KNOWLEDGE OF COUNTERFACTUALS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ana Smaranda Sandu

May 2021

KNOWLEDGE OF COUNTERFACTUALS

Ana Smaranda Sandu, Ph.D.

Cornell University 2021

We formalize agents' knowledge of counterfactuals in two different settings, players' behavior in extensive-form games and the process of agents' conditioning their beliefs.

For extensive-form games, we define the notion of subgame-rationality, where players best-respond to what they believe would happen at any subgame where they are due to play. This approach settles a well-known disagreement in the literature between Aumann and Stalnaker - supporting Aumann - regarding whether common knowledge of rationality leads to the backwards induction solution in perfect information games [2, 42, 41]. Subgame-rationality also makes it easier to relate epistemic characterizations of Nash equilibrium to those of subgame-perfect equilibrium. We also turn our attention to adding counterfactuals to agents' language, which leads to definitions of rationality which use iterated counterfactuals.

For conditional beliefs, we propose new public-announcement style semantics which factor out the act of conditioning, using two traditional modalities, beliefs and counterfactuals. We investigate the set of validities for these semantics. We also take a closer look at the relationship between traditional plausibility models for conditional beliefs within Dynamic Epistemic Logic and models which use, instead, explicit counterfactual shifts. We identify properties that counterfactual shifts need to satisfy in order to simulate plausibility models.

BIOGRAPHICAL SKETCH

Smaranda was born in Braila, Romania on June 9th, 1991. She graduated from Wellesley College with a double major in Mathematics and Computer Science in 2014. In 2018 she obtained her Master's in Computer Science from Cornell University. In 2021 she completed her PhD in Mathematics at Cornell University with this thesis. Smaranda has accepted a position as an Instructor in Science Laboratory in the Computer Science Department at Wellesley College in Wellesley, MA.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Anil Nerode, for taking me on, for his support, his optimism, and his unshakeable belief that I could finish even when I lost faith. The incredible breadth and depth of his mathematical knowledge made each of our conversations surprising and insightful.

I am deeply indebted to Adam Bjorndahl, without whom none of this would have happened. I started working with Adam around the time when I was starting to really doubt my ability to finish the PhD. I am grateful for his patience, advice, and understanding, but more than any of that, I am grateful for the generosity with which he shared his mathematical insight, the interest he showed in our conversations, and the amazing skill with which he extracted meaning from my confused questions.

I am very grateful to Joseph Halpern (who introduced me to this incredibly fun and vibrant research area), Michael Stillman and Richard Shore for their support and guidance.

Thank you, Noah Lloyd, for being an amazing pillar of steady and kind support, making me laugh and taking me on walks when the stress got too much to handle, and helping me feel grounded when so much seemed to make no sense. I am so fortunate to have had you by my side during these demanding times.

I also have to thank all the friends I was so lucky to make at Cornell, who have supported me and enriched my life in so many ways: from climbing sessions to - sometimes day-long - baking adventures, from working and chatting at Gimme Coffee to checking in on each other in Zoom-land, from hiking in the ridiculously beautiful Ithaca area to having hard and honest conversations about grad school, academia, teaching, and life in general. I didn't want to include any names due to the paralyzing fear I would forget someone, but sincere thanks to all my Cornell people.

Thank you to my Wellesley friends, my Wellesley host family, and my Wellesley

CS and Math professors, who've checked in with me and stayed by my side throughout grad school, have accepted me for who I am, and somehow never lost faith that I could do this.

My absolute deepest gratitude goes out to my family. There are no words (in either English or Romanian) to express how grateful I am for their love and support, which they've bestowed on me so generously throughout my life. Special thanks to my mom and sister, who made me believe that women can be anything they want to be in the world.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 A study of epistemic and counterfactual logics

Epistemic logics are concerned with studying knowledge and belief. This is a work that started with Aristotle, but continues to be an active area of research, as interesting applications into game theory and artificial intelligence are discovered and pursued. Studying these logics allows us to explore how we should define knowledge and belief, what differentiates the two, or how knowledge and belief allow us to characterize other concepts.

Modal logics, able to represent expressions like "it is necessary that..." or "it is possible that...", lend themselves to model knowledge and belief. Traditionally, we obtain an epistemic logic by adding to propositional logic a modal operator K and/or a modal operator B, allowing us to represent statements like $K\varphi$, read as "the agent knows φ", or $B\varphi$, read as "the agent believes φ". This is easily generalized to multiple agents, by including modal operators indexed by the set of agents we are considering. Epistemic logics are generally represented using possible worlds semantics, also known as Kripke models. For example, a statement like $K\varphi$ is modeled as φ holds at all the worlds the agent considers possible (so we can see the tight connection to necessity operators in modal logics [31]).

Our focus in this work starts with epistemic logic. We take a closer look at two of its applications, one in game theory and one in the logic of conditional belief. In both instances, we analyze how epistemic logics interact with logics for counterfactuals. Counterfactuals are conditionals of the form "if A were the case, then B would be the

case", or "if A had been the case, then B would have been the case". Such statements arise naturally when agents consider if doing something else would have left them better off (so in determining rationality in game theory), or when agents evaluate their beliefs if a certain event were to happen (so in the process of conditioning their beliefs).

To see this more clearly, let's look closely at Alice's beliefs as she's reasoning about her move in a game of Rock-Paper-Scissors with Bob. Suppose she is planning to play *Rock* and believes Bob for sure will play *Scissors*[1]. In fact, however, Bob is planning to play *Paper*. We can then ask, what does Alice believe would happen if she were to play *Scissors* instead? On the one hand, these beliefs can be informed by her current beliefs – namely, she believes she is at a world where she plays *Rock* and he plays *Scissors*. In this case, Alice believes that, if she were to switch to *Scissors*, it would fare worse than *Rock* against Bob's *Scissors*. Note that there is an important assumption we make here, namely, as Alice is engaging in this mental exercise of considering a different move, this doesn't affect Bob whatsoever, so he's still playing *Scissors* in this hypothetical scenario Alice is envisioning. This is a non-trivial but commonly taken assumption in game theory, also called "opaqueness" or "opacity" [29].

On the other hand, Alice's beliefs can be informed by the current state of the world: there is an actual state of the matter – she is at a world where she plays *Rock* and he plays *Paper* – and it can be that, in reality, if she were to play *Scissors*, Bob might detect a tell that she has when considering *Scissors*, so he would be playing *Rock* at that world. This definition for Alice's beliefs if she were to switch clearly leads to different beliefs than the ones above. Further, there is a subtle point to note about this second definition. In general, agents can have false beliefs about the world – like Alice believing Bob with play *Scissors* – and they might not even consider the actual world possible. Then,

[1]We make the assumption that if a player is planning to do a certain move, they will in fact do so.

this second definition for Alice's beliefs if she were to switch implies she can somehow identify (or access) the actual world as well as what would happen if something switched in the actual world. This doesn't seem reasonable, and we will delve deeply into this issue in Chapter 2. We can see the two different possible beliefs in Figure 1.1.

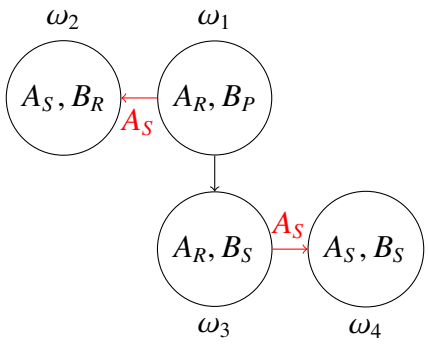

Figure 1.1: Possible counterfactual beliefs

Counterfactuals also arise if we consider conditional beliefs. Intuitively, these are beliefs that agents hold, conditional on other events happening. Suppose Alice is not sure what Bob will do in a round of Rock-Paper-Scissors, so she considers both him playing *Rock* and playing *Paper* possible. Conditional on him playing *Paper*, she believes he does not believe she's playing *Scissors* – for, otherwise, it wouldn't be rational for him to play *Paper* [2]. This act of conditionalization of one's beliefs, however, can easily be reinterpreted as, "if Bob were to play *Paper*, then Alice would believe he doesn't believe she plays *Scissors*."

Studying the interaction between beliefs and counterfactuals in these two settings leads, we argue, to a better understanding of these two concepts, namely, rationality and conditionalization.

[2] We turn to formalizing rationality in a later chapter but, for now, think of rationality as best-responding to your beliefs about what everyone else is doing (the classical game theory definition).

1.2 Fundamentals

In this section we go into technical detail in the formal structures we'll be using throughout this work, including important references for existing and related work. We start with modal logics, and then restrict our attention to a certain modal logic with an operator for knowledge; we describe the traditional Kripke semantics for such models. Further, since the interactions between beliefs and counterfactuals fuel many of the questions, we then turn to an exposition on counterfactual logics.

1.2.1 Modal logic

Modal logics have been studied for a long time, but the first formal approach appears to be the work of C.I. Lewis, at the beginning of the 20th century. For a more thorough discussion on modal logic, see [9, 15, 24]. There's applications of modal logic in fields like computer science, game theory, linguistics, or philosophy. Over the years, the family of modal logics has expanded, to include logics like temporal logics ([20]), logics for knowledge and belief [31], or dynamic logics [30, 46, 34].

Modal logic is an extension of classical propositional logic; by adding new "modal" operators, the language can capture concepts like possibility or necessity. The standard operator that is added is '$\Box$', and it stands for "it is necessary that". The operator $\Diamond$ (which stands for "it is possible that") is the dual of $\Box$, and is defined as $\Diamond\varphi = \neg\Box\neg\varphi$. There are many other possible interpretations for $\Box$, lending to multiple applications of modal logics, like "it will always be true that", or "the agent knows that".

Then we can define the language of modal logic. We start with a countable set of primitive propositions, $PROP$. We define the language $\mathcal{L}_{\Box}$ recursively, given by the

Backus-Naur form:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi,$$

where $p \in PROP$.

The most familiar logic for these statements results from adding to the axioms of propositional logic the distribution axiom (also known as K):

$$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$$

and adding to Modus Ponens, the inference rule for propositional logic, the Necessitation Rule:

From φ infer $\Box\varphi$.

The resulting logic (i.e. the set of formulas provable in this system) is called **K**. It is too weak to validate $\Box\varphi \to \varphi$. This is traditionally called the *knowledge axiom* since, if we were to interpret $\Box$ as knowledge, we might want an axiom that says that we don't know falsehoods (so we don't know φ unless it is true). By adding the knowledge axiom (also called the axiom scheme T) to **K** we obtain the logic denoted by **T**.

Note that these logics have no axioms formalizing iterating either the $\Box$ or $\Diamond$ operators. There's two particular axioms that are usually considered, traditionally called (4) and (5):

(4) $\Box\varphi \to \Box\Box\varphi$

(5) $\Diamond\varphi \to \Box\Diamond\varphi$

Adding axiom (4), also called the axiom of positive introspection, to **T** leads to the system **S4**. Note that in this system $\Box\varphi$ is equivalent to $\Box\Box\varphi$. Finally, adding axiom (5), also called axiom of negative introspection, to **S4** leads to the system **S5**.

1.2.2 Kripke models

Possible-world semantics are the most widespread semantics for modal logics, developed by Kripke [33]. The models are known as *Kripke models* or *Kripke structures*. Each world in the model constitutes a certain way the reality can be, and a binary relation embedded in the model reflects what is considered possible at each world.

Recently, modal logics have been used to interpret the epistemic notions of knowledge and belief. Hintikka developed this field, which has now become a very active area of research, with multiple applications [31]. For a more thorough introduction, some popular references are Fagin, Halpern, Moses and Vardi's foundational work (which also connects epistemic logic with computer science) "Reasoning about Knowledge" [22], Ditmarsch, Halpern, Hoek and Kooi's "An Introduction to Logics of Knowledge and Belief" [18], or Stanford Encyclopedia of Philosophy's entry on epistemic logic [39].

To see Kripke structures in action, we define these models for a modal logic where we use $\Box$ to interpret knowledge. Taking $\mathcal{L}_\Box$ as inspiration, we define the language $\mathcal{L}_K$ recursively, given by the Backus-Naur form:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid K\varphi,$$

where $p \in PROP$ and we read $K\varphi$ as "the agent knows φ". In order to define semantics associated with this language, we will use Kripke structures.

Definition 1. *Define a* Kripke structure *M over $PROP$ as a tuple $M = (W, V, R)$, where*

- *W is a set of possible worlds (also called "states"),*
- *V is a valuation (or interpretation) function, specifying, for each primitive proposition, the set of worlds at which it's true (so $V : PROP \to 2^W$)*

- *R is a function $R : W \to 2^W$, yielding a binary relation on W.*

We refer to R as a possibility, or accessibility relation according to the agent: if $\omega' \in R(\omega)$ (also written $\omega R \omega'$ or $R\omega\omega'$) we say ω' is *considered possible* or *accessible* at ω.

Consider a model $M = (W, V, R)$. We define what it means for a formula in $\mathcal{L}_K$ to be true at (M, ω), written $(M, \omega) \models \varphi$ inductively below:

$$(M, \omega) \models p \text{ iff } \omega \in V(p)$$

$$(M, \omega) \models \neg\varphi \text{ iff } \omega \not\models \varphi$$

$$(M, \omega) \models \varphi \wedge \psi \text{ iff } \omega \models \varphi \text{ and } \omega \models \psi$$

$$(M, \omega) \models K\varphi \text{ iff } \omega' \models \varphi \text{ for all } \omega' \in R(\omega).$$

Note that we model knowledge as "true in all the worlds considered possible". Kripke structures have a very intuitive visualization, since they can be viewed as *labeled directed graphs*. To see this, consider the following example. Suppose Matilda lives in Ithaca and is planning to go on a walk down Cascadilla Gorge. She's trying to best prepare for the weather. Denote by p the statement "it's raining" and by q the statement "it's windy". For the purposes of this example, as we can see in Figure 1.2, there's four possible ways the world can be.

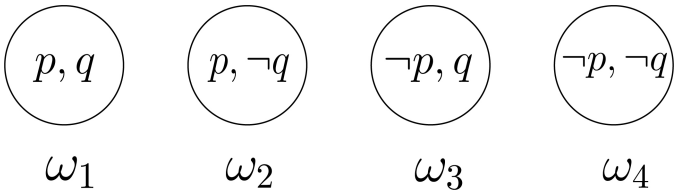

Figure 1.2: Possible worlds

Suppose it's actually raining and it's not windy. We can represent this by highlighting the actual world, as in Figure 1.3.

Now, in order to complete the model, we need to specify the relation R, which we used to model what Matilda considers possible. If $(\omega_2, \omega_1) \in R$ we usually represent this with an arrow from ω_2 to ω_1, read as "at ω_2, the agent considers ω_1 possible". We can see this in the Figure 1.3.

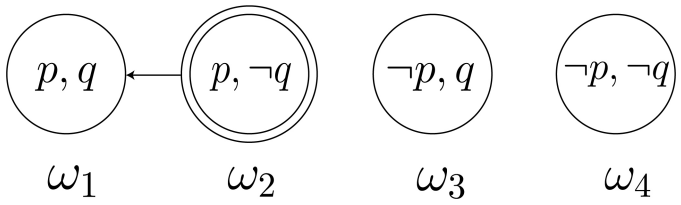

Figure 1.3: A relation R

Note that, in general, we don't have to impose any constraints on R, but if we are trying to use it to model knowledge, there are some restrictions needed. For example, in Figure 1.3, there's no outward edges at ω_1, so Matilda vacuously knows everything.

Then, by taking a look at certain restrictions for R, we can specify certain classes of Kripke models that model knowledge.

We say R is **reflexive** iff for all $\omega \in W$ we have $\omega \in R(\omega)$. In other words, at each world, the agent doesn't rule out that world. We depict one such relation in Figure 1.4. Further, note that if R is reflexive, then $K\varphi \to \varphi$ holds at every world.

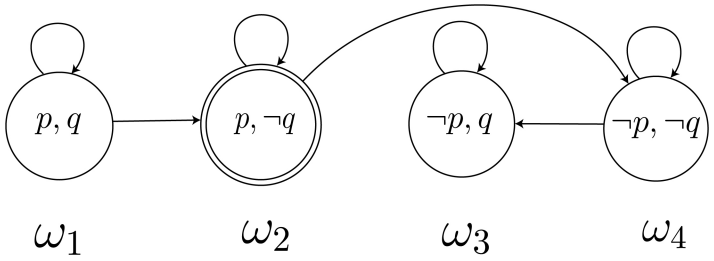

Figure 1.4: A reflexive relation R

We say R is **transitive** iff for all $w, v, u \in W$, if wRv and vRu, then wRu. Clearly, the empty relation is transitive, but that isn't very interesting. We depict another possible

transitive relation on W in Figure 1.5. Note that if R is transitive, $K\varphi \to KK\varphi$ (positive introspection) holds at every world.

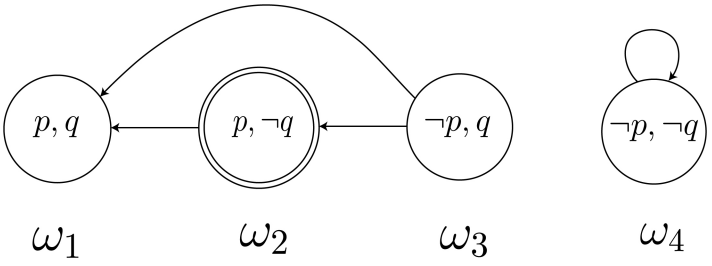

Figure 1.5: A transitive relation R

We say R is **Euclidean** iff for all $w, v, u \in W$, if wRv and wRu, then vRu. We depict a possible Euclidean relation in Figure 1.6. Note that if R is Euclidean, we have $\neg K\varphi \to K\neg K\varphi$ (positive introspection) holds at every world.

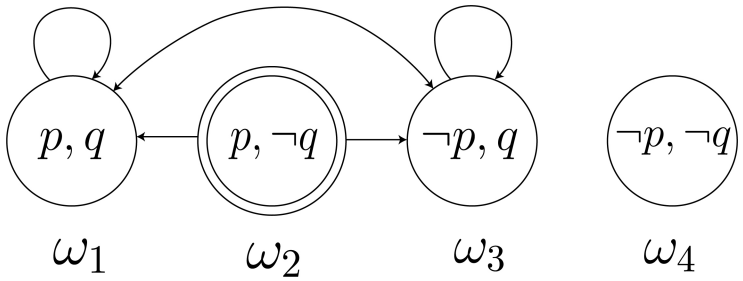

Figure 1.6: A symmetric relation R

Note that when a relation is reflexive, transitive, and Euclidean, then it's an equivalence relation. In this setting, we saw that we have axioms T, (4), and (5) valid – this is the system **S5**, the most popular model for knowledge. Notably, the relationship between the restrictions on the model and the types of axioms satisfied is basically the relationship between the *semantic* and *proof-theoretic* ways of defining a logic.

If S is a set of axioms and inference rules, we write $S \vdash \varphi$ to represent that φ can be deductively proven from S. The resulting logic is $\{\varphi \in \mathcal{L} : S \vdash \varphi\}$. On the other hand, if a formula φ is valid in every model M (written $M \models \varphi$) in a certain class of models $\mathcal{M}$,

we can define the set of formulas valid on the class of models to be $\{\varphi \in \mathcal{L} : \mathcal{M} \models \varphi\}$. An axiom system S is **sound** for a language $\mathcal{L}$ with respect to a class of models $\mathcal{M}$ if every formula provable in S is valid with respect to $\mathcal{M}$. An axiom system S is **complete** for a language $\mathcal{L}$ with respect to a class of models $\mathcal{M}$ if every formula in the language $\mathcal{L}$ that is valid with respect to $\mathcal{M}$ is provable in S. We say the axiom system characterizes the class of models if it provides a sound and complete axiomatization for the language with respect to that class of models (so $\{\varphi \in \mathcal{L} : S \vdash \varphi\} = \{\varphi \in \mathcal{L} : \mathcal{M} \models \varphi\}$).

Then we say **S5** provides a sound and complete axiomatization for $\mathcal{L}_K$ with respect to the class of models $M = (W, R, V)$ where R is an equivalence relation.

There are other properties of interest if we model belief instead of knowledge. Denote the language of belief by $\mathcal{L}_B$ where K is replaced by B and we read $B\varphi$ as "the agent believes φ". Semantics are defined the same way. Intuitively, beliefs can be false (so believing φ shouldn't imply φ), but we shouldn't believe falsehoods. There's different restrictions on R that we can impose.

We say R is **serial** iff for all $\omega \in W$ there exists a v such that $v \in R(\omega)$. Seriality implies $\Box\varphi \rightarrow \Diamond\varphi$, or, equivalently, $\neg\Box(false)$, axiom also denoted by D. When R is serial, transitive, and Euclidean, axioms K, D, (4), and (5) hold. In fact, **KD45** provides a sound and complete axiomatization of $\mathcal{L}_B$ with respect to the class of models $M = (W, R, V)$ where R is serial, transitive, and Euclidean.

1.2.3 Counterfactuals

Counterfactuals were introduced by C.I. Lewis. They are statements of the form "if A were the case, then B would be the case", or "if A had been the case, then B would

have been the case". The simple yet powerful characteristic that differentiates these statements from traditional conditional statements is that they can cover situations when the antecedent is false. This lends to their application in multiple fields, from philosophy to artificial intelligence, from decision theory to law. The most popular semantics were developed by Lewis [37] and Stalnaker [41]. For a more detailed exposition on counterfactuals, see [44].

Strict and similarity semantics for counterfactuals

The most popular semantics are similarity analyses and strict conditional analyses, usually stated in the possible worlds semantics for Kripke models that we have just introduced.

Suppose we have the language given by the following Backus-Naur form:

$$\varphi : p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \rightarrow \psi \mid \Box\varphi \mid \varphi \mathrel{\Box\!\!\rightarrow} \psi$$

where we read $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ as "if φ were the case, then ψ would be the case".

Semantics are given by the same kind of models $M = (W, R, V)$ used to interpret modal logic with one $\Box$ operator (so $\Box\varphi$ is interpreted using R, as φ holding in all R-accessible worlds). First, let's consider *strict analysis* semantics. The strict analysis interprets $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ as $\Box(\varphi \rightarrow \psi)$, so all R-accessible φ-worlds are ψ-worlds.

While this semantics relates counterfactuals to material conditionals, it fails to capture an important property for counterfactuals. For material conditionals (i.e. statements like "If A, then B"), the statement "if today is a day in January, then it snows in Ithaca" holding implies "if today is January 20, then it snows in Ithaca" holds; this property is known as antecedent monotonicity. For counterfactuals, the situation is more complicated. Consider the following example from [44], highlighted by Goodman:

(1) If I had struck this match, it would have lit.

(2) If I had struck this match and done so in a room without oxygen, it would have lit.

While the first, intuitively seems true, the second does not. Strict analysis semantics for counterfactuals, however, validates this principle. If $\varphi \boxrightarrow \psi$ holds, so must $\varphi \wedge \chi \boxrightarrow \psi$, since $[\![\varphi \wedge \chi]\!] \subseteq [\![\varphi]\!]$. By contrast, a type of semantics that does not - and is indeed more popular - is *similarity analysis*.

In a similarity analysis, $\varphi \boxrightarrow \psi$ is true at a world ω iff all φ-worlds most similar to ω are ψ-worlds. We can see, then, that models need to include another relation which models similarity. This can be done through a system of nested accessibility spheres, an ordering on worlds, or using a counterfactual shift function.

This semantics adds to models (W, R, V), a selection function f (also called a counterfactual shift function) which takes a world ω, a subset A of W, and returns the set of A-worlds most similar to ω. Formally, we have $f : W \times \mathcal{P}(W)^+ \to \mathcal{P}(W)^+$ (where $\mathcal{P}(W)^+ = \mathcal{P}(W) - \{\emptyset\}$), but in certain cases, subsets of the domain are chosen (e.g. if we are modeling an extensive-form game, we can choose a counterfactual shift which can only select the closest h-worlds, for any h a history in the game). In this semantics we have $(M, \omega) \models \varphi \boxrightarrow \psi$ iff $f(\omega, [\![\varphi]\!]_M) \subseteq [\![\psi]\!]_M$, where we define $[\![\varphi]\!]_M = \{\omega \in W : (M, \omega) \models \varphi\}$ (and we forego the subscript M when the model is clear).

Looking back on antecedent monotonicity, note that, since the most similar φ-worlds might not be the most similar $\varphi \wedge \psi$-worlds, if $\varphi \boxrightarrow \psi$ holds, this says nothing on whether $\varphi \wedge \chi \boxrightarrow \psi$. So the principle is not valid, in general, in this interpretation.

We continue by focusing on the similarity analysis, as this is also the semantics we use for counterfactuals in out models.

The logic generated by a similarity analysis depends on the restraints imposed on the counterfactual shift. Consider the following four constraints below, where $\omega \in W$, $A \subseteq W$:

1. $f(\omega, A) \subseteq A$ (success)
2. $f(\omega, A) = \{\omega\}$ if $\omega \in A$ (strong centering)
3. $f(\omega, A) \subseteq B$ and $f(\omega, B) \subseteq A$ implies $f(\omega, A) = f(\omega, B)$ (uniformity)
4. $f(\omega, A)$ contains at most one world (uniqueness)

Lewis adopts the first three constraints, while Stalnaker all four. Note that an axiom that we might want, namely conditional excluded middle ($\models \varphi \boxrightarrow \psi \vee \varphi \boxrightarrow \neg\psi$) holds once we adopt uniqueness, but not before.

The similarity semantics is more popular and seems to provide semantics closer to the use of counterfactuals in our daily lives, but it has some important weaknesses. One of the most important ones is an ambiguity on what "closest" should mean, or what "the most similar" worlds should be. There have been some proposals to make this notion of similarity more precise (e.g. Lewis' system of weights [36]).

Another important property for counterfactuals is called "Rational Monotonicity": $\varphi \boxrightarrow \chi, \neg(\varphi \boxrightarrow \neg\psi) \models (\varphi \wedge \psi) \boxrightarrow \chi$. This doesn't hold, in general, for either semantics, but it turns out to be very closely connected to Rational Monotonicity in Conditional Doxastic Logic (as we will see in a later chapter).

1.3 How the paper is organized

The rest of this paper is organized as follows. In Chapter 2 we delve into how knowledge, belief and counterfactuals interact in the process of defining rationality for extensive form games. The starting point is one of the best known disagreements in the field of epistemic characterizations of subgame perfect equilibria: Aumann claims common knowledge of rationality leads to the equilibrium, Stalnaker that it does not. Halpern provides a model that can represent both perspectives. Our work looks closely at their definitions and intuitions. With a careful eye on the relationship between counterfactuals and epistemic access, we provide a new definition and corresponding proof that ultimately supports Aumann's perspective.

In Chapter 3 we develop new semantics for conditional beliefs that use counterfactuals. We use a public announcement style syntax to "factor out" the act of "conditioning on φ" (or "supposing φ") syntactically, with a new family of modalities, each of which is interpreted as a kind of belief-changing model update. We study the resulting validities for the new semantics.

Chapter 4 contains some experimental work related to the projects in Chapters 2 and 3. In Section 4.1 we investigate the consequences of adding counterfactuals to the language agents use in their reasoning. We discover a direction that, to our knowledge, is new to literature, through which we define rationality in extensive-form games using iterated counterfactuals. In Section 4.2 we take a closer look at the relationship between counterfactual models and plausibility models for conditional beliefs. We identify the restrictions needed to be imposed on the counterfactual shift function that would lead to the same logic for conditional beliefs.

CHAPTER 2

A DEFENSE OF AUMANN'S RATIONALITY

2.1 Introduction

2.1.1 Subgame perfect equilibrium

Extensive form games provide a natural framework for analyzing a wide variety of situations where agents act sequentially. Interestingly, it is the sequential nature of the game that makes the concept of "optimal play" a bit more subtle. In particular, the prospect of "non-credible threats" leads naturally to a notion of *subgame rationality*.

To illustrate this, consider the following "market-entry" game ([38, Example 95.2]), depicted in Figure 2.1. Firm 1 considers entry into a market currently occupied by a monopolist, firm 2. If firm 1 decides to enter, firm 2 can either acquiesce, or fight. Fighting will leave them both worse off, while acquiescing will leave the monopolist worse off, but produce a much better outcome for the entrant.

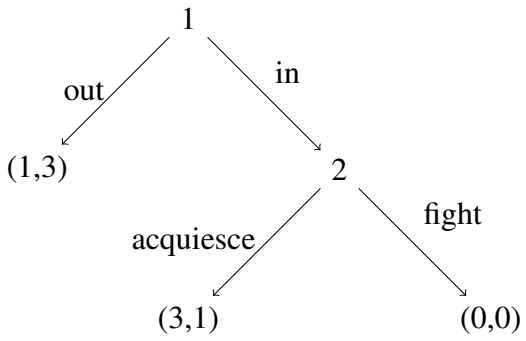

Figure 2.1: The market-entry game

Assuming the firms act rationally, there is a Nash equilibrium given by strategy profile (O, F), where firm 1's strategy is to stay out, and firm 2's is to fight. In this

case, since the challenger is staying out, switching between acquiescing and fighting is irrelevant for firm 2, and since the monopolist will fight if firm 1 enters, entering will leave them worse off. In epistemic terms, firm 1 believes firm 2's threat to fight if they decide to enter, and this makes it rational for them to decide to stay out. However, this equilibrium is not very satisfying, as it is hard to see how it models real-life behavior: firm 2's threat is "empty" in the sense that a rational player would never choose to fight in this game if they were truly due to play, since in that case fighting guarantees them a utility of 0 while acquiescing guarantees a utility of 1.

The standard way of addressing this kind of issue in extensive form games is to demand that players should be rational whenever they are (even hypothetically) due to play (see, e.g., [38]). The corresponding solution concept is called *subgame perfect equilibrium*. Thus, in the example above, since fighting is not rational in the subgame where firm 2 is due to play, the profile (O, F) is ruled out as a subgame perfect equilibrium.

2.1.2 Epistemic characterizations

There has been much discussion in the literature about the appropriate epistemic characterizations for subgame perfect equilibrium, which is equivalent to the backwards induction solutions in games of perfect information [5, 11, 13, 12, 6, 7, 8]. This has led to one of the best-known disagreements within the field, namely that between Aumann and Stalnaker: Aumann claims common knowledge of rationality leads to the backwards induction solution ([2]), and Stalnaker claims it does not ([41, 42]). Halpern, in an attempt to identify what is at the core of these very different characterizations, creates a model which is able to represent both Aumann's and Stalnaker's formalizations [27]. His conclusion is that the disagreement stems from a difference in how Aumann and

Stalnaker formalize rationality for extensive-form games.

As noted above, the relevant portion when considering extensive-form game subgame perfect equilibria is how to evaluate rationality at nodes players assume they will not reach.

Aumann, seems, at first glance, to take this entire discussion lightly. In order to see whether a player is best responding at a vertex they currently consider unreachable, he simply considers agents *suppose* that vertex already happened, and then they see whether continuing with the same strategy and beliefs on others' strategies is a best-response. With respect to this definition, common knowledge of rationality leads to playing the backwards induction solution (this is what we - and Halpern - will be referring to as *Aumann's result*).

For Stalnaker, reasoning about rationality in extensive-form games has an inherent counterfactual nature. In order to see, at the current world, if a player is rational in their response at a vertex currently considered unreachable, he argues for a need to (counterfactually) shift to the closest world where the vertex has indeed occurred, and see whether the player is best-responding to their beliefs there. In light of this definition, it turns out that common knowledge of rationality is not sufficient to ensure the backwards induction solution is played. Stalnaker points out that, since the vertex is considered unreachable at the current world, seeing what would happen if the player were to act at that vertex is not a simple matter, and it requires a belief revision policy - after all, it might imply that an opponent is doing something completely unexpected, and that needs to be included in updating the player's beliefs. This stems partly from the fact that, for Stalnaker, strategies themselves seem to have an epistemic component and a causal component, suggesting a need to separate strategies from syntactic specifications of a model; notably, however, he still includes them as primitives within his models.

In any case, as Halpern points out, some easily spottable differences between the two models are the lack of explicit counterfactual shifts, and the lack of an explicit belief revision policy for Aumann. And, as Stalnaker points out, these are not simple syntactic differences, but quite crucial elements inherent to reasoning in extensive-form games.

Halpern seems to agree with Stalnaker that a model without a belief revision policy is weaker in this context. His model, furthermore, is able to identify a coherency requirement between a player's beliefs at the current world and those at the closest-world where a certain vertex occurred. When this requirement is added to Stalnaker's models, Halpern proves common knowledge of rationality within these models is necessary and sufficient to ensure the backwards induction solution.

2.1.3 Our contributions

At a closer look at the models outlined above, however, there are a number of issues which arise.

To start, from a game theoretical perspective, the notion of rationality, as well as the subsequent notions of backwards induction or subgame perfect equilibrium, have very clear intuitions. It is suprising, then, that there are a number of differing epistemic characterizations of all of these notions. We claim that the reason for this is a certain large granularity at the syntactic level, which doesn't allow for transparency in the kinds of assumptions sometimes taken for granted within a game theoretical approach.

For example, Aumann's framework has him consider the outcome obtained by supposing vertex v happened and then following the path given by a certain strategy profile;

we can see that there is an implicit counterfactual shift to the closest-world where v indeed has happened. Stalnaker notes this and tries to include it in his models. A problem we note, however, is that Stalnaker evaluates rationality at a vertex v in the current world by considering an agent's beliefs at the *actual* closest world where v has occurred. A priori, there is no reason why an agent should have access to these beliefs. Furthermore, it defeats the purpose of evaluating rationality in these games, at least as far as subgame perfect equilibrium is concerned in classical game theory - the issue is not whether the player *would* actually be best-responding to her beliefs if v occurred, but if she is best-responding to her current beliefs about what *would* happen if v occurred.

Continuing the work of Stalnaker and Halpern, the models we propose are also based on a counterfactual framework. Our formalization of rationality, however, does not require this unmotivated epistemic access, and, instead, looks more carefully at the set of beliefs an agent currently has about what would happen at some counterfactual world. We make this more precise in Section 2.3.2. Interestingly, within our framework, we show that Aumann's intuition and formalization is the appropriate way to deal with rationality in perfect information extensive form games.

Furthermore, Stalnaker notes that, while strategies have a simple mathematical definition, they seem to encode both an epistemic component and a causal component. This would suggest a need to separate strategies from the syntactic specifications of a model, and have them recreated from other primitives. There is work in the existing literature pushing this issue forward. Bonanno emphasizes that defining strategies involves specification of actual behavior, as well as counterfactual behavior (e.g. specifying actions chosen at histories subsequent to actions not chosen, and as such, not reached) [12]. For him, this distinction between the epistemic and the causal components of strategies, and the consequences this has on rationality, lead him to recommend a *subjective*

counterfactual shift function, i.e. one for each player. This severely complicates the models and confuses a player's beliefs about her opponents with her beliefs about what would be the case if the game was continuing from a certain vertex. Samet defines a model which appears as a possible answer to Stalnaker's complaint ([40]). He introduces a separate modal operator to represent *hypothetical knowledge*, which he claims is unreproducible using counterfactuals (Halpern shows, later, the contrary [26]), and he reconstructs strategies as *hypothetical* objects. Samet's models, then, are quite far removed from the original ones of Aumann and Stalnaker.

We agree that, once we operate within a counterfactual framework, the notion of a strategy also cries out for a counterfactual treatment. Further, once we allow for models which no longer include strategies as primitives, the notion of iterating counterfactuals in reasoning about rationality in these games arises naturally. Hence, one of the most important contributions of the current paper is the construction of a framework through which we can isolate the assumptions that underlie some of the results in the models outlined above, and which seem, within these models, unmotivated. Within our framework, we can more easily study to what extent the assumptions above are reasonable, and whether they can be left out of models for extensive-form games.

2.2 Halpern's model

We begin by carefully introducing Halpern's model. As mentioned above, he constructs this model in order to compare Aumann's and Stalnaker's, so it is a great framework to start with for our analysis. Later, in section 2.3.3, we will enrich it in order to discuss models where strategies are no longer primitives, but much of our initial discussion uses his model, so we will dedicate this entire section to presenting his formalism carefully.

Fix a game Γ of perfect information for n players. The non-leaf vertices in the game tree are partitioned into n sets, $G_1, \ldots, G_n$; the vertices in G_i are said to belong to i, in that these are histories where i is due to move. We start by defining a **model** of Γ in Halpern's adaptation of Aumann's perspective.

Definition 2. *A* model *of* Γ *is a tuple* $(\Omega, (\mathcal{K}_i)_{i \in N}, \boldsymbol{s})$*, where* Ω *is a set of states of the world,* $\mathcal{K}_i$ *are information partitions, one for each player* i*, and* $\boldsymbol{s}$ *maps each world* $\omega \in \Omega$ *to a strategy profile* $\boldsymbol{s}(\omega) = (s_1, \ldots, s_n)$*, where* s_i *is* i*'s strategy in* Γ *at state* ω*.*

An **extended model** of Γ is Halpern's adaptation of Stalnaker's perspective. This is not precisely the one Stalnaker argues for, as Stalnaker has a probability measure on worlds instead of accessibility relations for semantics of belief, but, as Halpern outlines, for the purposes of outlining the intuition behind rationality, looking at this model suffices.

Definition 3. *Define* an extended model *of* Γ *to be a tuple* $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \boldsymbol{s}, f)$*, where* $(\Omega, (\mathcal{K}_i)_{i \in N}, \boldsymbol{s})$ *is a model of* Γ*, and* $f : \Omega \times G \to \Omega$ *is a closest-world shift function, and* $G = \bigcup_{i=1}^{n} G_i$*. There are some constraints on* f*:*

F1. v *is reached in* $f(\omega, v)$*; that is,* v *is on the path determined by* $\boldsymbol{s}(f(\omega, v))$*.*

F2. If v *is reached in* ω*, then* $f(\omega, v) = \omega$*.*

F3. $\boldsymbol{s}(f(\omega, v))$ *and* $\boldsymbol{s}(\omega)$ *agree on the subtree of* Γ *below* v*.*

A careful look at the previous models will outline their similarities. Clearly, an *extended-model* is just a *model* with an added counterfactual shift function, which is needed to interpret Stalnaker's definitions; the counterfactual shift function is not needed for Aumann since his model does use any explicit counterfactuals.

Both Aumann and Stalnaker assume players know their strategies, so, formally, if $\omega' \in \mathcal{K}_i(\omega)$, then $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$.

Define operator K_i on events as $K_i(E) = \{\omega : \mathcal{K}_i(\omega) \subseteq E\}$, where $K_i(E)$ is the event that i knows E. Let $A(E) = K_1(E) \cap \ldots \cap K_n(E)$, where $A(E)$ is the event that all players know E. Let $CK(E) = A(E) \cap A(A(E)) \cap \ldots$, where $CK(E)$ is the event that E is common knowledge. Let BI comprise all states where the backward induction solution is played.

Denote $h_i^v(s)$ to be i's payoff if the unique path in Γ determined by following strategy s starting at v. We can now define rationality at a vertex (where Aumann's and Stalnaker's views agree).

Definition 4. *Player i is said to be rational at vertex v in ω if for all strategies $s^i \neq \boldsymbol{s}_i(\omega)$, there exists $\omega' \in \mathcal{K}_i(\omega)$ such that*

$$h_i^v(\boldsymbol{s}(\omega')) \geq h_i^v(\boldsymbol{s}_{-i}(\omega'), s^i).$$

In other words, a player i is irrational if there exists another strategy they could be playing that would leave them better off in all the worlds they consider possible.

Ratonality in the overall game is where Stalnaker and Aumann disagree on, as we can see in Definition 5.

Definition 5. *A player i is Aumann-rational (or A-RAT) at ω if they are rational at every $v \in G_i$ in ω. A player i is Stalnaker-rational (or S-RAT) if they are rational at every $v \in G_i$ in $f(\omega, v)$.*

This difference, small as it may seem, is enough to generate the opposing perspectives on whether common knowledge of rationality is necessary and sufficient to ensure the backwards induction solution.

Theorem 1 (Aumann's Theorem)**.** *If* Γ *is a nondegenerate game of perfect information, then in all models of* Γ*,* $CK(A\text{-}RAT) \subseteq BI$*.*

Theorem 2 (Stalnaker's Theorem)**.** *There exists a nondegenerate game* Γ *of perfect information and an extended model of it (in which the selection function satisfies F1-F3) such that* $CK(S\text{-}RAT) \nsubseteq BI$*.*

Halpern notes that this difference is a formal reflection of the fact that Aumann's framework does not allow him to model players' revising their beliefs, while Stalnaker's does. He then identifies a formal requirement which, when imposed to Stalnaker's framework, is neccesary and sufficient to generate a model in which Aumann's result holds:

F4. For all i, v, if $\omega' \in \mathcal{K}_i(f(\omega, v))$ then there exists a state $\omega'' \in \mathcal{K}_i(\omega)$ such that $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree on the subtree of Γ below v.

This guarantees a certain coherency between players' beliefs about other players' strategies at the current world and those at the closest-world where players are, hypothetically, in a certain subgame.

This leads Halpern to a result similar, in spirit, to the one Stalnaker formulated in a response paper to Aumann [42]:

Theorem 3. *If* Γ *is a nondegenerate game of perfect information, then for every extended model of* Γ *in which the selection function satisfies F1-F4,* $CK(S\text{-}RAT) \subseteq BI$*. Moreover, there is an extended model of* Γ *in which the selection function satisfies F1-F4.*

2.3 Our approach

2.3.1 Motivation

There are a number of unmotivated assumptions in the existing models. Our goal is to generate a framework where the intuition directly guides our formalization, which will allow us to study how reasonable these assumptions truly are.

As we've already mentioned, we focus on the inherent counterfactual nature in reasoning about rationality in extensive-form games, so our models will start by adopting the general framework of Halpern's *extended-models*. First, we take a closer look at the existing definitions of rationality. Stalnaker's, in particular, seems to require an unreasonable *epistemic access*. Intuitively, if we say an agent has epistemic access to a certain state, this simply means that the reach of that agent's knowledge extends to that particular state.

We offer an alternative, which we claim is the appropriate way to reason about rationality within a counterfactual framework. Interestingly, once we do, we are able to see that Aumann's definition of rationality is not as unreasonable as Stalnaker and Halpern claim. Further, in this process, we observe that the bulk of the work in proving Aumann's result for extended-models is actually done by requirement F3, one of the three requirements imposed on a counterfactual shift function for extended-models. Halpern claims that the purpose of F3 is to capture the intuitive meaning of a strategy, insofar as this intuition can be given a voice through the language of counterfactuals: if the strategy at ω specifies choosing a at v, then at the closest v-world to ω, a should be played. However, as Halpern points out, requirement F3 says much more than this, since it requires this coherency at all vertices below v as well. We construct a model where the counter-

factual shift function does not satisfy F3, and does, instead, satisfy some well-motivated assumptions, linking counterfactual shifts to the inherent structure of an extensive-form game. These models will also no longer contain strategies as primitives, in an attempt to isolate further unmotivated assumptions within previous models. As it turns out, in these new models, we are able to prove that F3 holds.

2.3.2 Rationality

Intuitively, rationality requires best responding to one's beliefs. In the case of extensive form games, however, players' beliefs about their opponents' strategies are closely connected to their beliefs about what would happen after a certain node in the tree is reached. It will be helpful, then, to take a more general approach to defining rationality.

Definition 6. *We say player i is rational at vertex v in world ω **with respect to the belief set** S if, for all strategies $s^i \neq \boldsymbol{s}_i(\omega)$, there exists $\omega' \in S$ such that $h_i^v(\boldsymbol{s}(\omega')) \geq h_i^v(\boldsymbol{s}_{-i}(\omega'), s^i)$.*

Informally, this says i cannot do any better by using s^i (instead of $\mathbf{s}_i(\omega)$), considering her beliefs S at world ω regarding the strategies the other players are employing. Then Definition 4 appears as a special case, where we restrict the set of beliefs S to be the beliefs player i has at ω, namely the information set $\mathcal{K}_i(\omega)$. Interestingly, the definition above allows us to more clearly delineate between Aumann's definition of rationality in the overall game and Stalnaker's (as defined by Halpern).

Definition 7. *A player i is Aumann-rational at ω if, for all $v \in G_i$, she is rational at v in world ω with respect to $\mathcal{K}_i(\omega)$, i.e. for all $s^i \neq \boldsymbol{s}_i(\omega)$ there is $\omega' \in \mathcal{K}_i(\omega)$ such that $h_i^v(\boldsymbol{s}(\omega')) \geq h_i^v(\boldsymbol{s}_{-i}(\omega'), s^i)$.*

Definition 8. *A player i is Stalnaker-rational at ω if, for all $v \in G_i$, she is rational at v in world $f(\omega, v)$ with respect to $\mathcal{K}_i(f(\omega, v))$, i.e. for all $s^i \neq \boldsymbol{s}_i(f(\omega, v))$ there is $\omega' \in \mathcal{K}_i(f(\omega, v))$ such that $h_i^v(\boldsymbol{s}(\omega')) \geq h_i^v(\boldsymbol{s}_{-i}(\omega'), s^i)$.*

We can see the difference between these two definitions in Figure 2.2. *S-RAT* at ω (the world with a bold outline in the figure) is evaluated with respect to the belief set at the closest v-world to ω (marked with a red arrow in the figure), which could be in a different information set, while *A-RAT* at ω is evaluated with respect to the belief set at ω.

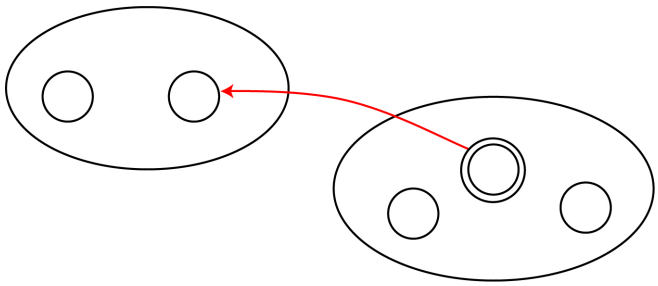

Figure 2.2: Rationality

Notably, Stalnaker-rationality at ω implies a player's access to beliefs at $f(\omega, v)$, the *actual* closest-world to the current world, which is something that the player might not have epistemic access to.

The way to think about rationality in extensive-form games is by seeing whether a player is best-responding to the beliefs they would have if they were at a certain point in the game tree. This is sometimes referred to as a player's *counterfactual beliefs*. Both English descriptions, however, are insufficiently precise. There are two different possible interpretations of *counterfactual beliefs*, or *beliefs I would have if v were to occur*. One interpretation is the set of beliefs at the counterfactual world closest to the actual one where v does occur, formally represented as $\mathcal{K}_i(f(\omega, v))$. Another interpretation is the set of beliefs you currently have about what would happen at the counterfactual sit-

uation where v occurs, in which case you have to consider the image of your beliefs under some counterfactual shift. As such, the expression *counterfactual beliefs* is too vague, in that it does not fully specify the scope of the belief operator with respect to the counterfactual shift function. Due to this ambiguity we will avoid using this term.

We claim the second interpretation is the appropriate one with respect to evaluating rationality in extensive-form games. However, we cannot represent this formally using existing notation. Then, we define the set of beliefs at ω about what *would* be the case if v **were** to occur [1] as

$$\mathcal{K}_i^v(\omega) = \{f(\omega', v) : \omega' \in \mathcal{K}_i(\omega)\}.$$

Basically, these beliefs are the *pushforward* of the player's *actual* beliefs at ω, via the counterfactual shift function. We can now define what it means for a player to believe they **would** (counterfactually) be best-responding in a given subgame. We call this *subgame-rationality*.

Definition 9. *A player i is subgame-rational at ω if, for all $v \in G_i$, she is rational at v in world ω with respect to $\mathcal{K}_i^v(\omega)$, i.e. for all $s^i \neq \boldsymbol{s}_i(\omega)$ there is $\omega' \in \mathcal{K}_i^v(\omega)$ such that $h_i^v(\boldsymbol{s}(\omega')) \geq h_i^v(\boldsymbol{s}_{-i}(\omega'), s^i)$.*

Figure 2.3 points out the main difference between *S-RAT* and our definition of *SUB-RAT*, since the definitions correspond to the two different interpretations we outlined above for the beliefs one *would* have if they *were* at a certain point in the game tree.

Further interesting points arise, after a careful look at the formalism of Definitions 7 through 9. In Aumann's definition, the superscript v in h_i^v is necessary, as it implicitly

[1] We agree that this is slightly awkward; as we have already noted, our natural language cannot capture this distinction, while the formal models can, emphasizing their importance

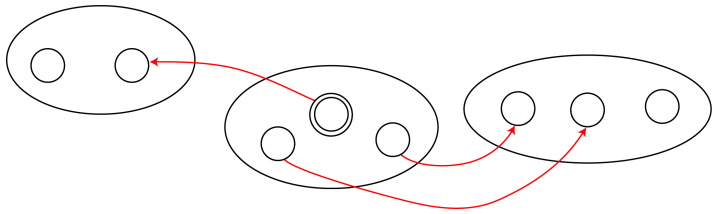

Figure 2.3: *S-RAT* and *SUB-RAT*

defines a counterfactual shift; it helps agents reason about what would happen in the subgame starting at v. Hence, even though Aumann does not have a counterfactual shift function, there is an implicit one in the notation used above.

In Stalnaker's definition, it might seem, at first glance, that v is not necessary, since, after all, we are looking at the beliefs at $f(\omega, v)$, which is a v-world by definition. However, a player i might consider several non-v-worlds possible at the closest v-world, so, in fact, v appears twice within the definition, and it is needed both times.

In our definition, the superscript v is superfluous: we are considering an agent's response to the pushforward of her beliefs, so we have already ensured that all worlds considered are v-worlds. As such, for all $\omega' \in \mathcal{K}_i^v(\omega)$, we have $h_i(\mathbf{s}(\omega')) = h_i^v(\mathbf{s}(\omega'))$ (where $h_i(\mathbf{s}(\omega'))$ simply is the outcome determined by the strategy profile $\mathbf{s}(\omega')$).

Interestingly, due to condition F3, we note that Definition 9 - which we claim is appropriate once we deal with a model with the added power of counterfactual shifts - is actually equivalent to Aumann's.

Proposition 1. *Let Γ be a nondegenerate game of perfect information. Then for every world ω and player i, player i is Aumann-rational at world ω iff they are subgame-rational at world ω.*

Proof. Let $\omega \in \Omega$, $i \in N$. We will prove that checking whether Aumann-rationality holds (as per Definition 7) coincides with checking whether subgame-rationality holds

(as per Definition 9).

Now consider $v \in G_i$. Using condition F3, $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ coincide below v, for any $\omega' \in \mathcal{K}_i(\omega)$ and $\omega'' \in \mathcal{K}_i^v(\omega)$. This implies $h_i^v(\mathbf{s}(\omega')) = h_i^v(\mathbf{s}(\omega''))$, and, in fact, these are also equal to $h_i(\mathbf{s}(\omega''))$. Further, we also have $h_i^v(\mathbf{s}_{-i}(\omega'), s^i) = h_i^v(\mathbf{s}_{-i}(\omega''), s^i)$ for any $s^i \neq \mathbf{s}_i(\omega)$, since $\mathbf{s}_i(\omega)$, $\mathbf{s}_i(\omega')$, and $\mathbf{s}_i(\omega'')$ all coincide below v (by F3 and players knowing their strategies). □

Let *SUB-RAT* comprise all states where all players are subgame-rational. Given this equivalence to Aumann's rationality, it is easy to see that $CK(\textit{SUB-RAT}) \subseteq \textit{BI}$.

Interestingly, however, once we consider subgame-rationality, a different perspective renders the common knowledge requirement obsolete. Let $\textit{Strat} = \{\omega : \forall i \; \forall \omega' \in \mathcal{K}_i(\omega), \; \mathbf{s}(\omega) = \mathbf{s}(\omega')\}$ be the set of states where all players know their strategies.

Proposition 2. *Consider an extended model for nondegenerate game* Γ *of perfect information where the selection function satisfies F1-F3. Then* $\textit{Strat} \wedge \textit{SUB-RAT} \subseteq \textit{BI}$. *Moreover, there exists an extended model of* Γ *in which the selection function satisfies F1-F3.*

Proof. The proof follows the structure of the proof outlined by Halpern.

Suppose $\omega \in \textit{Strat} \wedge \textit{SUB-RAT}$. We proceed by induction on k: if v is at height k in the tree, the move indicated by the backwards induction solution is played at v in ω.

For the base case, suppose v is at height 1. Suppose i moves at v and suppose, for a contradiction, i plays a', and not a, which is the backwards induction solution. Then, since a is the backwards induction solution (in this nondegenerate game), for all $\omega' \in \mathcal{K}_i(\omega)$, we have $h_i^v(\mathbf{s}(f(\omega', v)) < h_i^v(\mathbf{s}_{-i}(f(\omega', v)), a)$, which is a contradiction to i being *SUB-RAT* at v in ω.

Now for the inductive step, suppose v is at height $k+1$, and i moves at v. For a contradiction, suppose $\mathbf{s}_i(\omega')(v) = a' \neq a$, where a is the move indicated by the backwards induction solution. Since $\omega \in \textit{SUB-RAT}$, there exists $\omega' \in \mathcal{K}_i(\omega)$ such that

$$h_i^v(\mathbf{s}(f(\omega', v)) \geq h_i^v(\mathbf{s}_{-i}(f(\omega', v)), a).$$

Since players know their strategies, it must be that i plays a' at ω' and, by F3, she does the same at $f(\omega', v)$. Now, by the inductive hypothesis, we know that at all vertices below v (strictly), all players play the backwards induction solution in ω. Since $\omega \in \textit{Strat}$, we know the same holds for ω' and, by F3, for $f(\omega', v)$. This implies that $h_i^v(\mathbf{s}_{-i}(f(\omega', v)), a)$ is the outcome corresponding to the backwards induction solution at v, so the inequality above is equivalent to i getting a bigger payoff by playing a' than a, a contradiction (to the definition of backwards induction).

For the second half, we only note that the model constructed by Halpern requires f to satisfy F4 as well, so there clearly is a model where the function satisfies F1-F3.

□

In fact, common knowledge is not required even when we use Halpern's definition of Stalnaker-rationality and the requirement that the counterfactual shift function satisfies F1-F4.

Proposition 3. *Consider an extended model for a nondegenerate game Γ of perfect information where the selection function satisfies F1-F4. Then $\textit{Strat} \wedge \textit{S-RAT} \subseteq \textit{BI}$.*

Proof. Suppose $\omega \in \textit{Strat} \wedge \textit{S-RAT}$. We proceed by induction on k: if v is at height k, the move indicated by the backwards induction solution is played at v in ω.

For the base case, suppose v is at height 1. Since $\omega \in S\text{-}RAT$, we know i is rational at $f(\omega, v)$, and since v has height 1, this implies i makes the move dictated by the backwards induction solution at $f(\omega, v)$. By F3, i makes the same move at ω.

For the inductive case, suppose v is at height $k + 1$ and that i moves at v. Suppose for a contradiction that $\mathbf{s}_i(\omega)(v) = a' \neq a$, which is the backwards induction action.

By the inductive hypothesis, at every vertex below v, players plan to play the backwards induction solution at ω. By F3, we know i's strategy is to play a' at $f(\omega, v)$, and at every vertex below v players plan to play the backwards induction solution at $f(\omega, v)$. Since $\omega \in S\text{-}RAT$, we know i is rational at v in $f(\omega, v)$. Then there exists $\omega' \in \mathcal{K}_i(f(\omega, v))$ such that using $\mathbf{s}_i(f(\omega, v))$ does at least as well at ω' as using the backwards induction solution starting from v. By F4, there is $\omega'' \in \mathcal{K}_i(\omega)$ such that $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree below v. Since $\omega \in Strat$, we know $\mathbf{s}(\omega) = \mathbf{s}(\omega'')$, so, by the inductive hypothesis, the players are playing the backwards induction below v at ω''. Since $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree below v, the players are playing the backwards induction solution below v at ω'.

This implies that i does at least as well playing a' as she does playing a in ω', a contradiction to a being the backwards induction solution. □

It's worthwhile, then, to take a closer look at the $Strat$ requirement. At first glance, this seems a far too strong constraint on agents' epistemic access, since it seems that mutual knowledge of the strategy being played, and not common knowledge of rationality is what is needed. However, this distinction appears in a clear analog to the characterization of Nash Equilibrium for normal-form games. After all, when we are analyzing whether a strategy profile is a Nash equilibrium, it is irrelevant what one player believes about another, and, instead, the assumption is that players have mutual knowledge of the strategy being played. With this in mind, requiring that rationality (in the basic sense,

at the empty history of the game), together with mutual knowledge of the strategy being played, leads to the analog of Nash Equilibrium for Extensive Form Games. Now, as we argued in the introduction, this kind of rationality is not appropriate for reasoning about extensive form games, particularly due to empty threats. But if we are to change the definition of rationality to the suitable one, namely *SUB-RAT*, while maintaining the constraints for Nash Equilibrium for normal-form games, we obtain the subgame perfect equilibrium/backwards induction solution. Given this, the fact that $Strat$ is the necessary requirement for Halpern's model is no longer surprising - once the assumptions on the model are fixed to correct for the inappropriate definition of rationality, requiring mutual knowledge of the strategies being played is a reflection of the fact that the backwards induction solution is an analog of Nash Equilibrium. Notably, then, although our model can capture higher order beliefs, and can, thus, capture common knowledge of rationality, we only need the appropriate notion of rationality and counterfactuals in order to get the backwards induction solution.

2.3.3 Extended counterfactual models

We now turn to constructing a model where we forgo requirement F3 for the counterfactual shift, as well as strategies as primitives.

Recall that in Halpern's models, worlds encoded strategies. As we've already seen, they also implicitly associated an outcome with every world. We claim this is in fact the primitive agents truly reason about - the way the world can be - and so, while strategies will be recreated as counterfactual objects, outcomes become our primitives. In order to do so, first, we need to distinguish between the terminal history of a certain game, and a player's utility at that terminal history (what Halpern would refer to as *outcome*). As

such, let Z be the subset of G consisting of terminal histories, and $u_i : Z \to \mathbb{R}$ are as given by the game tree for any player i.

Definition 10. *Define an* extended-counterfactual-model $M = (\Omega, (\mathcal{K}_i)_{i \in N}, O, f)$*, where:*

(R1) Ω *is a nonempty, finite state space.*

(R2) $\mathcal{K}_i : \Omega \to (2^\Omega \setminus \emptyset)$ *are information partitions of* Ω*, one for each player* i*.*

(R3) $O : \Omega \to Z$ *assigns a unique outcome to every state.*

(R4) $f : \Omega \times G \to \Omega$ *is a "closest-world" function satisfying*

- *(S1)* v *is reached in* $f(\omega, v)$*; in other words, the sequence of moves leading to* v *is an initial history for* $O(f(\omega, v))$*.*
- *(S2) if* v *is reached in* ω *(i.e., the sequence of moves leading to* v *is an initial history for* $O(\omega)$*), then* $f(\omega, v) = \omega$*.*

These are the models Halpern considered, except we have removed strategies as primitives, and we have added an explicit outcome function. We also note that S1 and S2 are precisely F1 and F2 in Halpern's models. And while we do not include requirement F3, we do impose some further conditions on our counterfactual shift operator.

(C1) $\forall \omega \in \Omega$, $\forall i \in N$, $\forall v \in G_i$, if v', which is obtained by taking some available action a_i at v, is on the path specified by $O(f(\omega, v))$, then for all $\omega' \in \mathcal{K}_i(\omega)$, v' is on the path specified by $O(f(\omega', v))$.

(C2) $\forall \omega \in \Omega$, $\forall v \in G$, and for all v' successors of v (i.e. obtained by taking some available action a at v), we have

$$f(f(\omega, v), v') = f(\omega, v').$$

Condition (C1) essentially guarantees that players are sure of their own *strategies* (we make this precise in Section 2.3.4). Condition (C2) says, roughly speaking, that the counterfactual shift operator respects the temporal nature of the game in the sense that the closest v'-world to ω is also the closest v'-world to the closest v-world to ω (where v and v' are immediate successors).

It is worth noting that although it seems that condition (C2) is superficially restricted to a single move after a vertex, this coherency condition actually holds across all extensions, as established by the following lemma.

Lemma 1. *For all $\omega \in \Omega$ and $v \in G$, if $v' \in G$ is reachable from v via a sequence of moves h', we have*

$$f(f(\omega, v), v') = f(\omega, v').$$

Proof. We proceed by induction on the length of the sequence of moves h'. The base case, when the length is 1, follows by one application of (C2). Now suppose the statement holds for any v' reachable from v in k many steps, we want to prove it holds for a vertex reachable in $k + 1$ steps. Namely, suppose we have moves $a^1, \ldots, a^{k+1}$ such that by following them from vertex v, we obtain vertex $v' \in G$. We want to prove

$$f(f(\omega, v), v') = f(\omega, v').$$

For ease of notation, let v^k be obtained by following the set of actions $a_1, \ldots, a^k$ form v. Then v' is obtained by taking action a^{k+1} at v^k. Using (C2), we know $f(\omega, v') = f(f(\omega, v^k), v')$. Our inductive hypothesis then gives

$$f(\omega, v') = f(f(f(\omega, v), v^k), v'),$$

and another application of (C2) (to $f(\omega, v)$) gives the desired result. □

2.3.4 Representing strategies

In the classical setting, a *strategy for player i* is a function ρ_i mapping histories where player i is due to play to actions such that for each $v \in G_i$ we have $\rho_i(h)$ an available action in the game tree at vertex v. Informally, a strategy is a complete specification of what the player would do at every point in the game where they would be due to play. This might be thought of as a "plan of action", in a certain sense, but with the understanding that it includes a plan for how to act even at histories that the player is sure they won't reach.

As we've already mentioned, a profile of strategies (one for each player) uniquely determines an outcome, but the converse is not true in general (an outcome is typically compatible with more than one strategy). The models we have defined explicitly associate an outcome with each world via the function O, and they have all the necessary machinery to also uniquely pick out a strategy for each player at each world, derived using the counterfactual operator.

Fix a game Γ and let M be an extended-counterfactual-model for this game. Denote by Λ_i the set of all strategies for player i. For every world ω, each player i, and each vertex $v \in G_i$ at which player i is due to play, we can ask, "What would player i do if they were to find themselves at vertex v?" Provided the model M is sufficiently rich, the counterfactual operator provides a unique answer to this question by moving to the world closest to ω at which v actually occurs. Since f is, by definition, total and since f satisfies (S1), we note that for every node v of the tree, there is $\omega \in \Omega$ such that v is reached in ω. In particular, this implies O is surjective in counterfactual-extended-models.

Proposition 4. *Let* $M = (\Omega, (\mathcal{K}_i)_{i \in N}, O, f)$ *be a counterfactual-extended-model for* Γ.

For all $\omega \in \Omega$, for all $i \in N$, and for all nonterminal $v \in G_i$, there is a unique action in the set of actions available at v, call it $s_i(\omega)(v)$, such that the vertex reached by taking $s_i(\omega)(v)$ is also reached at $f(\omega, v)$.

Proof. Since O is surjective, there is $\omega' \in \Omega$ such that v is reached in ω', therefore $f(\omega, v)$ is defined. Since v is nonterminal, it follows that there exists an available action a at v such that action a is taken at $f(\omega, v)$. Now suppose two different available actions at v are taken at $f(\omega, v)$, leading to vertices v_1 and v_2 both reached at $f(\omega, v)$. This clearly contradicts O assigning a unique outcome to every vertex. □

As the notation suggests, we can think of each $s_i : \Omega \to \Lambda_i$ as a function assigning a strategy for player i to each world.

In the remainder of this section, we describe and prove a number of desirable properties these functions enjoy in our models. Fix a game Γ and an extended-counterfactual-model for it M. First, players are sure of their own strategies:

Proposition 5. *For every $\omega \in \Omega$ and each $i \in N$, we have that, for every $\omega' \in \mathcal{K}_i(\omega)$, $s_i(\omega') = s_i(\omega)$.*

Proof. Let ω' be as above, and consider nonterminal $v \in G_i$. By Proposition 4, $s_i(\omega)(v) = a$ for a unique a available at v. Then the vertex v' obtained by taking action a at v is reached at $f(\omega, v)$. By (C1), that implies that the sequence of moves leading to v' is an initial segment to $O(f(\omega', v))$ (i.e. v' is reached at $f(\omega', v)$). Again by Proposition 4, we know that $s_i(\omega')(v) = a'$ for a unique a' available at v, so the vertex v'' obtained by taking action a' at v is on the path determined by the terminal history $O(f(\omega', v))$, which implies that $a = a'$, i.e. $s_i(\omega)(v) = s_i(\omega')(v)$. Since this equality holds for arbitrary v, we have $s_i(\omega) = s_i(\omega')$ for every $\omega' \in \mathcal{K}_i(\omega)$, as desired. □

Now recall Halpern's notation $h_i^v(\mathbf{s}(\omega))$, which identified the outcome (in his case, the utility) for player i at the terminal history determined by following strategy profile $\mathbf{s}(\omega)$ after vertex v. Since we have had to separate the outcome of the game with a player's subsequent utility, we need a bit of notation.

For each $v \in G$, let $[v] \subseteq Z$ be the set of terminal histories that pass through v. We denote the unique terminal history in $[v]$ reached from v when the players play according to ρ by $[v]_\rho$. Then, in this notation, $h_i^v(\mathbf{s}(\omega))$ now becomes $u_i([v]_{\mathbf{s}(\omega)})$.

Further, $[\emptyset]_\rho$ is the outcome of the game induced by ρ. The next proposition ensures that the strategies associated with any given world actually induce the outcome that is assigned to that world.

Proposition 6. *For all $\omega \in \Omega$, the terminal history induced by $s(\omega) = (s_1(\omega), \ldots, s_n(\omega))$ is precisely $O(\omega)$. In other words, $[\emptyset]_{s(\omega)} = O(\omega)$.*

Proof. For a contradiction, suppose not; that is, suppose that $[\emptyset]_{s(\omega)} \neq O(\omega)$. Then there exists $v \in G$ such that v is the last common vertex in the paths determined by $s(\omega)$ and O. This implies v is not terminal. Let a be the available action at v taken at $O(\omega)$. Suppose player i is due to play at v. Then by assumption $s_i(\omega)(v) = a' \neq a$ (for some a'). This implies a' is the action taken at v in $f(\omega, v)$. Now, since v is on the path determined by $O(\omega)$, this implies v is reached at ω, so $f(\omega, v) = \omega$. This implies a' is the action taken at v in ω. Since O assigns a unique outcome to every state, it must be that $a' = a$, a contradiction to our assumption. □

Now, recall our definition of a strategy from Proposition 4. Note that it captures precisely the intuition that Halpern refers to in his motivation for F3: if a player's current plan is to play a if at the subgame determined by vertex v, then at the closest v-world, the player is choosing action a.

And, although we do not explicitly impose that strategies below v coincide at ω and the closest v-world to ω, as F3 does without much motivation, this arises naturally within our framework. They key condition here is (C2), which ensures the counterfactual shift respects the temporal nature of the game.

Proposition 7. *Suppose we have a extensive-counterfactual-model* $M = (\Omega, (\mathcal{K}_i)_{i \in N}, f, O)$ *of an extensive-form game* Γ. *Then* f *also satisfies F3.*

Proof. Let $\omega \in \Omega$ and $v \in G$. We want to show $s(f(\omega, v))$ and $s(\omega)$ agree at any vertex below v.

We proceed by strong induction on the length k of the sequence of moves following v which lead to a new vertex v'.

The base case, for $k = 0$, follows by definition.

For the inductive case, consider the vertex v'' obtained by starting at vertex v and following the sequence of moves $a_1, \ldots, a_k, a_{k+1}$); we want to show $s(f(\omega, v))$ and $s(\omega)$ agree at v''.

By definition, we know $s(f(\omega, v''))$ and $s(\omega)$ agree at v''. We want to show $s(f(\omega, v))$ and $s(f(\omega, v''))$ agree at v''. By Lemma **??**, we know $f(\omega, v'') = f(f(\omega, v), v'')$. Then we want to show $s(f(\omega, v))$ and $s(f(f(\omega, v), v'')))$ agree at v'', which holds by definition, once again. □

As such, while F3 seems a strong requirement in Halpern's original model, it arises as a natural consequence within our models.

What is more, there is a strong connection between the strategies primitive in an extended-model, and the ones recreated in an extended-counterfactual-model.

Consider game Γ and an extended-model for it $M = (\Omega, (\mathcal{K}_i)_{i\in N}, f, \mathbf{s})$. Generate an associated model M', using the same set of worlds, the same accessibility relations, and the same counterfactual shift; generate the outcome function defined by the outcome determined by the strategy profile at every world in the extended-model.

It is easy to see that R1-R4 hold for M', so M' is an extended-counterfactual-model. In fact, M' also satisfies C1.

Proposition 8. *Consider game Γ and an extended-model for it $M = (\Omega, (\mathcal{K}_i)_{i\in N}, f, \boldsymbol{s})$ which satisfies F1-F3. Generate an associated model M', using the same set of worlds, the same accessibility relations, and the same counterfactual shift; generate the outcome function defined by the outcome determined by the strategy profile at every world in the extended-model. Then M' is an extended-counterfactual model which also satisfies C1.*

Proof. The first part of the proof follows clearly - requirements R1-R4 are easy to check, so M' is an extended-counterfactual-model. To prove M' satisfies C1, let $\omega \in \Omega$, $i \in N$, and $v \in G_i$. Suppose v', which is obtained by taking action a at v, is on the path determined by $O(f(\omega, v))$. Then, by construction, this implies $\mathbf{s}_i(f(\omega, v))(v) = a$ (in M). Since F3 holds in M , this implies $\mathbf{s}_i(\omega)(v) = a$. Using the fact that players know their strategies in the models employed by Halpern, we get $\mathbf{s}_i(\omega')(v) = a$ for all $\omega' \in \mathcal{K}_i(\omega)$. Using F3 once again, this implies $\mathbf{s}_i(f(\omega', v))(v) = a$, so, by the definition of the outcome function, we get v' is also on the path determined by $O(f(\omega', v))$, for all $\omega' \in \mathcal{K}_i(\omega)$. □

Since this associated model satisfies C1, we can now reconstruct strategies at a world, as outlined in Section 2.3.4. In fact, the strategies reconstructed coincide with the ones in the original extended-model.

Proposition 9. *Consider game Γ and the associated extended-counterfactual-model for it $M = (\Omega, (\mathcal{K}_i)_{i\in N}, f, \boldsymbol{s})$ (constructed as in Proposition 8). Construct the associated*

strategy functions $s_i : \Omega \to \Lambda_i$ *for each* $i \in N$*, for* M'*. Then for every* $\omega \in \Omega$*,* $\boldsymbol{s}(\omega) = s(\omega)$.

Proof. Let $\omega \in \Omega$, $i \in N$, $v \in G_i$. We want to show $\mathbf{s}_i(\omega)(v) = s_i(\omega)(h)$. By construction, there exists a unique $a = s_i(\omega)(v)$ such that the vertex reached by taking a is also reached in $f(\omega, v)$. Then $\mathbf{s}_i(f(\omega, v))(v) = a$, so by F3, $\mathbf{s}_i(\omega)(v) = a$. Since this holds for arbitrary v and i, we have $\mathbf{s}(\omega) = s(\omega)$, for every $\omega \in \Omega$. □

2.3.5 Quantitative interpretation

Until now, we have used knowledge in our exploration of epistemic characterizations of solution concepts, in a similar spirit to Halpern. Furthermore, the entire discussion above is done through a qualitative lens. We note that we can easily consider a slightly more general model, in which we use belief for our epistemic characterization and, correspondingly, we look into a quantitative interpretation of rationality as expected utility maximization.

While all the results follow naturally, we include this section, since this quantitative version might be more familiar to some readers.

Define a *counterfactual-model* $M = (\Omega, (\mathcal{PR}_i)_{i \in N}, \mathcal{O}, f)$, where:

(R1) Ω is a nonempty, finite state space.

(R2) $\mathcal{PR}_i : \Omega \to \Delta(\Omega)$ assigns to each $\omega \in \Omega$ a probability measure $\mathcal{PR}_i(\omega)$ on Ω representing the beliefs of player i at ω.

(R3) $\mathcal{PR}_i(\omega)(\{\omega' : \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)\}) = 1$ (i.e., players are sure of their own beliefs).

(R4) $\mathcal{O} : \Omega \to Z$ assigns a unique outcome to every state.

(R5) $f : \Omega \times G \to \Omega$ is a "closest-world" function satisfying

(S1) v is reached in $f(\omega, v)$; in other words, the sequence of moves leading to v is an initial history for $O(f(\omega, v))$.

(S2) if v is reached in ω (i.e., the sequence of moves leading to v is an initial history for $O(\omega)$), then $f(\omega, v) = \omega$.

We note that, formally, the only change is replacing the information partitions with probability distributions on states, and ensuring coherency inherent within a partition (namely, players are sure of their beliefs). The added requirements appear without almost any changes:

(C1) $\forall \omega \in \Omega$, $\forall i \in N$, $\forall v \in G_i$, if v', which is obtained by taking some available action a_i at v, is on the path specified by $O(f(\omega, v))$, then for all $\omega' \in \Omega$ such that $\mathcal{PR}_i(\omega)(\omega') > 0$, v' is on the path specified by $O(f(\omega', v))$.

(C2) $\forall \omega \in \Omega$, $\forall v \in G$, and for all v' successors of v (i.e. obtained by taking some available action a at v), we have

$$f(f(\omega, v), v') = f(\omega, v').$$

Propositions 4 and 6 are unaffected by our change in the description of the model, and Proposition 5 follows very easily. Hence, we can also associate a unique strategy profile with every state in these counterfactual-models, such that players are sure of their own strategies and strategy profiles played at a world induce the outcomes associated with said world. Notably, since (C2) holds, it's easy to see that for any counterfactual-model, its counterfactual shift function satisfies $F3$.

And, once states are assigned probability measures representing agents' beliefs, we can capture agents' uncertainty over their opponents' strategies. Each probability mea-

sure on worlds induces a probability measure on opponents' strategies via the functions s_i. We can then compute each player's expected utility for playing their strategy at a given world:

Definition 11. *For every $i \in N$ and $\omega \in \Omega$, define*

$$EU_i(\omega, s_i(\omega)) = \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(\mathcal{O}(\omega')).$$

Now, using the same terminology as Aumann, we can easily find, for each world considered possible, the outcome (and hence, the utility) obtained by switching strategies, supposing all other players' strategies remain the same.

Definition 12. *For every $i \in N$, $\omega \in \Omega$, and $\rho_i \in \Lambda_i$, define*

$$EU_i(\omega, \rho_i) = \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i([\emptyset]_{(s_{-i}(\omega'), \rho_i)}).$$

When $\rho_i = s_i(\omega)$, Definition 11 clearly arises.

Let RAT_i comprise of all states where player i's expected utility function is maximized when playing the strategy associated with that state, and RAT comprise all states where all players are rational.

As discussed in the introduction, mere rationality falls short in articulating solution concepts for extensive form games, since it allows players to make bad decisions at histories that they assign probability 0 to. Thus, we want to extend our analysis to be able to reason about whether players' strategies constitute best responses *in every subgame*.

The existence of a probability distribution has some interesting consequences for the definition of subgame-rationality. Keeping the intuition on the appropriate notion of rationality in extensive-form games (from Section 2.3.2), we define the pushforward

of the player's actual beliefs at ω. We define a counterfactual probability distribution reflecting what player i believes *would* be the case if vertex v were to occur (this is precisely the notion of a player's counterfactual belief at state ω from [29], except adapted here to an extensive form setting). For every $i \in N$ and $\omega \in \Omega$, define

$$\mathcal{PR}_i^v(\omega)(\omega') = \sum_{\{\omega'' : f(\omega'', v) = \omega'\}} \mathcal{PR}_i(\omega)(\omega''). \tag{2.1}$$

Then we can define the expected utility of player i playing any given strategy, supposing history h has occurred:

Definition 13. *For every $i \in N$, $v \in G_i$, and $\rho_i \in \Lambda_i$, define*

$$EU_i^v(\omega, \rho_i) = \sum_{\omega' \in \Omega} \mathcal{PR}_i^v(\omega)(\omega') \cdot u_i([\emptyset]_{(s_{-i}(\omega'), \rho_i)}).$$

Note that F3 becomes crucial here. Since switching to the closest v-world does not change any of the players' strategies *from vertex v onward*, this definition of expected utility at a subgame actually computes the utilities associated with playing different strategies *in that subgame*, as we would want it to.

Superficially, Definition 13 looks quite different from Definition 12. However, we can show that when the history considered is $\emptyset$, corresponding to the beginning of the game, we obtain Definition 12 as a special case:

Proposition 10. *For every $i \in N$, $\omega \in \Omega$, and $\rho_i \in \Lambda_i$, we have $EU_i^{\emptyset}(\omega, \rho_i) = EU_i(\omega, \rho_i)$.*

Proof.

$$\begin{aligned} EU_i^{\emptyset}(\omega, \rho_i) &= \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(O(f(f(\omega', [\![\varphi_{\rho_i}]\!]), [\![\varphi_{\emptyset}]\!])), \text{ by Lemma 5} \\ &= \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(O(f(\omega', [\![\varphi_{\rho_i}]\!])), \text{ since } f(\omega'', [\![\varphi_{\emptyset}]\!]) = \omega'' \text{ for all } \omega'' \in \Omega \\ &= EU_i(\omega, \rho_i). \end{aligned}$$

□

We can now let $SUB\text{-}RAT_i$ denote the set of states where player i is maximizing their counterfactual expected utility whenever they are due to play, and *SUB-RAT* the set of states where all players are maximizing this counterfactual expected utility.

Note that, if a player is rational by maximizing their expected utility, then they must be rational using Aumann's qualitative definition. This is easier to formulate for irrationality: if a player is irrational since something else would leave them better off in all possible worlds, then they cannot be maximizing expected utility. The same holds for subgame-rationality. In particular, this implies all the results regarding epistemic characterizations of backwards induction solutions from Section 2.3.2 hold.

2.3.6 Discussion

To start, we want to address the proof of Proposition 7. The proof is quite easy, even more so since it follows closely the structure outlined by Halpern. Nevertheless, the point we want to emphasize is that it outlines that our assumptions are in line with the non-epistemic interpretation of Nash Equilibrium. While these assumptions are all quite natural, to the best of our knowledge, this result has not been discussed in the literature.

Further, there are a number of questions that we note are now open, and we leave for a future paper.

It is natural to study the full expressive power of counterfactuals within these models. As Bonanno points out, once we are doing counterfactual reasoning, it makes sense to include counterfactuals explicitly in the formal language agents are reasoning with [12]. This would allow us to adopt a more general framework for counterfactual shifts, in which we are no longer limited to shifting to a certain subgame. In the literature,

counterfactual shifts tend to be defined on subsets of the state space and, after all, shifting to the closest v-world is just a shorthand for shifting to the sets of worlds represented by vertex v. Once we can shift to any subsets of the state space, however, we can represent, formally, the notion of a player counterfactually switching their strategy. Now, since strategies are themselves counterfactual objects, this is a sort of "second-order" counterfactual shift, since it's a shift to a world where a conjunction of counterfactual statements hold (namely, the ones specifying the strategy). There is no existing work in the literature (that we have seen) that adopts this iterated counterfactual approach to model rationality.

Finally, while requirement F3 seemed too strong at an intuitive level, and unmotivated even though it also did much of the work in terms of Aumann's result, we were able to obtain it as a natural consequence of other, more palatable, primitive assumptions. In fact, this suggests that, if we are trying to find a more general model, it is not requirement F3 that we can relax, but instead, some of these more primitive assumptions.

At a first glance, both (C1) and (C2) seem crucial. We want players to know their strategies, and we want the counterfactual shift function to reflect our operating in an extensive-form game. However, we can see that our discussion on rationality, where we noticed the difference between counterfactually shifting from the actual world versus within the scope of your beliefs, also directly applies to (C1). Models without this requirement would allow us to study players who have false beliefs about what they would do. Halpern himself notes that, within extensive form games, strategies are very complicated, since, in the natural language, we seem to often conflate *what we would do* with *what we believe we would do*.

Further, in the context of extensive form games, relaxing this assumption seems to

illustrate a new way, not yet existing in the literature, of capturing agents' uncertainty in games. In the literature, this uncertainty is captured by randomization of strategies, or the probability distribution reflecting opponents' uncertainty of what the agent will do. The interpretation that we propose, however, can model agents that are uncertain about what they would do if they were to find themselves at v. We leave this discussion for future work.

In conclusion, we note that at the core of the disagreement in the literature on the epistemic characterization of the backwards induction solution lies confusion in the appropriate way to formalize game theoretic intuitions. Our framework is able to capture, formally, where these disagreements stem from. And while, in the end, our framework suggests that Aumann-rationality and his result are the appropriate ways to deal with rationality in extensive-form games, we note that it was the added counterfactual machinery that allowed us to identify this.

CHAPTER 3

ALTERNATE SEMANTICS FOR CONDITIONAL BELIEFS

3.1 Introduction

At the intersection of conditionals and beliefs lie statements like "conditional on φ, the agent believes ψ", statements which are also called "conditional beliefs". There are many different approaches in the literature to study conditional beliefs, and many semantics have been defined over the years [23, 17, 41, 10, 3]. It is easy to see that a counterfactual reading might also be applicable for such statements, namely "if φ were the case, then the agent would believe ψ". Using counterfactuals to interpret conditional beliefs is not new to the literature [41, 10]. We want to study the types of theories of conditional beliefs that arise by combining beliefs and counterfactuals, and where these theories stand within the ample literature on conditional beliefs.

Stalnaker points out the important difference between epistemic counterfactuals ("how players would revise their beliefs were they to learn they were mistaken") and causal counterfactuals ("a player considers what the consequences would be of his doing something he is not in fact going to do") [41]. He pushes this distinction forward and constructs plausibility models for belief revision in games. Board takes them as inspiration for a theory of conditional belief for multiple agents (BRSIC), using a language with an additional operator $B_i^{\psi}\varphi$, read as "i believes that φ upon learning that ψ" [10]. He calls them revised belief operators. Within the context of Dynamic Epistemic Logic, Baltag and Smets generate a single-agent version of his theory, namely Conditional Doxastic Logic (or CDL), where they introduce an additional modality, $B^{\psi}\varphi$, to interpret conditional belief. These models will serve as our inspiration.

Conditional beliefs arise naturally when we study how agents change (revise or update) their beliefs upon learning new information. Belief change is an active area of research in philosophy and artificial intelligence. There are two well-known theories which specify postulates that any structure modeling belief change has to satisfy. The distinction between - and the important implication of - static and dynamic belief change is well known in the literature [4, 45, 19]. There has been much work on the mechanisms for belief revision and update within the dynamic epistemic logic paradigm [4, 3, 45, 19, 46, 34]. The standard AGM theory is associated with belief revision, and with learning information about a static world [1], while the Katsuno and Mendelzon theory (also denoted KGM) is associated with belief update, and learning information about a changing world [32]. Many models have been proposed which model belief change according to these postulates, from connections to justification logic [4], to dynamic doxastic logic [34], or Halpern's belief-change systems [28]. Iterated belief change is also an active area of research [16, 35, 14]. Conditional beliefs can also be interpreted as suppositions, and there is a body of literature studying this approach [21].

We take a closer look at agents' conditional beliefs on a static situation. Using counterfactuals, however, we note there's an inherent "active" interpretation of conditional beliefs: *after* we consider φ being the case, *then* the agent believes ψ. In fact, broadly, given how the conditional belief is interpreted, we can divide existing models in two categories, *fixed* models which do not change as the agent conditions on the given information, and *kinetic* models which do. A "fixed" read is closer to the initial one offered, namely "conditional on φ, the agent believes ψ, while a "kinetic" read is closer to "the agent will believe ψ after revising her belief state by successfully incorporating the information that φ is true" [3]. While these interpretations are similar for non-epistemic facts, the difference appears in higher-order conditional beliefs. Our approach bridges the gap between these two interpretations, as we use counterfactuals only to change the

epistemic formulas at a world. It is easy to see that such an approach will have consequences on higher-order conditional beliefs, which becomes another departure from existing models for static belief change, which cannot capture higher-order beliefs.

3.1.1 Plausibility models

In this section we introduce some of the main components of plausibility models for conditional beliefs, as defined by Baltag and Smets. For a more detailed exposition, see [3, 4].

Suppose **PROP** is a set of propositional letters. A *plausibility model* is a structure $M = (W, \leq, V)$, where W is a set of worlds, $\leq$ is a reflexive and transitive binary relation on W, and $V : W \to 2^{\mathbf{P}}$ a valuation mapping each world to the set of propositional letters true at that world. For each $\omega \in W$, we define the set of worlds at least as plausible as ω as $\omega^{\downarrow} = \{x \in W \mid x \leq \omega\}$. For each $\omega \in W$, we define its *connected component* as $cc(\omega) = \{y \in W \mid x(\geq \cup \leq)^{+} y\}$, where $(\geq \cup \leq)^{+}$ is the transitive closure of $\geq \cup \leq$.

There are certain properties of interest that $\leq$ can satisfy, to which one usually restricts plausibility models. If $cc(\omega) = W$ for each $\omega \in W$, we say M is *connected*. We say $\leq$ is well-founded if, for each nonempty set $S \subseteq W$, the set $\min S = \{x \in S \mid \forall y \in S : y \not< x\}$ of minimal elements of S is nonempty, and we call the associated model *well-founded* also. We say $\leq$ is total on W iff for each $(x, y) \in W \times W$, we have $x \leq y$ or $y \leq x$, and we call the associated model M *total* as well. Finally, we say M is *well-ordered* when $\leq$ is well-ordered, namely when it is total and well-founded, and *locally well-ordered* when $\leq$ well-founded and total on each connected component.

Now consider the language of Conditional Doxastic Logic defined by the grammar:

$$\varphi ::= \bot \mid p \mid (\varphi \to \psi) \mid B^{\varphi}\psi$$

where $p \in$**PROP**. The formula $B^{\psi}\varphi$ is read "conditional on ψ, the agent believes φ". Intuitively, this means that each of the most plausible ψ-worlds satisfies φ.

Let $M = (W, \leq, V)$ be a locally well-ordered plausibility model. Semantics for $\mathcal{L}_{CDL}$ with respect to these models are defined the following way [3]:

- $[\![\varphi]\!]_M = \{v \in W \mid M, v \models \varphi\}$
- $M, \omega \not\models \bot$
- $M, \omega \models p$ iff $p \in V(\omega)$
- $M, \omega \models \varphi \to \psi$ iff $M, \omega \not\models \varphi$ or $M, \omega \models \psi$.
- $M, \omega \models B^{\psi}\varphi$ iff for all $x \in cc(\omega)$ we have

$$x^{\downarrow} \cap [\![\psi]\!] = \emptyset \text{ or } \exists y \in x^{\downarrow} \cap [\![\psi]\!] : y^{\downarrow} \cap [\![\psi]\!] \subseteq [\![\varphi]\!].$$

In other words, $B^{\psi}\varphi$ holds at ω iff for every world connected to ω that has an equally or more plausible ψ-world y, the ψ-worlds that are equally or more plausible than y satisfy φ.

Theory CDL_0 is a single-agent variant of Board's theory BRSIC, and is a sound and complete axiomatization of the language of conditional belief with respect to plausibility models [3]:

[(CL)] Schemes for Classical Propositional Logic

(K) $B(\varphi_1 \to \varphi_2 \mid \psi) \to (B(\varphi_1 \mid \psi) \to B(\varphi_2 \mid \psi))$

(Succ) $B(\psi|\psi)$

(IEa) $B(\varphi|\psi) \to (B(\chi|\psi \wedge \varphi) \leftrightarrow B(\chi|\psi))$

(IEb) $\neg B(\neg\varphi|\psi) \to (B(\chi|\psi) \to B(\chi|\psi \wedge \varphi))$

(PI) $B(\chi|\psi) \to B(B(\chi|\psi)|\varphi)$

(NI) $\neg B(\chi|\psi) \to B(\neg B(\chi|\psi)|\varphi)$

(WCon) $B(\bot|\psi) \to \neg\psi$

(MP) From φ and $\varphi \implies \psi$ infer ψ

(MN) From φ infer $[\psi]B\varphi$

(LE) From $\psi \leftrightarrow \psi'$ we can infer $B^{\psi}\varphi \leftrightarrow B^{\psi'}\varphi$

Baltag et. al also provide the equivalent theory CDL, which they use to give rise to a theory of Conditional Doxastic Logic with justifications:

[(CL)] Schemes for Classical Propositional Logic

(K) $B(\varphi_1 \to \varphi_2 \mid \psi) \to (B(\varphi_1 \mid \psi) \to B(\varphi_2 \mid \psi))$

(Succ) $B(\psi|\psi)$

(KM) $B(\bot|\psi) \to B(\bot|\psi \wedge \varphi)$

(RM) $\neg B(\neg\varphi|\psi) \to (B(\chi|\psi) \to B(\chi|\psi \wedge \varphi))$

(Inc) $B(\chi|\psi \wedge \varphi) \to B(\varphi \to \chi|\psi)$

(Comm) $B(\chi|\varphi \wedge \psi) \to B(\chi|\psi \wedge \varphi)$

(PI) $B(\chi|\psi) \to B(B(\chi|\psi)|\varphi)$

(NI) $\neg B(\chi|\psi) \to B(\neg B(\chi|\psi)|\varphi)$

(WCon) $B(\bot|\psi) \to \neg\psi$

(MP) From φ and $\varphi \implies \psi$ infer ψ

(MN) From φ infer $[\psi]B\varphi$

Note that this latter theory separates (IEa) and (IEb) into sub-axioms, and so (RM),

also called rational monotonicity, appears as a more basic principle within (IEb). We also note the similarity between this version of (RM) and (RM) for counterfactuals. In fact, Halpern notes that rational monotonicity (for conditionals) appears in the AGM postulates (in R8, in particular) [28].

Now, although plausibility models have some intuitive appeal, they validate some problematic principles. Let's start with (RM) itself. Consider the following statements:

(1a) After conditioning on me being a karate master, I consider it possible that I have a twisted ankle.

(1b) After conditioning on me being a karate master, I believe I would have won the match.

(1c) After conditioning on me being a karate master and me having a twisted ankle, I believe I would have won the match.

Supposing (1a) and (1b) somehow doesn't seem sufficient to then be able to conclude (1c).

Now let's take a closer look at axioms (PI) and (NI). These are clearly a mark of these models capturing agents' beliefs that are immutable in the face of conditioning. Notably, these axioms do not correspond to principles of AGM. One important question to answer is whether these are reasonable axioms when modeling conditioning. Consider the following two statements:

(2a) After conditioning on this match being struck, I believe the match will light.

(2b) After conditioning on this match being wet, I believe that, after I condition on this match being struck, I believe the match will light.

(2a) seems reasonable, and by positive introspection (and modus ponens), we would get to conclude (2b), which seems less reasonable.

A similar situation happens when we take a closer look at (NI). While this latter example is slightly more challenging to represent in natural language, it seems problematic to have an axiom that allows us to conclude (3b) from the very reasonable (3a):

(3a) It's not the case that after conditioning on this match being struck, I believe the match will light.

(3b) After conditioning on this match being lit, I believe that, it's not the case that after I condition on this match being struck, I believe the match will light.

We note that these examples are very similar to those in the literature on counterfactuals, which is unsurprising given our goal to outline the applicability of a counterfactual read.

3.2 Semantics

In this section we start outlining our response to some of the issues outlined above. To start, we define a language that isolates the process of conditioning, introducing a public announcement style operator which factors out "conditioning on φ". Formally, fix a finite set of countable primitive propositions PROP. The language of conditionalization, denoted $\mathcal{L}_C$, is defined recursively

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid B\varphi \mid [\psi]\varphi,$$

where $p \in PROP$ and $[\psi]\varphi$ is read, "after conditionalization on ψ, φ holds". As mentioned in the introduction, this language can be used to interpret conditional beliefs,

since $[\psi]B\varphi$ can be read as "after conditioning on φ, the agent believes ψ". We can recover Baltag and Smets' language of Conditional Doxastic Logic by considering the fragment of $\mathcal{L}_C$ where we can only apply $[\cdot]$ if it is followed by a belief operator (so the only formulas that contain $[\cdot]$ are formulas like $[\psi]B\varphi$).

Semantics for this language are given using **counterfactual models**.

Definition 14. *A* counterfactual model *is a tuple $M = (W, V, R, f)$, where:*

(P1) *W is a finite set of worlds*

(P2) *$V : PROP \to 2^W$ is a valuation function*

(P3) *$R \subseteq W \times W$ is a serial, transitive, and Euclidean binary relation, representing the set of worlds the agent considers possible*[1]

(P4) *$f : W \times \mathcal{P}(W)^+ \to \mathcal{P}(W)^+$* [2] *is a Lewis-style closest-world function, i.e. it satisfies the following three properties:*

(c1) *$f(\omega, A) \subseteq A$ (success)*

(c2) *$f(\omega, A) = \{\omega\}$ whenever $\omega \in A$ (strong centering)*

(c3) *If $f(\omega, A) \subseteq B$ and $f(\omega, B) \subseteq A$, then $f(\omega, A) = f(\omega, B)$ (uniformity)*

Although, in general, counterfactuals cannot qualify any relationships between the $A-$closest and the $A \wedge B-$closest worlds, uniformity does relate the two, in a special subcase when A and B themselves are related.

Lemma 2. *Let M be a counterfactual model. Consider $\omega \in W$ and $A, B \subseteq W$. If $f(\omega, A) \subseteq B$ then $f(\omega, A) = f(\omega, A \wedge B)$.*

[1] We write $\omega R \omega'$ as well as $\omega' \in R(\omega)$ to represent this accessibility relation.

[2] $\mathcal{P}(W)^+ = \mathcal{P}(W) - \{\emptyset\}$

Proof. We know $f(\omega, A) \subseteq A$ by success, so together with $f(\omega, A) \subseteq B$, we get $f(\omega, A) \subseteq A \wedge B$. Note that success also implies $f(\omega, A \wedge B) \subseteq A \wedge B \subseteq A$. Uniformity implies then that $f(\omega, A) = f(\omega, A \wedge B)$. □

For all $\omega \in W$ and $A \subseteq W$, we define **the agent's beliefs conditional on** A or A**-conditional beliefs** as $R^A(\omega) = \{\omega' \in f(\omega_i, A) \mid \omega_i \in R(\omega)\}$ (we will also be referring to this as the *pushforward of the agent's beliefs at* ω).

For each $M = (W, V, f, R)$ and $A \subseteq W$, we define the associated A**-conditional model** to be $M^A = (W, V, f, R^A)$, where the set of worlds the agent considers possible has been "conditionalized".

Formulas in the language are interpreted the natural way:

- $(M, \omega) \models p$ iff $p \in V(\omega)$ for $p \in PROP$.
- $(M, \omega) \models \neg\varphi$ iff $(M, \omega) \not\models \varphi$.
- $(M, \omega) \models \varphi \wedge \psi$ iff $(M, \omega) \models \varphi$ and $(M, \omega) \models \psi$.
- $(M, \omega) \models B\varphi$ iff $(M, \omega') \models \varphi$ for all $\omega' \in R(\omega)$.

Before we turn to semantics for the conditionalization operator, we need further notation. Stalnaker takes a deep dive on appropriate definitions and logics for knowledge and belief [41, 43]. A relation of interest for him is "epistemic indistinguishability", namely, worlds which the agent can't distinguish between with respect to her beliefs. In particular, consider our binary relation R on W through which we interpret belief. Then, for every $x \in W$, we can look at the set of "epistemically indistinguishable" worlds, namely $\{y : R(x) = R(y)\}$. Since the agent's beliefs coincide at these worlds, the agent can't tell these worlds apart. Also note that, in general, this is a superset for the agent's

beliefs at x. Epistemic indistinguishability is also an (easy to check) equivalence relation on W; we will call the equivalence classes brushes, and we denote them by $Br(x)$.

Recall that one complaint in the literature against similarity semantics for counterfactuals is the lack of a clear specification of what constitutes as *similar*. We claim that epistemic indistinguishability is a yet unexplored notion of closeness, but one that seems entirely reasonable. On the one hand, while counterfactuals are context-dependent, it is not obvious why this context should depend on the agent's beliefs. This kind of interaction between an agent's epistemic access and counterfactual relations in the model might sound unacceptable, if one is to interpret counterfactuals as reflecting some laws of the world we are modeling. On the other hand, given that we are modeling how agents condition their beliefs, assuming these are related to a notion of closeness brings us closer to "subjective counterfactuals", where each agent has a certain ordering on worlds, and this ordering determines the similarity semantics for that agent [12, 43].

We say the worlds where the agent has the same beliefs as at the worlds currently considered possible should be thought of as "closer" than any other worlds in the model. We want to study models where, *whenever possible*, we counterfactually shift within such a brush. Formally, we define the **subjective** property for a Lewis-style counterfactual shift function:

(c4) for all $\omega \in W$ and $A \subseteq W$, if $Br(\omega) \cap A \neq \emptyset$, then $f(\omega, A) \subseteq Br(\omega)$ (subjective)

A counterfactual model M whose counterfactual shift function satisfies $(c1) - (c4)$ is called **subjective**. We restrict our attention to subjective counterfactual models in the remaining discussion.

We propose two different semantics for the conditionalization operator. Note the dynamic style, which includes changing the model, similar to public announcements.

Also note the similarity in these semantics in the existence of a precondition.

1. $(M, \omega) \models [\varphi]\psi$ iff $[\![\varphi]\!] \neq \emptyset$ implies $(M^{\varphi}, \omega) \models \psi$; we will be referring to this semantics as **adventurous**, representing an agent that can update their beliefs, as long as the information they conditionalize on is remotely compatible with the rules of the world (i.e. the model).
2. $(M, \omega) \models [\varphi]\psi$ iff $Br(\omega) \cap [\![\varphi]\!] \neq \emptyset$ implies $(M^{\varphi}, \omega) \models \psi$; we will be referring to these semantics as **conservative**, representing an agent that updates their beliefs only if the information they conditionalize on is compatible with their epistemic indistinguishability set.

We note that in each semantics, when the precondition doesn't hold, we can represent a certain type of existential quantification, of different scopes. Define $\Diamond\varphi \equiv \neg[\varphi]\bot$.

1. For the first semantics, we have full existential quantifier over the domain: if $\omega \models [\varphi]\bot$ then the precondition fails, so there are no φ-worlds in the model; then $\Diamond\varphi$ is an existential quantifier over W.
2. For the second semantics, we have an existential quantifier over a brush: if $\omega \models [\varphi]\bot$ then the precondition fails, so there are no φ-worlds in ω's brush; then $\Diamond\varphi$ is an existential quantifier over $Br(\omega)$.

We can easily define universal quantification, using the dual. Define $\Box\varphi \equiv [\neg\varphi]\bot$. Then the first semantics can capture universal quantification over W, while the second can capture universal quantification over the brush.

Note that, if one can stay in the brush, we recover a belief relation, as we can see in the following claim.

Proposition 11. *Let $M = (W, R, V, f)$ be a subjective counterfactual model and let $A \subseteq W$. Consider $\omega \in W$ such that $Br(\omega) \cap A \neq \emptyset$. Then the restriction of R^A to $Br(\omega)$ is serial, transitive and euclidean.*

Proof. First, note that, since we are restricting to ω, where $Br(\omega) \cap A \neq \emptyset$ and f is subjective, the adventurous and conservative semantics are equivalent. Also note that, if $Br(\omega) \cap A \neq \emptyset$, we have $R^A(\omega) = R^A(\omega')$ for any $\omega' \in Br(\omega)$, so the agent's beliefs conditional on A are the same anywhere in the brush.

For seriality, let $\omega' \in Br(\omega)$. Since $Br(\omega) \cap A \neq \emptyset$ and f is well-defined, $R^A(\omega) = R^A(\omega')$ nonempty, so there exists v such that $v \in R^A(\omega')$.

For transitivity, let $w \in Br(\omega)$ and suppose wR^Av and vR^Au. Since $Br(\omega) \cap A \neq \emptyset$, we have $v \in Br(\omega)$, so vR^Au implies $u \in f(\omega_i, A) \subset Br(\omega)$ for some $\omega_i \in R(\omega)$, which implies wR^Au.

For euclidean, let $w \in Br(\omega)$ and suppose wR^Av and wR^Au. Since $Br(\omega) \cap A \neq \emptyset$, we have $u, v \in Br(\omega)$, so $R(u) = R(v) = R(\omega)$, which implies, as above, $R^A(u) = R^A(v)$, so vR^Au. □

Note that the adventurous and conservative semantics differ when we condition on an event whose satisfiability set intersects the brush trivially. Let $M = (W, R, V, f)$ be a subjective counterfactual model and let $A \subseteq W$. Consider $\omega \in W$ such that $Br(\omega) \cap A = \emptyset$. For conservative semantics, conditioning on A leads to the agent vacuously believing everything (including $\perp$). For adventurous semantics, there are still non-trivial validities when there are A-worlds in the model, but, even so, A-conditional beliefs need not be introspective. We can see this depicted in a possible representation of this situation, in Figure 3.1.

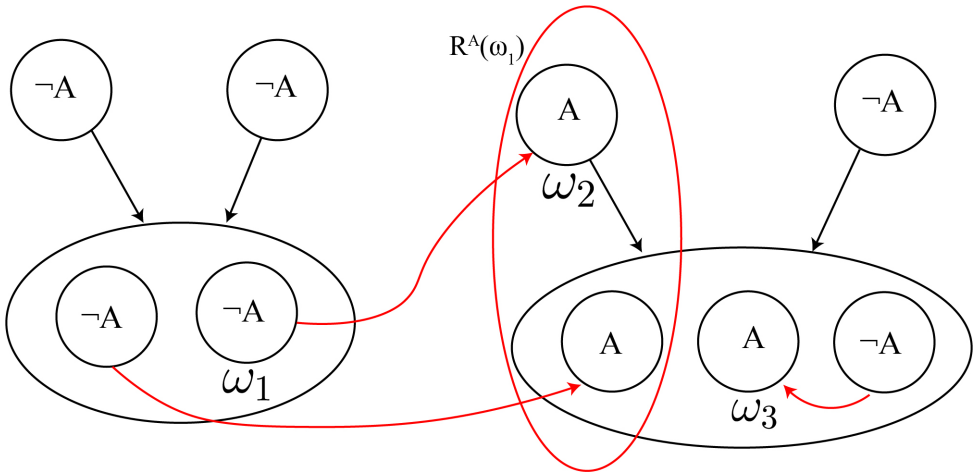

Figure 3.1: A relation R^A

The brush for world ω_1 contains no A-worlds, but W has another brush which does contain A-worlds. In the figure, the red arrows specify the closest-world counterfactual shift (so, for example, $f(\omega_1, A) = \{\omega_2\}$). In particular, note that the relation R^A is not transitive: $\omega_1 R^A \omega_2$ and $\omega_2 R^A \omega_3$, but $\omega_3 \notin R^A(\omega_1)$.

3.2.1 Soundness

In this section we study the soundness of CDL-theory and CDL_0-theory for our language, $\mathcal{L}_C$, with respect to our models. This will point out how our models differ from traditional models for conditional belief.

To start, we have the expected axiomatization for the universal and existential quantifiers defined in the previous section.

Proposition 12. *The $\Box$ and $\Diamond$ operator have an S5 axiomatization, where $\Diamond\varphi = \neg[\varphi]\bot$ and, its dual, $\Box\varphi = [\neg\varphi]\bot$:*

(K) $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$

(T) $\Box\varphi \to \varphi$

(4) $\Box\varphi \to \Box\Box\varphi$

(5) $\Diamond\varphi \to \Box\Diamond\varphi$

Proof. We prove the axioms above for an adventurous semantics - the proof for conservative semantics follows in the same manner.

(K): First, suppose $(M, \omega) \models \Box(\varphi \to \psi)$. Then $(M, \omega) \models [\varphi \wedge \neg\psi]\bot$ which implies all the worlds in the model are either $\neg\varphi$-worlds or ψ-worlds. Now suppose $(M, \omega) \models \Box\varphi$, i.e. $(M, \omega) \models [\neg\varphi]\bot$, so all worlds are φ-worlds. Since all worlds are either $\neg\varphi$-worlds or ψ-worlds, this implies all worlds are ψ-worlds, so $(M, \omega) \models \Box\psi$.

(T): Suppose $(M, \omega) \models \Box\varphi$, i.e. $(M, \omega) \models [\neg\varphi]\bot$, so all worlds are φ-worlds. Then it must be that $(M, \omega) \models \varphi$.

(4): Suppose $(M, \omega) \models \Box\varphi$. Suppose for a contradiction that there is a world ω' in the model where $(M, \omega') \models \neg\Box\varphi$, which would imply there exists a non-φ-world, a contradiction. Then $(M, \omega') \models \Box\varphi$ for all $\omega' \in W$, so, by definition, $(M, \omega) \models \Box\Box\varphi$.

(5): Suppose $(M, \omega) \models \Diamond\varphi$, so there exists a φ-world in the model. Using a similar reasoning as in (4), we get that $(M, \omega') \models \Diamond\varphi$ for all $\omega' \in W$, which implies $(M, \omega) \models \Box\Diamond\varphi$. □

Proposition 13. *The following axioms of CDL and CDL_0 hold for us (adapted to our notation):*

(CL)

(K) $[\psi]B(\varphi_1 \to \varphi_2) \to ([\psi]B\varphi_1 \to [\psi]B\varphi_2)$

(Succ) $[\psi]B\psi$

(KM) $[\psi]B\bot \to [\psi \wedge \varphi]B\bot$

(Comm) $[\varphi \wedge \psi]B\chi \to [\psi \wedge \varphi]B\chi$

(WCon) $[\psi]B\bot \to \neg\psi$

(IEa) $[\psi]B\varphi \to ([\psi \wedge \varphi]B\chi \leftrightarrow [\psi]B\chi)$

(MP) *From* φ *and* $\varphi \implies \psi$ *infer* ψ

Proof. We show that the axioms hold for a counterfactual model M and adventurous semantics for the conditioning operator. Adapting the proofs for a conservative semantics follows easily.

Axioms of (CL) clearly hold. So does (MP).

Consider axiom (K). Suppose $(M, \omega) \models [\psi]B(\varphi_1 \to \varphi_2)$. If $[\![\psi]\!]_M = \emptyset$, then there are no ψ-worlds in the domain, so $(M, \omega) \models [\psi]B\varphi_1$ and $(M, \omega) \models [\psi]B\varphi_2$ hold trivially too. Otherwise, we get $(M^\psi, \omega) \models B(\varphi_1 \to \varphi_2)$, so $(M^\psi, \omega') \models (\varphi_1 \to \varphi_2)$ for all $\omega' \in R^\psi(\omega)$ (*). Now suppose $(M, \omega) \models [\psi]B\varphi_1$, so, since we already know $W \cap [\![\psi]\!] \neq \emptyset$, we have $(M^\psi, \omega) \models B\varphi_1$. This is equivalent to $(M^\psi, \omega') \models \varphi_1$ for all $\omega' \in R^\psi(\omega)$. This together with (*) implies $(M^\psi, \omega') \models \varphi_2$ for all $\omega' \in R^\psi(\omega)$, so, by definition $(M, \omega) \models [\psi]B\varphi_2$.

Consider axiom (Succ). It holds by construction, since f satisfies success, so it must be that ψ holds at all closest ψ-worlds.

Consider axiom (KM). Suppose $(M, \omega) \models [\psi]B\bot$. This implies there are no ψ worlds in the model, so there are no $\psi \wedge \varphi$ worlds either, so $(M, \omega) \models [\psi \wedge \varphi]B\bot$.

Consider axiom (Comm). This is valid since $f(\omega, [\![\varphi \wedge \psi]\!]) = f(\omega, [\![\psi \wedge \varphi]\!])$ for any

$\omega \in \Omega$.

Consider axiom (WCon). Suppose $(M, \omega) \models [\psi]B\bot$ implies there are no ψ worlds in the model, so $(M, \omega) \models \neg\psi$.

Consider axiom (IEa). Suppose $(M, \omega) \models [\psi]B\varphi$. If $W \cap [\![\psi]\!] = \emptyset$, then $W \cap [\![\psi \wedge \varphi]\!] = \emptyset$ so it all holds trivially. Otherwise, we have $f(\omega', [\![\psi]\!]) \subseteq [\![\varphi \wedge \psi]\!]$ for all $\omega' \in R(\omega)$. By uniformity, this is equivalent to $f(\omega', [\![\psi]\!]) = f(\omega', [\![\psi \wedge \varphi]\!])$ for all $\omega' \in R(\omega)$. We then have $(M, \omega) \models [\psi \wedge \varphi]B\chi \leftrightarrow [\psi]B\chi$. [3]

□

Proposition 14. *The following axioms are valid for $\mathcal{L}_C$ with respect to subjective counterfactual models:*

Introspective conditioning	$\Diamond\varphi \rightarrow B\Diamond\varphi$
Serial conditioning	$B\varphi \rightarrow \Diamond\varphi$
Belief and conditioning	$[\psi]B\varphi \rightarrow B[\psi]B\varphi$

Proof. Once more, we consider adventurous semantics for the $[\cdot]$ operator, and note the proofs for the conservative semantics follow in the same manner.

Consider the axiom Introspective Conditioning. Suppose $(M, \omega) \models \Diamond\varphi$, i.e. $(M, \omega) \models \neg[\varphi]\bot$. If $[\![\varphi]\!] = \emptyset$, then we would have $(M, \omega) \models [\varphi]\bot$, so $[\![\varphi]\!] \neq \emptyset$. Suppose for a contradiction $(M, \omega) \models \neg B\Diamond\varphi$, so there exists $\omega' \in R(\omega)$ such that $(M, \omega') \models \neg\Diamond\varphi$, i.e. $(M, \omega') \models [\varphi]\bot$, which implies $[\![\varphi]\!] = \emptyset$, a contradiction. Hence $(M, \omega) \models B\Diamond\varphi$.

Consider the axiom Serial Conditioning. Suppose $(M, \omega) \models B\varphi$. Then we must have $[\![\varphi]\!] \neq \emptyset$, so $(M, \omega) \models \Diamond\varphi$.

[3] We actually have $[\psi]B\varphi \rightarrow ([\psi \wedge \varphi]\chi \leftrightarrow [\psi]\chi)$.

Consider the axiom Belief and Conditioning. Suppose $(M, \omega) \models [\psi]B\varphi$. If $[\![\psi]\!] = \emptyset$, then the consequent holds trivially. Otherwise, we have $(M^\psi, \omega) \models B\varphi$, and since $R^\psi(\omega) = R^\psi(\omega')$ for all $\omega' \in R(\omega)$, we have $(M^\psi, \omega') \models B\varphi$ for all $\omega' \in R(\omega)$. The consequent follows. □

Proposition 15. *The following axioms of CDL and CDL_0 don't, in general, hold for $\mathcal{L}_C$ with respect to subjective counterfactual models:*

(PI) $[\psi]B\chi \to [\varphi]B[\psi]B\chi$

(NI) $\neg[\psi]B\chi \to [\varphi]B\neg[\psi]B\chi$

(Inc) $[\psi \wedge \varphi]B\chi \to [\psi]B(\varphi \to \chi)$

(IE b) $\neg[\psi]B\neg\varphi \to ([\psi \wedge \varphi]B\chi \leftrightarrow [\psi]B\chi)$

(RM) $\neg[\psi]B\neg\varphi \to ([\psi]B\chi \to [\psi \wedge \varphi]B\chi)$

Proof. First, note that (PI) and (NI) are axioms reflecting the static nature of conditional belief semantics and, as argued in Section 3.1.1, we don't want these axioms to hold in our semantics. We can see that they are, indeed, not valid, by considering the following counterexamples.

For (PI), consider the model in Figure 3.2.

This clearly satisfies KD45 for R, success, subjectivity, and strong centering for f. The trickier property to prove is uniformity. Recall uniformity requires that, if $f(\omega, A) \subseteq B$ and $f(\omega, B) \subseteq A$ for any $\omega \in W$, and $A, B \subseteq W$, then $f(\omega, A) = f(\omega, B)$. Note that uniformity does not apply if $A \cap B = \emptyset$. Then we have to check the following cases:

Case 1. $A = \{\omega_1\}$, $B = \{\omega_1, \omega_2\}$ (using symmetry, we don't need to check when $A = \{\omega_1, \omega_2\}$, $B = \{\omega_1\}$). Now, for ω_1, we clearly have (by strong centering) that $f(\omega_1, A) = f(\omega_1, B) = \{\omega_1\}$. For ω_2, we have $f(\omega_2, B) = \{\omega_2\}$ (by strong centering), and $\omega_2 \notin A$, so uniformity does not apply.

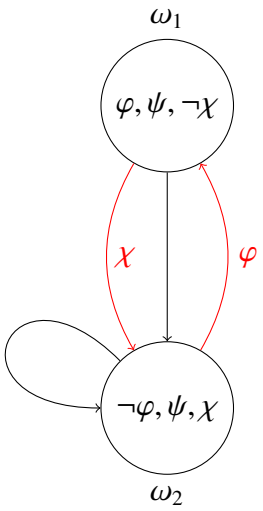

Figure 3.2: A counterexample for PI

Case 2. $A = \{\omega_2\}$, $B = \{\omega_1, \omega_2\}$ (using symmetry, we don't need to check when $A = \{\omega_1, \omega_2\}$, $B = \{\omega_2\}$). Now, for ω_2, we clearly have (by strong centering) that $f(\omega_2, A) = f(\omega_2, B) = \{\omega_2\}$. For ω_1, we have $f(\omega_1, B) = \{\omega_1\}$ (by strong centering), and $\omega_1 \notin A$, so uniformity does not apply.

So the model defined in Figure 3.2 is a subjective counterfactual model. Now consider adventurous semantics [4]. Note that at ω_1, we have $(M, \omega_1) \models [\psi]B\chi$, since the only world considered possible, ω_2, is already a $\psi \wedge \chi$-world; however, $(M, \omega_1) \not\models [\varphi]B[\psi]B\chi$, since ω_2 is not a φ-world, conditioning to the closest φ-world leads to ω_1, which is a ψ-world, but not a χ-world.

For (NI) consider the model in Figure 3.3.

This model clearly satisfies KD45 for R, success, subjectivity, and strong centering for f. As for (PI), uniformity is more tricky to prove, but the small number of worlds makes it as straightforward as for (PI) (with exactly the same cases considered).

So the model defined in Figure 3.3 is a subjective counterfactual model. Now con-

[4]The proof follows in the same manner for conservative semantics.

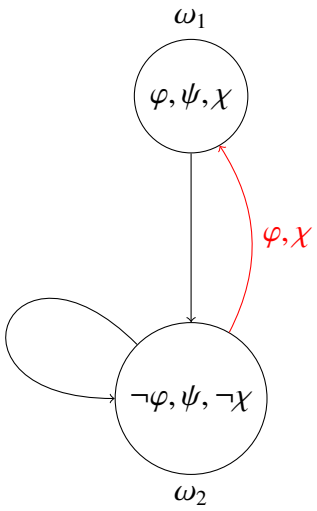

Figure 3.3: A counterexample for NI

sider adventurous semantics. Note that at ω_1, we have $(M, \omega_1) \models \neg[\psi]B\chi$, since the only world considered possible, ω_2, is a $\psi \wedge \neg\chi$-world; however, $(M, \omega_1) \not\models [\varphi]B\neg[\psi]B\chi$, since ω_2 is not a φ-world, conditioning to the closest φ-world leads to ω_1, which is a $\psi \wedge \chi$-world.

On the other hand, the axioms (Inc), (RM) and (IEb) don't hold, all for a similar reason, namely the counterfactual semantics for the conditionalization operator. In general, unless the closest ψ-worlds are all φ worlds (in which case, by Lemma 2, the closest ψ-worlds are precisely the closest $\psi \wedge \varphi$-worlds), the closest ψ-worlds need not be (even related to) the closest $\psi \wedge \varphi$-worlds. This is precisely the issue of antecedent monotonicity.

A restriction of (Inc) does hold (easy to check using Lemma 2): $[\psi]B\varphi \rightarrow ([\psi \wedge \varphi]B\chi \rightarrow [\psi]B(\varphi \rightarrow \chi))$. There are other special cases we could consider. To start, let $\psi = \top$. Then (Inc) becomes $[\varphi]B\chi \rightarrow [\top]B(\varphi \rightarrow \chi)$ which is valid in our models (both semantics). This shows us another case in which (Inc) holds, namely if $[\![\psi]\!] \subseteq [\![\varphi]\!]$. The resulting axiom is $[\psi]B\chi \rightarrow [\psi]B(\varphi \rightarrow \chi)$, which clearly holds. Slightly more generally, if $f(\omega, [\![\psi]\!]) \subseteq [\![\varphi]\!]$, then by Lemma 2, (Inc) holds again.

□

The axioms in Proposition 13 and 14 are all sound for our language with respect to subjective counterfactual models. We believe they also form a complete axiomatization:

Conjecture 1. *The axioms in Proposition 13 and 14 form a sound and complete axiomatization for the fragment of $\mathcal{L}_C$ that does not include iterated $[\cdot]$ operators, with respect to subjective counterfactual models.*

3.3 Discussion

We proposed adventurous and conservative semantics for conditional beliefs which continue the tradition of Dynamic Epistemic Logic. This approach for interpreting conditional beliefs focuses on a kinetic, counterfactual read, where "believing φ conditional on ψ" is read as, "if ψ were the case, the agent would believe φ". We saw the effects of this interpretation on validities for higher-order conditional beliefs in this logic, but a sound and complete axiomatization remains an open problem.

Further, we noticed that there are a number of axioms which hold intrinsically in existing theories for conditional beliefs, and don't in counterfactual models, like positive and negative introspection. We leave for future work the study of restrictions needed to be imposed on counterfactuals in order to obtain an equivalence with the existing plausibility models that interpret conditional beliefs.

Finally, we also turned our attention to the conditions under which A-conditional beliefs remain introspective (as unconditional beliefs are). There are a number of questions here that remain open, from whether we can preserve introspection with weaker conditions than subjectivity, to whether there is a place for adventurous models.

CHAPTER 4

EXTENSIONS FOR PREVIOUS RESULTS

4.1 The language of counterfactuals

We have seen that counterfactuals arise, naturally, when agents reason about the best thing to do in extensive form games [12, 42, 26]. Chapter 2 continued the work in this field by taking a closer look at how counterfactuals and beliefs interact in extensive-form games. A natural next step is to include counterfactuals explicitly in the language. While this seems like a small change, this cascades into interesting consequences. Consider the following game:

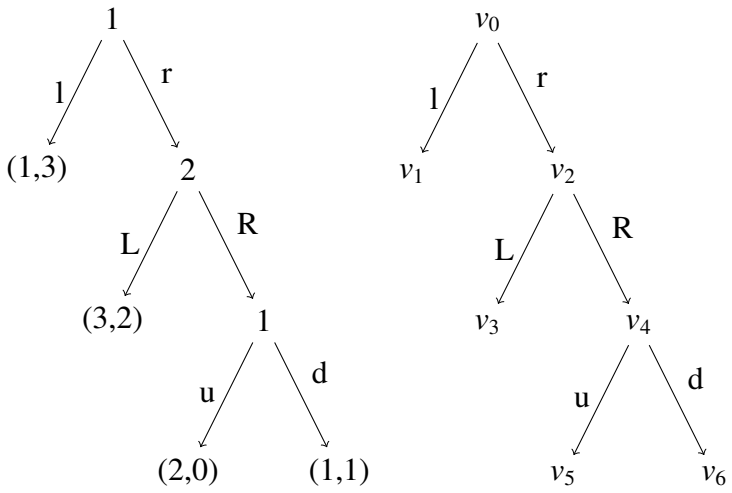

Figure 4.1: The game (left) and associated vertices (right)

Suppose player 1 believes player 2 would choose L if given the chance, so is planning to play l; they are trying to reason about what would happen if player 2 would choose R instead. Now, since both players have access to the game tree, a traditional representation of this setting in game theory, similar to the one in Chapter 2, would require associating a vertex with each node, like the figure on the right. Then player 1's

reasoning above basically asks her to consider what is the case at the closest v_2-world where player 2 chooses R, which requires primitive access, in the model, to player's strategies, or checking whether node v_4 is reached. It turns out that, with a more precise specification of the language agents use to reason about such a game, we no longer need to have the vertex-representation in mind when reasoning about the game, nor do we need primitive access to strategies. We are studying static settings, where agents reason a priori to the game starting about how the game will unfold, but we are not including time in our models. Hence, a state in our models specifies a way in which the game unfolds, i.e. an outcome. Outcomes are, simply, sequences of moves in the tree (all the way down to a leaf), so we want our agents to be able to specify, in the language, the moves they choose, and when they do so. Then, player 1 reasoning about the closest v_2-world where v_4 is reached can be expressed as the closest world where player 1 chooses r, and the subsequent move in whatever outcome is determined at said world is R.

We consider, then, a language which includes beliefs and counterfactuals. This can capture statements like, "player 2 believes player 1 will choose l at the onset of the game", or "if player 1 were to choose r at the onset, player 2 would choose L". Now, recall that in Chapter 2, we argue that, since strategies can be thought of as complete specifications of what a player would do at every history they are due to play, they are counterfactual objects. We outline below how easily strategies can be represented in our language. Further, once we turn to defining rationality in extensive-form games, our language enables us to take a closer look at how counterfactuals and beliefs interact. In particular, note that determining rationality implies determining whether switching strategies would leave someone better off. Switching implies a counterfactual shift, but since strategies are counterfactual objects, reasoning about switching strategies becomes reasoning about iterated counterfactuals. There is no existing work in the literature (that we have seen) that adopts this iterated counterfactual approach to model rationality.

The rest of this chapter goes as follows. To start, we adapt the model introduced in Chapter 2, Section 2.3.3 to a more general setting which includes counterfactuals in the language. Many of the results included in that chapter follow easily. The interesting changes occur once we look at definitions of rationality.

4.1.1 Counterfactual models for extensive-form games

An **extensive form game** is a tuple $\Gamma = (N, H, P, (u_i)_{i \in N})$ where:

(P1) $N = \{1, \ldots, n\}$ is the set of *players*.

(P2) H is a set of sequences called *histories*, intuitively describing paths in the game tree. A *terminal history* or *outcome* is a member of H that has no proper extension in H. Denote the set of terminal histories by Z.

(P3) $P : H \setminus Z \to N$ is the *player function*, where intuitively $P(h) = i$ means that player i is due to act at history h. Let $H_i = \{h \in H : P(h) = i\}$.

(P4) $u_i : Z \to \mathbb{R}$ is a *utility function* for every player $i \in N$.

For each $h \in H \setminus Z$, let $A(h) = \{a : (h, a) \in H\}$; in other words, $A(h)$ is the set of *actions* available to player $P(h)$ at history h. To streamline the presentation, we restrict our attention in this paper to *finite* games, that is, games where Z is finite.

Next we build a logical language for reasoning about extensive form games, starting with the set of primitive propositions

$$\Phi_\Gamma = \{move_i(h, a_i) : i \in N, a_i \in A(h), P(h) = i\}.$$

We read $move_i(h, a_i)$ as "player i moves from history h by taking action a_i". The requirement $P(h) = i$ ensures that it is player i's turn at history h, and $a_i \in A(h)$ ensures

that the move is legal, in that it generates a well-defined history $(h, a_i) \in H$. Closing this set under negation, conjunction, unary belief modalities B_i for $i \in N$, and the binary counterfactual modality $\Box\!\!\rightarrow$ generates $\mathcal{L}_{EFG}$. This is our object language.

By conjoining primitives, the language can describe not only individual moves but also whole histories of play, which will be particularly useful in representing players' plans of action at different histories of the game. For each $h \in H$, define $\varphi_h \in \mathcal{L}_{EFG}$ recursively: for the history $\emptyset$ let $\varphi_\emptyset = \top$; now, supposing φ_h is defined, define $\varphi_{(h,a)} = \varphi_h \wedge move_i(h, a)$, where $P(h) = i$. Intuitively, φ_h describes the moves in the game up to history h. For example, in the market-entry game, we have $\varphi_{(I,F)} = \top \wedge move_1(\emptyset, I) \wedge move_2((I), F)$.

Semantics for this language are given by a generalized version of the *counterfactual models* introduced in Chapter 2 where the closest-world function is defined on any nonempty subset of the state space, instead of, simply, sets which specify where certain vertices are reached. As motivated in the introduction, this is so that we can model agents who reason about switching their strategies, which for us will be counterfactual objects as well.

Consider an adaptation of *counterfactual-models* from Section 2.3.3 to this more general counterfactual shift, so we define $M = (\Omega, (\mathcal{PR}_i)_{i \in N}, \mathcal{O}, f)$:

(R1) Ω is a nonempty, finite state space.

(R2) $\mathcal{PR}_i : \Omega \to \Delta(\Omega)$ assigns to each $\omega \in \Omega$ a probability measure $\mathcal{PR}_i(\omega)$ on Ω representing the beliefs of player i at ω.

(R3) $\mathcal{PR}_i(\omega)(\{\omega' : \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)\}) = 1$ (i.e., players are sure of their own beliefs).

(R4) $\mathcal{O} : \Omega \to Z$ assigns a unique outcome to every state.

(R5) $f : \Omega \times (2^{\Omega} \setminus \emptyset) \to \Omega$ is a "closest world" function satisfying $f(\omega, T) \in T$ and such that if $\omega \in T$, then $f(\omega, T) = \omega$ (intuitively, f maps (ω, T) to the T-world most similar to ω).[1]

Formulas in the language $\mathcal{L}_{EFG}$ are interpreted in such models recursively, as follows:

$$\begin{aligned}
[\![move_i(h, a_i)]\!] &= \{\omega \in \Omega : (h, a_i) \text{ is an initial segment of } O(\omega)\} \\
[\![\neg\varphi]\!] &= \Omega \setminus [\![\varphi]\!] \\
[\![\varphi \wedge \psi]\!] &= [\![\varphi]\!] \cap [\![\psi]\!] \\
[\![B_i\varphi]\!] &= \{\omega \in \Omega : \mathcal{PR}_i(\omega)([\![\varphi]\!]) = 1\} \\
[\![\varphi \mathrel{\square\!\!\rightarrow} \psi]\!] &= \{\omega \in \Omega : f(\omega, [\![\varphi]\!]) \in [\![\psi]\!]\}.
\end{aligned}$$

Thus, worlds in our models encode outcomes via the function O, B_i is interpreted as probability 1 belief, and counterfactual expressions are interpreted in the standard way: $\varphi \mathrel{\square\!\!\rightarrow} \psi$ is true at ω just in case ψ is true at the φ-world that is "most similar" to ω, as picked out by the function f.

Conditions (C1) and (C2) introduced and motivated in Chapter 2 are easily rewritten in this new setting:

(C1) $\forall\omega \in \Omega$, $\forall i \in N$, $\forall h \in H_i$, if (h, a_i) is an initial segment of $O(f(\omega, [\![\varphi_h]\!]))$ for some a_i, then for all $\omega' \in \Omega$ such that $\mathcal{PR}_i(\omega)(\omega') > 0$, (h, a_i) is an initial segment of $O(f(\omega', [\![\varphi_h]\!]))$.

[1] It's common to include a general coherency requirement for counterfactual shift functions, namely that for all $\omega \in \Omega$, and for all $T, U \in 2^{\Omega}$, if $f(\omega, T) \in U$, then $f(\omega, T) = f(\omega, T \cap U)$. This simply says that if the closest T-world is a U world, then it is also the closest T-and-U-world. For the present work, we omit it for the sake of generality, though it could certainly be imposed.

(C2) $\forall \omega \in \Omega$, $\forall h \in H$, $\forall a \in A(h)$ we have

$$f(f(\omega, [\![\varphi_h]\!]), [\![\varphi_{(h,a)}]\!]) = f(\omega, [\![\varphi_{(h,a)}]\!]).$$

Recall that condition (C1) essentially guarantees that players are sure of their own *strategies*. Condition (C2) says, roughly speaking, that the counterfactual shift operator respects the temporal nature of the game in the sense that the closest (h, a)-world to ω is also the closest (h, a)-world to the closest h-world to ω; this will allow us to compute the expected utility of playing a strategy in a certain subgame. Also recall that, although it seems that condition (C2) is restricted to a single move after a history, this condition holds across all extensions, and an analagous result to Lemma 1 (Section 2.3.3) easily holds in this setting as well:

Lemma 3. *For all $\omega \in \Omega$ and $h \in H$, if h' is a sequence of moves such that $(h, h') \in H$, we have*

$$f(f(\omega, [\![\varphi_h]\!]), [\![\varphi_{(h,h')}]\!]) = f(\omega, [\![\varphi_{(h,h')}]\!]).$$

Denote the class of all models satisfying (R1)–(R5) and (C1)–(C2) by $\mathcal{M}_{EFG}$, and call these *models for* Γ.

Representing strategies

Our previous work already showed how we can represent strategies. We take the time in this section to reiterate our previous results in this new semantics.

Recall that a **strategy for player** i is a function ρ_i mapping histories where player i is due to play to actions such that for each $h \in H_i$ we have $\rho_i(h) \in A(h)$.

Fix a game Γ and let $M \in \mathcal{M}_{EFG}$ be a model for this game. Denote by Λ_i the set of all strategies for player i. Recall that we can use the counterfactual operator to answer

the question "What would player i if history h occurred?", by switching to the closest world where history h does occur – provided the model is rich enough. In Chapter 2, this richness was ensured by the fact that f was defined as a total function, but as f is defined in greater generality here, we need to consider a certain subset of the class of models. Say that a model for Γ is **history-rich** if O is surjective.[2]

Then we can recover strategies, as expected (similar to Proposition 4, Section 2.3.3):

Proposition 16. *Let $M = (\Omega, (\mathcal{PR}_i)_{i\in N}, O, f)$ be a history-rich model for Γ. For all $\omega \in \Omega$, for all $i \in N$, and for all nonterminal $h \in H_i$, there is a unique $s_i(\omega)(h) \in A(h)$ such that $\omega \models \varphi_h \boxdot\!\!\rightarrow move_i(h, s_i(\omega)(h))$.*

Proof. Since O is surjective, $[\![\varphi_h]\!] \neq \emptyset$, therefore $f(\omega, [\![\varphi_h]\!])$ is defined. Since h is nonterminal, it follows that there exists an $a \in A(h)$ such that $\omega \models \varphi_h \boxdot\!\!\rightarrow move_i(h, a)$. Now suppose there exist $a_i, a_i' \in A(h)$ such that $\omega \models \varphi_h \boxdot\!\!\rightarrow move_i(h, a_i)$ and $\omega \models \varphi_h \boxdot\!\!\rightarrow move_i(h, a_i')$. Then $f(\omega, [\![\varphi_h]\!]) \in [\![move_i(h, a_i)]\!]$, so (h, a_i) is an initial segment of $O(f(\omega, [\![\varphi_h]\!]))$, and similarly (h, a_i') is an initial segment of $O(f(\omega, [\![\varphi_h]\!]))$. By (R4), since O assigns a unique outcome to every state, we have $(h, a_i) = (h, a_i')$, so $a_i = a_i'$. □

Recall then that we can think of each $s_i : \Omega \to \Lambda_i$ as a function assigning a strategy for player i to each world. In light of the importance of these functions, we henceforth restrict attention to history-rich models.

It easily follows that players are sure of their own strategies, or, formally:

Proposition 17. *For every $\omega \in \Omega$ and each $i \in N$, we have $\mathcal{PR}_i(\omega)(\{\omega' : s_i(\omega') = s_i(\omega)\}) = 1$.*

[2]This richness condition can be weakened slightly: it is only necessary that every *non-terminal* node in the game tree be actualized at some world. Nonetheless, we opt for this more succinct and slightly stronger formulation since the extra richness will be required later, when we discuss rationality and strategy shifts.

Proof. The proof is analagous to that for Proposition 5, adapted to this setting.

It is sufficient to prove that for every $\omega' \in \Omega$ such that $\mathcal{PR}_i(\omega)(\omega') > 0$ we have $s_i(\omega') = s_i(\omega)$. Let ω' be as above, and consider nonterminal $h \in H_i$. By Proposition 16, $s_i(\omega)(h) = a$ for a unique $a \in A(h)$. Then $\omega \models \varphi_h \; \Box\!\!\rightarrow \; move_i(h, a)$, so (h, a) is an initial segment to $O(f(\omega, [\![\varphi_h]\!]))$. By (C2), that implies (h, a) an initial segment to $O(f(\omega', [\![\varphi_h]\!]))$. Again by Proposition 16, we know that $s_i(\omega')(h) = a'$ for a unique $a' \in A(h)$, so (h, a') an initial segment to $O(f(\omega', [\![\varphi_h]\!]))$, which implies that $a = a'$, i.e. $s_i(\omega)(h) = s_i(\omega')(h)$. Since this equality holds for arbitrary h, we have $s_i(\omega) = s_i(\omega')$ for every $\omega' \in \Omega$ such that $\mathcal{PR}_i(\omega)(\omega') > 0$, as desired. □

In Section 2.3.3 we identified the unique terminal history reached from a vertex given a strategy profile - we need a similar thing here, for histories instead. Then for each $h \in H$, let $[h] \subseteq Z$ be the set of terminal histories that extend h. Denote the unique terminal history in $[h]$ reached from h when the players play according to ρ by $[h]_\rho$. In particular, $[\emptyset]_\rho$ is the outcome of the game induced by ρ.

We also have the analog of Proposition 6 for this setting, so the strategies associated with any given world actually induce the outcome that is assigned to that world.

Proposition 18. *For all $\omega \in \Omega$, the terminal history induced by $s(\omega) = (s_1(\omega), \ldots, s_n(\omega))$ is precisely $O(\omega)$. In other words, $[\emptyset]_{s(\omega)} = O(\omega)$.*

Proof. The proof follows that for Proposition 6 precisely.

For a contradiction, suppose not; that is, suppose that $[\emptyset]_{s(\omega)} \neq O(\omega)$. Then there exists $h \in H$ such that h is a maximal initial segment for both of these terminal histories, but h is not itself terminal. Let a be such that (h, a) is an initial segment for $O(\omega)$. Then by assumption $s_{P(h)}(\omega)(h) = a' \neq a$ (for some a'). This means that $\omega \models \varphi_h \; \Box\!\!\rightarrow$

$move_{P(h)}(h, a')$, so $f(\omega, [\![\varphi_h]\!]) \in [\![move_{P(h)}(h, a')]\!]$. But $\omega \in [\![\varphi_h]\!]$, so $f(\omega, [\![\varphi_h]\!]) = \omega$, hence $\omega \in [\![move_{P(h)}(h, a')]\!]$, so (h, a') is an initial segment of $O(\omega)$, contradicting the fact that $a \neq a'$. □

Now then that by conjoining counterfactual statements, the language can encode full strategies. For each strategy $\rho_i \in \Lambda_i$, let φ_{ρ_i} denote the formula

$$\bigwedge_{h_i \in H_i} \varphi_{h_i} \boxdot\!\!\rightarrow move_i(h_i, \rho_i(h_i)).$$

Proposition 19. *For all $\omega \in \Omega$, we have $\omega \models \varphi_{\rho_i}$ iff $s_i(\omega) = \rho_i$.*

Proof. First suppose $\omega \models \varphi_{\rho_i}$. Let $h_i \in H_i$; then $\omega \models \varphi_{h_i} \boxdot\!\!\rightarrow move_i(h_i, \rho_i(h_i))$, so by definition $s_i(\omega)(h_i) = \rho_i(h_i)$. Since $h_i \in H_i$ was arbitrary, this shows that $s_i(\omega) = \rho_i$. Conversely, suppose that $\omega \not\models \varphi_{\rho_i}$; then there is some $h_i \in H_i$ such that $\omega \not\models \varphi_{h_i} \boxdot\!\!\rightarrow move_i(h_i, \rho_i(h_i))$, from which it follows that $s_i(\omega)(h_i) \neq \rho_i(h_i)$, hence $s_i(\omega) \neq \rho_i$. □

A standard assumption in classical game theory is that, as players consider switching strategies, their opponents are unaffected by this "mental exercise", and as such, their strategies do not change. It will be useful to have a name for models with this "opaqueness" property (cf. [29]): say that a model is **opaque** if it satisfies the following:

(C3) $\forall \omega \in \Omega$, $\forall i \in N$, we have $s_j(f(\omega, [\![\varphi_{\rho_i}]\!])) = s_j(\omega)$ for all $j \in N$ with $j \neq i$.

4.1.2 Rational Play

Best Responding

To start, we look at defining the notion of rationality in the classical setting within our models. Recall that a player is *rational* if they are best responding to their beliefs about

the strategies their opponents are using—that is, if their chosen strategy maximizes their expected utility across all their possible strategies. This is easy to define in our models. Each probability measure on worlds induces a probability measure on opponents' strategies via the functions s_i. This captures players' uncertainties about the strategies their opponents are using. We can then compute each player's expected utility for playing a given strategy at a world:

Definition 15. *For every $i \in N$, $\omega \in \Omega$, and $\rho_i \in \Lambda_i$, define*

$$EU_i(\omega, \rho_i) = \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(\mathcal{O}(f(\omega', [\![\varphi_{\rho_i}]\!])).$$

Thus, for each possible world ω' that player i considers possible, we compute the utility player i *would* get if they were to play ρ_i at that world (which is just the utility of the outcome associated with the closest ρ_i-world to ω', namely $f(\omega', [\![\varphi_{\rho_i}]\!])$); this value is then weighted by the probability player i assigns to ω'. Note that when $\rho_i = s_i(\omega)$, by Proposition 17, for each ω' with $\mathcal{PR}_i(\omega)(\omega') > 0$ we have $s_i(\omega') = s_i(\omega)$, and therefore $f(\omega', [\![\varphi_{\rho_i}]\!]) = \omega'$, so this definition reduces to player i's actual expected utility at ω:

$$EU_i(\omega, s_i(\omega)) = \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(\mathcal{O}(\omega')).$$

Observe that $EU_i(\omega, \rho_i)$ is not always defined—it crashes whenever $[\![\varphi_{\rho_i}]\!] = \emptyset$. Thus, to make sense of expected utility, we must assume this does not occur. Call a model in which each s_i is surjective **strategy-rich**. In light of the importance of strategy-richness for articulating the notion of best response, we henceforth restrict our attention to models that satisfy it. Extensive-form games can easily have very large strategy spaces, making this condition highly non-trivial; still, in order for a player to actually compute their expected utility, they must consider in turn how things would be if they were to play each and every one of their alternative strategies. For this reason we do not feel that

strategy-richness, as a condition on models, oversteps the bounds of what it is reasonable for models appropriate for representing rationality to encode.

We can collect at each world ω the set of strategies that maximize player i's expected utility,

$$BR_i(\omega) = \{\rho_i \in \Lambda_i \ : \ \forall \rho_i' \in \Lambda_i, \ EU_i(\omega, \rho_i) \geq EU_i(\omega, \rho_i')\},$$

and extend the object language with new propositional constants RAT_i, read "player i is rational", interpreted in the natural way: $[\![RAT_i]\!] = \{\omega \in \Omega : s_i(\omega) \in BR_i(\omega)\}$.

Subgame Rationality

We have already seen in Chapter 2 that the appropriate notion of rationality for extensive-form games seems to be subgame rationality. Interestingly, however, once we add counterfactuals to our language, defining subgame-rationality is no longer as straightforward.

Recall that we need to define the expected utility of playing a strategy after a given history has occurred; this will allow us to express whether a player's strategy is optimal not only at the start of the game, but after every possible history where they are due to play. This involves yet another kind of counterfactual shift: roughly speaking, for each world and each player, we want to check what *would* be the case if history h were to occur. Here an interesting subtlety arises. The concept of subgame perfect equilibrium might be read as implying a specific order of operations: first, we consider a given subgame h (which in our setting corresponds to a first-order counterfactual shift, moving to the closest h-world); next, player i considers switching strategies to check whether their current strategy is expected utility maximizing (which in our setting corresponds to second degree counterfactual shifts, moving to the closest ρ_i-worlds, as in Definition

15).

Roughly speaking, if we try to implement this in our setting—shifting first to the closest h-world, then to the closest ρ_i-world from there—it looks like $f(f(\omega, [\![\varphi_h]\!]), [\![\varphi_{\rho_i}]\!])$. However, the closest ρ_i-world to the closest h-world might not be an h-world, so we couldn't use this world to compute player i's expected utility of playing ρ_i after history h occurred. And while we do want to switch to a world where both history h has occurred and afterwards ρ_i is employed, shifting to the closest $\varphi_h \wedge \varphi_{\rho_i}$ world does not always make sense: some strategies are incompatible with certain histories occurring, in which case $[\![\varphi_h \wedge \varphi_{\rho_i}]\!] = \emptyset$ no matter how rich the model is.

Hence, in order to compute the expected utility of playing a given strategy after history h has occurred, we need a somewhat subtler approach, one that essentially reverses this order of operations. To begin, the following lemma will be useful. Essentially, it says that switching to the closest h-world does not change any of the players' strategies *from history h onward*. Note that this requirement looks very similar to F3 in Halpern's models from Chapter 2.

Lemma 4. *Let $h \in H$, $\omega \in \Omega$, and let $\omega' = f(\omega, [\![\varphi_h]\!])$. Then for every $i \in N$, we have $s_i(\omega)(h_i) = s_i(\omega')(h_i)$ for all $h_i \in H_i$ extending h.*

Proof. Let $i \in N$ and consider $h_i \in H_i$ for which h is an initial segment. By Lemma 3, we have $f(\omega', [\![\varphi_{h_i}]\!]) = f(\omega, [\![\varphi_{h_i}]\!])$. Then there exists a unique a_i such that $f(\omega', [\![\varphi_{h_i}]\!]), f(\omega, [\![\varphi_{h_i}]\!]) \in [\![move_i(h_i, a_i)]\!]$, so ω and ω' both satisfy $\varphi_{h_i} \boxdot\!\!\rightarrow move_i(h_1, a_i)$, from which it follows that $s_i(\omega)(h_i) = s_i(\omega')(h_i)$, as desired. □

We next define a counterfactual probability distribution reflecting what player i believes *would* be the case if they were to switch to another strategy (this is precisely the notion of a player's counterfactual belief at state ω from [29], except adapted here to an

extensive form setting). For every $i \in N$ and $\omega \in \Omega$, define

$$\mathcal{PR}_i^{\rho_i}(\omega)(\omega') = \sum_{\{\omega'': f(\omega'', [\![\varphi_{\rho_i}]\!]) = \omega'\}} \mathcal{PR}_i(\omega)(\omega''). \tag{4.1}$$

Note that $\mathcal{PR}_i^{\rho_i}(\omega)([\![\varphi_{\rho_i}]\!]) = 1$, since

$$\begin{aligned}
\mathcal{PR}_i^{\rho_i}(\omega)([\![\varphi_{\rho_i}]\!]) &= \sum_{\omega' \in [\![\varphi_{\rho_i}]\!]} \mathcal{PR}_i^{\rho_i}(\omega)(\omega') \\
&= \sum_{\omega' \in [\![\varphi_{\rho_i}]\!]} \sum_{\{\omega'': f(\omega'', [\![\varphi_{\rho_i}]\!]) = \omega'\}} \mathcal{PR}_i(\omega)(\omega'') \\
&= \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega').
\end{aligned}$$

Using these "counterfactual beliefs", we can define the expected utility of player i playing any given strategy, supposing history h has occurred:

Definition 16. *For every $i \in N$, $h \in H$, and $\rho_i \in \Lambda_i$, define*

$$EU_i^h(\omega, \rho_i) = \sum_{\omega' \in \Omega} \mathcal{PR}_i^{\rho_i}(\omega)(\omega') \cdot u_i(O(f(\omega', [\![\varphi_h]\!]))).$$

By Lemma 4, for all $j \in N$ and all $h_j \in H_j$ extending h, we have $s_j(f(\omega', [\![\varphi_h]\!]))(h_j) = s_j(\omega')(h_j)$, so the players' strategies at $f(\omega', [\![\varphi_h]\!])$ agree with their strategies at ω' *when restricted to histories extending h*. This means that $O(f(\omega', [\![\varphi_h]\!])) = [h]_{s(\omega')}$, so this definition of expected utility at a subgame actually computes the utilities associated with playing different strategies *in that subgame*, as we would want it to.

Notice that there are two counterfactual shifts at play in Definition 16—one for strategies, and one for histories. But we avoid the previous problem by, essentially, implementing the shift to the closest ρ_i-world first (in the definition of counterfactual belief), and then inside that context we shift to the closest h-world, relying on Lemma 4 to ensure that in so doing we do not change ρ_i below h. The following lemma makes this precise.

Lemma 5. *For all $i \in N$, $h_i \in H_i$, $\omega \in \Omega$, and $\rho_i \in \Lambda_i$, we have*

$$EU_i^h(\omega, \rho_i) = \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(O(f(f(\omega', [\![\varphi_{\rho_i}]\!]), [\![\varphi_{h_i}]\!])).$$

Proof.

$$\begin{aligned}
EU_i^{h_i}(\omega, \rho_i) &= \sum_{\omega' \in \Omega} \mathcal{PR}_i^{\rho_i}(\omega)(\omega') \cdot u_i(O(f(\omega', [\![\varphi_{h_i}]\!]))), \text{ using definition 16} \\
&= \sum_{\omega' \in \Omega} \left(\sum_{\{\omega'' : f(\omega'', [\![\varphi_{\rho_i}]\!]) = \omega'\}} \mathcal{PR}_i(\omega)(\omega'') \right) \cdot u_i(O(f(\omega', [\![\varphi_{h_i}]\!]))), \text{ using (1)} \\
&= \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i(O(f(f(\omega'', [\![\varphi_{\rho_i}]\!]), [\![\varphi_{h_i}]\!]))).
\end{aligned}$$

□

Superficially, Definition 16 looks quite different from Definition 15. However, we can show that when the history considered is $\emptyset$, corresponding to the beginning of the game, we obtain Definition 15 as a special case:

Proposition 20. *For every $i \in N$, $\omega \in \Omega$, and $\rho_i \in \Lambda_i$, we have $EU_i^{\emptyset}(\omega, \rho_i) = EU_i(\omega, \rho_i)$.*

Proof.

$$\begin{aligned}
EU_i^{\emptyset}(\omega, \rho_i) &= \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(O(f(f(\omega', [\![\varphi_{\rho_i}]\!]), [\![\varphi_{\emptyset}]\!]))), \text{ by Lemma 5} \\
&= \sum_{\omega' \in \Omega} \mathcal{PR}_i(\omega)(\omega') \cdot u_i(O(f(\omega', [\![\varphi_{\rho_i}]\!]))), \text{ since } f(\omega'', [\![\varphi_{\emptyset}]\!]) = \omega'' \text{ for all } \omega'' \in \Omega \\
&= EU_i(\omega, \rho_i).
\end{aligned}$$

□

The set of strategies that are a best response for player i at world ω given that history h has occurred is denoted $BR_i^h(\omega) = \{\rho_i \in \Lambda_i \ : \ \forall \rho_i' \in \Lambda_i, \ EU_i^h(\omega, \rho_i) \geq EU_i^h(\omega, \rho_i')\}$. From this we can define a notion of *subgame-rationality*, capturing those players who are best responding to their beliefs in every subgame where they are due to play. Formally, we add formulas of the form $SubRAT_i$ to the language (read as "player i is

subgame-rational"), and extend the valuation function so that $[\![SubRAT_i]\!] = \{\omega \in \Omega : \forall h \in H_i,\ s_i(\omega) \in BR_i^h(\omega)\}$.

Example

We now return to the "market-entry" game Γ_0 introduced in Section 2.1.1 (depicted by the game tree in Figure 2.1) as a concrete illustration of the framework we have developed. Denote by ρ_1 player 1's strategy to play O, by ρ_1' player 1's strategy to play I, by ρ_2 player 2's strategy to play F, and by ρ_2' player 2's strategy to play A. We define a strategy-rich model for Γ_0 (Figure 4.2).

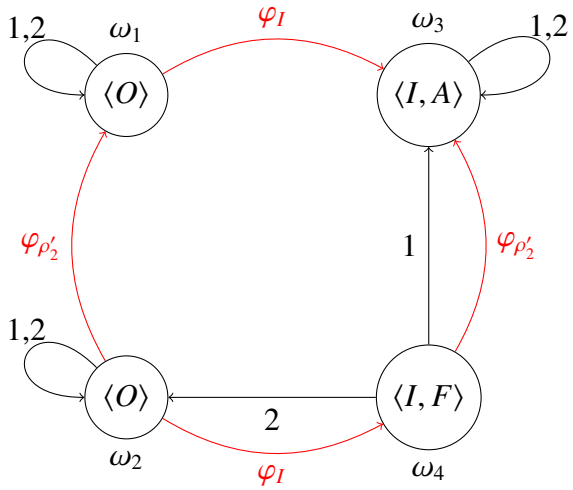

Figure 4.2: A model for Γ_0

The outcomes associated with each world are labeled in the obvious way. All beliefs in this model place 100% probability on a single world, as depicted by the black arrows. So, for example, the black arrow from world ω_4 to ω_2 labeled 2 indicates that at world ω_4, player 2 considers only world ω_2 possible and assigns it probability 1. The closest world function associated with this model is somewhat harder to illustrate compactly, but it can be viewed as arising from the following notion of "distance" on Ω (where

$d(\omega, \omega') = d(\omega', \omega)$ for all $\omega, \omega' \in \Omega$):

- $d(\omega, \omega) = 0$ for all $\omega \in \Omega$;
- $d(\omega_1, \omega_2) = d(\omega_3, \omega_4) = 1$;
- $d(\omega_1, \omega_3) = d(\omega_2, \omega_4) = 2$;
- $d(\omega_1, \omega_4) = d(\omega_2, \omega_3) = 3$.

In general, $f(\omega, T)$ returns the d-closest T-world to ω. For example, $f(\omega_1, [\![\varphi_I]\!]) = f(\omega_1, \{\omega_3, \omega_4\}) = \omega_3$, since $d(\omega_1, \omega_3) = 2 < 3 = d(\omega_1, \omega_4)$. We depict some relevant counterfactual shifts in Figure 4.2 using the red arrows. Thus we can see that in this model, player 2's strategy at ω_1 is to acquiesce, while at ω_2 it is to fight.

This model clearly satisfies requirements (R1)–(R5), and (C1) is easy to check as well. As for (C2), the only non-trivial equalities to check are whether $f(f(\omega_1, [\![\varphi_I]\!]), [\![\varphi_{(I,F)}]\!]) = f(\omega_1, [\![\varphi_{(I,F)}]\!])$ (which follows since there is a unique world associated with history (I, F)) and $f(f(\omega_2, [\![\varphi_I]\!]), [\![\varphi_{(I,A)}]\!]) = f(\omega_1, [\![\varphi_{(I,A)}]\!])$ (which likewise follows since there is a unique world associated with history (I, A)).

Finally, we observe that this is an opaque model for Γ_0. While the distance function defined above may seem artificial, it actually aligns with a certain way of counting the histories on which the strategy profiles associated with the worlds differ, with more weight given to differences that actually affect the outcome of the game. For example, we define the distance between ω_1 and ω_2 to be 1, since we only have player 2 switching between I and F, and it is on an unreachable history. But we define the distance between ω_1 and ω_3 to be 2, since player 1's switch from O to I also changes the outcome. Finally, world ω_4 is "farthest" from ω_1 since both players' strategies differ. From this we can see that the distance function is such that at the closest world where

a player changes strategies, the strategies for all the other players remain the same, and opaqueness follows.

One can check that in this model, $\omega_2 \models RAT_1 \wedge RAT_2$ (in fact, $RAT_1 \wedge RAT_2$ is common belief at ω_2!), but $\omega_2 \not\models SubRAT_2$. More generally, we can see that the formula $(\varphi_O \wedge RAT_1) \rightarrow \neg B_1 SubRAT_2$ is valid in all opaque models for Γ_0.

This result implies, in particular, that the strategy profile (O, F), which is a Nash equilibrium of the initial game, cannot be played at a world where there is common belief of subgame-rationality.

Characterizing Subgame Perfection

We first recall the definition of subgame perfect equilibrium, adapted to our notation.

Definition 17. *Let Γ be a finite game with perfect information. Then the strategy profile ρ is a subgame perfect equilibrium (SPE) if, for every player i, every $h_i \in H_i$, and every $\rho_i' \in \Lambda_i$, we have $u_i([h_i]_\rho) \geq u_i([h_i]_{\rho_i', \rho_{-i}})$.*

In other words, ρ is a SPE if, for every player i, at every history where they are due to play, ρ leads to an outcome at least as good as what could be reached by player i unilaterally changing their strategy.

For convenience we make use of the following abbreviations: $SubRAT \equiv SubRAT_1 \wedge \cdots \wedge SubRAT_n$ ("everyone is subrational"), $\varphi_\rho \equiv \varphi_{\rho_i} \wedge \cdots \wedge \varphi_{\rho_n}$ ("everyone is playing according to ρ"), and $E\varphi \equiv B_1\varphi \wedge \cdots \wedge B_n\varphi$ ("everyone believes φ"). We have the following characterization theorem, which establishes a connection between subgame-perfect equilibrium, subgame-rationality, and everyone knowing each other's strategies - similar to Proposition 2 in Section 2.3.2.

Theorem 1. *Let* Γ *be a finite extensive form game. Then the following are equivalent:*

1) ρ *is a subgame perfect equilibrium.*

2) there exists an opaque model M *for* Γ *and a state* ω *therein with* $(M, \omega) \models E\varphi_\rho \wedge SubRAT$.

Proof. ($\Leftarrow$) Consider an opaque model M for Γ such that $(M, \omega) \models E\varphi_\rho \wedge SubRAT$. Suppose for a contradiction that ρ is not a SPE. Then there exists $i \in N$ and $h_i \in H_i$ such that $u_i([h_i]_\rho) < u_i([h_i]_{\rho_i', \rho_{-i}})$ for some ρ_i'. Since M is opaque, note that for all ω' with $\mathcal{PR}_i^{\rho_i}(\omega)(\omega') > 0$, we have $j \neq i$ implies $s_j(\omega') = \rho_j$. Moreover, since $\omega \models E\varphi_\rho$, we know that

$$\mathcal{PR}_i(\omega)(\omega') > 0 \Rightarrow \omega' \models \varphi_\rho \qquad (*)$$

We have

$$\begin{aligned}
EU_i^{h_i}(\omega, \rho_i) &= \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i(O(f(f(\omega'', [\![\varphi_{\rho_i}]\!]), [\![\varphi_{h_i}]\!]))), \text{ by Lemma 5} \\
&= \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i(O(f(\omega'', [\![\varphi_{h_i}]\!]))), \text{ since } s_i(\omega') = \rho_i \text{ whenever } \mathcal{PR}_i(\omega)(\omega') > 0 \\
&= \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i([h_i]_\rho), \text{ using } (*) \text{ and Lemma 4} \\
&< \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i([h_i]_{\rho_i', \rho_{-i}}).
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
EU_i^{h_i}(\omega, \rho_i') &= \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i(O(f(f(\omega'', [\![\varphi_{\rho_i'}]\!]), [\![\varphi_{h_i}]\!]))) \text{ by Lemma 5} \\
&= \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i([h_i]_{\rho_i', \rho_{-i}})
\end{aligned}$$

Here, the last equality requires some explanation. Note that for every ω' with $\mathcal{PR}_i(\omega)(\omega') > 0$, $(*)$ together with opaqueness imply $\omega' \models \varphi_{(\rho_i', \rho_{-i})}$. This together with Lemma 4 implies that $O(f(f(\omega', [\![\varphi_{\rho_i'}]\!]), [\![\varphi_{h_i}]\!])) = [h_i]_{\rho_i', \rho_{-i}}$. Thus $EU_i^{h_i}(\omega, \rho_i) < EU_i^{h_i}(\omega, \rho_i')$, which contradicts i being subgame-rational at ω.

($\Rightarrow$) We begin with a state space consisting of one world for each strategy profile: $\Omega = \{\omega_\rho : \rho \in \Lambda\}$. For every $\omega_\rho \in \Omega$ and each $i \in N$, let $\mathcal{PR}_i(\omega_\rho) = \delta_{\omega_\rho}$, the point-mass probability measure concentrated on ω_ρ, and define $O(\omega_\rho) = [\emptyset]_\rho$.

In order to define the closest world function, we first define a notion of "distance" between worlds by setting

$$d(\omega_\rho, \omega_\rho') = |\{h : \rho_i(h) \neq \rho_i'(h) \text{ for some } i\}|.$$

Note that $d(\omega_\rho, \omega_\rho') = 0$ iff $\omega_\rho = \omega_\rho'$. Let $C_{\omega,T} \subseteq T$ be the subset consisting of those worlds in T with minimum distance to ω (as distances are always natural numbers, this set is nonempty). Now consider an enumeration $\{\omega_1, \omega_2, \ldots\}$ of all the worlds in Ω. Define the closest world function f by setting $f(\omega, T)$ to be the first element of $C_{\omega,T}$ that appears in this enumeration. Note that if $\omega \in T$, then $f(\omega, T) = \omega$ since it is the unique world at distance 0 from itself, and so the unique element of $C_{\omega,T}$.

Let $M = (\Omega, (\mathcal{PR}_i)_{i \in N}, O, f)$, and observe that this is a model for Γ satisfying (R1)–(R5). It is also clear that it is strategy- and history-rich. We wish to show that it also satisfies (C1)–(C3). Condition (C1) follows easily from the definition of the probability measures. For (C2) and (C3), the following lemma is useful.

Lemma 6. *For every $\omega_\rho \in \Omega$ and $h \in H$, we have $f(\omega_\rho, [\![\varphi_h]\!]) \models move_{P(h)}(h, \rho_{P(h)}(h))$. That is, the closest h-world to ω_ρ is a world where the player due to play at h is following his original plan of action given by ρ.*

Proof. First define, for every strategy profile ρ and history h,

$D_h^\rho = \{h' : h' \text{ is an initial segment of } h \text{ and } \rho_{P(h)}(h') \neq a, \text{ where } (h', a) \text{ is an initial segment of } h\}$.

Intuitively, $|D_h^\rho|$ tells us how far h is from the history that employing ρ would generate. Note that there is a unique ρ^h such that $d(\omega_\rho, \omega_{\rho^h}) = |D_h^\rho|$, obtained by changing ρ only at those histories that lie in D_h^ρ so as to agree with h. Note that for any other strategy profile ρ'' which generates h we have $d(\omega_\rho, \omega_\rho'') > d(\omega_\rho, \omega_{\rho^h})$, as it would have to differ from ρ on at least one additional history. It follows that $C_{\omega_\rho, [\![\varphi_h]\!]} = \{\omega_{\rho^h}\}$, so $f(\omega_\rho, [\![\varphi_h]\!]) = \omega_{\rho^h}$. Moreover, since ρ^h only differs from ρ at initial histories of h, we must have $\rho^h_{P(h)}(h) = \rho_{P(h)}(h)$, which establishes the desired result. □

From the proof of this lemma, we can also see condition (C2) holds. Let $h \in H$ and $a \in A(h)$. Let $i = P(h)$. As in the proof above, let $f(\omega_\rho, [\![\varphi_h]\!]) = \omega_{\rho^h}$. First, suppose $\rho_i(h) = a$. Then $f(f(\omega_\rho, [\![\varphi_h]\!]), [\![\varphi_{(h,a)}]\!]) = f(\omega_{\rho^h}, [\![\varphi_{(h,a)}]\!]) = \omega_{\rho^h}$ by construction. Further, since $D^\rho_{(h,a)} = D^\rho_h$, we know $\rho^h = \rho^{(h,a)}$, so $f(\omega_\rho, [\![\varphi_{(h,a)}]\!]) = \omega_{\rho^h}$. Now, suppose $\rho_i(h) = a' \neq a$. Then $D^\rho_{(h,a)} = D^\rho_h \cup \{(h, a)\}$. Then $\rho^h_i(h) = a'$, while $\rho^{(h,a)}_i(h) = a$, but ρ^h and $\rho^{(h,a)}$ coincide at every history in D^ρ_h, and coincide with ρ at every history in $H \setminus D^\rho_h$. Thus $f(\omega_{\rho^h}, [\![\varphi_{(h,a)}]\!]) = \omega_{\rho^{(h,a)}}$. Thus (C2) follows.

Observe also that for every $\rho = (\rho_1, \ldots, \rho_i, \ldots, \rho_n)$ and every ρ_i', we have $f(\omega_\rho, [\![\varphi_{\rho_i'}]\!]) = \omega_{\rho'}$, where $\rho' = (\rho_1, \ldots, \rho_i', \ldots, \rho_n)$. This is because among all $\varphi_{\rho_i'}$-worlds, $\omega_{\rho'}$ differs from ω_ρ only at the histories specified differently by ρ_i', whereas all other $\varphi_{\rho_i'}$-worlds must differ in at least one more history. From this it easily follows that M is opaque.

The preceding observations establish that $M \in \mathcal{M}^o_{EFG}$.

Now consider ω_ρ and, for ease of notation in the following calculation, denote it

by ω. We want to prove $(M, \omega) \models E\varphi_\rho \wedge SubRAT$. Note that $(M, \omega) \models E\varphi_\rho$ follows easily from the definition of the probability measures in M. Let $i \in N$. We know that $(M, \omega) \models SubRAT_i$ is equivalent to

$$EU_i^{h_i}(\omega, s_i(\omega)) \geq EU_i^{h_i}(\omega, \rho_i') \tag{4.2}$$

for all $h_i \in H_i$ and $\rho_i' \in \Lambda_i$. Recall $s_i(\omega) = \rho_i$ and $s_j(\omega) = \rho_j$ for all j.

Using Lemma 5, we get

$$\begin{aligned} EU_i^{h_i}(\omega, \rho_i) &= \sum_{\omega'' \in \Omega} \mathcal{PR}_i(\omega)(\omega'') \cdot u_i(O(f(f(\omega'', [\![\varphi_{\rho_i}]\!]), [\![\varphi_{h_i}]\!]))) \\ &= u_i(f(f(\omega, [\![\varphi_{\rho_i}]\!]), [\![\varphi_{h_i}]\!])) \text{ by definition of the probability distribution} \\ &= u_i([h_i]_\rho) \end{aligned}$$

and through a similar expansion, we have $EU_i^{h_i}(\omega, \rho_i') = u_i([h_i]_{\rho_i', \rho_{-i}})$. Then relation (4.2) is equivalent to $u_i([h_i]_\rho) \geq u_i([h_i]_{\rho_i', \rho_{-i}})$, which holds since ρ is a SPE. □

4.2 Transformations between plausibility and counterfactual models

In this section, we take a closer look at plausibility semantics for the language of conditional doxastic belief, defined in [4, 3]. As we discussed in Chapter 3, counterfactuals lend themselves easily to interpreting conditional beliefs. However, there were some axioms in the axiomatization for CDL that we argued against, like positive and negative introspection. In this section we try to identify what restrictions on counterfactual semantics would ensure a system like the one represented by plausibility models.

We start by recalling the plausibility semantics introduced in Chapter 3. We then define associated semantics using counterfactuals. We show that the new semantics defined

are equivalent - in terms of validities - and then note the restrictions we need to impose on counterfactuals in order to obtain these information-loss-free transformations. We end with a discussion on potential directions for future work.

Plausibility and counterfactual models

Consider the language $\mathcal{L}_{C\mathcal{D}\mathcal{L}}$ defined in Section 3.1.1.

First, we consider plausibility semantics for it. Consider a locally well-ordered plausibility model $M = (W, \leq, V)$. For every $\omega \in W$, recall its connected component is given by $cc(\omega) = \{\omega' \in W : \omega'(\geq \cup \leq)^+\omega\}$, where $(\geq \cup \leq)^+$ is the transitive closure of $\geq \cup \leq$. Since $\leq$ is well-founded, for every $S \subseteq W$, we know $\min S = \{x \in S \mid \forall\, y \in S : y \not< x\}$ is always nonempty ([3]).

Although we already outlined how the fomulas in the language $\mathcal{L}_{CDL}$ are interpreted recursively in M (in Section 3.1.1), we note that, when $\leq$ is total, the semantics for conditional belief simplify to:

$$[\![B^{\psi}\varphi]\!] = \{\omega \in W : min_{\leq}([\![\psi]\!]) \cap cc(\omega) \subseteq [\![\varphi]\!]\}.$$

Now, recall a *counterfactual model* $\hat{M} = (\hat{W}, \hat{V}, R, f)$ (as per Definition 14 in Section 3.2). Formulas in the language are interpreted the natural way, except we reinterpret conditional belief as

$$[\![B^{\psi}\varphi]\!] = \left\{\omega \in \hat{W} : \bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]) \subseteq [\![\varphi]\!]\right\}.$$

From Plausibility to Counterfactual Belief

There is a natural transformation of a plausibility model into a counterfactual belief model.

Suppose we have plausibility model $M = (W, \leq, V)$ where W is finite and $\leq$ is a total pre-order on W. In particular, this implies the connected component $cc(\omega) = W$ for every $\omega \in W$.

Given such a plausibility model M, we construct a counterfactual belief model $\hat{M}$. This new model has the same state space as M, and the valuation function acts the same as for the plausibility model on primitives.

The plausibility relation induces a partition on the state space, which can be interpreted as a series of concentric spheres, where more plausible worlds are on spheres of smaller radius [25]. Formally, there is a partition $\{T_0, T_1, \ldots, T_k\}$ of W (for some k) such that $T_0 = min_{\leq} W$, $T_1 = min_{\leq}(W \setminus T_0)$, ..., $T_j = min_{\leq}(W \setminus \bigcup_{0 \leq r < j} T_r)$. For every $\omega \in W$, let $R(\omega) = T_0$, so, at every world in the new model we are building, the agent considers possible only the most plausible worlds. Given this partition, define the function $rank : W \to \mathbb{N}$ where $rank(\omega)$ is the unique number such that $\omega \in T_{rank(\omega)}$.

For every $\omega \in W$ and $A \subseteq \mathcal{P}(W)^+$, let $f(\omega, A) = \begin{cases} \omega & \text{if } \omega \in A \\ min_{\leq} A & \text{otherwise} \end{cases}$. Note this function is the identity function on A and the constant function on A^c, for every $A \subseteq \mathcal{P}(W)^+$. Formally, denote this by property (CI):

(CI) For every $A \subseteq \mathcal{P}(W)^+$, $f(\omega, A) = \omega$ if $\omega \in A$ and $f(\omega, A) = C$ if $\omega \in A^c$ (where the constant $C = \min_{\leq} A$).

This counterfactual shift clearly satisfies success and strong centering. We take a closer look at the uniformity property. Note that we need only analyze the cases when ω is both an A- and a B-world, or when it is neither. If ω is an A-and B-world, by strong centering the result easily follows. If ω is neither an A- nor a B-world, suppose for a contradiction that there exists $u \in \min_{\leq} B$ such that $u \notin \min_{\leq} A$. Then $u \in A$ and $u \in T_r$ for some r such that $T_n \cap B = \emptyset$ for all $n < r$. Since $u \in A$ and $u \notin \min_{\leq} A$, there exists $m < r$ such that $T_m \cap A \neq \emptyset$. However, since $\min_{\leq} A \subseteq B$, we get $T_m \cap B \neq \emptyset$, a contradiction. Then this closest-world function satisfies all requirements for a Lewis-style selection function.

This definition also leads to an interesting connection between the binary relation encoding beliefs and the counterfactual shift function. Note we have $f(\omega, W \setminus \{\omega\}) = \begin{cases} T_0 & \text{if } \omega \notin T_0 \\ T_0 \setminus \{\omega\} & \text{otherwise} \end{cases}$, so applying the counterfactual shift function to every world in the state space allows us to recover the worlds believed possible (i.e. the most plausible ones). Formally, denote such a property by P_0:

(P_0) For all $\omega \in W$, we have $f(\omega, W \setminus \{\omega\}) = T_0$ if $\omega \notin T_0$ and $f(\omega, W \setminus \{\omega\}) = T_0$ otherwise.

There are two other important properties this counterfactual shift function satisfies. If $c : W \to \mathcal{P}(W)$ is a choice function, recall Sen's α and β:

(α) $A \subseteq B, x \in A, x \in c(B) \to x \in c(A)$

(β) $A \subseteq B, x, y \in c(A), x \in c(B) \to y \in c(A)$

For every $\omega \in W$, the function $f_\omega : \mathcal{P}(W)^+ \to \mathcal{P}(W)^+$ given by $f_\omega(A) = f(\omega, A)$ (i.e. the choice function associated to world ω) satisfies Sen's α and β.

First, we show it satisfies α.

Suppose $A, B \in \mathcal{P}(W)^+$, with $A \subseteq B$, $x \in A$ and $x \in f_\omega(B)$. Then $x \in T_r \cap B$ for some r such that $T_s \cap B = \emptyset$ for any $s < r$. For a contradiction, suppose $x \notin f_\omega(A)$, which implies $f_\omega(A) = T_s \cap A$ for some $s < r$. Since $x \in A \subseteq B$, we get $B \cap T_s \neq \emptyset$, a contradiction.

Now we show it satisfies β.

Suppose $A, B \in \mathcal{P}(W)^+$, with $A \subseteq B$, $x, y \in f_\omega(A)$ and $x \in f_\omega(B)$. Then $x \in T_r \cap B$ for some r such that $T_s \cap B = \emptyset$ for any $s < r$. In particular, this implies $rank(x) = r$. Since $x \in f_\omega(A) = T_k \cap A$ for a unique k, this implies $k = r$. Then $rank(y) = r$ and $y \in A \subseteq B$, so $y \in T_r \cap B = f_\omega(B)$.

Then this function satisfies Sen's α and β.

Denote a counterfactual belief model obtained in this manner by $M^{\leq} = (W, V, R^{\leq}, f^{\leq})$.

We note that, in these models, this notion of conditional belief gives unconditional belief the semantics of standard relational belief models. Formally, we have $R(\omega) \subseteq [\![\varphi]\!]$ iff $\omega \models B^\top \varphi$. This is easy to see, since

$$
\begin{aligned}
&\omega \models B^\top \varphi \text{ iff} \\
&\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\top]\!]) \subseteq [\![\varphi]\!] \text{ iff} \\
&\bigcup_{\omega' \in R(\omega)} \omega' \subseteq [\![\varphi]\!] \text{ iff} \\
&R(\omega) \subseteq [\![\varphi]\!].
\end{aligned}
$$

From Counterfactual Belief to Plausibility

We also consider one possible transformation from certain counterfactual belief models into plausibility models.

Consider a counterfactual belief model $\hat{M} = (\hat{W}, \hat{V}, R, f)$. Suppose $R(\omega) = R(\omega')$ for any $\omega, \omega' \in W$, and W is finite. Further, suppose the counterfactual shift function is such that it satisfies CI, P_0, and Sen's α and β (for every $\omega \in W$).

We construct a plausibility model M^f. This new model has the same state space as $\hat{M}$, and the valuation function acts the same on primitives.

Let $\omega \in W$ and $\tilde{\omega} \in R(\omega)$. Consider the sets $T_0 = R(\omega)$, $T_1 = \bigcup_{\omega' \in R(\omega)} f(\omega', W \setminus T_0) = f(\tilde{\omega}, W \setminus T_0)$ (since $f(\omega', W \setminus T_0) = f(\tilde{\omega}, W \setminus T_0)$ for all $\omega' \in R(\omega)$ by (CI)), and, in general, $T_{j+1} = \bigcup_{\omega' \in R(\omega)} f(\omega', W \setminus (\bigcup_{n=0}^{j} T_n)) = f(\tilde{\omega}, W \setminus (\bigcup_{n=0}^{j} T_n))$.

We claim these sets form a partition of W. First, since W is finite, the procedure above clearly terminates. Now suppose $T_i \cap T_j \neq \emptyset$, for some i, j with $i \neq j$. Without loss of generality, suppose $i < j$. Consider $\omega \in T_i \cap T_j$. Since $\omega \in T_i$, this implies $\omega \in \bigcup_{n=0}^{j-1} T_n$. Since $\omega \in T_j$, success for the counterfactual shift function implies $\omega \in W \setminus \bigcup_{n=0}^{j-1} T_n$, which is a contradiction. So the sets defined above form a partition for W.

Recall the function $rank : W \to \mathbb{N}$ where $rank(\omega)$ is the unique number such that $\omega \in T_{rank(\omega)}$. Define a preference relation $\leq^f$ on W so that $\omega \leq^f \omega'$ iff $rank(\omega) \leq rank(\omega')$. This is clearly a reflexive, transitive and total relation on W. Note, in particular, that we have $cc(\omega) = W$ for any $\omega \in W$.

Denote a plausibility model obtained in such a manner by $M^f = (W, V, \leq^f)$.

Transformations

We claim that the transformations outlined above actually occur without any information loss, as far as the language of conditional belief can tell.

Claim 1. *For every $\omega \in W$, for every $\varphi \in \mathcal{L}_{CDL}$, we have $(M, \omega) \models \varphi$ iff $(\hat{M}, \omega) \models \varphi$.*

Proof. We proceed by induction on the structure of φ. The base case when $\varphi = p \in$ PROP follows by construction of $\hat{M}$, and the Boolean connectives are easy to check. So suppose that the result holds for φ, ψ (namely, $[\![\varphi]\!]_M = [\![\varphi]\!]_{\hat{M}}$ and the analagous equality for ψ); we want to show it holds for $B^{\psi}\varphi$. Equivalently, we will prove $\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]) = min_{\leq}([\![\psi]\!])$.

We will prove the equality holds by considering two cases, one where $R(\omega) \cap [\![\psi]\!] = \emptyset$, and one where $R(\omega) \cap [\![\psi]\!] \neq \emptyset$.

Case 1. Suppose $R(\omega) \cap [\![\psi]\!] \neq \emptyset$. Then $min_{\leq}([\![\psi]\!]) = T_0 \cap [\![\psi]\!]$. We have $R(\omega) = \{\omega \in R(\omega) : \omega \in [\![\psi]\!]\} \cup \{\omega \in R(\omega) : \omega \in [\![\neg\psi]\!]\}$. Now note that

$$\bigcup_{\omega' \in R(\omega), \omega' \in [\![\psi]\!]} f(\omega', [\![\psi]\!]_{\hat{M}}) = \bigcup_{\omega' \in R(\omega), \omega' \in [\![\psi]\!]} \{\omega'\} = R(\omega) \cap [\![\psi]\!] = T_0 \cap [\![\psi]\!] = min_{\leq}[\![\psi]\!]_M,$$

while

$$\bigcup_{\omega' \in R(\omega), \omega' \in [\![\neg\psi]\!]} f(\omega', [\![\psi]\!]_{\hat{M}}) = \bigcup_{\omega' \in R(\omega), \omega' \in [\![\psi]\!]} min_{\leq}[\![\psi]\!]_M = min_{\leq}[\![\psi]\!]_M.$$

Then $\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]) = min_{\leq}([\![\psi]\!])$.

Case 2. Suppose $R(\omega) \cap [\![\psi]\!] = \emptyset$. Then

$$\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]_{\hat{M}}) = \bigcup_{\omega' \in R(\omega)} min_{\leq}[\![\psi]\!]_M = min_{\leq}[\![\psi]\!]_M.$$

□

In fact, we have a stronger result than just the fact that the transformation above occurs without information loss. The transformation sequence that starts with a plausibility model, constructs a counterfactual belief model, and then from this model ends with a plausibility model, is one that returns the original model the sequence started with. To prove that the plausibility ordering induced by this latter partition is precisely the one we started with, it suffices to show these two partitions are the same.

Claim 2. *Suppose $M = (W, V, \leq)$ is a plausibility model. The ordering $\leq$ induces a partition $\{T_0, \ldots, T_k\}$. Now generate $M^{\leq}$ as given above, and use this model to generate partition $\{T_0^f, \ldots, T_j^f\}$. These two partitions are the same.*

Proof. We prove by induction that $T_n = T_n^f$ for every $n \geq 0$. For the base case, note $T_0 = R(\omega) = T_0^f$ (for any $\omega \in W$). For the inductive case, suppose $T_n = T_n^f$ for every $n < k$. We want to show the result holds for k. Now, for $\omega \in T_0^f (= T_0)$, we have

$$\begin{aligned} T_k^f &= f\left(\omega, W \setminus \left(\bigcup_{0 \leq r < k} T_r^f\right)\right) \\ &= min_{\leq}\left(W \setminus \left(\bigcup_{0 \leq r < k} T_r\right)\right) \text{ by IH and definition of f} \\ &= T_k \end{aligned}$$

□

Therefore, to some extent, we can say the transformations outlined above are inverses of each other (when we start the process with a plausibility model).

From Counterfactual Belief to Plausibility

Suppose we have $\hat{M} = (W, V, R, f)$, where $R(\omega) = R(\omega')$ for any $\omega, \omega' \in W$, W is finite, and the counterfactual shift function is such that it satisfies CI, P_0, and Sen's α and β

(for every $\omega \in W$).

Instead of α and β, however, we will be using an equivalent property γ.

Claim 3. *For every $\omega \in W$, $f_\omega : \mathcal{P}(W)^+ \to \mathcal{P}(W)^+$ given by $f_\omega(A) = f(\omega, A)$ (i.e. the choice function associated to world ω) satisfies Sen's α and β iff it satisfies*

(γ) : Whenever $A \subseteq B$, $f_\omega(B) \cap A \neq \emptyset \implies f_\omega(A) = f_\omega(B) \cap A$.

Proof. ($\to$) Suppose $A \subseteq B$ and $f_\omega(B) \cap A \neq \emptyset$.

($\subseteq$) Let $u \in f_\omega(A)$. Then $u \in A$. For a contradiction, suppose $u \notin f_\omega(B)$. Since $f_\omega(B) \cap A \neq \emptyset$, there exists $v \in f_\omega(B) \cap A$. Now, since $A \subseteq B$, $v \in A$, and $v \in f_\omega(B)$, by Sen's α, we know $v \in f_\omega(A)$. Now, since $A \subseteq B$, $u, v \in f_\omega(A)$, and $v \in f_\omega(B)$, then by Sen's β we have $u \in f_\omega(B)$. Since $u \in A$, we get $u \in f_\omega(B) \cap A$.

($\supseteq$) Let $u \in f_\omega(B) \cap A$. Since $A \subseteq B$, $u \in A$, and $u \in f_\omega(B)$, then by Sen's α, we have $u \in f_\omega(A)$.

($\leftarrow$) Suppose $A \subseteq B$. Now let $x \in A$ and $x \in f_\omega(B)$. Then $f_\omega(B) \cap A \neq \emptyset$, so by γ we get $f_\omega(A) = f_\omega(B) \cap A$, so $x \in f_\omega(A)$. This proves Sen's α holds.

Now suppose $x, y \in f_\omega(A)$ and $x \in f_\omega(B)$. Then $x \in f_\omega(B) \cap A$, so by γ, we know $f_\omega(A) = f_\omega(B) \cap A$. Then $y \in f_\omega(B) \cap A$, so $y \in f_\omega(B)$, which proves Sen's β. □

From such a model $\hat{M}$, construct a plausibility model M^f as outlined in the previous section. This tranformation happens without any information loss, as the following claim shows.

Claim 4. *For every $\omega \in W$, for every $\varphi \in \mathcal{L}_{CDL}$, we have $(\hat{M}, \omega) \models \varphi$ iff $(M^f, \omega) \models \varphi$.*

Proof. We proceed by induction on the structure of φ. The base case when $\varphi = p \in$ PROP follows by construction of $\hat{M}$, and the Boolean connectives are easy to check. So suppose that the result holds for φ, ψ (namely, $[\![\varphi]\!]_{\hat{M}} = [\![\varphi]\!]_{M^f}$ and the analagous equality for ψ); we want to show it holds for $B^{\psi}\varphi$. Equivalently, we will prove $\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]) = min_{\leq}([\![\psi]\!])$.

We will prove the equality holds by considering two cases, one where $R(\omega) \cap [\![\psi]\!] = \emptyset$, and one where $R(\omega) \cap [\![\psi]\!] \neq \emptyset$.

Case 1. Suppose $R(\omega) \cap [\![\psi]\!] \neq \emptyset$. Then $min_{\leq}([\![\psi]\!]) = T_0 \cap [\![\psi]\!]$. Then the equality we want to prove becomes $\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]) = T_0 \cap [\![\psi]\!]$.

We have $R(\omega) = \{\omega \in R(\omega) : \omega \in [\![\psi]\!]\} \cup \{\omega \in R(\omega) : \omega \in [\![\neg\psi]\!]\}$. Now note that

$$\bigcup_{\omega' \in R(\omega), \omega' \in [\![\psi]\!]} f(\omega', [\![\psi]\!]_{\hat{M}}) = \bigcup_{\omega' \in R(\omega), \omega' \in [\![\psi]\!]} \{\omega'\} = R(\omega) \cap [\![\psi]\!] = min_{\leq}[\![\psi]\!]_M.$$

Now, for every $\omega' \in T_0 \cap [\![\neg\psi]\!]$, we have $[\![\psi]\!] \subseteq W \setminus \{\omega'\}$, and $f(\omega', W \setminus \{\omega'\}) = T_0 \setminus \{\omega'\}$ (by P_0), and since there exist ψ-worlds in T_0, this implies $f(\omega', W \setminus \{\omega'\}) \cap [\![\psi]\!] \neq \emptyset$. Then by γ we have $f(\omega', [\![\psi]\!]) = f(\omega', W \setminus \{\omega'\}) \cap [\![\psi]\!] = T_0 \cap [\![\psi]\!] = min_{\leq}[\![\psi]\!]$.

Then

$$\begin{aligned}\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]_{\hat{M}}) &= \bigcup_{\omega' \in R(\omega), \omega' \in [\![\psi]\!]} f(\omega', [\![\psi]\!]_{\hat{M}}) \cup \bigcup_{\omega' \in R(\omega), \omega' \in [\![\neg\psi]\!]} f(\omega', [\![\psi]\!]_{\hat{M}}) \\ &= min_{\leq}[\![\psi]\!]\end{aligned}$$

Case 2. Suppose $R(\omega) \cap [\![\psi]\!] = \emptyset$. By CI, we know $f(\omega', [\![\psi]\!]) = f(\omega'', [\![\psi]\!])$ for any $\omega', \omega'' \in T_0$.

Say $min_{\leq}([\![\psi]\!]) = T_r \cap [\![\psi]\!]$ for some $r > 0$, where, by CI, $T_r = f(\omega', (T_0 \cup \ldots \cup T_{r-1})^c)$. Then $[\![\psi]\!] \subseteq (T_0 \cup \ldots T_{r-1})^c$. Then $f(\omega', (T_0 \cup \ldots \cup T_{r-1})^c) \cap [\![\psi]\!] \neq \emptyset$ for any $\omega' \in R(\omega)$.

Then by γ, we have $f(\omega', [\![\psi]\!]) = T_r \cap [\![\psi]\!] = min_{\leq}([\![\psi]\!])$ for any $\omega' \in R(\omega)$. Thus $\bigcup_{\omega' \in R(\omega)} f(\omega', [\![\psi]\!]) = min_{\leq}([\![\psi]\!])$.

□

CHAPTER 5

CONCLUSION

This paper focuses on agents' knowledge and belief of counterfactuals by considering two applications of this epistemic concept, rationality in extensive-form games and the process of conditioning agents' beliefs.

We saw that at the center of this discussion lies the difference between an agent's beliefs on A if B were to occur and her beliefs on A if B actually did occur. In Chapter 2, we saw that these two different beliefs leads to a lack of consensus in the literature on what the appropriate epistemic characterization of subgame perfect equilibrium should be. Our work points out that when a language and a model are sophisticated enough to capture the important distinction above, the appropriate characterization arises naturally.

Chapter 3 on conditional beliefs points out another important application of agents' beliefs of counterfactuals. Using counterfactuals to interpret the process of conditioning our beliefs outlines how some of the existing models for conditional beliefs don't really align with our intuition on how conditional beliefs "work". Our approach is at the intersection of counterfactual logic, dynamic epistemic logic and conditional doxastic logic, as we employ counterfactuals to actively change the epistemic fomulas at a world.

Chapter 4 contains some experimental work related to Chapters 2 and 3. We show some advances in some of the open questions enumerated earlier, like defining rationality in extensive-form games using iterated counterfactuals, or finding appropriate restrictions on counterfactual shifts to simulate plausibility models for Conditional Doxastic Logic.

A number of important open questions remain in this topic. On the one hand, counterfactual analysis could lend itself to characterizing other equilibria, like rationalizabil-

ity, or the notion of sequential rationality for perfect Bayesian equilibria. On the other, we saw that a sound an complete axiomatization for the language of conditionalization $\mathcal{L}_C$ remains open, as does finding other conditions under which we can preserve introspection for conditional beliefs.

BIBLIOGRAPHY

[1] Carlos E. Alchourron, Peter Gardenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *J. Symbolic Logic*, 50(2):510–530, 06 1985.

[2] Robert J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8(1):6 – 19, 1995.

[3] Alexandru Baltag, Bryan Renne, and Sonja Smets. Revisable justified belief: Preliminary report. *CoRR*, abs/1503.08141, 2015.

[4] Alexandru Baltag and Sonja Smets. *A Qualitative Theory of Dynamic Interactive Belief Revision*, pages 813–858. Springer International Publishing, Cham, 2016.

[5] Alexandru Baltag, Sonja Smets, and Jonathan Zvesper. Keep 'hoping' for rationality: A solution to the backward induction paradox. *Synthese*, 169:301–333, 07 2009.

[6] Pierpaolo Battigalli and Marciano Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2):356–391, 2002.

[7] Pierpaolo Battigalli, Alfredo Di Tillio, and D. Samet. Strategies and interactive beliefs in dynamic games. 2011.

[8] Ken Binmore. Interpreting knowledge in the backward induction problem. *Episteme*, 8(3):248–261, 2011.

[9] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 2001.

[10] Oliver Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49(1):49–80, 2004.

[11] Giacomo Bonanno. A dynamic epistemic characterization of backward induction without counterfactuals. *Games and Economic Behavior*, 78:31–43, 2013.

[12] Giacomo Bonanno. *Reasoning About Strategies and Rational Play in Dynamic Games*, pages 34–62. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[13] Giacomo Bonanno. Behavior and deliberation in perfect-information games: Nash equilibrium and backward induction. *International Journal of Game Theory*, 47:1–32, 09 2018.

[14] Craig Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25(3):263–305, 1996.

[15] Brian F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

[16] Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. In RONALD FAGIN, editor, *Theoretical Aspects of Reasoning About Knowledge*, pages 5–23. Morgan Kaufmann, 1994.

[17] Alfredo Di Tillio, Joseph Y. Halpern, and Dov Samet. Conditional belief types. *Games and Economic Behavior*, 87:253–268, 2014.

[18] Hans Ditmarsch, Joseph Halpern, Wiebe Hoek, and Barteld Kooi. An introduction to logics of knowledge and belief. 1, 03 2015.

[19] Hans P. Van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147(2):229–275, 2005.

[20] E. Emerson and Joseph Y. Halpern. Decision procedures and expressiveness in the temporal logic of branching time. *J. Comput. Syst. Sci.*, 30:1–24, 1985.

[21] Benjamin Eva, Ted Shear, and Branden Fitelson. Four approaches to supposition, November 2020.

[22] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, Cambridge, MA, USA, 2003.

[23] Nir Friedman and Joseph Y. Halpern. Plausibility measures: A user's guide. *CoRR*, abs/1302.4947, 2013.

[24] James Garson. Modal Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

[25] Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170, 1988.

[26] Joseph Y. Halpern. Hypothetical knowledge and counterfactual reasoning. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, page 307–316, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[27] Joseph Y. Halpern. Substantive Rationality and Backward Induction. Game Theory and Information 0004008, University Library of Munich, Germany, November 2000.

[28] Joseph Y. Halpern. *Reasoning About Uncertainty*. MIT Press, 2003.

[29] Joseph Y. Halpern and Rafael Pass. Game theory with translucent players. *CoRR*, abs/1308.3778, 2013.

[30] David Harel, Dexter Kozen, and Jerzy Tiuryn. Dynamic logic. In *Handbook of Philosophical Logic*, pages 497–604. MIT Press, 1984.

[31] Jaakko Hintikka. Knowledge and belief: An introduction to the logic of the two notions. *Studia Logica*, 16:119–122, 1962.

[32] Hirofumi Katsuno and Alberto O. Mendelzon. On the difference between updating a knowledge base and revising it. pages 387–394. Morgan Kaufmann.

[33] Saul A. Kripke. Semantical analysis of modal logic i normal modal propositional calculi. *Mathematical Logic Quarterly*, 9(5-6):67–96, 1963.

[34] Hannes Leitgeb and Krister Segerberg. Dynamic doxastic logic: Why, how, and where to? *Synthese*, 155:167–190, 02 2007.

[35] Isaac Levi. Iteration of conditionals and the ramsey test. *Synthese*, 76(1):49–81, 1988.

[36] David Lewis. Counterfactual dependence and time's arrow. In *In Philosophical Papers, 3251*. University Press, 1986.

[37] D.K. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, Massachusetts, 1973.

[38] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, Massachusetts and London, England, 1993.

[39] Rasmus Rendsvig and John Symons. Epistemic Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019.

[40] Dov Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, 17(2):230 – 251, 1996.

[41] R.C. Stalnaker. Knowledge, belief and counterfactual reasoning in games. volume 12, pages 133–163, 1996.

[42] Robert Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36(1):31–56, July 1998.

[43] Robert Stalnaker. On logics of knowledge and belief. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 128(1):169–199, 2006.

[44] William Starr. Counterfactuals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

[45] Johan van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

[46] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. Dynamic epistemic logic. In *Internet Encyclopedia of Philosophy*. 2016.