Foundations and Philosophical Applications of Game Theory

by

Cosmo Grant

MMathPhil, University of Oxford (2013)
BA, University of Oxford (2012)

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Signature redacted

Author ....
Department of Linguistics and Philosophy
August 28, 2019

Signature redacted

Certified by.........
Vann McGee
Professor of Philosophy
Thesis Supervisor

Signature redacted

Accepted by .....
Bradford Skow
Laurence S. Rockefeller Professor of Philosophy
Chair of the Committee on Graduate Students

Foundations and Philosophical Applications of Game Theory

by

Cosmo Grant

Submitted to the Department of Linguistics and Philosophy
on August 28, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

I investigate three questions. The first belongs to game theory: When will people play a Nash equilibrium? The second, to decision theory: Why maximize expected value? The third, to the philosophy of language: How should we work out the meaning of a sentence? What unites my dissertation is a decision-theoretic approach to games, and a game-theoretic approach to meaning.

An epistemic characterization of a solution concept shows under what epistemic conditions the players behave as the solution concept describes. Chapter 1 is about epistemic characterizations of Nash equilibrium. First, I argue that theorists have slipped between two interpretations of Nash equilibrium: *strategic* and *doxastic*. As a result, they've drawn unwarranted conclusions from the characterizations. Second, following a broader discussion of the role of solution concepts, I assess doxastic equilibrium on its own merits. I argue that it doesn't deserve the attention it's received.

A key theme of Chapter 1 is the decision-theoretic approach to games: asking what you should do in a game is just a special case of asking what you should do in a decision problem. But what should you do in a decision problem? A standard answer is that you should maximize expected value, because maximizing expected value does best in the long-run. In Chapter 2, I adapt an idea well-known in economics but little-known in philosophy—maximizing expected growth rate—to argue that the long-run defense of maximizing expected value isn't sound.

In Chapter 3, I take for granted the decision-theoretic approach to games and apply it in the philosophy of language. David Lewis showed how conventions arise from repeated *coordination games*, and, as a special case, how meanings arise from repeated *signaling games*. I build on Lewis's framework. I construct coordination games in which the players can be wrong about their conventions, and signaling games in which the players can be wrong about their messages' meanings. The examples put pressure on the Elicitation Method, a typical method in semantic fieldwork, according to which we should work out the truth-conditions of a sentence by eliciting speakers' judgments about its truth-value in different situations.

Thesis Supervisor: Vann McGee
Title: Professor of Philosophy

Acknowledgments

But for many people's help, I wouldn't be here, of course. I'm grateful to all of them, but I'll single out some proximate causes.

Among the faculty at MIT, particular thanks to Alex Byrne, Vann McGee, Agustín Rayo, Roger White, and my advisor Robert Stalnaker. Thanks also to Daniel Rothschild, who supervised me during a productive semester in London. Among the students, past and present, particular thanks to David Boylan, Thomas Byrne, Kevin Dorst, Matt Mandelkern, Milo Phillips-Brown, and Kirsi Teppo.

But for these people's help, I wouldn't be here; but for my parents' help, I wouldn't be. Thank you to both of them, for providing the favorable background conditions for everything I do.

Finally, thank you to Alanna Cole-Baker. This thesis is poor recompense for our spending six years apart, but that was an impossible task.

Contents

Chapter 1

When will people play a Nash equilibrium?

1.1 Overview

In classical game theory, the main analytical tool is the *solution concept*, a function which takes a game and returns a set of *strategy profiles*, where a strategy profile is a tuple of strategies, one for each player. A famous solution concept is Nash equilibrium, which returns those strategy profiles in which no player improves in expectation by unilaterally changing her strategy. Solution concepts are often taken to describe how people will or ought to play.

In epistemic game theory, the main analytical tool is the *game model*, a representation of a particular play of the game. Game models represent not just what the players do but also their epistemic states (for example, what they believe about each other's actions, rationality and beliefs). They provide a formal framework in which to think about how assumptions about players' epistemic states (for example, that there is common belief in rationality) constrain how the players behave.

A bridge between the two approaches to game theory—classical and epistemic—is provided by *epistemic characterizations* of solution concepts, which show under what epistemic conditions the players behave as described by a given solution concept. Epistemic characterizations are important. They reveal the *scope* of a solution concept: that is, the conditions under which its predictions are accurate or its prescriptions apt. This paper is about epistemic characterizations of Nash equilibrium.

In classical game theory, it's standard to suppose that players can *randomize* over their actions using their pocket randomizing devices. A randomization for player i may be represented by a *mixed strategy*, a probability distribution over i's actions. Some solution concepts, such as Nash equilibrium, return strategy profiles involving mixed strategies. But randomizing doesn't fit easily into epistemic game theory, because, from that point of view, randomizing looks bizarre: players never have any reason to randomize, for example. How, then, to give an epistemic characterization of Nash equilibrium?

A standard response to the problem is to reinterpret mixed strategies. On the old interpretation, a mixed strategy for player i represents player i's randomization, and Nash equilibria are *strategic* equilibria. On the new interpretation, a mixed strategy for player i represents the other players' beliefs about i's action, and Nash equilibria are *doxastic* equilibria. Strategic equilibria involve randomizing; doxastic equilibria don't. So switching from the one to the other, by reinterpreting mixed strategies, avoids the problem. Epistemic characterizations of doxastic equilibrium have been given—notably, by Aumann and Brandenburger (1995).

Nash equilibrium is a fundamental concept in game theory. Theorists often claim that people will or ought to play a Nash equilibrium. Under what conditions are the predictions accurate or the prescriptions apt? An epistemic characterization can tell us. But we have to interpret the characterization correctly: if we slip between strategic and doxastic equilibrium, as people have done, we will misunderstand the scope of our predictions and prescriptions.

First, I argue that the difference between strategic and doxastic equilibrium hasn't been appreciated in the literature. As a result, people have drawn unwarranted conclusions about when people will play a Nash equilibrium.

Second, I assess doxastic equilibrium on its own merits. A solution concept, like any analytical tool, should be assessed by how well it performs its role. So I try to get clearer about the role of solution concepts, and, as a byproduct, about randomization.

In light of that, I argue that doxastic equilibrium doesn't deserve the attention it has received.

Here's the plan. In Section 2, I contrast decision theory, classical game theory, and epistemic game theory. I also introduce the game models—*type spaces*—which I use throughout. In Section 3, I describe why people have thought that randomizing doesn't fit easily into epistemic game theory. In Section 4, I clarify and contrast strategic and doxastic equilibrium. In Section 5, I describe Aumann and Brandenburger's characterization of doxastic equilibrium. In Section 6, I argue that theorists, including Aumann and Brandenburger, have drawn unwarranted conclusions from that characterization about strategic equilibrium. In Section 7, I discuss the role of solution concepts. In Sections 8 and 9, I argue that doxastic equilibrium doesn't deserve the attention it has received. Section 10 sums up.

1.2 What is epistemic game theory?

Epistemic game theory is a *decision-theoretic* and *descriptive* approach to games. What does that mean?

1.2.1 Decision theory

Consider a simple decision problem. You have an appointment across town. You can either walk or take the bus. If it stays dry, you'd prefer to walk than take the bus; if it rains, you'd prefer to take the bus than walk. What should you do? Well, it depends. It depends on the strengths of your preferences and your beliefs about the weather.

A decision theorist answers the question in three steps. First, she represents the decision problem in a regimented form, say by using a payoff matrix to represent numerically the strengths of your preferences. Second, she represents your beliefs, say by using a probability matrix. The payoff and probability matrices make a *formal* decision problem, a representation of all relevant features of your flesh-and-blood decision problem.

	dry	rain
walk	2	0
bus	1	1

payoff matrix

	dry	rain
walk	2/3	1/3
bus	2/3	1/3

probability matrix

Figure 1-1: Walk or Bus

Third, she applies a *decision rule*, say Maximize Expected Utility, which identifies what you should do. In our example, the expected utility of *walk* is $\frac{2}{3} \cdot 2 + \frac{1}{3} \cdot 0 = \frac{4}{3}$ and of *bus* is $\frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 1 = 1$. Since $\frac{4}{3} > 1$, Maximize Expected Utility says you should walk. So far, so familiar.

1.2.2 From decision theory to game theory

Now consider a simple *interactive* decision problem, or *game*. You and I are going hunting. We can either hunt stag or hare. On the one hand, each of us prefers to catch a stag than a hare; on the other hand, to catch a stag we must work together, but either can catch a hare by herself. What should you do? Well, it depends.

As before, let's first represent the game in a regimented form, using a payoff matrix, where the first components represent your preferences and the second components mine.

	Stag	Hare
Stag	2, 2	0, 1
Hare	1, 0	1, 1

Figure 1-2: Payoff matrix in the Stag Hunt.

You have analogous preferences over outcomes in the Stag Hunt as you did in Walk or Bus. However, in Walk or Bus, you were uncertain about the weather, whereas in the Stag Hunt you are uncertain about my choice. Perhaps that means we shouldn't continue as before, by supplementing the payoff matrix with a probability matrix and applying a decision rule. Why not? Two reasons:

First. Perhaps you can't assign probabilities to my actions. You ask yourself, (a) *What will he do?* Relevant to what I'll do is what I think you'll do. So before deciding (a), you need to decide, (b) *What does he think I'll do?* Relevant to that is what I think you think I'll do. So before deciding (b), you need to decide, (c) *What does he think I think he'll do?* Relevant to that...And so on. Each of us is thinking about the other, in the knowledge that the other is doing the same. Perhaps the reflexive nature of our reasoning means that you can't assign probabilities to my actions.

Second. Suppose, somehow or other, you do assign probabilities to my actions, say you believe with probability $\frac{2}{3}$ that I'll hunt stag and with probability $\frac{1}{3}$ that I'll hunt hare. Then, similarly to Walk or Bus, Maximize Expected Utility says you should hunt stag. But the probability matrix leaves out interesting features of your epistemic state. For example, it doesn't represent whether you think that I'm rational, nor what you think I think you'll do. Perhaps such features are worth making explicit so that we can see how they constrain what you do. A probability matrix may be enough to answer the question, *What will or should you do?*, but it's blind to other interesting features of the situation, features worth making explicit.

For whatever reason—see Section 7 for further discussion—classical game theorists don't follow decision theorists by supplementing the payoff matrix with a probability matrix and applying a decision rule. Instead, they proceed more directly: they identify strategy profiles and argue that people will or ought to do their bits of one of the profiles.[1] The central analytical tools of classical game theory, then, are *solution concepts*: functions which take a game and return a set of strategy profiles.

1.2.3 From game theory to epistemic game theory

Epistemic game theory is a *decision-theoretic* and *descriptive* approach to games. It's decision-theoretic because it assumes that players do assign probabilities to each other's actions. It's descriptive because it makes explicit other features of the agents' epistemic states too, such as their beliefs about the others' rationality, and investigates how these constrain what the players do.

[1] The prediction or prescription could take two forms. Is there a strategy profile such that everyone ought to do her bit of *that*? Or ought each player do her bit of some strategy profile or other, maybe different profiles for different players?

Let's say that a *play* of a game is a complete way the game can turn out. For example, a play of the Stag Hunt fixes not only what we do, but also: what each believes about the other; what each believes the other believes about her; what each believes about what the other believes about what the other believes about her, and so on; what each knows about the other; what each knows the other knows about her, and so on; our belief revision policies; our hopes, dreams, fears, and regrets; where we play; the number of hairs on your head; the distance to the third-nearest canary.[2]

Some features of a play are relevant to predicting or evaluating the players' actions. Others aren't. We only care about features which are relevant to prediction or evaluation, so we may identify plays which agree about all such features.[3]

Let S_G be the set of all plays of a game G. We can think about interesting subsets of S_G, perhaps picking out a subset using an epistemic concept (say, the set of plays in which everyone believes that everyone is rational), or using concepts from classical game theory (say, the set of plays in which no players' action is weakly dominated), or using a mixture of the two. It's particularly interesting to compare how subsets picked out using concepts from classical game theory relate to subsets picked out using epistemic concepts.

In order to investigate the structure of S_G, we need to find some way to *represent* plays of a game, or, at least, those features of plays of a game which are relevant to evaluating the players' actions. That's what a game model does.

1.2.4 Game models

There are many kinds of game models in the literature. I use the models from Aumann and Brandenburger (1995) (henceforth AB). AB's models are not the best way to represent plays: for example, they don't represent the players' belief revision policies.

[2] Games are played over time: it takes time to decide what to do, carry out that decision and enjoy the consequences. But a play is a snapshot. So it's fair to ask, *When do we click the shutter?* That's a hard question. Let's say it's the point at which the players have all decided but before their decisions have been revealed, and not worry here about the problems with that formulation.

[3] It's controversial which features are relevant, and how best to represent them. For example, it is, or was, controversial whether the players' ***belief revision policies*** are relevant, and how to represent them. See Aumann (1995), Stalnaker (1998) and, for a helpful summary, Halpern (1999).

But their kind of model is common in the literature, and the kind of model we use is irrelevant to this paper's claims.

Let $G = \langle N, \{A_i, u_i\}_{i \in N} \rangle$ be a game in strategic form, where $N = \{1, \ldots, n\}$ represents the players, A_i represents player i's actions, $A = \prod_i A_i$ represents the set of action profiles, and $u_i : \prod_i A_i \to \mathbf{R}$ represents player i's utility function. A *model* for G represents a particular play of G. A model consists of:[4]

a. for each player i, a set S_i (i's *types*);

b. a designated member $s \in S_1 \times \ldots \times S_n$ (the *actual state*);

and for each type s_i of i:

c. a probability distribution on the set S^{-i} of types of the other players (s_i's *theory*), and

d. an action a_i of i (s_i's *action*).

A *state* is a tuple of types, one for each player, or in other words a member of $S = S_1 \times \ldots \times S_n$. An *event* is a set of states. Let $[s_i]$ be the set of states where i's type is s_i. Each type's theory may be extended to a distribution $p(\cdot; s_i)$ over S, as follows: for any event E, $p(E; s_i)$ is the probability s_i's theory assigns to $\{s^{-i} \in S^{-i} : (s_i, s^{-i}) \in E\}$.[5] We say that i is *rational* at s if $a_i(s)$ maximizes i's expected payoff given his theory.[6] We say that i *believes* an event E at s if she assigns E probability 1, i.e. $p(E; s_i) = 1$.[7] An event E is mutual belief at s if everyone believes E at s, and is common belief if, at s, everyone believes it, everyone believes that everyone believes it, and so on.

[4] AB's models also specify, for each type, a function $g_i : \prod_i A_i \to \mathbf{R}$ (s_i's *payoff function*), which lets them represent the players' uncertainty about the game itself. I drop this structure because it's irrelevant to my point and clutters the exposition. But see n.12.

[5] As usual, if τ is a tuple of actions or action sets or mixed strategies or whatever, then τ_{-i} is the result of deleting the ith component, and $(s; \tau_{-i})$ is the result of replacing the ith component with s.

[6] $a_i(s)$ is i's action at state s, as determined by s_i, i's type at s.

[7] AB interpret this as *knowledge*; but it's better interpreted as *belief with probability 1*, or just *belief* for short, since it isn't factive. At any rate, I'll use the term *belief* without comment wherever AB use *knowledge*, as do e.g. Bach and Tsakas (2014).

A state is a tuple of types, one for each player. Each type is associated with an action. So each state is associated with a tuple of actions, one for each player. In particular, the actual state is associated with a tuple of actions, the *outcome.* The rest of the model represents the players' beliefs: about actions, about rationality, and about beliefs.

A type's theory is a distribution on the set of the others' types. To describe a model we have to describe each type's theory. Writing down lots of theories is tedious. Occasionally, we can save time and ink. A distribution P on S is a *common prior* just if the conditional distribution of P given $[s_i]$ is s_i's theory, for all types s_i. If a model admits a common prior P then instead of writing down each type's theory we can just write down P.[8] Not all models admit a common prior, so we can't always use this shortcut. But as it happens the only models I need in the paper do admit a common prior, so that's what I write down. To work out a type's theory from the common prior, conditionalize the prior on the type.

1.2.5 Example

Figure 3 is a model of the Stag Hunt—a representation of a way the game could go, or at least, of all those features relevant to evaluation.

	t_s	t'_s	t_h
t_S	2/9	0	1/9
t_H	1/9	1/9	0
t'_H	0	1/9	**3/9**

Figure 1-3: Model of the Stag Hunt.

You have three types: t_S, which hunts stag, and t_H, t'_H, which hunt hare. I also have three types: t_s and t'_s, which hunt stag, and t_h, which hunts hare. There are $3 \times 3 = 9$ states, one for each combination of types. The actual state is $\langle t'_H, t_h \rangle$, in bold. We both hunt hare, and get a payoff of 1.

[8] To say that a model admits a common prior P is not to say that P represents the agents' beliefs at some earlier point in time, nor that the agents conditionalized the prior on the type. A common prior, when it exists, is merely a compact way of specifying all the types' theories at once.

The numbers describe our common prior, a distribution on the 9 states, from which each type's theory is derived by conditionalizing on the type. For example, t'_H's theory, which is your actual theory, is derived by conditionalizing the prior on $[t'_H]$, or in other words, setting all numbers not in the bottom row to 0 and re-proportioning the numbers in the bottom row so that they add up to 1. Thus according to the model you believe with probability 3/4 that I am t_h, and so hunt hare, and with probability 1/4 that I am t'_s, and so hunt stag. Similarly, t_h's theory, which is my actual theory, is derived by conditionalizing the prior on $[t_h]$, or in other words, setting all numbers not in the right-hand column to 0 and re-proportioning the numbers in the right-hand column so that they add up to 1. According to the model, I believe with probability 3/4 that you're t'_H, and so hunt hare, and with probability 1/4 that you're t_S, and so hunt stag. We each maximize expected utility given our theories, so we're rational.

Do I believe that you are rational? Well, I believe that you are either t_S, which hunts stag, or t'_H, which hunts hare. The theories of t_S and t'_H are derived by conditionalization. Both types are rational, since they maximize expected utility given their theories. Hence I believe that you're rational. Do you believe that I'm rational? Well, you assign positive probability to my type t'_s. But t'_s hunts stag and believes that you hunt hare. So t'_s isn't rational. Hence you don't believe that I'm rational.

We can continue working through the model, working out our higher-order beliefs about the other's action, rationality and beliefs.

1.3 How to fit mixed strategies into epistemic game theory?

1.3.1 Mixed strategies

In the Stag Hunt, each of us has two possible actions: hunt stag or hunt hare. In classical game, it's standard to suppose that players can *randomize*: they can take out their pocket randomizing devices, choose any probability distribution over their actions, and commit to doing whatever action the device selects. The players choose not between actions, but between randomizations.

A randomization for i may be represented by a *mixed strategy*, a distribution over A_i.[9] Each player's utility function is defined on action profiles, but we can extend it to mixed strategy profiles by taking its expectation:

Expected utility. $U_i(\sigma) = \sum_{a \in A} u_i(a) \cdot \prod_j \sigma_j(a_j)$

Perhaps randomizing seems an outlandish idea.[10] In later sections, I discuss the motivations behind it.

1.3.2 The problem

In standard game models players choose actions, not randomizations, in all states. So standard game models don't represent plays in which the players randomize. We could easily modify the models so as to represent plays in which the players randomize. In type spaces, for example, we could specify, for each type s_i of i, a mixed strategy rather than a particular action. But should we?

Models represent plays of a game, or ways a game could go. Suppose the players *can't* randomize—that is, there aren't any plays in which they *do* randomize—perhaps because they don't have pocket randomizing devices, or because they don't regard randomizing as a live option. Then adding mixed strategies to the models isn't helpful: it allows us to represent impossible plays, to no good end. It might be answered that the player *can* randomize, since some *deus ex machina* might hand them pocket randomizing devices. It's implausible, but no less plausible than some plays we already admit; and besides, what's possible outstrips what's plausible.[11]

[9] A mixed strategy is a mathematical object: a probability distribution. A randomization is an interpretation of that object. The term 'mixed strategy' pushes us towards interpreting them as randomizations, which is confusing when other interpretations are in the air. Later in the paper I will use 'distribution' instead of 'mixed strategy'. Since 'mixed strategy' is standard, I'll stick with it for the moment.

[10] It does happen. For example, poker players sometimes randomize. How? By using a watch (call if the second hand is between 0 and 48 and fold otherwise), or a random number generator on their computer, or whatever.

[11] The question of whether players can randomize is a special case of a more general question: *What are an agent's options?* See Hedden (2012) for discussion.

Even if the players can randomize, why would they? The expected utility of a randomization is an average of the expected utilities of the actions in its support, so never yields higher expected utility than choosing some action. So even if players can randomize, perhaps they won't.[12] However, this argument is unconvincing. Even though there's never an incentive to randomize, that's irrelevant to whether the players can randomize. (There's never an incentive to choose a strictly dominated strategy either, but we still represent plays in which people do.) Epistemic game theory should represent every way a game can go, however strange.

In any case, the standard view is that in epistemic game theory, with its emphasis on plays, mixed strategies are out of place. The problem, then, is how to make sense of the solution concepts of classical game theory, which often involve mixed strategies, in epistemic game theory.

1.3.3 A solution?

Stalnaker (1994: 57–8) describes the standard solution:

> We should follow the suggestion of Bayesian game theorists, interpreting mixed strategy profiles as representations, not of players' choices, but of their beliefs.

Let σ_i be a distribution over player i's action set, A_i. On the *classical* interpretation, σ_i represents player i's strategy: a randomization, using his pocket randomizing device, over his actions. On the *doxastic* interpretation, σ_i represents the other players' beliefs about i's action. Stalnaker, following the suggestion of Bayesian game theorists, endorses the doxastic interpretation.

When we reinterpret mixed strategies, we thereby reinterpret solution concepts. The next section explains how.

[12] This seems to be what Stalnaker has in mind (1994: 57): "One could easily weaken this assumption [that players don't randomize], allowing players to choose blindly, withholding knowledge of their own choices from themselves by turning the choice over to a randomizing device, but while it might be harmless to permit this, players satisfying the cognitive idealizations that game theory and decision theory make could have no motive for playing a mixed strategy."

1.4 Strategic equilibrium and doxastic equilibrium

1.4.1 Uninterpreted formalism

Fix a game $G = \langle N, \{A_i, u_i\}_{i \in N} \rangle$ in strategic form. A *distribution profile* is a tuple $\sigma = \langle \sigma_1, \ldots, \sigma_n \rangle$, where σ_i is a distribution over A_i.[13] Don't think yet of a distribution profile as representing the players' randomizations or beliefs. Think of it as a mathematical object awaiting interpretation. A distribution profile σ is a *Nash equilibrium* just if:

(A) For any i and any distribution μ_i over A_i, $U_i(\mu_i; \sigma_{-i}) \leq U_i(\sigma)$.

Nash equilibrium is a property, or set, of mathematical objects, namely, distribution profiles of a given game. We may also think of Nash equilibrium as a function which takes a game and returns the set of distribution profiles which satisfy (A).

It's well-known and easily proved that (A) is equivalent to:

(B) For any i, if $\sigma_i(a) > 0$ then $a \in \text{argmax}_{a' \in A_i} U_i(a'; \prod_{j \neq i} \sigma_j)$.

So we may say, equivalently, that a distribution profile is a Nash equilibrium just if it satisfies (B).

A distribution is *pure* if it assigns probability 1 to some action and *mixed* otherwise; a distribution profile is pure if all its distributions are pure, and mixed otherwise. Not every game has a pure Nash equilibrium. But, remarkably, every finite game has at least one Nash equilibrium, either pure or mixed (Nash 1950).[14]

1.4.2 Classical interpretation

Fix a play of the game G. A play determines each player's strategy, and a strategy for player i determines a distribution over A_i. Hence the play determines a distribution profile $\langle \sigma_1, \ldots, \sigma_n \rangle$, the *strategic profile*, where σ_i represents i's strategy.

[13] Earlier I called this mathematical object a mixed strategy profile. But we'll consider two interpretations of it, one to do with strategies and the other to do with beliefs. It will help avoid confusion if the terminology doesn't favour one interpretation over the other.

[14] A game is finite if it involves finitely many players and each player has finitely many actions.

We may consider the set of plays in which the strategic profile is a Nash equilibrium. Call that *the set of plays with a strategic equilibrium*, or STRATEGIES. We can pick out that set of plays, STRATEGIES, in an equivalent but more illuminating way. The most familiar way is to interpret (A) using the classical interpretation of distributions, yielding:

> **Deviation Doesn't Pay.** The set of plays in which no player gains in expectation by unilaterally changing her strategy.

This is the classical interpretation of Nash equilibrium.

1.4.3 Doxastic interpretation

Fix a play of the game G. Let's say that i's *overall conjecture* is her belief about her opponents' actions and i's *individual conjecture about j* is her belief about j's action. So i's overall conjecture determines a distribution over A_{-i} and for each $j \neq i$, i's individual conjecture about j determines a distribution over A_j.

Suppose that the players' individual conjectures happen to agree.[15] Then the play determines a distribution profile $\langle \sigma_1, \ldots, \sigma_n \rangle$, the *doxastic profile*, where σ_i represents the individual conjecture of i's opponents about i. Clearly, a play may determine one strategic profile and another doxastic profile, and the distributions in the doxastic profile may be mixed even when the players don't randomize.

We may consider the set of plays where the players' individual conjectures agree and the doxastic profile is a Nash equilibrium. Call that *the set of plays with a doxastic equilibrium*, or CONJECTURES.

It would be nice to pick out CONJECTURES in an equivalent but more illuminating way, just as we did for STRATEGIES with Deviation Doesn't Pay. But how? The best I've managed is to interpret (B) using the doxastic interpretation of distributions, yielding:

[15] That is, for any players i, j, k, all different, i's individual conjecture about k is the same as j's individual conjecture about k.

Optimal Support. The set of plays such that the players' individual conjectures agree and for each player i, if i's opponents assign positive probability to i's action a then a is optimal given the overall conjecture canonically determined by σ_{-i}.

This seems a little more illuminating. Perhaps there's a better way to pick out CONJECTURES, one which makes it clear why we should be interested in it, just as Deviation Doesn't Pay makes clear why we should be interested in STRATEGIES. But I doubt it: see Sections 8 and 9.

1.4.4 Sets and functions, models and plays

Recall that, as defined here, Nash equilibrium is a set of *distribution profiles* of a given game. We may also think of it as a function which takes a game and returns a set of distribution profiles in that game.

By contrast, STRATEGIES and CONJECTURES are sets of *plays* of a given game. We may, harmlessly, think of STRATEGIES and CONJECTURES as sets of *models* instead, since models represent all the relevant features of plays. And we may also think of each as a *function* which takes a game and returns a set of plays, or models, of that game. How to think of them (sets or functions, of plays or models) should be clear from the context.

1.4.5 Examples

Battle of the Sexes has three Nash equilibria: $\langle U, L\rangle$, $\langle D, R\rangle$ and $\langle \frac{2}{3} \cdot U + \frac{1}{3} \cdot D, \frac{1}{3} \cdot L + \frac{2}{3} \cdot R\rangle$. I'll give two models for the game: the first is in STRATEGIES but not CONJECTURES and the second is in CONJECTURES but not STRATEGIES.

	L	R
U	2, 1	0, 0
D	0, 0	1, 2

Figure 1-4: Payoff matrix in Battle of the Sexes

In the first model, the actual state is $\langle t_U, t_L \rangle$, in bold. So Rowena does U and Colin does L. The strategic profile is a Nash equilibrium, and the model is in STRATEGIES. Since Battle of the Sexes is a two-player game, individual conjectures and overall conjectures come to the same thing. Rowena's conjecture is $\langle \frac{3}{4} \cdot L + \frac{1}{4} \cdot R \rangle$. Colin's conjecture is $\langle \frac{3}{4} \cdot U + \frac{1}{4} \cdot D \rangle$. Thus the doxastic profile isn't a Nash equilibrium, and the model isn't in CONJECTURES.

	t_L	t_R
t_U	**3/8**	1/8
t_D	1/8	3/8

Figure 1-5: First model for Battle of the Sexes.

In the second model, the actual state is $\langle t_U, t_R \rangle$, in bold. So Rowena does U and Colin does R. The strategic profile isn't a Nash equilibrium, and the model isn't in STRATEGIES. Rowena's conjecture is $\langle \frac{1}{3} \cdot L + \frac{2}{3} \cdot R \rangle$. Colin's conjecture is $\langle \frac{2}{3} \cdot U + \frac{1}{3} \cdot D \rangle$. So the doxastic profile is a Nash equilibrium, and the model is in CONJECTURES.

	t_L	t_R
t_U	1/6	**2/6**
t_D	2/6	1/6

Figure 1-6: Second model for Battle of the Sexes.

1.4.6 Taking stock

(A) and (B) are equivalent: they pick out the same distribution profiles. By interpreting (A) using the classical interpretation of distributions, we get STRATEGIES. By interpreting (B) using the doxastic interpretation of distributions, we get CONJECTURES. STRATEGIES and CONJECTURES are not equivalent: they don't pick out the same plays.

(A) and (B) are two ways of looking at the same thing. STRATEGIES and CONJECTURES are got by interpreting (A) and (B). So perhaps STRATEGIES and CONJECTURES, even though not equivalent, are also two ways of looking at the same thing, in which case an epistemic characterization of CONJECTURES will shed light on STRATEGIES. Not so. STRATEGIES and CONJECTURES are not two ways of looking at the same thing, and epistemic characterizations of CONJECTURES don't shed light on STRATEGIES. This hasn't been properly appreciated in the literature; as a result, people have drawn unwarranted conclusions from characterizations of CONJECTURES about the role of common belief in STRATEGIES.

For a given game, there are two domains in the air. One is the domain of distribution profiles; the other is the domain of plays. (A) and (B) are equivalent over distribution profiles. STRATEGIES and CONJECTURES are got by interpreting (A) and (B). STRATEGIES and CONJECTURES are not equivalent over plays. If you focus on distributions it's easy to think, wrongly, that STRATEGIES and CONJECTURES are two ways of looking at the same thing. When you focus on plays, you realize that they aren't.

1.5 Epistemic characterizations of Nash equilibrium

To give an epistemic characterization of a solution concept F on the *classical* interpretation of distributions is to prove a theorem of the form: For any game G and model M, if M satisfies such-and-such conditions, then the *strategic* distribution in M is in $F(G)$; and for any distribution profile in $F(G)$, there exists a model M' satisfying the conditions which has that strategic distribution.[16]

To give an epistemic characterization of a solution concept F on the *doxastic* interpretation of distributions is to prove a theorem of the form: For any game G and model M, if M satisfies such-and-such conditions, then the players' individual conjectures in M agree and the *doxastic* distribution is in $F(G)$; and for any distribution profile in $F(G)$, there exists a model M' satisfying the conditions which has that doxastic distribution.

Either kind of characterization might more briefly be described just as an epistemic characterization of F. But that invites confusion. If we leave implicit how

[16] This the standard way to characterize a solution concept in epistemic game theory, but there are others. See Perea (2007).

the distributions are interpreted, we may confuse the two kinds of characterization, drawing conclusions from one that would only be licensed by the other. Indeed, the next section shows that that's what people have done.

AB give an epistemic characterization of Nash equilibrium on the doxastic interpretation of distributions, or, for short, an epistemic characterization of CONJECTURES. Here are their landmark results:

Theorem 1. *For any two-person game G and model M, suppose that the rationality of the players and their overall conjectures are mutual belief in M. Then the doxastic distribution is a Nash equilibrium (that is, M is in* CONJECTURES*). And for any Nash equilibrium, there exists a model M', satisfying the conditions, in which that's the doxastic distribution.*

Theorem 2. *For any n-person game and model M, suppose the players have a common prior, that it's mutual belief that they're rational, and that their overall conjectures are common belief. Then the players' individual conjectures agree and the doxastic distribution is a Nash equilibrium (that is, M is in* CONJECTURES*). And for any Nash equilibrium, there exists a model M', satisfying the conditions, in which the players' individual conjectures agree and that's the doxastic distribution.*

AB also show that their results are tight. That is, if you delete any of the conditions, or relax them in a natural way, the resulting claims are false.

1.6 Common belief and Nash equilibrium

Remember the motivating question: *When will people play a Nash equilibrium?* Theorists often claim that people will or ought to play a Nash equilibrium. To answer the question is to find out the scope of these predictions or prescriptions. But once we distinguish between the classical and doxastic interpretation of distributions, the question resolves into two: *When will the play be in* STRATEGIES*?* and *When will the* *play be in* CONJECTURES*?* The distinction hasn't been appreciated in the literature. AB's characterization answers the second question. But people have taken it to answer the first, and so have drawn unwarranted conclusions about when the strategic profile will form a Nash equilibrium. This section describes one such unwarranted conclusion, to do with the role of common belief.

Common belief plays a minor role in AB's characterizations of CONJECTURES: in Theorem 1, it plays no role at all; in Theorem 2, common belief is assumed, but only common belief in *conjectures*, not *rationality*.[17] In subsequent generalizations of their result (Barelli 2009; Bach and Tsakas 2014), common belief plays an even smaller role.

AB and others take this to be a significant and surprising result.

> In Theorem 1 [...] common belief plays no role. This is worth noting, in view of suggestions that have been made that there is a close relation between Nash equilibrium and common belief—of the game, the players' rationality, their beliefs, and/or their choices. [...] [In Theorem 2] common belief enters the picture after all, but in an unexpected way, and only when there are at least three players. Even then, what is needed is common belief in the players' conjectures, not of the game or of the players' rationality. (1162–3)

Or take Barelli (2009: 373): "One immediate comment [about the minor role of common belief in his characterization] is that equilibrium behavior is not, after all, too demanding in terms of epistemic conditions." Or Bach and Tsakas (2014: 48–9): "AB's result challenged the widespread view that common belief in rationality was essential for Nash equilibrium. [...] We reinforce Aumann and Brandenburger's intuition about common belief in rationality *not* being essential for Nash equilibrium,

[17] Actually, the situation is more subtle. AB's game models specify, for each type, a payoff function, which lets them represent uncertainty about the game being played. Polak (1999) pointed out that when the game is common belief, AB's conditions, even dropping the common prior condition, entail common belief in rationality. I dropped the payoff functions for ease of exposition. Hence AB's conditions do entail common belief in rationality in all my models. In any case, subsequent generalizations of AB's results do *not* entail common belief in rationality, even when the game is common belief.

by showing that even mutual belief in rationality is not a crucial component." Or Stalnaker (1994: 59): "Surprisingly, one can construct models for which common belief in rationality fails, but that satisfy the epistemic conditions we have stated for Nash equilibrium."

AB and other commentators have misinterpreted the result. They showed that common belief isn't bound up with CONJECTURES; they didn't show that common belief isn't bound up with STRATEGIES. It was thought that common belief was bound up with STRATEGIES; it was not thought (why would it be?) that common belief was bound up with CONJECTURES. In short: people did think that common belief was bound up with STRATEGIES, but AB haven't shown otherwise; they did show that common belief isn't bound up with CONJECTURES, but who thought it was?

The mistake, which is common in the literature, is to assume, wrongly, that STRATEGIES and CONJECTURES are two ways of looking at the same thing, and to draw conclusions from a characterization of one that would only be licensed by a characterization of the other. I'll give two more examples. AB again:

> On the face of it, such a relation [between common belief and Nash equilibrium] sounds not implausible. One might have reasoned that each player plays his part of the equilibrium "because" the other does so; he, in turn, also does so "because" the first does so; and so on ad infinitum. This infinite connection does sound related to common knowledge; but the connection, if any, is murky. (1162–3)

AB have sketched a reason to believe that common belief is connected to STRATEGIES, not CONJECTURES. Their result shows that common belief plays a minor role in CONJECTURES, not STRATEGIES.

Or take Pacuit and Roy (2015):

> There is another important lesson to draw from AB's epistemic characterization result. The widespread idea that game theory "assumes common knowledge of rationality", perhaps in conjunction with the extensive use

> of equilibrium concepts in game-theoretic analysis, has lead to the misconception that the Nash equilibrium either requires common knowledge of rationality, or that common knowledge of rationality is sufficient for the players to play according to a Nash equilibrium. [...] The above result shows that both of these ideas are incorrect.

As is clear from their talk of "playing according to a Nash equilibrium", Pacuit and Roy have confused STRATEGIES and CONJECTURES. The lesson they draw assumes that AB's result is about STRATEGIES, not CONJECTURES; in fact, the result is about CONJECTURES, not STRATEGIES.

1.7 What role for solution concepts?

(A) and (B) are equivalent. By interpreting (A) using the classical interpretation of distributions, we get STRATEGIES. By interpreting (B) using the doxastic interpretation of distributions, we get CONJECTURES. Still, it's a fallacy to infer that since STRATEGIES has such-and-such property, so does CONJECTURES. For example, it's a fallacy to infer that since STRATEGIES is interesting, so is CONJECTURES.

Fallacious arguments can have true conclusions. We should assess CONJECTURES on its own merits. A solution concept, like any analytical tool, should be assessed by how well it performs its role. So in order to assess CONJECTURES, we should first get clearer about the role of solution concepts. That's no easy task. For despite being the main analytical tool of classical game theory, their role is obscure. In this section, I discuss some possible roles. By examining this broader issue, we also shed light on a key concept of the paper: randomization.

1.7.1 Starting point

A useful starting point is the contrast between decision theory and game theory, which I described in Section 2, using a simple decision problem, Walk or Bus, and a simple game, the Stag Hunt. The contrast between the standard ways of analyzing the two situations is striking. After all, your situation in Walk or Bus is very similar to your situation in the Stag Hunt. In both cases, you are choosing between two options (walk or bus, hunt stag or hunt hare), you are uncertain about which of two states actually obtains (dry or rain, I hunt stag or I hunt hare), and you have analogous preferences over the four outcomes. And yet the standard ways of analyzing the situations are very different.

In Walk or Bus, the payoff matrix by itself is taken to under-specify your situation. The payoff matrix settles some features of your situation. But it leaves open other features. What you should do depends on those other features: on some ways of settling them, you should walk; on other ways, you should take the bus. Until we settle the features, there is no answer to the question of what you should do. (Compare: Given that a triangle has base 5cm, what is its area? What time is it in Europe? How long is a piece of string?) Decision theorists settle the features by supplementing the payoff matrix with a probability matrix. Then they apply a decision rule, a function from the pair of matrices to actions.

In the Stag Hunt, by contrast, the payoff matrix by itself is not taken to under-specify your situation. On the standard approach, game theorists don't supplement the payoff matrix with a probability matrix and then apply a decision rule. Instead, they just apply a solution concept, a function from the payoff matrix alone to strategy profiles.

Why the contrast? In their textbook, Kevin Leyton-Brown and Yoav Shoham explain it as follows (2008: 9):

> In single-agent decision theory the key notion is that of an *optimal strategy*, that is, a strategy that maximizes the agent's expected payoff for a given environment in which the agent operates. [...] However, the situation is even more complex in a multiagent setting. In this case the environment includes—or, in many cases we discuss, consists entirely of—other agents, all of whom are also hoping to maximize their payoffs. Thus the notion of an optimal strategy for a given agent is not meaningful; the best strategy depends on the choices of others.

Leyton-Brown and Shoham point out the defining difference between decision problems and games: in a game, the environment includes other agents; in a decision problem, it doesn't. But it's not clear why it follows, as they suggest, that the notion of an optimal strategy makes sense in decision problems but not games. They say that the best strategy in a game depends on the choices of others. But equally well the best action in a decision problem depends on the state of the environment. They must think it makes a difference whether or not the environment includes other agents, but they have yet to explain why that makes a difference. The contrast between decision theory and game theory, and the role of solution concepts, still stand in need of explanation.

1.7.2 Do solution concepts pick out solutions?

Here's the first idea:

> Games have solutions. Solutions are optimal strategies. They are determined by the game itself—the payoff matrix. The role of a solution concept is to pick out the solutions of a game.

The term 'solution concept' encourages this idea, for what would a solution concept aim to pick out if not solutions, and what could solutions be if not optimal strategies? The folk often talk about 'optimal strategies' and 'optimal outcomes' of a game.[18] Game theorists sometimes do it too.[19] The idea is worth thinking about seriously.

First, I'll suggest how the idea lets us make sense of some aspects of game theory. Then I'll describe some games which, at first glance, do seem to have solutions. But

[18] For example: "von Neumann had also demonstrated mathematically that the optimum strategy in poker is simply to place your bets in proportion to the odds" (Hoyle 1994: 276); "the Nash Equilibrium is a concept of game theory where the optimal outcome of a game is one where no player has an incentive to deviate from his chosen strategy after considering an opponent's choice." (From Investopedia, a popular website about investment strategies and financial news. See investopedia.com/terms/n/nash-equilibrium.asp.)

[19] For example: "Cornerstone of Zermelo's (1913) proof that chess has optimal pure strategies, [backward induction] subsequently played a vital role in the development of perfect equilibrium." (Aumann 1995: 6)

I'll argue that, on closer inspection, they don't, at least not in a sense which can sustain the idea. So I think the idea is wrong.

A solution concept *always exists* just if, applied to any game, it returns a non-empty set of strategy profiles. When thinking about a new solution concept, game theorists typically care whether it always exists. Always existing is taken to be a virtue of a proposed solution concept. Without Nash's existence theorem, Nash equilibrium would not have become such a central concept in game theory. The idea above can make sense of this attitude. For if you think that the concept of a solution makes sense, it's natural also to think that every game has a solution. Just as you might think there are no decision-theoretic dilemmas (decision problems where no action is optimal), so too you might think there are no game-theoretic dilemmas (games where no strategies are optimal). If games always have solutions, and the role of a solution concept is to pick them out, then always existing is a virtue of a solution concept.

As noted earlier, not every game has a pure Nash equilibrium. But every finite game has at least one Nash equilibrium, either pure or mixed. Perhaps, then, the idea also reveals a motive for assuming the players can randomize. For if you think that the solutions of a game are Nash equilibria, and every game has a solution, then you might be tempted to assume that players can randomize, for else some games have no solution.

But does the concept of a solution make sense? Let's think through some examples.

First example. In Nim, there are several piles of pebbles. Two players, in turn, choose a pile and remove one or more pebbles from it. The winner is whoever removes the very last pebble. Take a particular case: there are 3 piles consisting of 14, 5 and 7 pebbles. An elegant argument shows that the player who moves first has a winning strategy: if she follows the strategy then, whatever the other player does, she'll win (Binmore 2007: 56). More generally, the argument shows which of the two players has a winning strategy, depending on the initial set-up. In that sense, there are optimal strategies in Nim.

The argument about Nim doesn't just show which player has a winning strategy.

It shows what that strategy is. Contrast, say, Hex (Binmore 2007: 57–9) or chess. In Hex, it's known that the player who moves first has a winning strategy, but it's not known what that strategy is. In chess, it's known that one or other player has a non-losing strategy, but it's not known which player, let alone what the strategy is. In any case, for a range of games, of which Nim, Hex and chess are well-known examples, it's been shown that there are winning, or non-losing, strategies for some player. In that sense, the games have optimal strategies.

Second example. My roommate reliably beats me at poker. Professional players would reliably beat both of us. A recent AI, Pluribus, reliably beats even the top professionals (Brown and Sandholm 2019). It seems harmless to rephrase this as follows: my friend is better at poker than I am, professionals are far better at poker than either of us, Pluribus is even better at poker than top professionals. Surely for one person to be better at poker than another just is for the one's strategy to be better than the other's. So some strategies are better than others. Perhaps, then, some strategy is best, or more cautiously, some strategy is such that no other strategy is better than it. That is an optimal strategy.

Third example. In the Coin Game, you and I each privately toss a fair coin and then guess how the *other's* coin landed. If either of us guesses correctly, we win; else, we lose. We're allowed to agree on a strategy ahead of time. What should we do? For example, one strategy is for both of us to guess heads, in which case the probability we win is $\frac{3}{4}$. Another strategy is to guess at random, in which case the probability we win is again $\frac{3}{4}$. You might suspect that we can't do better. But we can. For suppose you guess the same as your coin and I guess the opposite of mine. If the coins land the same way, you'll guess correctly; if they don't, I'll guess correctly. Either way, we'll win.[20] In that sense, the Coin Game has an optimal strategy. For

[20] Note that the probability you guess correctly is still $\frac{1}{2}$, as is the probability I guess correctly. We've boosted the probability that one or other of us guesses correctly, without boosting the probability that you guess correctly or that I guess correctly. Aside: Here's another surprising instance of a similar phenomenon. Suppose a referee tosses a coin three times. You and I guess how it lands each time. Your strategy is to always guess heads. My strategy is to guess heads, unless the previous toss landed heads in which case I guess tails. The winner is whoever gets the most right. You might suspect we're equally likely to win. But in fact I'll win with probability 3/8, draw with probability 3/8 and loses with probability 2/8.

a more sophisticated example along the same lines, involving the popular card game *Hanabi*, see Cox et. al. (2015).

On closer inspection, I don't think that these games have solutions—optimal strategies—at least not in a sense which can sustain the idea above. I'll take the examples in turn.

First, Nim, Hex, chess and so on. In these games, there are winning, or non-losing, strategies for some player and they are determined by the structure of the game. A perfectly reasonable aim is to discover what the strategies are—to discover, for example, how to force a win or draw in chess. Early formal work in game theory tended to focus on such games. However, such games are atypical. They make up only a negligible fraction of the vast range of situations which game theorists hope to analyze. If we take 'optimal strategy' to mean 'winning strategy' or 'non-losing strategy', then games typically don't have optimal strategies and we fail to make sense of game theorists' practice of using solution concepts. Anyway, we clearly should not assess CONJECTURES by this standard.

It's worth emphasizing, too, that a player shouldn't necessarily play a non-losing strategy. Take noughts-and-crosses. It might be that if I play the first few moves of the non-losing strategy, my opponent will force a draw; but if I play the first few moves of some other strategy, a strategy which opens me up in principle to defeat, my opponent will slip up and I'll win. (When I used to play my cousin, that was sometimes what happened.) In that case, I shouldn't play the non-losing strategy.[21]

Second, poker. For any two strategies in poker, σ and μ, there is a well-defined probability, $\Pr(\sigma, \mu)$, that someone who uses the first beats someone who uses the second. Whether the one beats the other is probabilistic, not categorical, because of poker's chance element: how the cards happen to fall. (I'm imagining poker between two players, played until one player goes bust. But the same ideas hold of more complicated versions.) Suppose Tom uses σ_1, Emma uses σ_2, and $\Pr(\sigma_1, \sigma_2) = 90\%$. So once in a while, when the cards fall her way, Emma beats Tom, but typically Tom

[21] You might complain that when analyzing a game we should assume rationality, or even common belief in rationality, in which case the scenario I described is irrelevant. See later sections and Stalnaker (1998).

beats Emma. Sick of losing her money, Emma reads a poker tutorial and switches to using σ_3, where $\Pr(\sigma_3, \sigma_1) = 70\%$. Emma now typically beats Tom. In that sense, her strategy has improved. She might hope to improve it further, increasing her chance of beating Tom, by reading another tutorial.

Is σ_3 better than σ_2? The question is ambiguous. On one reading, it asks: Is $\Pr(\sigma_3, \sigma_1) > \Pr(\sigma_2, \sigma_1)$? The answer is yes, because $70\% > 10\%$. On another reading, it asks: Is $\Pr(\sigma_3, \sigma_2) > 50\%$? That is, is someone who uses σ_3 more likely than not to beat someone who uses σ_2? The answer isn't settled by what I've said. You might think there is a third reading, not comparing σ_3 and σ_2 against σ_1 or against each other, but overall.

Compare: Suppose you're discussing three former snooker world champions, Ronnie O'Sullivan, Stephen Hendry and Steve Davis. You might ask: Is Ronnie O'Sullivan better than Stephen Hendry? The question is ambiguous. On one reading, it asks: Is O'Sullivan more likely to beat Davis than Hendry is? On another reading, it asks: Is O'Sullivan more likely than not to beat Hendry? On a third reading, it asks: Is O'Sullivan overall better than Hendry? (Whether O'Sullivan is more likely than not to beat Hendry is relevant but may not settle the third question. After all, O'Sullivan might be just as likely as not to beat Hendry, or Hendry might tend to beat people who beat O'Sullivan but lose to O'Sullivan himself.) The picture lying behind the third reading is of some yardstick of quality: measure O'Sullivan, measure Hendry, and compare the numbers.

Poker strategies aren't snooker players. The third reading may well make sense for O'Sullivan and Hendry but it doesn't make sense for σ_3 and σ_2. In poker, there is no yardstick of quality. The two-place function, Pr, makes sense: it takes two strategies and returns the probability that someone who uses the first beats someone who uses the second. A one-place function, which takes a strategy and returns a measure of its quality, does not.

Don't try to extract a one-place function, $\text{Quality}(\sigma)$, from Pr, say by averaging $\Pr(\sigma, \mu)$ over all strategies μ. No doubt formally it can be done, but it would be done to no purpose: a strategy which maximized $\text{Quality}(\sigma)$ wouldn't be an optimal strategy, in any reasonable sense of 'optimal'.

In some special cases, it does make sense to say that one strategy is overall better than another. For example, it might be that for any strategy ρ, $\Pr(\sigma, \rho) > \Pr(\mu, \rho)$. In other words, σ dominates μ. In that case, it's reasonable to say that σ is better than μ. But the case is atypical. It does not vindicate the concept of optimal strategies.

The point I'm making is straightforward. If it seems otherwise, it's because of poker's complexity: chance moves, bluffs, bankroll management and so on. Take a simple game instead, which I picked more or less at random:

	L	*R*
U	1, 2	0, 1
M	0, 2	1, 3
D	0, 3	−1, 0

Figure 1-7: Payoff matrix in the Simple Game.

Is, say, playing U better than playing M? If the column-chooser plays L, it is. If the column-chooser plays R, it isn't. But what about overall? The problem is how to make sense of that question. We can say that playing U is overall better than playing D, since U dominates D: no matter whether column-chooser plays L or R, you do better by playing U than D. But that doesn't show that it also makes sense to ask whether playing U is overall better than playing M. Poker, with all its complexity, obscures the problem. But the problem is no easier in poker than here.

Finally, the Coin Game. In the Coin Game, our interests coincide: if either correctly guesses how the other's coin landed, we win; else, we lose. That feature is unremarkable: it's perfectly kosher for players' interests to coincide. In the games underlying paradigm conventions, for example, they typically do (Lewis 1969). When the players' interests coincide, we can say, reasonably enough, that a strategy profile is optimal just if it leads to the maximum payoff. Multiple strategy profiles may be optimal on that definition. However, in typical games, the definition doesn't make sense, because what yields the maximum payoff for one player doesn't for another. Even when the definition does make sense, it's of little theoretical value. For example, saying that players in a coordination game do their bits of some or other optimal outcome may be false as a prediction and misguided as a prescription.[22]

The Coin Game illustrates a further point. Not only do our interests coincide but we're allowed to agree on a strategy ahead of time. The second feature, like the first, is unremarkable: it's perfectly kosher to suppose the players have a chance to agree on a strategy ahead of time, as long as we remember that the proper object of analysis becomes the broader situation, which includes the pre-play communication. However, in combination the two features—coincidence of interest and pre-play communication—make it tempting to treat the situation as a decision problem, not a game. The earlier analysis of the Coin Game yields to that temptation. The analysis treats us, not as two interacting agents, but as a single agent. Our uncertainty on that analysis is solely uncertainty about the environment, not about each other. Furthermore, our uncertainty is derived from objective chances: how fair coins will land. In such situations, the concept of an optimal strategy does make sense: it is nothing more than the idea of maximizing expected utility. The strategy arrived at above—you guess the same as your coin and I guess the opposite of mine—is just a strategy which maximizes expected utility in the implicit decision problem, a decision problem disguised as a game. The analysis of the Coin Game has little to tell us about games.

So much for the idea that the role of solution concepts is to pick out solutions. Except for a special class of games, such as Nim, Hex and chess, we've yet to see how to make sense of the idea.

[22] An example I heard of from Robert Stalnaker: Suppose you say to the students in your class, "Each of you write down the name of a state. If you all write down the same one, I'll give you each $2, except if you all write down 'Massachusetts', in which case I'll give you each $1; else, you get nothing." The optimal outcomes are those in which, for some state other than Massachusetts, everyone writes down the name of that state. But to predict or prescribe that the students do their bits of some or other optimal outcome would be a mistake. On natural ways of fleshing out the scenario, each student will, and should, write down 'Massachusetts'.

1.7.3 Do solution concepts describe how people will play under particular conditions?

Here's a second idea:

> The contrast between decision theory and game theory is superficial: in both decisions and games, what the agent(s) will or should do is not determined by the payoff matrix alone. However, when analyzing a game, game theorists implicitly make assumptions about the epistemic states of the players. The assumptions, in combination with the payoff matrix, constrain what the players will or should do. The role of a solution concept is to characterize the constraints: that is, to characterize how the players might behave, given the implicit epistemic assumptions.

In decision theory, implicit assumptions about the epistemic state of the agent, in combination with the payoff matrix, typically don't constrain what the agent will or should do, unless the assumptions are outrageously specific. Take Walk or Bus. In order to constrain what you should do, the assumptions would have to settle whether walking or taking the bus maximizes expected value. But any such assumptions would need to bring in the structure of Walk or Bus. The assumptions could not be *general*: the kind of thing that decision theorists could assume across the board, independent of any particular payoff matrix.

In game theory, it's more plausible that implicit assumptions about the epistemic states of the players—assumptions which don't bring in the structure of the game at hand—do constrain how the players behave. After all, players in a game have richer beliefs: beliefs not just about what the others will do, but also about their rationality and their beliefs. There is more scope for rich, general epistemic assumptions.

In epistemic game theory, we can represent players' epistemic states and check how assumptions about their states constrain what they do. On one way of understanding the idea above, epistemic game theory merely formalizes classical game theory: classical game theorists made informal implicit assumptions about the players' epistemic states and reasoned informally about how those assumptions constrain what the players do; epistemic game theory formalizes that practice. On that picture, epistemic game theory stands to classical game theory as modern calculus stands to the calculus of Newton and Leibniz.[23]

I'm inclined to think that that picture is correct. However, the picture seems not to reflect game theorists' actual attitudes, for epistemic game theory has not superseded classical game theory, as modern calculus superseded the calculus of Newton and Leibniz. That suggests that people who endorse the idea understand it differently. The implicit epistemic assumptions they have in mind are not, it seems, the sort of assumptions formalized in epistemic game theory.

What are they then? I'll briefly discuss two options. First, perhaps game theorists' implicit epistemic assumption is that the players suffer from *Knightian uncertainty* about each other (Knight 1921; see also Weatherson 2016). The idea is that the players have no information about each other, beyond perhaps common knowledge of rationality and of the payoff matrix, so their uncertainty about the other players' strategies cannot be captured in the standard Bayesian way. (Compare: Did Paul Scholes score the winner in the 1998 FA Cup Final? Is the number I just wrote on my whiteboard prime? Does neutrinoless double beta decay ever happen? You might worry that your uncertainty about these claims cannot be captured in the standard Bayesian way.) The role of a solution concept, on this picture, is to characterize how players might behave, given the assumption of Knightian uncertainty, an assumption which cannot be captured in epistemic game theory. Epistemic game theory and classical game theory are in different lines of work.

I'm not convinced that this picture makes sense. For one thing, it seems to confuse the position of the *theorist* with the position of the players she is theorizing *about*: the theorist may not assume anything about the players, but it doesn't follow that the players themselves are similarly uncertain. Besides, the picture seems unstable:

[23] Stalnaker (1994) suggests another analogy: epistemic game theory stands to classical game theory as model theory stands to axiom systems in modal logic. Model theory helps us better understand axiom systems, just as epistemic game theory helps us better understand solution concepts. On Stalnaker's analogy, epistemic game theory doesn't supersede, but complements, classical game theory. However, the analogy seems to me to be over-generous to classical game theory. For axiom systems can be fruitfully investigated by other means too and are of interest beyond just the class of models to which they correspond.

the theorist assumes that the players are radically uncertain about what the others will do, and argues on that basis to what the players might do; but if that approach is coherent, then the players seem to be in a position to make the argument for themselves, and so escape their radical uncertainty, undermining a premise of that very argument. Even if the picture did make sense, solution concepts would turn out to have very limited application. For on this picture solution concepts describe what players might do in a special kind of situation, a situation which rarely obtains in practice. After all, normally in games we are not radically uncertain about what the other players will do. Kasparov didn't cease to play chess just because he was confident and correct about what his opponent would do; I don't cease to play chess just because I'm confident and incorrect about my opponent will do. Games aren't played in a vacuum.

So much for the first option. Now for the second: perhaps game theorists' implicit epistemic assumption is that the players are rational, not in the familiar decision-theoretic sense of maximizing expected value, but in a distinctive game-theoretic sense. This seems to be what Binmore (2007: 43) has in mind:

> If Eve is rational, then she reasons optimally, and so Adam has only to figure out his opponent's optimal line of reasoning to know precisely what she will be thinking. If he has trouble in doing so, he can look the answer up in a game theory book. Psychological questions therefore have no place in a discussion of the rational play of games. If everybody played poker rationally, there wouldn't be a world poker championship because the winners and losers would be entirely determined by what cards the players were lucky enough to be dealt.

By 'rational' Binmore presumably doesn't mean 'maximizes expected value'. For if he did, then what he says would be obviously wrong: it's simply false that assuming poker players are rational (in the sense of maximizing expected value) determines their strategies. You can write down game models for poker in all of which Adam and Eve are rational (in fact, have common knowledge of rationality), and yet they do different things in different models. Binmore must have another sense of 'rational' in mind: a distinctive game-theoretic sense.

I deny that there is a distinctive game-theoretic sense of 'rational'. I've found no argument for it in the literature. (Binmore doesn't supply one.) I'm also skeptical that the distinction could be developed coherently. (Take cases where an agent's uncertainty is generated by both the environment and other agents, or where the agent is unsure whether the environment includes other agents.) I agree with Stalnaker (1998: 36): "Explaining behavior in [a game] should require, not a new theory [of rationality], but an application of the general theory to a specific situation."

1.7.4 Solution concepts and unexploitable advice

Here's a third idea:

> The role of solution concepts is to identify *unexploitable advice.*

What is unexploitable advice? I'll build up to it.

An aim of game theory is to advise players in various games about what they should do. Books about poker, for example, offer such advice. That's why poker players buy the books. (Chen and Ankenman (2006) is a popular choice.) In principle, game theorists could offer the advice in secret to one person only, so that no one else knew what the advice was or who had received it. However, in practice, that's not what happens: game theorists' advice tends to be public. That poses a problem. For if you advise a player to play a particular strategy, and the other players know that you have done so, then they may exploit the situation to their gain and your advisee's cost.

Let's look at a simple example. Suppose Tom and Emma are playing Matching Pennies. Each has a penny. They simultaneously put their pennies on the table, either heads up or tails up: if the pennies match Tom wins; else, Emma wins.

Suppose you are trying to advise Tom how to play in Matching Pennies. If you advise Tom to play heads, and Emma finds out that you have done so, then she may play tails, so that when Tom follows your advice, Emma will win. Similarly if you

	heads	tails
heads	1, 0	0, 1
tails	0, 1	1, 0

Figure 1-8: Payoff matrix in Matching Pennies

advise Tom to play tails. Tom would be foolish to blindly follow either bit of advice in Matching Pennies, when the advice is public.

Of course, Tom might be able to exploit the situation himself. For example, if you advise Tom to play heads, he might let Emma think that he was going to follow the advice, so that she would play tails, but in fact himself play tails, so that he wins. In this case, receiving the advice benefits Tom indirectly: not by following the advice, but by exploiting Emma's expectation that he'll follow it.

Tom's attempt might backfire. For Emma might see through his pretense and herself play heads, so that she wins. And so on. The situation can ramify, with Tom and Emma each trying to outguess the other about whether Tom will follow the advice, in the knowledge that the other is doing the same, just as, in the initial situation, Tom and Emma were each trying to outguess the other about how they'd place their pennies, in the knowledge that the other was doing the same.

The situation is similar if the game theorist offers advice, not just to Tom, but to both Tom and Emma. For whatever the advice is—⟨h,h⟩, ⟨h,t⟩, ⟨t,h⟩, ⟨t,t⟩—a player who blindly follows it leaves himself open to exploitation.

The moral is not that public advice has no effect or should not be followed or will never help the advisee: as we have seen, it might have an effect, in some situations it should be followed, and it might help the advisee. The moral is instead that public advice is merely a move in a broader game. Instead of allowing the player to delegate the strategic thinking to a game theorist, the game theorist's advice is just another feature of the player's situation, about which the player himself needs to think strategically.

I've focused on problems which arise when the game theorist's advice is public. But similar problems might arise even if the advice is private. For if you advise a player, privately, to play a particular strategy, and the player follows your advice every time she plays the game, then, if she plays the game often, the other players may work out her strategy and exploit it to their gain and your advisee's cost.

Back to Matching Pennies. Suppose you advise Tom, privately, to play heads. If Tom and Emma play Matching Pennies repeatedly and Tom follows your advice every time, then Emma will soon catch on. She will start playing tails and winning. And similarly if you advise Tom to play heads. Tom would be foolish to persistently follow either bit of advice, even though the advice was private.

Tom might be able to exploit the situation himself. For example, he might follow the advice to play heads just long enough that Emma starts to play tails, and then himself switch to tails. And he might continue playing tails just long enough that Emma switches back to heads, and then himself also switch back to heads. But Tom's attempt might backfire. For Emma might anticipate when Tom will switch and switch herself. And so on. The situation can ramify, just as before.

The problem is that private advice, when followed persistently, becomes public. When it has become public, the advisee is obliged to re-shoulder the burden of thinking strategically. The aim of delegating all strategic thinking to the game theorist is frustrated.

An aim of game theory is to advise players in various games about what they should do. An obstacle to realizing the aim, as we have seen, is that public advice is merely a move in a broader game and private advice becomes public when followed persistently. Either way, the game theorist fails to relieve the players of the burden of thinking strategically. Still, we might hope to realize the aim by some more subtle means. Perhaps we can come up with a special kind of advice, a kind of advice which avoids the obstacle and does relieve the players of the burden of thinking strategically. Call it *unexploitable advice.*

Back to Matching Pennies again. Matching Pennies has a unique Nash equilibrium in strategies: Tom and Emma both randomize, choosing heads with probability $\frac{1}{2}$ and tails with probability $\frac{1}{2}$, giving each expected payoff $\frac{1}{2}$. Suppose you advise Tom to do his bit of the equilibrium. If Tom follows the advice, his expected value is $\frac{1}{2}$, no matter what Emma does. It makes no difference to Tom's expected value whether the advice is public or how persistently he follows it. In that sense, Tom is relieved of the burden of thinking strategically and the advice seems to be a good candidate for unexploitable advice.

Unexploitable advice needn't be the best advice. For example, Tom might be confident that Emma will play heads. (Perhaps Emma tends to play the opposite of what she played last time. Perhaps Tom has an informant in Emma's camp. Emma might even randomize, playing heads when the second hand on her watch is in the range 0–30, but Tom knows this and stole a glance at her watch as they prepared to play.) In that case, Tom shouldn't follow the advice and randomize. He should just play heads. The idea is not that unexploitable advice is the best advice but that unexploitable advice relieves Tom of the burden of thinking strategically. (Tom needn't worry, for example, about whether Emma planted the informant, or wants him to think that she planted the informant, or wants him to think that she wants him to think that she planted the informant.)

The randomization is essential: there is no pure strategy such that, if Tom uses it, his expected value is insensitive to what Emma does. So randomization emerges naturally on this picture. But it occupies a delicate position: not as something that players have reason to do but as something that game theorists have reason to advise them to do, if the game theorist's aim is to relieve the players of the burden of thinking strategically.

Binmore (2007: 19) expresses a similar idea and points out the key role for strategic equilibrium:

> Why should anyone care about Nash equilibria? [A] game theory book can't authoritatively point to a pair of strategies (s, t) as the solution of a game unless it is a Nash equilibrium. Suppose, for example, that t weren't a best reply to s. Eve would then reason that if Adam follows the book's advice and plays s, then she would do better not to play t. But a book can't be authoritative on what is rational if rational people don't play as it predicts.

Binmore refers to the 'solution' of a game and to how 'rational' people will play it. As I've argued, these ideas are confused: games don't have solutions in Binmore's sense; there is no distinctive game-theoretic sense of 'rational'; in the decision-theoretic sense of 'rational', no more can be said about how rational players will play than that they won't play a dominated strategy. However, when we strip away those confusions, what remains is an interesting claim: that advising players to do their bit of a particular strategy profile is unexploitable only if the profile forms a strategic equilibrium.

I described the concept of unexploitable advice using Matching Pennies, a two-player zero-sum game. It seems to me that such games are the most favorable setting for making sense of the concept of unexploitable advice. If the concept makes sense for any games, it makes sense for two-player zero-sum games. When we move to multi-player games or games not of pure conflict, the concept is harder to make sense of.

Remember the purpose of the section: to get clearer about the role of solution concepts, the better to evaluate CONJECTURES. The role under consideration is that solution concepts identify unexploitable advice. The concept of unexploitable advice, although still murky, is clear enough for that purpose. For we can already see that CONJECTURES is not even a candidate for identifying unexploitable advice. Why not? Because CONJECTURES doesn't generate advice, let alone unexploitable advice. CONJECTURES is the set of plays with a doxastic equilibrium. Game theorists cannot advise players to do their bits of a doxastic equilibrium, for a doxastic equilibrium is a pattern of beliefs, not of strategies. Even if this role for solution concepts makes sense, CONJECTURES isn't fit to perform it. We should look for some other standard by which to evaluate CONJECTURES.

1.7.5 Do solution concepts yield non-self-defeating predictions?

Roger Myerson (1991: 4) describes a fourth idea:

> When we analyze a game, as game theorists or social scientists, we say

> that a player in the game is *intelligent* if he knows everything that we know about the game and he can make any inferences about the situation that we can make. In game theory, we generally assume that players are intelligent in this sense. Thus, if we develop a theory that describes the behavior of intelligent players in some game and we believe that this theory is correct, then we must assume that each player in the game will also understand this theory and its predictions.

There's something attractive about Myerson's idea. If we, the theorists, can predict the players' behavior, then surely the players themselves can do the same. If we, the theorists, don't assume that the players we're theorizing about are as good at predicting the players' behavior as we are, then we're treating ourselves as exceptions to our own theories. That's an uncomfortable situation. Game theorists' theories should apply to butchers, bakers, candlestick makers—and game theorists themselves. Game theorists are people too.

Myerson uses the idea to justify the emphasis on strategic equilibrium. Let $\sigma = \langle \sigma_1, \ldots, \sigma_n \rangle$ be a strategy profile which is not a strategic equilibrium. Suppose you, the theorist, argue that outcome will be σ. Suppose further that your argument is sound and the players are intelligent. Then each player can make that argument for himself and arrive at the same conclusion. Since, by assumption, σ is not a strategic equilibrium, some component strategy σ_i is not a best response to the others. So player i won't play σ_i. (That step assumes that player i is rational. We might flatter ourselves and take rationality to follow from intelligence or we might take it as an independent assumption. I won't worry about that here.) So the prediction is incorrect. So your argument for it is not sound. Contradiction. Hence: your argument is not sound, if the players are intelligent. Generalizing, for any non-equilibrium strategy profile σ, no argument that the outcome will be σ is sound, if the players are intelligent.

Take a game, such as Matching Pennies, with no strategic equilibrium in pure strategies. The theorist has two options on this picture: either we cannot predict what intelligent players will do in the game, or the players can randomize. The second option might look more attractive than the first: to take the first option looks like admitting defeat at the first sign of difficulty; to take the second option looks like overcoming a technical obstacle by theoretical ingenuity. Thus if you're convinced by Myerson's idea you might be led to suppose that the players can randomize.

Myerson's idea is sensible: game theorists typically shouldn't assume they are cleverer than the players they are analyzing. However, in developing the idea, Myerson combines it with another assumption: that substantive things can be said, based on the payoff matrix alone (perhaps plus the players' rationality), about what players will do. In previous subsections, I've argued that that's a mistake: nothing can be said, based on the payoff matrix alone, about what players will do; if the players are rational, all that follows is that they won't play dominated strategies. Myerson assumes, wrongly, that more can be said. That leads him to overestimate the consequences of the idea.

For example, take Myerson's conclusion above: for any non-equilibrium strategy profile σ, no argument that the outcome will be σ is sound, if the players are intelligent. Fully spelled out, the conclusion is in fact: (a) for any non-equilibrium strategy profile σ, no argument *based on the payoff matrix alone* that the outcome will be σ is sound, if the players are intelligent. That conclusion is correct. Now suppose for the sake of argument that: (b) for some strategy profile σ^*, there exists a sound argument based on the payoff matrix alone that the outcome will be σ^*. From (a) and (b) it follows that: (c) the outcome will be a strategic equilibrium, if the players are intelligent. In short, when we combine Myerson's idea with a strong assumption about what we can predict based on the payoff matrix, we're led to conclude, wrongly, that intelligent players will play a strategic equilibrium. You might go on to speculate, as people do, whether we can further narrow down how they'll play.

Now, (b) is false. I don't mean to suggest that Myerson endorses it. But he does seem to endorse a weakened version of (b), perhaps something like: (b') for some small set of strategy profiles Σ, there exists a sound argument based on the payoff matrix alone that the outcome will be in Σ. From (a) and (b'), (c) doesn't strictly follow but it's very tempting to make the jump anyway. In short, even when we weaken the assumption about what we can predict based on the payoff matrix, when we combine it with Myerson's idea, it's tempting to conclude again, wrongly, that intelligent players will play a strategic equilibrium, and to go on to speculate whether we can further narrow how they'll play.

To assume the players are intelligent is to assume that they're as clever as the theorist. That's a slippery assumption, mixing up as it does the roles of the theorist and the people she is theorizing about, but looks admirably humble: whatever we, the theorists, can predict will happen based on the payoff matrix, the players themselves can predict too. Fine, but when you also assume that you can make substantive predictions based on the payoff matrix, you're led to overestimate the consequences of this idea.

You might worry that if you treat games in the same way as decision problems, by supplementing the payoff matrix with a probability matrix or similar, then you can't assume the players are intelligent, because the theorist will know more about the situation than the players. (For example, the theorist will know what each player believes but the players typically won't know what each player believes.) That's a mistake. To restrict your attention to cases where the players are in the same position as you are is to slip from humility into egoism, for you are in effect refusing to talk about anyone other than yourself.

1.7.6 Do solution concepts pick out interesting outcomes?

Leyton-Brown and Shoham claimed that it doesn't make sense to say that each player in a game should maximize her expected payoff given her beliefs. That, I think, is a mistake, as the development of epistemic game theory shows. In any case, Leyton-Brown and Shoham also make a positive suggestion (2008: 9): "Game theorists deal with this problem by identifying certain subsets of outcomes, called *solution concepts*, that are interesting in one sense or another."

The role of solution concepts, on their view, is open-ended, for they don't try to specify what qualifies as interesting. (Interesting to whom—topologist? philosopher? poker player?) But I agree that some solution concepts pick out interesting outcomes, strategic equilibrium being a notable example. Payoff matrices in games have a richer structure than payoff matrices in decision problems. Therefore there's more scope for picking out, based on the payoff matrix alone, interesting outcomes.

Aumann (2008: 464) suggests a similar role for solution concepts:

> The relationship of solution concepts to games is similar [to the relationship of summary statistics to probability distributions.] Like the median and mean, they in some sense summarize the large amount of information present in the formal description of a game. The definitions themselves have a certain fairly clear intuitive content, though they are not predictions of what will happen. Finally, the relations between a game and [various kinds of solutions of it] is best revealed by seeing where these solution concepts lead in specific games and classes of games.

Aumann's suggestion, like Leyton-Brown and Shoham's, is open-ended. He invites us to evaluate solution concepts case-by-case, with more emphasis on practice than grand theory. I'm skeptical about whether Aumann's suggestion reflects the general attitude among game theorists. For one thing, it clashes with remarks in plenty of popular textbooks, several of which I've quoted here. For another, it doesn't account for game theorists' emphasis on strategic equilibrium or existence.

The aim of this section was to get clearer about the role of solution concepts so that we could assess doxastic equilibrium. Whether or not Leyton-Brown and Shoham's or Aumann's suggestions reflect game theorists' actual attitudes, a role for solution concepts along the lines they suggest does at least make sense. In the rest of the paper, I evaluate doxastic equilibrium by that standard.

1.8 Doxastic equilibrium and correlated conjectures

I argue that CONJECTURES is not an interesting set of plays, and doesn't deserve the attention it has received: it presupposes, objectionably, that players' beliefs about the others are independent; and doxastic equilibria, unlike strategic equilibria, are in no way self-enforcing. I also apply the doxastic interpretation of mixed strategies to another solution concept, Lewis's *coordination equilibrium*, in order to illustrate that reinterpreting mixed strategies as conjectures can turn an interesting solution concept into an arbitrary one.

1.8.1 Independence of conjectures

Recall that i's overall conjecture is represented by a distribution over A_{-i} and, for each j, i's individual conjecture about j is represented by a distribution over A_j. Note that i's individual conjectures about her opponents canonically determine *an* overall conjecture, namely, by taking the product: $\sigma = \Pi_{j \neq i} \sigma_j$. But σ needn't be i's overall conjecture, since i's beliefs about her opponents' actions may be correlated. If i's overall conjecture is the one canonically determined by her individual conjectures, then i's conjectures are *independent*; else, they are *correlated.*

1.8.2 Example

$$\underset{W}{\begin{array}{c|c|c|} & L & R \\ \cline{2-3} U & 1 & 0 \\ \cline{2-3} D & 0 & 1 \\ \cline{2-3} \end{array}} \qquad\qquad \underset{E}{\begin{array}{c|c|c|} & L & R \\ \cline{2-3} U & 0 & 1 \\ \cline{2-3} D & 1 & 0 \\ \cline{2-3} \end{array}}$$

Figure 1-9: Payoff matrix in the Three-Player Game.

$$\underset{t_W}{\begin{array}{c|c|c|} & t_L & t_R \\ \cline{2-3} t_U & 1/6 & 1/12 \\ \cline{2-3} t_D & 1/12 & 1/6 \\ \cline{2-3} \end{array}} \qquad\qquad \underset{t_E}{\begin{array}{c|c|c|} & t_L & t_R \\ \cline{2-3} t_U & \mathbf{1/12} & 1/6 \\ \cline{2-3} t_D & 1/6 & 1/12 \\ \cline{2-3} \end{array}}$$

Figure 1-10: Model for the Three-Player Game.

Figure 9 is a three-player game between Rowena, Colin and Mattea. Figure 10 is a model for it. According to the model, the actual state is $\langle t_U, t_L, t_E \rangle$, in bold. The outcome is $\langle U, L, E \rangle$, in which everyone gets payoff 0. Rowena's overall conjecture is

$\langle \frac{1}{3}LW + \frac{1}{6}RW + \frac{1}{6}LE + \frac{1}{3}RE \rangle$; and similarly for Colin and Mattea.[24]

Each player's individual conjectures assign equal weight to the opponent's actions. For example, Rowena's individual conjecture about Colin is $\langle \frac{1}{2}L + \frac{1}{2}R \rangle$ and about Mattea is $\langle \frac{1}{2}W + \frac{1}{2}E \rangle$. Hence the doxastic distribution is a Nash equilibrium, and the model is in CONJECTURES. The overall conjecture canonically determined by Rowena's individual conjectures is $\langle \frac{1}{4}LW + \frac{1}{4}LE + \frac{1}{4}RW + \frac{1}{4}RE \rangle$, which isn't Rowena's overall conjecture. So her conjectures are correlated, as are Colin's and Mattea's.

1.8.3 Optimal support and doxastic equilibrium

Is CONJECTURES an interesting set of plays? Perhaps Optimal Support shows that it is, for Optimal Support looks at first glance like an interesting property.

Not so. Take a model in CONJECTURES such that some player i's conjectures are correlated—Figure 10, for example. Since the model is in CONJECTURES, if i's opponents assign positive probability to an action a of i's, then a is optimal given the overall conjecture determined by σ_{-i}. That overall conjecture is not i's overall conjecture, since i's conjectures are correlated. So why care if i's opponents only assign positive probability to actions optimal given that overall conjecture?

Optimal Support seems to reveal why CONJECTURES is worth caring about. But when we remember that players' conjectures may be correlated, so that the overall conjecture canonically determined by σ_{-i} needn't be i's overall conjecture, we see that it doesn't.

It's easy to forget that the players' conjectures may be correlated. Consider Perea's (2007: 252) description of CONJECTURES: "[The doxastic interpretation] states that player i's belief about player j's choice should only assign positive probability to choices that are optimal for player j, given j's beliefs about the other players' choices." Perea's description is incorrect. The set he describes includes plays in which the doxastic distribution is not a Nash equilibrium. He's overlooked the fact that the players'

[24] The overall conjectures are not common belief. For example, Colin assigns positive probability to $\langle t_D, t_L, t_E \rangle$, in which case Rowena's overall conjecture is $\langle \frac{1}{6}LW + \frac{1}{3}RW + \frac{1}{3}LE + \frac{1}{6}RE \rangle$, which isn't her actual overall conjecture. Hence AB's Theorem 2 doesn't apply.

conjectures may be correlated, and so makes CONJECTURES look more interesting than it is.

1.8.4 Assuming and entailing independence

Let INDEPENDENCE be the set of plays in which the players's conjectures are independent. To switch attention from CONJECTURES to CONJECTURES ∩ INDEPENDENCE is to jump from the frying pan into the fire, for independence is an artificial restriction. Correlated conjectures are not at all exotic, even when players' choices are causally independent.[25]

AB's epistemic conditions for CONJECTURES entail that all the players' conjectures are independent. So do the subsequent weakenings of their conditions by Barelli (2009) and then Bach and Tsakas (2014).[26] Therefore we may view their results as characterizing either CONJECTURES or CONJECTURES ∩ INDEPENDENCE. When you are worried about overall conjectures, as you should be, you may view the results in the second way; when you are worried about independence, as you should be, you may view the results in the first way. But both worries must be confronted at once. Switching points of view disguises, but does not avoid, the problem: either way, the set of plays being characterized isn't interesting.

AB are well aware that independence is an artificial restriction. Consider what they say about an alternative epistemic characterization of Nash equilibrium, one which assumes independence and that players' individual conjectures agree:

> We consider this result of limited interest in the context of this paper; neither assumption has the epistemic flavor that we are looking for. *Moreover, in the current subjectivist context, we find independence dubious as*

[25] See Stalnaker (1998: 43-4) for discussion of this point. To modify one of his examples, suppose my partner and I are in our voting booths on election day. How she votes is causally independent of how I vote. You may have no idea how either of us will vote, but still be confident (and justifiably so) that, however we vote, we'll vote the same way.

[26] Bach and Tsakas (2014: 49) pay lip service to the idea that the players' conjectures may be correlated, saying "Note that [i's overall conjecture] is not necessarily a product measure." But their epistemic conditions, like AB's, do entail that the conjectures are independent. They show this in Appendix A, but don't mention it in the body of the paper.

an assumption (though not necessarily as a conclusion). (1177, my italics)

I see little space between assuming independence and assuming things which entail independence. (Would AB attach the same significance to their result if it assumed, not common belief in rationality, but things which entailed common belief in rationality? Surely not.) Since assuming independence is dubious, so is assuming things which entail independence.

Perhaps AB are distinguishing not between assuming independence and assuming things which entail independence, but between characterizing CONJECTURES and characterizing CONJECTURES ∩ INDEPENDENCE. They are advising us to take their result as characterizing CONJECTURES. Fine, but the criticism stands: however we view the result, the set of plays being characterized isn't interesting.

1.9 Self-enforcing behaviour and coordination equilibrium

1.9.1 Self-enforcing behaviour

In what sense, if any, are strategic equilibria and doxastic equilibria self-enforcing? To fix ideas, recall Figures 5 and 6, two models for Battle of the Sexes. Figure 5 is in STRATEGIES; Figure 6 is in CONJECTURES.

First consider Figure 5. Rowena plays U; Colin plays L. Suppose either player learns what the other plans to do, or what both plan to do is publicly announced. In either case, it won't change what they plan to do. The strategic profile will remain a Nash equilibrium. Or suppose Rowena and Colin make a non-binding pre-play agreement to do their bits of $\langle U, L\rangle$. The agreement is self-enforcing: to agree to $\langle U, L\rangle$ is a reason to do $\langle U, L\rangle$. Or suppose a third-party recommends that they play $\langle U, L\rangle$. The recommendation is not self-defeating. In some sense, then, the strategic equilibrium is self-enforcing.

Now consider Figure 6. Rowena's conjecture is $\langle \frac{1}{3} \cdot L + \frac{2}{3} \cdot R\rangle$; Colin's conjecture is $\langle \frac{2}{3} \cdot U + \frac{1}{3} \cdot D\rangle$. If Rowena learns what Colin's conjecture is, she will learn that the state is $\langle t_U, t_R \rangle$, and that Colin will play R. The doxastic profile will become $\langle 1 \cdot R, \frac{2}{3} \cdot U + \frac{1}{3} \cdot D \rangle$, which is not a Nash equilibrium. And similarly if Colin learns what Rowena's conjecture is. If one player learns the other's conjecture, the doxastic equilibrium is destroyed. Or suppose their conjectures are publicly announced. There's no reason to expect that the resulting doxastic profile will be a Nash equilibrium. Finally, note that talk of *agreeing* to a doxastic profile, or *recommending* a doxastic profile, is misguided. You can't form beliefs by agreement, nor sensibly recommend what the players should believe. In no sense, then, is the doxastic equilibrium self-enforcing.

Strictly speaking, these claims go beyond the models. The models represent what the players do and believe. They don't represent what the players would do or believe if they learnt the other's strategy or conjecture, nor how they would respond to recommendations or agreements. However, on natural ways of fleshing out the models, the claims hold.

Figure 5 illustrates a general phenomenon: there is a connection between STRATEGIES and self-enforcing behaviour. The connection may be subtle.[27] But theory and practice both show that there is some connection. Figure 6 also illustrates a general phenomenon: we have no reason to expect any connection between CONJECTURES and self-enforcing behaviour. Neither theory nor practice suggests there is one. Strategic equilibria are in some sense self-enforcing; doxastic equilibria are not.

1.9.2 Applying the doxastic interpretation to other solution concepts

Many solution concepts involve mixed strategies, not just Nash equilibrium. It's instructive to apply the doxastic interpretation of mixed strategies to them too, and see what we get.

A case study: Lewis's concept of *coordination equilibrium.* A distribution profile is a coordination equilibrium just if:

(C) For any i and any distribution μ_i over A_i, $U_j(\mu_i; \sigma_{-i}) \leq U_j(\sigma)$, for all j.

[27] Aumann (1990) argued that non-binding pre-play agreements to play Nash equilibria aren't always self-enforcing.

It's easily proved that (C) is equivalent to:

(D) For any i, if $\sigma_i(a) > 0$ then $a \in \text{argmax}_{a' \in A_i} U_j(a'; \prod_{j \neq i} \sigma_j)$, for all j.

Consider the set of plays in which the strategic profile is a coordination equilibrium. Call that *the set of all plays with a strategic coordination equilibrium*, or CE-STRATEGIES. We can pick out that set of plays in an equivalent but more illuminating way, by interpreting (C) using the classical interpretation of distributions, yielding:

> **No Resentment.** The set of plays in which no player gains in expectation if *any* player unilaterally changes her strategy.

This yields David Lewis's way of thinking about coordination equilibrium, which plays a large role in his theory of convention and meaning. In STRATEGIES, no one regrets her choice, given the others' choices; in CE-STRATEGIES, additionally, no one resents another player's choice, given the choices of that player's opponents.

Now consider instead the set of all plays in which the players' individual conjectures agree and the doxastic profile is a coordination equilibrium. Call that *the set of plays with a doxastic coordination equilibrium*, or CE-CONJECTURES. It would be nice to pick out that set in an equivalent but more illuminating way. How? Let's try interpreting (D) using the doxastic interpretation of distributions, yielding:

> **Universally Optimal Support.** The set of plays such that, for each player i, if i's opponents assign positive probability to i's action a, then for each player j, a maximizes j's expected utility given the overall conjecture canonically determined by σ_{-i}.

CE-CONJECTURES is a strange set of plays. It has no intuitive interest. CE-STRATEGIES, by contrast, is a natural set of plays, of considerable intuitive interest. Reinterpreting mixed strategies as conjectures turns an interesting solution concept into an arbitrary one. As with CE-STRATEGIES and CE-CONJECTURES, so with STRATEGIES and CONJECTURES.

1.10 Conclusion

Reinterpreting a solution concept is not merely a change in point of view. It's the plays that matter, not the distribution profiles. A *play concept* is a function which takes a game and returns a set of models of the game. A solution concept F resolves into multiple play concepts, as many as there are ways to interpret distributions. For example, Nash equilibrium resolves into STRATEGIES and CONJECTURES, via the classical and doxastic interpretations. Better, I suggest, not to think of *many* interpretations of *one* solution concept, but simply of *many* play concepts. Working with play concepts, rather than different interpretations of solution concepts, reminds you that a change in interpretation is a substantive change, not merely a change in point of view.

Doxastic equilibrium has been taken to be an approximation of strategic equilibrium in epistemic game theory. That's a mistake: it's a quite different concept, and not an interesting one. Besides, strategic equilibrium doesn't need to be approximated. Either players can randomize or they can't. If they can, then game models should represent randomizations. If they can't, then better to forget about mixed strategies than reinterpret them. Why cling to an old formalism when forced to give up the interpretation for which the formalism was invented?

Epistemic characterizations of solutions concepts are a bridge between classical game theory and epistemic game theory. They import the analytical tools of classical game theory into the epistemic framework. To think about the classical theory in the epistemic framework is helpful. That's not what we do when we reinterpret mixed strategies as conjectures.

Chapter 2

Against long-run defenses of decision rules

2.1 The Gamble

I offer Mac a gamble on a coin toss. Here's how it works. Mac bets some amount between \$0 and \$100. Then I flip a fair coin. If it comes up heads she gets back what she bet plus twice the same again. If it comes up tails I keep what she bet. That is The Gamble. For example, suppose she bets \$50. If the coin comes up heads she gets back her \$50, plus an extra \$100. If it comes up tails I keep her \$50. Mac's utility function is linear in dollars.[1] How much should she bet?

The *expected value of betting m dollars* is $0.5 \cdot 2m - 0.5 \cdot m$, which is $0.5 \cdot m$. The amount between \$0 and \$100 which maximizes this quantity is \$100. So to maximize expected value in The Gamble is to bet \$100.

But should Mac maximize expected value in The Gamble? Some say she should, because maximizing expected value does best in the long-run.[2] I argue that the

[1] I assume a linear utility function for convenience but nothing hangs on it.

[2] In Ray Briggs's survey in the Stanford Encyclopedia of Philosophy (2014), they discuss only two defenses of maximizing expected value: the long-run defense and a defense based on the representation theorems. That suggests that the long-run defense is generally taken seriously. Looking through popular textbooks (e.g. Hacking (2001: 81–2) or Bertsekas and Tsitsiklis (2008: 90)) provides further evidence. Explicit appeals to the long-run defense are less common in recent research articles than in surveys or textbooks. But they do happen. For example, after a careful and illuminating discussion, Kenny Easwaran (2008: 633) concludes that "it might be reasonable to believe that the

long-run defense isn't sound.

In Section 2, I describe the long-run defense. In Section 3, I suggest why it's worth taking seriously, so is worth refuting. In Section 4, which is the heart of the paper, I adapt an idea well-known in economics but little-known in philosophy—maximizing expected growth rate—to show that a rival bet also has a long-run defense. The long-run defenses are parallel but come to incompatible conclusions, so neither is sound—or so I argue. In Section 5, I show how to formalize a new conjecture, a conjecture with an interesting philosophical upshot: that many bets have a long-run defense, so long-run defenses are cheap. In Section 6, I describe some work towards resolving the conjecture.

2.2 The long-run defense of betting \$100

Betting \$100 in The Gamble does best in the long-run. Let's unpack what that means.

In the *horizontal iteration* of The Gamble, you face The Gamble repeatedly: on Day 1, Day 2, Day 3, and so on forever. You have to bet the same amount between \$0 and \$100 every day. Suppose you bet m dollars. By the Weak Law of Large Numbers, the probability that [the number of wins by Day n divided by n is within ϵ of 0.5] tends to 1 as n tends to infinity, for any $\epsilon > 0$. Therefore the probability that [your average winnings by Day n are within ϵ of the expected value] tends to 1 as n tends to infinity, for any $\epsilon > 0$. Betting \$100 maximizes expected value. Therefore—and this is the key consequence—in the horizontal iteration of The Gamble, someone who bets \$100 eventually tends to make more money than someone who doesn't.

To state the key consequence more formally. Suppose in the horizontal iteration you bet \$100 and I bet some other amount. Then for any $\delta > 0$, there exists an N such that for all $n > N$ the probability that you've made more than me by Day n is at least $1 - \delta$.[3]

value of an individual game is constrained by the long-run payout of repeated plays of the game."

[3] If we apply the Strong Law instead of the Weak Law, we can show something stronger: that with probability 1, eventually someone who bets \$100 makes more money than someone who doesn't. However, to apply the Strong Law we have to imagine a completed infinity of bets, because the Strong Law requires assigning probabilities to infinite sequences. To apply the Weak Law, we need

So much for the horizontal iteration of The Gamble. How does it bear on The Gamble itself? Distinguish three claims:

(1) In the horizontal iteration of The Gamble, someone who bets \$100 eventually tends to make more money than someone who doesn't.

(2) Mac should bet \$100 in the horizontal iteration of The Gamble.

(3) Mac should bet \$100 in The Gamble.

(1) is true. If (1) is true, so is (2). If (2) is true, so is (3). Therefore (3) is true. That's the long-run defense of betting \$100 in The Gamble.

You might argue that the long-run defense isn't sound, on the basis that hypothetical repetitions are irrelevant to a decision problem. After all, no decision will actually be repeated indefinitely, most won't even be repeated often, some can't be repeated often, and even those which will be repeated aren't independent. But I'll argue on a quite different basis that the long-run defense isn't sound.

2.3 Motivating the long-run defense

In the next section, I argue that the long-run defense isn't sound. In this section, I suggest why it's worth taking seriously, so is worth refuting, and draw connections to broader issues about long-run defenses in other fields, dominance, and dynamic choice.

2.3.1 Long-run defenses in other fields

Perhaps the long-run defense of betting \$100 looks silly. But similar long-run defenses are used in a wide range of other fields too. In Bayesian confirmation theory, the fact that differences in priors are washed out as the number of data points tends to infinity

only imagine a potential infinity of bets (for each n, a sequence of n bets) because the Weak Law only requires assigning probabilities to finite sequences. In short, applying the Strong Law yields a stronger result, but also asks more of our imagination.

is taken to ease worries about subjectivity now.[4] In Neyman-Pearson hypothesis testing, how a rule of statistical inference performs as the number of applications goes to infinity is taken to bear on whether we should apply it in a particular instance.[5] In complexity theory, how an algorithm performs as the input size tends to infinity is taken to bear on how it performs given an input of size, say, 25.[6] In game theory, what a player should do when the number of rounds tends to infinity is taken to bear on what he should do when the game is just played once.[7] And so on.

Are these long-run defenses all on a par with the long-run defense of betting $100 in The Gamble? If they are, then a great deal hangs on properly evaluating any of them. If they aren't, then the status of a long-run defense must depend on details of the particular case. Either way, the long-run defense of betting $100 is worth taking seriously.

To fix ideas, I'll spell out in more detail a long-run defense from another field: estimation theory in frequentist statistics.

How can we estimate the bias of a coin, $\theta \in [0,1]$? Here's what the frequentist statistician says. Flip the coin, so that the outcome X is either 1 (heads) with probability θ or 0 (tails) with probability $1-\theta$. If you flip the coin n times then the outcome is a sequence $X_1, X_2, \ldots, X_n$. An *estimator* is a function of the outcome. For example, one estimator is $\hat{\theta}_n := \frac{X_1+\ldots+X_n}{n}$, the proportion of heads in n flips. The *estimate* is the value yielded by the estimator after actually flipping the coin. For example, take the estimator just defined and suppose you toss the coin five times. If

[4] Take Strevens (2017: 84): "The most common and in many ways the most effective Bayesian response to the subjectivity objection is the convergence response: although in the short term scientists may disagree on the significance of the evidence, in the longer term their subjective probabilities will converge on the same hypotheses, and so a consensus will emerge."

[5] Take Neyman and Pearson themselves (1933: 291): "Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong."

[6] Take Aaronson (2013: 265): "The polynomial-exponential distinction is open to obvious objections: an algorithm that took 1.00000001^n steps would be much faster in practice than an algorithm that took n^{10000} steps! [...] But empirically, polynomial time *turned out* to correspond to "efficient in practice," and exponential time to "inefficient in practice," so often that complexity theorists became comfortable making the identification."

[7] Osborne and Rubinstein (1994: 38–9) discuss this idea, among others, when considering how to interpret and justify the concept of Nash equilibrium.

you happen to get, say, HTTHT then the estimate of the bias is $\frac{1+0+0+1+0}{5} = \frac{2}{5}$.

A key question for the frequentist is: Which estimator should you use? Say that an estimator is *consistent* if it eventually tends to give estimates close to the bias, no matter what the bias is.[8] As it happens, the estimator above is consistent.

Frequentist statisticians say, among other things, that you should use a consistent estimator.[9] Why?

To bring out the issue more clearly, I'll first focus on a more specific question. Say that an estimator is *anti-consistent* if it tends in probability to something other than θ, for all θ. Why prefer consistent to anti-consistent estimators? The idea seems to be that consistent estimators do better in the long-run. Let's unpack what that means.

To estimate the bias of the coin, you'll flip it finitely many times—100, say—and then make an estimate. That is an estimation problem. In an *iterated* estimation problem, you flip the coin repeatedly: on Day 1, Day 2, Day 3, and so on forever. You estimate the bias each day based on the flips so far.

Now distinguish three claims:

(A) In an iterated estimation problem, the estimate yielded by a consistent estimator eventually tends to be closer to the bias than the estimate yielded by an anti-consistent estimator, no matter what the bias is.

(B) A consistent estimator is preferable to an anti-consistent estimator in an iterated estimation problem.

(C) A consistent estimator is preferable to an anti-consistent estimator in an estimation problem.

(A) is true. There's no denying that: it follows from the definitions of 'consistent' and 'anti-consistent'. If (A) is true, so is (B). If (B) is true, so is (C). Therefore

[8] That is, an estimator is consistent if it tends in probability to θ, for all θ.

[9] Take Fisher (1950: 11): inconsistent estimators "should be regarded as outside the pale of decent usage". Or Neyman (1952: 188): "it is definitely not profitable to use an inconsistent [estimator]."

(C) is true. That's the long-run defense of preferring consistent to anti-consistent estimators.

The original question was: Why prefer consistent over *non*-consistent estimators? I focused on a more specific question: Why prefer consistent over *anti*-consistent estimators? Focusing on the more specific question brings out the issue more clearly. But we can answer the original question in a similar way. The only difference will be in (A). When we expand the class of rivals to consistent estimators, from anti-consistent to inconsistent, we have to weaken (A). But the argument still takes a similar form.[10]

The long-run defense of preferring consistent estimators in an estimation problem has the same form as the long-run defense of betting $100 in The Gamble. If you reject one, it seems you should reject the other too.

You might well reject frequentist estimation theory for other reasons. In that case, you won't mind rejecting the long-run defense of preferring consistent estimators, since you'll think that the whole frequentist approach is wrong-headed. But similar long-run defenses come up in other fields as well. To reject long-run defenses across the board is a radical move. That's a reason to take seriously the long-run defense of betting $100.

2.3.2 Dominance

Consider the following claim:

> If with high probability option X yields higher utility than option Y then X is preferable to Y.

The claim is false of course. For example, an option which yields $1000 with probability 0.1 and nothing with probability 0.9 is preferable to an option which yields $1 no matter, contrary to the claim. Furthermore, it's possible that with high probability X yields higher utility than Y, and similarly for Y and Z, and similarly

[10] Alternatively, we could dismiss some estimators on other grounds, and then use a long-run defense to argue that among the remaining estimators we should prefer consistent ones. The key point is that a long-run defense will come in at some stage or other.

for Z and X, in which case the claim implies, absurdly, that X is preferable to Y, and Y to Z, and Z to X.[11]

But compare:

> If with probability 1 option X yields higher utility than option Y then X is preferable to Y.

This claim is surely true. Replacing 'high probability' by 'probability 1' avoids the earlier counterexamples.

Now, it might seem that a premise of the long-run defense—if (1) is true, so is (2)—is an instance of the first claim, and therefore suspect. For the premise seems to be saying: betting \$100 is preferable to betting, say, \$50 in the horizontal iteration, because with high probability betting \$100 yields more money than betting \$50.

But in fact the premise, although not strictly an instance of either claim, is more similar in spirit to the second, which is true, than the first, which is false. For the premise actually says: betting \$100 is preferable to betting, say, \$50, because the probability that betting \$100 yields more money than betting \$50 by Day n tends to 1 as n tends to infinity. That's another reason to take the long-run defense seriously.

2.3.3 Dynamic and one-off choices

When someone faces a sequence of choices, we can either view her as facing one big decision problem or as facing many little decision problems. A natural idea, at first glance, is that which perspective we take doesn't matter: whatever she should do in the big problem will decompose into what she should do in each little problem, and vice versa.

In fact, things are more complicated. Sometimes, it seems, the perspective does matter: what the agent should do in the big problem comes apart from what she should do in each little problem.[12] There are various theories about what the agent should do in such cases. For a formal treatment, see McClennen (1990). Still, even if

[11] This is the phenomenon of non-transitive dice.

[12] For a survey, see Andreou (2017). For examples involving infinite sequences of choices, as in the horizontal iteration of The Gamble, see Rayo (2019: 66–69).

the perspective sometimes matters, it doesn't always matter. If it turns out that the perspective doesn't matter in the horizontal iteration of The Gamble, what follows?

Well, consider an abstract procedure for working out how much you should bet in The Gamble. Suppose that the amount you should bet in The Gamble depends only on X, Y, Z. Call that amount, whatever it is, a. Now imagine an infinite sequence of gambles, all like The Gamble with respect to X, Y, Z. First, view the sequence as many little decision problems. Conclude in each case that you should bet a. After all, each case is like The Gamble with respect to X, Y, Z. Next, view the sequence as one big decision problem. Suppose you work out, somehow or other, that you should bet b in every case. So, assuming that the perspective—many little problems or one big problem—doesn't matter, conclude that $a = b$.

We're assuming, for the sake of argument, that the perspective doesn't matter in the horizontal iteration of The Gamble. So, I suggest, we can apply this abstract procedure in order to support the long-run defense of betting \$100, by supporting the premise that if (2) is true then so is (3).

Indeed, assume for conditional proof that (2) is true. How much Mac should bet in The Gamble—call it a—depends only on the payoffs, win probability and her utility function. The horizontal iteration of The Gamble is an infinite sequence of gambles, all like The Gamble with respect to these properties. So, viewing the sequence as many little problems, on each day Mac should bet a. But, viewing the sequence as one big problem, by assumption Mac should bet \$100 every day. Given that the perspective doesn't matter, $a = \$100$. Mac should bet \$100 in The Gamble. In other words, (3) is true. Discharging our assumption, if (2) is true then (3) is true.

In short, a natural idea is that what does best in a series of choices coheres with what does best in each component choice. That idea doesn't always hold, but if it does hold for the horizontal iteration of The Gamble, then it supports a premise of the long-run defense of betting \$100. That's a third reason to take the long-run defense seriously.

2.4 The long-run defense of betting proportion 0.25

The *expected growth rate of betting proportion m of $100* is $(1+2m)^{0.5} \cdot (1-m)^{0.5}$. The proportion of $100 which maximizes this quantity is 0.25. So to maximize expected growth rate in The Gamble is to bet proportion 0.25 of $100, which is $25.

But should Mac maximize expected growth rate in The Gamble? You might say she should, because maximizing expected growth rate does best in the long-run.[13] Let's unpack what that means.

In the *vertical iteration* of The Gamble, you face The Gamble repeatedly: on Day 1, Day 2, Day 3, and so on forever. You have to bet the same proportion of your bankroll every day.

What's your bankroll? Think about it like this. Before Day 1, you set aside $100 from your savings and put it in a pot. You resolve to draw all your bets from and deposit all your winnings into this pot. Your bankroll is the money in the pot. Using the pot means you won't get carried away and end up losing your savings, because your maximum loss, as compared to your initial wealth, is the amount you put in the pot to start with: $100. For that reason, gamblers often follow this approach.

For example, suppose you bet proportion 0.4. On Day 1, your bankroll is $100. So you bet $40 and win, say. Your bankroll on Day 2 is $180. So you bet $72 and lose, say. Your bankroll on Day 3 is $108. So you bet $43.20 and win, say. And so on.

Suppose you bet proportion m. By the Weak Law of Large Numbers, the probability that [the number of wins by Day n divided by n is within ϵ of 0.5] tends to 1 as n tends to infinity, for any $\epsilon > 0$. Therefore the probability that [your growth rate on Day n is within ϵ of the expected growth rate] tends to 1 as n tends to infinity, for any

[13] J. L. Kelly (1956) and H. A. Latané (1959) independently came up with the idea of maximizing expected growth rate. The idea has generated a lot of discussion in economics. MacLean et al. (2011) is a useful collection. Discussion of the idea in economics has focused, so far as I know, on whether you should maximize expected growth rate when you actually face a long sequence of decisions. (Think, for example, about an investor who has to decide how to manage their portfolio over the years.) I'm using the idea for a different purpose: to give a long-run defense of maximizing expected growth rate in a one-off decision. So I'm not pointing out a new result, but using an old result for a new purpose.

$\epsilon > 0$. Betting proportion 0.25 maximizes expected growth rate. Therefore—and this is the key consequence—in the vertical iteration of The Gamble, someone who bets proportion 0.25 eventually tends to make more money than someone who doesn't.

To state the key consequence more formally. Suppose in the vertical iteration you bet proportion 0.25 and I bet some other proportion. Then for any $\delta > 0$, there exists an N such that for all $n > N$ the probability that you've made more than me by Day n is at least $1 - \delta$.[14]

So much for the vertical iteration of The Gamble. How does it bear on The Gamble itself? Distinguish three claims:

(α) In the vertical iteration of The Gamble, someone who bets proportion 0.25 eventually tends to make more money than someone who doesn't.

(β) Mac should bet proportion 0.25 in the vertical iteration of The Gamble.

(γ) Mac should bet proportion 0.25 in The Gamble.

(α) is true. If (α) is true, so is (β). If (β) is true, so is (γ). Therefore (γ) is true. That's the long-run defense of betting proportion 0.25 in The Gamble.

In the horizontal iteration, you bet the same amount between \$0 and \$100 every day. Betting \$100 is the *replicable* choice. In the vertical iteration, you bet the same proportion of your bankroll every day. Betting proportion 0.25 is the *sustainable* choice.

The long-run defenses of betting \$100 and betting proportion 0.25 in The Gamble are parallel. If one is sound, so is the other. But their conclusions are incompatible. Therefore neither is sound.

To endorse one long-run defense over the other, you must break the symmetry: to find a difference which makes a difference. I don't think that can be done.

You might say: "In The Gamble, you bet an amount between \$0 and \$100. In the horizontal iteration, you do too, repeatedly. But in the vertical iteration, you don't:

[14] As before, applying the Strong Law yields a stronger result (that with probability 1, eventually someone who bets proportion 0.25 makes more money than someone who doesn't) but also asks more of our imagination.

the amount you bet typically changes from day to day, and needn't be between \$0 and \$100. That breaks the symmetry. The iteration which matters when it comes to The Gamble is the horizontal, not the vertical."

But that line of thought is wrong. When you face The Gamble, we can parameterize your options in two ways: as betting any amount between \$0 and \$100, or as betting any proportion of your bankroll. The difference is in the representation, not the represented. So you might equally well say: "In The Gamble, you bet a proportion of your bankroll. In the vertical iteration, you do too, repeatedly. But in the horizontal iteration, you don't." If the line of thought above seems plausible, it's because you're privileging amounts over proportions. But there's no reason to do that.

2.5 How cheap are long-run defenses?

A bet's having a long-run defense isn't a decisive reason in its favour. For at least two rival bets (\$100, proportion 0.25) have long-run defenses. Still, perhaps a bet's having a long-run defense, even if not a decisive reason in its favour, is still a *pro tanto* reason in its favour.

The more bets which have long-run defenses, the less defensible is that view. I conjecture that many bets have long-run defenses. In this section, I formalize the conjecture, turning it into a conjecture in probability theory.

A preliminary step. By definition, in the horizontal iteration of The Gamble, you face The Gamble repeatedly: on Day 1, Day 2, Day 3, and so on forever. Let me change that definition slightly. The change doesn't affect the substance of anything I've said, but only the wording, and will make the coming material easier to follow. Here's the change. In the horizontal iteration of The Gamble, an infinite sequence of people A_1, A_2, A_3, ... each face The Gamble in turn: A_1 on Day 1, A_2 on Day 2, A_3 on Day 3, and so on forever. They all bet the same amount between \$0 and \$100.

On the old definition, one person faces The Gamble many times, bets the same amount every time, and we care about how much money that person makes. On the new definition, many people face The Gamble once, all betting the same amount, and we care about how much money the group makes. In the rest of this section, I'll use the new definition.

Recall the key facts. On Day n in the horizontal iteration of The Gamble, Agents 1 to n have each bet once. People who bet \$100 eventually tend to make more money than people who don't. On Day n in the vertical iteration of The Gamble, Agent 1 has bet n times. Someone who bets proportion 0.25 eventually tends to make more money than someone who doesn't. I conjecture that on other ways to iterate The Gamble, other bets do best.

Now to formalize the conjecture. An *iteration function* takes a number, representing the day, and returns a list of numbers, representing how many times each agent has bet on that day. All agents bet the same proportion of their bankroll each day, and the same proportion as each other.[15]

For example, the horizontal iteration is represented in this scheme by the function:

$$H : n \mapsto [1, ..., 1]$$

which takes a number n and returns a list of length n, representing that at the end of Day n, $A_1, ..., A_n$ have each bet once.

And the vertical iteration is represented in this scheme by the function:

$$V : n \mapsto [n]$$

which take a number n and returns a list of length 1, representing that at the end of Day n, A_1 has bet n times.

Other iterations suggest themselves. For example, consider a function:

[15] Formally, an iteration function h takes a number and returns a list of numbers. The i^{th} entry of $h(n)$ represents the number of times A_i has bet at the end of Day n. We insist that for $n \leq m$, the length of $h(n)$ is less than or equal to the length of $h(m)$, and each entry of $h(n)$ is less than or equal to the corresponding entry of $h(m)$. The constraints reflect the facts that as the days pass the number of agents who have bet doesn't decrease and the number of times an agent has bet doesn't decrease. The definition also enforces a helpful book-keeping constraint: that A_i starts betting before A_{i+1}.

$$f : n \mapsto [n, n-1, n-2, ..., 1]$$

which takes a number n and returns a list of length n, representing that at the end of Day n, A_k has bet $n+1-k$ times.

Imagine a grid of points $(i, j) \in \mathbb{N}^2$. Circling point (i, j) represents that A_i has bet at least j times. We can visualize an iteration function as an expanding collection of points on the grid. The horizontal iteration is an initial segment of the grid's bottom row. The vertical iteration is an initial segment of the grid's leftmost column. The example just above is a triangle of points, with vertices at $(n, 1)$, $(1, n)$, and $(1, 1)$.

We've now defined a wide class of ways to iterate The Gamble. You can think of them as 'weighted averages' of the horizontal and vertical iterations.

Fix an iteration function, f. Let's say that proportion m_1 *does better* than proportion m_2 in f just if the probability that [on Day n, people who bet m_1 have made more money than people who bet m_2 in f] tends to 1 as n tends to infinity. Let's also say that proportion m *does best* in f just if for every other proportion m', proportion m does better than proportion m' in f. We know that proportion 1 does best in H and proportion 0.25 does best in V.

We can now state the conjecture formally:

> **Conjecture.** For many proportions m, there exists an iteration function f such that m does best in f.

If the conjecture is true, then long-run defenses are cheap, in which case a bet's having a long-run defense is only a weak reason in its favour, or perhaps no reason at all.

2.6 Towards resolving the conjecture

The conjecture is about how bets do in the long-run: on Day n, as n tends to infinity. You can get a sense, short of a proof, of how a bet does in the long-run by seeing how it does in the short-run: up to Day 1000, say. I wrote two simple computer programs for that purpose. In the first program, the user specifies an iteration function, a number n, and some bets, and then the program simulates n coin tosses and returns a plot showing how each bet did from Day 1 to Day n. The user can also change the win probability and payoff odds in order to see how bets do, not only in iterations of The Gamble, which has win probability $\frac{1}{2}$ and payoff odds 2:1, but also in iterations of other gambles. In the second program, the user specifies an iteration function, a number n, two bets and a number of trials t, and then the program simulates n coin tosses t times and returns the number of times the first bet was doing better than the second on Day n. When t is large, that number is a guide to the probability that the first bet does better than the second by Day n. You can access the programs here: github.com/cosmo-grant/long-run-and-short-run.

Of course, short-run behaviour doesn't settle long-run behaviour: one bet can tend to do better than another on Day 100, or Day 1000, or Day 10,000, or whatever, but still do worse asymptotically. You might question whether short-run behaviour is even *evidence* of long-run behaviour. For example, suppose that out of 100 trials, proportion 0.4 was doing better than proportion 0.6 on Day 1000 a majority of times. Is that evidence that proportion 0.4 does better than proportion 0.6 in the long-run?

This paper has been about whether the long-run bears on the short-run. Thinking about the simulations brings out the converse question: whether the short-run bears on the long-run. All I'll add here, about the simulations, is that in a wide range of fields people do take the short-run to bear on the long-run, just as they take the long-run to bear on the short-run. So if you take the simulations to bear on the conjecture, you're in good company.

Still, proofs are better than simulations. Let f be an iteration function such that, at the end of Day n, A_1, ..., A_n have each bet n times. (So, thinking about the grid of points $(i, j) \in \mathbb{N}^2$, f is an expanding square, with side-length incrementing by 1 each day.) Here's a proof, due to Ewain Gwynne, that proportion 0.25 does best in f. Let $B^m_{n,i}$ be A_i's bankroll at the end of Day n assuming she bets proportion m (and, for convenience, assuming her initial bankroll is \$1 instead of \$100). By applying Hoeffding's Inequality to the number of times A_i has won by the end of Day n, we have that with probability at least $1 - 2e^{-2\epsilon^2 n}$:

$$(1-m)^{n(\frac{1}{2}+\epsilon)}(1+2m)^{n(\frac{1}{2}-\epsilon)} \leq B^m_{n,i} \leq (1-m)^{n(\frac{1}{2}-\epsilon)}(1+2m)^{n(\frac{1}{2}+\epsilon)}$$

for any ϵ, n, and each $i = 1, \dots, n$. By the union bound, we have that with probability at least $1 - 2ne^{-2\epsilon^2 n}$, all the $B^m_{n,i}$ lie within these bounds simultaneously. So, letting S^m_n be the agents' total bankroll at the end of Day n, with probability at least $1 - 2ne^{-2\epsilon^2 n}$:

$$n(1-m)^{n(\frac{1}{2}+\epsilon)}(1+2m)^{n(\frac{1}{2}-\epsilon)} \leq S^m_n \leq n(1-m)^{n(\frac{1}{2}-\epsilon)}(1+2m)^{n(\frac{1}{2}+\epsilon)}$$

for any ϵ, n. Furthermore, if $(1-m)(1+2m) > (1-m')(1+2m')$ then for sufficiently small ϵ (depending on m, m') and any n:

$$n \cdot (1-m)^{n(\frac{1}{2}+\epsilon)}(1+2m)^{n(\frac{1}{2}-\epsilon)} > n \cdot (1-m')^{n(\frac{1}{2}-\epsilon)}(1+2m')^{n(\frac{1}{2}+\epsilon)}$$

That is, the lower bound for S^m_n is higher than the higher bound for $S^{m'}_n$. So with probability at least $1 - 2ne^{-2\epsilon^2 n}$, we have $S^m_n > S^{m'}_n$. Finally, noting that for any ϵ, $1 - 2ne^{-2\epsilon^2 n} \to 1$ as $n \to \infty$ and $\frac{1}{4}$ maximizes $(1-m)(1+2m)$ over $m \in [0,1]$, we have the result.

Note that Gwynne's form of argument establishes two more general facts as well. First, it establishes what proportion does best in f when we vary the win probability and payoff odds: it's the proportion which maximizes expected growth rate. Second, take any iteration function such that at the end of Day n, $g(n)$ agents have each bet n times. (So, thinking about the grid of points again, g is an expanding rectangle, with height increasing by 1 each day and length controlled by g.) Then some $m \neq \frac{1}{4}$ does best only if g grows at least exponentially. So Gwynne's result narrows down the search space for resolving the conjecture.

The informal conjecture is that many bets have a long-run defense. In the previous section, I formalized that conjecture, turning it into a conjecture in probability theory, and therefore opening it up to simulations and to proofs like Gwynne's. But let's briefly return to the informal conjecture.

Remember the horizontal iteration: at the end of Day n, A_1, ..., A_n have each bet m dollars once. Let k be the number of agents who win. Then their total bankroll is:

$$k \cdot (100 + 2m) + (n - k) \cdot (100 - m)$$

and their average bankroll is:

$$\frac{k}{n} \cdot (100 + 2m) + \frac{n-k}{n} \cdot (100 - m)$$

By the Weak Law of Large Numbers, $\frac{k}{n}$ eventually tends to be close to $\frac{1}{2}$, so the average bankroll eventually tends to be close to $\frac{1}{2} \cdot (100 + 2m) + \frac{1}{2} \cdot (100 - m)$. It follows that what does best in the horizontal iteration is what maximizes that quantity, which is $100.

Now the vertical iteration: at the end of Day n, A_1 has bet n times, betting proportion m of her bankroll each time. Let k be the number of times she wins. Then her final bankroll is:

$$100 \cdot (1 + 2m)^k \cdot (1 - m)^{n-k}$$

and her growth rate is:

$$(1 + 2m)^{\frac{k}{n}} \cdot (1 - m)^{\frac{n-k}{n}}$$

As before, by the Weak Law of Large Numbers, $\frac{k}{n}$ eventually tends to be close to $\frac{1}{2}$, so the growth rate eventually tends to be close to $(1 + 2m)^{0.5} \cdot (1 - m)^{0.5}$. It follows that what does best in the vertical iteration is what maximizes that quantity, which is 0.25.

The recipe is clear: imagine a way of iterating The Gamble (horizontal, vertical); find a suitable measure of the agents' performance (average, growth rate); apply the Weak Law of Large Numbers to find what does best ($100, 0.25). So instead of trying to show what does best in this or that iteration function, as per the formalized conjecture, we can try to follow this recipe. The challenge is to find fresh ingredients.

Chapter 3

Mistakes about conventions and meanings

3.1 Overview

The *Standard View* is that, other things equal, speakers' judgments about the meanings of sentences of their language are correct. After all, we *make* the meanings, so how wrong can we be about them? I put pressure on the Standard View: for quite straightforward reasons, speakers can be radically mistaken about meanings.

Lewis (1969) gave a theory of convention in a game-theoretic framework. He showed how conventions could arise in repeated *coordination games.* He also introduced a special kind of coordination game, a *signaling game*, and showed how, as a special case of his theory, conventional meanings could arise in repeated signaling games.

I put pressure on the Standard View by building on Lewis's framework. I construct coordination games in which the players can be wrong about their own conventions. The key idea is simple: knowing your own strategy and payoff needn't determine what the others do, so leaves room for false beliefs about the convention. Guided by the coordination games, I consider the special case of signaling games, and construct signaling games in which the players can be wrong about their messages' meanings. We make the meanings, but we can still be wrong about them.

Perhaps we already know that speakers can be wrong about meanings, because of Twin Earth cases, or semantic deference, or the like (Putnam 1975; Burge 1979). Still, the examples I give are interesting: they are simple, explicit, new in kind, and based on an independently plausible meta-semantic story.

Section 2 describes Lewis's game-theoretic framework. Sections 3–4 describe and discuss coordination games which leave room for mistakes about conventions. Section 5 shows that the Standard View is no straw man. Sections 6–7 describe and discuss signaling games which leave room for mistakes about meanings. Section 8 sums up.

3.2 Lewis's game-theoretic framework

3.2.1 A paradigm coordination game

Each day at noon, you and I play a game. You're the row-chooser: you play U or D. I'm the column-chooser: I play L or R. If we play UL or DR, we get lunch that day; else, we go hungry. Call this the *Simple Game.*

	L	*R*
U	1	0
D	0	1

Figure 3-1: Payoff matrix in the Simple Game

Imagine the first day we play the game. What will we do? No strategies suggest themselves. If I knew that you would play U, I would play L. If you knew that I would play R, you would play D. But neither of us knows how the other will play. More or less at random, you play U and I play R. We go hungry. Each thereby learns about the other. For example: you learn that I played R and I learn that you played U.[1]

The next day, we play again. What will we do? What we learned yesterday might help. If you think I'm stubborn, sticking with R, you will switch to D. If I think you're accommodating, switching to D, I will stick with R. But it might not help. If

[1] Perhaps we learn more than this. You learn that I learn that you played U; I learn that you learn that I played R. And so on, up the hierarchy.

each of us is stubborn, or each thinks the other is accommodating, each of us will stick, and we'll go hungry again. If each of us is accommodating, or each thinks the other is stubborn, each of us will switch, and we'll go hungry again. Each of us uses what she learned yesterday, together with what she already knows about the other, to predict how the other will play, knowing that the other is doing the same.

Eventually, by good luck and good sense, we coordinate: you play U and I play L. The next day the memory of our success is fresh in our minds, and we repeat it. Success breeds success. We soon cease to worry. We coordinate day after day, each happy with her own choice and confident of the other's. We have established a convention: *play UL.*

3.2.2 A paradigm signaling game

Nature flips a coin and shows you the result. You then slip a note under my door, reading either # (pound) or & (amp). Then I have to guess how the coin landed. If I'm right, we get lunch that day; else, we go hungry. Each of us wants me to guess correctly. We must attempt this indirectly, by coordinating our strategies. Call this the *Coin Game.*

The first day, Nature flips the coin and it lands heads. You must either send me # or &, trying to signal how it landed so that I'll guess correctly. But which to send?

Imagine it. Neither message will be much help. So you send one at random, # say. Now I must guess. Your message was no help. So I guess at random, tails say. We go hungry. As before, we each thereby learn about the other. And as before, each of us uses what she learns, together with what she already knows about the other, to predict how the other will play, knowing that the other is doing the same.

Eventually, more or less by chance, we coordinate: the coin comes up heads, you send #, and I guess heads. We eat. When the coin comes up heads again the next day, our success is fresh in our minds, and we repeat it. When the coin comes up tails, what will we do? If we're sensible, surely you'll switch your message and then I'll switch my response. Success breeds success. Through good luck and good sense, we've fixed on complementary strategies: you send # when heads and & when tails;

I guess heads given # and tails given &. We soon cease to worry. We coordinate day after day, each happy with her own choice and confident of the other's.

When we started playing the game we chose our moves more or less at random. In the course of playing the game over and over again, we have established a convention, and the messages have acquired meanings: # means *the coin landed heads* and & means *the coin landed tails.*

3.2.3 Taking Lewis's theory as a starting point

Lewis (1969) showed how conventions could arise from repeated coordination games, and, as a special case, how meanings could arise from repeated signaling games. I use Lewis's game-theoretic framework, and take his theory of convention as a starting point. But I don't endorse all the details of the theory. What I say about the examples in Section 3 conflicts with his theory, and the signaling games in Section 6 generalize his. I flag the differences as we go along. See Section 4.6 in particular. Here, I'll briefly discuss two worries about taking Lewis's theory as a starting point.

First worry. You might worry that small-scale interactions, like those in the Simple Game or the Coin Game, don't give rise to conventions. After all, if I make the coffee and my partner makes the eggs every morning, though we could equally well have done the reverse, it's unnatural to say that we have established a convention.

Reply. First, I agree it's unnatural to say that my partner and I have established a convention. But I suggest that it's unnatural, not because of the scale (two people, low stakes), but because of other features, perhaps that, unlike in the Simple Game or the Coin Game, neither my partner nor I had to reason about the other in deciding what to do. Second, a range of theorists do think that even small-scale interactions, like those in the Simple Game and Coin Game, can be conventions. Take, for example, Hume's rowers (2000, 3.2.2) or Margaret Gilbert's walkers (1990b). Third, even if small-scale interactions don't give rise to conventions, that doesn't undermine the main claims of this paper. For the examples in the paper can easily be scaled-up, so that they involve lots of people and high-stakes situations. Fourth, I say with Lewis

(p. 3) that "what I call convention is an important phenomenon under any name."[2]

Second worry. When we communicate, we do more than play signaling games—much more.

Reply. Meaning in signaling games is a simple example of linguistic meaning. A simple example is still an example; a special case is still a case. Signaling games are not *simplified models* of meaning; they are *simple cases* of meaning. (Compare: heredity in pea plants is a simple case of heredity, not a simplified model of heredity.) In studying signaling games, we have not changed the subject.

Here's Lewis:

> If we endow a hypothetical community with a great many [. . .] signaling conventions [where the messages are written or spoken] for use in various activities, with verbal expressions suitably chosen ad hoc, we shall be able to simulate a community of language users—say, ourselves—rather well. An observer who stayed in the background watching these people use conventional verbal messages as they went about their business might take a long time to realize that they were not ordinary language users. [. . .] Yet it remains true that our hypothetical verbal signalers do not do anything we do not do. We just do more. Their use of language duplicates a fragment of ours. (pp. 142–3.).

Lewis's hypothetical signalers don't do anything we don't do. We just do more. If the hypothetical signalers can be wrong about the meanings of their own messages, so can we. If they can be radically wrong, so can we.

3.3 Mistakes about conventions in coordination games

Conventions arise in repeated coordination games. As a special case, meanings arise in repeated signaling games. Before focusing on the special case of meanings, I consider conventions in general. This section shows how players can be wrong about their own

[2] All orphan page numbers refer to Lewis (1969)

conventions. I describe four games—the *Three-Player Game*, the *Five-Player Game*, the *Nature Game*, and the *Cycle Game*—and how each might turn out. I assume throughout that the structure of the situation is common knowledge, the players play repeatedly, and each player knows her own move and payoff but doesn't observe the other players' moves.

Each game leaves room for the players to be wrong about their own convention: on some ways things might turn out, the players establish a convention but are wrong about which. The key idea is simple: knowing your own move and payoff needn't determine what the others do, so leaves room for false beliefs about the convention.

The four examples are abstract. That helps make them clear and concise. But ordinary practical situations have the same structure as the examples. See Section 4.3. The examples are not mere theoretical curiosities.

3.3.1 First example: the Three-Player Game

Rowena, Colin and Mattea are playing a coordination game. Rowena is the row-chooser: she plays Up or Down (U or D). Colin is the column-chooser: he plays Left or Right (L or R). Mattea is the matrix-chooser: she plays West or East (W or E). If they play ULW or DRW or DLE or URE, each gets lunch; else, nothing.[3] Call this the *Three-Player Game*.

$$
\begin{array}{c|c|c|}
 & L & R \\
\hline
U & 1 & 0 \\
\hline
D & 0 & 1 \\
\hline
\multicolumn{3}{c}{W}
\end{array}
\qquad\qquad
\begin{array}{c|c|c|}
 & L & R \\
\hline
U & 0 & 1 \\
\hline
D & 1 & 0 \\
\hline
\multicolumn{3}{c}{E}
\end{array}
$$

Figure 3-2: Payoff matrix in the Three-Player Game

Imagine the first day Rowena, Colin and Mattea play the game. What will they do? No strategies suggest themselves, and they play more or less at random: URW, say. They go hungry. Each thereby learns about the others. For example, Rowena

[3] I give the actions different labels: 'U' or 'D' for Rowena, 'L' or 'R' for Colin, 'W' or 'E' for Mattea. That makes things easier to follow. But different labels needn't mean different actions. For example, U, L, and W can all be the same action. See Section 4.3 for an example.

learns that Colin and Mattea either played RW or LE.[4]

The next day they play again. What will they do? What they learned yesterday might help. If Rowena knows that Colin and Mattea are stubborn, sticking with their strategies, then she will switch to D, so that they will coordinate next time. But it might not help. If Rowena doesn't know how Colin and Mattea will react, then she won't know how to react either.

They play the game day after day, learning about each other as they go. Eventually, by good luck and good sense, they coordinate: Rowena plays U, Colin plays L, Mattea plays W. The next day, the memory of their success is fresh in their minds, and they repeat it. Success breeds success. They coordinate day after day, each happy with her own choice and confident of the others'. They have established a convention: *play ULW.* So far, so familiar.

The structure of the game is common knowledge. Each player knows her move and payoff. But that's not enough for each player to work out what the others are doing. Take Rowena. She knows that she plays U and that she gets lunch every time. But for all she knows, Colin and Mattea could be playing LW or RE. And similarly for Colin and Mattea.

None of the players knows what the convention is. Still, they may have beliefs about it. Suppose that, for one reason or another, Rowena believes that the convention is *play URE*; Colin, that it's *play DLE*, Mattea, that it's *play DRW.* Section 4.1 explains why they might have these beliefs.

Take Rowena again. If you asked her what the convention is, she'd say, "it's *play URE*"; if she were to switch roles with Colin, she would play R and expect him to play U; she would bet at long odds that Colin and Mattea play RE.

[4] A warning. After the players make their moves, each gets a payoff—lunch, for example. The numbers in the matrices—***utilities***—represent the players' preferences over the payoffs. Utilities are coarser-grained than payoffs: if a player is indifferent between two payoffs (chicken and beef, say), then those payoffs have the same utility, even though the player may be able to tell apart the outcome in which they get the one (chicken) from the outcome in which they get the other (beef). Now, for the games in this paper, it matters which outcomes the players can tell apart. So the payoffs matter, not just the utilities. Therefore—and this is the key point—in all the games interpret the utilities in the matrices as normal, as representing the players' preferences over the payoffs, except also assume that where the utilities are the same, the payoffs are the same too.

As it happens, they're all wrong, for the convention is *play ULW.* Every player is wrong about the convention.

3.3.2 Second example: the Five-Player Game

A, B, C, D, and E are playing a coordination game. Each player has two actions: 0 or 1. The outcome 10101, for example, is the outcome in which A, C, E play 1 and B, D play 0. If they play 00000, 11000 or 11111, each gets positive payoff; else, nothing. Call this the *Five-Player Game.*

They play day after day. Eventually they coordinate on 00000, each happy with her own choice and confident of the others'. They have established a convention: *play 00000.*

The structure of the game is common knowledge. Each player knows her own move and payoff. The only outcome consistent with A's move and payoff is 00000, and similarly for B. A and B know that the convention is *play 00000.* But C, D, E don't. For all they know, they could be playing 00000 or 11000.

C, D, E don't know what the convention is, but they may still have beliefs about it. Suppose that, for one reason or another, they believe, wrongly, that the convention is *play 11000.* Then a majority of players (C, D, E) have the same mistaken belief about their own convention.

3.3.3 Third example: the Nature Game

Roland and Col are playing a coordination game, but they're not sure which. Nature chooses the West Game or East Game at random. The game, once chosen, is fixed. Then Roland and Col play that game repeatedly. Roland chooses U or D; Col chooses L or R. Call this the *Nature Game.* The Nature Game is the two-player analog of the Three-Player Game, where Mattea's role has been taken by Nature, and Nature only gets to choose once.[5]

[5] The Nature Game is known as a Bayesian game, because Roland and Col have incomplete information about the payoffs.

	L	R
U	1	0
D	0	1

West Game

	L	R
U	0	1
D	1	0

East Game

Figure 3-3: Payoff matrices in the Nature Game

As it happens, Nature chooses the West Game. Roland and Col play day after day. Eventually they coordinate on UL, each happy with her choice and confident of the other's. They have established a convention: *play UL.*

The structure of the situation is common knowledge.[6] Each player knows her move and payoff. As before, that's not enough for either to work out what the other is doing. So neither player knows the convention. But they may still have beliefs about it. Suppose that, for one reason or another, they believe they're in the East Game, so Ro believes the convention is *play UR* and Col believes it's *play DL.* Both are wrong.

3.3.4 Fourth example: the Cycle Game

Rosie and Colt are playing the Cycle Game.[7] Rosie chooses U, M or D; Colt chooses L, C or R.

	L	C	R
U	1	1	0
M	0	1	1
D	1	0	1

Figure 3-4: Payoff matrix in the Cycle Game

They play day after day. Eventually they coordinate on UL, each happy with her choice and confident of the other's. Have they established a convention? Unlike the previous examples, either player can unilaterally deviate from UL (Rosie, by playing D; Colt, by playing C) without losing out. Perhaps that means that *play UL* is not

[6] The situation includes the initial choice by Nature. The game is either the West Game or East Game. The structure of the *game* is not common knowledge, since the players don't know which game Nature chooses. The structure of the *situation* is common knowledge.

[7] Daniel Rothschild suggested this kind of example, but I don't mean to imply that he agrees with what I say about it.

a convention. (Lewis thought so.) If so, strike out the example and jump to the next section. If not, I say that Rosie and Colt have established a convention: *play UL*.

As before, neither player knows the convention, but they may still have beliefs about it. Suppose that, for one reason or another, Rosie believes the convention is *play UC* and Colt believes that the convention is *play DL*. Both are wrong.

3.3.5 Further examples

The key idea behind all four examples is that knowing your own move and payoff needn't determine what the others do, so leaves room for false beliefs about the convention. The examples exploit that idea differently: in the Three-Player Game, every player is wrong, although wrong in different ways; in the Five-Player Game, a majority of players are wrong in the same way, although some are right; in the Nature Game, both players are wrong, because they are wrong about the payoff structure; in the Cycle Game, both players are wrong, assuming that the Cycle Game can give rise to conventions.

The idea applies generally. For example, can we find a game which leaves room for every player to be wrong about the convention *and* most of the players to have the same mistaken belief about the convention? Yes, it's not hard.[8] The recipe is simple. Pick your pattern of mistakes and cook up a game to match.

3.4 Discussion of examples

3.4.1 Why the false beliefs?

In each example, some of the players don't know the others' choices. Take Rowena in the Three-Player Game: for all she knows, Colin and Mattea could be playing LW or RE. Even so, in the scenario I've described, she believes that they're playing RE.

[8] Suppose, for example, that there are eight players, each with two actions, 0 or 1. If an even number of players choose 1, all get lunch; else, nothing. Suppose they all eventually play 0, getting lunch every time, but, for one reason or another, two believe that they both play 0 and the other six play 1, while the six believe that those two play 1 and they all play 0.

Why? Her belief seems to be unreasonable. And as with the Three-Player Game, so with the others.

Not so. On perfectly ordinary ways of fleshing out the scenarios, the players' beliefs about the conventions, although false, are reasonable. I'll focus on Rowena but what I say carries over to the other players and the other games.

Carelessness

Perhaps Rowena is careless, not realizing that her evidence is consistent with LW as well as RE. The more complicated the game, the more excusable the carelessness.

Careless people can establish conventions. If you and I reasoned sloppily in the Simple Game in coming to coordinate, that does not undermine our convention *play UL*. If Rowena, Colin and Mattea reason sloppily in the Three-Player Game, that does not undermine their convention *play ULW*.

False antecedent beliefs

Perhaps Rowena started the game with false beliefs about Colin and Mattea. For example: that Colin is stubborn and Mattea is accommodating; or that Colin wrongly thinks that Mattea is accommodating. Her antecedent beliefs, even if false, may be justified. Perhaps they're all friends. In any case, her antecedent beliefs, together with how things happen to go in the early rounds, lead her to believe that the convention is URE. Successive iterations reinforce her belief.

Or Rowena might have false beliefs, not about her opponents' *styles* of play (accommodating, or stubborn, or whatever), but about what *choices* they'll make. For example: that Colin will play R and Mattea will play E; or that Colin and Mattea will either play UR or LW.[9] Her antecedent beliefs may be justified. Perhaps one outcome is salient. In any case, as before, they lead her to believe that the convention is URE.

Sometimes in game theory we assume that the players start from a position of

[9] The belief that Colin and Mattea will either play UR or LW won't lead Rowena to a false belief about the convention. I mention it just to point out that a player may have correlated beliefs about her opponents' actions. In other games, correlated beliefs may lead to false beliefs about the convention.

radical uncertainty about each other. What exactly that assumption amounts to isn't always clear, but in any case it excludes the sort of antecedent beliefs I've been describing. However, the assumption is optional, not required. It's perfectly legitimate to suppose that the players have antecedent beliefs about each other. Kasparov didn't cease to play chess because he had true antecedent beliefs about what his opponent would do; I don't cease to play chess because I have false antecedent beliefs about what my opponent will do. We're not going beyond standard game theory by imagining that Rowena has the sort of antecedent beliefs I've been describing.

Even Rowena's being sure of what her opponents will do or having correlated beliefs about her opponent's choices, are perfectly consistent with standard game theory. In particular, such beliefs are perfectly consistent with the assumption that the players' choices are causally independent (and that causal independence is common knowledge).[10]

False antecedent beliefs about the others don't undermine a convention. If when we play the Simple Game you wrongly think I'm accommodating, or that I'll play R, that doesn't undermine our convention. Given Rowena's false antecedent beliefs, impeccable reasoning might lead her to believe that the convention is URE. Her antecedent beliefs don't undermine the convention. Impeccable reasoning doesn't undermine it either. So nor does the end result: her mistaken belief that the convention is URE.

3.4.2 Are the regularities conventions?

If the regularities in the examples are not conventions, then the examples don't show that people can be mistaken about their own conventions, for there are no conventions for them to be mistaken about. But the regularities in the examples are conventions. Again, I focus on the Three-Player Game but what I say carries over to the other examples.

[10] See Stalnaker (1998: 43–4) for discussion of this point. To borrow one of his examples, suppose my partner and I are in our voting booths on election day. How she votes is causally independent of how I vote. You may have no idea how either of us will vote, but still be confident (and justifiably so) that, however we vote, we'll vote the same way.

First argument

If Rowena, Colin and Mattea established some convention or other in the Three-Player Game, then the regularity *play ULW* is a convention. They did establish some convention or other. Therefore the regularity *play ULW* is a convention.

To defend the first premise, suppose Rowena, Colin and Mattea established a convention other than ULW. What is it? A convention is, in particular, a regularity in behaviour. No outcome other than ULW is a regularity in behaviour. So no outcome other than ULW is a convention. Could the convention be some regularity other than an outcome of the game? I don't see what. If they established some convention or other, the regularity *play ULW* is a convention.

To defend the second premise, consider the similarities between the Simple Game and the Three-Player Game: the players' interests coincide; the players initially choose more or less at random; each uses what she learns, together with what she already knows about the others, to predict how the others will play in the knowledge that they are doing the same; several outcomes yield the preferred payoff; eventually, the players settle on actions, each happy with her own and confident of the others'; if anyone deviates, nobody benefits.

The Simple Game leads to a paradigm convention. The Three-Player Game resembles the Simple Game both in structure and game-play. The Three-Player Game leads to a convention too.

Second argument

Here's a happier way the Three-Player Game might turn out. Each player believes that the convention is ULW. Their beliefs are justified, for they started the game with justified antecedent beliefs about the others. Contrast that with how things actually turn out. Rowena wrongly believes the convention is URE; Colin, that it's DLE; Mattea, that it's DRW. Perhaps they were careless; or perhaps they had false antecedent beliefs about the others.

In the happier case, each player believes that *play ULW* is the convention. They're correct, and not by luck. Each started the game with true beliefs about the others (say, that Colin is stubborn and Mattea is accommodating). Their true antecedent beliefs, together with how things happen to go in the early rounds, lead them to believe that *play ULW* is the convention. Perhaps they're not in a position to know that *play ULW* is the convention. Nevertheless, their beliefs are reasonable. It's not a requirement on conventions that you only believe what you're in a position to know. That would be absurdly demanding. The regularity *play ULW* is a convention in the happier case. Don't punish the players for reasoning well about each other.

Back to the actual case: each player is mistaken about the convention, because of carelessness or false antecedent beliefs or whatever. Conventions don't depend on how attentive the players are, nor on how well they know each other. If the regularity *play ULW* is a convention in the happier case, it's a convention in the actual case too.

Putting things together, I conclude that the regularity *play ULW* is a convention in the actual case, as required.

3.4.3 Concrete examples

The four examples are abstract. That makes them clean, clear and concise. They may also seem contrived. But they aren't. For ordinary practical situations have the same structure as the abstract examples and could easily turn out in the same way. To dismiss the examples as mere theoretical curiosities is a mistake.

Concrete example for the Three-Player Game

Take the Three-Player Game. Suppose Rowena, Colin and Mattea all work in the same restaurant kitchen. At about 6pm each day, the chef puts two trays of squash in the oven to roast, one on top and one on bottom. After twenty minutes or so the trays need to be swapped, else the top one will burn and the bottom one won't caramelize. Swapping the trays is up to Rowena, Colin or Mattea, but it's one task among hundreds in a hectic kitchen and it isn't clear which of them will do it.

Imagine it. If none of them swaps the trays, the squash will be ruined. If just one of them does, it'll be fine. More is true. If two of them swap the trays in turn, neither realizing the other's plan, the trays will end up back where they started and the squash will be ruined. Similarly, if all three of them swap the trays in turn, it'll be fine. Of course, if all three of them swap the trays some effort is wasted, but that's small fry compared to roast squash.

The situation has the same structure as the Three-Player Game. U and D correspond to Rowena's swapping and not swapping the trays, and similarly for L and R, and W and E. In four outcomes (URE, DLE, DRW, where only Rowena or Colin or Mattea swap the trays, and ULW, where all three do), the squash is fine. In the other four outcomes (DRE, where none of them does, and URW, DLW, ULE, where two of them do), the squash is ruined.

What might happen? Here's one way things might go. At first, each of Rowena, Colin and Mattea thinks that one of the others will swap the trays. So no one does, and the squash is ruined. When the chef shouts at them they realize what happened and in the heat of the kitchen each goes away thinking that she alone will swap the trays from now on. The next day all three swap them, and the squash is fine. The kitchen is hectic and none of them sees the others do it. Their mistaken beliefs are reinforced. And so it goes on, each happily swapping the trays and confident that she alone is doing so. They have established a convention, *all swap the trays*, but each is wrong about what the convention is.

Concrete example for the Nature Game

Or take the Nature Game. Suppose Roland and Col are each trying to schedule a departmental reading group on Mondays. Roland's can start at 9am or 2pm; Col's at 11am or 4pm. Reading groups last two hours. Roland and Col are rivals and, out of pride, won't attend each other's group nor even coordinate times directly.

If the groups meet at 9am and 4pm, or 11am and 2pm, there'll be a break in between. If they meet at 9am and 11am, or 2pm and 4pm, there won't be. Roland and Col aren't sure which is best. On the one hand, four hours is a long time to concentrate. So if there is a break, maybe attendance will be higher, because most people will attend both. On the other hand, it's useful to have an uninterrupted morning or afternoon. So if there's a break, maybe attendance will be lower, because few people will attend both.

The situation has the same structure as the Nature Game. U and D correspond to 9am and 2pm; L and R correspond to 11am and 4pm. If a break is better, Roland and Col are in the East Game; if no break is better, they're in the West Game. Whether or not a break is better depends on people's preferences, which Roland and Col aren't sure about.

What might happen? Here's one way things might go. As it happens, no break is better. The first week, Roland's group meets at 9am and Col's at 4pm. Attendance is low. Roland suspects, wrongly but reasonably, that a break is better and that Col's group met at 11am. But, being stubborn, he sticks with 9am. Col suspects, wrongly but reasonably, that a break is better and that Roland scheduled his group for 2pm. And, being pragmatic, she switches to 11am. The next week, attendance is high. Roland and Col's mistaken beliefs are reinforced. And so it goes on, each happy with her own group's time and confident of the other's. They have established a convention, *meet at 9am and 11am*, but each is wrong about what the convention is.

3.4.4 Belief about regularities and belief about conventions

Distinguish two claims: Rowena believes that the regularity is *play URE*; Rowena believes that the convention is *play URE.* I've glossed over the difference. But in fact the claims are independent. Rowena might believe that the convention but not the regularity is *play URE*, because she might mistakenly believe that a convention doesn't require a regularity. She might believe that the regularity but not the convention is *play URE*, because she might mistakenly believe that the structure of the Three-Player Game rules out conventions.

I assume that whenever a player believes the regularity is *play such-and-such*, the player believes the convention is *play such-and-such.* The assumption isn't true in general. All I claim is that the four examples might turn out that way. The examples leave room for mistakes about conventions. They don't force mistakes.

3.4.5 Mistakes about conventions on the cheap

Suppose again that Rowena believes wrongly that the structure of the Three-Player Game rules out conventions. Perhaps she's a philosopher in the grip of a false theory of convention. And suppose as before that each player, including Rowena, does her bit of URE, happy with own choice and confident of the others'. Then the regularity *play URE* is a convention but Rowena believes it isn't. She's mistaken about her own convention.

Or suppose that each day when it's time to make her move Rowena plays U but afterwards forgets her choice, believing she played D. Perhaps conventions can survive this selective forgetfulness. Except when she makes her move, Rowena believes the convention is, say, *play DLE*; in fact, it's *play ULW*. Most of the time she's mistaken about her own convention.

Or suppose Rowena lacks the concept of a convention, so although she correctly believes that the regularity is *play ULW*, she doesn't believe that it's a convention.

These are mistakes about conventions on the cheap, relying on fussy details or exotic situations. The examples in Section 3 are abstract in order to make the structure clear. They are not fussy; they are not exotic.

3.4.6 How do the examples fit with Lewis's theory?

Are my claims about conventions sanctioned by Lewis's theory? No. This section spells out the details of his theory and shows why, according to it, my examples are not examples of conventions. So Lewis's theory is wrong.

We need some preliminary definitions. A *strategy profile* is a tuple of strategies, one for each player. A strategy profile is a *Nash equilibrium* if, for each agent, if she alone had done otherwise, she would be no better off. A strategy profile is a *coordination equilibrium* if, for each agent, if she alone had done otherwise, no one would be better off. In a Nash equilibrium, no one wishes that she alone had done otherwise ('no regret'); in a coordination equilibrium, no one wishes, of any one else, that she alone had done otherwise ('no resentment'). A strategy profile is a *proper coordination equilibrium* just if, for each agent, if she alone had done otherwise, no one would be better off *and someone would be worse off.*[11] A *coordination problem* is a situation of interdependent decision by two or more agents in which their interests largely coincide and which has two or more proper coordination equilibria.

Here is Lewis's first pass at a theory of convention.[12] A regularity R in agents' behaviour when in a recurrent situation S is a *convention* if and only if, in any instance of S,

(1) everyone conforms to R;

(2) everyone expects everyone else to conform to R;

(3) everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a proper coordination equilibrium in S.

How do my examples fit with Lewis's theory? Well, all four are situations of interdependent decision by two or more agents in which interests coincide. In the Cycle Game, there are three coordination equilibria, but none is proper; and the definition of coordination equilibrium doesn't apply to the Nature Game. So the regularities in these games are not conventions, on Lewis's theory.

In the Three- and Five-Player Games, there are two or more proper coordination equilibria. But the regularities aren't conventions, according to Lewis's theory, for another reason.

Take the Three-Player Game. Everyone conforms to the regularity *play ULW.* So (1) is true. And everyone prefers to conform to that regularity on condition that the

[11] Gilbert (1981) pointed out that the term 'proper coordination equilibrium' is ambiguous, and Lewis didn't make clear which he intended. Gilbert reports that Lewis clarified in private communication that this is what he had in mind.

[12] p. 42

others do. So (3) is true. But not everyone expects everyone else to conform to it. For example, Rowena expects Colin and Mattea to do RE, not LW. (Rowena does expect the others to conform to what she takes to be the actual regularity, URE, but she doesn't expect the others to conform to what is in fact the actual regularity, ULW.) So (2) is false. Therefore the regularity *play ULW* is not a convention, on Lewis's theory.

In short: according to Lewis's first pass at a theory of convention, none of the regularities in my four examples is a convention. Lewis's final theory is more complicated.[13] But the complications don't change things: on his final theory, too, the regularities are not conventions.

If Lewis's theory is correct, the examples don't show you can be wrong about your own conventions, for the examples are not examples of conventions. But Lewis's theory isn't correct. As I've argued, the regularities in the examples are conventions. Lewis provided a clear and simple framework, and he brought to light significant features of conventions. But not all the details of his theory are correct.

I don't have a replacement in mind. One could look for a minimal departure from Lewis's theory according to which the examples are examples of conventions. But that's not a profitable line of inquiry, since Lewis's theory is questionable in other respects too, like his insisting on a proper coordination equilibrium. (See Gilbert (1981, 1983) and Vanderschraaf (1998) for prominent criticisms.) Developing a theory of convention would support my argument, but it isn't essential.

3.5 Motivating The Standard View

In Sections 3–4, I described coordination games which, for quite straightforward reasons, leave room for mistakes about conventions. In Sections 6–7, I consider the special case of signaling games, and describe signaling games which leave room for mistakes about meanings. These games put pressure on the *Standard View* that, other things equal, speakers' judgments about the meanings of sentences of their lan-

[13] p. 78

guage are correct. This section motivates the next two by showing that the Standard View is not a straw man.

3.5.1 Language mavens

Steven Pinker devotes a chapter of his book *The Language Instinct* to criticizing language mavens, those self-appointed authorities on usage, who pull people up for using 'who' instead of 'whom', saying 'very unique', confusing 'disinterested' and 'uninterested', and the like. Pinker is concerned with syntax, not semantics, but the issues are parallel. Here's how the chapter starts:

> Imagine that you are watching a nature documentary. The video shows the usual gorgeous footage of animals in their natural habitats. But the voiceover reports some troubling facts. Dolphins do not execute their swimming strokes properly. White-crowned sparrows carelessly debase their calls. Chickadees' nests are incorrectly constructed, pandas hold bamboo in the wrong paw, the song of the humpback whale contains several well-known errors, and monkeys' cries have been in a state of chaos and degeneration for hundreds of years. Your reaction would probably be, What on earth could it mean for the song of the humpback whale to contain an "error"? Isn't the song of the humpback whale whatever the humpback whale decides to sing?

He continues:

> To a linguist or psycholinguist, of course, language is like the song of the humpback whale. The way to determine whether a construction is "grammatical" is to find people who speak the language and ask them. (Pinker 1995: 370)

As with syntax, so with semantics: to determine whether a construction is grammatical, find people who speak the language and ask them; to determine what a sentence means, find people who speak the language and ask them.

To find out the meaning of ‘literally’, or ‘decimate’, or ‘enormity’, or the other favourites of the language mavens, don’t argue from the armchair, nor fixate on etymology—just ask people! Only a language maven would disregard the judgments of ordinary speakers. After all, we make the meanings, so how wrong can we be about them? Other things equal, implies Pinker, speakers’ judgments about meanings are correct.

3.5.2 The Elicitation Method

How should we work out the truth-conditions of a sentence? Here is the *Elicitation Method*: Describe scenarios and ask many speakers whether the sentence is true relative to each scenario. If the speakers judge that the sentence is true relative to a scenario, the sentence *is* true relative to that scenario, or in other words the scenario does belong to the sentence’s truth-conditions. If the speakers judge that the sentence isn’t true relative to a scenario, the sentence *isn’t* true relative to that scenario, or in other words the scenario doesn’t belong to the sentence’s truth-conditions.

We could refine the Elicitation Method in all sorts of ways: ask only native speakers; instead of asking the speakers for a binary judgment (whether the sentence is true relative to a scenario), ask them for a graded judgment (how well the sentence fits a scenario); instead of asking about one sentence, ask about lots of sentences of the same form; instead of describing the scenarios in the same language as the sentence, use another language, or use pictures or videos; help the speakers distinguish infelicity from falsehood; avoid asking speakers who are corrupted by theory (semanticists, for example); add filler questions so as to obscure the experiment’s purpose; randomize the order of the questions...

The Elicitation Method, or some refinement of it, is a standard method in semantic fieldwork. Take Altshuler et. al. (2019, Chapter 1): “We will judge our progress in terms of how closely the system we develop tracks the intuitions speakers have about the truth of a sentence in different situations.” Or Winter (2016, p. 16): “Just as intuitive judgments about sentence grammaticality have become a cornerstone in syntactic theory, intuitions about entailments between sentences are central for natural language semantics." (Winter uses judgments about entailments, not about truth-conditions, but the approaches are equivalent.) Or Matthewson (2004, p. 369): "direct elicitation (including asking consultants for judgments) is an indispensable methodological tool." For thorough discussion, see Matthewson (2004) or Bochnak and Matthewson (2015).

The Elicitation Method is not the only way to work out the truth-conditions of a sentence. Other techniques are available. For example, you might gather texts or record conversations and extract truth-conditions from patterns of use. Or you might ask bilingual speakers to translate a sentence of the language under study into another language. And so on. Still, there is no question that eliciting speakers' judgments is a standard method in semantic fieldwork. The Standard View justifies the Elicitation Method. By better understanding how speakers' judgments about meanings can go wrong, we will better understand the limits of the Elicitation Method.

3.5.3 Lewis on knowledge of conventions

According to Lewis, participants in a convention are in a position to know what the convention is. As with conventions in general, so with conventions of language in particular: speakers of a language are in a position to know what their linguistic conventions are, or in other words, to know the meanings of their terms.

Lewis tempers the claim by pointing out that you may be in a position to know the convention without actually knowing it, that you may not be able to put what you know into words, and that snap judgments about the convention, like snap judgments about anything, may be wrong.[14] Still, other things equal, speakers' judgments about meanings are correct.

A view held by Lewis is a view worth taking seriously. In Sections 3–4, I argued that Lewis's view is wrong: participants in a convention may fail to know, or even be in a position to know, what the convention is; they may believe of some other regularity that it's the convention; they may easily state their mistaken belief verbally; they

[14] He describes a further qualification, too, involving Abelard's distinction between beliefs and expectations *in sensu composito* and *in sensu diviso*. See pp. 64–8.

may stand by their mistaken belief even after careful reflection. In Sections 6–7, I argue that Lewis's view is wrong about conventions of language in particular.

3.6 From coordination games to signaling games

It's easy to come up with coordination games which leave room for mistakes about conventions. Remember the key idea: knowing your own move and payoff needn't determine what the others do, so leaves room for false beliefs about the convention.

A signaling game is a special kind of coordination game. Signaling games are particularly interesting, since they give rise to meanings. In this section I apply the key idea to the special case of signaling games. I construct signaling games which leave room for mistakes about meanings.

Just as the examples from Section 3 go beyond the Simple Game, our paradigm coordination game, so too the examples in this section go beyond the Coin Game, our paradigm signaling game. We must be careful, when we go beyond the Coin Game, to ensure that a Lewis-style analysis of the messages' meanings still applies. Our examples must balance two desiderata: on some ways the game can turn out, the messages have meanings; and the players can be mistaken about the meanings. The first desideratum pulls us towards the Coin Game, for a Lewis-style analysis of meanings is most straightforward in games like that. The second desideratum pushes us away from the Coin Game, for games like that leave little room for mistakes about meanings. See Section 7.1 for further discussion.

I describe four games—the *ABC Game*, the *Two-Sender Coin Game*, the *Coin Game with Nature*, the *Signaling Cycle Game*—and how each might turn out. I assume as in Section 3 that the structure of the situation is common knowledge, the players play repeatedly, and each player knows her own move and payoff but doesn't observe the other players' moves. Each game leaves room for the players to be wrong about the messages' meanings: on some ways things might turn out, the messages acquire meanings but some players are wrong about some meanings.

Remember Lewis's hypothetical verbal signalers. They don't do anything we don't

State	**Sienna**	**Reg**	**Rae**	**Roy**	**Payoff**
A	!	A	C	B	1
B	&	C	B	A	1
C	#	B	A	C	1

Table 3.1: Actual strategies in the ABC Game.

do. We just do more. If the hypothetical signalers can be wrong about the meanings of their own terms, so can we. If they can be radically wrong, so can we.

3.6.1 The ABC Game

Sienna, Reg, Rae, and Roy are playing a signaling game. Sienna is the sender; Reg, Rae and Roy are receivers. Nature chooses one of three states—A, B, C—at random. Sienna observes the state. She sends one of three messages, ! (bang), & (amp), # (pound), to the receivers (the same message to each). Then the receivers independently guess the state. If at least one guesses correctly, they all get positive payoff; else, nothing.

They play the game day after day. Eventually, they coordinate. Their strategies are represented in Table 1. For example: Sienna sends ! when A, & when B, # when C; Reg guesses A given !, C given &, B given #. Each receiver guesses correctly given one of the messages (Reg given !, Rae given &, Roy given #) but incorrectly given the other two. Since someone guesses correctly no matter what the state, everyone always gets positive payoff. Each is happy with her own strategy and confident of the others'. The messages have acquired meanings: ! means A, & means B, # means C.

No receiver knows what Sienna is doing. Still, they may have beliefs about it, and corresponding beliefs about the messages' meanings. Suppose each is cocky, believing he always guesses correctly. For example, Reg believes Sienna sends ! when A, # when B, & when C, and so believes that ! means A, # means C, & means B. And similarly for Rae and Roy. (See Section 7.2 for further discussion.)

Sienna knows the meanings: she knows her own strategy and the meanings are determined by that. (See Section 7.3 for further discussion.) The receivers don't.

Reg is right about the meaning of ! and wrong about the meanings of & and #. And similarly for Rae and Roy. All receivers are wrong about the meanings of two messages. For each message, a majority of receivers (two of three) are wrong about the message's meaning.

Since there are as many receivers as states, if the receivers guess differently, one of them is bound to guess correctly. If any receiver guesses correctly, all get the preferred payoff. So, if the receivers could confer among themselves, they could make sure to guess differently given each message, and so ensure they get the preferred payoff every time, regardless of Sender's strategy. If that were what happened, perhaps the messages wouldn't be meaningful. But that's not what happens. The receivers don't confer among themselves. The fact (if it is a fact) that if they were to confer the messages wouldn't be meaningful doesn't undermine the claim that the message are meaningful, given that the receivers don't confer.

3.6.2 The Two-Sender Coin Game

Nature flips two coins. Sender 1 sees how the first coin landed and sends a message, m_1 or m_2, to Receiver; Sender 2 sees how the second coin landed and sends a message, m_3 or m_4, to Receiver. The senders act independently. After receiving the messages, Receiver guesses how each coin landed. If she gets both right, everyone gets positive payoff; else, nothing.

They play day after day. Eventually, they coordinate: Sender 1 sends m_1 when heads, m_2 when tails; Sender 2 sends m_3 when heads, m_4 when tails; and Receiver's strategy complements theirs, so she always guesses correctly. The messages have acquired meanings: m_1 and m_3 mean *the coin landed heads*, m_2 and m_4 mean *the coin landed tails*.[15]

The structure of the game is common knowledge. Each player knows her strategy and payoffs. That's enough for Receiver to work out the senders' strategies. And it's enough for each sender to work out how Receiver responds to his messages. But

[15] Or perhaps m_1 means *the first coin landed heads* and m_2 means *the second coin landed heads*, and so on. We needn't decide the matter here.

it's not enough for either sender to work out the other sender's strategy, nor how Receiver responds to the other sender's messages. For all Sender 1 knows, Sender 2 might send m_4 when heads and m_3 when tails, and Receiver guess heads given m_4 and tails given m_3. And similarly for Sender 2.

Still, each sender may have beliefs about the other sender's strategy, and corresponding beliefs about the messages' meanings. Suppose each sender flips the other's strategy, so Sender 1 believes that m_3 means tails and m_4 means heads, and Sender believes that m_1 means tails and m_2 means heads. Each is wrong about the meanings of the other's messages.

3.6.3 The Coin Game with Nature

Sender and Receiver are playing a signaling game, but they're not sure which. Nature chooses the West Game or the East Game at random. The game, once chosen, is fixed. Then Sender and Receiver play that signaling game repeatedly.

In either game, Nature flips a coin and Sender sees the result. Sender sends a message, # or &, to Receiver, who then guesses how the coin landed. In the East Game, the players are rewarded if Receiver guesses correctly; in the West Game, the players are rewarded if Receiver guesses incorrectly.[16]

	guess H	guess T
H	0	1
T	1	0

West Game

	guess H	guess T
H	1	0
T	0	1

East Game

Figure 3-5: State-response correspondences in the Coin Game with Nature

As it happens, Nature chooses the East Game. Sender and Receiver play day after day. Eventually they coordinate: Sender sends # when heads and & when tails; Receiver guesses heads given # and tails given &. Each is happy with her strategy and confident of the other's. The messages have acquired meanings: # means *the coin landed heads* and & means *the coin landed tails.*

[16] Note that the matrices don't represent the games in strategic form; rather, they represent the payoffs for each state-act pair, from which the strategic forms may be derived, given a probability distribution over the states.

As before, neither player knows the other's strategy. Still, they may have beliefs about it, and corresponding beliefs about the meanings. Suppose they believe Nature chose the West Game, so Sender believes Receiver guesses tails given # and heads given & and Receiver believes Sender sends # when tails and & when heads. In short: each player is wrong about the game and flips the other's strategy.

Sender, despite his mistake, knows that # means heads and & means tails, since he knows his own strategy. Receiver is not so lucky: she believes that # means tails and & means heads.

3.6.4 The Signaling Cycle Game

Nature chooses one of three states (s_1, s_2, or s_3) at random. Sender observes the state and sends a message (m_1, m_2, or m_3) to Receiver. Then Receiver chooses a response (r_1, r_2, or r_3). The payoffs for each state-response pair are given below:

	r_1	r_2	r_3
s_1	1	1	0
s_2	0	1	1
s_3	1	0	1

Figure 3-6: State-response correspondence in the Signaling Cycle Game

Sender and Receiver play day after day. Eventually they coordinate: Sender sends m_2 when s_1, m_3 when s_2, m_1 when s_3; Receiver does r_1 given m_1, r_2 given m_2, r_3 given m_3. Each is happy with her strategy and confident of the other's.

Have the messages acquired meanings? Unlike the previous examples, for each state there are two responses which yield the preferred payoff. Therefore each player could unilaterally change her strategy without losing out. Perhaps that means the messages don't acquire meanings. If so, strike out the example and jump to the next section. If not, I say that m_1 means s_3, m_2 means s_1, m_3 means s_2.

As above, neither player knows what the other is doing, but they may have beliefs about it, and corresponding beliefs about the meanings. Suppose each player permutes the other's strategy. Sender, despite his mistake, knows that m_1 means s_3, m_2 means s_1, m_3 means s_2, since he knows his own strategy. Receiver is not so lucky:

she believes that m_1 means s_1, m_2 means s_2, m_3 means s_3. She is wrong about the meaning of every message.

3.7 Discussion of examples

3.7.1 Why go beyond basic signaling games?

In a *basic signaling game*, like the Coin Game, there are two players: Sender and Receiver. One of n possible states, $s_1, \ldots, s_n$, is chosen by Nature with equal probability. Sender observes the state; Receiver doesn't. Sender sends one of n possible messages, $m_1, \ldots, m_n$, to Receiver, who then chooses one of n responses, $r_1, \ldots, r_n$. The payoffs depend on the state and the response. If Receiver does r_i in s_i, Sender and Receiver each get equal positive payoff. Otherwise, each gets nothing.[17] A strategy for Sender is a function from states to messages. A strategy for Receiver is a function from messages to responses. Both players want Receiver to guess correctly. They must attempt this indirectly, by coordinating their strategies.

The four examples in Section 6 go beyond basic signaling games in all sorts of ways: in the ABC Game, there is one sender and several receivers, only one of whom has to guess right; in the Two-Sender Coin Game, there are two senders and one receiver, who has to guess how both coins landed; in the Coin Game with Nature, the players are unsure about the state-response correspondence; the Signaling Cycle Game relaxes the payoff structure.

Why go beyond basic signaling games? Because basic signaling games leave little room for mistakes about meanings. Take the Coin Game, a basic signaling game with two states, messages and responses. The messages acquired meanings. You know your strategy and payoffs. That is enough to work out my strategy. I know my strategy and payoffs. That is enough to work out your strategy. Since each of us can work out the other's strategy, each of us can work out what the messages mean. The Coin

[17] Isn't it obvious what the players should do, namely, send m_i in s_i and do r_i given s_i? No. That confuses a property of our representation (how we label the states, messages and responses) with a property of what we're representing.

Game doesn't leave room for mistakes about meanings.

Now consider a larger basic signaling game, say with ten states, messages and responses. Here's one way things might turn out. Sender and Receiver play day after day, learning about each other as they go. They eventually settle on strategies, each happy with her own and confident of the other's. In states $s_1, \ldots, s_8$ their strategies match up: Sender sends m_i in s_i and Receiver does r_i given m_i. These messages have acquired meanings: m_i means s_i $(i = 1, \ldots, 8)$. In states s_9 and s_{10}, their strategies don't match up: Sender sends m_9 in s_9 and m_{10} in s_{10}, but Receiver does r_{10} given m_9 and r_9 given m_{10}. By good luck, Nature doesn't choose s_9 or s_{10}. The players believe, wrongly but with good reason, that their strategies would match up no matter the state.

It's not clear whether m_9 and m_{10} are meaningful. If they are meaningful, then m_9 means s_9 and m_{10} means s_{10} (although see Section 7.5). Whether meaningful or not, we may suppose that Receiver believes, wrongly, that m_9 means s_{10} and m_{10} means s_9. Basic signaling games do, thus, leave room for mistakes about meanings.

I don't rely on examples like this. The example depends on an unlikely event (that Nature doesn't choose two of the states). More importantly, the players' beliefs are not robust: with probability 1, eventually Nature will choose s_9 or s_{10}, and then the players will correct their mistakes. If they're mistaken about meanings, they won't be for long. The examples in Section 6, by contrast, don't depend on unlikely events, nor need the players ever realize their mistakes.

When conditions are strange enough (the players are fantastically unlucky, or selectively forgetful, or dazed and confused, or philosophers), no doubt they can be mistaken about meanings. Examples like that aren't interesting. The aim is not just to find signaling games which leave room for mistakes about meanings, but to find games which leave room for mistakes about meanings in a simple, straightforward way. You don't need outlandish set-ups to be mistaken about meanings. The examples in Section 6 are artificial, in order to make the structure clear. They are not fussy; they are not exotic.

3.7.2 Belief about regularities and belief about meanings

Consider, say, Reg in the ABC Game. Distinguish two claims: (a) Reg believes Sienna sends ! when A, # when B, & when C; (b) Reg believes that ! means A, # means B, & means C. I've glossed over the difference. But in fact the claims are independent: you might have (a) without (b), if Reg (mistakenly) believes that messages in a signaling game can't acquire meanings; you might have (b) without (a), if Reg (mistakenly) believes that meanings don't require a regularity.

I've assumed that by the time the players coordinate, each happy with her own strategy and confident of the others', if a player believes that a sender sends message m in state s, she also believes that m means s. The assumption isn't true in general. All I claim is that the four examples might turn out that way. The examples leave room for mistakes about meanings; they don't force mistakes.

3.7.3 Sender is not wrong about meanings

The meaning of a message, given that it's meaningful, is determined by Sender's strategy. Look back at the examples: the meaning of a message, given that it's meaningful, is the state in which Sender sends that message.

Suppose m means s. Since m is meaningful, the players coordinate, each happy with her own strategy and confident of the others'. Since m means s, Sender sends m whenever the state is s. Sender knows his strategy. Hence Sender knows that he sends m in s. Given the assumption stated in the previous subsection, it follows that Sender believes that m means s. Generalizing, Sender is right about the meanings of his own messages, if they're meaningful at all. The key idea—that knowing your own strategy and payoff needn't determine what the others do—doesn't leave room for Sender to be wrong about the meanings of his own messages.

When there are multiple Senders, each Sender may be wrong about the meanings of the *other* Senders' messages, as in the Two-Sender Coin Game. But each Sender still knows the meanings of his own messages. That is a limitation of the examples.

3.7.4 Switching roles

In all the examples, the players' roles are fixed. In particular, senders never become receivers and receivers never become senders. In actual languages, of course, things aren't like that. People sometimes speak and sometimes listen. No problems arise in actual languages from switching roles. The conventions of meaning in actual languages are robust to role-switches.

Is the same true of the conventions of meaning in the examples? The short answer is: typically not. In each case I assume, naturally enough, that a player behaves in a new role as she believes the player she is taking over from behaved. Now, take the Signaling Cycle Game. Suppose Sender and Receiver swap roles. Then Receiver, now in Sender's role, will send m_1 when s_1, m_2 when s_2 and m_3 when s_3. And Sender, now in Receiver's role, will do r_3 given m_1, r_1 given m_2 and r_2 given m_3. The result? They'll get zero payoff no matter the state. Swapping roles leads them to anti-coordinate. The situation is similar for the Two-Sender Coin Game and the ABC Game.[18] As it happens, in the Coin Game with Nature, the players can switch roles without any problem arising. But typically the conventions of meaning in the examples are not robust to role-switches. That is another limitation of the examples.

Still, it's worth noting that the coordination game examples from Section 3 are more robust to role-switches. Take the Three-Player Game. If any *two* players switch roles (Colin becoming the matrix-chooser and Mattea becoming the column-chooser, say) then no problems will arise. But if all *three* players switch roles, they'll get zero payoff. Similarly, the Five-Player Game is robust to many, but not all, role-switches. The Nature Game is robust to the two players' switching roles. The Cycle Game isn't.

The conventions in the coordination games are often, but not always, robust to role-switches, even when the players are wrong about the convention. That is evidence—even if only weak evidence—that conventions of meaning can be robust to

[18] For the ABC Game, things are not completely straightforward, because we don't know what Sienna will do as a receiver. But if we suppose that she behaves the same way no matter which receiver she swaps with, then things will go wrong.

role-switches too, even when the players are wrong about the meanings.

3.7.5 Indicative meanings and imperative meanings

Recall the Coin Game: you send # when heads and & when tails; I guess heads given # and tails given &. We coordinate day after day, each happy with her own strategy and confident of the other's. I said that # means *the coin landed heads* and & means *the coin landed tails.* These are the *indicative* meanings of the messages.

As Lewis pointed out, there's an alternative interpretation: # means *guess heads* and & means *guess tails.* These are *imperative* meanings of the messages. An indicative meaning gives information about the state. An imperative meaning gives an instruction about the response.

You might take the imperative meaning of a message to be determined by Sender's strategy, just like the indicative meaning: the imperative meaning of m is to make the appropriate response, whichever it is, to the state in which Sender sends m. Or you might take the imperative meaning to be determined by Receiver's strategy: the imperative meaning of m is to respond however Receiver does respond given m. In basic signaling games, these coincide; in more complicated games, they may not. Taking the meanings to be indicative, Sender is in a privileged position; taking the meanings to be imperative, Receiver is in a privileged position.

When should we interpret messages as indicative and when as imperative? It's not clear. In these simple settings, either interpretation may be acceptable.[19] If we interpret the messages one way, we may get one pattern of mistakes about the meanings; if we interpret them the other way, we may get another.

The distinction between indicative and imperative meanings might help leave room for Sender to be wrong about the meanings of his own messages, despite knowing his own strategy, for he could be wrong about the *imperative* meanings, as determined by *Receiver's* strategy. But better to find examples which don't rely on choosing between indicative and imperative meanings. And better to find meanings which every player

[19] Lewis suggests when we should interpret them as indicatives and when as imperatives. See pp. 143–7. See also Millikan (1995).

is wrong about (the same meanings for all), rather than find, for each player, a type of meaning which that player is wrong about.

3.8 Conclusions

People can be wrong about their own conventions; in particular, people can be wrong about the meanings of their own messages. The examples are simple, explicit, new in kind and based on an independently plausible meta-semantic story. They're artificial, in order to make the structure clear, but not fussy and not exotic. If you want to observe mistakes about conventions or meanings first-hand, then grab some friends, feed them a little misleading information, and let them play the games in the paper.

Imagine a large community of signalers who are radically wrong about the meanings of their messages. Send a field linguist among them to discover the semantics of their language using standard techniques like the Elicitation Method. The linguist will be radically misled.

Suppose Method X is taken to be a reliable test for Disease D. It's then discovered that Method X is unreliable when the subject has Condition C. Condition C is a common-or-garden condition, not involving genetic quirks or strange diets or radiation exposure. How should we react? I say: we shouldn't endorse Method X as confidently as before in cases where Condition C doesn't obtain; we should re-examine the reliability of Method X even when Condition C doesn't obtain. As with Method X and Disease D, so too with the Elicitation Method and meanings.

I don't say that we should give up the Elicitation Method; nor that the key idea behind the examples is responsible for mistakes about meanings in natural languages. I do say that if speakers can be wrong about meanings for such straightforward reasons, then we should reconsider the reliability of speakers' judgments about meanings more generally.

Bibliography

[1] Scott Aaronson. Why Philosophers Should Care About Computational Complexity. In B. Jack Copeland, Carl J. Posy, and Oron Shagrir, editors, *Computability: Turing, Gödel, Church, and Beyond*. The MIT Press, 2013.

[2] Daniel Altshuler, Terence Parsons, and Roger Schwarzschild. *A Course in Semantics*. The MIT Press, Cambridge, MA, September 2019.

[3] Chrisoula Andreou. Dynamic Choice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[4] Robert Aumann. Agreeing to Disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.

[5] Robert Aumann. Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, 55(1):1–18, 1987.

[6] Robert Aumann. Nash Equilibria Are Not Self-Enforcing. In J. J. Gabszewicz, J.-F. Richard, and L. Wolsey, editors, *Economic Decision Making: Games, Econometrics, and Optimisation: Essays in Honor of Jacques Dreze*, pages 201–206. Elsevier Science Publishers, Amsterdam, 1990.

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

[8] Robert Aumann. Game Theory. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008.

[9] Robert Aumann and Adam Brandenburger. Epistemic conditions for Nash Equilibrium. *Econometrica*, 63(5):1161–1180, 1995.

[10] Christian W. Bach and Elias Tsakas. Pairwise epistemic conditions for Nash equilibrium. *Games and Economic Behaviour*, 85(C):48–59, 2014.

[11] Paulo Barelli. Consistency of beliefs and epistemic conditions for Nash and correlated equilibria. *Games and Economic Behaviour*, 67(2):363–375, 2009.

[12] Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, Belmont, Massachusetts, 2nd edition, July 2008.

[13] Ken Binmore. *Playing for Real: Game Theory*. Oxford University Press, February 2007.

[14] M. Ryan Bochnak and Lisa Matthewson, editors. *Methodologies in Semantic Fieldwork*. Oxford University Press, 2015.

[15] Rachael Briggs. Normative Theories of Rational Choice: Expected Utility. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[16] Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, July 2019.

[17] Tyler Burge. Individualism and the Mental. *Midwest Studies in Philosophy*, 4(1):73–122, 1979.

[18] Bill Chen and Jerrod Ankenman. *The Mathematics of Poker*. Conjelco, Pittsburgh, PA, November 2006.

[19] Christopher Cox, Jessica De Silva, Philip Deorsey, Franklin H. J. Kenter, Troy Retter, and Josh Tobin. How to Make the Perfect Fireworks Display: Two Strategies for Hanabi. *Mathematics Magazine*, 88(5):323–336, 2015.

[20] Kenny Easwaran. Strong and Weak Expectations. *Mind*, 117(467):633–641, 2008.

[21] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 11th edition, 1950.

[22] Ronald A. Fisher. *Statistical Methods and Scientific Inference*, volume 9. Oliver and Boyd, 1956.

[23] Margaret Gilbert. Game Theory And Convention. *Synthese*, 46(1):41–93, 1981.

[24] Margaret Gilbert. Agreements, Conventions, and Language. *Synthese*, 54(3):375–407, 1983.

[25] Margaret Gilbert. Rationality, Coordination, and Convention. *Synthese*, 84(1):1–21, 1990.

[26] Margaret Gilbert. Walking Together: A Paradigmatic Social Phenomenon. *Midwest Studies in Philosophy*, 15(1):1–14, 1990.

[27] Margaret Gilbert. *Social Convention Revisited*. Oxford University Press, November 2013.

[28] Ian Hacking. *An Introduction to Probability and Inductive Logic*. Cambridge University Press, 2001.

[29] Joseph Y. Halpern. Substantive Rationality and Backward Induction. *Games and Economic Behavior*, 37(2):425–435, November 2001.

[30] John Harsanyi. Rejoinder to Professors Kadane and Larkey. *Management Science*, 28(2):124–125, 1982.

[31] John Harsanyi. Subjective Probability and the Theory of Games: Comments on Kadane and Larkey's Paper. *Management Science*, 28(2):120–124, 1982.

[32] Brian Hedden. Options and the Subjective Ought. *Philosophical Studies*, 158(2):343–360, 2012.

[33] Fred Hoyle. *Home is where the Wind Blows: Chapters from a Cosmologist's Life*. University Science Books, 1994.

[34] David Hume. *A Treatise of Human Nature*. Oxford Philosophical Texts. Oxford University Press, Oxford, New York, February 2000.

[35] Joseph B. Kadane and Patrick D. Larkey. Reply to Professor Harsanyi. *Management Science*, 28(2):124–124, 1982.

[36] Joseph B. Kadane and Patrick D. Larkey. Subjective Probability and the Theory of Games. *Management Science*, 28(2):113–120, 1982.

[37] J. L. Kelly. A New Interpretation of Information Rate. *Bell System Technical Journal*, 35(4):917–926, 1956.

[38] Frank H. Knight. Risk, Uncertainty and Profit. SSRN Scholarly Paper ID 1496192, Social Science Research Network, Rochester, NY, 1921.

[39] Henry Allen Latané. Criteria for Choice Among Risky Ventures. *Journal of Political Economy*, 67(2):144–155, 1959.

[40] Henry Allen Latané. Rational Decision-Making in Portfolio Management. *The Journal of Finance*, 14(3):429–430, 1959.

[41] David Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

[42] David Lewis. Languages and Language. In Keith Gunderson, editor, *Minnesota Studies in the Philosophy of Science*, pages 3–35. University of Minnesota Press, 1975.

[43] Kevin Leyton-Brown and Yoav Shoham. Essentials of Game Theory: A Concise Multidisciplinary Introduction. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2(1):1–88, January 2008.

[44] Leonard C. MacLean, Edward O. Thorp, and William T. Ziemba, editors. *The Kelly Capital Growth Investment Criterion*, volume 3 of *World Scientific Handbook in Financial Economics*. World Scientific, 2011.

[45] Lisa Matthewson. On the Methodology of Semantic Fieldwork. *International Journal of American Linguistics*, 70(4):369–415, 2004.

[46] Edward F. McClennen. *Rationality and Dynamic Choice: Foundational Explorations.* Cambridge University Press, 1990.

[47] Ruth G. Millikan. Pushmi-Pullyu Representations. *Philosophical Perspectives*, 9:185–200, 1995.

[48] Roger Myerson. *Game Theory: Analysis of Conflict.* Harvard University Press, 1991.

[49] John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[50] Jerzy Neyman. *Lectures and conferences on mathematical statistics and probability.* Graduate School, US Department of Agriculture Washington, DC, 1952.

[51] Jerzy Neyman and Egon Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, February 1933.

[52] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory.* The MIT Press, Cambridge, Mass, July 1994.

[53] Eric Pacuit and Olivier Roy. Epistemic Foundations of Game Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2015.

[54] Andrés Perea. A one-person doxastic characterization of Nash strategies. *Synthese*, 158(2):251–271, 2007.

[55] Steven Pinker. *The Language Instinct.* Harper Perennial, 1995.

[56] Ben Polak. Epistemic Conditions for Nash Equilibrium, and Common Knowledge of Rationality. *Econometrica*, 67(3):673–676, 1999.

[57] Hilary Putnam. The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, 7:131–193, 1975.

[58] Agustín Rayo. *On the Brink of Paradox: Highlights from the Intersection of Philosophy and Mathematics.* MIT Press, April 2019.

[59] Paul A. Samuelson. The "Fallacy" of Maximizing the Geometric Mean in Long Sequences of Investing or Gambling. *Proceedings of the National Academy of Sciences of the United States of America*, 68(10):2493–2496, 1971.

[60] Brian Skyrms. *Signals: Evolution, Learning, and Information.* OUP Oxford, 2010.

[61] Robert Stalnaker. *Inquiry.* Cambridge University Press, 1984.

[62] Robert Stalnaker. On the Evaluation of Solution Concepts. *Theory and Decision*, 37(1):49–73, 1994.

[63] Robert Stalnaker. Knowledge, Belief and Counterfactual Reasoning in Games. *Economics and Philosophy*, 12(2):133–163, 1996.

[64] Robert Stalnaker. Belief Revision in Games: Forward and Backward Induction. *Mathematical Social Sciences*, 36(1):31–56, 1998.

[65] Robert Stalnaker. Extensive and Strategic Forms: Games and Models for Games. *Research in Economics*, 53(3):293–319, 1999.

[66] Michael Strevens. Notes on Bayesian Confirmation Theory. June 2017.

[67] Peter Vanderschraaf. Convention as Correlated Equilibrium. *Erkenntnis*, 42(1):65–87, 1995.

[68] Peter Vanderschraaf. Knowledge, Equilibrium and Convention. *Erkenntnis*, 49(3):337–369, 1998.

[69] Brian Weatherson. Games, Beliefs and Credences. *Philosophy and Phenomenological Research*, 92(2):209–236, 2016.

[70] Yoad Winter. *Elements of Formal Semantics: An Introduction to the Mathematical Theory of Meaning in Natural Language.* Edinburgh University Press, Edinburgh, July 2016.