine x', what would their chances of winning have been had they chosen some other x instead?" This query is precisely that captured by the Effect of the Treatment on the Treated (ETT) as introduced in Chapter 2 (Def. 2.4.2), in which we are interested in finding $P(Y_x|x')$.

To witness why this quantity is desirable, let us first assume that we have the fully-specified model, M_3, and know the reactive payout policy designed by the casino in Table 3.1. We can use the three-step counterfactual computation algorithm described in Theorem 2.4.1 to compute $P(Y_x|x')$.

Example 3.3.1. Let us, without loss of generality, demonstrate how to compute $P(Y_{x_0} = 1|x_1)$.

1. **Abduction**: Update $P(b, d)$ by evidence x_1 to obtain $P(b, d|x_1)$. Here, we would update $P(b, d|x_1)$ to produce the corresponding row of Table 3.2: $P(b_0, d_1|x_1) = P(b_1, d_0|x_1) = 0.5$ and $P(b_0, d_0|x_1) = P(b_1, d_1|x_1) = 0$.

2. **Action**: Modify M by the action $do(x_0)$ to obtain the submodel M_{3x}; this amounts to deleting the equation f_x from M_3, or removing the incumbent arrows to X in $G(M_3)$, and forcing f_x to the constant x_0. Let M'_{3x} be the intervened submodel having forced $f_x = x_0$ and updated $P_{M'_{3x}}(b, d) = P(b, d|x_1)$.

3. **Prediction**: With M'_{3x} in hand (in concert with the casino's payouts from Table 3.1), we can now compute our target $P(Y_{x_0} = 1|x_1)$:

$$\begin{aligned} P(Y_{x_0} = 1|x_1) &= P_{M'_{3x}}(Y = 1) \\ &= \sum_{b,d} P_{M'_{3x}}(Y = 1|b, d) P_{M'_{3x}}(b, d) \\ &= P_{M'_{3x}}(y_1|b_1, d_0) P_{M'_{3x}}(b_1, d_0) + P_{M'_{3x}}(y_1|b_0, d_1) P_{M'_{3x}}(b_0, d_1) \\ &= 0.5 * 0.5 + 0.4 * 0.5 \\ &= 0.45 \end{aligned}$$

Note that in the computation above, that $P_{M'_{3x}}(b_0, d_0) = P_{M'_{3x}}(b_1, d_1) = 0$ from our update in step 1, which explains why these terms are missing from the summation (they've been nullified). Repeating this process for all other combinations of $x, x' \in X$, we would obtain the following table:

$P(Y_x = 1\|x')$	$x' = 0$	$x' = 1$
$x = 0$	0.15	0.45
$x = 1$	0.45	0.15

Table 3.3: Results of computing the counterfactual $P(Y_x = 1|x')\ \forall\ x, x' \in X$ in the Greedy Casino example using the fully-specified model M_3 and its associated $P(x, y, b, d)$.

The interpretation of these counterfactual computations is clear: in observational circumstances when an agent in the system chose arm x', it was much better off (i.e., 3 times more likely to receive a reward) choosing the opposite arm x instead; an intelligent agent presented with this information would suffer the human experience of regret, knowing that the action alternative to the one chosen would have had a more likely desirable outcome. That said, as mentioned earlier in the text, the human notion of regret (as represented in Table 3.3) is useful for learning insofar as it allows for changes in future behavior that avoids the regretted actions under the same circumstance $U = u$ in which it was experienced. As such, if we reinterpret Table 3.3 as indication that the circumstances in which x' was chosen by one's *natural inclinations*[3] (i.e., $X = f_x(pa_x, u_x)$) were those in which $x \neq x'$ was a better choice, then those natural inclinations can provide information about the state of the environment that decided them (as evidenced by the information provided by the observed arm choice X about the state of U in Table 3.2). This presents the ETT as a promising quantity to consider optimizing in MABUC settings, given that in the present MABUC problem of the Greedy Casino, there is no optimal arm from either the experimental or observational perspective (Table 3.1b).

However, we must note that the computations in Example 3.3.1 are performed post-hoc, meaning that the observed evidence is from actions taken in the past and with the aid of a fully-specified model. In order for an active learning agent to be able to use the ETT as a tool in a MABUC scenario, it must surmount two challenges: (1) it must be able to compute the ETT in the absence of a fully-specified model (and thus, in the absence of knowledge

[3]Throughout the text, we will use terms like "natural," "observational," or "intended" choice to indicate a decision that was made under intervention-free circumstances (i.e., a choice X made by $X = f_x(pa_x, u_x)$), like in model M_3 of the Greedy Casino example.

about the state of any unobserved factors), and (2) it must be able to use the observational arm choice as *predictive* evidence for a choice that is about to made rather than as *reflective* evidence about choices that were made in the past.

To address these challenges, we will first define a new type of SCM that will be useful for modeling a learning task like that in a MABUC problem, and which bridges an important gap between SCMs used to model the offline domain. For our purposes, traditional models of the observational and experimental facets of MABUCs like the Greedy Casino (M_3 and M_{3x}, respectively) have a number of shortcomings:

1. **They are models of offline aggregates**. In other words, they represent offline data collected from many subjects and so do not explicitly model agents as individuals learning through active experimentation (in which experience gathered by trial t in some learning task will affect decisions in future trials $t + \delta$).

2. **They do not explicitly model changing decision policies or learning over time.** To illustrate, they do not give prescriptions for how an individual can at trial t_i act by whim (i.e., as a function of their observed and unobserved environments for choice $X = f_x(pa_x, u_x)$), at trial t_k act experimentally (e.g., by the outcome of a coin flip that disrupts their natural predilections), or at another trial t_m acknowledge their whim and choose to act differently (for $t_i \neq t_k \neq t_m$). In other words, active agents must be able to discern which of their actions (and their associated outcomes) were performed by whim, by experiment, or by a counterfactual impetus (to be formalized shortly), the distinction between which will be useful for learning.

As such, we define a SCM for an individual, active learning agent as a *Structural Decision Model*.

Definition 3.3.1. (Structural Decision Model (SDM)) A Structural Decision Model (SDM) models a sequential decision scenario for a learning agent over some T trials, and is a 2-tuple, $M^{\Pi} = \langle M, \Pi \rangle$ where:

1. M is a Structural Causal Model (Def. 2.1.1) of the un-intervened system (i.e., the observational model).

2. Π is a set $\{\Pi_1, \Pi_2, ...\}$ of *decision variables*, which are endogenous variables in M such that $\Pi \subseteq V$.

Definition 3.3.2. (Decision Variables) A *decision variable* $\Pi_i \in \Pi$ is an endogenous variable in a SDM with the following attributes:

1. The learning agent has direct, manipulable control over variables in Π (i.e., all variables in Π are amenable to external intervention).

2. All decision variables are functionally decided by a *policy* mapping any observed covariates, and the agent's decision history H_t (to be defined) up to the present trial t (where $0 < t < T$), to a decision: $\Pi_{i,t} = f_{\pi_i}(pa_{\Pi_i,t},\ u_{\Pi_i,t},\ h_t)$.

In graphical representations of SDMs, there will be only one key difference from SCMs: we will, by convention, model decision variables as closed squares, rather than circles, to distinguish them from covariates over which the agent has no control. For instance, in MABUC problems, arm selection X will qualify as a decision variable because the agent has manipulable control over it (and can, importantly, choose to randomly select an arm to mimic an experimental trial, a la $do(X = x)$). However, covariates like sex and age will not qualify as decision variables, because the agent cannot actively select their values. Decisions in learning environments are also functions of the agent's past experience with those decisions, which are modeled in the agent's decision history, defined next:

Definition 3.3.3. (Decision History) A *decision history* H_t is a sequence of all covariates, decisions, and outcomes of past trials (up to trial t) in a sequential decision task modeled by a SDM such that for covariates Z, decisions X, and outcomes Y, $H_t = \{Z_0, X_0, Y_0, ...Z_{t-1}, X_{t-1}, Y_{t-1}\}$. It is assumed that this data structure can distinguish decisions that were made observationally (i.e., $X_t = f_{x_t}(pa_{x,t}, u_{x,t})$) versus those made experimentally (i.e., via do-intervention like a coin flip, $X_t = do(x_t)$).

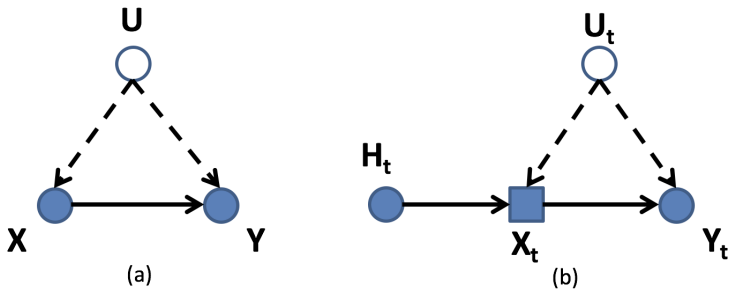

Figure 3.3: (a) Depiction of a confounded decision-making scenario for decision variable X as modeled by a SCM, like the Greedy Casino MABUC model M_3. (b) Depiction of the same scenario using a SDM for a learning MABUC agent.

Put succinctly by analogy, SCMs are to population modeling as SDMs are to individual, learning-over-time agent modeling – a distinction highlighted by a comparison between the SCM in Figure 3.3(a) versus the same system modeled from the perspective of a learning agent in the SDM in Figure 3.3(b). The distinction can be further expanded as SCMs capturing a system bereft of learning nor attempts to maximize some reward (i.e., *irrational* decision-making) and SDMs capturing a system with these objectives in place (i.e., *rational* decision-making); modeling the influence of an agent's experiential history plays a key role in this distinction.

Armed with the new tool of modeling an agent's learning process in a MABUC scenario using a SDM, we now rejoin our original goal: to find a means of mitigating u-regret experienced by agents in a MABUC scenario. In Example 3.3.1, we illustrated that computing the ETT from a fully specified model of the Greedy Casino scenario allowed us to retrospectively disambiguate[4] the "best" arm choice in a given trial using the action that the agent chose (observationally) in practice. As such, we identified this counterfactual quantity as one that, were it computable prospectively, could guide our agent to an optimal decision-making policy even when confounded by unobserved factors. In the following section, in concert with our

[4]Recall that from the observational ($P(Y|X)$) and experimental ($P(Y|do(X))$) reward distributions in the Greedy Casino (Table 3.1(b)), neither arm appears to ever be a superior choice to the other. In other words, e-regret is 0 for any arm choice, but the same cannot be said for u-regret.

definition of SDMs, we demonstrate that this is indeed possible, and take our first steps in attaining empirical estimations of counterfactual quantities.

3.4 Intent-specific Decision-making

To review the limitations faced by an agent in an arbitrary MABUC setting, we assume that the learner (a) has no access to the fully-specified model of the system, and as a corollary, (b) does not know the state of the unobserved confounders (UCs) at every trial. Yet, we wish to be able to estimate the ETT $P(Y_x|x')$ to gain evidence about the state of the UCs, and so choose an arm that is superior for a given trial's circumstance $U = u$. As such, if our agent can neither know nor measure U, we turn instead to any observable factors that might serve as proxies for its state.

A solution thus presents itself: an agent that is influenced by some UCs may not know the state of these unobserved factors, but it *will* know the arm choice that is *suggested* by the UCs. Consider an inebriated gambler ($D = 1$) approaching some blinking slot machines ($B = 1$) in the Greedy Casino; the individual may not know *why* they desire to play at the $X = 0$ model (i.e., they may not know that the blinking lights of $X = 0$ are palatable to inebriates), but they do know *that* they want to play at the $X = 0$ model. In this capacity, the agent's observationally *intended* arm choice (i.e., the one suggested by environmental factors via $X = f_x(pa_x, u_x)$) serves as evidence for the state of the environmental factors, observed or otherwise, that suggested that choice. Moreover, the information about the environment that is carried by the agent's intended, but not necessarily yet enacted, arm choice will be the same as that observed from an arm choice that was observationally enacted in the past, like those used to compute the ETT from a fully-specified model in Example 3.3.1.

We can now begin to formulate a new decision strategy for agents in MABUC settings that fuses the information about the environment curried by observational arm choices while being able to choose an arm that is *contrary* to the one suggested by the agent's predilections. In practical terms, an agent's intent serves as a proxy for the state of the environment, which

the agent can use as a context in which to better inform their decisions. Informally, what we will deem *intent-specific decision-making (ISDM)* considers decisions to be a two-step process: (1) in which the agent assesses its predilections in the current circumstance, and then (2) acts in a way that is a function of the outcomes of actions that were observed in the same circumstance in the past, using their intended action as an indication of circumstantial similarity. The steps of ISDM can be sketched as follows:

1. An agent about to make a decision observes its intended arm choice (i.e., its observational choice).

2. The agent then pauses, not necessarily enacting that intent.

3. The agent makes a final arm choice that is a function of its history and intended arm choice.

We will soon demonstrate the benefits of ISDM, but first must formalize some of its aspects, starting with the notion of an agent's intent.

Definition 3.4.1. (Intent) For all decision variables (Def. 3.3.2) $\Pi_i \in \mathbf{\Pi}$ in a SDM (Def. 3.3.1) $M^{\mathbf{\Pi}}$, let the agent's intended arm choice $I_{\Pi_i,t} = i_{\Pi_i,t} \in \Pi_i$ be the choice that the agent would make observationally at unit/trial t for the present unit's/trial's configuration of UCs $U_t = u_t$. Formally, let $I_{\Pi_i,t} = f_{\Pi_i}(pa_{\Pi_i,t},\ u_{\Pi_i,t})$.

Definition 3.4.2. (Intent-specific Decision-making (ISDM)) For any decision variable $\Pi_i \in \mathbf{\Pi}$ in a SDM, an agent whose decision policy f_{Π_i} is a function of I_{Π_i} is said to be practicing *intent-specific decision-making*, because any decisions made will be considered (as well as their outcomes recorded) within the strata of a particular intent condition.[5] Formally, for each decision $\Pi_i \in \mathbf{\Pi}$, $\Pi_i = f_{\Pi_i}(pa_{\Pi_i}, i_{\Pi_i}, h_t)$[6]

[5]This also means that, as an observed context, intent is recorded in the agent's SDM history, $I_{\Pi_i} \subseteq Z_t \in H_t$.

[6]Note that choices made in observational settings, as captured by SCMs without interventions (Def. 2.2.1), are a special case of ISDM such that the agent's intent is always followed, viz., $\Pi_{i,t} = I_{\Pi_i,t} = i_{\Pi_i,t}$.

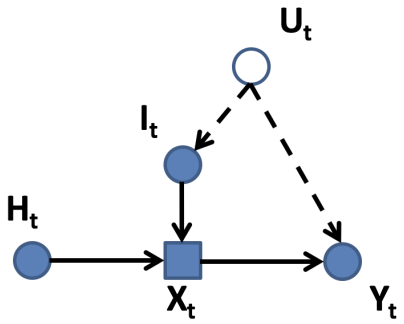

Figure 3.4: Depiction of intent-specific decision-making (ISDM) in the prototypical MABUC SDM with History H_t, decision variable X_t, intent I_t, unobserved confounder U_t, and outcome Y_t.

We can depict the SDM of an agent practicing ISDM in the Greedy Casino MABUC in Figure 3.4. Importantly, we note that what was once an unblocked back-door path via U between the decision variable X and the reward Y is now blocked by the rules of d-separation (Def. 2.1.3) given that the agent's intent I is assumed to always be observed for each decision node in the graph.[7]

3.4.1 ISDM for Reinforcement Learning

We now return to the reinforcement learning setting of the Multi-Armed Bandit Problem with Unobserved Confounders (MABUC), in which our objective was to find a means of achieving sub-linear u-regret (Def. 3.2.2). We will soon demonstrate that ISDM provides such a means, but will first place it in the context of the decision theory literature, and demonstrate how it provides a bridge between existing, and competing, theories.

Decision theory makes distinctions between two categories of decision-making tasks [ELH15]:

1. The *dualistic* category assumes that the deciding agent and the environment in which it is making decisions as separate entities. The agent only affects the environment

[7]This is not a particularly strong assumption, given that, by virtue of each decision node in a SDM being labeled as such, it is assumed that the agent already possesses some (possibly confounded) choice policy that dictates its action for that particular decision, which can thus be used as its intent.

through its actions, but the environment provides no information about the agent. A typical example of a dualistic model is for an agent playing chess.

2. The *physicalist* category assumes that the agent is embedded in the environment that it is also affecting with its actions, and so the environment may have a hidden state that mutually affects the agent. A typical example of a physicalist model is in robotic exploration.

Plainly, the MABUC setting, on which the present work focuses, belongs to the physicalist perspective, given that agents are assumed to be affected by UCs in the environment. Through this perspective, agents still attempt to maximize reward (or, in the decision theory vocabulary, agents attempt to maximize utility), but the target quantity that a policy should optimize is an ongoing debate. The two established camps in this debate fall within Evidential Decision Theory and Causal Decision Theory, defined and reviewed next:[8]

Definition 3.4.3. (Evidential Decision Theory (EDT)) [Bri17, Ahm14] Evidential Decision Theory states that an agent's action X may both influence and be evidence of the state of its environment, U. EDT agents thus maximize the reward Y from an *observational* (Def. 2.2.1) perspective, such that the optimal action $x^* \in X$ is defined as:

$$x^* = \underset{x \in X}{\operatorname{argmax}} P(Y|x) \tag{3.11}$$

Of note, evidential decision theory is typically considered within scenarios in which the agent possesses knowledge of the observational reward distribution prior to decision-making. The optimal arm choice under this context is thus the one that maximizes the observational distribution apart from any sort of empirical sampling technique to derive that distribution. As such, though EDT is well defined in these traditional contexts (wherein the agent begins with some amount of omniscience pertaining to the scenario at hand), its analogy to MABUC

[8]EDT and CDT are traditionally discussed under metrics of utility and expected utility, which we simplify herein to be directly comparable to the MABUC problem (with Bernoulli reward) without loss of generality.

problems is not clear since we define observational choices as those that are congruent with intent, and therefore not amenable to experiential maximization. For this reason, we cannot compare its performance as an optimization criteria in the MABUC, though we may discuss observational rewards as those amounting from environmental influences. By contrast, causal decision theory provides an optimization metric that may involve an individual agent's experiential history, though one that ignores the effects of any UCs.

Definition 3.4.4. (Causal Decision Theory (CDT)) [Wei16, SS90] Causal Decision Theory states that an agent's action X is a rational choice that is not indicative of the state of the environment; choices are made as by intervention $do(X = x)$. CDT agents thus maximize the reward Y from an *experimental* (Def. 2.2.3) perspective, such that the optimal action $x^* \in X$ is defined as:

$$x^* = \underset{x \in X}{\operatorname{argmax}} P(Y|do(x)) \tag{3.12}$$

EDT and CDT boast a rich history of debate in which proponents advocate their strategy across a variety of philosophical problems like Newcomb's Paradox [Noz69]. However, as we have earlier noted, neither EDT nor CDT provide sufficient maximization criteria for MABUC instances, in which the goal is to minimize u-regret; in the Greedy Casino Example 3.1.1, recall that neither the observational (as would be maximized by EDT) nor experimental (as would be maximized by CDT) reward distributions tout an optimal arm choice (Table 3.1(b)), nor does either distribution minimize the u-regret (Figure 3.1). However, a *fusion* of these two theories can yield a means of approaching MABUC problems: treat the agent's intent (i.e., the X from EDT) as a context in which to then interventionally (i.e., the $do(X)$ from CDT) make a final decision. This is precisely the quantity encoded in the counterfactual ETT (Def. 2.4.2) and a new target optimization quantity that we call Regret[9] Decision Theory, defined below.

[9]Herein, the term "Regret" is intended to highlight both the counterfactual nature of this decision theory and bind it to the vocabulary of reinforcement learning problems.

Definition 3.4.5. (Regret Decision Theory (RDT))[10] [BFP15] Regret Decision Theory states that an agent's intended action $I \in X$ serves as evidential context for the state of its environment, in which it may then interventionally act. RDT agents thus maximize the reward Y from a *counterfactual* (Def. 2.4.1) perspective, such that the optimal action $x^* \in X$, conditioned on the intended action x', is defined as:

$$x^* = \underset{x \in X}{\operatorname{argmax}} P(Y_x|x') \tag{3.13}$$

Note that Eq. 3.13 simply makes the ETT (Def. 2.4.2) the new maximization target for MABUC agents, as we intimated that it should be from Example 3.3.1. All that remains is to find a means of estimating the ETT *empirically* (i.e., without needing to compute a counterfactual quantity requiring a fully-specified model), which we will demonstrate can be done using ISDM. The key contribution of this dissertation exploits the equivalence between an agent's observationally / naturally chosen action (Def. 2.2.1) and its intent (Def. 3.4.1), and proves that intent facilitates an empirical measurement of the ETT.

Theorem 3.4.1 (Empirical Counterfactual Estimation)**.** [FPB17] Let X be a decision variable (Def. 3.3.2) in a SDM (Def. 3.3.1) with measured outcome Y, and let I be the agent's intent (Def. 3.4.1) for X. A counterfactual quantity $P(Y_x|x')$ for evidence x' and antecedent x (where $x, x' \in X$ and x need not be equivalent to x') can be estimated empirically using ISDM (Def. 3.4.2). Formally, we may write the counterfactual query in interventional notation such that

$$P(Y_x|x') = P(Y|do(X = x), I = x') \tag{3.14}$$

Proof. See appendix for proof of Theorem 3.4.1. □

Because the RDT is equivalently an interventional quantity using ISDM, we have also shown that the ETT, a counterfactual expression of the same format, can be estimated

[10] We have interchangeably referred to RDT as the Regret Decision Criteria (RDC) in other work.

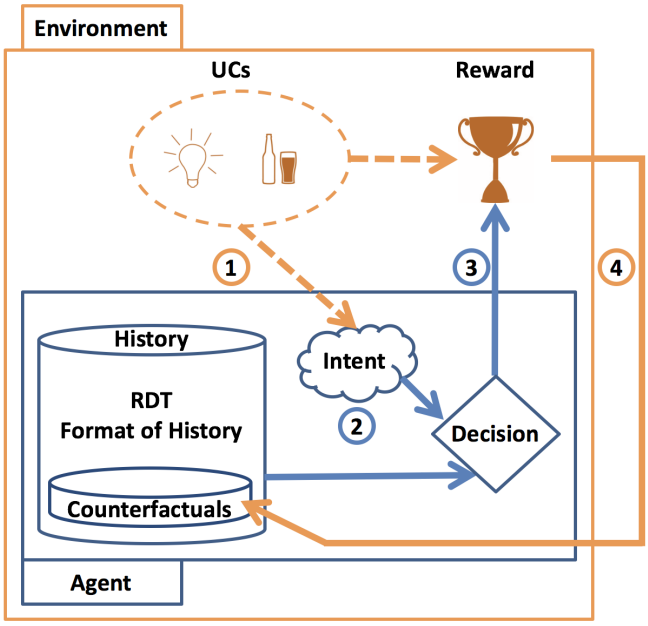

Figure 3.5: Depiction of the intent-specific decision-making (ISDM) process during a single trial of a MABUC sequential decision learning task.

empirically. Thus, we now possess a target maximization quantity with the means of reducing u-regret in a MABUC problem.[11]

To visualize the process of an ISDM agent operating in a MABUC scenario, consider the diagram in Figure 3.5, the steps of which are detailed below.

1. Unobserved confounders are realized in the environment, though their states are unknown to the agent.

2. From these UCs and any other observed features in the environment, the agent's heuristics suggest an action to take, i.e., its intent.

3. Based on its intent and history of context-action-reward triplets (in which intent is considered a member of context), the agent commits to a final action choice, "pulling" a selected arm.

[11]It is understood that the ETT can be computed for binary decisions or when the backdoor criterion holds [Pea00, Ch. 8], but it was not believed to be estimable for arbitrary decision-models nor dimensions prior to the development of RDT.

	$I = x_1$	$I = x_2$	...	$I = x_K$
$X = x_1$	$P(Y_{x_1} \mid x_1)$	$P(Y_{x_1} \mid x_2)$	...	$P(Y_{x_1} \mid x_k)$
$X = x_2$	$P(Y_{x_2} \mid x_1)$	$P(Y_{x_2} \mid x_2)$	...	$P(Y_{x_2} \mid x_k)$
⋮	⋮	⋮	⋱	...
$X = x_K$	$P(Y_{x_k} \mid x_1)$	$P(Y_{x_k} \mid x_2)$	⋮	$P(Y_{x_k} \mid x_k)$

Figure 3.6: An ISDM agent's counterfactual history in which rewards are recorded by intent-context I (columns) and final-arm-choice X for an arbitrary K-armed MABUC instance.

4. The action's response in the environment (i.e., its reward) is observed, and the collected data point is added to the agent's counterfactual history (as a consequence of Theorem 3.4.1).

Specifically, an ISDM agent's counterfactual history will be recorded in a tabular data structure akin to Figure 3.6.

Before we provide empirical support for ISDM in both simulations (Sec. 3.5) and human-subject trials (Ch. 4), we will prove some of its theoretical guarantees in the following section.

3.4.2 ISDM Theoretical Guarantees

In the previous section, we intimated that intent (Def. 3.4.1) serves as an observable proxy for the state of any unobserved confounders between an agent's action and its associated reward, thus giving the agent a means of reducing u-regret (Def. 3.7) in a MABUC problem. However, the extent to which intent-specific decision-making (Def. 3.4.2) can mitigate u-regret depends on the particular MABUC instance's functional relationship between the UCs and the decision X, i.e., $U \to X$ (and thus, for ISDM, $U \to I$). Indeed, as demonstrated in Table 3.2, the ability of intent to provide information about the state of U depends on both the true reward Y parameterization, $Y = f_Y(U, X)$, and the joint distribution over $P(I, U)$. This suggests that there are MABUC instances in which ISDM may fully mitigate u-regret (as in the Greedy Casino Example 3.1.1, demonstrated in the simulations to follow), but

other (arguably highly artificial) instances where it may not.[12]

To characterize the instances in which the version of ISDM presented in this chapter may not fully reduce u-regret, we must first introduce a new metric of regret specific to ISDM agents.

Definition 3.4.6. (Intent-specific Decision-maker Regret (i-Regret)) For a MABUC problem with time horizon T, decision variable $X \in \{x_1, ..., x_k\}$ (where $K = |X| \in \mathbb{N}, K \geq 2$ represents the number of choices), reward Y, and intent $I \in \{x_1, ..., x_k\}$ (where I is the intent experienced for decision X), the *optimal action* $x^*(i)$ is considered the one that maximizes expected reward under intent state $I = i$, defined as:

$$x^*(i) = \underset{x \in X}{\operatorname{argmax}} P(y_x|i) \tag{3.15}$$

The *i-regret* experienced by an agent using choice policy π at trial $0 < t < T$ is defined as:

$$r_t^i = P(y_{x^*(i_t)}|i_t) - y_{x_t^\pi} \tag{3.16}$$

The *cumulative i-regret* experienced by an agent across all T trials is thus:

$$R_T^i = \sum_{t=1}^{T} r_t^i = \sum_{t=1}^{T} P(y_{x^*(i_t)}|i_t) - y_{x_t^\pi} \tag{3.17}$$

Equipped with this definition, we demonstrate that ISDM is superior to traditional decision-making strategies in MABUC problems.

Theorem 3.4.2 (RDT u-regret Reduction is Superior to CDT)**.** Let R_t^u be the cumulative u-regret (Def. 3.2.2) experienced by an agent in a MABUC problem by trial t. If $R_t^u(CDT)$ represents the u-regret experienced by a CDT agent and $R_t^u(RDT)$ represents the u-regret experienced by an RDT agent, then as $t \to \infty$, $R_t^u(RDT) \leq R_t^u(CDT)$ for all possible MABUC parameterizations.

Proof. See appendix for proof of Theorem 3.4.2. □

[12]This issue is addressed in Ch. 6.

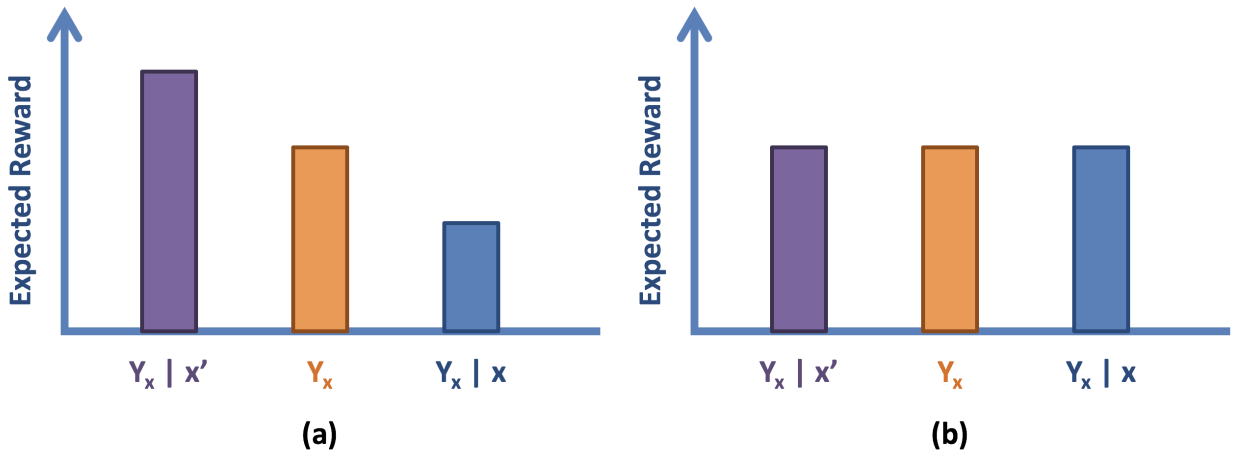

Figure 3.7: Graphical depictions of expected reward in intent-specific strata of Y_x. Counterfactual quantities are displayed in purple, experimental in orange, and observational in blue. (a) Demonstrates a scenario wherein RDT provides a superior maximization target (as in the Greedy Casino Example), and (b) depicts one in which RDT does no better, but no worse, than CDT.

The result of Theorem 3.4.2 is that, in terms of converging to an optimal choice policy and mitigating u-regret, RDT agents are always as good, if not better, than CDT agents (the traditional approach in MAB problems). This result can be visualized in Figure 3.7, wherein intent-specific reward quantities are always more or equivalently informative as experimental ones (by virtue of the interventional reward distribution being a probability-weighted sum over intent-specific arm rewards; see proof). The outcomes of employing a CDT vs. RDT maximization target are illustrated in Table 3.4.

Strategy (↓) & Scenario (→)	MAB	MABUC
CDT	Converge	¬ Converge
RDT	Converge	Converge

Table 3.4: Outcomes of employing a CDT vs. RDT maximization target in both MAB and MABUC scenarios; "Converges" indicates that the strategy will converge to the optimal choice policy.

Convergence may be slower for an RDT agent in a MAB setting than a CDT agent, but as Theorem 3.4.2 demonstrates, RDT agents obtain strictly more information during play, and could theoretically be equipped to distinguish a setting of no-confounding and then switch to a CDT optimization.

Lastly, we consider conditions under which RDT can minimize u-regret.

Theorem 3.4.3 (Sufficiency of i-regret Minimization for u-regret Minimization)**.** Let R_t^i be the cumulative i-regret (Def. 3.4.6) and R_t^u be the cumulative u-regret (Def. 3.7) experienced by an ISDM agent in a MABUC problem by trial t. As $t \to \infty$, if $R_t^i = O(1)$ then $R_t^u = O(1)$ if the following equivalence holds:

$$x^*(u_t) = \underset{x \in X}{\operatorname{argmax}} P(y_x|u_t) = \underset{x \in X}{\operatorname{argmax}} P(y_x|i_t) = x^*(i_t) \ \forall \ u_t \tag{3.18}$$

In words, sub-linear cumulative i-regret will imply sub-linear cumulative u-regret if the optimal action under known confounder state $U = u_t$ is the same as the optimal action under experienced intent $I = i_t$ for all trials $t \in T$.

Proof. See appendix for proof of Theorem 3.4.3. □

In the Greedy Casino Example 3.1.1, observe that when the criteria in Theorem 3.4.3 holds, minimizing i-regret likewise minimizes u-regret due to agreement in optimal arm choices.

$P(y_x\|D,B);P(y_x\|I)$	$D=0$		$D=1$	
	$B=0$	$B=1$	$B=0$	$B=1$
	$I=0$	$I=1$	$I=1$	$I=0$
$X=0$	$0.10; 0.15$	$0.50^*; 0.45^*$	$0.40^*; 0.45^*$	$0.20; 0.15$
$X=1$	$0.50^*; 0.45^*$	$0.10; 0.15$	$0.20; 0.15$	$0.40^*; 0.45^*$

Table 3.5: Table illustrating the Greedy Casino MABUC parameterization under which $x^*(u) = x^*(i) \ \forall \ u$, implying that i-regret is equivalent to u-regret. Optimal arm choices, based on maximal expected reward, are indicated by asterisks (*).

3.5 MABUC Simulations

With Intent-specific Decision-making (ISDM) defined, and its applications to reinforcement learning formalized through Regret Decision Theory (RDT), we now aim to verify that it works in practice. Towards this goal, we will demonstrate that ISDM mitigates u-regret in a variety of MABUC scenarios, firstly (in this section) through simulation support, and then (in Chapter 4), in a human-subjects study. The first will establish the ground-truth, illustrating the merit of ISDM in a model known to the experimenter, and the second will showcase its merit in a real-world MABUC scenario.[13]

3.5.1 Simulation Interpretation

Before we detail the specifics of the MABUC simulation procedure, we will describe two similarly modeled scenarios to which the simulations are expected to apply. We should also note that the following scenarios are both modeled by the prototypical MABUC SDM (Def. 3.3.1) displayed in Figure 3.4. These scenarios distinguish between what entities in the environment are the *learning agent* (or simply, *agent*, for short) and those that are the *actors*. An agent can be defined as the learning, rational decision-maker who is maintaining a history of intents, actions, and payouts as well as a policy that maps each trial's intent to a finally-chosen action (including all of the balancing between exploration and exploitation as might be assumed within a dynamic experiment or traditional MAB problem). An actor can be defined as the entity who is experiencing the intent (and who is assumed to be under the influence of confounding factors between the choice and feedback variables) and who is making a final arm choice.

There are reasonable scenarios that are captured when the agent and the actor are the *same* entity, as well as those in which they are *separate*. We will dissect these different scenarios in reference to Figure 3.8.

[13]Following [BFP15], other studies have shown that ISDM successfully applies to reinforcement learning tasks wherein trials are not independent in what are known as Markov Decision Processes with Unobserved Confounders [ZB16].

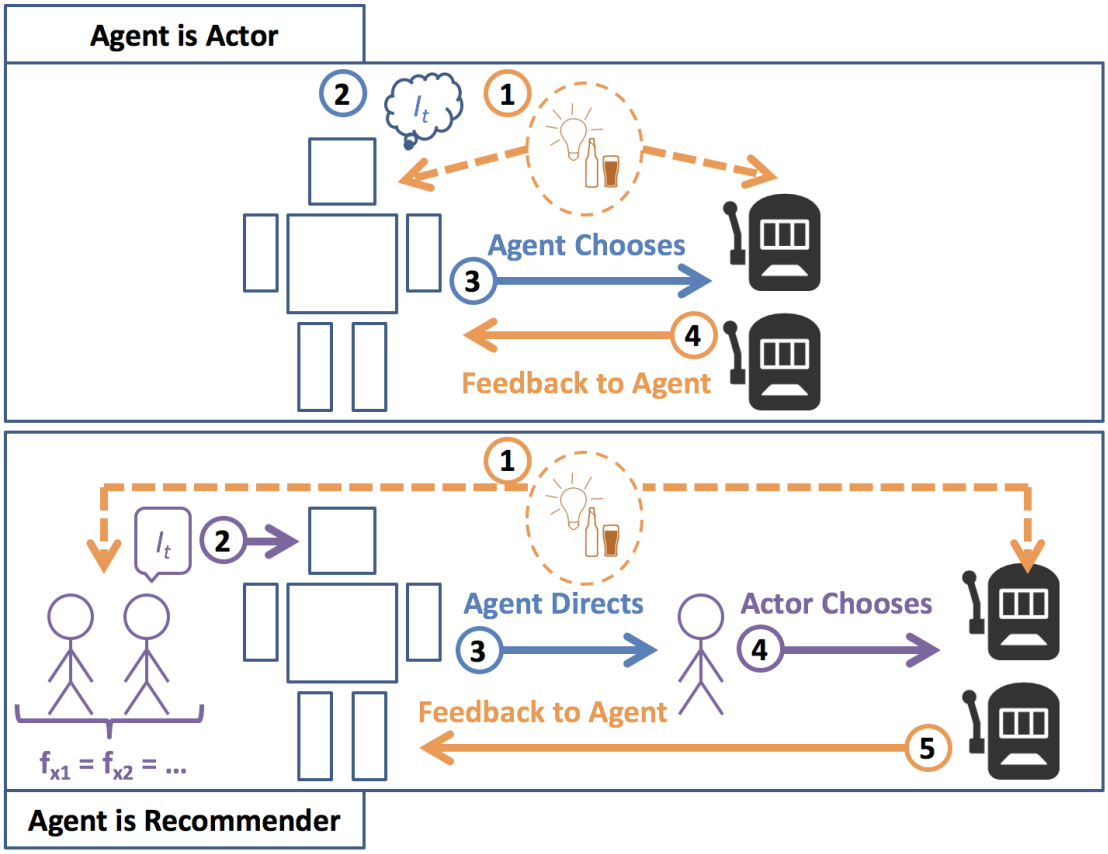

Figure 3.8: Interpretations of the MABUC simulations that employ the same SDM, but may have separate agents and actors. Pictured: [top] the agent (blue) and the actor (also blue) are the same entity; [bottom] the agent (blue) and actor(s) (purple) are distinct entities. In both panels, the environment's states and actions are drawn in orange.

Scenario 1: Agent is the Actor. This scenario presents ISDM as a means of self-reflection or meta-cognition, in which the learner is the same entity as the one that is affected by UCs but is also recording its own intent, action, reward histories. In the Greedy Casino Example 3.1.1, this would be akin to a gambler on the casino floor practicing ISDM as a rational decision-maker that is cognizant of their own intents being subject to confounding influences. Other examples of this scenario might include: support for legacy autonomous decision systems (in which certain policies may not scale to perturbations in the environment since their inception) or for human decision-makers in MABUC scenarios like the Greedy

Casino, or other reinforcement learning tasks that require reflection on one's own decision-making mechanisms (see Chapter 4 for an experiment that fits this scenario).

In Figure 3.8, Scenario 1 is depicted in the "Agent is Actor" panel with the following steps:

1. Environmental state of background factors (including any UCs) is realized for a particular trial in the MABUC.

2. Based on the state of those factors, the agent develops an intended choice (which is salient to it).

3. Based on its current intent, and history of recorded intents, actions, and rewards, the agent commits to a final arm choice.

4. The environment provides feedback to the agent based on that choice, which is then recorded in its history along with the intent and action that accompanied it.

Scenario 2: Agent is the Recommender. This scenario presents ISDM as a recommender system, in which the rational agent is the learner attempting to maximize rewards for (not necessarily rational) actors. In the Greedy Casino Example 3.1.1, this would be akin to a recommender machine (which, itself, is not directly affected by any UCs) on the casino floor, to which actors could divulge their intended arm choice, and the agent would provide an arm recommendation for the actor to pull. The distinction is that actors may be sensitive to confounding influences in the environment, resulting in irrational (or non-optimal) decisions, but the agent is not, and can address the same exploration vs. exploitation trade-off in its recommendations to actors to achieve an optimal intent-specific policy. Other examples of this scenario might include a recommender system for doctors making treatment decisions during an outbreak of an unknown ailment, under which a condition could present as a separate one, but have harmful treatment consequences if the conditions were conflated.

Herein, there are several important assumptions that, if broken, could distinguish Scenario 2 from Scenario 1: (1) We assume that every actor shares the same intent-function, f_x (not necessarily the same intent i_t at every trial), namely, that the SDM function deciding

the observational arm choice is the same for all actors (though changes in the environment between actors may vary the realized values of those functions, like one actor being drunk and another being sober in the Greedy Casino). Formally, we refer to two actors having the same intent-function as having *homogeneous intent*, defined as:

Definition 3.5.1. (Homogeneous Intent) Let A_1 and A_2 be two agents within a MABUC instance, and $M_{A_1}^{\Pi}$ be the SDM associated with the choice policies of A_1 and likewise $M_{A_2}^{\Pi}$ be the SDM associated with the choice policies of A_2. For any decision variable $X \in \Pi_M$ and its associated intent $I_x = f_x$, the agents are said to have homogeneous intent if $f_{I_x}^{A_1} \in F_{M_1}$ and $f_{I_x}^{A_2} \in F_{M_2}$ are equivalent, viz., if $f_{I_x}^{A_1} = f_{I_x}^{A_2}$.

In Chapter 6, we relax this assumption, and demonstrate how *heterogeneous intent* can be a boon rather than a restriction. (2) We assume perfect honesty and perfect compliance, meaning that actors will always divulge their true intent, and will always carry out the recommendation of the agent. (3) Trials do not represent repeated experiments with the same actor, but rather, that a new actor has approached the agent with its intent (which, by assumption (1), is licensed because actors are assumed to be exchangeable), and the agent then provides its recommendation.

In Figure 3.8, Scenario 2 is depicted in the "Agent is Recommender" panel with the following steps:

1. Environmental state of background factors (including any UCs) is realized for a particular trial in the MABUC.

2. Based on the state of those factors, the actor develops an intended choice (which is salient to it), and divulges this intended choice to the agent.

3. Based on the actor's expressed intent, and the agent's history of recorded intents, actions, and rewards, the agent recommends a final arm choice to the actor.

4. The actor complies with that recommendation (which was made via the agent's choice policy, which may be leveraging exploration and exploitation), and pulls the recommended arm.

5. The agent observes feedback from the environment based on that choice, which is then recorded in its history along with the actor's intent and action that accompanied it.

	Agent is Actor	Agent is Recommender
History is maintained by...	Agent	Agent
Intent is experienced by...	Agent	Actor
Arm being chosen according to...	Agent	Agent
Arm being pulled by...	Agent	Actor
Each trial involves...	The same agent	A separate actor

Table 3.6: Summary of two scenarios comparing the identities of the agent and actor in a MABUC; simulation results can be interpreted as a consequence of either scenario.

In summary, Table 3.6 provides a comparison for differences in how we might interpret facets of a MABUC problem between the "Agent is the Actor" vs. "Agent is the Recommender" scenarios, but the SDM and simulation results will be the same for each. With these interpretations in place, we describe the simulation procedure and results in the following segment.

3.5.2 Simulation Procedure & Results

Intent-specific Decision-making is a decision-making framework that can be flexibly applied to existing MAB learning algorithms. Note that it gives a prescription for how to reduce u-regret in decision-making scenarios with unobserved confounders, but it does not provide a separate algorithm for leveraging exploration and exploitation of arms *within* each intent condition in the dynamic experiment sense; for this aspect of the learning problem, we may consult the MAB literature for approaches on which ISDM can be layered atop. Towards this end, we chose to use a Thompson Sampling (TS) bandit player (see [OB10, CL11, AG11] for a review of TS and its results as a competitive MAB learning algorithm) as the basis for the present simulations.[14]

[14] All simulation source code for Chapter 3 can be found at:
https://github.com/Forns/ucla-forns/tree/master/projects/dissertation/ch3.

Apropos, we next detail two algorithms that will be useful for interpreting the results of simulations: (1) the algorithm for the actual simulation procedure is described in Algorithm 1 and (2) the algorithm for the RDT-enhanced TS bandit player is described in Algorithm 2. All choice policy algorithms (TS^{RDT} included) are invoked for a decision in the MABUC-Sim algorithm for every policy action selection step (Line 7 in Algorithm 1).

Algorithm 1 MABUC Simulation

1: **procedure** $MABUC - Sim(\mathrm{T})$
2: $R^u \leftarrow 0$ (initialize cum. u-regret)
3: $H \leftarrow \{\}$ (initialize history)
4: **for** $t = [1, ..., T]$ **do**
5: $u_t \leftarrow f_u(...)$ (realize environmental factors for trial)
6: $i_t \leftarrow f_x(u_t)$ (intent is initialized for trial)
7: $x_t \leftarrow f_\Pi(i_t, h_t)$ [15](policy selects final decision)
8: $y_t \leftarrow f_y(x_t, u_t)$ (reward is observed from chosen arm)
9: $H \leftarrow H \cup \{i_t, x_t, y_t\}$ (history is updated)
10: $r_t^u \leftarrow P(Y_{x_t^*}|u_t) - y_t$ (u-regret is logged)
11: $R^u \leftarrow R^u + r_t^u$ (cum. u-regret is updated)

Algorithm 2 RDT Thompson Sampling

1: **procedure** $TS^{RDT}(i_t, h_t)$
2: $s_t \leftarrow [\#Y_{x_0} = 1|i_t, ..., \#Y_{x_k} = 1|i_t]_{h_t}$ (count number of successes for each intent-arm)
3: $f_t \leftarrow [\#Y_{x_0} = 0|i_t, ..., \#Y_{x_k} = 0|i_t]_{h_t}$ (count number of failures for each intent-arm)
4: $A_t \leftarrow [\beta(s_t[1], f_t[1]), ..., \beta(s_t[k], f_t[k])]$ (sample from beta-dists. of each intent-arm)
5: $x_t \leftarrow \mathrm{argmax}_{x \in [1,k]} A_t$ (choose max)
6: **return** x_t

Procedure. All reported simulations are partitioned into rounds of $T = 1000$ trials averaged over $N = 1000$ Monte Carlo repetitions. In brief, at each trial in a single repetition, (1) values for the unobserved confounders u_t and resultant intent i_t are instantiated by their

[15]Agents maximizing via RDT will consider the intent, i_t, but others will not.

respective structural equations (see Example 3.1.1), (2) the player chooses an arm based on their given strategy to maximize reward (depending on whether each algorithm invokes RDT, CDT, or EDT), and finally, (3) the player receives a Bernoulli reward $Y \in \{0, 1\}$ and (4) records the outcome in the history. In this section, we conducted experiments across 2 reward parameterizations (described below), but the findings generalize across choices of payout parameters (as proven in the previous section).

Candidate algorithms. *CDT Thompson Sampling* (TS^{CDT}) attempts to maximize rewards based on the CDT optimization criteria, ignoring intent; since other traditional CDT algorithms will perform more or less equivalently to TS^{CDT} in MABUC scenarios (see Figure 3.1), we omit their performance for clarity. *RDT Thompson Sampling* (TS^{RDT}) operates according to Algorithm 2, maximizing the counterfactual ETT. For baseline comparison, we also display the performance of an irrational, "observational" player (*Obs.*) who abides by intent at every trial, and an irrational "experimental" player (*Exp.*) who chooses arms at random.

Evaluation metrics. We assessed each algorithms' performances with MABUC evaluation metrics: (1) the probability of choosing the optimal arm under each round's confounder state $U_t = u_t$, and (2) the cumulative u-regret. As in traditional bandit problems, these measures are recorded as a function of the time step t averaged over all N round repetitions. Note that this metric is available to us as the simulation designers, but would be rarely available in reality without access to the fully-specified model (including knowledge of the states of the UCs).

Experiment 1: "Greedy Casino." The Greedy Casino parameterization (specified in Table 3.1) illustrates the scenario where each arm's payout appears to be equivalent under the observational and experimental distributions alone. Only when we concert the two distributions and condition on a player's intent can we obtain the optimal policy. Simulations for Experiment 1 support the efficacy of ISDM (see Figure 3.9). Analyses revealed a significant difference in the u-regret experienced by TS^{RDT} ($M = 11.07, SD = 16.34$) compared to TS^{CDT} ($M = 149.22, SD = 14.37$), $t(1998) = 200.75, p < .001$.

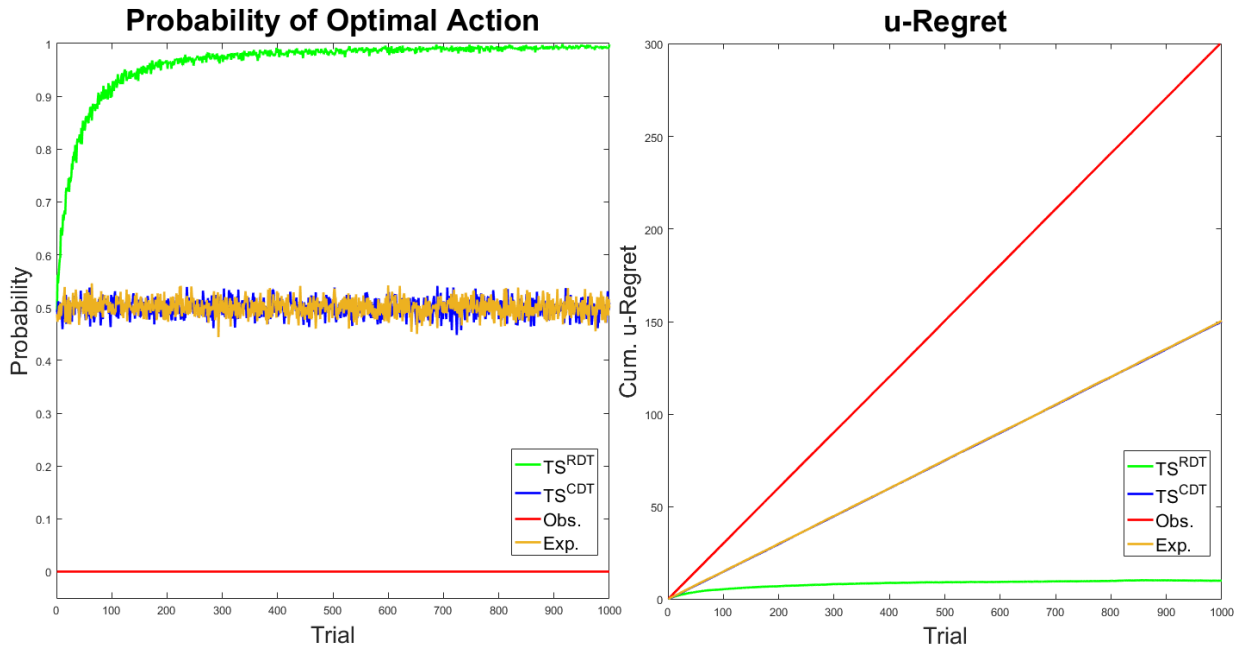

Figure 3.9: Simulation results for Experiment 1, the Greedy Casino scenario.

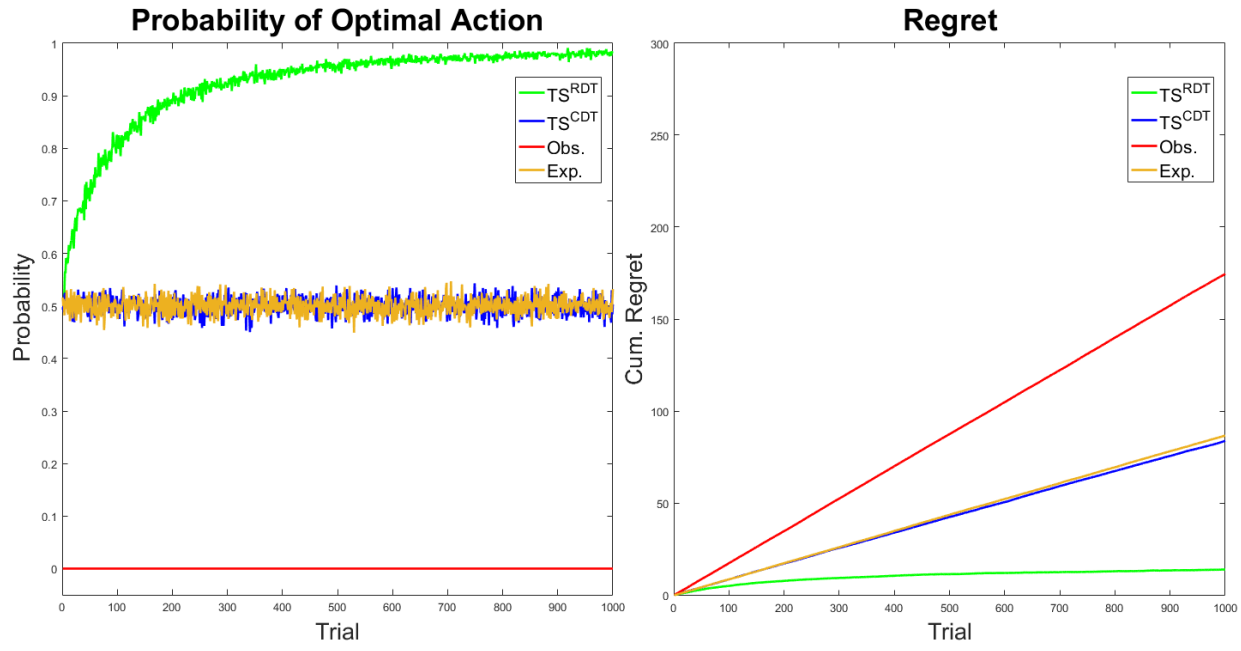

Figure 3.10: Simulation results for Experiment 2, the Paradoxical Switching scenario.

(a)	$D=0$		$D=1$	
$P(y\|X,D,B)$	$B=0$	$B=1$	$B=0$	$B=1$
$X=0$	$^i0.40$	0.30	0.30	$^i0.40$
$X=1$	0.60	$^i0.10$	$^i0.20$	0.60

(b)	$P(y\|X)$	$P(y\|do(X))$
$X=0$	0.40	0.35
$X=1$	0.15	0.375

Table 3.7: Paradoxical Switching: (a) Payout rates decided by reactive slot machines as a function of arm choice, sobriety, and machine conspicuousness. Players' natural arm choices under D, B are indicated by the superscript i, to indicate intent. (b) Payout rates according to the observational, $P(Y=1|X)$, and experimental $P(Y=1|do(X))$, distributions, where $Y=1$ represents winning (shown in the table), and 0 otherwise.

Experiment 2: "Paradoxical Switching." The Paradoxical Switching parameterization (specified in Table 3.7) illustrates the scenario where one arm ($X=0$) appears superior in the observational distribution, but the other arm ($X=1$) appears superior in the experimental. Again, we must use RDT to resolve this ambiguity and obtain the optimal policy. Simulations for Experiment 2 also support the efficacy of ISDM (see Figure 3.10). Analyses revealed a significant difference in the regret experienced by TS^{RDT} ($M=13.11, SD=16.81$) compared to TS^{CDT} ($M=84.24, SD=15.89$), $t(1998)=97.25, p<.001$.

3.6 Conclusion

In this chapter, we demonstrated the problems introduced by UCs in confounded decision-making scenarios, as motivated by the Greedy Casino example. Moreover, this demonstration showed that experimental measures of treatment efficacy average the influence of the UCs over the outcome, and do not consider their state at any given trial to determine the best unit-level intervention. Counterfactual quantities provide the desired unit-level granularity, but can only be computed in certain scenarios or in possession of the fully-specified causal model, including the distribution over UC states. When we do not possess such a model, but can condition on a deciding agent's observational arm choice before it is chosen (i.e., the agent's intent), we can empirically estimate the ETT that uses intent as an observed

proxy for the UC state. Using these new, empirically estimable counterfactual quantities, intelligent agents can then maximize the efficacy of interventions under the context of intent by the tenets of a new decision-making theory called Regret Decision Theory (RDT). We then proved the superiority of RDT versus traditional agents that maximize experimental rewards in confounded decision-making tasks, and corroborated our theoretical results in simulated MABUC scenarios.

CHAPTER 4

Human-Subjects Intent-Specific Decision-Making

The cognitive sciences have studied human counterfactual reasoning in application of a variety of cognitive tasks (see [Byr16] for an overview). Included in this variety are counterfactuals as rationalizations of past events, such as a student failing an exam and regretting that they *would have* done better *had they* studied more, or gotten more sleep, etc. [McC08, TMK12]. Others consider how counterfactuals help humans make causal inferences; for example, that a drug associated with certain side effects would not have hindered those that took it *had they taken* another drug without those effects [SEB05]. Still others have investigated applications tangential to the present work, but include counterfactual reasoning as modulators of emotional experiences (as through survivor's regret [TKT11]) and moral judgments (for example, associating blame, sympathy, and punishment [MMG93, MD05, SG14].

The present work, however, focuses on application of counterfactual reasoning towards forward thinking, planning, and decision-making. Previous efforts have studied human counterfactual logic in pursuit of these goals, finding that people tend to leverage past regrets as motivators for changes in future behavior such as wishing to study more for an upcoming exam assuming that they *would have* performed better on past exams *had they* studied more [FGS13]. Viewing counterfactual outcomes from past experiences also appears to be tied to a priming for intentions about future instances of similar scenarios [SM12]. fMRI studies have lent support for this claim, finding that brain regions associated with episodic memory recall and planning were activated during exercises with counterfactual thought [SBD15, DSM15].

Chapter 4 is an extended version of [FWB].

This latter binding of counterfactual inference to experiential history may support the model proposed by the tenets of intent-specific decision-making (ISDM), since they are based, on only a partially-specified causal model (i.e., without knowing the precise functions that decide the value of a particular variable or knowledge of the identities and states of any unobserved factors). The plausibility of the ISDM framework as explanation for how humans learn through counterfactual reasoning is a chief query in the present chapter, but other theories have also tried to explain the mechanisms of retrospect. Several suggest that humans focus on counterfactual "fault lines," common additive counterfactuals that add an extra proposition to the event in question, such as "If he had only worn a seat-belt, he would not have been injured in the car crash" (where the fault line is the addition of the seat belt to the past scenario) due to those additions being within the agent's locus of control [ER08]. Support for the effect of regret on learning is also found in lesion studies, wherein participants with damage to the prefrontal cortex fail to learn from past errors and do not contextualize these mistakes with admissions of what they could have done differently – a common occurrence in healthy individuals [BGA05]. Moreover, studies examining the counterfactual reasoning capacities of participants with impaired memory had difficulties with what-if scenarios depending upon spatial scenes in the participant's past [MM14].

Evidence, like the above, surveyed from the existing literature strongly suggests that semantic episodic memories are integral components of human counterfactual reasoning. However, as a recent survey of this literature states, there exist few computational theories or models of human cognitive processes in which learning occurs from counterfactual reasoning, as well as how the fusion of experience and hypothesis combine to produce behavioral change [Byr16]. As such, the present work seeks to propose ISDM as a model for how humans may experientially compute counterfactuals with the added difficulty of invisible counterfactual "fault lines" (i.e., when unobserved confounders (UCs) exist between some action and some outcome, and the agent's intent serves as the only indication for what could have been done differently in the past). To the best of our knowledge, it is the first work to propose a model for computing empirical counterfactuals in the face of UCs, and explain how learning will result.

With the formalisms and models of ISDM in place from Chapter 3, we now seek to supply some empirical support for the efficacy of ISDM in a real-world, human-agent, reinforcement-learning task. In addition to addressing the questions posed above, there are a variety of motivations for this study surrounding the formalizations of ISDM; in particular, our research objectives are to determine that:

1. *Intent is an isolable signal.* A fundamental tenet of ISDM is that agents (either as the actors themselves or as recommenders to those experiencing intent, as depicted in Figure 3.8) are able to isolate their intended action choice and either employ it as a contextual variable for decision-making or express it to a learning system that will. If humans *are* able to isolate their intent, then we should consequently verify that they are able to employ it in pursuit of optimizing the quality of their decisions. If humans are *not* capable of determining which action corresponds with their intent, then the application of ISDM to human confounded decision-making tasks is seriously compromised.

2. *Intent is reactive.* In the formalization of ISDM from Chapter 3, a key modelling assumption is that formation of intent is independent from the agent's experiential history. In other words, we assume that intent remains an "impulse" to environmental conditions via System 1 cognitive processes, rather than deliberative ones implicit in System 2. If intent is *not* independent from history, then at two trials t_1, t_2, we may have $P(U|I_{t_1}) \neq P(U|I_{t_2}) \Rightarrow P(Y_x|I_{t_1}) \neq P(Y_x|I_{t_2})$, meaning that intent-specific learning may not occur.

3. *ISDM is a naturally employed human counterfactual reasoning mechanism.* If humans exhibit patterns of learning by regret (i.e., by counterfactual reasoning) that follows the tenets of ISDM, then it is possible that the present work has modeled an important aspect of cognition and learning. As such, we wish to determine if humans naturally use their intent as a reasoning mechanism (i.e., without outside instruction to do so), or if not, that they can improve the quality of their decisions when they are told to employ ISDM (as by the tenets of RDT in this reinforcement learning scenario).

Towards these ends, we sought to create a learning task for humans that could be modeled as a MABUC problem, in which UCs affect both the participant's intended action choice and the resulting outcome. Furthermore, we wished to incorporate experiential history as a key component of the learning task, but which would be independent of any participant's prior knowledge. Well suited to developing confounded intents in the MABUC specification is a word association task, wherein, given a "cue" word, participants produce the first word that comes to mind. Databases of these word associations, including the strengths of associations between a wide variety of cues and their responses, demonstrate that English speakers produce predictable responses when presented with certain cue words [NMS04].

Leveraging this fact for the research question at hand, we presented participants with a series of cue words and asked them to choose from a selection of two answer (or "target") words, one with a strong association with the cue and one with a weak association with the cue. The answer that was considered correct was always the target that was the least associated with the cue, while the incorrect answer was the target with the strongest association with the cue. In three separate experimental groups, participants were either, 1) directly told the intent-specific strategy they needed to implement to maximize correct responses, 2) indirectly told the intent-specific strategy, or 3) given no information about the intent specific strategy. Participants who successfully navigated the task would thus have to acknowledge their intended choice, but then finally select the opposite.

4.1 Methods

The quiz was conducted using Amazon's Mechanical Turk, a web service through which workers can accept various jobs like surveys. Workers in the United States' Mechanical Turk population have been found to be relatively representative of the nation as a whole, with a few demographic exceptions (workers tend to be predominantly female and with lower income levels) [RZI09]. We restricted quiz participants to those residing in the United States, with English as a first language, and over the age of 18. We recruited 180 participants to take the quiz, but 15 needed to be discarded after Mechanical Turk's screening mechanism failed

to prevent non-US residents and non-first-language English speakers from taking it (which were later caught after collecting participant demographics at the end of the quiz). Of the remaining 165 participants, 69 were male, 96 were female, and were primarily middle-aged ($M = 37.87, SD = 11.76$).

Participants were self-selected as workers on Amazon's Mechanical Turk who chose to take part in the study advertised as a "Psychological test investigating aspects of decision-making." Additionally, they were offered incentives of \$0.10 for their participation, and a bonus of \$0.01 for every question that they answered correctly. We emphasize that this compensation is what makes the task a reinforcement learning problem, and is important for ensuring that participants are motivated to improve their policy over time.

4.1.1 Materials

The MABUC task was constructed from 50 cue words, each with 2 accompanying associational answer (i.e., target) words from the University of South Florida Free Association Norms database [NMS04]. Cue-target triplets were selected via the following guidelines: (1) We attempted to avoid overlap of semantics between any two triplets such that priming was generated within, but not between, any triplets. (2) We attempted to choose cues that possessed a disparity in associational strength between the strongly and weakly associated target of approximately 0.3; the average associational strength of the strong targets was 0.34 with a standard deviation of 0.03, and the average strength of the weak targets was 0.04 with a standard deviation of 0.02. (3) Lastly, we attempted to choose cues that spanned the alphabet, choosing 2 cues beginning with each letter, except for X and W, which had too few cue-target triplets matching our criteria (X had no matching cues, and W only 1), and S, which had 1 extra to make up the deficit. In each question, the target word that was strongly associated with the cue was always the "incorrect" choice, and the one that was weakly associated with the cue was always "correct." For example, when presented with the cue "dart," the answer choices were "board," the strongly associated target, and "throw," the weakly associated (correct) target. See Table B.1 in the appendix for a full list of the

cue-target triplets and their associational strengths.[1]

4.1.2 Procedure

Briefing. To begin the MABUC task, workers would click on the link, with participation contingent upon them acknowledging that they were a native English speaker, over the age of 18, and residing in the United States (the Mechanical Turk engine allowed us to confirm these final two demographics through their system). Along with this briefing, participants were given the contact information of the experimenters as well as the details for informed consent: they were told that their participation was completely voluntary and that they could withdraw from the study without penalty at any time. Having agreed to the terms of the study, participants clicked a button to continue to the instructions for the quiz. See Figure B.1 in the appendix for the exact briefing screen presented to participants.

Experimental Conditions. Upon volunteering participation in the previous section, participants were then randomly assigned to one of three experimental conditions, designed to test the capacity of humans to employ ISDM without instruction, as well as their capacity to carry it out when instructed to. The intervention in each of the three conditions was merely how much of a hint towards using ISDM the participant was given, and are detailed as follows:

1. *No hint*. In the no hint condition, participants were left to determine their own reward maximizing strategy on their own, with no indication that their intent would serve as a useful decision-making mechanism.

2. *Weak hint*. In the weak hint condition, participants were told that "the hidden rule is the same for each question in the quiz and is either that: the answer choice you feel like choosing first will likely (1) always be correct OR (2) never be correct," and that

[1]The quiz itself was implemented as a web application whose mechanics were built upon the TurkSuite Template Generator [Mor14]; for quiz source code, visit:
https://github.com/Forns/ucla-forns/tree/master/projects/dissertation/ch4
To take the quiz itself, visit:
https://rawgit.com/Forns/ucla-forns/master/projects/dissertation/ch4/index.html

they should take this into account before making their final choice.

3. *Strong hint.* In the strong hint condition, participants were explicitly told that "the hidden rule is the same for each question in the quiz and is that the answer choice you feel like choosing first will likely never be correct," and that they should take this into account before making their final choice.

Depending upon their assigned hint condition, participants would see the hint displayed in the instructions and repeated after selecting an answer on each of the 50 questions that followed.

Instructions. After the briefing, participants were presented with the instructions screen, which provided the details of the quiz to follow. In particular, they were told:

1. At the top of each page, they would be shown a word and should consider the first thing that comes to mind when they see that word. After this consideration, they will click a button that will reveal two additional words that will be their answer choices, and were informed that the cue and target words would be presented to them in random order (to prevent them from trying to find order effects since there should be none).

2. They were then told that one of these two answer options would be considered the "correct" one by some hidden rule, and the other will be considered "incorrect." Once they had made a decision, they should click the answer they believe to be correct, but had to do so within a 30 second window in order to attain the bonus of $0.01 for a correct answer.

3. They were then given their goal: to develop their own decision-making strategy to answer as many questions correctly as possible.

4. Depending on their assigned experimental condition, they were then given their hint as to what the hidden rule entailed (if provided with a hint at all).

5. Finally, they were shown an example of a cue word with two sample answer choices, asked to do the experiment in one sitting without taking any breaks, and then shown

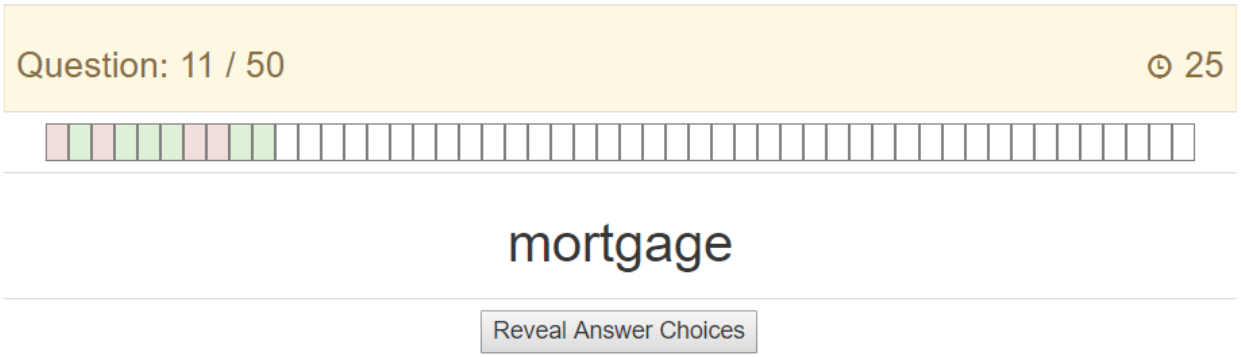

Figure 4.1: Sample pre-answer phase question in the quiz depicting the 11th cue word "mortgage" before the participant has revealed its answer choices. The pictorial representation of the participant's answer history is above the cue, with red blocks indicating incorrect answers, and green correct ones. The time remaining is shown at the top-right next to the small clock.

a button for them to begin when ready.

Quiz. Each of the 50 quiz questions were presented with randomly ordered cues and their accompanying target answer choices in randomly counterbalanced order. Each question also had had a pre- and post-answer phase, at which point the participant was shown the following information: In the *pre-answer phase*, a user would be presented with (1) the cue at the top of the screen, (2) a button to reveal the two answer choices, (3) a pictorial representation of their history of correct and incorrect answer choices, and (4) a timer that began at 30 seconds and counted down until the user provided an answer. If the participant ran out of time on the question, they would still be able to answer and be told if they were correct or incorrect, but would not receive the bonus \$0.01 if correct. A sample pre-answer phase question is depicted in Figure 4.1.

Having taken some time to consider the first thing that came to mind after seeing the cue, participants would then (1) click the "Reveal Answers" button (a feature created with the purpose of making users develop an intended answer from the cue before revealing any biasing choices), (2) make a final answer choice by clicking on one of the two revealed answers,

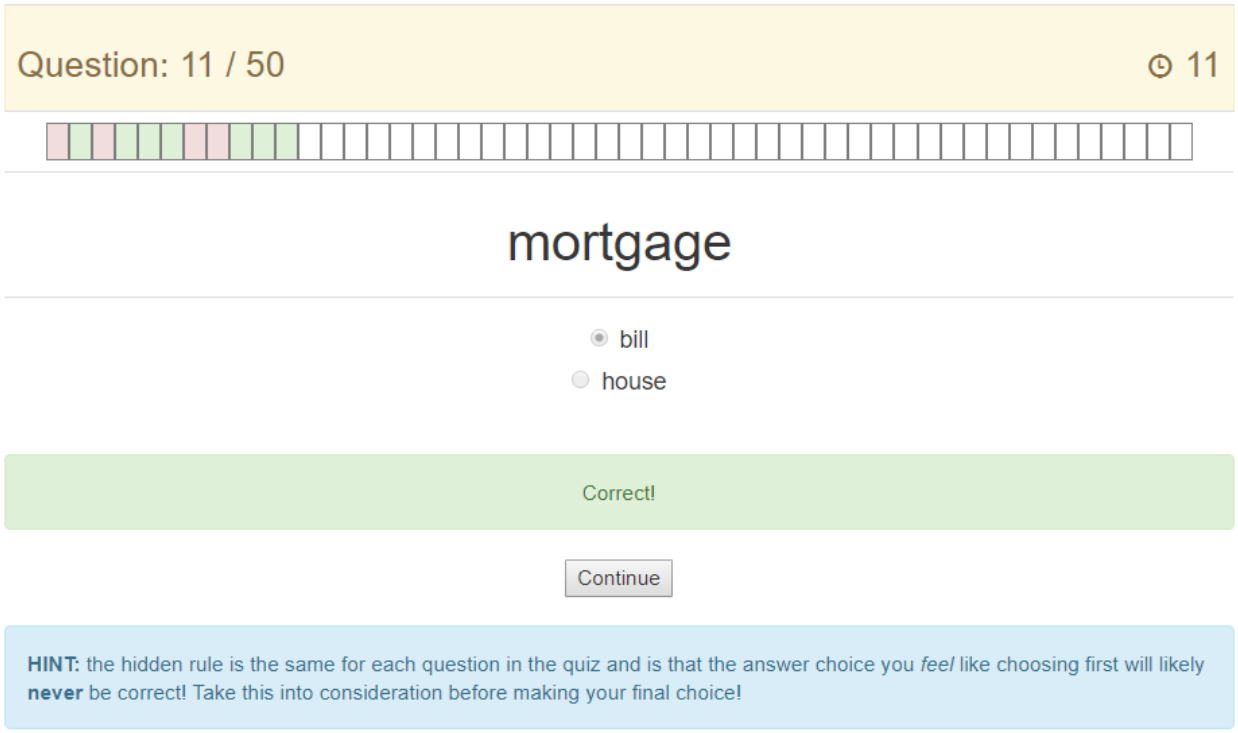

Figure 4.2: Sample post-answer phase question in the quiz after the participant correctly chose the weakly associated target "bill" to the cue "mortgage." Feedback is provided to the user in the form of a large "Correct!" box, followed by the (in the present example, a strong) hint to remind the participant of their objective. In the no hint condition, this box is absent.

and (3) view the reward from their final choice (correct or incorrect). In this *post-answer phase*, if participants belonged to either the weak or strong hint conditions, the hint would be repeated below their feedback, and all participants would then be given a separate button to continue to the next question. For each trial, we recorded the target chosen as well as the reaction time between the presentation of the cue and selection of the target. A sample post-answer phase question is depicted in Figure 4.2.

Rationale & Demographics. Following the 50 questions, we had two remaining sections for participants to complete: (1) they were asked to provide a brief description of their decision-making strategy, and (2) were asked to provide their basic demographic information, including age, gender, country of residence, first language, and political leaning. Having

provided this information, participants were shown their total answered correctly (and the accompanying bonus compensation that they earned), were thanked for their participation, and clicked a final submission button to forward their answers.

4.1.3 Analysis

Based upon the hidden associative rule that we employed, for each participant, we first computed the cumulative regret ($R = \sum_t (1 - Y_t)$)[2] and average reaction time across all $t \in T = 50$ trials in each of our three experimental conditions. We expected that if intent is an isolable signal, any or all of the experimental groups should have success in maximizing the correct responses. However, if participants were extracting the counter-intent rule, we would expect these participants to take longer to respond to the trial. We further expected that if intent is reactive, then participants in this study would not change their strategy over time and would continue to use the intent-specific strategy to their benefit. Lastly, if intent is naturally employed by humans, we should expect that even in the no hint condition, participants would be able to identify the intended response and respond with the counter-intent choice.

To verify that the experimental conditions translated to a strategy, we asked participants to explain their strategy at the conclusion of the quiz, and codified each response into one of four categories:

1. The *guessing* approach coded participants who clearly indicated that they had no strategy or were simply picking answers at random; sample responses in this class include "Random" and "Switched from top to bottom back and forth."

2. The *counter-intent* approach (the correct one) coded participants who clearly indicated that they were picking answers that were contrary to their first choice; sample responses in this class include "I tried to go with the opposite of my gut reaction" and "tried to pick the one that didnt [sic] feel right."

[2]Since we, the experimenters, do not possess the fully-specified SCM for this task, nor the intents of each participant, we cannot analyze the u-regret (Def. 3.2.2) experienced at each round.

3. The *intent* approach coded participants who clearly indicated that they were picking answers that were aligned with their first choice; sample responses in this class include "Picked the first that came to mind" and "I went by instinct."

4. The *other* approach coded participants who gave incoherent responses, or detailed strategies that were incorrect, but were not specifically related to their intent; sample responses in this class include tangential strategies like "I had various rules that I came up with throughout the study like synonyms, whether the word was a part of a larger group, or whether they were similar but different in some way" and incoherent responses like "Where I would find something."

We will first discuss the cumulative regret and reaction time results across all participants in each experimental group, regardless of the strategy employed. We will then present the results from the codification of the written responses that described their response strategy. Lastly, we will focus on participants that explicitly responded that they were using intent or counter-intent strategies, regardless of the hint that allowed them to develop their approach.

4.2 Results

Experimental Groups. After computing the average cumulative regret across all 50 trials, we found that the experimental groups significantly differed in cumulative regret, $F(2, 162) = 9.37, p < .001$. The Strong Hint group had significantly lower final cumulative regret average than the No Hint group, $t(108) = 4.38, p = .077$ and the Weak Hint group, $t(108) = 2.44, p = .016$. However, the No Hint group did not significantly differ from the Weak Hint group in cumulative regret, $t(108) = 1.78, p = .077$ (see Figure 4.4). There were no significant differences between experimental groups in average per trial reaction time (see Figure 4.3).

Next, we tested the rate at which each experimental group accumulated regret. Using a timeseries regression across the 50 trials, we found that cumulative regret accumulated significantly more rapidly in the No Hint group compared with the Weak Hint and the Strong Hint groups. Specifically, the slope of cumulative regret in the No Hint group ($\beta = .50$) was

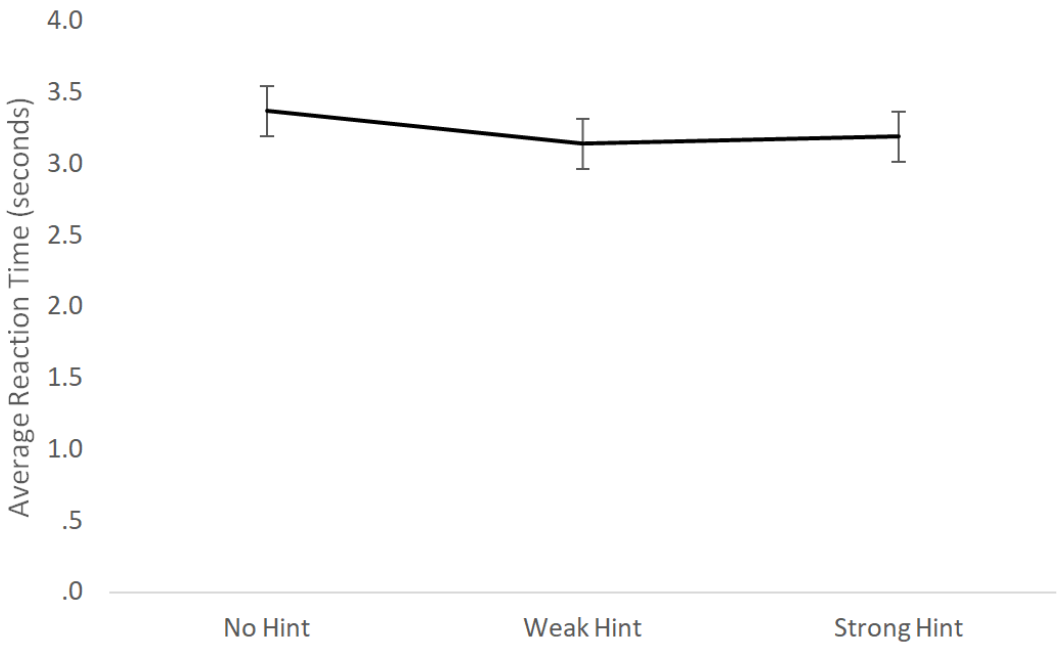

Figure 4.3: Average reaction time by participants in each experimental condition. Error bars represent standard errors about the mean.

significantly steeper than the slope in the Weak Hint group ($\beta = .46), t(96) = 14.04, p < .001$, and the Strong Hint group ($\beta = .39), t(96) = 35.64, p < .001$. Additionally, the slope of cumulative regret in the Weak Hint group was significantly steeper than the slope in the Strong Hint group, $t(96) = 28.50, p < .001$ (see Figure 4.4).

Similarly, we calculated the probability of the correct response at each trial based upon its trial number in the quiz. We averaged these in 10 trial increments to create a smoothed time series of correct responses (see Figure 4.5).

Strategies. After codifying participants' written responses on their decision-making strategy, a chi-square test of independence determined that experimental group was not independent from the strategy chosen. Unsurprisingly, participants in the Strong Hint group were more likely to use the counter-intent strategy than the intent strategy (Table 4.1) while those in the Weak and No Hint groups were less likely to use the counter-intent strategy $\chi^2(6) = 12.78, p = .047$. This effect was more apparent in a separate chi-square test that only included the counter-intent and intent strategies across the three groups,

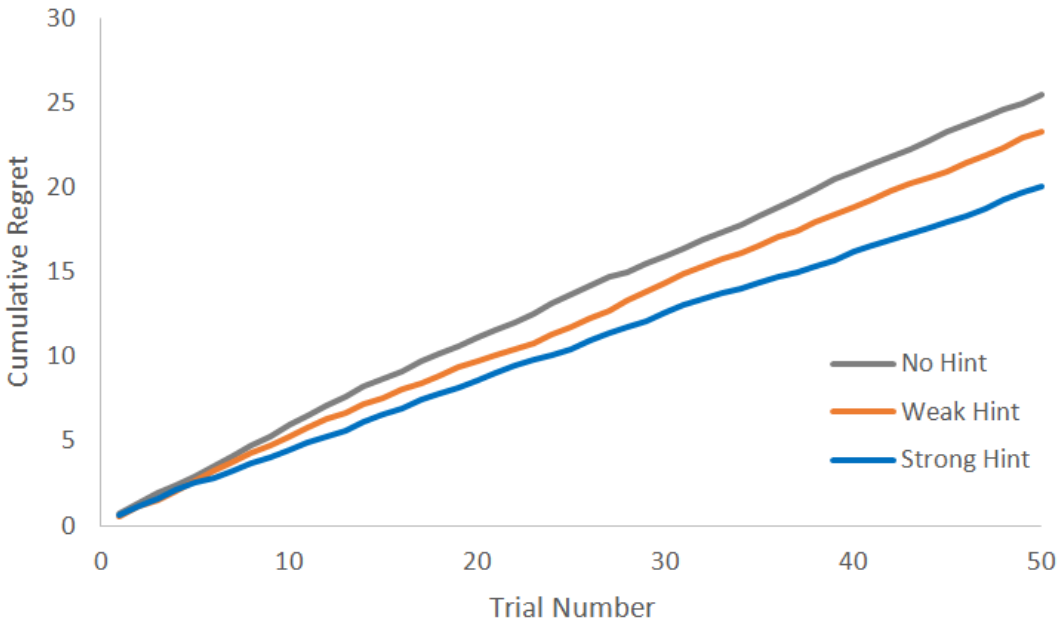

Figure 4.4: Timeseries of average cumulative regret experienced by participants in each experimental condition.

$\chi^2(2) = 11.90, p = .047$. The Weak and No Hint were more likely to use their intended response and less likely to use the counter-intent response than the Strong Hint group, suggesting that our main experimental manipulation elicited identification of the intent-specific strategy needed to obtain the correct answer.

Within the strategy types, there was a significant difference in the average final cumulative regret, such that those who used the intent strategy had overall higher cumulative regret than those who used the counter-intent strategy, $t(101) = 7.47, p < .001$. Further, comparing the slopes of each strategy in cumulative regret across trials, those who used the intent strategy accumulated regret at a significantly higher rate than those who used the counter-intent strategy, $t(96) = 48.96, p < .001$ (see Figure 4.6). Figure 4.7) also depicts the average cumulative regret experienced by those employing each of these two strategies within each condition.

We also found a significant difference in reaction times between counter-intent and intent

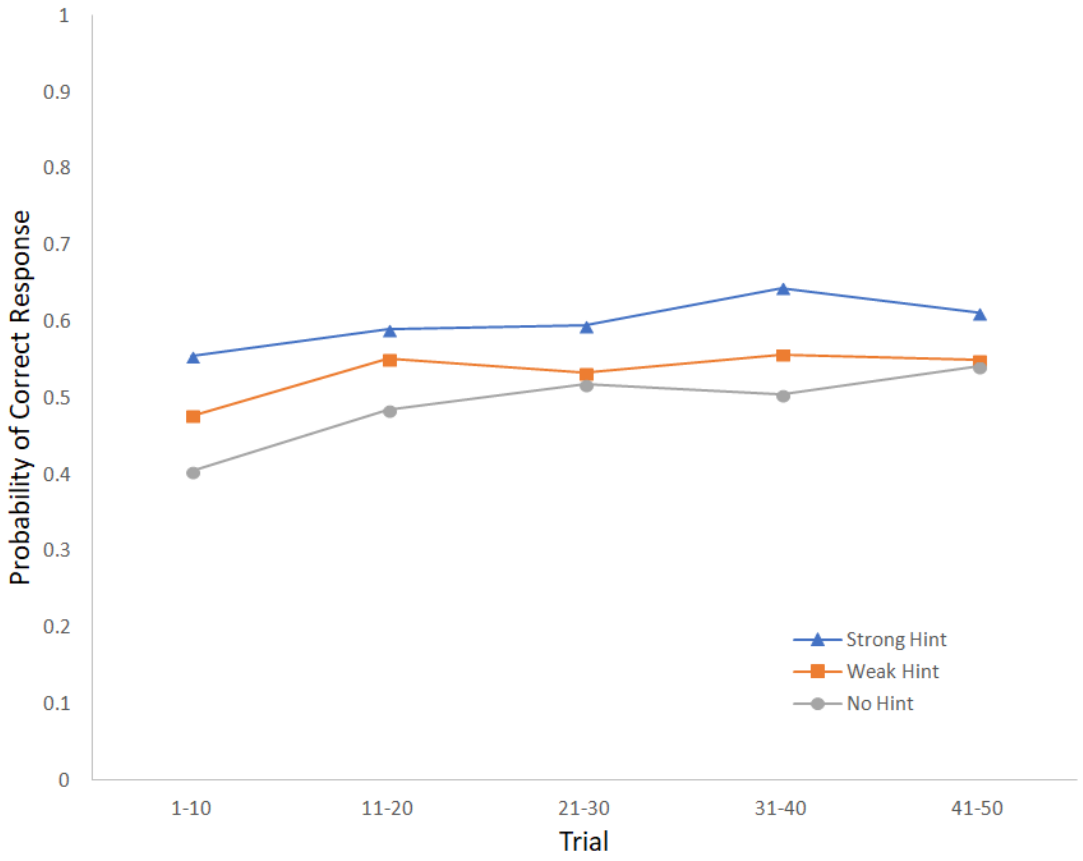

Figure 4.5: Probability of a correct response for each experimental condition within 10 trial increments across all 50 trials.

strategies, $t(101) = 3.06, p = .003$, such that those who used counter-intent strategies took significantly longer to respond to the cue than those who used the intent strategy. This finding supports the idea that participants using the counter-intent strategy were suppressing their intended response in order to respond counter to their intent. Importantly, when comparing cumulative regret and reaction time for each strategy across groups, we find similar patterns. Lastly, while we have outlined the main effects of the strategy and of the experimental group in both cumulative regret and reaction time, we note that there was no interaction effects between experimental groups and strategy (see Figure 4.8).

	Strategy			
Group	Random	Counter-intent	Intent	Other
No Hint	10(33%)	18(28%)	15(39%)	12(38%)
Weak Hint	11(37%)	17(26%)	18(47%)	9(28%)
Strong Hint	9(30%)	30(46%)	5(13%)	11(34%)

Table 4.1: Number of participants in each experimental group ($n = 55$ per group) that utilized various decision-making strategies. Numbers in parentheses indicate column percentages.

4.3 Discussion

Summarizing our results, we find that the experimental interventions were effective means of influencing each participant's policy formation and determining whether or not certain policies were formed naturally or required the intervention. Across experimental groups, those who adopted the RDT strategy experienced significantly less regret than those who did not; moreover, the group that was given the strongest suggestion to adopt the RDT approach (i.e., the Strong Hint condition) experienced significantly less regret than those left to their own policy formation. Reaction time was shown to be inversely correlated with regret; in other words, those who spent more time on each question tended to answer more correctly. This delay can be explained by the RDT strategy that demands participants first consider their intended choice, pause, and then make a final choice that is conditional upon the intent. With these results in hand, we return to answer our research questions:

Is intent an isolable signal? Our results strongly suggest that intent is an isolable signal from support that 98 participants between experimental conditions mentioned using their intent (or vocabulary that would be considered equivalent to the present work's notion of intent) in some way to inform their decision strategy. Whether or not they employed their intent *correctly* in the present reinforcement learning task is a separate question, but because we also witnessed a significant difference in cumulative regret between strategies that followed intent vs. those that disobeyed it, we assert that intent as a signal is indeed

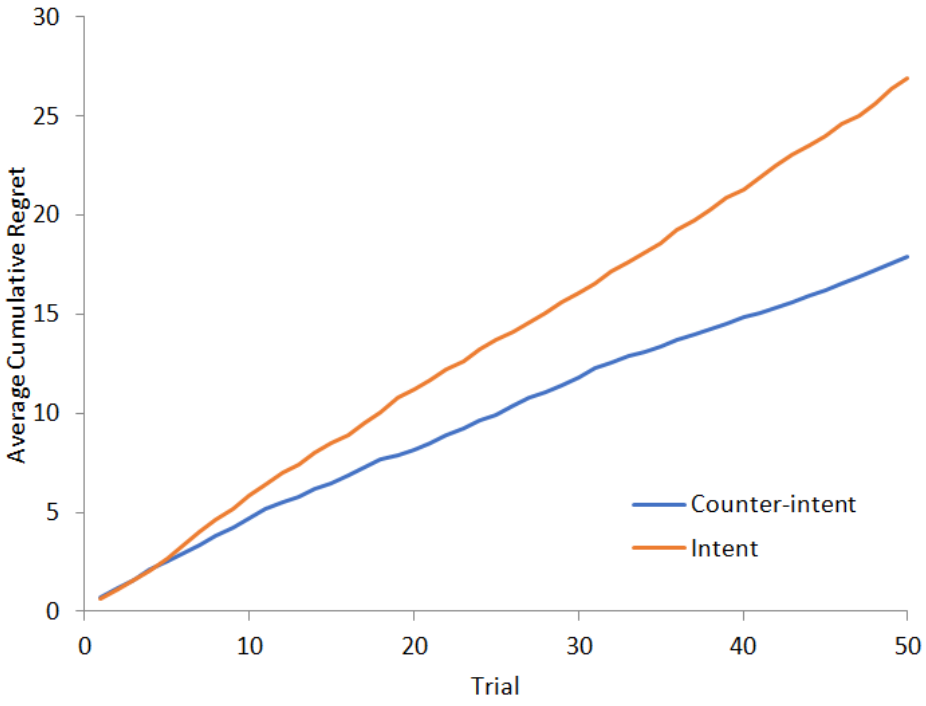

Figure 4.6: Average cumulative regret by trial between those who explicitly stated that they used the "counter-intent" ($n = 65$) vs. the "intent" ($n = 38$) strategy.

salient to humans.

Is intent reactive? Once again, our results suggest that intent, as an isolable signal, is indeed reactive to environmental factors (observed or otherwise) and is unaffected by experiential history. This point assuages the concern that, for trials $t_i \neq t_j$, UCs U, and Intent I, $P(U_{t_i}|I_{t_i}) \neq P(U_{t_j}|I_{t_j})$; in other words, if the conditional distribution over U given I changes between trials, then ISDM may not be an asymptotically optimal strategy. That said, we find no evidence for this concern, because if intent *was* sensitive experiential history, then later trials in intent- or counter-intent specific strategies should have exhibited a rise and drop in accuracy, respectively. The statistics reveal no such behavior, and can be verified graphically in Figures 4.5 and 4.6. Though it is possible that the quiz simply did not continue long enough to expose such an effect, the evidence from the 50 trials suggests that intent is reactive to the environment alone.

Is ISDM a naturally employed human counterfactual reasoning mechanism? The results

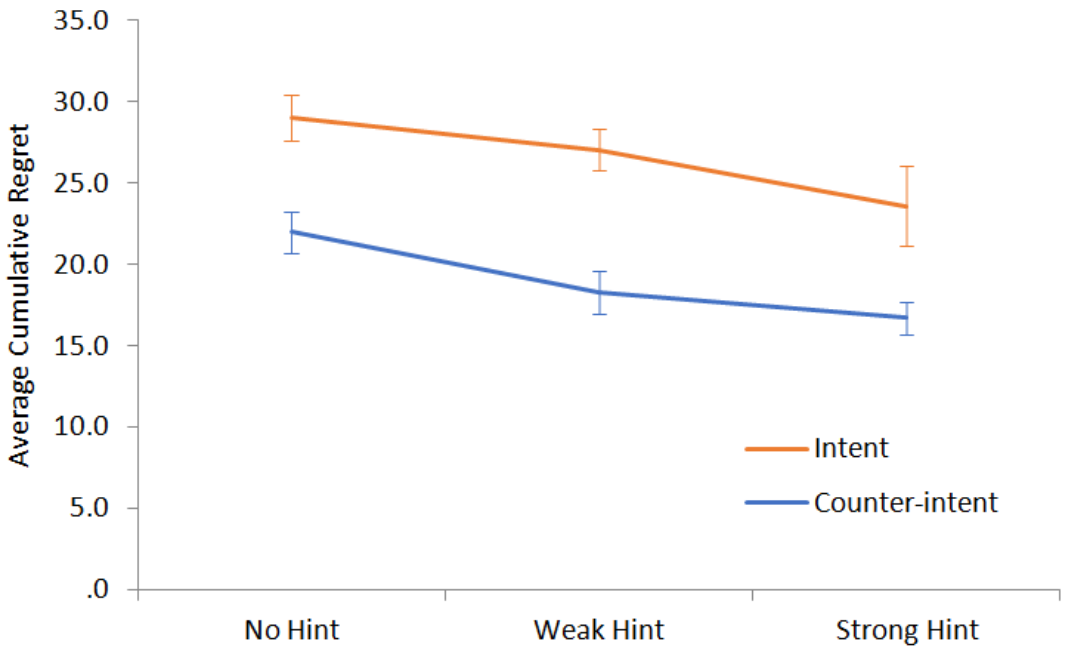

Figure 4.7: Average cumulative regret by experimental group between those who explicitly stated that they used the "counter-intent" ($n = 65$) vs. the "intent" ($n = 38$) strategy.

do not clearly answer this question, though it appears that ISDM is natural *for some* individuals but not others. In order for us to conclude that ISDM is a "naturally" employed reasoning mechanism, we should have witnessed a large proportion of individuals in the No Hint and Weak Hint conditions discover the counter-intent strategy to be the superior policy, as would be discovered by the tenets of an RDT agent. Instead, only 33% participants in the No Hint and 31% in the Weak Hint groups discovered the proper strategy, and a tepid 55% from the Strong Hint condition, in which participants were essentially told to use RDT. These results suggest that ISDM is not natural to *most* humans, but once equipped with its rules as a reasoning strategy, they can become more resilient to cognitive bias.

Apropos, the non-trivial proportion of participants who followed their intents in the No and Weak Hint conditions (27% and 33% respectively), despite it hindering their rewards, illustrates the propensity of individuals to abide by their "gut" instinct and ignore alternatives. This point not only emphasizes the prevalence of cognitive biases, but highlights their danger (and the merit of ISDM) as well.

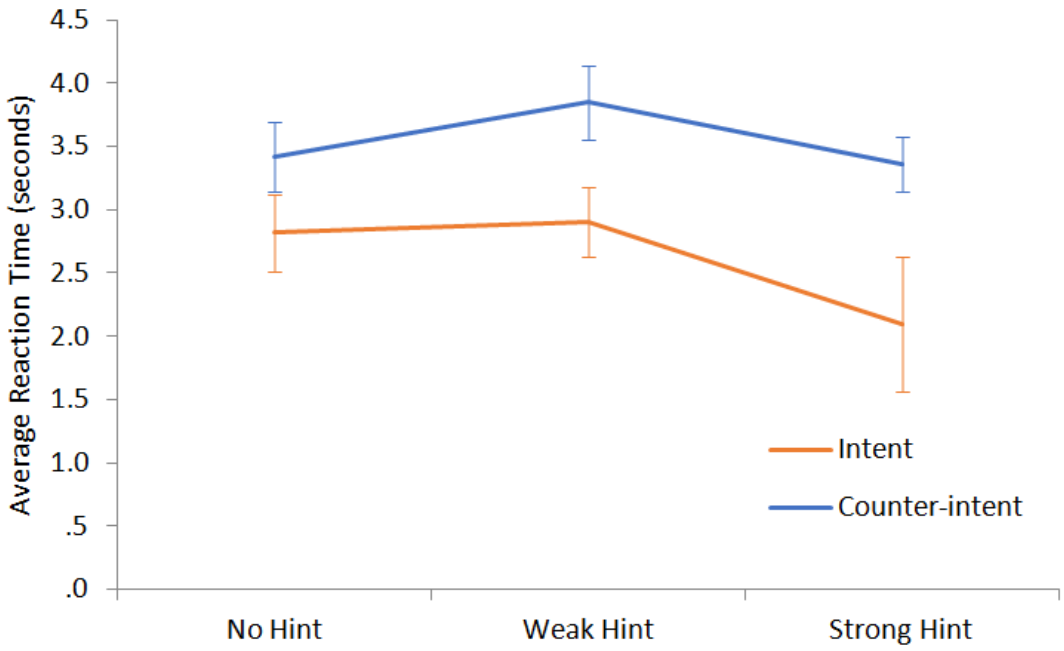

Figure 4.8: Average response times of participants employing opposite strategies of following intent vs. disobeying intent.

4.4 Conclusion

In this chapter, we verified several of the theoretical assumptions surrounding ISDM that were made in Chapter 3 and tested the strategy's efficacy in a human-subject MABUC scenario. Our results corroborated the assumptions that intent is reactive to the environment, unaffected by the agent's experiential history, and that agents can isolate the signal of their intent and employ it as a context for decision-making. Though a minority of participants were capable of discovering the ISDM strategy on their own, it appears that it is not a naturally occurring decision-making tactic in the majority of the population. That said, when instructed to use ISDM, participants were capable of using their intent as a context for their decisions, and were successfully able to improve their task performance as a result.

Acknowledgements

Special thanks go to Patricia Cheng, Elias Bareinboim, and Chéla Willey for their advice and feedback on the development of this study.

CHAPTER 5

Counterfactually Enabled Data-Fusion

As active learning agents become increasingly integrated into real-world environments, they gain new sources of information related to their tasks at hand. Not only do these agents possess the ability to interact with their environments (choosing actions, receiving feedback on the quality of their choices, and then modifying future actions accordingly), they may also observe other agents doing the same. However, with opportunities to adjust policies from sources other than personal experimentation come new challenges of "transfer" in learning. In particular, agents should be wary of how observed behavior generalizes (i.e., transfers) to them, how these observations should be combined with the agent's own experience, and how such a combination can be robustly maintained in the face of changing environmental factors.

In this chapter, we consider how data collected by an online agent under various conditions (e.g., experimental vs. non-experimental settings) can be combined to improve performance in a reinforcement learning task. This challenge is not without precedent, as recent studies have investigated dataset transportability, though in offline domains [BP16]. Others have studied scenarios in which agents learn from expert teachers in what are known as inverse reinforcement learning problems [AN04, HLM16]. These efforts can be broadly categorized as those of *data-fusion*, or the ability to take data-sets with different causal assumptions (i.e., different models of the data-generating process akin to the difference between observational settings (e.g., M_3 of Figure 2.1(a)) and experimental ones (e.g., M_{3x} of Figure 2.1(b)), and interpret them in a unified manner. Data-fusion is thus desirable because, when

Chapter 5 is an extended version of [FPB17].

possible, it allows for efficient use of existing data, thus reducing the need to collect more for a given task (which can be expensive). in the context of reinforcement learning, data-fusion can lead to an acceleration of the learning process, requiring fewer trials for an agent to converge to the optimal choice policy.

Environments for which an agent (1) observes all state variables and (2) possesses a fully specified model (in which all factors relating contexts, actions, and their associated rewards are known) are trivial from a data-fusion perspective; in such scenarios, collected data is homogeneous because all factors that may introduce bias between samples can be controlled. Conversely, in this chapter, we rejoin the focus of Chapter 3 in which the challenges that arise due to *unobserved confounders* (UCs), namely, unmeasured variables that influence an agent's natural action choice as well as the feedback from that action, can complicate both rational decision-making (as illustrated in the Greedy Casino Example 3.1.1) as well as data-fusion. Such factors are particularly subtle when left uncontrolled due to their invisible nature and potential to introduce *confounding bias* [Pea00, Chs. 3,6].

Because our agent's goal is to quickly learn an optimal policy by consolidating data collected from observing other agents and data collected through its own experience, UCs pose a fundamental challenge: the results from *seeing* another agent performing an action are not necessarily *interchangeable* with those from *doing* the action itself. As such, throughout this paper, we will differentiate three classes of data that may be employed by an autonomous agent to inform its rational decision-making:

1. **Observational data** is gathered through (1) passive examination of the actions and rewards of actors other than the agent (but for whom the agent is assumed to be exchangeable, i.e., that acting and observed agents have homogeneous intent (Def. 3.5.1)) or (2) from the agent's history in which arm choice abided by intent.

2. **Experimental data** is gathered through randomization, or from fixed policies that are not reactive to the environmental state.

3. **Counterfactual data** represents the rewards associated with actions under a particular (or "personalized") configuration of the UCs. Counterfactual data points are

generated by Intent-specific Decision-making (ISDM, Def. 3.4.2).

In the remainder of this chapter, we demonstrate how these disparate data types can be fused to facilitate learning in the Multi-Armed Bandit problem with Unobserved Confounders (MABUC), as introduced in Chapter 3. Note that the previous presentation of a MABUC problem (the Greedy Casino Example) illustrated that neither observational nor experimental reward quantities should be maximized in order to reduce u-regret (Def. 3.2.2), the regret experienced by a MABUC agent under the knowledge of each trial's UC state. Note also that observational, experimental, and counterfactual data points cannot be naively combined, treated as though they are sampled from the same distribution (see, e.g., Table 3.1(a) vs (b)). Consequently, given that we also desire to maximize the counterfactual reward distribution in a MABUC (as by the tenets of RDT, (Def. 3.4.5)), we might be tempted to discard any observational and experimental data given that they do not conform to our optimization metric, but this chapter will detail the means by which they can aid an ISDM agent nonetheless. Just as the development of ISDM in Chapter 3 allowed algorithms to converge to the optimal u-regret reduction in MABUC settings, so will this one accelerate that process through a counterfactually-enabled data-fusion technique.

Though the data-fusion problem is an ongoing exploration in the data sciences [BP16, Men14, Cou13], this chapter presents the first to study online learning techniques in MABUC settings that combine data sampled under disparate conditions. Specifically, its contributions are as follows:

1. We demonstrate how observational, experimental, and counterfactual datasets can be combined through a heuristic for MABUC agents.

2. We then develop a variant of the RDT Thompson Sampling algorithm that implements this new heuristic.

3. We run extensive simulations illustrating its faster convergence rates compared to the current state-of-the-art.

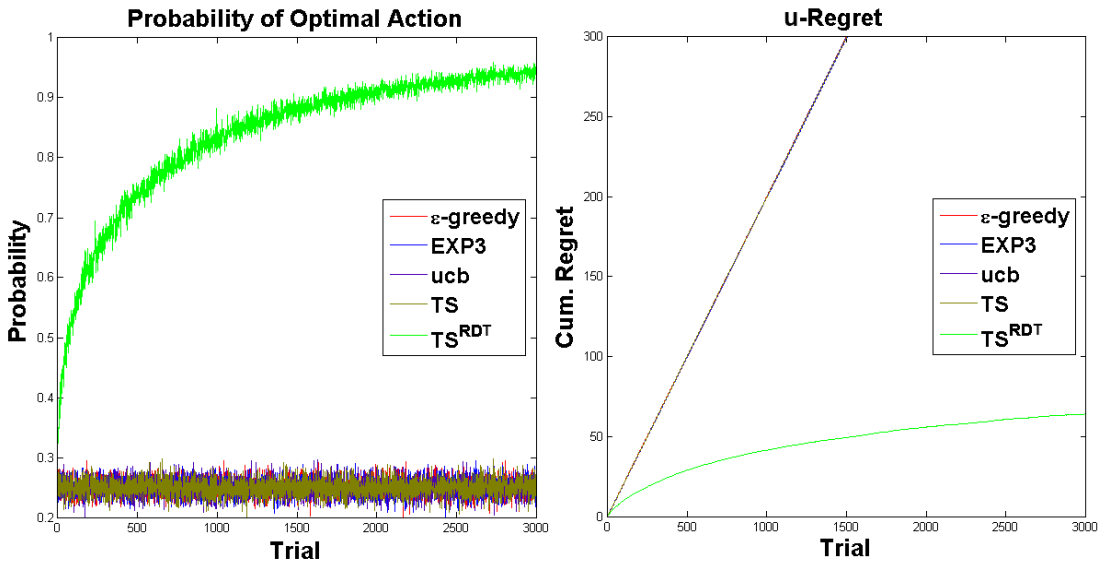

Figure 5.1: Plots of CDT MAB algorithms' performance vs. an RDT Thompson Sampling agent in the Greedier Casino scenario. Note that all algorithms but TS^{RDT} experience linear u-regret, but convergence in this 4-arm scenario takes much longer than in the 2-arm MABUC problem.

5.1 Motivating Example: The Greedier Casino

Example 5.1.1. In this section, we consider an expanded version of the Greedy Casino Example 3.1.1 from Chapter 3. Now aware that certain observant gamblers had learned to thwart their predatory payout policy (using ISDM), the executives for the Greedy Casino met to discuss alternative means of preying upon the predilections of their gamblers. Given that the vast majority of their patrons are *not* practicing ISDM, they wish to preserve the reactive slot machine payouts while making it more difficult to obtain an optimal ISDM policy. As such, they decide to expand the number of slot machine types from two to four (thus increasing the number of intent-action combinations required to learn, which would take more trials to converge to an optimal policy), and tune the payout policy to the predilections of gamblers to these expanded choices.

Apropos, in its new floor's configuration, the Greed*ier* Casino has crafted four new themed slot-machines (instead of the two used in the previous version) and wishes to make

(a)	$D=0$		$D=1$	
$P(y_1\|X,B,D)$	$B=0$	$B=1$	$B=0$	$B=1$
$X=0$	^{i}0.20	0.30	0.50	0.60
$X=1$	0.60	^{i}0.20	0.30	0.50
$X=2$	0.50	0.60	^{i}0.20	0.30
$X=3$	0.30	0.50	0.60	^{i}0.20

(b)	$P(y_1\|X)$	$P(y_1\|do(X))$
$X=0$	0.20	0.40
$X=1$	0.20	0.40
$X=2$	0.20	0.40
$X=3$	0.20	0.40

Table 5.1: (a) Payout rates decided by reactive slot machines as a function of arm choice X, sobriety D, and machine conspicuousness B. Players' natural arm choices ($f_x = B + 2D$) under D, B are indicated by superscript i. (b) Payout rates according to the observational, $P(y_1|X)$, and experimental $P(y_1|do(X))$, distributions, where $Y = y_1$ represents winning (shown in the table).

them as lucrative as possible. After running a battery of preliminary tests, the executives once more discover that the two traits from their previous iteration well predict which of the four machines that a gambler is likely to play: whether or not the machines are blinking (denoted $B \in \{0, 1\}$), and whether or not the gambler is drunk (denoted $D \in \{0, 1\}$). After consulting with their team of psychologists and statisticians, the casino learns that any arbitrary gambler's natural machine choice can be modeled by the structural equation: $X \leftarrow f_x(B, D) = B + 2 * D$ if the four machines are indexed as $X \in \{0, 1, 2, 3\}$. The casino also knows that its patrons have an equal chance of being drunk or not (i.e., $P(D = 1) = 0.5$) and decide to program their new machines to blink half of the time (i.e., $P(B = 1) = 0.5$).

Recall that a gambling law stipulates that all slot machines in the state must maintain a minimum 30% win rate. Wishing to leverage their gamblers' machine choice predilections while conscious of this law, the casino implements a reactive payout strategy for their machines, which are equipped with sensors to determine if their gambler is drunk or not (assume that the sensors are perfect at making this determination). As such, the machines are programmed with the payout distribution illustrated in Table 5.1.

After the launch of the new slot machines, some observant gamblers note that players appear to be winning only 20% of the time, and report their suspicions to the state gambling

commission. Once more, the investigator is sent to the casino to determine the merit of these complaints, and begins recruiting random gamblers from the casino floor to play at randomly selected machines, despite the players' natural predilections. Surprisingly, he finds that players in this experiment win 40% of the time, and declares that not only has the casino committed no crime, but appears to be paying its patrons generously above the law-mandated minimum. Meanwhile, the casino continues to exploit players' gambling predilections, paying them 10% less than the minimum. Still, most gamblers are unaware of being manipulated by the UCs B, D, and of the predatory payout policy that the casino has constructed around them. The collected data is summarized in Table 1b; the second column ($P(y_1|X)$) represents the observations drawn from random observations on the casino's floor while the third ($P(y_1|do(X))$) represents the randomized experiment performed by the state investigator (both assumed to boast large sample sizes).

In an attempt to find a better gambling strategy, a handful of players decides to run a battery of experiments using standard MAB algorithms (e.g., ϵ-greedy, UCB, Thomson Sampling), which, unsurprisingly, result in winnings that are no different from the state inspector's findings. However, one observant habitué, who has been recording the abysmal winnings of those playing by intent and the only incrementally better winnings of those following CDT (Def. 3.4.4) optimization algorithms, wonders if she might devise a superior strategy. As she is well-versed in the interplay of causal inference and reinforcement learning, she follows the ISDM implementation of TS^{RDT} as described in Algorithm 2 from Chapter 3, and maximizes reward by Regret Decision Theory (RDT, Def. 3.4.5). The results of her experiments, compared to those of her unenlightened peers, are depicted in Figure 5.1. Noting the differences in the payout rates between the observational, experimental, and counterfactual techniques, she realizes that the convergence of her approach is still somewhat slow (with the addition of 2 arms) and ponders how the failings of her peers could have better informed her superior strategy.

5.2 Background & Existing Techniques

The Greedier Casino Example 5.1.1 illustrates a Multi-Armed Bandit problem with Unobserved Confounders (MABUC), akin to that of the Greedy Casino Example 3.1.1. However, in this new problem, our learning agent gains access to additional side-information before beginning play (i.e., before beginning its first trial). In particular, our agent may possess heterogeneous (1) Observational Data D_{obs}, eliciting a reward distribution like $P(y_1|X)$ in Table 5.1(b), and (2) Experimental Data D_{exp}, eliciting a reward distribution like $P(y_1|do(X))$ in Table 5.1(b). Plainly, in a MABUC setting, these are heterogeneous quantities because $P(y_1|X) \neq P(y_1|do(X)) \ \forall \ X$ (as exemplified in Table 5.1(b)), and so data from each of these collection techniques (i.e., random sampling from the observational case and random experimentation from the experimental) cannot be naïvely combined. Furthermore, as demonstrated by the Greedy Casino example and successful application of RDT to reach an optimal choice policy, neither observational nor experimental optimization quantities are optimal in a MABUC setting.

As such, we might be tempted to ignore these datasets in pursuit of the counterfactual optimization quantity demanded by RDT. Indeed, the goal of RDT to estimate intent-specific rewards to each final arm choice will not be different in the present setting, but we should avoid haste to discard observational and experimental data that might aid in that process. To wit, [Pea00, Ch. 7] demonstrated that the counterfactual ETT can be estimated without a fully-specified model from these two datasets, but only when the treatment / action choice is binary. To be specific, estimating $P(Y_x|x')$ for binary $x, x' \in X$ can be accomplished via the decomposition:

$$\begin{aligned} P(Y_x) &= P(Y_x|x)P(x) + P(Y_x|x')P(x') \\ &= P(Y|x)P(x) + P(Y_x|x')P(x') \\ P(Y_x|x') &= \frac{P(Y_x) - P(Y|x)P(x)}{P(x')} \end{aligned} \tag{5.1}$$

The above tactic exploits the fact that $P(Y_x|x) = P(Y|x)$ by the *consistency axiom* (Def. 2.4.3), which translates the counterfactual "$P(Y)$ had X been x, given that it was (in reality)

x" to the observational equivalent, since the observed x and hypothesized antecedent x are the same. Thus, it is possible to estimate the counterfactual $P(Y_x|x')$ without a fully-specified model from experimental $P(Y_x)$ and observational $P(Y|x)$ quantities in the binary treatment case because there is 1 unknown ($P(Y_x|x')$). However, as soon as the treatment choices expand to 3 or more options, this technique does not scale; consider the same decomposition for $x_0, x_1, x_2 \in X$ and a desire to estimate (without loss of generality) $P(Y_{x_0}|x_1)$ with only observational and experimental data:

$$\begin{aligned} P(Y_{x_0}) &= P(Y_{x_0}|x_0)P(x_0) + P(Y_{x_0}|x_1)P(x_1) + P(Y_{x_0}|x_2)P(x_2) \\ &= P(Y|x_0)P(x_0) + P(Y_{x_0}|x_1)P(x_1) + P(Y_{x_0}|x_2)P(x_2) \\ P(Y_{x_0}|x_1) &= \frac{P(Y_{x_0}) - P(Y|x_0)P(x_0) - P(Y_{x_0}|x_2)P(x_2)}{P(x_1)} \end{aligned} \tag{5.2}$$

Here, we have not 1 but 2 unknown counterfactual quantities, (1) the query $P(Y_{x_0}|x_1)$ on the LHS and (2) $P(Y_{x_0}|x_2)$ on the RHS. While traditional, offline causal inference would conclude that $P(Y_{x_0}|x_1)$ is thus unidentifiable (in the absence of a fully-specified model), the ability to practice ISDM in the online decision-making domain allows us to empirically sample counterfactual data-points (as a consequence of Theorem 3.4.1) and surmount this problem for non-binary arm choices. That said, with the additional constraint that online domains like reinforcement learning value not only convergence to an optimal choice policy (which can be translated in the MABUC to estimating $P(Y_x|x')\ \forall\ x, x' \in X$), but also the speed with which said convergence takes place, our strategies that attempt to leverage any observational and experimental data must do so with regards to finite-sample concerns. In the following section, we will develop such a strategy, but first, must model the scenario at hand.

Apropos, we will extend the model associated with our prototypical MABUC from Chapter 3 to accommodate side-information in the form of observational and experimental data. In this updated Structural Decision Model (SDM, Def. 3.3.1), we explicitly indicate that learning agents may possess, and treat as distinct samples, information regarding obs. and exp. rewards. Figure 5.2 demonstrates this modified SDM, with the side-information contributing to the agent's experiential history. As a reminder, the remaining components of

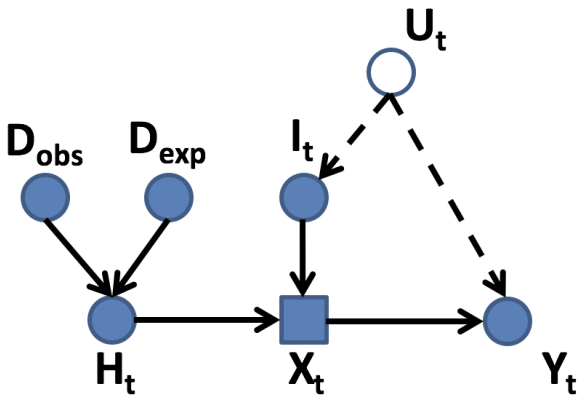

Figure 5.2: SDM of a prototypical MABUC instance with side-information in the form of observational D_{obs} and experimental D_{exp} data. This information is incorporated into the ISDM learning agent's experiential history H_t and used to better inform its decision-making.

this SDM are:

1. *Unobserved confounders:* U_t represents the unobserved confounders instantiated to $U_t = u_t$ at trial t.

2. *Intent:* $I_t \in \{x_1, ..., x_k\}$ represents the agent's intended arm choice at round t (prior to its final choice, X_t) such that $I_t = f_i(pa_{x_t}, u_t)$.

3. *Decision:* $X_t \in \{x_1, ..., x_k\}$ denotes a decision variable (Def. 3.3.2), which indicates a rational choice made as a function of the agent's history and current intent, $f_\pi(h_t, i_t)$.

4. *History:* $H_t = \{Z_0, X_0, Y_0, ...Z_{t-1}, X_{t-1}, Y_{t-1}\}$ denotes the agent's recorded history of contexts (including intent), final arm choices, and rewards at each trial up to t. In the present setting, the History also contains qualitatively separate data points for those originating from observations D_{obs} and experiments D_{exp}.

5. *Reward:* $Y_t \in \{0, 1\}$ represents the Bernoulli reward (0 for losing, 1 for winning) from choosing arm x_t under UC state u_t as decided by $y_t = f_y(x_t, u_t)$.

In the next section, we will formalize how the agent can leverage this extra data to accelerate learning in a MABUC instance like the Greedier Casino.

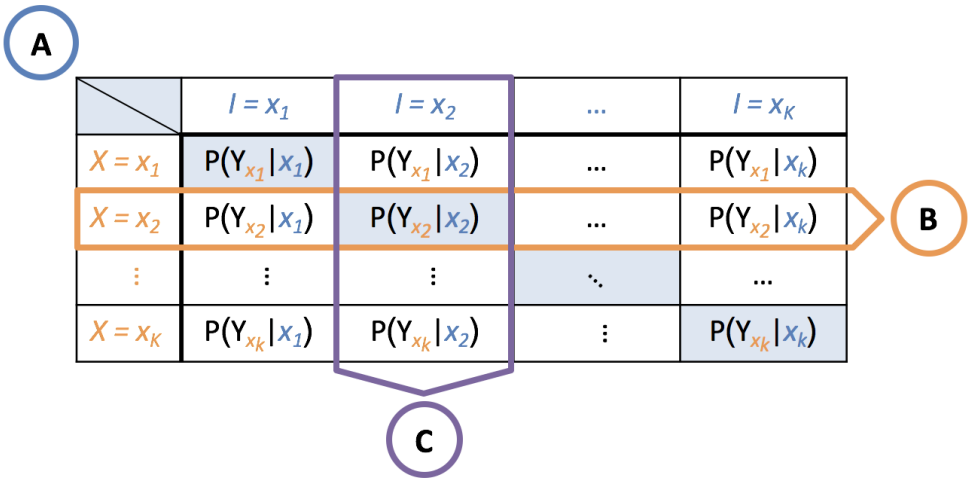

Figure 5.3: An ISDM agent's counterfactual history in which rewards are recorded by intent-context I (columns) and final-arm-choice X for an arbitrary K-armed MABUC instance (replicated from Table 3.6) but with illustrations of data-fusion Strategies A (blue, along diagonal), B (orange, across intents), and C (purple, across arms).

5.3 Counterfactual Data-fusion for Online Reinforcement Tasks

Suppose our agent assumes the role of the observant gambler in the Greedier Casino Example 5.1.1 and possesses (1) observations of arm choices and payouts from players gambling by intent in the casino (with whom the agent is assumed to have homogeneous intent (Def. 3.5.1)), (2) the randomized experimental results from the state investigator or the CDT gamblers, and (3) the knowledge to use ISDM for by optimizing the counterfactual reward advocated by RDT. In other words, the agent begins the MABUC problem with large samples of observations ($P(Y|X)$) and experimental results ($P(Y|do(X))$), and will maximize the counterfactual RDT ($P(Y_{X=a} = 1|X = i)$) because it recognizes the presence of UCs (viz. $P(Y|X) \neq P(Y|do(X))$. The agent now seeks to employ the obs. and exp. data to speed this optimization process. In this section, we will detail several, separate approaches that attempt this goal, and finally, how they may all be combined to form a complete data-fusion algorithm.

Strategy A: Observational-Counterfactual Consistency

In the previous section, we mentioned that the consistency axiom (Def. 2.4.3) can be used to equate counterfactual quantities of the format $P(Y_x|x')$ (where $x = x'$) with observational quantities $P(Y|x)$. In other words, the counterfactual expression measuring the "probability of Y had X been x given that X was *observed to be* x in reality" is not a counterfactual at all – the antecedent and the observation agree in this statement, and so the observed response of Y to the observed $X = x$ will not differ when we hypothesize about what $X = x$ (the same value) would have done to Y in that scenario. As such, for ISDM agents maximizing by the RDT, we can immediately employ our observational data D_{obs} to provide information about all intent-action payouts where the intent and action agree.

Recall the ISDM agent's counterfactual reward history given in Chapter 3, and replicated herein (Figure 5.3) with some additional highlights for the coming data-fusion techniques. Note that the diagonal (i.e., cells shaded in blue, and tagged by the circled A) encodes all intent-action payouts for which the intent and action are the same value. In the presence of observational data, our agent may immediately populate this diagonal and obtain the true values for $P(Y_x|x)$. From this simple incorporation, the agent reduces a MABUC problem with K^2 separate intent-specific arm reward parameters to learn down to $(K-1)^2$.

Strategy B: Cross-Intent (XInt) Information Leakage

With the remaining $P(Y_x|x')$ $x \neq x'$ in Figure 5.3 to learn via ISDM, and because a MABUC is an online learning problem in which each of these cells must be explored sufficiency, the next two strategies exploit the obs. and exp. datasets' relationship to the counterfactual targets while managing the uncertainty implicit in a MAB learning scenario.

As such, consider Eq. 5.1 once again (which decomposes the experimental $P(Y_x)$ into constituent observational and counterfactual terms) but for the general, non-binary treatment case:

$$P(Y_x) = \sum_{x' \in X} P(Y_x|x')P(x') \tag{5.3}$$

Now, consider a single cell in our counterfactual experiential history Figure 5.3, say $P(Y_{x_r}|x_w)$, which we can solve and rewrite as:

$$P(Y_{x_r}) = P(Y_{x_r}|x_1)P(x_1) + ... + P(Y_{x_r}|x_w)P(x_w) + ... + P(Y_{x_r}|x_K)P(x_K) \tag{5.4}$$

$$P(Y_{x_r}|x_w) = \frac{P(Y_{x_r}) - P(Y_{x_r}|x_1)P(x_1) - ... - P(Y_{x_r}|x_K)P(x_K)}{P(x_w)} \tag{5.5}$$

$$P_{XInt}(Y_{x_r}|x_w) = \frac{P(Y_{x_r}) - \sum_{i \neq w}^{K} P(Y_{x_r}|x_i)P(x_i)}{P(x_w)} \tag{5.6}$$

Here, Eq. 5.6 provides a systematic way of learning about arm x_r payouts across intent conditions, which is desirable because an arm pulled under one intent condition now provides knowledge about the payouts of that arm under other intent conditions. This can be depicted graphically, as shown by the flow across an example row B in Figure 5.3 – information about Y_{x_r} flows from intent conditions $x_i \neq x_w$ to intent x_w (what has been referred to as a form of *information leakage*, wherein information about rewards associated with arms in one condition inform those in another [SSD17]).

Strategy C: Cross-Arm (XArm) Information Leakage

Consider any three arms, x_r, x_s, x_w such that $r \notin \{s, w\}$ and assume we are interested in estimating the value of $P(Y_{x_r}|x_w)$ (our query, for short). Considering again the equations induced by Eq. (5.3), we have,

$$P(Y_{x_r}) = \sum_{i}^{K} P(Y_{x_r}|x_i)P(x_i) \tag{5.7}$$

$$P(Y_{x_s}) = \sum_{i}^{K} P(Y_{x_s}|x_i)P(x_i) \tag{5.8}$$

Note that each of Eqs. (5.7, 5.8) share the same intent priors on our query intent $P(x_w)$, so we can solve for $P(x_w)$ in both equations using simple algebra, which yields,

$$\begin{aligned} P(x_w) &= \frac{P(Y_{x_r}) - \sum_{i \neq w}^{K} P(Y_{x_r}|x_i)P(x_i)}{P(Y_{x_r}|x_w)} \\ &= \frac{P(Y_{x_s}) - \sum_{i \neq w}^{K} P(Y_{x_s}|x_i)P(x_i)}{P(Y_{x_s}|x_w)} \end{aligned} \tag{5.9}$$

Using Eq. (5.9) and solving for the query in terms of our paired arm x_s, $\forall\ r \neq s$ we have

$$P(Y_{x_r}|x_w) = \frac{\left[P(Y_{x_r}) - \sum_{i \neq w}^{K} P(Y_{x_r}|x_i)P(x_i)\right] P(Y_{x_s}|x_w)}{P(Y_{x_s}) - \sum_{i \neq w}^{K} P(Y_{x_s}|x_i)P(x_i)} \tag{5.10}$$

Eq. (5.10) illustrates that any non-diagonal cell from the table in Figure 5.3 can be estimated through pairwise arm comparisons with the same intent. Put differently, Eq. (5.10) allows our agent to estimate $P(Y_{x_r}|x_w)$ from samples in which any arm $x_s \neq x_r$ was pulled under the same intent x_w.

In practice, the online nature of the MABUC learning problem can make some of these pairwise computations noisy due to sampling variability when x_r is an infrequently explored arm. To obtain a more robust estimate of the target quantity, this pairwise comparison can be repeated between the query arm and all other arms with the same intent, and then pooled together. This can be seen as information about $Y_{x_r}|x_w$ flowing from arm $x_s \neq x_r$ to x_r (under intent x_w) – for example, column C in Figure 5.3.

One such pooling strategy is to take the *inverse-variance-weighted* average.[1] Formally, we can consider a function $P(Y_{x_r}|x_w) = h_{XArm}(x_r, x_w, x_s)$ such that h_{XArm} performs the empirical evaluation of the RHS of Eq. (5.10). Additionally, let $\sigma^2_{x,i} = Var_{samp}[Y_x|i]$ indicate the empirical payout variance for each arm-intent condition (as from the reward successes and failures captured by the agent in Table 5.3). To estimate our query from all other arms in the same intent through inverse-variance weighting, we have our now complete, third strategy:

$$P_{XArm}(Y_{x_r}|x_w) = \frac{\sum_{i \neq r}^{K} h_{XArm}(x_r, x_w, x_i)/\sigma^2_{x_i,x_w}}{\sum_{i \neq r}^{K} 1/\sigma^2_{x_i,x_w}} \tag{5.11}$$

[1]This strategy follows from the fact that we have Bernoulli rewards for each arm-intent condition, and as the number of samples increases for these distributions, the variance diminishes, meaning that arm-intent conditions with smaller variances are more reliable than those with larger ones.

The Combined Approach

The payout estimates for an ISDM algorithm maximizing rewards via RDT can be estimated from three different sources: (1) $P_{samp}(Y_{x_r}|x_w)$, the *sample* estimates collected by the agent during the execution of the algorithm. (2) $P_{XInt}(Y_{x_r}|x_w)$, the computed estimate using *cross-intent* learning. (3) $P_{XArm}(Y_{x_r}|x_w)$, the computed estimate using *cross-arm* learning. Naturally, these three quantities can be combined to obtain a more robust and stable estimate to the target query.

Once again, we employ an inverse-variance weighting scheme so as to leverage these three estimators, and so we must formulate a metric for the payout variance associated with each strategy's computed estimate. To do so, we define an average variance for each strategy, which is the average over each sample estimate's variance (i.e., $\sigma^2_{x,i}$) used in the computation. Specifically, for the cross-arm approach (Eq. 5.11), we have two summations over sample payout estimates $P(Y_{x_r}|x_i)$, $P(Y_{x_s}|x_i)$ $\forall i \neq w$ which involve $2(K-1)$ terms, plus the numerator's $P(Y_{x_s}|x_w)$, giving us a total of $2(K-1)+1 = 2K-1$ variances to average. The same is true for the cross-intent apprach (Eq. 5.6), which involves $K-1$ sample variances to average. When estimating $P(Y_{x_r}|x_w)$, we can write the corresponding variances:

$$\sigma^2_{XArm} = \frac{1}{2K-1}\Big[\big[\sum_{i\neq w}^{K} \sigma^2_{x_r,x_i}\big] + \big[\sum_{i\neq w}^{K} \sigma^2_{x_s,x_i}\big] + \sigma^2_{x_s,x_w}\Big]$$

$$\sigma^2_{XInt} = \frac{1}{K-1}\sum_{i\neq w}^{K} \sigma^2_{x_r,x_i}$$

Finally, to estimate $P(Y_{x_r}|x_w)$ using our combined approach, we have:

$$\alpha = P_{samp}[Y_{x_r}|x_w]/\sigma^2_{x_r,x_w} + P_{XInt}[Y_{x_r}|x_w]/\sigma^2_{XInt} + P_{XArm}[Y_{x_r}|x_w]/\sigma^2_{XArm}$$

$$\beta = 1/\sigma^2_{x_r,x_w} + 1/\sigma^2_{XInt} + 1/\sigma^2_{XArm}$$

$$P_{combo}[Y_{x_r}|x_w] = \frac{\alpha}{\beta} \tag{5.12}$$

To visualize the data-fusion process discussed here, consider the diagram in Figure 5.4.

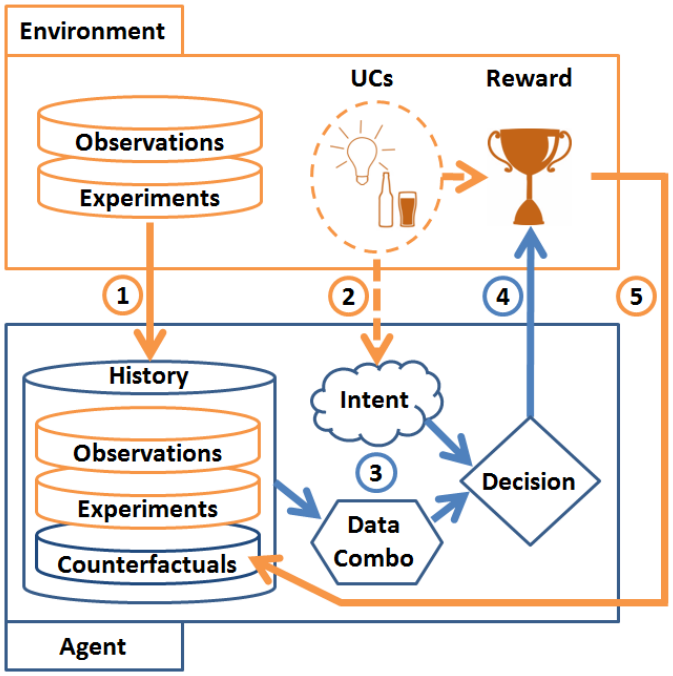

Figure 5.4: Illustrated data-fusion process.

1. In this scenario, we consider that our agent has collected large samples of experimental and observational data from its environment (e.g., in the Greedier Casino, the agent might observe other gamblers to comprise its observational data and incorporate experimental findings from the state investigator's report).

2. Unobserved confounders are realized in the environment, though their labels and values are unknown to the agent.

3. From these UCs and any other observed features in the environment, the agent develops its intent. With its intent known, the agent combines the data in its history (in this work, by the prescription of the Combined Strategy above) to better inform its decision-making.

4. Based on its intent and combined history, the agent commits to a final action choice.

5. The action's response in the environment (i.e., its reward) is observed, and the collected data point is added to the agent's counterfactual dataset (as a consequence of Theorem 3.4.1).

5.4 MABUC (with Side Information) Simulations & Results

In this section, we validate the efficacy of the strategies discussed in the previous section through simulations. To make a fair comparison to previous ISDM bandit players (as demonstrated in Chapter 3), we will follow the first implementation of an RDT reward maximizing algorithm that used Thompson Sampling (TS) as its basis, embedding the strategies described in the previous section within a TS player called (TS^{RDC*}).

5.4.1 Simulation Interpretation

The interpretation for the MABUC simulation with Side Information is identical to the one presented in Chapter 3, with the same assumption of homogeneous intent (Def. 3.5.1) between observed actors; the sole difference is that, in the current scenario, the reasoning agent may have access to observational and experimental data before play. The same distinctions between agent and actor apply, and are depicted in Figure 5.5.

5.4.2 Simulation Procedure & Results

The algorithm for (1) the MABUC scenario with side-information and (2) the TS^{RDC*} ISDM data-fusion player are described in Algorithm 3 and 4, respectively.[2]

In brief, TS^{RDC*} agents perform the following at each round: (1) Observe the intent i_t from the current round's realization of UCs, u_t. (2) Sample $\hat{P}_{samp}(Y_{x_r}|i_t)$ from each arm's (x_r) corresponding intent-specific beta distribution $\beta(s_{x_r,i_t}, f_{x_r,i_t})$[4] in which s_{x_r,i_t} is the number of successes (wins) and f_{x_r,i_t} is the number of failures (losses). (3) Compute each arm's i_t-specific score using the combined datasets via the Combined Strategy (Eq. 5.12). (4)

[2]All simulation source code for Chapter 5 can be found at:
https://github.com/Forns/ucla-forns/tree/master/projects/dissertation/ch5.

[3]Different agents will employ the available datasets D_{obs} and D_{exp} according to their policies, with some of the more naïve variants (like CDT TS) ignoring them entirely.

[4]The parameters for these distributions are decided by the agent's history (see Figure 5.3), including contributions from observational data for cells in which action and intent agree.

Algorithm 3 MABUC Simulation (with Side Information)

procedure $MABUC - Sim+(\mathrm{T},\ D_{obs},\ D_{exp})$

$R^u \leftarrow 0$ (initialize cum. u-regret)

$H \leftarrow \{\}$ (initialize history)

for $t = [1, ..., T]$ **do**

$u_t \leftarrow f_u(...)$ (realize environmental factors for trial)

$i_t \leftarrow f_x(u_t)$ (intent is initialized for trial)

$x_t \leftarrow f_\Pi(i_t, h_t, D_{obs}, D_{exp})$ [3](policy selects final decision)

$y_t \leftarrow f_y(x_t, u_t)$ (reward is observed from chosen arm)

$H \leftarrow H \cup \{i_t, x_t, y_t\}$ (history is updated)

$r_t^u \leftarrow P(Y_{x_t^*}|u_t) - y_t$ (u-regret is logged)

$R^u \leftarrow R^u + r_t^u$ (cum. u-regret is updated)

Algorithm 4 RDT Thompson Sampling (with Side Information)

procedure $TS^{RDT*}(i_t, h_t, D_{obs},\ D_{exp})$

$s_t \leftarrow [\#Y_{x_0} = 1|i_t, ..., \#Y_{x_k} = 1|i_t]_{h_t}$ (count number of successes for each intent-arm)

$f_t \leftarrow [\#Y_{x_0} = 0|i_t, ..., \#Y_{x_k} = 0|i_t]_{h_t}$ (count number of failures for each intent-arm)

$A_t \leftarrow [\beta(s_t[1], f_t[1]), ..., \beta(s_t[k], f_t[k])]$ (sample from beta-dists. of each intent-arm)

$A_t* \leftarrow f_{combo}(A_t, P_{XInt}, P_{XArm}, P_{Samp})$ (data-fusion weighting)

$x_t \leftarrow \mathrm{argmax}_{x \in [1,k]}\, A_t*$ (choose max)

return x_t

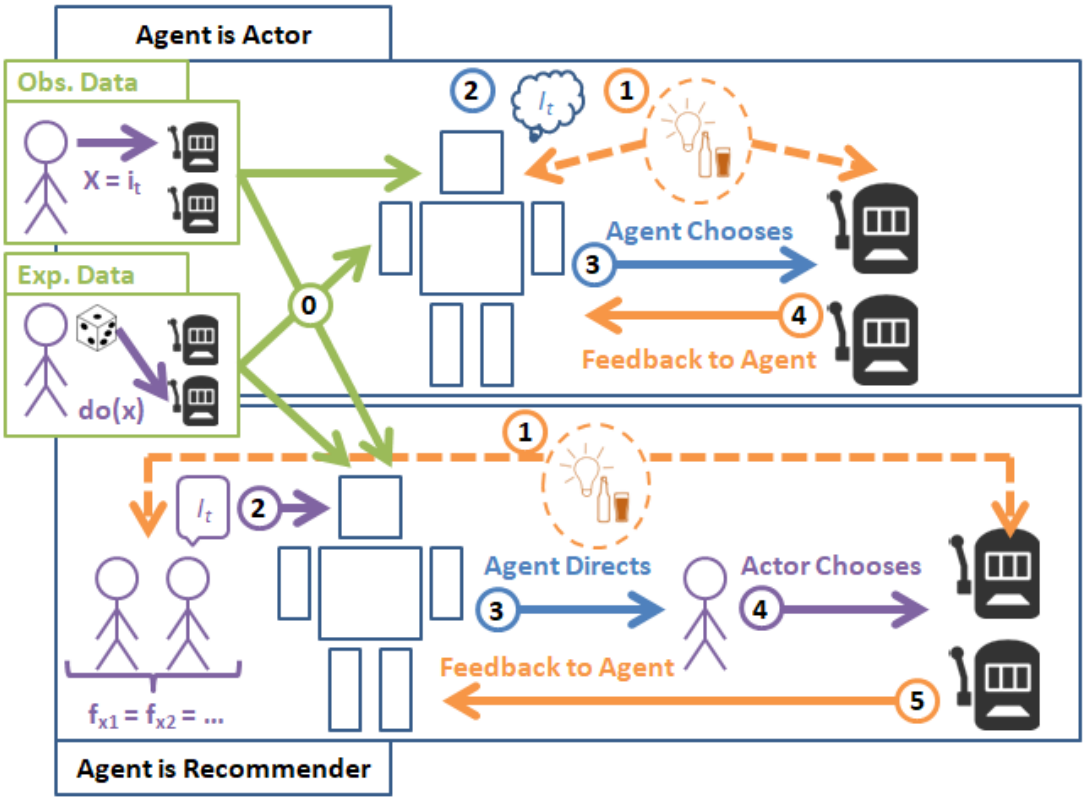

Figure 5.5: Interpretations of the MABUC simulations that employ the same SDM, but may have separate agents and actors. Pictured: [top] the agent (blue) and the actor (also blue) are the same entity; [bottom] the agent (blue) and actor(s) (purple) are distinct entities. In both panels, the environment's states and actions are drawn in orange, and side information available to the agent is drawn in green.

Choose the arm, x_a, with the highest score computed in previous step. (5) Observe result (win / loss) and update $\hat{P}_{samp}(Y_{x_a}|i_t)$.

Procedure. Simulations were performed on the 4-arm MABUC problem, with results averaged across $N = 1000$ Monte Carlo repetitions, each $T = 3000$ rounds in duration. To illustrate the robustness of each proposed strategy, we performed simulations spanning across a wide range of payout parameterizations (see Appendix XXX for a complete report of experimental results).

Compared Algorithms. Each simulation compares the performance of four variants

Algorithm	Cf. Data	Obs. Data	Exp. Data
TS^{RDT*}	✓	✓	✓
TS^{RDT+}	✓	✓	
TS^{RDT}	✓		
TS			✓

Table 5.2: Data-sets employed by the compared TS variants.

of Thompson Sampling, described below and with the data-sets employed by each indicated in Table 5.2:

1. TS is the traditional Thompson Sampling bandit algorithm that attempts to maximize the interventional quantity $P(y|do(x))$, and does not condition on intent.

2. TS^{RDT} is the ISDM TS player that uses RDT, but employs no additional observational or experimental data in its play.

3. TS^{RDT+} is TS^{RDT} that also incorporates observational data via Strategy A, but does not incorporate experimental data nor exploit the relationship between data types via the combined approach.

4. TS^{RDT*} follows Algorithm 4 and uses the data-fusion strategy described in the previous section.

Evaluation. Each algorithm's performance is evaluated using two standard metrics: (1) the probability of optimal arm choice under the state of each round's confounders $U_t = u_t$ and (2) cumulative u-regret (Def. 3.2.2), both as a function of t averaged across all N Monte Carlo simulations.

Experiment 1: "Greedier Casino." The Greedier Casino parameterization, as described in Table 5.1, exemplifies the scenario where all arms are both observationally equivalent ($P(Y|x) = P(Y|x'), \forall x, x'$) and experimentally equivalent ($P(Y|do(x)) = P(Y|do(x')), \forall x, x'$), but distinguishable within intent conditions ($P(Y_x|x')$). In this reward parameterization,

TS^{RDC*} experienced significantly less regret ($M = 42.23$) than its chief competitor, TS^{RDC+}, ($M = 65.04$), $t(1998) = 13.25, p < .001$.

(a)	$D=0$		$D=1$	
$P(y_1\|X,B,D)$	$B=0$	$B=1$	$B=0$	$B=1$
$X=0$	$^i0.90$	0.20	0.45	0.45
$X=1$	0.30	$^i0.40$	0.50	0.40
$X=2$	0.10	0.35	$^i0.60$	0.35
$X=3$	0.10	0.10	0.30	$^i0.60$

(b)	$P(y_1\|X)$	$P(y_1\|do(X))$
$X=0$	0.90	0.50
$X=1$	0.40	0.40
$X=2$	0.60	0.35
$X=3$	0.60	0.20

Table 5.3: (a) Payout rates decided by reactive slot machines as a function of arm choice X, sobriety D, and machine conspicuousness B. Players' natural arm choices under D, B are indicated by superscript i. (b) Payout rates according to the observational, $P(y_1|X)$, and experimental $P(y_1|do(X))$, distributions, where $Y = y_1$ represents winning (shown in the table).

Experiment 2: "Paradoxical Switching." The Paradoxical Switching parameterization (see Table 5.3 for parameters) exemplifies a curious scenario wherein $P(Y_{x_1}) = 0.5 > P(Y_{x'}), \forall x' \neq x_1$, but for which x_1 is the optimal arm choice in only one intent condition ($I = x_1$). Agents unempowered by RDT will face a paradox in that the arm with the highest experimental payout is not always optimal. Again, TS^{RDT*} experienced significantly less regret ($M = 36.91$) than its chief competitor, TS^{RDT+}, ($M = 64.70$), $t(1998) = 22.43, p < .001$.

The accelerated learning enjoyed by RDT^* is not localized to these parameter choices alone. In Appendix C, we show that TS^{RDT*} consistently experiences significantly less regret than its competitors across a wide range of reward parameterizations.

5.5 Conclusion

In this chapter, we examined the Greedier Casino scenario, a more difficult version of the Greedy Casino MABUC scenario from Chapter 3 in which our agent was tasked with learning the optimal policy between four arms in the presence of UCs. Unlike in the previous

scenario, the agent also possessed side-information at the start of the "game" in the form of observational and experimental arm-specific rewards. Due to the presence of UCs in the system, these obs. and exp. datasets are not exchangeable, and (due to the tenets of RDT established in Chapter 3) neither represent the proper counterfactual maximization target in the MABUC scenario. Although tempting to discard these datasets as useless for a MABUC learning task, we developed a strategy that employs obs. and exp. data in pursuit of learning the counterfactual intent-specific rewards. Once again, we find that ISDM strategies are not only superior to the traditional, experimental maximization approaches, but can be accelerated by the incorporation of obs. and exp. side-information. In other words, while the development of ISDM in Chapter 3 allowed our learning agents to experience sub-linear u-regret in MABUC scenarios, the present chapter detailed a strategy that can accelerate its convergence.

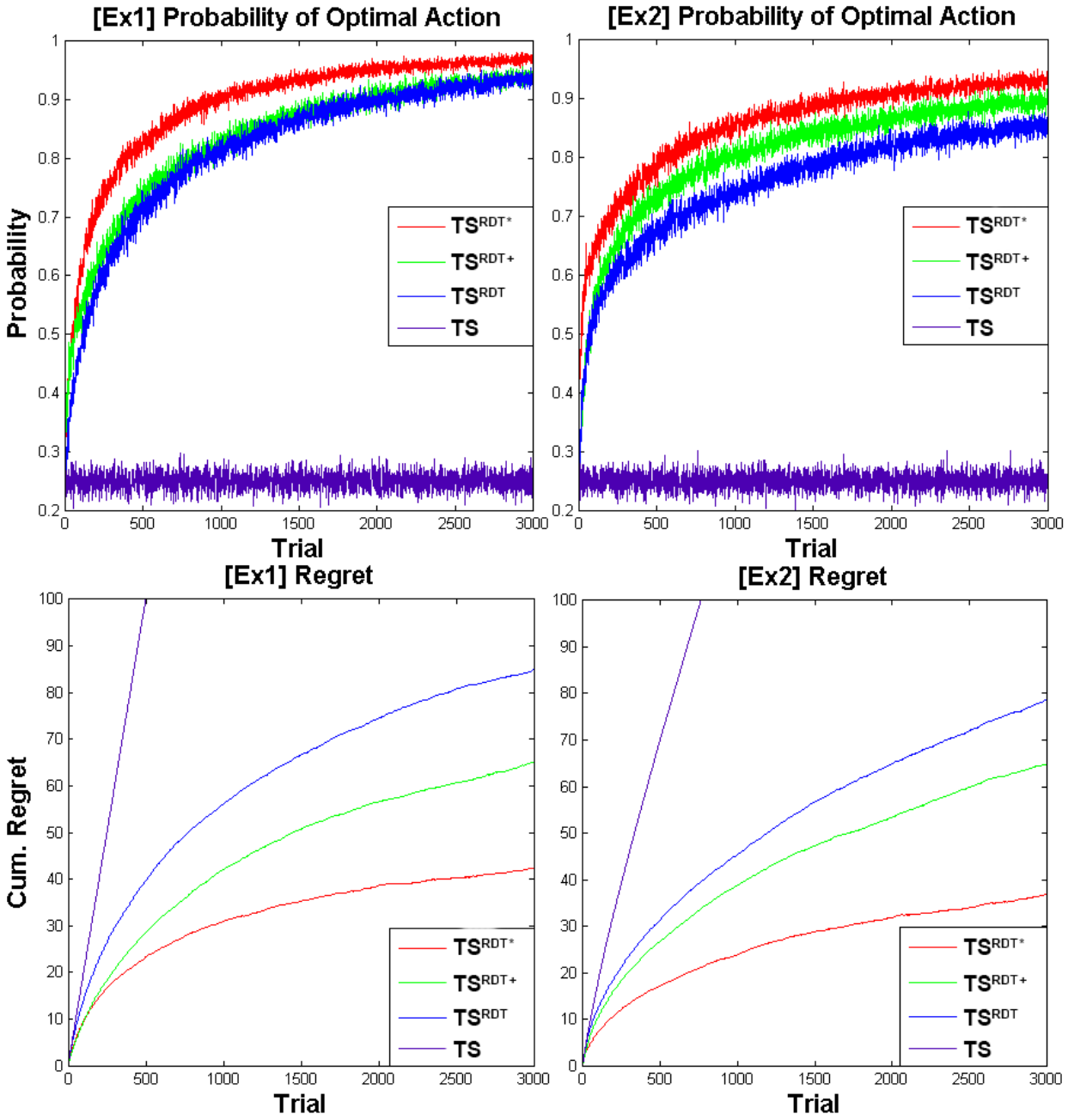

Figure 5.6: Plots of TS variant performances in the Greedier Casino [Ex1] and Paradoxical Switching [Ex2] scenarios. Optimal actions are considered those that minimize u-Regret.

CHAPTER 6

Heterogeneous Intent-Specific Decision-Making

In the previous chapters, we have detailed the tools surrounding, and examples involving, agents and actors practicing intent-specific decision-making (ISDM, Def. 3.4.2) under the key assumption that all agents are exchangeable. We formalized this quality by definition of homogeneous intent (Def. 3.5.1), which supposes that all intent-generating *functions* in the environment are the same, though the realization of a given instance's intent may change due to changing factors in the environment. For instance, in the Greedy Casino Example 3.1.1, though confounded agents may experience different intents from trial to trial (due to changes in the environment, such as when the machine lights stop blinking), the underlying function that dictates the intended response to that environment is considered equivalent between agents. This assumption was key to the different interpretations of the Multi-Armed Bandit problems with Unobserved Confounders (MABUC) scenarios depicted in Chapters 3 and 5, and in particular, was required for the data-fusion incorporation of observational data from actors other than the agent.

In this chapter, we relax this assumption and accept that agents within MABUC environments may experience heterogeneous intent (to be formalized shortly), meaning that the functions deciding their observational arm choices (and thus their intents) may be different. While this loosened restriction re-opens issues of data-fusion (in that one agent's observational and intent-specific rewards are no longer exchangeable), we also demonstrate that diversity of intent functions can be instrumental as a lens with higher sensitivity to the state of any unobserved confounders (UC) in the environment. This goal is not without

Chapter 6 is an extended version of [FBP].

precedent; in the introduction, we mention a real-world case depicting decision-makers with heterogeneous intents: in the recidivism example, comparisons in offenders' propensities to recommit crimes were measured between rulings of bail vs. incarceration from what were labeled as *strict* vs. *lenient* judges in the court system [KLL17]. We also noted studies finding that implicit biases (related to a patient's race or socio-economic status) influenced certain physicians' treatment decisions and patient interactions, but not others [Els99, GCP07]. Plainly, in the general case, different *classes* of decision-makers exist in the same population of actors, and with diversity of intent functions come new challenges and opportunities for ISDM.

We begin this endeavor with a motivating example that demonstrates both potential and pitfall with regards to managing agents with heterogeneous intents. We first use this example to motivate applications of heterogeneous ISDM in a dynamic experiment for online learners. Later, we discuss an offline data collection approach whereby heterogeneous intents may give rise to a new, more informative, experimental design that improves upon the age-old Randomized Clinical Trial (RCT); herein, we combine the offline data collected in an RCT with the online intents of agents in the target environment of the RCT. Finally, we formalize heterogeneous intent, relate it to the theories developed in Chapter 3, and provide simulation support for all of the above.

6.1 Motivating Example: The Confounded Physicians

Example 6.1.1. We begin by considering a motivating example depicting UCs in medical decision-making. In this scenario, physicians regularly prescribe one of two FDA-approved drugs to treat a certain condition. Each of the drugs, denoted $X \in \{0, 1\}$, have been shown to be equally effective at treating the condition in a randomized clinical trial (RCT); specifically, for patient recovery $Y \in \{0, 1\}$ where $Y = 1 = y_1$ indicates recovery, the study found a 70% recovery rate for each drug, i.e., $P(y_1|do(x)) = 0.7 \; \forall \; x \in X$. In reviewing her own patient records, one physician confirms this recovery rate, noting that the recovery rates of each patient she has treated are also recovering at the experimentally reported rates, i.e.

$P(y_1|x) = 0.7 \; \forall \; x \in X$. However, upon attending a conference with other physicians from a variety of backgrounds, she learns that some of her colleagues are not witnessing the same recovery rates at their practices.

Supposing that patient populations between physicians (at least on metrics relevant to the current condition) are exchangeable, it is plausible that the differences in witnessed recovery rates could be explained by treatment assignments that are confounded with patient recovery. This possibility is not without precedent, as recent studies that have investigated the complex mechanisms of treatment selection have implicated a rich tapestry of interactions between patient, physician, and healthcare system that, either through direct or indirect pathways, ultimately confound treatment with recovery [BSG10]. From the physician's side, treatment decisions may not only be based on their subjective perception of the patient's prognosis, but also of their opinions of (or experiences with) the available treatments, their assessments of the patient's ability, willingness, or financial capacity to comply with the treatment, and a variety of other factors. From the patient's side, in what is known as *adherence bias*, compliance to treatment could covary with other healthy lifestyle choices, which ultimately account for recovery (either by these choices alone or their undocumented interaction with the treatment) [Whi05]. Furthermore, *direct-to-consumer advertising (DTCA)* is a practice that allows pharmaceutical companies to advertise their drugs directly to patients; though a prescription should, in theory, be based solely on objective metrics of per-patient applicability, advertising has been shown to increase the sale of drugs, indicating a patient-requested effect on their selection. Such requests may not be recorded in patient histories (or if they are, may not be considered a diagnostically relevant factor), and so any influences on the final treatment assignment that amount from these requests are not available for analysis [Lyl02, Ven11].

In the present example, we will, for simplicity, consider only two such possible unobserved confounding factors. The first is the patient's socio-economic status (SES) that we will encode as either low-SES ($S = 0$) or high-SES ($S = 1$). A patient's SES may be heuristically assessed by the physician (for example, through anecdotal indicators or appearance of the patient) and influence their treatment based on differences between the short- or long-term

expenses of different therapies. Consider also that SES may covary with certain nutritional quality, such that higher SES patients may have access to better or more diverse meals that interact with the given treatments in different ways. The second UC will be the patient's treatment request, which can be influenced by DTCA. In particular, a patient may request one treatment ($R = 0$) over another ($R = 1$), which may influence a physician's decision if they decide to accommodate such requests. Consider also that an indirect pathway may link the medication requested to certain recovery covariates; for instance, it is possible that a drug advertised on a sports station will be observed by patients who tend to get better exercise, and thus have better cardiovascular health (which then may interact with assigned treatment). Plainly, there are many such influencing factors that may act as UCs in treatment assignment, but we will demonstrate the procedures herein using R and S for illustrative purposes.

Returning to our physicians' conference at which they are comparing recovery rates for drugs $X = 0$ and $X = 1$, suppose that different physicians have different assignment policies. In particular, consider that more accommodating physicians will attempt to honor their patients' requests for one medication over the other, but are also influenced by their perception of each patient's SES. Physicians of this "type" assign treatment by the structural equation, $X \leftarrow f_X^{P_1}(S, R) = XOR(S, R)$. Now, suppose another type of physician is aware of the influences of DTCA, and consciously refuses to let patient requests influence their decisions; as such, these physicians' treatments can be modeled by the structural equation $X \leftarrow f_X^{P_2}(S) = S$.

Modeling the reality of this scenario from an omniscient viewpoint, we note that there is an even patient distribution over SES and requesters for each drug i.e., $P(r) = P(s) = 0.5 \; \forall \; r \in R, s \in S$. As such, the *true* probabilities of recovery from the condition under each confounder state are listed in Table 6.1(a). Also of note, the recovery rates (derived from this "true" distribution) witnessed in the FDA's experimental study are shown in Table 6.1(b) along with the observational recovery rates of the accommodating physicians of type P_1 and those of the stringent physicians P_2 (where $f_X^{P_1}(S, R) = XOR(S, R)$ and $f_X^{P_2}(S) = S$).

Scrutinizing this data, we see that the observational treatment policy of physician P_1 represents a case of *invisible confounding*; namely, $P(Y|do(X)) = P^{P_1}(Y|X) \; \forall \; x \in X$, yet

(a)	$S=0$		$S=1$	
$P(y_1\|X,S,R)$	$R=0$	$R=1$	$R=0$	$R=1$
$X=0$	$^{P_1,P_2}0.70$	$^{P_2*}0.80$	0.60	$^{P_1*}0.70$
$X=1$	$^{*}0.90$	$^{P_1}0.70$	$^{P_1,P_2*}0.70$	$^{P_2}0.50$

(b)	$P(y_1\|do(X))$	$P^{P_1}(y_1\|X)$	$P^{P_2}(y_1\|X)$
$X=0$	0.70	0.70	0.75
$X=1$	0.70	0.70	0.60

Table 6.1: (a) Recovery rates as a function of drug choice X, patient SES status S, and patient treatment request R. The observational treatment assigned by physicians of type 1 are indicated by P_1, and those by type 2 are indicated by P_2 (where $f_X^{P_1}(S,R) = XOR(S,R)$ and $f_X^{P_2}(S) = S$). The optimal treatment under each configuration of S, R are indicated by asterisks. (b) Recovery rates according to the FDA experiment, $P(y_1|do(X))$, the observations of physician 1 $P^{P_1}(y_1|X)$, and the observations of physician 2 $P^{P_2}(y_1|X)$, where $Y = y_1$ represents recovery (shown in the table).

there are indeed confounding factors present in the system that the statistical distribution over recovery does not reveal alone. The plight of physician 2 is not entirely better; while the recovery rates associated with the ostensibly optimal drug $X = 0$ are superior in two configurations of S, R, and it appears as though P_2 receives more discriminant information about the UCs compared to P_1 (since $P(Y|do(X)) \neq P^{P_2}(Y|X)\ \forall\ x \in X$) we can see from Table 6.1(a) that there exist conditions under which $X = 1$ is actually the optimal assignment choice.

Having now compared their notes and observed recovery rates using each of the drugs, physicians P_1 and P_2 consider how they might repair for the influence of confounding factors. After some research, they discover that dynamic experiments using intent-specific decision-making (ISDM, Def. 3.4.2) may be appropriate. Each returns to their respective practices and collects data on the intent-specific recovery rates of each drug. The results of their experiments are displayed in Table 6.2. Perhaps surprisingly, the intent-specific recovery rates of P_1 appear to be no different than the observational and experimental recovery rates

for each drug. Using the Regret Decision Criteria (RDT, Def. 3.4.5) as a maximization criteria, the expected recovery rates of any arbitrary patient of P_1 will be 70% – no different than the results of a Causal Decision Theory (CDT, Def. 3.4.4) maximization. Even the results of the ISDM experiment from P_2 make marginal improvements over the CDT average such that the recovery rates of an arbitrary patient of P_2 under RDT maximization will be 72.5%.

$P^{P_1}(Y_x = 1\|x')$	$x' = 0$	$x' = 1$
$x = 0$	0.70	0.70
$x = 1$	0.70	0.70

$P^{P_2}(Y_x = 1\|x')$	$x' = 0$	$x' = 1$
$x = 0$	0.75	0.65
$x = 1$	0.80	0.60

Table 6.2: Results of ISDM dynamic experiments conducted by physicians P_1 (left) and P_2 (right). The intent-specific recovery rates witnessed by P_1 are illustrative of *invisible confounding*.

With these latest results in hand, the two physicians once again compare notes. They face a perplexing situation in which the results of physician P_1's ISDM experiment suggest that no confounding exists, yet P_2's seems to suggest that there does. Even so, the improvement in recovery rates witnessed by P_2 in the ISDM experiment appear to be only marginal improvements over the experimental average of 70%. The two physicians ponder whether to conclude that confounding is present or not, and more importantly, whether they might still be able to improve the recovery rates of their patients.

6.2 Formalizing Heterogeneous Intent

The Confounded Physicians Example 6.1.1 demonstrates a decision-making scenario with several noteworthy characteristics:

1. Although under the influence of confounding, physician P_1's intent-specific recovery rates show no indications of any unobserved factors that might distinguish observational, experimental, or counterfactual recovery distributions (see Tables 6.1(b), 6.2).

2. Yet, physician P_2 *does* exhibit some traditional, statistical indications of confounding, such that recovery differences are manifest between observational, experimental, and counterfactual distributions (see Tables 6.1(b), 6.2).

In the present section, we will first formalize the distinguishing features of this scenario from past examples in the Greedy Casino, then demonstrate how our new way to model the scenario can be used in not only an online, dynamic experiment, but can also improve the traditional offline Randomized Clinical Trial (RCT).

We begin by explaining the statistical phenomena experienced by physician P_1 such that confounding exists despite equivalence between observational, experimental, and counterfactual quantities. We refer to this type of scenario as one with "invisible confounding."

Definition 6.2.1. (Invisible Confounding) For some decision variable X (Def. 3.3.2) and some measured outcome of that decision Y, we say that X and Y are subject to invisible confounding whenever

$$P(Y|X) = P(Y|do(X)) = P(Y_x|x') \neq P(Y_x|U) \ \forall \ x, x' \in X, \ u \in U \tag{6.1}$$

Invisible confounding is possible in any setting with unobserved confounders, though requires a careful tuning of outcome parameters to be manifest.[1] Invisible confounding is particularly subtle in settings wherein all reasoning agents are of homogeneous intent (Def. 3.5.1). Were all physicians to possess the same observational choice policy as P_1, all agents would (from an omniscient perspective) experience linear u-regret (Def. 3.2.2), never converging to an optimal policy. Mercifully, in the present setting, different agents possess different observational decision-making functions, which we call *heterogeneous intents*.

Definition 6.2.2. (Heterogeneous Intents) Let A_1 and A_2 be two agents within a MABUC instance, and $M^{\Pi}_{A_1}$ be the SDM (Def. 3.3.1) associated with the choice policies

[1]For this reason, we concede that invisible confounding is a strongly artificial phenomena, but one that the present technique will help us address nonetheless.

of A_1 and likewise $M^{\Pi}_{A_2}$ be the SDM associated with the choice policies of A_2. For any decision variable $X \in \Pi_M$ and its associated intent $I = f_x$, the agents are said to have heterogeneous intent if $f_I^{A_1} \in F_{M^{\Pi}_{A_1}}$ and $f_I^{A_2} \in F_{M^{\Pi}_{A_2}}$ are distinct, viz., if $f_I^{A_1} \neq f_I^{A_2}$.

In the Confounded Physicians Example 6.1.1, physicians P_1 and P_2 are said to have heterogeneous intents, since a data point from one of the physician's observational / intent-specific recovery distributions are not exchangeable (i.e., are not necessarily sampled from the same configuration of background variables U) with the other's. This is because $f_I^{P_1} = XOR(S, R) \neq S = f_I^{P_2}$, and so an intent to treat with, say, drug $X = 0$ provides different indications of S, R depending on whether it was P_1 or P_2 who experienced it. As a consequence, $P^{P_1}(Y|X) \neq P^{P_2}(Y|X)$ and $P^{P_1}(Y_x|x') \neq P^{P_2}(Y_x|x')$ (as demonstrated in Tables 6.1, 6.2). Though this relationship may appear to be a modeling complication, we will demonstrate that it can be exploited to yield a choice policy that is more successful than either agent individually.

6.2.1 Online Heterogeneous Intent-specific Decision-making

We can structure the learning problem of the Counfounded Physicians as a dynamic experiment in which the disparate predilections of each physician are concerted to yield a superior choice policy. As such, suppose our confounded physicians attempt to determine whether or not they are subject to confounding by conjoining their practices for some period of time, and begin hosting joint diagnostic sessions. By doing so, not only do they ensure homogeny of patient populations, but also that any confounding factors that might be to blame for differences in treatment affects can be controlled through ISDM by affecting both of them in the same way at the same time for every data point. The procedure that they agree to is as follows:

1. A patient visits the clinic suffering from the condition in question, at which point both physicians will be jointly present for the consult.[2]

[2]Plainly, this is not a feasible requirement for the average physician; for the present example, however, we will demonstrate its utility and later, how it can be applied in a more realistic scenario.

2. Having heard the patient's complaints, run any preliminary tests, and asked any diagnostic questions (per each physicians' usual routine), each physician develops an intended treatment for the patient. Note that the key assumption here is that each physician is exposed to the same configuration of the UCs, but, due to their heterogeneous intents, may react differently to them.

3. Each physician then submits their intended treatment to a learning system, which then decides the final treatment choice for the patient.

4. The resulting recovery (or lack thereof) for the patient is recorded alongside the treatment choice and the intended treatment of each physician individually.

With this procedure in mind, the physicians decide to model the dynamic experiment as a MABUC problem. They note that the original formalization of a Structural Decision Model (Def. 3.3.1) would be appropriate for this task, but while SDMs are *capable* of accommodating scenarios with heterogeneous intents, it will be useful to explicitly denote that the scenario at hand may contain disparate intent functions for the treatment decision. Thus, they formalize the notion of a Heterogeneous Intent Structural Decision Model.

Definition 6.2.3. (Heterogeneous Intent Structural Decision Model) A Heterogeneous Intent Structural Decision Model (HI-SDM) represents a *composite* of individual Structural Decision Models (SDMs, Def. 3.3.1) wherein Decision Variables (Def. 3.3.2) can be a function of distinct intent functions. HI-SDMs are denoted M^{Π_A} where A is the set of heterogeneous agents in the system. Formally, we consider that $A = \{A_1, A_2, ..., A_a\}$ denotes the heterogeneous *intent equivalence classes* of actors in the model, such that for at least one decision variable $X \in \Pi$, $A_i \neq A_j \Leftrightarrow f_I^{A_i} \neq f_I^{A_j}\ \forall\ A_i, A_j \in A$. An HI-SDM connects the agent-specific SDMs (i.e., $M^{\Pi_{A_1}}, ..., M^{\Pi_{A_a}}$) such that the decision variables, outcome variables, and UCs in each SDM all correspond to the same unit, t.

Definition 6.2.4. (Intent Equivalence Class (IEC)) In an HI-SDM M^{Π_A}, we say that any two actors $A_i \neq A_j$ belong to separate *equivalence classes* of intent functions f_I for a particular decision variable X and outcome variable Y if $P(Y_x|I^{A_i}) \neq P(Y_x|I^{A_j}) \neq P(Y_x|I^{A_i}, I^{A_j})\ \forall\ x, i^{A_i}, i^{A_j}$, thus $f_I^{A_i} \neq f_I^{A_j}$.

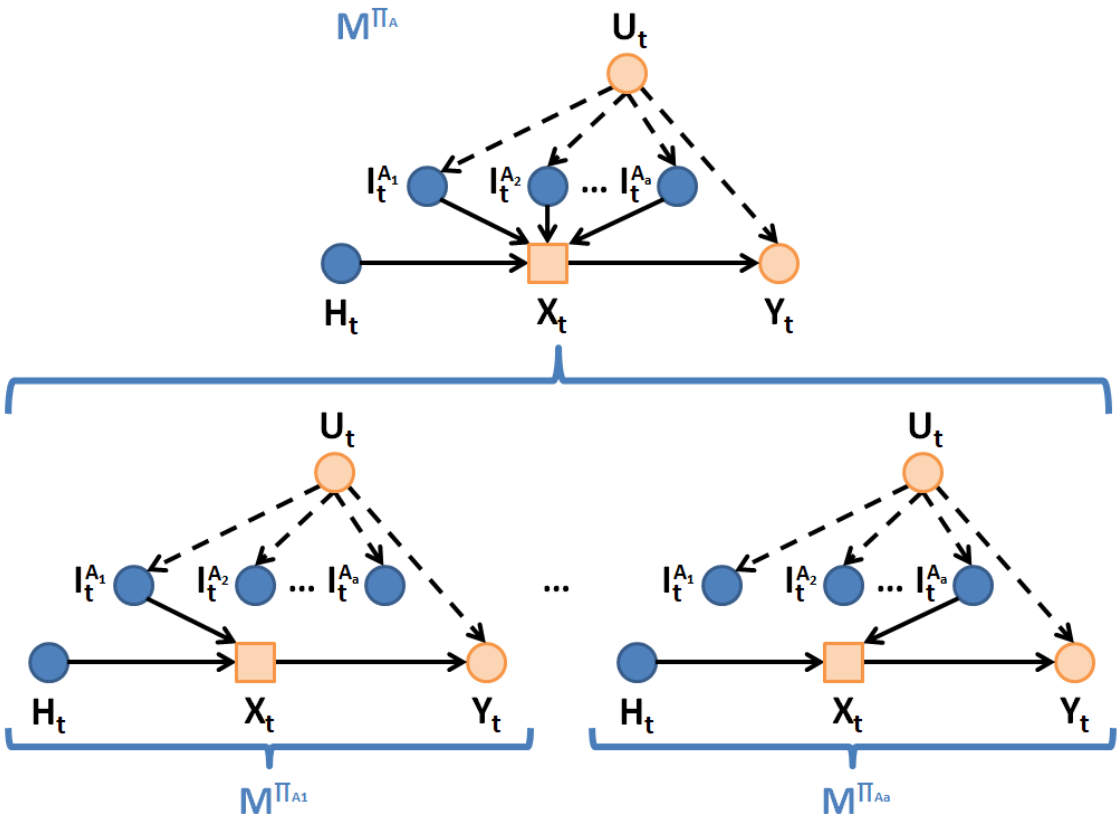

Figure 6.1: Graphical model of prototypical HI-SDM M^{Π_A} as a composite of individual IEC SDMs $M^{\Pi_{A_1}}, ..., M^{\Pi_{A_a}}$. Variables shared between each model that correspond to a particular unit t are highlighted in orange (viz., U_t, X_t, Y_t).

Figure 6.1 depicts the interpretation for the prototypical HI-SDM, M^{Π_A}, which represents a composite of SDMs for each agent IEC $A = \{A_1, A_2, ..., A_a\}$ such that $M^{\Pi_{A_i}}$ is a homogeneous-intent SDM for a single agent. This decomposition allows us to preserve intra-agent observational outcomes of the format $P(Y|X) = P(Y|I)$ and counterfactuals of the ISDM format $P(Y_x|X = x') = P(Y_x|I = x')$ while also capturing inter-agent treatment outcomes, discussed shortly. Importantly, what allows us to holistically discuss the HI-SDM and its constituent SDMs is the idea that, although each agent IEC's response represents a different functional relationship with the confounder state $U_t = u_t$, finally assigned treatment $X_t = x_t$, and outcome Y_t correspond to the same unit t in each model. For instance, if we consider a "unit" to be a particular patient in the Confounded Physicians example, then

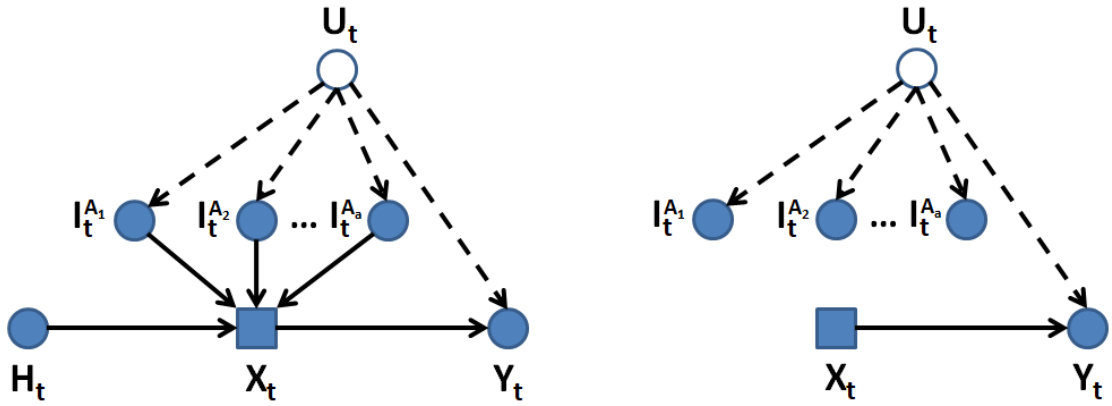

Figure 6.2: Juxtaposition of graphical models for online vs. offline HI-SDMs. (Left) Graphical model of a prototypical HI-SDM M^{Π_A} for an online MABUC instance with decision variable X_t, outcome Y_t, unobserved confounders U_t, and agent history H_t. (Right) Graphical model of a prototypical HI-RCT ($M_x^{\Pi_A}$) wherein treatment is assigned at random, but results can be enriched by conditioning on distinct IECs.

it is assumed that U_t corresponds to that particular patient's features, X_t corresponds to that particular patient's final treatment assignment, and Y_t corresponds to that particular patient's recovery.

Note that the prototypical SDM employed in previous chapters is merely a special case of the heterogeneous intent SDM such that a homogeneous intent SDM is a model $M^{\Pi_{A_1}}$ for a single actor intent class A_1. Additionally, we see that a particular trial's configuration of UCs is linked to each actor's intent for that unit; in this way, each heterogeneous intent function can provide a more complete picture of the UC state (when considered in concert) than any individual intent function in isolation. This fact provides the impetus for our modeling decision to design approaches that consider a final decision X_t that is conditional upon *all* IECs, as demonstrated in the HI-SDM itself.

Because each actor's intent provides a potentially separate piece of the UC's "puzzle," the way our learning agent should record each trial changes only slightly from the method described in Figure 3.6. In particular, the agent will be able to record intent-specific rewards for not only each actor individually (i.e., for each agent's SDM yielding $P(Y_x|I^{A_1}), ..., P(Y_x|I^{A_a})$,

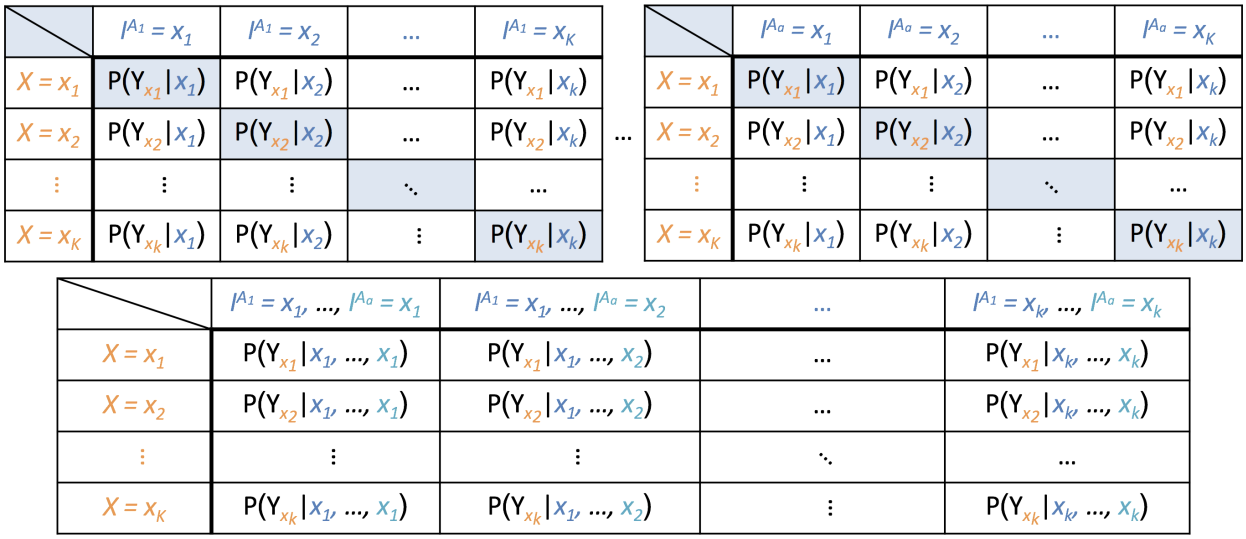

Figure 6.3: Tabular reward histories of (top) individual actor intent-specific rewards and (bottom) combined actor intent-specific reward distributions.

Figure 6.3 (top)) but also a Combined table, containing the rewards of each arm sampled under all combinations of all actors' intents, i.e., $\forall\ I^{A_i} \in I^A$ (i.e., $P(Y_x|I^{A_1}, ..., I^{A_a})$, Figure 6.3 (bottom)).

To visualize the merit of this organization of reward distributions, we return to the individual treatment success rates of the physicians in the Confounded Physicians Example 6.1.1. Recall that the chief benefit of conditioning upon intent is the information that it yields about the state of the UCs (as demonstrated from the Greedy Casino Example in Table 3.2). In the Confounded Physicians Example, by comparison, the information about S, R provided by I^{P_1}, I^{P_2} is via the distributions $P(S, R|I^{P_1}), P(S, R|I^{P_2})$, respectively. The conditional distribution over each individual actors' intent is shown in Tables 6.3(a, b). Note that the precise configuration of UCs indicated by either actors' individual intents is split between a possibility of 2 states; this ambiguity is the source of the invisible confounding (Def. 6.2.1) manifest in the example. However, when each actors' intents are considered together, this ambiguity is resolved, and each combination of the heterogeneous intents points to a precise configuration of UCs (see Table 6.3(c)).

Although the distributions detailed in Table 6.3 would be unavailable to any reasoning

(a)	$S=0$		$S=1$	
$P(S,R\|I^{P_1})$	$R=0$	$R=1$	$R=0$	$R=1$
$I^{P_1}=0$	0.50	0.00	0.00	0.50
$I^{P_1}=1$	0.00	0.50	0.50	0.00

(b)	$S=0$		$S=1$	
$P(S,R\|I^{P_2})$	$R=0$	$R=1$	$R=0$	$R=1$
$I^{P_2}=0$	0.50	0.50	0.00	0.00
$I^{P_2}=1$	0.00	0.00	0.50	0.50

(c)		$S=0$		$S=1$	
$P(S,R\|I^{P_1},I^{P_2})$		$R=0$	$R=1$	$R=0$	$R=1$
$I^{P_1}=0$	$I^{P_2}=0$	1.00	0.00	0.00	0.00
	$I^{P_2}=1$	0.00	0.00	1.00	0.00
$I^{P_1}=1$	$I^{P_2}=0$	0.00	1.00	0.00	0.00
	$I^{P_2}=1$	0.00	0.00	0.00	1.00

Table 6.3: Probability of each UC state $\{S=s,R=r\}$ given the intent of each actor (physician). (a) depicts the probability of each UC state for P_1 individually, and (b) for P_2 individually. (c) Probability of each UC state for concerted intents.

agent (because their derivation depends on knowledge of the fully-specified model), the empirical benefit of concerting heterogeneous intents is tangible. In particular, we can consider the heterogeneous intent-specific recovery rates of each drug in Table 6.4. Viewing Table 6.4, we can make several key remarks: (1) though an extreme case, we see that conditioning on heterogeneous intents has reproduced the parameters in the "true" recovery distribution (Table 6.1(a)) without ever having to know the states of S,R; (2) as a consequence (and what will be formalized briefly), agents that condition on the heterogeneous intents of P_1 and P_2 in the present example will reach the optimal choice policy, and minimize u-regret; and (3) the optimal policy from combining heterogeneous intents will experience a higher recovery rate (77.5%) than either actor's ISDM recovery rates individually (70% for P_1 and 72.5% for P_2; see Table 6.2).

	$I^{P_1}=0$		$I^{P_1}=1$	
$P(Y_x=1\|I^{P_1},I^{P_2})$	$I^{P_2}=0$	$I^{P_2}=1$	$I^{P_2}=0$	$I^{P_2}=1$
$X=0$	0.70	0.60	0.80	0.70
$X=1$	0.90	0.70	0.70	0.50

Table 6.4: Recovery rates for each drug given the intents of both physicians P_1 and P_2.

These results provide us with an extension to the Regret Decision Theory (RDT, Def. 3.4.5) in which the intents of actors in the environment are heterogeneous. We describe this extension as Heterogeneous Intent Regret Decision Theory, which follows from considering actors (and their corresponding intent functions) as members of certain IECs.

In words, if two actors possess the same intent function for a particular decision variable (i.e., their observational action choice predilection), then we would expect that both (1) their intents and (2) their counterfactual reward quantities (at any given time in a confounded decision-making task) will coincide. However, we should note that, while this definition is true in one direction (i.e., that if two actors belong to the same IEC, that their intents and counterfactual rewards will agree), the opposite is true only to a degree of *observational equivalence*. Because a fundamental assumption of the confounded decision-making scenarios is that the reasoning agent does *not* possess the fully-specified SCM of the environment, it is possible to observe two actors eliciting the same intents and the same counterfactual rewards, but still possess two separate intent functions. For our purposes in the present task, however, observational equivalence of actor intents will be sufficient, given that we merely wish to obtain some information about the state of the UCs through the proxies of intent (a la Table 6.3). If the final condition for two actors' intents to be considered entities of the same equivalence class holds (i.e., $P(Y_x|I^{A_i}) = P(Y_x|I^{A_j}) = P(Y_x|I^{A_i}, I^{A_j})$), then we gain nothing from conditioning on both agents' intents separately.

As such, we can now specify our new optimization criteria for agents to maximize in a heterogeneous intent confounded decision task, which applies to environments in which actors belong to distinct IECs.

Definition 6.2.5. (Heterogeneous Intent Regret Decision Theory (HI-RDT)) Het-

erogeneous Intent Regret Decision Theory (HI-RDT) states that, for all distinct IECs (Def. 6.2.4) $A = \{A_1, A_2, ..., A_a\}$ of actors in a Heterogeneous Intent Structural Decision Model M^{Π_A} (Def. 6.2.3), conditioning on the intents of all actors provides evidential context for the state of the environment that is richer than the context of any individual intent. HI-RDT agents thus maximize the reward Y from a distribution over the action space X given the intents of actors in each equivalence class. The optimal action $x^* \in X$ is thus defined as:

$$x^* = \underset{x \in X}{\operatorname{argmax}} P(Y_x | I^{A_1}, I^{A_2}, ..., I^{A_a}) \tag{6.2}$$

In words, HI-RDT prescribes that agents should maximize reward within the context of actors' intents of heterogeneous IECs to learn as much about the U-specific reward distribution using each I^A as a proxy for the state of the UCs. We can now depict the workflow of a HI-RDT agent in a Heterogeneous Intent MABUC scenario, as shown in Figure 6.4.

1. We begin by considering the utility of any observational and experimental data from the environment that may aid in the heterogeneous intent MABUC learning process, as by tenets of counterfactually-enabled data-fusion presented in Chapter 5. *Observational data* could exist for each actor's past experience with each action (or drug choice, in the case of the Confounded Physicians Example), though under the effects of their personal IEC. As such, any observational data points can be used to populate each individual IEC's intent-specific rewards, and does not provide information about the combined-intent reward. *Experimental data*, on the other hand, may provide information about the average treatment effect of each action in both individual IEC and combined IEC reward tables (e.g., as by the results of the FDA study). Though the possibility of employing such side information to speed learning in a heterogeneous intent MABUC exists (for the same reasons it helped in Chapter 5), we will not consider it in the present work.

2. From the current configuration of UCs (in the Confounded Physicians Example: S, each physician's perception of the patient's socio-economic status, and R, the specific drug request of the patient), each actor develops an intent I^{A_i}.

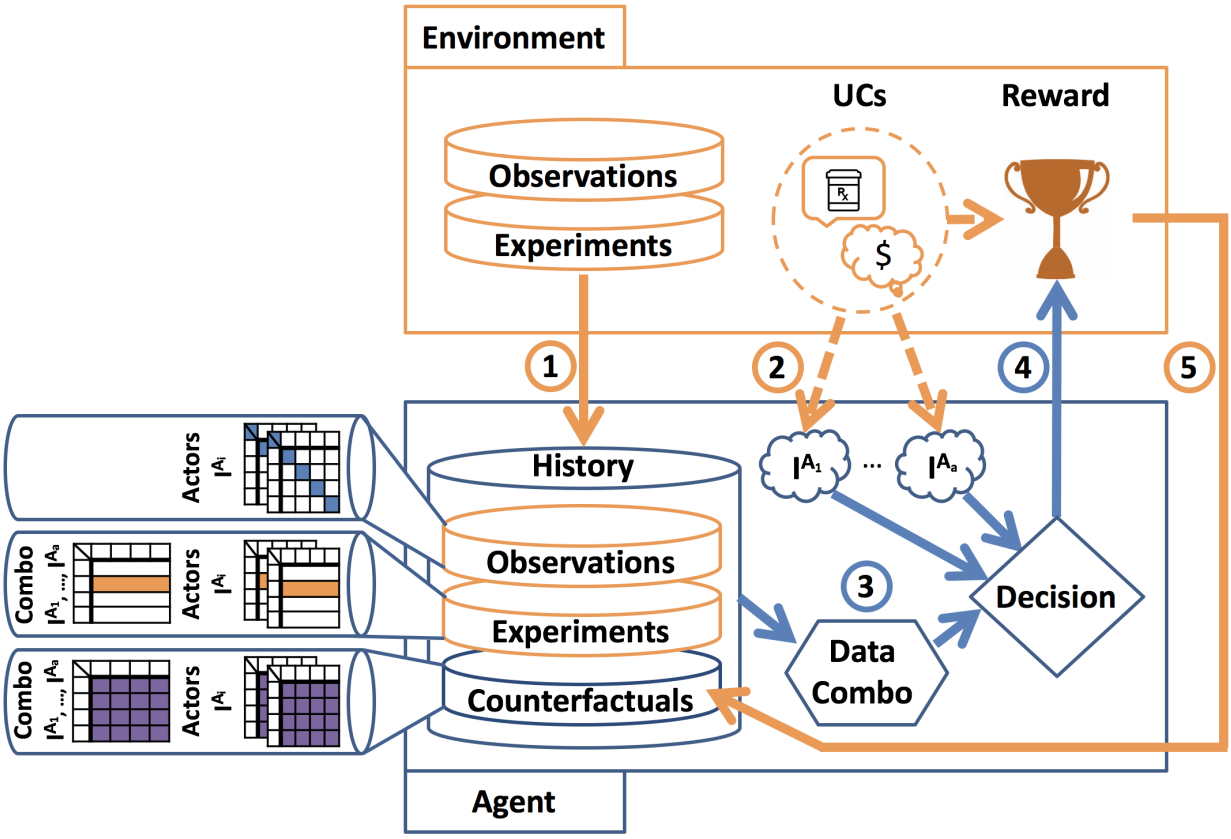

Figure 6.4: Depiction of an online heterogeneous intent MABUC scenario.

3. Using the HI-RDT learning agent's history of IEC-specific rewards, a data combination is accomplished whereby each actor's submitted intent is classified into a set of heterogeneous IECs $A = \{A_1, A_2, ..., A_a\}$ (i.e., actors believed to be in a homogeneous IEC are summarized in a single intent condition), which serve as the context in which to make a final decision.

4. Based on this data combination, the HI-RDT agent makes a final action decision, and pulls the chosen arm.

5. The environment provides a reward to the agent, which then remembers the outcome as a data-point, populating its distribution over both the combined, HI-IEC space, $P(Y_x|I^{A_1}, I^{A_2}, ..., I^{A_a})$, and the individual, counterfactual IEC space $P(Y_x|I^{A_1}), ..., P(Y_x|I^{A_a})$.

6.2.2 Offline Heterogeneous Intent-specific Decision-making

In the traditional randomized clinical trial (RCT) experiment, control of confounding factors is done through random assignment of participants to experimental conditions with the expectation that any influence of unmodeled factors is averaged in each condition. For instance, in a Food and Drug Administration (FDA) experiment to test the efficacy and safety of new prescriptions, participants are typically randomly assigned to either an experimental condition in which the drug is administered, or a placebo group where it is not. The rates of recovery between these groups is compared after the duration of the study, and a drug's case for approval will be loosely rooted in its improved treatment efficacy over the placebo, as well as considerations for side-effects incurred and a variety of other factors.

However, while an RCT may appear to be an effective, offline strategy to nullify any unmodeled influences between drug assignment and recovery, it also suffers a central weakness: any unmodeled confounding factors that would be manifest during real-world treatment assignment (such as S, R in the Confounded Physicians Example) are not discovered until after the drug is in the hands of physicians prescribing it. Worse yet is that these factors may remain undetected if the confounding effects exhibit cases of invisible confounding (Def. 6.2.1) like they did for physician P_1 in the Confounded Physicians Example. In the best case, these confounding factors lead to an improvement in efficacy over the experimental results; in the worst, they actually may impede treatment effectiveness. Ideally, we would like to discover the identities and states of any confounding factors for drug assignment, but before that, we should be wary to determine whether or not there are any such UCs in the system to begin with.

Since RCTs will not reveal the presence of any confounders alone, and even observational follow-ups with physicians may obscure the presence of confounders (e.g., through invisible confounding), suppose instead that we collect data by sampling the intended treatments for each patient from a representative sample of physicians, and determine (by sampling the HI-RDT reward distribution) if confounders exist for certain physician IECs, and if so, to then isolate and identify these unmodeled influences. That said, HI-RDT is a reward maximization

criteria for an online learning agent in a dynamic experiment like a MABUC; the analogy between a MABUC scenario and an RCT study breaks down on several levels: (1) a MABUC agent attempts to determine the optimal arm choice as soon as possible, and then continues to exploit that arm after reaching a degree of confidence about its optimality, whereas an RCT attempts to ascertain some treatment efficacy between groups to a certain degree of confidence; (2) there may be ethical ramifications surrounding treatment assignment that is a function of any actor's input that are manifest in MABUC scenarios (by virtue of intent-specific decision-making) but should not be in an RCT; (3) data may be expensive or prohibitive to collect in some RCT, whereas MABUC scenarios do not always consider a cost associated with treatment at every trial. As it is, FDA RCTs undergo several phases of experimental drug testing before the drug ever appears at market; to require yet another round of HI-RDT experiments atop the existing requirements would impede an already saturated timeline for drug approval.

Instead, let us consider how we might marry the application of HI-RDT agents in the online heterogeneous decision-making domain towards improving traditional RCTs in the offline experimental design domain. The contribution we will make here is a consequence of the measure of Heterogeneous Intent Empirical Counterfactual Estimation, Theorem 6.2.1; viz., that counterfactual outcomes of a particular treatment can be measured empirically by conditioning on the treaters' intents. In confounded decision-making scenarios with heterogeneous actor intent, the same empirical estimability applies (see Eq. 6.3). We refer to the application of HI-SDM in the offline, experimental design domain as a Heterogeneous Intent Randomized Clinical Trial.

Definition 6.2.6. (Heterogeneous Intent Randomized Clinical Trial (HI-RCT)) Let X be the treatment of a Randomized Clinical Trial (RCT) in which all participants are randomly assigned to some experimental condition via $do(X = x)$ with measured outcome Y. Furthermore, let $A = \{A_1, A_2, ..., A_a\}$ be the set of all IECs for administrators of X in the un-intervened HI-SDM M^{Π_A} for which the RCT is meant to apply. A Heterogeneous Intent RCT (HI-RCT) is an RCT wherein treatments are still randomly assigned to each participant, but in addition, the HIs of sampled administrator IECs are collected for each

participant. In an RCT, data is collected over the distribution $P(Y_x)$; in a HI-RCT, data is collected over $P(Y_x|i^{A_1}, ..., i^{A_a})$. The graphical model of a HI-RCT is the mutilated subgraph of its associated SDM such that all inbound edges to the randomized decision variable X are severed, producing $M_x^{\Pi_A}$. A depiction of the prototypical HI-RCT is displayed in Figure 6.4 (right).

One of the more subtle results of a HI-RCT model is that, in the context of an RCT, the collection of each actor's intent can be done before the treatment is assigned and outcome recorded, *or* after, so long as the outcome is not an input to an intent function. To visualize this detail, consider the prototypical heterogeneous intent MABUC model depicted in Figure 6.2 (right). In a traditional RCT model, all incumbent edges to the treatment X are severed because random assignment represents a forced assignment (as by the interventional *do*-operator in the sub-model $M_x^{\Pi_A}$). Though the treatment is no longer a function of each intent I^{A_i}, this does not mean that actors exposed to the same environment U that would typically decide each I^{A_i} cannot still reproduce the desired heterogeneous intent specific rewards. In other words, when treatment is randomized, the information about the outcome Y provided by pre-treatment intents will be the same as that provided by post-treatment intents, assuming each actor is exposed to the same U in both cases (an assumption that is encoded in the SDM, via the edges between U and I in both the online and offline models, Figure 6.2 left and right, respectively). We show this formally in Theorem 6.2.1.1.

An example HI-RCT procedure is depicted in Figure 6.5. For each participant t in some RCT with discrete experimental conditions X (in the figure, demonstrated as two separate treatment assignments), the typical RCT procedure is followed (Figure 6.5 (bottom)): (1) each participant's descriptive information is collected (demographics, medical records, and other data that is deemed relevant to the measured treatment outcome Y). (2) Participants are randomly assigned to a particular treatment condition, as by the operator $do(X_t = x_t)$, indicating that the influences of any causal mechanisms that would otherwise confound treatment with outcome (in the unintervened system) are severed. (3) An outcome for the randomly assigned treatment is recorded for participant t, generating an *experimental* data point in the space of Y_x for assigned treatment $do(X_t = x_t)$.

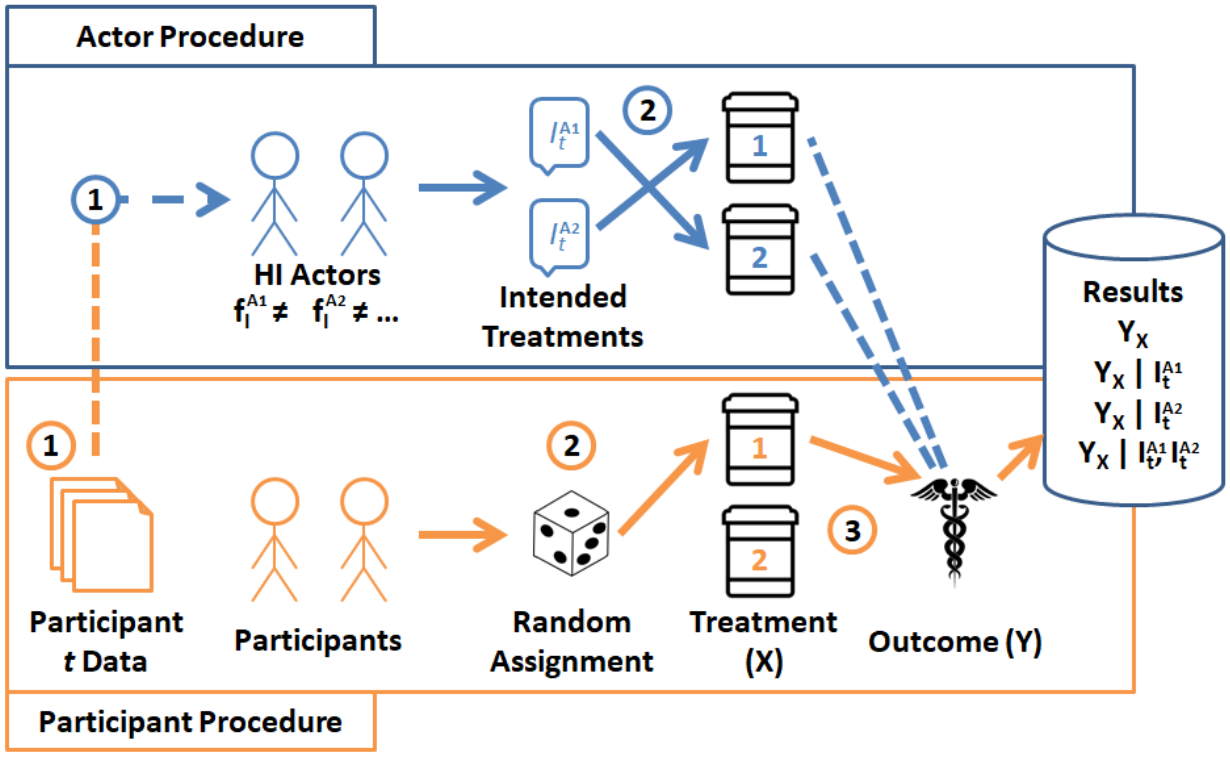

Figure 6.5: Depiction of a HI-RCT in the medical RCT domain.

However, suppose we now model the HIs of actors who would normally be responsible for treatment assignment, e.g., physicians who would be tasked with prescribing the treatments in their practices. For each participant in the original RCT, a selected set of HI physicians will be tasked with the following (Figure 6.5 (top)): each physician is provided with all of the relevant treatment data for participant t such as volunteered medical records, including (as the experiment may see fit) the ability to conduct diagnostic interviews or any other procedural aspects of each doctor's usual pre-treatment routine. Note that each physician is blind to the participant's randomly assigned treatment condition, and is equally blind to their treatment outcome (e.g., recovery or no recovery). (2) From these patient features, each physician makes a recommended treatment (i.e., submits his or her intended treatment) such that for each physician IEC $A = \{A_1, ..., A_a\}$, we possess a corresponding intent $I_t^{A_1}, ..., I_t^{A_a}$ for unit t. Note also that these intents will be evidence for both observed and unobserved outcome covariates per each physician's subjective diagnostic criteria, but will either agree

or disagree with the ultimately randomly assigned (and administered) treatment.

By pairing each physician's intended treatment with the randomly assigned one (and its resulting outcome), we obtain several more informative data points atop the experimental results:

1. $Y_{x_t}|I_t^{A_j} = x_t$ for each actor A_j represents an *observational* data point for A_j, since the outcome corresponding to the treatment that was randomly assigned to participant t, x_t, coincides with the intended treatment $I_t^{A_j} = x_t$. These data are equivalently over the space of $Y|x_t$, and are useful for: (a) comparing with the experimental results to detect confounding (lest it be invisible), and (b) if confounding does exist, can be used to identify high / low actor performance and address its causes.

2. $Y_{x_t}|I_t^{A_j} = x'_t$ for each actor A_j representing a single IEC's *counterfactual* data point for A_j, since the outcome corresponding to the treatment that was randomly assigned to participant t, x_t, contrasts with the intended treatment $I_t^{A_j} = x'_t$. These data are useful for: (a) comparing with the experimental and observational results to detect confounding, and (b) identifying intents that lead to superior / inferior outcome rates compared to the experimental average.

3. $Y_{x_t}|I_t^{A_1} = x_t^{A_1}, ..., I_t^{A_a} = x_t^{A_a}$ for all IECs provides a HI data point, wherein each individual IEC's intent is free to agree or contrast with the administered treatment x_t. These data are useful for: (a) detecting confounding across actor IECs providing the strongest chance to find invisible confounding, (b) identifying superior treatment policies in the case where some combination of HI conditions may lead to better outcomes than observational, experimental, or single actor IECs alone.

Note that because the above procedure collects actor intents atop an existing RCT, it is not necessary to perform any additional experimental trials than are already involved in the traditional approach; the only added component is that actor intents be paired with each participant's randomized treatment and resulting outcome. Thus, an HI-RCT marries observational and experimental studies, with the added piece of providing not only counterfactual

data for individual IECs, but of the HIs as well; this procedure may lead to more robust treatment policies in which previously unforeseen confounding factors can be controled.

6.2.3 Theoretical Results

Just as we provided theoretical proof that ISDM delivers upon its promise to empirically evaluate counterfactual quantities of interest in SDMs, and that doing so will always yield as much or more information than experimental quantities (Chapter 3), so too will we echo these theoretical guarantees for the case of heterogeneous intents. In scenarios involving agents with homogeneous intents, Theorem 3.4.1 demonstrated that counterfactuals of the ETT format $P(Y_x|x')$ could be empirically estimated using intent such that $P(Y_x|x') = P(Y|do(x), I = x')$. We begin with a theorem that is analogous to that of homogeneous intent-specific decision-making such that the counterfactual query of the format $P(Y_{x'}|X^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a})$ is empirically estimable through use of an HI-SDM.

Theorem 6.2.1 (Heterogeneous Intent Empirical Counterfactual Estimation)**.** Let X be a decision variable in a heterogeneous intent SDM M^{Π_A} (Def. 6.2.3) with measured outcome Y, and let $I^{A_1}, ..., I^{A_a}$ be the heterogeneous intents for X of actors in the IECs $A = \{A_1, ..., A_a\}$ in M^{Π_A}. A HI-specific outcome quantity $P(Y_{x'}|I^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a})$ is equivalent to a counterfactual for a single IEC $A_j \in A$, $P(Y_{x'}|X^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a})$, and can be estimated empirically for observed intents $i^{A_1}, ..., i^{A_a}$, and antecedent $X = x'$ (where $X = x'$ indicates the antecedent for *any* of the individual IEC SDMs as well as the HI-SDM since, by assumption, $do(X^{A_j}) = do(X^{A_i})$ for any two IECs $A_i \neq A_j$). Formally, we may write the counterfactual query in interventional notation such that

$$P(Y_{X=x'}|X^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a}) = P(Y_{X=x'}|I^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a}) \tag{6.3}$$

$$= P(Y|do(X = x'), I^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a}) \tag{6.4}$$

Proof. See appendix for proof of Theorem 6.2.1. □

The significance of Theorem 6.2.1 is twofold: (1) the collection of heterogeneous IEC intents does not impede the empirical counterfactual estimation for any individual IEC

SDM and (2) collecting the HI-SDM outcomes provides counterfactual outcomes for each IEC SDM individually; for example, for 2 IECs, the following equivalence holds as a result:

$$P(Y_x|i^{A_1}, i^{A_2}) = P(Y_x|x^{A_1}, i^{A_2}) = P(Y_x|i^{A_1}, x^{A_2}) \tag{6.5}$$

A consequence of Theorem 6.2.1 is the following corollary, which asserts that in an HI-RCT, pre-treatment intent sampling holds the same information about the treatment-specific outcome as does post-treatment intent sampling.

Corollary 6.2.1.1 (Equivalence of Pre- and Post-Assignment Intent Sampling)**.** In an HI-RCT (Def. 6.2.6) with randomly assigned treatment X, measured outcome Y, IECs $A = \{A_1, ..., A_n\}$, and intended treatments of actors in each IEC $I^A = \{I^{A_1}, ..., I^{A_a}\}$, empirical estimation of IEC-specific treatment outcomes can be accomplished by the tenets of the Heterogeneous Intent Empirical Counterfactual Estimation, Theorem 6.2.1. Because HI-RCTs randomize treatment assignment, IECs that are sampled before treatment assignment yield equivalent information about the assigned treatment's outcome Y_x as do those that are sampled after, or formally:

$$P(Y_x|i^{A_1}, ..., i^{A_a}) = P(Y_x|i_x^{A_1}, ..., i_x^{A_a}) \tag{6.6}$$

Proof. See appendix for proof of Corollary 6.2.1.1. □

The significance of Theorem 6.2.1.1 is that, assuming the preservation of any data relevant to each unit in the RCT, IEC intents can be collected either before or after the RCT has been conducted in order to obtain HI-RCT data (i.e., Figure 6.3).[3] This finding may lead to a proliferation of new studies and more informative results from existing RCTs, with only modest additional requirements of labor.

In the online setting, we must also be careful to distinguish the u-regret that would be experienced by an HI-RDT agent (which is knowable only to the omniscient modeller, in possession of the fully-specified model) and the regret that would be experienced under context

[3]Note that this claim assumes that the unit-specific UC state $U_t = u_t$ is invariant to treatment assignment (i.e., $U_{t,x} = U_t$), as is the case in the canonical HI-RCT depicted in Figure 6.2 (right).

of each heterogeneous intent. We call this latter, observable regret the Heterogeneous-Intent-Specific Decision-Maker Regret (hi-regret).

Definition 6.2.7. (Heterogeneous-Intent-Specific Decision-Maker Regret (hi-Regret)) For a MABUC problem with time horizon T, decision variable $X \in \{x_1, ..., x_k\}$ (where $K = |X| \in \mathbb{N}, K \geq 2$ represents the number of choices), reward Y, and heterogeneous IECs $I^A = \{I^{A_1}, ..., I^{A_a}\}$ (Def. 6.2.4) (where I^{A_j} is the intent experienced by an actor of the IEC A_j for decision X), the *optimal action* $x^*(I^A)$ is considered the one that maximizes expected reward under HI state $I^A = \{I^{A_1} = i^{A_1}, ..., I^{A_a} = i^{A_a}\}$, defined as:

$$x^*(I^A) = \underset{x \in X}{\operatorname{argmax}} P(y_x | I^A) \tag{6.7}$$

The *hi-regret* experienced by an agent using choice policy π at trial $0 < t < T$ is defined as:

$$r_t^{i^A} = P(y_{x^*(i_t^A)} | i_t^A) - y_{x_t^\pi} \tag{6.8}$$

The *cumulative hi-regret* experienced by an agent across all T trials is thus:

$$R_T^{i^A} = \sum_{t=1}^{T} r_t^{i^A} = \sum_{t=1}^{T} P(y_{x^*(i_t^A)} | i_t^A) - y_{x_t^\pi} \tag{6.9}$$

Equipped with this definition, we demonstrate that HI-RDT is superior to RDT decision-making strategies in MABUC problems.

Theorem 6.2.2 (HI-RDT u-regret Reduction is Superior to RDT)**.** Let R_t^u be the cumulative u-regret (Def. 3.2.2) experienced by an agent in a MABUC problem by trial t. If $R_t^u(RDT)$ represents the u-regret experienced by an RDT agent and $R_t^u(HIRDT)$ represents the u-regret experienced by an HI-RDT agent, then as $t \to \infty$, $R_t^u(HIRDT) \leq R_t^u(RDT)$ for all possible MABUC parameterizations.

Proof. See appendix for proof of Theorem 6.2.2. □

Given that Theorem 6.2.2 establishes HI-RDT as a strategy that reduces at least as much u-regret as RDT (and usually more in HI-MABUC scenarios), we next consider the sufficient conditions under which HI-RDT does indeed minimize u-regret.

Theorem 6.2.3 (Sufficiency of hi-regret Minimization for u-regret Minimization)**.** Let $R_t^{i^A}$ be the cumulative hi-regret (Def. 6.2.7) and R_t^u be the cumulative u-regret (Def. 3.2.2) experienced by an HI-SDM agent in a HI-MABUC problem by trial t. As $t \to \infty$, if $R_t^{i^A} = O(1)$ then $R_t^u = O(1)$ if the following equivalence holds:

$$x^*(u_t) = \underset{x \in X}{\operatorname{argmax}}\, P(y_x|u_t) = \underset{x \in X}{\operatorname{argmax}}\, P(y_x|i_t^A) = x^*(i_t^A) \;\forall\; u_t \tag{6.10}$$

In words, sub-linear cumulative hi-regret will imply sub-linear cumulative u-regret if the optimal action under known confounder state $U_t = u_t$ is the same as the optimal action under experienced HIs $I^A = i_t^A$ for all trials $t \in T$.

Proof. See appendix for proof of Theorem 6.2.3. □

With these theoretical results in place, we now see that the simulation results that follow support both the offline and online formulations, and corroborate the theoretical premises laid out above.

6.3 Heterogeneous Intent MABUC Simulations

We now demonstrate the efficacy of HI-RDT in the online MABUC domain, though the results herein can be interpreted to mutually support the procedure of an HI-RCT.[4]

Candidate Algorithms. To make a fair comparison to the RDT agents presented in Chapter 3, we examined variants of Thompson Sampling (TS) bandit players in the Confounded Physicians MABUC reward parameterization. Following the motivating example, we present the TS^{RDT} agents of P_1 and P_2 individually and compare their cumulative u-regret to a TS^{HIRDT} agent that conditions on both actors' (i.e., physicians') HIs; we refer readers to Algorithms 1 and 2 for information on the simulation and TS^{RDT}.

Procedure. The simulation was composed of $N = 1000$ Monte Carlo repetitions of $T = 2000$ trials per repetition. At each trial, t, the state of the UCs $U_t = u_t$ was instantiated,

[4]All simulation source code for Chapter 6 can be found at:
https://github.com/Forns/ucla-forns/tree/master/projects/dissertation/ch6.

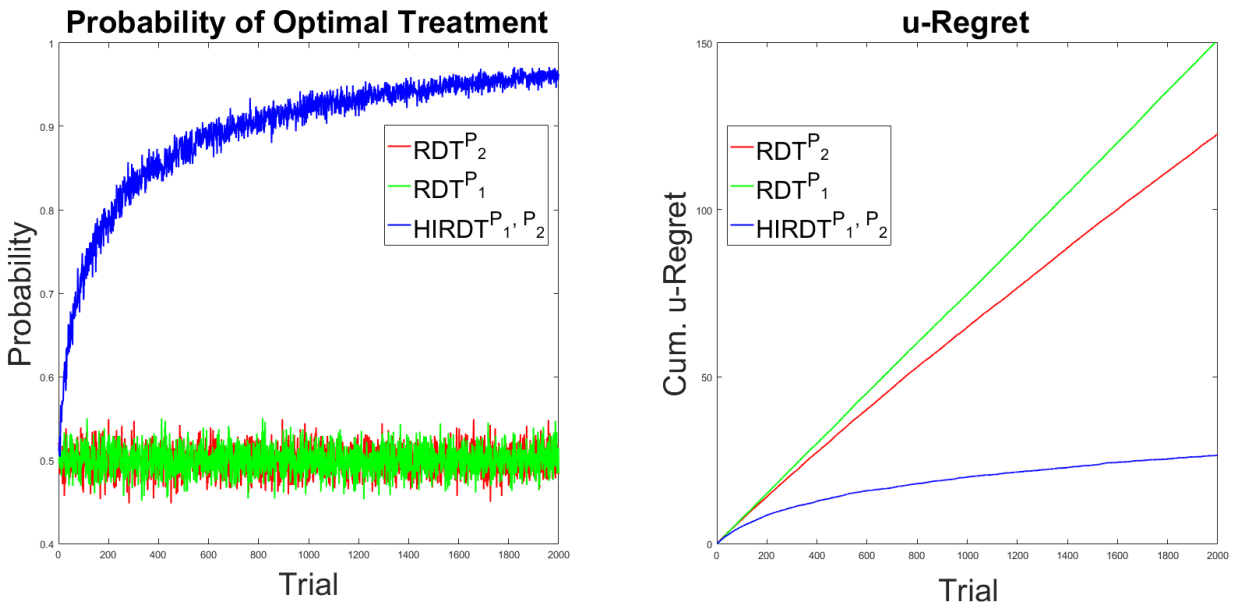

Figure 6.6: Simulation results for the 2-arm Confounded Physicians MABUC scenario.

the intents of each actor $I_t^A = i_t^A$ was then instantiated, the agent made its arm choice $X_t = x_t$, and then received a reward $Y_t = y_t$.

Results. The results of the experiment are depicted in Figure 6.6. Notably, neither RDT approach alone converges to the optimal policy that experiences sub-linear u-regret, but the HI-RDT agent does. These results do not undercut the value of RDT; note that the RDT^{P_1} agent experienced equivalent rewards as a traditional, experimental bandit player would, with the highest u-regret of the three ($M = 150.86, SD = 20.10$), but the next best RDT^{P_2} experienced significantly less ($M = 122.60, SD = 23.59$), $t(1998) = 28.82, p < .001$. The HI-RDT approach, as hypothesized, performed significantly better than either individual RDT agent, experiencing significantly less u-regret than its chief competitor RDT^{P_2} ($M = 26.51, SD = 22.78$), $t(1998) = 129.42, p < .001$.

6.4 Conclusion

In this chapter, we examined a more general version of a confounded decision-making scenario wherein actors possess heterogeneous intent (HI) functions, i.e., different reactionary

criteria to environmental factors. We motivated this scenario by discussing physicians that may possess subjective diagnostic criteria that are different between distinct actor intent equivalence classes (IECs). Prior to this chapter, we assumed the less general case of all actors belonging to the same IEC, but demonstrated that heterogeneous IECs may actually lead to a higher sensitivity version of ISDM. We then defined the HI analog of the Regret Decision Theory (RDT) deemed the HI-RDT, demonstrated that it represents a composite of individual IEC counterfactual expressions, and proved that it yields strictly more information than a single IEC ISDM alone (with simulations to support these theoretical results). Finally, we demonstrated how conditioning on heterogeneous IECs can add a layer to a traditional randomized clinical trial (RCT) in what we deemed an HI-RCT. Compared to a standard RCT, which generates only experimental data, we demonstrated that collecting actors' intended treatments alongside each randomly assigned treatment yields observational, experimental, *and* counterfactual results, with no added experimental cost. The added layer of discrimination afforded by the comparisons of these datasets can better inform policy making and individualized treatment like personalized medicine.

CHAPTER 7

Concluding Remarks

At the onset of this work, we sought to demonstrate the merit of counterfactual reasoning in three disciplines: artificial intelligence, cognitive science, and experimental design. Broadly speaking, the thread that connects these subjects and the present thesis is a study of how intelligent agents learn about their environments through the lens of how their own heuristics and biases interact with it. Although past work has shown that exploration a key component of rational learning, the present study examines the ability to learn from mistakes and correct for a maladaptive policy. Through the formalizations of empirically estimable counterfactuals detailed in the previous chapters, we have shown the mechanisms by which agents can accomplish intent-specific decision-making (ISDM) and the benefits of doing so, but have yet to discuss the implications and significance of this strategy in each of our focal disciplines.

In this chapter, we will discuss the higher-order impacts of the theories presented in the rest of the work. We begin by discussing the significance of our findings to the broader fields of artificial intelligence, cognitive science, and experimental design. We then make an honest assessment of the limitations of ISDM in application to these fields, and discuss possible remedies to some of its shortcomings. Finally, we conclude with a synopsis of future directions for counterfactual reasoning to contribute to these important fields of scientific inquiry.

7.1 Broader Significance

Significance to Artificial Intelligence. The significance of ISDM to the field of artificial intelligence must be examined from two perspectives, as highlighted in Chapter 3: (1) the

capacity of ISDM to behoove an AI agent that is itself confounded in some decision-making task, and (2) the capacity of ISDM to behoove an AI recommender system, that is capable of better informing the decisions of confounded humans in a consultatory nature.

Given the centrality of self-reflection to human learning, it stands to reason that equipping artificial agents with similar reflective capacities may be an important step in the evolution of artificial, general intelligence (AGI). Self-regulating agents have been studied in the AI community for some time, though none have approached the problem from a counterfactual perspective [Doy88]; indeed, the benefit of self-reflection is to better inform future behavior, which implies a change from an existing policy to another. This change is well encapsulated by the counterfactual nature of regret: agents examine their actions in the past that, had they chosen differently, would have resulted in a superior outcome. We posit that ISDM represents an advance on this front, such that an agent's intended action (derived from an existing policy) serves as a self-reflection mechanism should an action counter to intent be discovered as a superior choice. In the preceding chapters, we have already seen evidence for how ISDM informs superior policy formation, and assert that it may feature prominently in future self-regulating systems.

Correcting for human cognitive biases is an ongoing investigation at the national scale: the Office of the Director of National Intelligence's IARPA project (Intelligence Advanced Research Projects Activity) lists a variety of human-centric data science investigations to be used in improvement of policy making and defense. One sponsored competition, the "Hybrid Forecasting Competition," suggests that "Human-generated forecasts may be subject to cognitive biases and/or scalability limits," and that AI systems will be necessary to correct for these biases using more data-driven approaches. The previous chapters have demonstrated ISDM's capacity to control for biases in data when intents are collected alongside other relevant covariates. We foresee ISDM providing an important tool for disentangling the subjective human biases that are manifest in data-sets meant to objectively inform policy making.

Lastly, we highlight the benefit of ISDM to the causal inference community. As presented in the previous chapters, counterfactual quantities are often desirable components of scientific

inquiry in many model-based systems. However, the requirement of a fully-specified structural causal model (SCM) for their traditional means of computation may not be feasible in some settings. ISDM provides a means of computing these counterfactual quantities with only modest modeling assumptions, and takes a data-driven perspective for their empirical estimation.

Significance to Cognitive Science. In the domain of cognitive science, we have, in Chapter 4, seen further support for the idea that humans perform some experientially-based counterfactual reasoning. ISDM may serve as a lens through which cognitive scientists can understand the underpinnings of human bias formation. We have also observed that ISDM does not appear to be a naturally employed human decision-making tactic, though once learned, can lead to superior choice policies even in the presence of unobserved confounders. Apropos, ISDM may serve as a launchpad for understanding mindfulness-based approaches to cognitive-behavioral therapy and bias repair. Mindfulness-based approaches to therapy, which saw a surge in proliferation in the early 2000s, have been shown to improve a variety of mental and physical ailments by improving the salience of, and reactions to, certain bodily and mental signals [AA06] – of which, intent may add an important signal to consider for patients with maladaptive instincts (e.g., addiction). Other efforts from cognitive science have attempted to gamify bias correction, to which intent-specific counterfactual quantities may provide a new metric of success [SBQ14].

Significance to Experimental Design. Randomization has long been the established means of controlling for the influence of UCs in experimental design. As the present work has demonstrated throughout, randomizing treatments is a coarser solution than computing counterfactuals formatted as the Effect of Treatment on the Treated (ETT) – the latter of which is strictly more informative, but requires additional tools to compute; prior to this work, arbitrary ETT computation required a fully-specified causal model. ISDM provides unit-level (e.g., for a particular trial or patient) causal effects without requiring a fully-specified model or even knowledge of the state of any UCs. This quality provides a unifying perspective between observational data (i.e., treatments in accordance with intent) and experimental data (i.e., treatments that are forced assignments like randomization, which are

merely a summation across all intent conditions), while adding an additional empirically estimable counterfactual layer (i.e., treatments forcibly assigned within intent conditions) that can better inform policy making. The prescriptions for HI-RCTs in Chapter 6 may prove instrumental in combatting confounding bias in medicinal contexts. Other approaches have attempted to unify observational and experimental data, like *propensity scoring*, which attempts to obtain causal effects from observational data by examining covariates thought to predict treatment assignment [RR83]. However, these methods do not provide the unit-level accuracy of the intent-specific counterfactual quantities endorsed herein, and are susceptible to the influences of UCs [Pea09]. Whenever the intents of deciders can be collected, we assert that HI-RCTs should be used in favor of RCTs, ushering in new analytic opportunities for the empirical sciences.

7.2 Limitations

Although ISDM is backed by many of the established guarantees of tools from causal inference and the reinforcement learning domains, we would be remiss in our duty as scientists if we did not examine some of its limitations. We will discuss several global limitations of ISDM and several discipline-specific ones that may even represent avenues for future exploration.

Global limitations. The most obvious limitation of ISDM, as defined herein, is the size of the sample space for large action spaces. Due to the tabular intent-specific histories endorsed throughout the work, the number of action $\times$ intent outcomes that need to be sampled grows quickly without stronger, simplifying assumptions. For each action $x \in X$ such that $|X| = K$, there exists a full compliment of intent-specific results such that $|I| = K$, meaning that there are K^2 intent-specific outcome quantities that must be sampled. In finite sample scenarios where the goal is to obtain every intent-specific action outcome, this can be further complicated by intents with a low probability (which are beyond the experimenter's control). Even more fundamental is the assumption that an agent's intent can be reliably captured at all, an issue which is somewhat ameliorated when the collection of intent would behoove the reasoning agent; for instance, the cash bonuses offered to participants

in the MABUC task from Chapter 4 can incentivize intent collection, or if physicians were persuaded that their honest intents might better treat a patient like in Chapter 6. Note that none of these issues compromise the asymptotic guarantees of ISDM, but can represent challenges in real-world application.

Furthermore, from the broader consideration of causal inference, ISDM gives prescriptions for empirical estimation of counterfactuals over *decision variables* (Def. 3.3.2) on which intent may be collected, but computations of arbitrary counterfactuals $P(Y_z|z')$, where Z is *not* a decision variable, may not be estimable (and require modeling assumptions to measure). That said, ISDM's prescription for the measurement of ETT-like counterfactuals are of widespread utility in the empirical sciences, and span a wide variety of important applications detailed throughout this work.

Domain-specific limitations. Interpreting ISDM as a useful means of self-reflection for an artificial agent can be challenging, given that we generally suppose that the agent's inputs have been curated by a programmer and the notion of an "unobserved" input that mutually affects the action-choice and outcome might seem impossible. That said, with the proliferation of deep-learning approaches in many facets of AI, it is conceivable that agents trained on observational data may exhibit the same confounded decision-making that humans would by influence of spurious correlations contaminating causal effects. Moreover, humans (still the gold-standard of general intelligence) are *known* to be influenced by UCs, so it stands to reason that our artificial agents (if or when comparable) may need to likewise navigate the same challenges to decision-making that cognitive biases pose to humans. Our brief examination of ISDM in human decision-making via Chapter 4 speaks to a similar concern: although humans do not appear to *naturally* employ ISDM in confounded decision-making tasks, it is both a theoretically and empirically superior policy. That said, questions remain regarding how humans, at a finer-granularity, form their intents and can often transform UCs to observed covariates that are attended to in future decisions. Although ISDM provides a high-level explanation for how humans employ regret as a learning tool, this work opens other questions of how humans compute counterfactuals, which are out of scope in this thesis but present opportunities for further exploration in the cognitive sciences.

7.3 Future Directions

Counterfactual reasoning is one of the cornerstones of the human intellectual advantage, and should represent a focal investigation for AI practitioners. Although the present work provides a full treatment for empirically estimable intent-specific counterfactuals in a variety of domains, there exist several immediate opportunities to use ISDM as a springboard into other important problems. In an adjacent-possible exploration, which would likely involve a collaborative effort with fields of feature detection, reinforcement learning, and causal inference, the algorithmization of UC discovery would represent a large step toward AGI. Just as humans seek to learn the identities of UCs and treat them as observed contexts (whether it be through active learning or scientific inquiry), so too is it important that artificial agents become autonomous scientists. Important counterfactual statements such as "it would not happen but for X" are central to scientific inquiry and personalized decision-making, and so equipping agents with the capacity to not only answer, but *ask*, these types of questions will be important for developing the next generation of AIs.

Concretely, future explorations from this work might begin by taking a similar approach to policy iteration, whereby confounded policies can be incrementally deconfounded by ISDM and be used to expose the identities and states of the UCs; for example, if background variables tend to covary with intent, but were previously thought to be independent of the decision or outcome, they may be employed in future iterations of the agent's policy. In the HI-RDT domain, addressing the sample space size problem mentioned in the previous section, a worthwhile investigation may address how slightly different heterogeneous intent functions can be smoothed to attain approximations of HI-SDMs in noisier, real-world systems. As a related effort, in the offline experimental design domain, confirmation of the tenets proposed for HI-RCTs would be a worthwhile investigation should a setting arise where the intents of practitioners can be collected alongside a traditional RCT.

7.4 Conclusion

In present work, we demonstrated that unobserved confounders (UCs) present a significant obstacle to causal inference from statistical data, which can complicate policy making and machine learning. We showed that the traditional approach to control of UCs, viz. randomization, operates by averaging the influence of UCs between treatment groups, thus providing population-level outcomes of each treatment. However, population data does not always best inform personalized decisions, in which the optimal solution in the population *on average* may not be optimal for a particular unit of that population (i.e., a particular trial or patient). To determine the optimal unit-level treatments under the influence of UCs, counterfactual quantities must be compared. These counterfactuals, though strictly more informative than experimental data, required a fully-specified model (including a probability distribution of the confounder states) to compute. However, with the invent of intent-specific decision-making (ISDM), we have demonstrated that counterfactual quantities for some decision variable can be empirically estimated when the agent's intent (i.e., its observational decision) is given.

Having formalized the theoretical requirements and results of ISDM, we demonstrated its applicability in a variety of real-world and synthetic problems. In the online reinforcement learning domain, we showed that ISDM leads to superior choice policies in Multi-Armed Bandit problems with Unobserved Confounders (MABUCs), outperforming traditional approaches that maximize experimental, rather than counterfactual, reward targets. We corroborated these findings in a human-subject experiment, wherein we determined that, although humans do not appear to use ISDM naturally, the quality of their decisions can be improved by its employment. We then discussed how ISDM can be applied in domains in which agents' intents are not exchangeable (as previously assumed), and how the solution in these domains can be used to empower offline, traditional, randomized clinical trial experiments. Finally, we discussed the broader implications of ISDM to artificial intelligence, cognitive science, and the empirical sciences, and suggested avenues for future exploration.

APPENDIX A

Supplementary Material for Chapter 3

Theorems and Proofs in Chapter 3

Theorem 3.4.1 (Empirical Counterfactual Estimation)**.** [FPB17] Let X be a decision variable (Def. 3.3.2) in a SDM (Def. 3.3.1) with measured outcome Y, and let I be the agent's intent (Def. 3.4.1) for X. A counterfactual quantity $P(Y_x|x')$ for evidence x' and antecedent x (where $x, x' \in X$ and x need not be equivalent to x') can be estimated empirically using ISDM (Def. 3.4.2). Formally, we may write the counterfactual query in interventional notation such that

$$P(Y_x|x') = P(Y|do(X = x), I = x') \tag{3.14}$$

Proof. Recall that the values of $x \in X$ and $i \in I$ are equivalent, and so let $a, i \in X, I$ wherein a (the antecedent) and i (the observed intent) need not be equivalent. We start by writing the corresponding expansion of the counterfactual, summing over all possible intents, i':

$$P(Y_{X=a}|X = i) \tag{A.1}$$

$$= \sum_{i'} P(Y_{X=a}|X = i, I = i')P(I = i'|X = i) \tag{A.2}$$

$$= \sum_{i'} P(Y_{X=a}|I = i')P(I = i'|X = i) \tag{A.3}$$

$$= \sum_{i'} P(Y_{X=a}|I_{x=a} = i')P(I = i'|X = i) \tag{A.4}$$

$$= \sum_{i'} P(Y|do(X = a), I = i')P(I = i'|X = i) \tag{A.5}$$

$$= \sum_{i'} P(Y|do(X = a), I = i')1(i' = i) \tag{A.6}$$

$$= P(Y|do(X = a), I = i) \tag{A.7}$$

Eq. (A.2) expands the counterfactual using the law of total probability to sum over all intent conditions. Eq. (A.3) follows from the conditional independence $Y_x \perp\!\!\!\perp X|I$ that holds, allowing us to remove $X = i$. Eq. (A.4) follows because $I_x = I$ given that $(I \perp\!\!\!\perp X)_{G_x}$, where G_x is the interventional submodel where all causal parents of X are severed (as represented by the counterfactual antecedent notation). Eq. (A.5) is a notational re-arranging because all variables (Y_x and I_x) are in terms of the interventional submodel M_x (and thus G_x), licensing us to express the quantity using the $do(x)$ notation. Eqs. (A.6, A.7) follow from the fact that, observationally, an agent's final arm choice will always coincide with their intent (i.e., $P(i|x) = 1\ \forall i = x, 0$ otherwise), which nullifies all summed expressions where the two differ. □

Theorem 3.4.3 (Sufficiency of i-regret Minimization for u-regret Minimization)**.** Let R_t^i be the cumulative i-regret (Def. 3.4.6) and R_t^u be the cumulative u-regret (Def. 3.7) experienced by an ISDM agent in a MABUC problem by trial t. As $t \to \infty$, if $R_t^i = O(1)$ then $R_t^u = O(1)$ if the following equivalence holds:

$$x^*(u_t) = \underset{x \in X}{\operatorname{argmax}}\, P(y_x|u_t) = \underset{x \in X}{\operatorname{argmax}}\, P(y_x|i_t) = x^*(i_t)\ \forall\ u_t \tag{3.18}$$

In words, sub-linear cumulative i-regret will imply sub-linear cumulative u-regret if the optimal action under known confounder state $U = u_t$ is the same as the optimal action under experienced intent $I = i_t$ for all trials $t \in T$.

Proof. The conditions under which $R_t^u = O(1)$ are when there exists some t' such that for all trials $t^+ \in [t', T]$, $E[R_{t^+}^u] = 0$. In other words, for some choice policy that converges to the optimal policy after t', the optimal action chosen for all t^+ will be $x(t^+) = x^*(u_{t^+})$. Were $x(t^+) \neq x^*(u_{t^+})$, then there would be some $\epsilon = x^*(u_{t^+}) - x(t^+)$ over which $R_{t^+}^u = \sum_{t \in [t', T]} \epsilon \neq 0$. As such, if $x^*(u_{t^+}) = x^*(i_{t^+})$, then $x(t^+) = x^*(i_{t^+}) = x^*(u_{t^+}) \Rightarrow E[R_{t^+}^u] = 0$. □

Theorem 3.4.2 (RDT u-regret Reduction is Superior to CDT)**.** Let R_t^u be the cumulative u-regret (Def. 3.2.2) experienced by an agent in a MABUC problem by trial t. If $R_t^u(CDT)$ represents the u-regret experienced by a CDT agent and $R_t^u(RDT)$ represents the u-regret experienced by an RDT agent, then as $t \to \infty$, $R_t^u(RDT) \leq R_t^u(CDT)$ for all possible MABUC parameterizations.

Proof. Note that e-regret is defined over the experimental reward space, as would be maximized by a traditional, Causal Decision Theory (CDT, Def. 3.4.4) agent, such that $x^*_{CDT} = \text{argmax}_{x \in X} P(Y_x)$. Conversely, i-regret is defined over the counterfactual reward space, as would be experienced by a Regret Decision Theory agent (RDT, Def. 3.4.5), such that $x^*_{RDT}(i) = \text{argmax}_{x \in X} P(Y_x|i)$. We can next note that:

$$P(Y_x) = \sum_i P(Y_x|i)P(i) \tag{A.8}$$

To show that $R^u_t(CDT) < R^u_t(RDT)$ it is sufficient to show that the rewards collected by a CDT-maximizing agent would be strictly greater than those collected by an RDT agent. Assume to the contrary that this is the case, such that if W_{CDT} represents the expected winnings of the CDT agent and W_{RDT} represents the expected winnings of the RDT agent, $W_{CDT} > W_{RDT}$. We thus have:

$$W_{CDT} = P(Y_{x^*_{CDT}}) = \sum_i P(Y_{x^*_{CDT}}|i)P(i) \tag{A.9}$$

$$W_{RDT} = \sum_i P(Y_{x^*_{RDT}(i)}|i)P(i) \tag{A.10}$$

In other words, CDT chooses the x that maximizes the probability-weighted reward sum over intents, whereas RDT chooses the x that maximizes within-intent reward, summed over the probability-weighted priors of each intent. However, because in either case, each $P(i)$ will be the same, we have:

$$W_{CDT} > W_{RDT} \Rightarrow \sum_i P(Y_{x^*_{CDT}}|i)P(i) > \sum_i P(Y_{x^*_{RDT}(i)}|i)P(i) \tag{A.11}$$

$$\sum_i P(Y_{x^*_{CDT}}|i) > \sum_i P(Y_{x^*_{RDT}(i)}|i) \tag{A.12}$$

Contradiction: $P(Y_{x^*_{RDT}(i)}|i)$ is, by definition, the largest reward possible under each intent, meaning that even if x^*_{CDT} is the maximizing arm in all intent conditions, $W_{CDT} = W_{RDT}$, and $W_{CDT} \not> W_{RDT}$. Thus, we are guaranteed that $W_{CDT} \not> W_{RDT}$ and so $R^u_t(CDT) \not< R^u_t(RDT)$. The cases where $R^u_t(CDT) > R^u_t(RDT)$ and $R^u_t(CDT) = R^u_t(RDT)$ are shown in Examples 5.1.1 and 6.1.1, respectively.

□

APPENDIX B

Supplementary Material for Chapter 4

Cue Word	Strong Association Target	Weak Association Target
accelerate	speed: 0.386	gas: 0.029
adjective	noun: 0.333	English: 0.043
bell	ring: 0.399	school: 0.041
bandage	cut: 0.331	hurt: 0.071
cow	milk: 0.352	pasture: 0.042
chair	table: 0.314	sofa: 0.077
dig	shovel: 0.32	grave: 0.031
dart	board: 0.358	throw: 0.081
extinct	dinosaur: 0.32	animal: 0.033
enrage	mad: 0.304	temper: 0.014
frost	cold: 0.37	jack: 0.036
fur	coat: 0.324	warm: 0.047
gain	weight: 0.26	acquire: 0.016
glue	sticky: 0.371	paper: 0.053
hoop	hula: 0.392	earring: 0.039
hand	finger: 0.358	glove: 0.048
injection	needle: 0.331	drug: 0.047
imagine	dream: 0.336	fantasy: 0.075
juggler	circus: 0.362	act: 0.039
jazz	music: 0.367	blues: 0.048
keyboard	piano: 0.355	play: 0.033

knife	fork: 0.327	spoon: 0.051
lobby	hotel: 0.345	lounge: 0.034
lung	breathe: 0.362	smoke: 0.057
mortgage	house: 0.349	bill: 0.024
mansion	house: 0.326	huge: 0.036
nucleus	atom: 0.316	science: 0.053
noise	loud: 0.34	ear: 0.058
outcome	end: 0.31	future: 0.021
orchestra	music: 0.309	conductor: 0.052
peer	friend: 0.325	group: 0.039
picture	frame: 0.316	camera: 0.051
quantity	amount: 0.379	many: 0.043
roof	house: 0.307	tar: 0.024
ray	sun: 0.362	beam: 0.047
scold	yell: 0.32	anger: 0.02
scheme	plan: 0.392	sneaky: 0.028
sailing	boat: 0.359	swim: 0.021
task	job: 0.37	duty: 0.055
thief	steal: 0.388	crook: 0.091
universe	world: 0.385	everything: 0.014
used	old: 0.358	worn: 0.061
virus	sick: 0.351	germ: 0.026
visitor	guest: 0.365	relative: 0.061
wrist	watch: 0.345	bracelet: 0.061
weird	strange: 0.312	normal: 0.049
yummy	good: 0.34	sweet: 0.02
year	month: 0.321	annual: 0.045
zero	none: 0.338	number: 0.065
zucchini	vegetable: 0.331	broccoli: 0.034

Table B.1: List of quiz questions in the human-subjects RCT experiment.

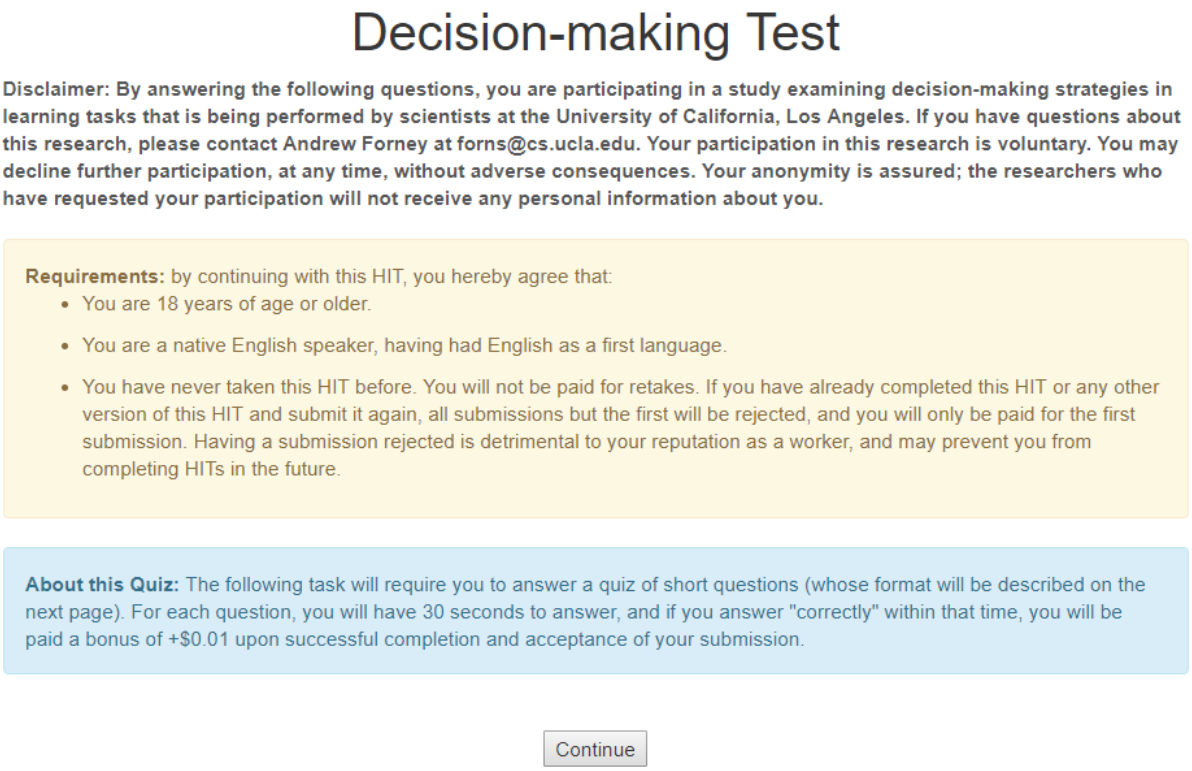

Figure B.1: Image of the informed consent screen presented to participants before beginning the quiz.

APPENDIX C

Supplementary Material for Chapter 6

Theorem 6.2.1 (Heterogeneous Intent Empirical Counterfactual Estimation)**.** Let X be a decision variable in a heterogeneous intent SDM M^{Π_A} (Def. 6.2.3) with measured outcome Y, and let $I^{A_1}, ..., I^{A_a}$ be the heterogeneous intents for X of actors in the IECs $A = \{A_1, ..., A_a\}$ in M^{Π_A}. A HI-specific outcome quantity $P(Y_{x'}|I^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a})$ is equivalent to a counterfactual for a single IEC $A_j \in A$, $P(Y_{x'}|X^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a})$, and can be estimated empirically for observed intents $i^{A_1}, ..., i^{A_a}$, and antecedent $X = x'$ (where $X = x'$ indicates the antecedent for *any* of the individual IEC SDMs as well as the HI-SDM since, by assumption, $do(X^{A_j}) = do(X^{A_i})$ for any two IECs $A_i \neq A_j$). Formally, we may write the counterfactual query in interventional notation such that

$$P(Y_{X=x'}|X^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a}) = P(Y_{X=x'}|I^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a}) \tag{6.3}$$

$$= P(Y|do(X = x'), I^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a}) \tag{6.4}$$

Proof. This proof employs the causal assumptions implicit in the prototypical HI-SDM, depicted graphically in Figure 6.1 with each HI-SDM's constituent individual IEC SDMs. The proof for the empirical estimation of heterogeneous intent follows from the analogous one for empirical counterfactual estimation for homogeneous intent. Recall from the theorem statement that $A_j \in A$ is a single IEC. Also, we note that by definition of a HI-SDM (Def. 6.2.3), $do(X^{A_j} = x^{A_j})$ is considered an equivalent intervention to some other IEC $A_s \in A$, meaning: $do(X^{A_j} = x^{A_j}) = do(X^{A_s} = x^{A_s}) = do(X = x)$. We begin by writing our counterfactual query $P(Y_{X=x'}|X^{A_j} = i^{A_j}, ..., I^{A_a} = i^{A_a})$ and then demonstrate that it can be

written in strictly interventional notation.

$$P(Y_{X=x'}|X^{A_j}=i^{A_j},...,I^{A_a}=i^{A_a})) \tag{C.1}$$

$$= \sum_{i^{A'_j}\in I^{A_j}} P(Y_{X=x'}|X^{A_j}=i^{A_j}, I^{A_j}=i^{A_j}...,I^{A_a}=i^{A_a})P(I^{A_j}=i^{A'_j}|X^{A_j}=i^{A_j}, I^{A\setminus A_j}=i^{A\setminus A_j}) \tag{C.2}$$

$$= \sum_{i^{A'_j}\in I^{A_j}} P(Y_{X=x'}|I^{A_j}=i^{A_j}...,I^{A_a}=i^{A_a})P(I^{A_j}=i^{A'_j}|X^{A_j}=i^{A_j}, I^{A\setminus A_j}=i^{A\setminus A_j}) \tag{C.3}$$

$$= \sum_{i^{A'_j}\in I^{A_j}} P(Y_{X=x'}|I_x^{A_j}=i^{A_j}...,I_x^{A_a}=i^{A_a})P(I^{A_j}=i^{A'_j}|X^{A_j}=i^{A_j}, I^{A\setminus A_j}=i^{A\setminus A_j}) \tag{C.4}$$

$$= \sum_{i^{A'_j}\in I^{A_j}} P(Y|do(X=x'), I^{A_j}=i^{A_j}...,I^{A_a}=i^{A_a})P(I^{A_j}=i^{A'_j}|X^{A_j}=i^{A_j}, I^{A\setminus A_j}=i^{A\setminus A_j}) \tag{C.5}$$

$$= \sum_{i^{A'_j}\in I^{A_j}} P(Y_{X=x'}|I^{A_j}=i^{A_j}...,I^{A_a}=i^{A_a})1(i^{A'_j}=i^{A_j}) \tag{C.6}$$

$$= P(Y|do(X=x'), I^{A_j}=i^{A_j}...,I^{A_a}=i^{A_a}) \tag{C.7}$$

Eq. (C.2) expands the counterfactual using the law of total probability to sum over the intent conditions of A_j. Eq. (C.3) follows from the conditional independence $Y_{X=x} \perp\!\!\!\perp X^{A_j}|I^{A_j}$ that holds in $M^{\Pi_{A_j}}$, allowing us to remove $X^{A_j}=i^{A_j}$. Eq. (C.4) follows because $I_x^{A_j} = I^{A_j}$ given that $(I^{A_j} \perp\!\!\!\perp X)_{G_x}$, where G_x is the interventional submodel where all causal parents of X are severed (which can be considered for either $M^{\Pi_{A_j}}$ individually, or the HI-SDM M^{Π_A}, as represented by the counterfactual antecedent notation). Eq. (C.5) is a notational re-arranging because all variables (Y_x and I_x^A) are in terms of the interventional submodel M_x (and thus G_x), licensing us to express the quantity using the $do(x)$ notation. Eqs. (C.6, C.7) follow from the fact that, observationally, an agent's final arm choice will always coincide with their intent (i.e., $P(i^{A_j}|x^{A_j},...,i^{A_a}) = 1\ \forall i^A = x^A, 0$ otherwise, regardless of the value of i^{A_a}), which nullifies all summed expressions where the two differ. □

Corollary 6.2.1.1 (Equivalence of Pre- and Post-Assignment Intent Sampling)**.** In an HI-RCT (Def. 6.2.6) with randomly assigned treatment X, measured outcome Y, IECs $A = \{A_1,...,A_n\}$, and intended treatments of actors in each IEC $I^A = \{I^{A_1},...,I^{A_a}\}$, empirical

estimation of IEC-specific treatment outcomes can be accomplished by the tenets of the Heterogeneous Intent Empirical Counterfactual Estimation, Theorem 6.2.1. Because HI-RCTs randomize treatment assignment, IECs that are sampled before treatment assignment yield equivalent information about the assigned treatment's outcome Y_x as do those that are sampled after, or formally:

$$P(Y_x|i^{A_1}, ..., i^{A_a}) = P(Y_x|i_x^{A_1}, ..., i_x^{A_a}) \tag{6.6}$$

Proof. The proof for Theorem 6.2.1.1 follows immediately from the graphical assumptions of an RCT, namely, that treatment X is randomized via the semantics of an intervention $do(X)$. In the canonical HI-RCT depicted in Figure 6.2 (right), we note that $I^{A_j} \perp\!\!\!\perp X \Rightarrow I^{A_j} = I_x^{A_j} \; \forall \; A_j \in A$ by the rules of do-calculus. Therefore:

$$P(Y_x|i^{A_1}, ..., i^{A_a}) = P(Y_x|i_x^{A_1}, ..., i_x^{A_a}) = P(Y|do(x), i^{A_1}, ..., i^{A_a})$$

□

Theorem 6.2.2 (HI-RDT u-regret Reduction is Superior to RDT)**.** Let R_t^u be the cumulative u-regret (Def. 3.2.2) experienced by an agent in a MABUC problem by trial t. If $R_t^u(RDT)$ represents the u-regret experienced by an RDT agent and $R_t^u(HIRDT)$ represents the u-regret experienced by an HI-RDT agent, then as $t \to \infty$, $R_t^u(HIRDT) \leq R_t^u(RDT)$ for all possible MABUC parameterizations.

Proof. This proof also follows from Theorem 3.4.2, in which the u-regret experienced by an RDT agent was shown to be always lesser than or equal to a CDT agent. Consider a heterogeneous-intent MABUC problem (HI-MABUC) in which actors belong to some number of intent equivalence classes $I^A = \{I^{A_1}, ..., I^{A_a}\}$ (IECs, 6.2.4). Note that i-regret 3.4.6 is defined over the counterfactual RDT reward space for a single IEC, as would be maximized by a Regret Decision Theory (RDT, Def. 3.4.5) agent, such that $x^*_{RDT}(i^{A_j}) = \text{argmax}_{x \in X} P(Y_x|i^{A_j})$ for an individual actor belonging to the IEC A_j. By extension, hi-regret is defined over the counterfactual reward space for IECs, as would be experienced by a Heterogeneous-Intent Regret Decision Theory target (HIRDT, Def. 6.2.5), such that

$x^*_{HIRDT}(i^A) = \text{argmax}_{x \in X} P(Y_x|i^A)$ and i^A is a vector of IEC intents $i^A = \{i^{A_1}, ..., i^{A_a}\}$. We can next note the key relationship between RDT and HI-RDT maximization targets:

$$P(Y_x|I^{A_j} = i^{A_j}) = \sum_{i' \in I^{A \setminus A_j}} P(Y_x|I^{A_j} = i^{A_j}, I^{A \setminus A_j} = i')P(I^{A \setminus A_j} = i'|I^{A_j} = i^{A_j}) \tag{C.8}$$

In words, an IEC actor's RDT maximization target is simply a probability-weighted sum over the superset of HI-specific reward targets, and so HI-RDT agents record strictly more information than RDT agents do. This implies that any RDT quantity can be computed from adequately sampled HI-RDT quantities.

To show that $R^u_t(RDT) < R^u_t(HIRDT)$ it is sufficient to show that the rewards collected by any RDT-maximizing agent in a HI-MABUC would be strictly greater than those collected by the HI-RDT agent. Assume to the contrary that this is the case, such that if W_{RDT} represents the expected winnings of any RDT agent A_j and W_{HIRDT} represents the expected winnings of the HI-RDT agent, $W_{RDT} > W_{HIRDT}$. We thus have:

$$W_{RDT} = \sum_{i \in I^{A_j}} P(Y_{x^*_{RDT}(i)}|i)P(i) = \sum_{i' \in I^{A \setminus A_j}} \sum_{i \in I^{A_j}} P(Y_{x^*_{RDT}(i)}|i, i')P(i|i') \tag{C.9}$$

$$W_{HIRDT} = \sum_{i' \in I^{A \setminus A_j}} \sum_{i \in I^{A_j}} P(Y_{x^*_{HIRDT}(i,i')}|i, i')P(i|i') \tag{C.10}$$

In other words, RDT chooses the x that maximizes within-intent reward, summed over the probability-weighted priors of each intent, whereas HI-RDT chooses the x that maximizes between-intent reward, summed over the probability-weighted priors of all IEC intents. However, because in either case, each $P(i|i')$ will be the same, we have:

$$W_{RDT} > W_{HIRDT} \Rightarrow \tag{C.11}$$

$$\sum_{i' \in I^{A \setminus A_j}} \sum_{i \in I^{A_j}} P(Y_{x^*_{RDT}(i)}|i, i')P(i|i') > \sum_{i' \in I^{A \setminus A_j}} \sum_{i \in I^{A_j}} P(Y_{x^*_{HIRDT}(i,i')}|i, i')P(i|i') \tag{C.12}$$

$$\sum_{i' \in I^{A \setminus A_j}} \sum_{i \in I^{A_j}} P(Y_{x^*_{RDT}(i)}|i, i') > \sum_{i' \in I^{A \setminus A_j}} \sum_{i \in I^{A_j}} P(Y_{x^*_{HIRDT}(i,i')}|i, i') \tag{C.13}$$

Contradiction: $P(Y_{x^*_{HIRDT}(i,i')}|i, i')$ is, by definition, the largest reward possible under *all* IEC intents, meaning that even if x^*_{RDT} is the maximizing arm in all IEC intent conditions, $W_{RDT} = W_{HIRDT}$, and $W_{RDT} \not> W_{HIRDT}$. Thus, we are guaranteed that $W_{RDT} \not> W_{HIRDT}$

and so $R_t^u(RDT) \not< R_t^u(HIRDT)$. The case where $R_t^u(RDT) > R_t^u(HIRDT)$ is shown in Example 6.1.1.

□

Theorem 6.2.3 (Sufficiency of hi-regret Minimization for u-regret Minimization)**.** Let $R_t^{i^A}$ be the cumulative hi-regret (Def. 6.2.7) and R_t^u be the cumulative u-regret (Def. 3.2.2) experienced by an HI-SDM agent in a HI-MABUC problem by trial t. As $t \to \infty$, if $R_t^{i^A} = O(1)$ then $R_t^u = O(1)$ if the following equivalence holds:

$$x^*(u_t) = \underset{x \in X}{\operatorname{argmax}} P(y_x|u_t) = \underset{x \in X}{\operatorname{argmax}} P(y_x|i_t^A) = x^*(i_t^A) \; \forall \; u_t \tag{6.10}$$

In words, sub-linear cumulative hi-regret will imply sub-linear cumulative u-regret if the optimal action under known confounder state $U_t = u_t$ is the same as the optimal action under experienced HIs $I^A = i_t^A$ for all trials $t \in T$.

Proof. The conditions under which $R_t^u = O(1)$ are when there exists some t' such that for all trials $t^+ \in [t', T]$, $E[R_{t^+}^u] = 0$. In other words, for some choice policy that converges to the optimal policy after t', the optimal action chosen for all t^+ will be $x(t^+) = x^*(u_{t^+})$. Were $x(t^+) \neq x^*(u_{t^+})$, then there would be some $\epsilon = x^*(u_{t^+}) - x(t^+)$ over which $R_{t^+}^u = \sum_{t \in [t', T]} \epsilon \neq 0$. As such, if $x^*(u_{t^+}) = x^*(i_{t^+})$, then $x(t^+) = x^*(i_{t^+}^A) = x^*(u_{t^+}) \Rightarrow E[R_{t^+}^u] = 0$. □

[AA06] Melbourne Academic Mindfulness Interest Group and Melbourne Academic Mindfulness Interest Group. "Mindfulness-based psychotherapies: a review of conceptual foundations, empirical evidence and practical considerations." *Australian and New Zealand Journal of Psychiatry*, **40**(4):285–294, 2006.

[ACF95] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. "Gambling in a rigged casino: The adversarial multi-armed bandit problem." In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pp. 322–331, Oct 1995.

[ACF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem." *Mach. Learn.*, **47**(2-3):235–256, May 2002.

[ACF03] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. "The Nonstochastic Multiarmed Bandit Problem." *SIAM J. Comput.*, **32**(1):48–77, January 2003.

[AG11] S. Agrawal and N. Goyal. "Analysis of Thompson Sampling for the multi-armed bandit problem." *CoRR*, **abs/1111.1797**, 2011.

[Ahm14] A. Ahmed. *Evidence, Decision and Causality.* Cambridge University Press, 2014.

[AMS09] J. Y. Audibert, R. Munos, and C. Szepesvri. "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits." *Theoretical Computer Science*, **410**(19):1876 – 1902, 2009. Algorithmic Learning Theory.

[AN04] P. Abbeel and A. Y. Ng. "Apprenticeship learning via inverse reinforcement learning." In *Proceedings of the twenty-first international conference on Machine learning*, p. 1. ACM, 2004.

[BC12] S. Bubeck and N. Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems." *Foundations and Trends in Machine Learning*, **5**:1–122, 2012.

[BFP15] E. Bareinboim, A. Forney, and J. Pearl. "Bandits with unobserved confounders: A causal approach." In *Advances in Neural Information Processing Systems*, pp. 1342–1350, 2015.

[BGA05] M. G. Beldarrain, J. C. Garcia-Monco, E. Astigarraga, A. Gonzalez, and J. Grafman. "Only spontaneous counterfactual thinking is impaired in patients with prefrontal cortex lesions." *Cognitive Brain Research*, **24**(3):723–726, 2005.

[BK10] R. Busa-Fekete and B. Kégl. "Fast boosting using adversarial bandits." In T. Joachims J. Fürnkranz, editor, *27th International Conference on Machine Learning (ICML 2010)*, pp. 143–150, Haifa, Israel, June 2010.

[BP16] E. Bareinboim and J. Pearl. "Causal inference and the data-fusion problem." *Proceedings of the National Academy of Sciences*, **113**(27):7345–7352, 2016.

[Bri17] R. Briggs. "Normative Theories of Rational Choice: Expected Utility." In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[BS12] S. Bubeck and A. Slivkins. "The best of both worlds: stochastic and adversarial bandits." *CoRR*, **abs/1202.4473**, 2012.

[BSG10] M. A. Brookhart, T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss. "Confounding control in healthcare database research: challenges and potential approaches." *Medical care*, **48**(6 0):S114–S120, jun 2010.

[Byr16] R. M. J. Byrne. "Counterfactual thought." *Annual review of psychology*, **67**:135–157, 2016.

[CL11] O. Chapelle and L. Li. "An Empirical Evaluation of Thompson Sampling." In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pp. 2249–2257. Curran Associates, Inc., 2011.

[Cou13] National Research Council et al. *Frontiers in massive data analysis*. National Academies Press, 2013.

[DHK11] M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. "Efficient Optimal Learning for Contextual Bandits." *CoRR*, **abs/1106.2369**, 2011.

[Doy88] J. Doyle. "Knowledge, Representation, and Rational Self-Government." *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 345–354, 01 1988.

[DSM15] F. De Brigard, R. N. Spreng, J. P. Mitchell, and D. L. Schacter. "Neural activity associated with self, other, and object-based counterfactual thinking." *Neuroimage*, **109**:12–26, 2015.

[Ebb13] H. Ebbinghaus. *On memory: A contribution to experimental psychology.* Teachers College, 1913.

[ELH15] T. Everitt, J. Leike, and M. Hutter. "Sequential Extensions of Causal and Evidential Decision Theory." *International Conference on Algorithmic DecisionTheory*, 2015.

[Els99] A. S. Elstein. "Heuristics and biases: selected errors in clinical reasoning." *Academic Medicine*, **74**(7):791–4, 1999.

[EMM06] E. Even-Dar, S. Mannor, and Y. Mansour. "Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems." *J. Mach. Learn. Res.*, **7**:1079–1105, December 2006.

[ER08] K. Epstude and N. J. Roese. "The functional theory of counterfactual thinking." *Personality and Social Psychology Review*, **12**(2):168–192, 2008.

[Far73] J. L. Farr. "Response requirements and primacy-recency effects in a simulated selection interview." *Journal of Applied Psychology*, **57**(3):228, 1973.

[FBP] A. Forney, E. Bareinboim, and J. Pearl. "Counterfactual Randomization for Clinical Trials." in prep.

[FGS13] D. Ferrante, V. Girotto, M. Stragà, and C. Walsh. "Improving the past and the future: A temporal asymmetry in hypothetical thinking." *Journal of Experimental Psychology: General*, **142**(1):23, 2013.

[Fis51] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 6th edition, 1951.

[FPB17] A. Forney, J. Pearl, and E. Bareinboim. "Counterfactual Data-Fusion for Online Reinforcement Learners." In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1156–1164, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[FWB] A. Forney, C. Willey, E. Bareinboim, and J. Pearl. "Regret as a Counterfactual Learning Mechanism in Human Decision-Making." in prep.

[FY75] J. L. Farr and M. C. York. "Amount of Information and Primacy-Recency Effects in Recruitment Decisions." *Personnel Psychology*, **28**(2):233–238, 1975.

[GCP07] A. R. Green, D. R. Carney, D. J. Pallin, L. H. Ngo, K. L. Raymond, L. I. Iezzoni, and M. R. Banaji. "Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients." *Journal of General Internal Medicine*, **22**(9):1231–1238, Sep 2007.

[HC02] C. B. Holroyd and M. G. H. Coles. "The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity." *Psychological review*, **109**(4):679, 2002.

[Hec92] J. J. Heckman. "Randomization and social policy evaluation." In C. Mansi and I. Garfinkle, editors, *Evaluations: Welfare and Training Programs*, pp. 201–230. Harvard University Press, Cambridge, MA, 1992.

[HG97] S. Highhouse and A. Gallo. "Order Effects in Personnel Decision Making." *Human Performance*, **10**(1):31–46, mar 1997.

[HLM16] M. K. Ho, M. Littman, J. MacGlashan, F. Cushman, J. Austerweil, and J. L. Austerweil. "Showing versus doing: Teaching by demonstration." In *Advances In Neural Information Processing Systems*, pp. 3027–3035, 2016.

[Kah11] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus, and Giroux, 2011.

[KK09] D. Kahneman and G. Klein. "Conditions for intuitive expertise: a failure to disagree." *American psychologist*, **64**(6):515, 2009.

[KLL17] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, D. Abrams, M. Alsdorf, M. Cohen, A. Crohn, G. R. Cusick, T. Dierks, J. Donohue, M. Dupont, M. Egan, E. Glazer, J. Gottschall, N. Hess, K. Kane, L. Kellam, A. Lascala-Gruenewald, C. Loeffler, A. Milgram, L. Raphael, C. Rohlfs, D. Rosenbaum, T. Salo, A. Shleifer, A. Sojourner, J. Sowerby, C. Sunstein, M. Sviridoff, E. Turner, and J. John. "Human Decisions and Machine Predictions." NBER Working Paper Series Human Decisions and Machine Predictions, 2017.

[LCL10] L. Li, W. Chu, J. Langford, and R. E. Schapire. "A Contextual-bandit Approach to Personalized News Article Recommendation." In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. ACM.

[LR85] T. L. Lai and H. Robbins. "Asymptotically efficient adaptive allocation rules." *Advances in Applied Mathematics*, **6**(1):4 – 22, 1985.

[Lyl02] A. Lyles. "Direct Marketing of Pharmaceuticals to Consumers." *Annual Review of Public Health*, **23**(1):73–91, may 2002.

[LZ08] J. Langford and T. Zhang. "The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information." In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pp. 817–824. Curran Associates, Inc., 2008.

[McC08] S. M. McCrea. "Self-handicapping, excuse making, and counterfactual thinking: consequences for self-esteem and future motivation." *Journal of personality and social psychology*, **95**(2):274, 2008.

[MD05] D. R. Mandel and M. K. Dhami. "What I did versus what I might have done: Effect of factual versus counterfactual thinking on blame, guilt, and shame in prisoners." *Journal of Experimental Social Psychology*, **41**(6):627–635, 2005.

[MDD06] S. Milberger, R. M. Davis, C. E. Douglas, J. K. Beasley, D. Burns, T. Houston, and D. Shopland. "Tobacco manufacturers' defence against plaintiffs' claims of cancer causation: throwing mud at the wall and hoping some of it will stick." *Tobacco control*, **15 Suppl 4**(Suppl 4):iv17–26, dec 2006.

[Men14] X. Meng. "A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it)." *Past, Present, and Future of Statistical Science*, 2014.

[MM14] S. L. Mullally and E. A. Maguire. "Counterfactual thinking in patients with amnesia." *Hippocampus*, **24**(11):1261–1266, 2014.

[MMG93] C. N. Macrae, A. B. Milne, and R. J. Griffiths. "Counterfactual thinking and the perception of criminal behaviour." *British Journal of Psychology*, **84**(2):221–226, 1993.

[Mor14] A. Morris. "TurkSuite Template Generator.", 2014. http://mturk.mit.edu/template.php.

[NMS04] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. "The University of South Florida free association, rhyme, and word fragment norms." *Behavior Research Methods, Instruments, & Computers*, **36**(3):402–407, 2004.

[Noz69] R. Nozick. "Newcombs problem and two principles of choice." In *Essays in honor of Carl G. Hempel*, pp. 114–146. Springer, 1969.

[OB10] P. A. Ortega and D. A. Braun. "A Minimum Relative Entropy Principle for Learning and Acting." *J. Artif. Int. Res.*, **38**(1):475–511, May 2010.

[Pea95] J. Pearl. "Causal diagrams for empirical research." *Biometrika*, **82**(4):669–710, 1995.

[Pea98] J. Pearl. "Why there is no statistical test for confounding, why many think there is, and why they are almost right." Technical Report R-256, Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, 1998.

[Pea00] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. Second ed., 2009.

[Pea09] J. Pearl. "Understanding propensity scores." *Causality: models, reasoning, and inference*, pp. 348–352, 2009.

[Pea12] J. Pearl. "The Do-Calculus Revisited." *AUAI Press*, **4**(11), 2012.

[Pea13] J. Pearl. "The Curse of Free-will and the Paradox of Inevitable Regret." *Journal of Causal Inference*, **1**(2):255–257, 2013.

[Pea14] J. Pearl. "Comment: Understanding Simpson's Paradox." *The American Statistician*, **68**(1):8–13, 2014.

[PGJ16] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[PV91] J. Pearl and T. Verma. "A Theory of Inferred Causation." In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441–452. Morgan Kaufmann, San Mateo, CA, 1991.

[Rei56] H. Reichenbach. "The Direction of Time." *Univ. California Press*, 1956.

[Rob52] H. Robbins. "Some aspects of the sequential design of experiments." *Bull. Amer. Math. Soc.*, **58**(5):527–535, 09 1952.

[RR83] P. R. Rosenbaum and D. B. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, **70**(1):41–55, 1983.

[RZI09] J. Ross, A. Zaldivar, L. Irani, and B. Tomlinson. "Who are the turkers? worker demographics in amazon mechanical turk." *Department of Informatics, University of California, Irvine, USA, Tech. Rep*, 2009.

[SB98] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[SBD15] D. L. Schacter, R. G. Benoit, F. De Brigard, and K. K. Szpunar. "Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions." *Neurobiology of learning and memory*, **117**:14–21, 2015.

[SBQ14] C. Symborski, M. Barton, M. Quinn, C. K. Morewedge, K. S. Kassam, and J. H. Korris. "Missing: A Serious Game for the Mitigation of Cognitive Biases." *Interservice/Industry Training*, 2014.

[Sco10] S. L. Scott. "A modern Bayesian look at the multi-armed bandit." *Applied Stochastic Models in Business and Industry*, **26**(6):639–658, 2010.

[SEB05] C. Santamaría, O. Espino, and R. M. J. Byrne. "Counterfactual and semifactual conditionals prime alternative possibilities." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **31**(5):1149, 2005.

[SG14] B. A. Spellman and E. A. Gilbert. "Blame, cause, and counterfactuals: The inextricable link." *Psychological Inquiry*, **25**(2):245–250, 2014.

[Sli14] A. Slivkins. "Contextual Bandits with Similarity Information." *J. Mach. Learn. Res.*, **15**(1):2533–2568, January 2014.

[SM12] R. Smallman and K. C. McCulloch. "Learning from yesterday's mistakes to fix tomorrow's problems: When functional counterfactual thinking and psychological distance collide." *European Journal of Social Psychology*, **42**(3):383–390, 2012.

[SS90] C. W. Savage and M. C. P. Science. *Scientific Theories*. Number v. 14 in Minnesota studies in the philosophy of science. University of Minnesota Press, 1990.

[SSD17] R. Sen, K. Shanmugam, A. Dimakis, and S. Shakkottai. "Identifying Best Interventions through Online Importance Sampling ." *International Conference on Machine Learning*, p. to appear, 2017.

[Sze10] C. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

[TK75] A. Tversky and D. Kahneman. "Judgment under uncertainty: Heuristics and biases." In *Utility, probability, and human decision making*, pp. 141–162. Springer, 1975.

[TKT11] K. H. Teigen, A. B. Kanten, and J. A. Terum. "Going to the other extreme: Counterfactual thinking leads to polarised judgements." *Thinking & Reasoning*, **17**(1):1–29, 2011.

[TMK12] M. P. Tyser, S. M. McCrea, and K. Knüpfer. "Pursuing perfection or pursuing protection? Self-evaluation motives moderate the behavioral consequences of counterfactual thoughts." *European journal of social psychology*, **42**(3):372–382, 2012.

[Ven11] C. L. Ventola. "Direct-to-Consumer Pharmaceutical Advertising: Therapeutic or Toxic?" *P & T : a peer-reviewed journal for formulary management*, **36**(10):669–84, oct 2011.

[Wai89] H. Wainer. "Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions." *Journal of Educational Statistics*, **14**:121–140, 1989.

[Wei16] P. Weirich. "Causal Decision Theory." In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[Whi05] H. D. White. "Adherence and outcomes: it's more than taking the pills." *The Lancet*, **366**(9502):1989–1991, 2005.

[ZB16] J. Zhang and E. Bareinboim. "Markov decision processes with unobserved confounders: A causal approach." Technical report, Technical Report R-23, Purdue AI Lab, 2016.