Convention-Affirming Equilibrium

by

Aristos Hudson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor David Miller, Chair
Associate Professor Shaowei Ke
Professor Scott Page
Professor Tanya Rosenblat

Aristos Hudson

aahuds@umich.edu

ORCID iD: 0009-0001-3003-6958

Acknowledgments

I would like especially to thank my primary thesis advisor David Miller for frequent and consistently helpful discussions and suggestions throughout the development of these chapters. I would like to thank Scott Page for helpful discussions and comments on this material, and to thank the other members of my committee, Shaowei Ke and Tanya Rosenblat, for helpful feedback. I would like also to thank Tilman Börgers and Amanda Friedenberg for useful comments on earlier versions of some of this material. Finally, I would like to thank Jaron Cordero for many stimulating discussions on walks around Ann Arbor over the years, which provided support and inspiration throughout this process, in all seasons.

Table of Contents

List of Figures

Abstract

I introduce the concept of convention-affirming equilibrium (CAE) and study its implications in a number of settings. A convention-affirming equilibrium represents a stationary outcome in a population of agents who do not know the strategies of other agents, but who correctly understand the process they are embedded in and the reasoning and possible beliefs of other agents, and who observe a sample of the outcomes of play around them. In a convention-affirming equilibrium, all agents believe that they are in some possible convention-affirming equilibrium, but may be uncertain which is the true one. Because of the circularity in this, convention-affirming equilibrium is defined setwise, with sets of convention-affirming equilibria (CAE sets) being the primitive concept.

In the first chapter, I analyze the convention-affirming equilibria for a simple 2x2 coordination game. If agents have a common prior, the set of convention-affirming equilibria coincides with the set of Nash equilibria. If agents may think others have a different prior than they do, there is a larger set of convention-affirming equilibria, which I characterize.

In the second chapter, I study the convention-affirming equilibria of a simple, two-stage game of effort choice on a joint project. I focus on cases where most agents play according to one of the four pure symmetric paths of play, and show that each such case can be a convention-affirming equilibrium for some values of the payoff parameters. The results highlight the importance of strategic uncertainty about the outcome of deviation, as distinct from a known punishment path, as a reason why agents may conform to cooperative behavior

widespread around them. They also highlight how the convention-affirming equilibrium concept can give rise to forward induction-like reasoning in a way that arises naturally and from first principles.

In the third chapter, I apply the convention-affirming equilibrium concept to a simple model of matching and bargaining between two populations. I characterize a natural solution set in which all matches result in immediate agreement on either the high wage or the low wage. In the high-wage case, all agents are able to infer each other's equilibrium strategies from the available information. In the low-wage case, their equilibrium behavior instead reflects their subjective confidence in the demands they can successfully make, which is prior-dependent due to limited information. The results thus highlight how both strong strategic inferences and subjective confidence (or lack thereof) in the absence of such inferences can support convention-affirming equilibrium bargaining outcomes, and clarifies which outcomes can be supported by one or the other.

Introduction

In the three chapters that follow, I introduce the concept of convention-affirming equilibrium and then apply it to two economic applications of interest. The first chapter characterizes the behavior of the concept in simple 2×2 games. The second applies the concept to a simple game of joint effort and seeks to understand the emergence (or lack thereof) of cooperative behavior. The third applies it to a simple matching-and-bargaining framework and studies the outcomes that can arise there.

One primary contribution of the present work is the introduction of this solution concept itself, along with the illustration of the kind of outcomes it can give rise to provided by the applications. A second is the results for the applications themselves, which are of interest in their own right and may illuminate the further insights that may be gained by the study of further applications.

Chapter 1

Convention-Affirming Equilibrium

Abstract

I introduce the concept of convention-affirming equilibrium, and study its implications in 2×2 games. In a convention-affirming equilibrium, all agents in a population believe they are in some convention-affirming equilibrium, and the choices they make on the basis of such beliefs in fact lead to one. In the case where agents have a common prior, convention-affirming equilibrium coincides with Nash equilibrium. When they may think other agents have different priors than they do, non-Nash outcomes are possible, within limits.

1.1 Introduction

We often interpret equilibrium concepts as representing conventions that might prevail in some society. In games with multiple equilibria, this interpretation naturally extends to thinking about multiple possible conventions which might prevail in different societies. The presumption is that behavior in any given society will be governed by some convention or other, a view which casual observation tends to support. The question is which patterns of behavior are likely to arise as conventions, and which conditions favor the emergence of one over another.

It is reasonable to expect that people entering some society will share this same presumption. Some convention or other already prevails, and one's task is to figure out which it is by observing the behavior of others. People who understand that the prevailing convention has emerged through the choices of other people who faced the same situation they do may plausibly also reason like game theorists, attempting to identify plausible candidates for the actual convention by reasoning about which hypothetical conventions could plausibly emerge in a world where everyone reasons and behaves in this way.

I introduce the concept of *convention-affirming equilibrium* to capture these ideas. In a convention-affirming equilibrium (CAE), all agents in a population believe the population plays according to some convention-affirming equilibrium (though they do not know which), and choose their own actions to maximize their expected payoff given this belief and some observations of the actions chosen by others in the population, which they use to Bayes update their prior about which convention-affirming equilibrium they are in. As this description would suggest, CAE is defined circularly and setwise; a distribution of play is a CAE if there is a set of distributions containing it all of whose elements can be equilibrium conventions for a population that views it as the set of possible conventions.

The definition of convention-affirming equilibrium differs from that of Nash equilibrium in that it does not require that agents best respond to the actual distribution of play. On the other hand, it builds in both a correct understanding of the payoffs and reasoning of others comparable to rationalizability and a presumption of being in some (convention-affirming) equilibrium compatible with them. It thus occupies a middle ground between strategic sophistication with no further structure and equilibrium compatible with best response to the true situation. The ability of agents to observe what others are doing lends predictive power to the concept, by allowing them to react not just to the empirical distribution, but to facilitate potentially stronger inferences about 'which (convention-affirming) equilibrium they are in'.

The assumption that agents are informed about the state of play in their society by seeing the actions and outcomes of others is a plausible one in many settings. Many games are played in public view, or at least in such a way that people will learn about their typical outcomes through gossip, reading, education, and so on. It is of interest to see both whether 'misunderstandings' can persist even for large numbers of such observations, and whether even small numbers suffice to strongly restrict outcomes when combined with the interactive reasoning aspect of the concept. As we will see, both can happen, in a certain sense and under the right conditions.

Alternative sources of empirical information could also be considered, such as learning from one's own past matches, actively choosing how much information to gather, or observing a sample that is not representative in some way. These alternatives are plausible but would add technical complexity; I focus on the simplest case of passive observation of others to more easily isolate the effect of the reasoning about potential equilibria that is the main contribution of this concept.

Much of the value of this concept lies in the way it allows us to consider the joint effects of learning from observation, strategic reasoning about the beliefs and motives of others, and the presumption of a world in equilibrium that is plausible in many contexts, without imposing further prior knowledge about which equilibrium has been realized in a way that is hard to justify. Agents are still assumed to have a correct understanding of the overall process, but this is more plausible in many settings than assuming a correct understanding of the particular outcome the process has led to. This strikes a balance between concepts where agents know more than is often plausible (as in Nash equilibrium and its refinements), and those where their ability to engage in strategic reasoning or understand the nature of the process they are engaged in is artificially curtailed (as in models of simple learning rules in games), and arguably corresponds more closely to the informal game-theoretic intuitions we are inclined toward in many real-world settings than either.

I study the consequences of the CAE concept primarily in 2×2 coordination games, the simplest context in which interesting questions arise about whether this concept selects for a strict subset of multiple Nash equilibria, and whether some non-Nash outcomes can also be CAE (in Section 1.6 I consider also some modest extensions). I study two different versions of the CAE concept: 'unitary' CAE, in which agents have a common prior over which CAE is the true one, and 'diverse' CAE, in which each agent thinks the other agents in the population may have a different prior than they do, and are uncertain about which one it is. These lead to starkly different predictions.

In Section 1.4, I characterize the set of unitary CAE in 2×2 coordination games, and show it coincides with the set of Nash equilibria. The reason is structural: if agents are best responding, the frequency with which each action is played will change faster than one-to-one with the probability of observing each action, and there can thus be at most one interior CAE if agents have a common prior, which can only be the mixed Nash.

In Section 1.5 I show that the diverse CAE concept with unrestricted beliefs allows any outcome to be a CAE. With mild restrictions on beliefs, which essentially require agents not to ignore their observations completely, the set of possible outcomes narrows, with different possibilities depending on the sample size and the minimal responsiveness of allowed priors to data.

It is desirable that a set of CAE is also *complete*: not only is each playable as an equilibrium convention given a belief in the set, but all outcomes playable in that sense are included. If a set of CAE is incomplete, non-CAE play is possible for agents who believe in it, so we cannot be as confident in it as a prediction. I will show that complete CAE sets exist, for some specifications of the priors agents are allowed to have.

1.1.1 Related Literature

The notion of equilibrium developed here is related to the sampling equilibrium of Osborne and Rubinstein (2003) and the more general 'sampling equilibrium with statistical inference' concept of Salant and Cherry (2020). As with these concepts, my CAE concept involves agents who observe a random sample of the actions of others, and a notion of equilibrium in which the distribution of actions is reproduced by the distribution of choices they make after processing this sample.

The major difference is that I study agents who reason explicitly about which possible equilibria they might be in, and ground this reasoning in strategic thinking about the information, payoffs, and choice process of others, while the preceding papers focus on a notion of equilibrium in solely data-driven decision processes without this *a priori* strategic reasoning, and compare alternative ways agents may process their data. The focus in my context is thus on which sets of candidate equilibria can survive the restrictions imposed by agents' reasoning about one another's reasoning, with the observation of samples being one of the factors that generates these restrictions, and not on properties of or comparisons between alternative observation structures or inference procedures.

Conceptual questions very similar to those of the present work are studied by Young (1993a), as well as the substantial subsequent literature on equilibria selected by population learning processes with a permanent stochastic component (such as Kandori et al. (1993), which covers much of the same ground as Young, Ellison (1993), which studies the effect of different interaction structures, and Bergin and Lipman (1996) and Binmore and Samuelson (1997), which consider the effects of alternative specifications of the random component). As in the present work, the focus is on which equilibria will emerge as conventions in a population in which agents observe a sample of past play. It differs in that agents mechanically best-respond to the sample average, rather than reasoning in a fully Bayesian way about the process and each other. The focus in this literature is also on studying the ergodic distribution of an explicit dynamic, while I formulate a static equilibrium concept.

Models of observational learning in which agents learn about the value of choices available to them by observing or talking to others, either by sophisticated inferences (e.g. Banerjee (1992), Bikhchandani et al. (1992)), or by following simple heuristics (e.g. Ellison and Fudenberg (1993), Ellison and Fudenberg (1995)), also consider a setting in which agents make choices after observing others in the population; in these models, they observe something about the frequency with which other agents take different actions, the payoffs they receive from certain actions, or both. While agents are influenced by the choices of others in these settings, the focus is generally on whether or not agents will end up adopting the 'right' action, according to some exogenously given activity; this is in contrast to the present setting, where there is direct interdependence of payoffs because agents face each other in a game.

The convention-affirming equilibrium concept can also be compared to the concept of self-confirming equilibrium (Fudenberg and Levine, 1993), especially the variants with a 'rationalizability' component (Dekel et al. (1999), Fudenberg and Kamada (2015), Fudenberg and Kamada (2018)). Both describe a notion of equilibrium among agents who only partly observe the play of others, and the latter variants also share with the present context the attempt to combine this with a priori restrictions in the spirit of rationalizability.

One difference between the two is that self-confirming equilibrium describes a setting in which agents observe the outcomes of their own strategy, whereas in a convention-affirming equilibrium agents observe the outcomes of strategies commonly played by others. Another is that convention-affirming equilibrium allows degrees of frequent or infrequent observation, whereas in a self-confirming equilibrium everything is either fully known or not observed.[1]

[1]These first two differences also explain why self-confirming equilibrium is an extensive-form concept (knowing the outcomes of your own strategy is not enough to know the whole distribution of play only if your choices prevent you from observing part of your opponent's strategy), while convention-affirming

For the rationalizable variants, there is the additional difference that in these papers the rationalizability-style requirement is imposed at the level of strategies in the game itself, whereas in a convention-affirming equilibrium it is imposed on the larger process mapping observations to action choices from which the actions chosen in the game are derived.

Finally, the convention-affirming equilibrium concept connects to two broader literatures – the literature on conventions in philosophy, and the literature on culture in political economy.

Lewis (1969) proposes an analysis of 'convention' according to which a convention is an equilibrium of a coordination game – one of multiple ways in which a group of agents might accomplish some goal that is in all of their interests – which has become commonly known to be the typical way of accomplishing this, and thus self-sustaining (see also Vanderschraaf (2001), which provides both an alternative analysis of 'convention' as a kind of correlated equilibrium, and an explicit evolutionary game theory model of how conventions might arise). This notion of 'convention' – picking out a certain kind of equilibrium in a certain kind of game – is narrower and more specialized than the sense I have in mind in 'convention-affirming equilibrium', where agents observe which outcomes are 'conventional' in the weaker, colloquial sense of 'widespread'. An equilibrium 'convention', in my sense, is thus a near-synonym for 'equilibrium' (in the game-theoretic sense) in general – some outcome in a strategic situation which is stable once established – and need not refer to the particular sort of equilibrium considered in the literature following Lewis (1969). The CAE concept could possibly be useful in studying 'conventions' in the latter sense, as well, though I have not as of this writing made a serious attempt to explore this possibility.

Bednar and Page (2007) define 'cultural behavior' as behavior that is consistent across activities and across individuals within a given culture, different across cultures, persistent even in the face of changing circumstances, and potentially suboptimal. Other attempts

equilibrium is nontrivial even in normal-form games.

in the literature to understand cultural behavior include Greif (1994), which contrasts the historical evolution of institutions across different cultures, Fearon and Laitin (1996), which studies the strategies different ethnic groups may develop to avoid conflict with each other, and Henrich et al. (2001), which studies the differences in play of experimental games of subject groups from different cultures. See also Alesina and Giuliano (2015) for a survey of the literature on this topic.

There is plausibly potential for the CAE concept to connect and contribute to the study of questions of this kind, as it offers a general, flexible, fully formal way of modeling the possible range of 'typical' behaviors which might perist when agents observe them and reason about how they will influence the expectations and thence choices of others. But, any such attempt would require a larger model capable of expressing ideas about e.g. different cultural groups, changing circumstances, perceived links across games and choice situations, etc. Accordingly such connections belong for now to the realm of future research.

1.2 Model

1.2.1 Action Distributions

Fix a symmetric, normal form game with two players (i and j) and finite set of actions A. An *action distribution* in this game is an probability measure $p \in \Delta A$ which assigns a probability to the event that each action $a \in A$ is played. Write p_a for the probability of a under p.

I take each action distribution to represent an outcome of play in the game, in a setting where agents drawn at random from the same population are matched to play each other. I assume that each player chooses each action a with the probability p_a, and that these choices are independent, so that the probability of a given action profile (a_i, a_j) being the outcome of the game is just the product $p_{a_i} \cdot p_{a_j}$.

The convention-affirming equilibrium concept will select a subset of action distributions as predicted equilibria. Note that more familiar equilibrium notions can also be written this way: we can identify a symmetric pure Nash equilibrium with the action distribution placing probability one on the equilibrium action, or a symmetric mixed Nash equilibrium with the action distribution assigning the same probability to each action as does the mixed action it prescribes for both players.

1.2.2 Samples and Sample Distributions

Fix an integer N, assumed to be not too small. Each agent, before choosing their action, will observe the realized action profiles of N matches between players whose actions are drawn from the true action distribution p. Since actions are chosen independently, and agents understand this, they see as informative only the frequency with which each action is chosen, and thus view what they observe as a sample of $2N$ action choices, as follows.

A *sample* is a tuple $s = (s_a)_{a \in A}$, where each $s_a \in \{0, 1, ...2N\}$ represents the number of times action a was observed and $\sum_{a \in A} s_a = 2N$, so that each sample contains exactly $2N$ observations in total. Let S denote the collection of all possible samples.

Every action distribution p induces a *sample distribution* $f_p \in \Delta S$, which describes the probability of observing each possible sample in a population playing according to that action distribution. Since each sample contains $2N$ iid draws from the fixed distribution $p \in \Delta A$, the sample distribution is multinomial, with probability mass function

$$f_p(s) = \frac{(2N)!}{\prod_{a \in A} s_a!} \prod_{a \in A} p_a^{s_a}$$

1.2.3 Reaction Functions and Fixed Points

A *reaction function* is a map $r : S \to \Delta A$. That is, a reaction function specifies, for each sample, the mixed action a player will choose after observing that sample.[2] We will be interested in action distributions which are reproduced by a given reaction function; that is, where the distribution of play resulting from applying the reaction function r to the distribution over samples f_p reproduces the original action distribution p. Thus:

Definition 1 (Fixed Point)**.** *An action distribution $p \in \Delta A$ is a fixed point for reaction function r if for each $a \in A$,*

$$p_a = \Sigma_{s \in S} f_p(s) \cdot r(s)(a)$$

where $r(s)(a)$ is the probability assigned to a by the mixed action r prescribes at s.

1.3 Convention-Affirming Equilibrium

1.3.1 Beliefs and Best Responses

I now consider reaction functions describing the behavior of agents who use their sample to Bayes update some prior over possible action distributions, and then choose an action to maximize the expected payoff given their posterior belief. The reaction functions that can be thus chosen will be called best responses to a given prior belief, and are the building blocks of the equilibrium concept to follow.

Fix a non-empty, finite subset $\mathcal{P}$ of ΔA – the set of action distributions which agents consider possible – and consider a prior $\mu \in \Delta\Delta A$ which has support $\mathcal{P}$.[3] For each sample

[2]In cases where many players behave according to the same reaction function, we have a choice of interpretations. We can think of individual players choosing mixed actions, or of their choosing pure actions in a way that 'adds up' to the mixed frequencies prescribed.

[3]That is, which assigns positive probability to all and only those points in $\mathcal{P}$, and zero probability to the complement of $\mathcal{P}$ in ΔA.

s which has positive probability under f_p for some $p \in \mathcal{P}$, the posterior belief μ_s of an agent with prior μ who observes s assigns to each $p \in \mathcal{P}$ the probability:

$$\mu_s(p) = \frac{\mu(p) f_p(s)}{\sum_{p' \in \mathcal{P}} \mu(p') f_{p'}(s)}$$

Definition 2 (Best Response to μ)**.** *Reaction function r is a best response to prior μ if for each sample s with positive probability under f_p for some $p \in \mathcal{P}$ and for each $a_s \in A$ assigned positive probability by the mixed action $r(s)$, we have*

$$a_s \in \arg\max_{a_i \in A} \sum_{a_j \in A} \left(u(a_i, a_j) \sum_{p \in \mathcal{P}} p_{a_j} \mu_s(p) \right)$$

where $u(a_i, a_j)$ is the payoff to player i of the action profile (a_i, a_j), and the argmax is thus taken over expected payoffs of each possible action a_i for player i, given the posterior μ_s.

1.3.2 Convention-Affirming Equilibrium

In a convention-affirming equilibrium (CAE), all players must believe they are in some CAE, though they need not know which. The concept of CAE is thus defined setwise as follows.

Definition 3 (Convention-affirming equilibrium)**.** *A non-empty, finite subset $\mathcal{P}^*$ of ΔA is:*

A unitary set of convention-affirming equilibria (or unitary CAE set) if there exists a single prior μ with support $\mathcal{P}^*$ and a reaction function r which is a best response to μ, such that each $p \in \mathcal{P}^*$ is a fixed point for r.

A diverse set of convention-affirming equilibria (or diverse CAE set) if there exists, for each $p \in \mathcal{P}^*$, a prior μ with support $\mathcal{P}^*$ and a reaction function r which is a best response to μ, such that p is a fixed point for r.

Say that an action distribution p^* is a *unitary convention-affirming equilibrium* if it belongs to a unitary CAE set and a *diverse convention-affirming equilibrium* if it belongs to a diverse CAE set. Note that every unitary CAE is also a diverse CAE, but not vice versa.

In both cases, agents agree on – and are, in a sense, 'right' about – the set of possible action distributions which can occur as equilibria. In both cases too, all such action distributions can be fixed points of a reaction function which is a best response to beliefs allowed under the equilibrium definition. The difference is that in the 'unitary' case, agents have a common prior about the probability that each action distribution they consider possible is the true one – and thus 'know the true reaction function', since everyone reacts in the same way they themselves do – while in the latter case, agents may be uncertain of the prior probabilities others assign to different action distributions, and thus about which reaction function is the true one.[4]

1.3.3 Interpretation

Although I do not model the dynamics of any out-of-equilibrium process here, we can think of a CAE as representing the limiting outcome of such an adjustment process, after an initial phase between when a society comes into existence and when it arrives at some equilibrium convention. The assumption that agents believe they are already in some CAE might correspond to the belief that most societies are old, and agents don't know exactly how old theirs is (so that the possibility they were born in the initial adjustment phase can be dismissed as unlikely). In such a process we would also have in mind that the sample agents receive comes from learning about the outcomes of past matches (at an earlier time but from the same equilibrium distribution), and that they update based on this before playing their match.

[4] In general, we would want to consider possible distributions over priors, and hence over reaction functions. But since, as we will see in Section 1.5, the set of fixed points consistent with a class of allowed beliefs is defined by the most extreme possible reaction functions, it is without loss of generality to use the definition given above.

It is natural in this motivating story to think of agents who see the game played by others many times but do not play many times themselves; games associated with a given life or career stage which occurs only once, for example. In the present setting, this is less restrictive than it may appear. An agent who observes an action played by their own opponent receives exactly the same information as if they observed the same action played in a match they were not a part of. This is in contrast to, say, an extensive game, in which a player with an unusual strategy might learn about different aspects of their opponent's strategy than would be revealed in that opponent's matches with others. Only the total sample size of observed actions matters in our setting, not their source. A model where agents play many times would thus differ only in that their sample size would increase over a lifetime; in the case where their sample of others' matches is large relative to the number of times they themselves play, it would be well approximated by the present case.

The notion of 'society' involved here can be flexible, corresponding perhaps to a whole society in the typical sense of the term, or perhaps to a particular community or organization within one. The functional definition is something like 'the unit within which an equilibrium convention is determined for the game in question'; hence, all matches within the bounds of a given 'society' are governed by the same convention, whereas crossing the boundary from one 'society' to another may result also in a change of convention. Which way of drawing these boundaries, and thus specifying the population whose conventions we are studying, is most plausible will depend on the application.

The unitary version of the concept corresponds to a setting in which there is prior agreement (and hence, presumably, accurate information) about the relative likelihood of different conventions across all possible societies, even though agents may not know which of these has been realized in their own society. This is plausible in settings where the sort of 'society' considered is relatively small (e.g. individual organizations or communities), and there is a substantial amount of ambient, accurate information about the distribution of conventions across all such societies available to the agents in each (through, say, word-of-mouth or widely publicized studies) which fixes ideas about relative frequencies of conventions in general before agents attempt to determine the particular convention they face. It is also plausible if societies have a publicly observable reputation for what kind of internal conventions they are believed to have, which coordinates expectations among new agents before they enter (and thus before they receive their samples as 'insiders' and update about whether the public reputation was accurate or not).

The diverse version probably makes more sense in the converse cases – when the 'society' in question is large and without many alternatives to compare it to, or where widely disseminated information about the distribution of conventions or the reputations of specific societies is not available, or where such information as is available is not considered credible or trustworthy.[5]

To fix ideas, imagine a world in which there is some number of firms employing workers, some of whom begin their careers at some such firm in any given period. Within each firm, some convention prevails, and those about to enter any such firm know which conventions are possible. Prior to entering a firm, each worker forms a prior belief about which convention will prevail there, based on whatever information they have as an 'outsider'. Once they enter it, and become an 'insider', they hear candid details about the interactions among others there and their outcomes, and use this to update their belief about which convention really prevails there before playing themselves.

As candidates for the unitary concept, we could imagine cases where there are many similar firms drawing on the same pool of workers, and some widely publicized recent survey finding, say, that 75% of them follow convention A and 25% convention B, but with the identity of the

[5]This difference in plausible applications across the two versions of the concept, together with the different predictions associated with the two versions which will be established below, could in principle be the basis for a testable empirical claim.

specific firms following each kept confidential. Alternatively, we could imagine that a given firm has, say, received recent media attention over claims that it follows convention A. In either case, we may imagine that people who anticipate entering these firms take an interest in this information, and believe others have taken an interest in it, and believe others believe others have..., leading to coordination on a common prior based on this public information.

As candidates for the diverse concept, we might imagine instead a setting dominated by a few firms whose inner workings are opaque and not reported on, or where such reports as exist are believed to be influenced by internal disputes or interests resulting in an unknown bias. People who anticipate entering these firms have much less clear information to go on before actually entering, and thus may more reasonably form different priors, and believe others will form different priors, and...

1.4 Unitary Convention-Affirming Equilibria

In this section, I will show that the unitary CAE concept coincides, in 2×2 coordination games, with the concept of Nash equilibrium. This is interesting in itself, and also motivates the study of the diverse CAE concept in the following section, in which non-Nash outcomes can also occur.

1.4.1 Pure Nash Equilibria

I start by establishing that symmetric pure Nash equilibria are also CAE. The arguments and constructions involved are very simple, but also illuminating regarding the context for later results to come. Identify each pure (symmetric) Nash equilibrium of the game with the action distribution p assigning probability 1 to the equilibrium pure action a^*, and each mixed (symmetric) Nash equilibrium with the p prescribing the same action frequencies as the equilibrium mixed action; hence when I speak of Nash equilibria in what follows I refer to these action distributions.

Call the sample in which only action a is observed the *extreme sample for* a, and the set of such samples for each $a \in A$ the *extreme samples*. The following is an obvious but important observation.

Proposition 1 (Extreme Samples Suffice)**.** *An action distribution placing probability one on action* a *is a fixed point of reaction function* r *if and only if* r *prescribes* a *with probability one after observing the extreme sample for* a

Proof. At such an action distribution, only the extreme sample for a is observed; thus it is a fixed point for r if and only if the mixed action r prescribes at this extreme sample reproduces the action distribution. □

We can then settle the pure Nash question accordingly.

Corollary 1. *Any* $\mathcal{P}^*$ *containing only pure Nash equilibria is a unitary CAE set.*

Proof. Any reaction function which assigns to each extreme sample associated with some pure Nash equilibrium in the set the corresponding pure Nash action makes each point in the set a fixed point. Such a reaction function must be a best response because at each such sample agents can rule out all points in $\mathcal{P}^*$ except the one they are at, and this is assumed to be a Nash equilibrium. □

The reaction functions used in the above are unrestricted except at the extreme samples, since an agent observing any non-extreme sample will rule out any pure Nash equilibrium (if two or more actions are being played, the true action distribution cannot place probability one on any single action). This will change as we now turn to considering sets that include action distributions which place positive weight on more than one action, starting with mixed Nash equilibria.

1.4.2 Mixed Nash Equilibria

For the remainder of Section 1.4 and Section 1.5, I will study the following 2×2 game, which I shall henceforth call *the simple coordination game.*[6]

	A	B
A	$1, 1$	$0, 0$
B	$0, 0$	$1, 1$

Figure 1.1: The Simple Coordination Game

Since there are only two actions, specifying the probability of A, p_A, and the number of times A was observed, s_A, is equivalent to specifying p and s. I will for convenience henceforth work with p_A and s_A rather than p and s when dealing with 2×2 games. The sample distribution f_p over s_A is then binomial with parameters $2N$ and p_A.

I start by showing that the set $\mathcal{P}_{Nash} = \{0, \frac{1}{2}, 1\}$ consisting of the two pure Nash equilibria $p_A = 0$ and $p_A = 1$, along with the unique mixed Nash equilibrium $p_A = \frac{1}{2}$, is a unitary CAE set for the simple coordination game.

Proposition 2. *$\mathcal{P}_{Nash}$ is a unitary CAE set for the simple coordination game.*

Proof. From Proposition 1, above, we know that $r(0) = B$ and $r(2N) = A$ are necessary and sufficient to make the two pure equilibria fixed points. The question is then how to fill in the other values of r to make the mixed Nash a fixed point as well, and to show that such an r is a best response to some valid prior.

Let r be defined so that $r(s_A) = B$ if $s_A < N$, and $r(s_A) = A$ if $s_A > N$.[7] Let $r(N)$ be the action distribution which places probability $\frac{1}{2}$ on both A and B.

[6] A more general 2×2 coordination game is considered in the extensions.

[7] That is, the distribution of actions played among players who observe samples $s_A \in \{0, ..., N-1\}$ places probability 1 on B, and that among players who observe $s_A \in \{N+1, ...2N\}$ places probability 1 on A.

The sample distribution $f_{\frac{1}{2}}$ associated with the mixed Nash assigns to each s_A the probability $\binom{2N}{s_A}(\frac{1}{2})^{s_A}(\frac{1}{2})^{2N-s_A} = \binom{2N}{s_A}(\frac{1}{2})^{2N}$. In particular, for $s_A < N$, $\pi_{p_A=\frac{1}{2}}(s_A) = \pi_{p_A=\frac{1}{2}}(2N - s_A)$. Hence, the total probability of the set of samples for which all players choose A is the same as that of the set where all choose B. Since A and B are also chosen equally often at the one sample not in either set, $s_A = N$, we have that A and B are chosen equally often overall. So, $p_A = \frac{1}{2}$ is also a fixed point for r.

I now turn to showing r as constructed is rational for a prior μ supported on $\mathcal{P}_{Nash}$. Let μ assign prior probability $\frac{1}{3}$ to each of the three Nash equilibria. (It is easy to see the argument I am about to make will work for any prior assigning positive probability to all three points; I make this specific choice for simplicity.)

Because the sample distributions for the two pure Nash equilibria place probability one on the extreme samples, players seeing any sample $s_A \in \{1, ..., 2N - 1\}$ will assign probability zero to both of them. Thus, players observing these other samples will be certain they are at the mixed Nash equilibrium. Since they are thus indifferent between A and B, any distribution of actions can be chosen, so the distributions assigned under the reaction function r I constructed in particular are consistent with r being a best response.

It remains to check the two extreme samples. $s_A = 0$ has positive probability under $p_A = \frac{1}{2}$ and $p_A = 0$ but not $p_A = 1$. The posterior belief is thus a convex combination of these with strictly positive weights and thus strictly below $\frac{1}{2}$, which makes the choice of B a best response. The analogous argument works for $s_A = 2N$. □

Because agents know they are not at either of the pure equilibria whenever they see both actions being played, their actions after such samples depend only on the probabilities they assign to action distributions under which both actions have positive probability. When the only such distribution they consider possible is the mixed Nash, they are certain that this is the true distribution after any such sample. They are thus indifferent between their actions at all such samples, so it is rather easy to construct a reaction function accommodating all three Nash equilibria which is also a best response to some prior.

1.4.3 Unitary CAE Must Be Nash

Thus far, we have considered only sets of Nash equilibria, and the inference problem of players observing non-extreme samples has been rather trivial because they consider only one interior[8] action distribution possible. So, let us consider sets $\mathcal{P}$ containing more than one interior action distribution; we thus now also consider $\mathcal{P}$ including some action distributions which are not Nash equilibria.

Definition 4. *A reaction function r is a cutoff reaction function if there exists an integer $n \in 1, ..., 2N-1$ and a number $x \in [0,1]$ such that*

$$r(s_A) = B \text{ for all } s_A < n$$

$$r(s_A) = A \text{ for all } s_A > n$$

$$r(s_A)(A) = x, \text{ for } s_A = n$$

Call the number $n - x \in [0, 2N-1]$ the index of the reaction function.

That is, a cutoff reaction function assigns a mixed action with probability x on A to some non-extreme sample, and the pure action A (B) to all samples above (below) this one. Note that this includes the case where only pure actions are played, since x can be 0 or 1. It is easy to see that each cutoff reaction function has a unique index, and each index picks out a unique cutoff reaction function.

[8]Interior in the sense of p_A strictly between the endpoints 0 and 1.

Let $0 < p_1 < \frac{1}{2} < p_2 < 1$. p_1 and p_2 are understood as two different probabilities with which players in the population will play A, which will remain fixed in what follows.

Proposition 3. *Suppose $p_1, p_2 \in \mathcal{P}$ and r is a best response to some prior μ with support $\mathcal{P}$. Then either r is a cutoff reaction function, or it is such that no $p \in (0,1)$ is a fixed point for r.*

Proof. From Proposition 1 in Milgrom (1981) and the fact that binomial samples s_A have the monotone likelihood ratio property for the parameter p_A, it follows that players' posterior probability of their opponent playing A, $\sum_{p_A \in \mathcal{P}} p_A \; \mu_{s_A}(p_A)$, must be strictly increasing in their sample s_A, for any such prior. Since their best response is to play A when this estimate is strictly above $\frac{1}{2}$, B when it is strictly below, and either action otherwise, any r which is a best response to some prior must either be a cutoff reaction function, or it must assign action A or B with probability one to all but perhaps one extreme sample, and assign a strictly mixed action to the latter if the action assigned to it is different from the others. In the latter case, there can be no interior fixed point, because the frequency of extreme samples is always less than that of the corresponding actions, as already noted. □

Remark 1. *The statement of the previous result rests on the fact that we have defined cutoff reaction functions to always assign play of only B at the smallest sample and only A at the largest. Higher and lower cutoffs than this are uninteresting, in light of this result.*

The next two results work toward the claim that each cutoff reaction function makes at most one interior action distribution a fixed point. Together with the preceding, this implies that there cannot be unitary CAE sets containing more than one interior point, and thus also that there cannot be non-Nash unitary CAEs.

Proposition 4. *For each $p_A \in (0,1)$, there exists exactly one cutoff reaction function r such that p_A is a fixed point for r.*

Proof. Note first that the right-hand side of the fixed point equation $p_A = \sum_{s_A} f_{p_A}(s_A) r(s_A)(A)$, is continuous and strictly decreasing in the index of r for fixed, interior p_A (a higher index and hence a higher cutoff means A is played weakly less often at every sample, and strictly less often at some sample). As observed above, the frequencies of the two extreme samples are always less than the frequencies of corresponding actions at any interior p. Thus, in particular, the cutoff reaction functions corresponding to the highest and lowest index assign probabilities of A that are too low and too high to be fixed points, respectively. By the intermediate value theorem, there thus exists an interior index such that the fixed point equation is satisfied. Because the RHS is strictly decreasing, there can be at most one such value. □

In light of the preceding, we can talk unproblematically about a function $g : (0,1) \to [0, 2N-1]$ taking each action distribution p_A to the index of the reaction function making it a fixed point, denoted $g(p_A) \in [0, 2N-1]$. Clearly g is continuous. We have also the following.

Proposition 5. *g is strictly increasing.*[9]

Proof. I will show that, for any fixed reaction function r, changing the ratio $\frac{p_A}{1-p_A}$ by a fixed scalar multiple $\gamma \approx 1$ leads to a continuous and strictly larger change in the ratio of frequency of A to frequency of B played under r. Since equality of these ratios is equivalent to the fixed point condition, it follows that, starting from some p_A and the corresponding $g(p_A)$, a slightly larger (smaller) p_A can be a fixed point only for r with a slightly higher (lower) cutoff, since $g(p_A)$ itself over-(under-)shoots the target in these cases.

It is a basic property of the binomial distribution[10] that the sample frequencies $f_{p_A}(s_A)$ can be written recursively given the value $f_{p_A}(0)$ as

$$f_{p_A}(k+1) = \frac{2N-k}{k+1}\frac{p_A}{1-p_A}f_{p_A}(k)$$

[9] An increasing index implies an increasing cutoff, and thus a decreasing frequency of A played at each p. Since the frequencies played increase faster than those generating the sample at a given equilibrium, the cutoff must increase somewhat, *decreasing* the frequency with which A is played at each p, in order to make some slightly higher p an equilibrium instead.

[10] See e.g. Krishnan (2006) p. 39

Suppose r has a cutoff at sample s_n, $n = 1, ...2N-1$, with fraction x playing A at s_n (equivalently, suppose the index of r is $n - x$). The ratio of the frequency of A played to the frequency of B played under r at p is

$$\frac{xf_{p_A}(n) + f_{p_A}(n+1) + ... + f_{p_A}(2N)}{f_{p_A}(0) + ... + f_{p_A}(n-1) + (1-x)f_{p_A}(n)}$$

This is the ratio of the total frequency of observed samples where A is played to the total frequency of observed samples where B is. Let $\beta_n = (\frac{p_A}{1-p_A})^n \prod_{k=0}^{n-1} \frac{2N-k}{k+1}$. We can rewrite this, in light of the above, as

$$\frac{x\beta_n f_{p_A}(0) + \beta_{n+1} f_{p_A}(0) + ... + \beta_{2N} f_{p_A}(0)}{f_{p_A}(0) + ... + \beta_{n-1} f_{p_A}(0) + (1-x)\beta_n f_{p_A}(0)} = \frac{x\beta_n + \beta_{n+1} + ... + \beta_{2N}}{1 + ... + \beta_{n-1} + (1-x)\beta_n}$$

Suppose a small change from p_A to $p_A^{'}$ satisfies $\frac{p_A}{1-p_A} = \gamma \frac{p_A^{'}}{1-p_A^{'}}$, $\gamma \approx 1$. Let α denote the corresponding multiplicative change in the frequency of the smallest sample, so that $f_{p_A^{'}}(0) = \alpha f_{p_A}(0)$. Then we have

$$\frac{xf_{p_A^{'}}(n) + f_{p_A^{'}}(n+1) + ... + f_{p_A^{'}}(2N)}{f_{p_A^{'}}(0) + ... + f_{p_A^{'}}(n-1) + (1-x)f_{p_A^{'}}(n)} =$$

$$= \frac{x\alpha\beta_n\gamma^n f_{p_A}(0) + \alpha\beta_{n+1}\gamma^{n+1} f_{p_A}(0) + ... + \alpha\beta_{2N}\gamma^{2N} f_{p_A}(0)}{\alpha f_{p_A}(0) + ... + \alpha\beta_{n-1}\gamma^{n-1} f_{p_A}(0) + (1-x)\alpha\beta_n\gamma^n f_{p_A}(0)} =$$

$$= \frac{x\beta_n\gamma^n + \beta_{n+1}\gamma^{n+1} + ... + \beta_{2N}\gamma^{2N}}{1 + ... + \beta_{n-1}\gamma^{n-1} + (1-x)\beta_n\gamma^n}$$

If $\gamma > 1$, we then have

$$\frac{x\beta_n\gamma^n + \beta_{n+1}\gamma^{n+1} + ... + \beta_{2N}\gamma^{2N}}{1 + ... + \beta_{n-1}\gamma^{n-1} + (1-x)\beta_n\gamma^n} > \gamma \frac{\gamma x\beta_n + \gamma^2(\beta_{n+1} + ... + \beta_{2N})}{(1 + ... + \beta_{n-1}) + \gamma(1-x)\beta_n} \equiv \frac{\gamma x\beta_n + \gamma^2 B_H}{B_L + \gamma(1-x)\beta_n}$$

We want to show that this is greater than γ times the original ratio, that is, that

$$\frac{\gamma x\beta_n + \gamma^2 B_H}{B_L + \gamma(1-x)\beta_n} > \gamma \frac{x\beta_n + B_H}{B_L + (1-x)\beta_n}$$

Rearranging and canceling yields the equivalent inequality

$$x(1-x)\beta_n^2 < B_L B_H$$

Since $x(1-x) \leq \frac{1}{4}$ and B_L and B_H are sums of all the betas above and below β_n (and there is always at least one of each), a sufficient condition for this is

$$\beta_n^2 < 4\beta_{n-1}\beta_{n+1}$$

Plugging in the definition of β_k, we have

$$(\frac{p_A}{1-p_A})^{2n}(\prod_{k=0}^{n-1}\frac{2N-k}{k+1})^2 < 4(\frac{p_A}{1-p_A})^{n-1}(\prod_{k=0}^{n-2}\frac{2N-k}{k+1})(\frac{p_A}{1-p_A})^{n+1}(\prod_{k=0}^{n}\frac{2N-k}{k+1})$$

which reduces to

$$\frac{n-1}{n}\frac{2N-(n-1)}{2N-(n-2)} < 4\frac{n}{n+1}\frac{2N-n}{2N-(n-1)}$$

Since $\frac{n-1}{n} < \frac{n}{n+1}$ and the ratio of $\frac{2N-n}{2N-(n-1)}$ to $\frac{2N-(n-1)}{2N-(n-2)}$ is maximized for $n = 2N-1$, where it is $\frac{1}{2}/\frac{2}{3} = \frac{3}{4}$, the inequality holds for all n.

Thus, a scalar change of $\gamma > 1$ in $\frac{p_A}{1-p_A}$ leads to a scalar change strictly larger than γ in the ratio of frequencies prescribed by r.

An exactly analogous argument holds for $\gamma < 1$. □

We then have the following, because g is strictly increasing, and accordingly no two interior p_A can be made fixed points by the same cutoff reaction function.

Corollary 2. *If $\mathcal{P}^*$ is a unitary CAE set, then either $\mathcal{P}^*$ contains no $p_A \in (0, \frac{1}{2})$, or $\mathcal{P}^*$*

contains no $p_A \in (\frac{1}{2}, 1)$.

Observing that any interior point above (below) the mixed Nash implies a strict preference for A (B), and that agents will thus be certain that one action is strictly preferable to the other at any non-extreme sample if $\mathcal{P}$ contains some $p \in (0, \frac{1}{2})$ but no $p \in (\frac{1}{2}, 1)$ or vice versa, we have also the following.

Theorem 1. *If* $\mathcal{P}^*$ *is a unitary CAE set, then all of its elements are Nash equilibria.*

In other words, the set $\mathcal{P}_{Nash}$ which we showed was a unitary CAE set earlier is in fact the largest possible unitary CAE set. Note that this result does not depend much on the game, beyond the cutoff structure of best responses that all non-trivial 2×2 games share. It derives from a structural property of reaction functions – they lead to frequencies of play that change more quickly than those they are reacting to around any interior equilibrium point, forcing uniqueness.

1.4.4 Discussion

Let us now pause and take stock. Starting with the unitary CAE concept, which assumed a common prior about the likelihood of possible conventions but no knowledge of the actual one, we have recovered Nash equilibrium. Note that the preceding result does not depend in any way on agents having or accumulating large amounts of information; it is consistent with sample sizes in the single digits. Accordingly, it is the interactive reasoning process itself, including reasoning about the observations and updating of others and their implications, which allows us to arrive at such a restricted result.

Taken back to our interpretation of unitary CAE as a case where agents are informed of the distribution of conventions across different societies but not the realized convention in their own, or where their society has some public reputation observed by all who enter it before they receive their idiosyncratic samples, we may predict that in such cases the *set* *of* Nash equilibria is likely to be a good prediction of realized behavior. That is, we may expect behavior in any such society to correspond to one of the Nash equilibria, though we cannot predict in advance which one, and agents themselves may never learn for sure which one they face (though in the mixed case many will).

As we shall see in the following section, the diverse CAE concept makes quite different predictions, and in particular does not coincide with Nash equilibrium. Accordingly, another benefit of the preceding result is clarifying that it is the presence of non-common beliefs, specifically, that breaks the connection with Nash equilibrium when we consider the diverse version of the concept. Limited sample size, per se, and the according inability of agents to directly estimate the action distribution to high confidence, is not the issue.

1.5 Diverse Convention-Affirming Equilibria

Our inability to get multiple interior points, and thus non-Nash points or interesting inference problems for interior samples, is a direct consequence of the assumption of a single prior, and thus a single reaction function, which must simultaneously justify all points in a unitary CAE set. This motivates study of the more general diverse CAE concept.

First I show that the diverse version, in the absence of further restrictions, pushes us to the opposite extreme of 'anything goes'.

Proposition 6. *Let $\mathcal{P}^*$ be any finite set of points in $[0,1]$ containing p_1, p_2 s.t. $0 < p_1 < \frac{1}{2} < p_2 < 1$. Then $\mathcal{P}^*$ is a diverse CAE set for the simple coordination game.*

Proof. We need to construct, for each $p \in \mathcal{P}^*$, a cutoff reaction function making it a fixed point, which is a best response to some prior with support $\mathcal{P}^*$. The two pure Nash equilibria, if included, are fixed points for any reaction function prescribing the corresponding action at the corresponding extreme sample; such a reaction function is a best response to any prior putting probability sufficiently close to one on the pure Nash equilibrium in question, and only a small amount on each other point.

From the above, any interior $p \in \mathcal{P}^*$ is a fixed point for some cutoff reaction function. It thus suffices for the rest to show that any cutoff reaction function is a best response to some prior with support $\mathcal{P}$. I do this by constructing such priors which place probability arbitrarily close to one on just the two points p_1 and p_2.

Since $0 < p_1 < \frac{1}{2} < p_2 < 1$, B is a best response if p_1 is the true distribution, and A is a best response if p_2 is. Assume the total weight placed by the prior on all points other than p_1 and p_2 is at most ϵ, and let ϵ approach zero. With the conditional distribution over points other than p_1 and p_2 fixed as $\epsilon \to 0$, the overall posterior for any given interior sample is less than $\frac{1}{2}$ when almost all prior weight is on p_1, greater when almost all prior weight is on p_2, and strictly increasing in the fraction of the weight on both which is placed on p_2. To justify any given cutoff reaction function, then, it suffices to set these prior weights in such a way that the posterior is exactly $\frac{1}{2}$ at the sample where actions may be mixed. $\square$

1.5.1 Restricted Priors

This is, perhaps, a little too unrestrictive. The proof relies heavily on constructing priors that effectively write off all but a few possible action distributions. It is thus natural to wonder next whether forcing priors to place at least some minimum weight on each possible distribution would lead to a narrower set of restrictions. The answer is yes, at least if the sample size N is also sufficiently large.

Definition 5. *$\mathcal{P}^*$ is an ϵ-skeptical diverse CAE set if it is a diverse CAE set and each $p \in \mathcal{P}^*$ is a fixed point for some reaction function which is a best response to a prior μ with support $\mathcal{P}^*$ satisfying $\mu(p) \geq \epsilon$ for every $p \in \mathcal{P}^*$.*

That is, an ϵ-skeptical diverse CAE is one that can be played even by agents who are not ‘dogmatic’, in the sense that they assign some minimum probability to each distribution they consider possible. We have the following.

Proposition 7. *For each $\epsilon > 0$ and each $\delta > 0$ there exists N s.t. for all larger sample sizes, any ϵ-skeptical diverse CAE set can contain only interior p within distance δ or less of the mixed Nash equilibrium.*

Proof. This is immediate from the fact that the posterior will concentrate on those subjectively possible points nearest the sample average, for ϵ fixed and in the limit of large N. There is thus concentration on a set including the true p, which if it is not close to the mixed Nash contains only points making A or B a strict best response. □

This has an interesting interpretation if we think of the sample size representing how commonly played a game is in agents' society, and thus how many times they are likely to observe the outcome of others playing it. Any game, no matter how often played, can be played with positive frequencies on both actions that do not coincide with the mixed Nash. But, very eccentric such distributions survive only if the game is rare, and agents thus have limited data about it; in games played very frequently, we do get convergence to Nash outcomes, in the sense that the true frequencies are close to the Nash ones.

1.5.2 Completeness

This still leaves us with a large and somewhat unclear set of possibilities. There is a reasonable refinement which can help us narrow the set of possibilities. I will say that a CAE set is *complete* if it contains every action distribution which can be a fixed point for a reaction function which is a best response to some allowed prior over it. I will make this precise in what follows.

But first, since as we will see finding complete CAE sets requires looking at infinite sets of action distributions, I will introduce a more general way of specifying admissible sets of action distributions and allowed priors over them, and a new, more general definition of (diverse) CAE sets relative to such choices of allowed classes. I will then show that the class we have been considering contains no complete CAE sets, and will specify an alternative class, in which there are complete CAE sets that I will characterize.

Fix a collection of admissible subsets $\mathcal{P} \subseteq \Delta A$ of action distributions and, for each of these, a collection $\mathcal{M}(\mathcal{P}) \subseteq \Delta\Delta A$ of allowed priors over each $\mathcal{P}$. For example, in the original definition of a diverse CAE set the admissible $\mathcal{P}$ were the finite subsets, and all priors supported on each were allowed. At the end of the previous subsection, we narrowed the set of allowed priors to contain only those satisfying the ϵ-skepticism condition.

Given a choice of the admissible subsets and allowed priors, the *population response map* is defined as follows.

Definition 6. *The population response map R takes each admissible $\mathcal{P}$ to the set of $p \in \Delta A$ which can be fixed points for some reaction function r which is a best response to some prior $\mu \in \mathcal{M}(\mathcal{P})$. Say that $(\{\mathcal{P}\}, \{\mathcal{M}(\mathcal{P})\})$ is response-closed if for each admissible $\mathcal{P}$, $R(\mathcal{P})$ is also admissible.*

For obvious reasons, we prefer to work with response-closed classes of sets. We can then recast the definitions of diverse CAE and completeness, relative to a fixed choice of allowed sets, as follows.

Definition 7. *A diverse CAE set for allowed sets $\{\mathcal{P}\}$ and $\{\mathcal{M}(\mathcal{P})\}$ is an admissible $\mathcal{P} \subseteq \Delta A$ with the property $\mathcal{P} \subseteq R(\mathcal{P})$.*

It is complete if it also satisfies $R(\mathcal{P}) \subseteq \mathcal{P}$ (equivalently $\mathcal{P} = R(\mathcal{P})$).

This is a new definition, because it works with general classes of allowed sets, not just the particular choices made originally. I now show that there are no non-trivial finite complete CAE sets for allowed priors satisfying the ϵ-skepticism condition, which will motivate studying a different $(\{\mathcal{P}\}, \{\mathcal{M}(\mathcal{P})\})$ for the remainder of this section. Since the class of allowed priors over any finite $\mathcal{P}$ satisfying the ϵ-skepticism condition is convex, the following implies that no finite $\mathcal{P}$ can be a complete CAE set for this class of allowed priors.

Proposition 8. *If $\mathcal{M}(\mathcal{P})$ is convex, then $R(\mathcal{P}) \bigcap (0,1)$ is convex.*

Proof. If $R(\mathcal{P}) \bigcap (0,1)$ is empty or contains only a single p, this holds trivially. Suppose it contains two or more p.

Fix $p_1, p_2 \in R(\mathcal{P}) \bigcap (0,1)$. By definition, there exist $\mu_1, \mu_2 \in \mathcal{M}(\mathcal{P})$ and r_1, r_2 which are best responses to them such that p_1 is a fixed point for r_1 and p_2 is a fixed point for r_2. Consider the collection of convex combinations $y\mu_1 + (1-y)\mu_2$, $y \in [0,1]$; by assumption each is also is an allowed prior.

The map from y to the new prior thus defined, the map from this prior to its posterior under each possible sample, and the map from each such posterior to the associated point estimate of the true p are all continuous, so their composition is as well. Thus, by the intermediate value theorem, every possible point estimate of the true p after a given sample between that prevailing under μ_1 and that prevailing under μ_2 occurs for some y. In particular, if for a given sample this is above $\frac{1}{2}$ for μ_1 and below it for μ_2, it is exactly $\frac{1}{2}$ for some y.

Consider any reaction function r with an index i between that of r_1 and r_2. By assumption $\lceil i \rceil$ is weakly smaller than the ceiling of the higher of the two indices for r_1 and r_2, and weakly larger than that of the lower. It follows that the associated point estimates for the sample $s_a = \lceil i \rceil$ under μ_1 and μ_2 are such that one is weakly higher and the other weakly lower than $\frac{1}{2}$. So, by the above, the prior associated with some y makes it exactly $\frac{1}{2}$, and r is thus a best response to it. The result then follows from Propositions 4 and 5. $\square$

Accordingly, if we wish to find complete CAE sets, we have to consider collections of allowed $\mathcal{P}$ that include intervals. For the remainder of this section, I will take the admissible $\mathcal{P}$ to be all sets of the form $[q, 1-q] \bigcup \{0,1\}$ for $q \in [0, \frac{1}{2})$. That is, an admissible $\mathcal{P}$ is a set of action distributions consisting of the two extreme distributions together with a closed interval which is symmetric around $\frac{1}{2}$. Let $\mathcal{P}_q$ denote the admissible set containing the interval $[q, 1-q]$; it is easy to see that the mapping from q to this admissible set is bijective, so that we can take the class of admissible $\mathcal{P}$ to be indexed by $q \in [0, \frac{1}{2})$.

We want the class of allowed priors over each $\mathcal{P}_q$ to capture the same idea as the ϵ-skepticism assumption did in the preceding section, in a way appropriate for priors with continuous support. For simplicity, and since we are interested here in which interior equilibria are possible, I will assume agents attach no weight to either of the two extreme distributions (given our previous results about cutoff reaction functions and extreme samples, these will be fixed points in any case where interior fixed points are also possible). The set of allowed priors $\mathcal{M}(\mathcal{P}_q)$, which are thus supported on the interval $[q, 1-q]$, will then be those with a density[11] bounded by upper and lower multiples of the uniform distribution on $[q, 1-q]$.

Fix a parameter $\eta \in (0, 1)$, which determines how strict the bounds on allowed beliefs are, with higher η corresponding to stricter bounds. I then define $\mathcal{M}(\mathcal{P}_q)$ to consist of all probability measures supported on $[q, 1-q]$ with densities whose range is bounded between $\eta\frac{1}{1-2q}$ and $\frac{1}{\eta}\frac{1}{1-2q}$. Since the uniform distribution on $[q, 1-q]$ has constant density $\frac{1}{1-2q}$, this just says that lower bound on allowed prior densities is η times the uniform density, and the upper bound is $\frac{1}{\eta}$ times the uniform density. Note that as η approaches zero, we move toward a case where all densities are allowed prior densities, whereas as η approaches one, we approach a state in which only the uniform density is allowed.

In seeking complete CAE sets, we have an additional restriction. Most fixed points, as we have seen, arise in connection with reaction functions that mix at some sample. But since mixing requires indifference, such points cannot define the endpoints of the interval of possible action distributions; more extreme mixtures can be best responses to the same

[11] I do not require this to be continuous, as allowing discontinuities simplifies some of the following proofs.

beliefs. We need accordingly to restrict our search to intervals with the 'right' class of endpoints. This is formalized as follows.

Definition 8. *A pure cutoff reaction function r is a cutoff reaction function with the property that $x = 0$ or $x = 1$. Let r_m denote the cutoff reaction function with the property that $r_m(s_A) = A$ iff $s_A > m$.*

Let p_m denote the unique interior fixed point associated with r_m. Note that $m_1 < m_2$ implies $p_{m_1} < p_{m_2}$ (from Proposition 5).

We thus further restrict attention to $\mathcal{P}_q$ for which $q = p_m$ for some m.[12] Write $\mathcal{P}_m$ for an element of this class (which consists of finitely many sets). We have the following.

Proposition 9. $(\{\mathcal{P}_m\}, \{\mathcal{M}(\mathcal{P}_m)\})$ *is response-closed.*

Proof. By Proposition 8, $R(\mathcal{P}_m) \bigcap (0, 1)$ must be an interval. By symmetry, it must be a symmetric interval. The results from Section 1.4.3 establishing the relation between cutoff reaction functions and fixed points, and that every best response is a cutoff reaction function, apply unchanged to this setting, so by the above reasoning about pure cutoffs it must be a closed interval whose endpoints are equal to some p_m. The two extreme action distributions are fixed points for any cutoff reaction function. The pure cutoff reaction function r_N is a best response to the uniform prior, which is always allowed, so existence is not an issue. □

We are now in a position to prove the existence of complete CAE sets. They exist only for some values of the parameter η; this is a consequence of the finite set of possible endpoints, which may fail to 'line up' in any case with the interval of possible fixed points for reactions functions which are best responses to allowed beliefs over any one of them. But, we can always choose to work with η for which they do exist, if analyzing complete CAE sets is desired.

[12]Note, by symmetry, that $q = p_m$ iff $1 - q = p_{m'}$, for some m'.

Our goal is to prove the following.

Theorem 2. *For each N, $2 < m < N$, there exists an interval $\eta(N,m) \subseteq (0,1)$ s.t. if $\eta \in \eta(N,m)$, then $\mathcal{P}_m$ is a complete CAE set.*

It will be helpful to prove the following claim first, and then use it to prove Theorem 2.

Proposition 10. *Fix $\mathcal{P}_m$, a sample s, and $0 < \eta_1 < \eta_2 < 1$. Then the posterior point estimates of p agents may hold after observing s form an interval whose endpoints are strictly between p_m and $1-p_m$ for the classes of allowed beliefs associated with both η_1 and η_2, and the highest and lowest allowed posterior point estimates associated with η_1 are strictly higher and lower, respectively, than those associated with η_2.*

Proof. The first claim is a direct consequence of the fact that priors are bounded below by a strictly positive density, together with the convexity property used in the proof of Proposition 8.

For the second, I will show, for each allowed belief under η_2, a way of constructing allowed beliefs under η_1 that give strictly higher and lower posterior point estimates.

Note first that, if a density μ over $[a,b] \subseteq [p_m, 1-p_m]$ satisfies the bounds in the definition of an allowed prior for η_2, then the density $\alpha\mu$ satisfies the bounds for η_1 for any multiplicative factor $\alpha \in [\frac{\eta_1}{\eta_2}, \frac{\eta_2}{\eta_1}]$.

Given an allowed density μ under η_2, construct a new density μ' whose restriction to the interval $[p_m, \frac{1}{2}]$ is α_L times that of μ, and whose restriction to the complementary interval is α_H times that of μ, for some choice of α_L and α_H within the interval $[\frac{\eta_1}{\eta_2}, \frac{\eta_2}{\eta_1}]$ which make μ' also a probability measure. Clearly there exist such choices, some satisfying $\alpha_L < 1 < \alpha_H$ and some the reverse.

The posterior point estimate at s given μ can be written as a weighted sum of the conditional expectations under the posterior μ_s of p in the intervals above and below $\frac{1}{2}$, weighted by the total probability of these upper and lower intervals. This is

$$\frac{\int_{p_m}^{\frac{1}{2}} f_p(s)d\mu}{\int_{p_m}^{1-p_m} f_p(s)d\mu}\frac{\int_{p_m}^{\frac{1}{2}} pf_p(s)d\mu}{\int_{p_m}^{\frac{1}{2}} f_p(s)d\mu}+\frac{\int_{\frac{1}{2}}^{1-p_m} f_p(s)d\mu}{\int_{p_m}^{1-p_m} f_p(s)d\mu}\frac{\int_{\frac{1}{2}}^{1-p_m} pf_p(s)d\mu}{\int_{\frac{1}{2}}^{1-p_m} f_p(s)d\mu} \equiv W_L \cdot E_L + W_H \cdot E_H$$

Since E_L and E_H are unchanged by any multiplicative change in the density over their respective intervals (it factors out of both sides of the ratio and cancels), the change in overall posterior point estimate is exhausted by the change in the weights. Since these sum to one, it is without loss to focus only on W_L.

We can rewrite $W_L = \frac{L}{L+H}$, where $L = \int_{p_m}^{\frac{1}{2}} f_p(s)d\mu$, $H = \int_{\frac{1}{2}}^{1-p_m} f_p(s)d\mu$. After rescaling by α_L and α_H, this becomes $\frac{L}{L+\frac{\alpha_H}{\alpha_L}H}$, which is strictly less than L if $\frac{\alpha_H}{\alpha_L} > 1$ and strictly greater if $\frac{\alpha_H}{\alpha_L} < 1$. Thus, from the above there exist choices of α_L and α_H which move these weights, and thus the overall posterior point estimates, in either direction.

This proves the result.

□

Proof of Theorem 2. Since agents are indifferent when their point estimate is exactly $\frac{1}{2}$, a given cutoff is possible if and only if it is possible for the point estimate to equal $\frac{1}{2}$ at the associated sample. Since the posterior point estimates for any given prior are strictly increasing in the sample, the lowest and highest such samples for which this is possible define the set of possible cutoffs and thus fixed points. By symmetry, these endpoints are of the form $p_m, 1 - p_m$.

From Proposition 10, reducing η expands the set of possible point beliefs at each sample in a continuous and strictly monotone way. As η approaches zero or one, it is eventually the case that the posterior point estimate at any $2 < m < N$ can be $\frac{1}{2}$ or that it cannot, respectively (as priors with arbitrary loose bounds can essentially ignore samples of any fixed size, and the limiting uniform prior can only have posterior point estimate $\frac{1}{2}$ for $m = N$). Therefore, there exists an intermediate range of η for which $m, 2N - m$ are the smallest and largest samples for which indifference is possible. □

1.5.3 Discussion

The concept of complete diverse CAE has thus given us a different predicted set. The Nash equilibria are all still possible, but in addition so are a range of possible 'misunderstandings', in which agents, on average, best respond to an estimate of the convention they face that is on the opposite side of the mixed Nash from the convention they actually face. The possibility of such misunderstandings makes them rational: if agents believe other agents can guess wrong in this way, it opens up the support of their priors in a way that allows them to guess wrong themselves (by coming to believe in an interior action distribution less than $\frac{1}{2}$ when the true distribution is greater, or vice versa). The sample size also now affects the predicted set, forcing possible misunderstandings to be smaller the more agents observe, whereas previously the exact Nash result held even for quite small samples.

The intermediate nature of this prediction – there can be misunderstandings, but only sufficiently small misunderstandings – is intuitive given the intermediate nature of the concept. Agents don't completely know the convention they face, so they cannot be expected to avoid misunderstandings entirely. But they also see enough, and know others see enough, not to be fooled into going along with a misunderstanding that is too extreme, and the more they see, the harder they are to fool. This is, arguably, realistic as a description of the situation of actual agents in the settings we have described.

1.6 Extensions

1.6.1 The General Coordination Game

Consider the following parameterized game matrix, which I call *the general coordination game.*

	A	B
A	a, a	$x, 0$
B	$0, x$	$x + b, x + b$

Figure 1.2: The General Coordination Game

Here, a is the gain from successfully coordinating if one plays A, b is the gain from successfully coordinating if one plays b, and x is a decision-irrelevant shift parameter that represents the payoff gain from one's opponent playing B rather than A, independent of one's own action. Assume $a, b \geq 0$ and $x \in \mathbb{R}$. Note that the simple coordination game is a special case of the general coordination game with $a = b = 1$ and $x = 0$.

The unique mixed Nash equilibrium of the general coordination game is at $p = \frac{b}{a+b}$. That is, the fraction playing A is equal to the ratio of the loss from miscoordinating while playing B to the sum of the losses from miscoordinating on A and on B; in particular, the fraction playing A must go down when the relative loss from miscoordinating on B goes up, to keep players indifferent between the two actions. (A, A) and (B, B) are also still pure Nash equilibria, since $a, b \geq 0$.

Let $\mathcal{P}_{Nash}$ again denote the set consisting of the three Nash equilibria of the general coordination game. We can extend the result for the simple coordination game as follows.

Proposition 11. *$\mathcal{P}_{Nash}$ is a CAE set of the general coordination game.*

Proof. As before, we need $r(0) = B$ and $r(2N) = A$ to accommodate the two pure Nash equilibria. Under the sample distribution associated with the mixed Nash, there are thus at least fraction $(\frac{b}{a+b})^{2N}$ playing A and fraction $(\frac{a}{a+b})^{2N}$ playing B. Observe that these are strictly smaller than the total fractions playing A and B at the mixed Nash, $\frac{b}{a+b}$ and $\frac{a}{a+b}$.

We thus have room to specify actions at the remaining samples which allow the mixed Nash to be a fixed point (for any values of a and b).

Complete the specification of r by setting, for all $s_A \in \{1, ...2N-1\}$, $r(s_A)$ to be the action distribution with

$$p_A = \frac{\frac{b}{a+b} - (\frac{b}{a+b})^{2N}}{1 - (\frac{b}{a+b})^{2N} - (\frac{a}{a+b})^{2N}}$$

This makes the overall distribution of actions played under r at $p_A = \frac{b}{a+b}$ equal to those of the mixed Nash, so the mixed Nash is also a fixed point for r.

As before let us use the prior assigning equal weight to all three elements, though any prior would do. As before, players who see any sample where both actions are played are certain they are at the mixed Nash, so they are indifferent and the distribution we assign them after such samples is consistent with r being a best response. When they see an extreme sample, they assign weight to both the mixed Nash and the corresponding pure Nash and thus strictly prefer to match the action in the sample they see, also as before. □

The construction here largely matches the previous one. The additional insight is that the assignments of actions to the extreme samples needed to make the two pure Nash equilibria fixed points for r never interferes with our ability to also make the mixed Nash a fixed point for r, no matter where the mixed Nash is located.

1.6.2 3×3 Games

All the foregoing focused on the special case of 2×2 games, which can be analyzed in one dimension and thus exhibit many special properties. As a first investigation of how much of the preceding results generalize, I consider now two simple 3×3 games. The first is a direct generalization of the simple coordination game to three dimensions, which in this setting I will call the 3×3 coordination game, or just the coordination game.

	A	B	C
A	$1,1$	$0,0$	$0,0$
B	$0,0$	$1,1$	$0,0$
C	$0,0$	$0,0$	$1,1$

Figure 1.3: The 3×3 Coordination Game

The second is Shapley's game,[13] in which agents, instead of wanting to match their opponent's action, seek to play the action alphabetically after it: B against A, C against B, A against C.

	A	B	C
A	$0,0$	$0,1$	$1,0$
B	$1,0$	$0,0$	$0,1$
C	$0,1$	$1,0$	$0,0$

Figure 1.4: Shapley's Game

Both these games have the property that one's best response depends only on one's subjective belief about which action is most commonly played in the population; the relative frequencies of the two less common actions do not matter. The coordination game has four Nash equilibria – three in pure strategies and a mixed one where each action is played a third of the time. Shapley's game has only one, mixed, Nash equilibrium, also with equal frequencies.

The following is an easy generalization of the construction used above in 2×2 games.

[13]See e.g. Fudenberg and Levine (1998), p. 34.

Proposition 12. *The sets* $\mathcal{P}_{Nash}$ *containing all and only the Nash equilibria are unitary CAE sets in the* 3×3 *coordination game and in Shapley's game.*

Proof. In Shapley's game, agents are certain of the mixed Nash and thus indifferent after every observation, so it is very easy to construct a reaction function that works.

In the 3×3 coordination game, players who see one of the three extreme samples believe in some mixture between the corresponding pure Nash and the mixed Nash, giving them a strict preference to match the pure Nash; since any r which assigns the matching actions to the extreme samples makes the pure Nash outcomes fixed points, this takes care of all but the mixed Nash.

For the mixed Nash, note again that the extreme samples are less frequent at it than the frequencies of corresponding actions it prescribes, and there is thus room enough left to use indifference at all other samples to make the mixed Nash an equilibrium too. □

Remark 2. *It is not hard to see how the proof which has worked for this setting and the previous one could continue to work in general, for any game in our class and collection of points consisting of all the pure symmetric Nash equilibria and some mixed Nash equilibrium in which every action is played. What is less clear, and may be difficult to settle in general, is the status of sets of all Nash equilibria, including perhaps multiple mixed Nash equilibria placing positive probability on different subsets of actions.*

This takes care of the question of whether the Nash outcomes are still supportable in this context. What is less clear is whether the arguments against non-Nash outcomes for a single prior still apply in this setting, and whether the implications of the diverse CAE concept are different here.

1.7 Conclusion

I have introduced the concept of convention-affirming equilibrium and characterized the outcomes it predicts in 2×2 coordination games. In the unitary version, this provided a justification for expecting Nash equilibrium to prevail in settings in which the relative frequencies of different conventions across societies is commonly known. In the diverse version, it provided a reason to expect non-Nash conventions featuring a degree of 'misunderstanding' to arise, but also to expect the scale of misunderstandings to be limited by how much is observed.

The predictions in both cases are clear-cut and illuminating, and provide reason for optimism about the value of this solution concept in more general settings. One natural direction for future work is to see if the basic predictions – Nash equilibrium in the unitary case, some limited set of additional possible misunderstandings in the diverse case – extend to general games in the class I study here. Another is applying the concept to extensive games, where limits on how much of the strategy distribution is observed come into play, and it may lead to interesting equilibrium selection results. A third is to study directly out-of-equilibrium dynamics that may lead to CAE.

Chapter 2

Deterred by Not Knowing: Equilibrium Cooperation Without Strategic Certainty

Abstract

I apply the concept of convention-affirming equilibrium (Hudson, 2023) to a simple, two-stage game of effort choice on a joint project. I study four classes of strategy distributions in the game, in which most agents play according to one of the four pure symmetric paths of play, and show that each can be a convention-affirming equilibrium under certain circumstances. The results highlight the importance of strategic uncertainty about the outcome of deviation, as distinct from a known punishment path, as a reason why agents may conform to cooperative behavior widespread around them. They also highlight the dependence of the possible equilibrium outcomes on the payoff parameters of the game.

2.1 Introduction

Cooperative behavior is ubiquitous, and plays a fundamental role in a wide range of human phenomena. For self-interested agents, engaging in cooperative behavior that is individually suboptimal in the present requires that their expectation of future outcomes for themselves be contingent on their current cooperation. As such, explanations of the origins and persistence of such expectations, and predictions about the form they will take, are also of fundamental importance.

The incentive to cooperate entails not just that one's own cooperation will in fact be rewarded by more favorable future behavior by others, but that one is subjectively confident enough of this at the time of one's own decision to form the basis for choosing to cooperate. In order for cooperation among self-interested individuals to reliably arise in practice, most of them must hold beliefs about the future according to which cooperating in the present will make them better off in the long run than not doing so.

In studying cooperative behavior, we may ask both how it arises in the first place, and why it may be expected to persist once present. We might call these 'the evolutionary question' and 'the stability question', respectively. Both have been studied extensively in the literature, and both can motivate predictions about which sort of behavior is most likely. The present paper is about the stability question, with a particular thesis about the evolutionary question playing a motivating role in the background.

I apply the concept of *convention-affirming equilibrium* (Hudson, 2023) to a simple two-player game in which the value of a joint project is determined by the effort choices of both players across two periods. In the first stage, each player's effort increases the value to both players of the eventual outcome – albeit by less than the individual effort cost – in a way that is additively separable from the value added in the second stage. The second stage is a coordination game, with one of the two coordination outcomes being better than the other, but also more costly to miscoordinate on. This captures many of the key general features of settings in which forward-looking incentives are necessary to support initial cooperation, notably the need for expected outcomes in the second stage to be (subjectively) contingent on initial behavior.

Convention-affirming equilibrium predicts play in the game that can persist across time in a setting where it is played by successive cohorts of agents, none of whom know the whole distribution of strategies played by others, but all of whom have observed some sample of the outcomes of past play and all of whom also understand which strategy distributions are possible convention-affirming equilibria, and believe that the distribution they face must be one of those. There is thus a circularity in the definition, which I address by defining sets of convention-affirming equilibria as the primitive concept, in the formal definition below. The background thesis about the evolutionary question is thus that agents always expect to be entering a world in equilibrium, and make their inferences and choices accordingly. The stability question then becomes: which outcomes reproduce themselves in the face of a population of agents who think this way?

In the results, where I construct sets of convention-affirming equilibria for the game (or CAE sets, as I will usually call them), or where I show that such sets cannot be found for certain cases, I focus on sets that are, in a certain sense, maximal. If it is possible to expand a CAE set slightly, by adding more strategy distributions 'at the margin', and arrive at something that is also a CAE set, then there is a sense in which the original was 'unnaturally small' – restricting agents to believe in a smaller set of similar possibilities than they reasonably might. I get around this by focusing, in the results, on what I call *locally complete CAE classes* (LCCCs for short). A LCCC is a *collection* of (generally 'similarly shaped') sets with the property that one of them is CAE and all of them are such that a population of agents who believe one of them is the set of possible strategy distributions cannot itself play a strategy distribution that is close to, but not within, the union of sets in the class.

The goal is to establish properties of CAE sets which are the 'largest', and hence the most interesting, within a given region. Since characterizing the largest sets themselves would involve prohibitive technical difficulties, I instead look at classes of sets constructed so as to include all CAE sets in a given region, along with some other sets that are not CAE. In proving that some property holds for all sets in the class, or for all sets in the class that are CAE, I also prove that this property holds for the 'largest' CAE sets within the class, without first needing to establish which exactly these are. The predictions of interest are still these largest CAE sets; the LCCC construction is just a device for indirectly proving things about them that would be very hard to prove directly.

The LCCC concept in general is very permissive; there can be LCCCs including many sets that are very different from any CAE set they contain. It is accordingly reasonable to state general criteria by which to judge whether a LCCC constitutes an 'interesting' or 'reasonable' prediction – whether it is 'tight', in a qualitative sense at least, around the largest CAE sets it contains. For LCCCs whose sets are all contained in some small neighborhood of strategy distribution space – those that all agree some pure path of play is almost always played, for example – a reasonable criterion is that all sets in the LCCC should agree on which actions, at histories reached with positive probability under strategy distributions in sets in the class, are played commonly, which rarely, and which never (i.e. each strategy distribution in each set in the LCCC should designate common, rare, and never played status to the same actions, and this designation should also be the same across all sets in the LCCC). In addition, if a history is never reached within the class, there should be no restrictions on the distribution of actions that can be played there. This last is an expression of the principle that agents' beliefs about what is possible should be restricted only by their observations and knowledge of the process they are embedded in (including the reasoning of other agents).[1] I call this last condition 'empiricism' in the main text, and the condition that there is agreement on which actions are never played 'same possible actions'; both are defined formally in Section 2.3.5. Agreement on what is 'common' or 'rare' is a fuzzier condition, and I do not attempt to formally define what this means in general. In the results, common actions are (implicitly)

[1]It is unproblematic to express this by considering any behavior possible because in the present setting there are no dominated actions in the second stage.

functionally defined as actions common enough that the distribution of play at the history immediately following them can be estimated fairly precisely from agents' samples, and rare actions as actions rare enough that they cannot.

We may of course wish to study CAE sets in which strategy distributions within several different neighborhoods are all possible. In this case, we can consider LCCCs whose elements are unions of sets in each of the neighborhoods, and put conditions on the class of sets possible for each neighborhood (the 'components' of the LCCC) like those stated above for an LCCC in a single neighborhood. This is related also to the 'local' part of the LCCC definition. If we wish to consider predictions outside of a given neighborhood, we ought to consider a class of sets large enough to include them and ask if it is LCCC, rather than worrying about whether agents who believe only strategy distributions in a certain neighborhood are possible may nevertheless manage to play according to a distribution well outside the neighborhood; the latter approach would be against the spirit of the concept, according to which agents understand accurately the set of possibilities they face and attempt to discern the case they are in accordingly.

In the cases I consider, agents will be entering a world in which most other agents follow a single, symmetric path of play in the game, which I will call the *conventional path.* I use the word 'convention' to pick out the collection of strategy distributions which have a given conventional path. Such a collection corresponds to the event that following this path is typical behavior, and agents observe that it is typical. It is 'affirmed' – roughly speaking – if a population of agents who observe this continue to play along with it, and in a convention-affirming equilibrium agents will believe only in strategy distributions that are *affirmable* (by some population or other, not necessarily the one they are in). As such, 'convention' or 'conventional' in the present context merely denotes a way of playing the game which is typical or widespread. There is no normative meaning attached to it, and no descriptive content beyond merely saying that other ways of playing are uncommon.[2]

There are four possible choices of conventional path – high effort in both periods, low effort in both periods, low effort followed by high effort, and high effort followed by low effort. I will show that all four can arise in convention-affirming equilibrium under certain conditions on payoff parameters, which I characterize. There is a sense in which these are the only 'natural' candidate outcomes. If more than one path were observed frequently, then agents would have enough information to assess the relative payoffs of the different kinds of strategies frequently played by others, and would thus not rationally be able to play according to all of them except in knife-edge cases of near-indifference, which I do not consider.

In such cases, then, agents understand the consequences of playing as others typically do, but will have at most very limited observations about the consequences of playing differently, with the off-path play of others accordingly being substantially uncertain and assessed according to an agent's prior belief, even when they have many observations of conventional play. In assessing whether a given strategy distribution can be a convention-affirming equilibrium or not, we must thus consider the reasoning process of agents confronted with such an environment, who observe what others usually do and perhaps a few examples of unusual behavior as well, but who remain uncertain of what would happen if they behaved in an unusual way themselves. This necessarily involves a payoff comparison between the largely known outcome of conventional behavior and the largely unknown – and from an empirical standpoint, largely unknowable – outcome of other possible courses of action.

In the results, which largely concern which conventional paths can arise for populations facing particular values of the payoff parameters of the game, three general kinds of cases

[2]This minimalist use of the word in particular diverges from some other established uses in the literature on game theory and philosophy. See, for example, Rescorla (2019), or the discussion of these connections in Hudson (2023).

will arise. In the first case, playing along with a given conventional path, so long as most others are doing so, is strictly better or strictly worse than any other course of action no matter what would happen if one did otherwise. This case arises for conventional paths prescribing high effort initially when the gain from cheating on an opponent who exerts high effort exceeds any possible loss in the second round (relative to the conventional path). It also arises for conventional paths prescribing low effort initially when the loss from exerting high effort when one's opponent does not exceeds any possible second round gain (again, relative to the conventional path). These are the simplest cases, as they do not require us to consider beliefs off the conventional path.

In the second kind of case, it is possible in principle to gain by deviating, but the presence of deviators undermines play of the convention itself. In any case where deviation is not ruled out by knowledge of the conventional path alone, the assumptions of 'same possible actions' and 'empiricism' imply that any LCCC must allow some agents to deviate and others not to; if either behavior is then not optimal for some belief, there cannot be an LCCC for the convention.

This case arises for the conventional paths featuring high effort in both periods and low effort in both periods, for payoff parameters making first-period deviation worthwhile if and only if one expects to coordinate on high effort in the second stage after deviating. Under these payoff parameters, one can only deviate if one expects then to coordinate on high second-period effort, and other agents, understanding this, must accordingly respond to a deviator with high effort. This undermines the convention, as such a deviation then yields a higher payoff for the conventional path.

This case represents a more complicated way in which a convention can fail to be supportable in LCCC – not because the path itself is inherently suboptimal, but because there is no set of strategy distributions associated with it which can satisfy all the conditions of the solution concept simultaneously. There is a forward induction-like logic to such cases: reasoning about the possible motivations of deviators restricts the possible responses to deviators in a way that then undercuts the motivation to play according to the convention itself. This emergence of forward induction-like reasoning – organically and from first principles, without imposing forward induction – is one of the virtues of the LCCC solution concept. I discuss this and related issues more in Sections 2.6.3 and 2.6.4.

In the third kind of case, we construct a CAE set (and associated LCCC) in which it is possible to believe that deviation is worthwhile, but also that it is not; where most agents can in fact choose not to deviate for some of the beliefs they might hold, and where the actual frequency of deviation must be sufficiently low. (If deviation were too frequent, most agents would have samples containing enough observed deviations to estimate post-deviation play confidently, and either non-deviators or some kind of deviator would learn from this that their strategy is suboptimal). This case arises for the two conventional paths featuring high effort in the first period, for payoff parameters under which deviation can be motivated by expecting reliable coordination on either second-stage outcome. In such cases, a deviator can reasonably play either second-stage action, and one may respond to a deviator with either second stage action. *Because* both of these second-stage outcomes are possible, it is also possible to consider both responses to a deviation relatively likely, in which case there is a fear of miscoordinating in the second round if one deviates at first. If most agents in the population do in fact have such beliefs, it will seem better in expectation to follow the conventional path, so that high initial effort will remain stable.

The fear of miscoordination under strategic uncertainty is thus a source of forward-looking incentives that can sustain initial effort in this case. It is, in fact, the *only* way in which high initial effort can be sustained in a LCCC for this game. An outcome where most agents switch from high to low second-stage effort in response to a first-stage deviation is ruled out by the forward induction-like logic discussed above. The threat of miscoordination works precisely because no one knows how to get out of it; since agents are genuinely confused about what to expect at post-deviation histories, the threat is credible even if neither party would be willing to carry it out if they knew enough to avoid it. For agents who are not confident, for whatever reason, that they can predict their opponent's response to a deviation, it is safer not to rock the boat. It is for this reason also that high effort followed by low effort can be a convention-affirming equilibrium outcome; one might make an effort even with no expectation of getting the better second-stage outcome for fear that not doing so would lead to a misunderstanding that would be even worse.

I define the model and the solution concept in sections 2.2 and 2.3, respectively. Section 2.4 presents results, first for agents whose beliefs are concentrated on each conventional path separately and then for the case when all four are possible. Section 2.5 compares these results to those induced by other solution concepts in the same game. Sections 2.6 and 2.7 discuss the results and related issues in light of the broader literature.

2.2 Model

2.2.1 The Game

Stage 1

	H_1	L_1
H_1	$1, 1$	$-l, 1+g$
L_1	$1+g, -l$	$0, 0$

Stage 2

	H_2	L_2
H_2	V_H, V_H	$0, e$
L_2	$e, 0$	V_L, V_L

Figure 2.1: The Game

I study the two-stage, simultaneous-move extensive game of perfect information depicted in Figure 2.1. Parameter values are such that $l, g > 0$, and $0 < e < V_L < V_H$. Final payoffs are the sum of payoffs in the two stages, with no discounting.

The first stage is the standard normalized prisoner's dilemma. The second stage is a coordination game, in which the action associated with the better pure Nash equilibrium is also the one giving a lower payoff if one's partner fails to coordinate. It becomes a pure coordination game in the limit as e approaches 0, and a stag hunt game as e approaches V_L (so that, in the cases we consider, it is intermediate between the two).

A natural interpretation is as follows. Two players are tasked to work together on some joint project across two periods, and can choose high or low effort in each.

In the second period, the project is due shortly and must be completed. There are three possible outcomes. Both players may coordinate on a more ambitious completion (both play H_2), or a less ambitious one (both play L_2). The effort cost of the former for each player exceeds that of the latter by e, but the net payoff for both players is higher. Players may also miscoordinate; this leads to the worst outcome (i.e. ambitious and unambitious completions involve mutually exclusive lists of sub-tasks that can't be mixed), and is more costly for the ambitious player.

In the first period, the final outcome of the project is not determined, but the players can do some work in advance which makes each of the three possible completions in the second round better or worse, depending on effort choices in the first round. This takes the form of shifting the value of each possible completion in the second round by a fixed constant. That is, if both players make high effort in the first round, say, the total value associated with a more ambitious completion is $1 + V_H$ for each player; if both made low effort at first it would be just V_H. The assumption of additive separability across rounds here is presumably not realistic in all applications, but is convenient for our purposes.

Assuming that the cost of effort depends only on one's own choice of H_1 or L_1, the case $l \geq g$ corresponds to a 'convex final outcome improvement technology' (the gain from high effort of both players is more than double that of only one, so that the gain from slacking off when one's opponent does not is less than when one's opponent also does), while the opposite case $l \leq g$ corresponds to a 'concave' one (slacking off is more valuable when one's opponent makes an effort). In either case, (H_1, H_1) is efficient but L_1 is stage-game dominant.

2.2.2 The Four Unitary Conventions

In what follows, I will mainly focus on the four cases in which most agents play strategies leading to the same, symmetric path of play. I call this the *conventional path* and the outcomes (distributions of strategies in the population) in which most agents play according to a given conventional path *unitary conventions*. There are four possible unitary conventions, corresponding to the four possible choices of (H_1, H_1) or (L_1, L_1) in the first round and (H_2, H_2) or (L_2, L_2) in the second round. As we will see, all four of them can occur as convention-affirming equilibrium outcomes, though the conditions under which they can are different in informative ways. I briefly describe each here, and preview the later results about which payoff parameters allow each to be a convention-affirming equilibrium outcome.

'Collective Procrastination' In this case, agents start by playing (L_1, L_1), but then play (H_2, H_2) in the second round. That is, they do not avail themselves of opportunities to improve their total payoff in the early stages, but when the deadline comes they work hard and do the best they can within the payoff possibilities they have left themselves. Agents in this case conventionally play a stage-game dominant strategy followed by the best possible second-stage outcome; this is a strict best response to a case where most others are doing so no matter what would have happened had they worked hard in the first round instead. Unsurprisingly, this convention is possible for all values of payoff parameters. It is also, casual observation might suggest, a very common tacit arrangement in practice.

'Always Low (Effort)' In this case agents play (L_1, L_1) and then (L_2, L_2). This is the case where agents avoid effort in both periods and coordinate on the unambitious outcome.

They face uncertainty here about what would happen if they chose H_1 in the first period. Obviously, they want to do so only if they have a high probability of coordinating on the ambitious outcome in the second round, to compensate them for the up-front cost of effort in the first round, *and* to minimize the risk of playing H_2 in the second round, where the opponent's action is uncertain. This convention is only possible when the loss from unilaterally deviating to high effort in the first period is higher than the difference in value between the two second-stage outcomes; if it is not, deviating this way 'signals' an intent to make high effort in the second round, and this destroys the incentive to play along with the convention.

'Always High (Effort)' In this case agents play (H_1, H_1) and then (H_2, H_2). This is Pareto efficient, and presumably the convention we would most like to encourage. Agents work hard in the first round to maximize the value of their second-stage options, then coordinate on the most ambitious completion. The uncertainty agents face in this case is about what their opponent would do if they chose L_1 in the first round. This convention is possible only for intermediate values of the gain from 'cheating' in the first period, g. If g is too high, all agents want to cheat no matter what, but if it is too *low*, then an agent can only choose to 'cheat' if they still expect H_2 in the second round, which as in the previous case signals this expectation to the opponent and destroys the incentive to follow the convention. For intermediate values of g, when coordinating on either second-round outcome with certainty would motivate cheating, the convention can be followed if most agents are sufficiently subjectively uncertain which second-round outcome to expect. Cheating is deterred in this case by strategic uncertainty and the risk of miscoordination; this is credible because both players are uncertain about the belief of the other about play after a deviation, so that neither has the means to prevent miscoordination even if they wished to.

'High, Then Low' This is the opposite case from 'collective procrastination': agents start with (H_1, H_1) and then play (L_2, L_2). That is, they work at first to improve their final payoff, but then coordinate on the unambitious completion. It is tempting to assume that this cannot happen for reasons similar to those explaining the robustness of 'collective procrastination' – agents are playing a dominated stage game strategy, and yet not being rewarded by coordinating on the better second-stage outcome, so why would they work hard in the first stage? But that turns out to be false. Agents cannot be motivated to work hard at first by the threat of moving to a worse second stage *coordination* outcome, but strategic uncertainty and the risk of miscoordination can still deter them from deviating from this convention in some cases. The upper bound on g is tighter in this case than the previous (since agents have less to lose in the second round relative to the conventional outcome), but there is no lower bound, since it is always the case that an agent will wish to deviate from this convention in the first round if they can predict either second stage action with sufficient confidence.

2.2.3 Strategy Distributions and Population Behavior Strategies

The set of histories in the game, $\mathcal{H}$, consists of the empty history $\emptyset$, the four length-one histories (H_1, H_1), (H_1, L_1), (L_1, H_1), (L_1, L_1), and the sixteen length-two or terminal histories consisting of the concatenation of one of the former with (H_2, H_2), (H_2, L_2), (L_2, H_2), or (L_2, L_2). Denote the set of terminal histories by $\mathcal{Z}$, and for generic elements write $h \in \mathcal{H}$ and $z \in \mathcal{Z}$.

A (pure) strategy in the game is an element $s \in S = \{H_1, L_1\} \times \{H_2, L_2\}^4$. It specifies a choice of H_1 or L_1 at the empty history, and a choice of H_2 or L_2 at each length-one history. I will assume all agents play pure strategies.[3] For each terminal history z, let $S_i(z)$ and $S_j(z)$ be the subsets of S that are consistent with reaching z for player roles i and j, respectively.

[3] As we are dealing with a continuum population, this is without loss of generality with respect to the distributions over outcomes that can occur.

I seek to predict aggregate play in a continuum population, from which agents are randomly matched in pairs to play the game. Since the game is symmetric, and the agents in both player roles are drawn from the same population, the possible outcomes of aggregate play can be identified with the space ΔS of possible aggregate probabilities with which each strategy is chosen; the strategies of the two players are drawn iid from some such distribution, and the distribution over terminal histories is induced by this distribution over strategy pairs.

A strategy distribution is an element $p \in \Delta S$. Denote the probability of strategy s under p by p_s. In light of the above, we can think of each strategy distribution as a mixed strategy for the population, so that the aggregate play of the population induces the same distribution over terminal histories as would a single pair of players both literally playing the mixed strategy p. In particular, we can identify symmetric Nash equilibria and other outcomes we might naturally express by a choice of some such strategy with their population strategy distribution analogues.

By Kuhn's Theorem, each such 'population mixed strategy' is realization-equivalent to a behavior strategy for the population, and we will prefer in most of what follows to work with the latter. A population behavior strategy (PBS) is a vector $\sigma \in \Sigma \equiv [0,1]^5$, with coordinates written as $\sigma = (\sigma(\emptyset), \sigma(H_1, H_1), \sigma(H_1, L_1), \sigma(L_1, H_1), \sigma(L_1, L_1))$. That is, each σ assigns a number in the unit interval to each non-terminal history, interpreted as the proportion of agents in the population who play the 'cooperative' strategy at that history (H_1 at $\emptyset$, H_2 at the others) as a fraction of the total proportion whose strategies allow them to reach that history.[4]

[4]For example, if half of agents choose H_1 at first, and 30% of *those agents* then choose H_2 at (H_1, L_1), $\sigma(H_1, L_1)$ is also just 30%; the fraction able to reach (H_1, L_1) doesn't enter the calculation, and neither does the probability that their opponent plays L_1 initially (so that (H_1, L_1) is actually reached). Note that this probability is for the first player role; the probability for the second in this case would be given by $\sigma(L_1, H_1)$, which may be different.

Write the PBS induced by $p \in \Delta S$ as σ_p. It pins down the probability of each choice at each history and also the distribution over terminal histories which will determine what agents observe. As such, it is sufficient to focus on PBSs, both from the point of view of predicting the probability of outcomes and as the possible states of play that agents reason about; no information not contained in σ_p is relevant for agent updating or optimization, or for pinning down the play that occurs.

The probability of a given terminal history $z \in \mathcal{Z}$ under PBS σ, denoted $\sigma(z)$, can be straightforwardly calculated in terms of the four action probabilities involved; for example, the probability of $z = ((H_1, L_1), (L_2, L_2))$ under σ is $\sigma(z) = \sigma(\emptyset)(1-\sigma(\emptyset))(1-\sigma(H_1, L_1))(1-\sigma(L_1, H_1))$.

2.2.4 Samples and Sample Distributions

We need now to specify what agents observe before choosing their strategies and playing the game. Fix an integer N, which is commonly known. I assume that each agent, before choosing their own strategy, observes a random sample of N terminal histories reached by other pairs of agents in the population, drawn from a distribution induced by the prevailing aggregate strategy distribution p, through population behavior strategy σ_p.

A sample is a tuple $x = (x_z)_{z \in \mathcal{Z}}$, with $x_z \in \{0, 1, ...N\}$ for all z and $\sum_{z \in Z} x_z = N$. That is, x_z is the number of times terminal history z was observed under sample x, and the total number of such observations must add up to the sample size N. Write X for the space of all possible samples. We may assume for concreteness that all observed terminal histories are written from the perspective of the first player role, though the perspective taken is irrelevant to inference.

As noted above, the prevailing population behavior strategy σ_p induces a probability $\sigma_p(z)$ with which each terminal history z is reached. The associated sample distribution is thus multinomial, with outcome space $\mathcal{Z}$ and probabilities $\sigma_p(z)$. The probability mass function is

$$f_{\sigma_p}(x) = \frac{N!}{\prod_{z\in\mathcal{Z}} x_z!} \prod_{z\in\mathcal{Z}} \sigma_p(z)^{x_z}$$

2.2.5 Reaction Functions and Aggregate Play

A (pure) reaction function is a mapping $r : X \to S$, which assigns to each possible sample a choice of strategy. A reaction function is thus a sort of super-strategy, a complete contingent plan specifying which strategy to choose after every possible sample. Write R for the set of all pure reaction functions; note in particular that it is finite.

A mixed reaction function is an element $\rho \in \Delta R$. We will think of mixed reaction functions as describing populations in which different agents may have different pure reaction functions, in which the probabilities that a given agent has each possible pure reaction function are given by ρ. I work in what follows primarily with mixed reaction functions; this includes point masses on pure reaction functions as a special case and is thus without loss.

The aggregate play induced by ρ and σ is the PBS $\alpha(\sigma,\rho) \in \Sigma$ resulting from a population described by mixed reaction function ρ and making observations generated by the (possibly different) PBS σ. ρ and σ naturally define a strategy distribution $p(\sigma,\rho)$ by the formula

$$p(\sigma,\rho)(s) = \sum_{x\in X} f_\sigma(x)\rho(\{r : r(x) = s\})$$

That is, the aggregate frequency of each $s \in S$ is the probability of the event that a given agent in the population has a sample and a reaction function leading them to play s. The aggregate play is then just

$$\alpha(\sigma,\rho) = \sigma_{p(\sigma,\rho)}$$

Definition 9. *A population behavior strategy* σ^* *is a fixed point for mixed reaction function* ρ *if* $\sigma^* = \alpha(\sigma^*, \rho)$.

2.3 Convention-Affirming Equilibrium

2.3.1 Priors and Updating

Let C be a closed subset of Σ, understood as the subset of PBSs an agent considers possible, and let $\mu \in \Delta C$ be an agent's prior. Agents update their prior using their sample and Bayes' rule. For each sample x with positive probability under the prior, their posterior belief μ_x after seeing sample x assigns to each Borel subset B of C the probability

$$\mu_x(B) = \frac{\int_B f_\sigma(x) d\mu}{\int_C f_\sigma(x) d\mu}$$

In the case of a zero-probability sample, I assume the agent will stick with their prior.

I will assume in what follows that an agent's prior and their sample are statistically independent. Thus, the distribution of posteriors under a given sample distribution and a given distribution of priors will be that resulting from independent draws from each followed by application of the above formula.

2.3.2 Best Responses

I assume agents always choose strategies which are best responses to their posterior point estimates of the true PBS after observing their sample. We will say a reaction function is a best response to an agent's prior if the strategies it prescribes are all optimal given the posteriors derived from the prior after each sample, as follows.

Definition 10. *The posterior point estimate of* σ *under posterior* μ_x*, denoted* $\hat{\sigma}(\mu_x)$*, is*

$$\hat{\sigma}(\mu_x) = \int_C \sigma d\mu_x$$

Definition 11. *A pure reaction function r is a best response to a prior μ if for each sample x,*

$$r(x) \in \arg\max_{s \in S} v(s, \hat{\sigma}(\mu_x))$$

where $v(s, \hat{\sigma}(\mu_x))$ is the expected payoff of strategy s playing against the behavior strategy $\hat{\sigma}(\mu_x)$.

2.3.3 Convention-Affirming Equilibrium

Given a closed subset $C \subseteq \Sigma$, let $\mathcal{M}(C) \subseteq \Delta C$ be a set of possible priors over C, called the *allowed priors* over C. A convention-affirming equilibrium set will be a pair $(C, \mathcal{M}(C))$ with the property that each $\sigma \in C$ could be the aggregate play of a population of agents who all hold allowed priors $\mu \in \mathcal{M}(C)$.

Which PBSs could be the aggregate play for a given allowed set is formalized as follows.

Definition 12. *A population behavior strategy σ is affirmable for $(C, \mathcal{M}(C))$ if there exists a mixed reaction function ρ such that σ is a fixed point for ρ and each pure reaction function in the support of ρ is a best response to some prior $\mu \in \mathcal{M}(C)$. Write $A(C, \mathcal{M}(C))$ for the set of σ which are affirmable for $(C, \mathcal{M}(C))$.*

Definition 13. *$(C, \mathcal{M}(C))$ is a convention-affirming equilibrium set if $C \subseteq A(C, \mathcal{M}(C))$.*

This definition leaves open the possibility that there may be additional affirmable PBSs which are not included in the CAE set. If this is the case, there is a danger of arbitrariness in the use of CAE as a prediction, since some outcomes ruled out seem reasonable in exactly the same sense those predicted do. One solution is to focus on CAE sets that contain all affirmable PBSs; I call these CAE sets *complete.*

Definition 14. *$(C, \mathcal{M}(C))$ is a complete convention-affirming equilibrium set if $C = A(C, \mathcal{M}(C))$.*

In a convention-affirming equilibrium, agents understand the convention-forming process, and have correct beliefs about which conventions can actually occur, but are uncertain which convention actually holds. They are uncertain also about the distribution of priors over the true convention held by other agents, which may differ from their own and one another's, and possible hypotheses about the true convention are tied to hypotheses about the distributions of priors for which it can occur. A complete CAE set does not restrict these hypotheses beyond requiring that they be possible for some populations holding allowed beliefs over the set, while an incomplete CAE set incorporates additional restrictions imposed by the analyst on which conventions, and hence belief distributions, agents consider possible, and also the restriction that one of these is the convention which actually occurs.

It is obviously attractive to work with complete CAE sets if possible, since the set of outcomes they predict is transparently self-consistent, with no additional restrictions. The collection of all CAE sets includes also many which are too small to be reasonable – e.g. subsets of larger CAE sets with some affirmable points arbitrarily deleted, or singleton sets which are best responses to themselves only because they exclude most off-path possibilities.

But working with complete CAE sets is not always possible. The most fundamental problem is that, for many sets of allowed priors, complete CAE sets may fail to exist. The reason is that the set of affirmable PBSs $A(C, \mathcal{M}(C))$ is tied to the subset of pure reaction functions which can be best responses for agents with beliefs in $\mathcal{M}(C)$. Since R is finite, and thus has finitely many subsets, there are accordingly also finitely many sets which can be $A(C, \mathcal{M}(C))$ for some C and $\mathcal{M}(C)$, and in general none of them may coincide with any of the C for which this is true. Even in cases where a complete CAE set does exist, it may be very difficult to characterizes, for technical reasons of limited economic interest.

In the following section, I will accordingly define a modified solution concept which coincides with completeness when a single complete set exists, but allows us to focus on a class of 'similar' sets rather than a single set in other cases. This sidesteps most of the thorniest technical problems associated with completeness. If this class is well-chosen, the interpretation of showing that it is a solution to this modified concept will have a similar economic interpretation to showing a single CAE set is complete.

2.3.4 Locally Complete CAE Classes

I will say a collection of subsets of Σ is a *locally complete CAE class* (LCCC) if it contains a CAE set, and if each affirmable PBS for each set in the class is either contained in some other set in the class, or is some minimal distance from all sets in the class.

Definition 15. *Let $\mathcal{C}$ be a class of closed subsets $C \subseteq \Sigma$ and let $\mathcal{M}(C)$ denote the set of allowed priors over each $C \in \mathcal{C}$. Say that $(\mathcal{C}, (\mathcal{M}(C))_{C\in\mathcal{C}})$ is a locally complete CAE class if*

(i) There exists $C \in \mathcal{C}$ s.t. $(C, \mathcal{M}(C))$ is a CAE set

(ii) There exists some $\epsilon > 0$ s.t. for each $C' \in \mathcal{C}$,[5] *$A(C', \mathcal{M}(C')) \bigcap B_\epsilon(\bigcup_{C\in\mathcal{C}} C) \subseteq \bigcup_{C\in\mathcal{C}} C$*

Note that a singleton class consisting of a complete CAE set is necessarily a LCCC, since by definition it is CAE and contains all points affirmable for itself. General LCCCs differ from this special case in two ways. First, I require that any PBS affirmable for a set in the class is *either* contained in some set in the class or is not 'close' to the class. This is the 'local' part; there is no 'global' requirement that all affirmable points are in the class. Second, the 'completeness' criterion applies within the class, rather than within a single set

[5] Where $B_\epsilon(\bigcup_{C\in\mathcal{C}} C)$ is the ball of radius ϵ around the union, according to the supremum metric – the set of all σ whose coordinates are all within ϵ of those of some σ' in some $C \in \mathcal{C}$.

– nearby affirmable points for one set in the class must be contained in another, but not necessarily in the first set itself.

Locally complete CAE classes, in general, do not inherently impose non-trivial restrictions; the class of all closed subsets of Σ, each endowed with the set of all possible priors over them, is always trivially a locally complete CAE class, for example.[6] Rather, the quality of the prediction is tied to the choice of class, particularly the extent to which all sets in the class occupy a 'similar' region of Σ (so that the weakening of the concept relative to that of a single complete CAE set is less severe). In the following section, I explain the kinds of classes I will investigate.

2.3.5 Maintained Assumptions

I will mostly be interested in classes $(\mathcal{C}, (\mathcal{M}(C))_{C\in\mathcal{C}})$ for which all $C \in \mathcal{C}$ are closed, convex subsets of Σ, and which also satisfy the following two properties:

Same possible actions: If some $C \in \mathcal{C}$ is such that $\sigma(h) = 0$ or $\sigma(h) = 1$ for all $\sigma \in C$ and some non-terminal history h, then the same is true for all other $C' \in \mathcal{C}$.

Empiricism: If h is reached with probability zero under all $\sigma \in C$ for some $C \in \mathcal{C}$ (equivalently for all $C \in \mathcal{C}$ in light of the 'same possible actions' condition), then $\sigma \in C$ implies $\sigma' \in C$, where σ' is any PBS s.t. $\sigma(h') = \sigma'(h')$ for all $h' \neq h$.

The first condition says that an action cannot be played with positive probability under one set in the class but not another. The second says that there can be restrictions on possible behavior only at histories which might be reached (where play might thus have actually been observed by someone, so that there is an empirical basis for these restrictions somewhere in the population). Thinking of $\mathcal{C}$ as the class of possible supports of agents'

[6] There is always at least one CAE set because any singleton consisting of a Nash equilibrium is a CAE set, and the game has a Nash equilibrium.

beliefs, these conditions say that agents consider the same subset of actions in the game subjectively possible under each $C \in \mathcal{C}$, and moreover that they are totally agnostic about play at histories which they are *ex ante* convinced are never reached. Note that the latter includes histories that they may contemplate reaching themselves by deviating in a way they are convinced no one else ever does.

Fix $\eta \in (0,1)$, a parameter of the model commonly known among the agents. For each $C \in \mathcal{C}$, let $\mathcal{M}(C)$ be the set of all priors over C with densities bounded by upper and lower multiples of the uniform prior[7] on C, with the upper and lower multiples being η and $\frac{1}{\eta}$. In what follows, I will sometimes abuse notation by writing C for $(C, \mathcal{M}(C))$ and $\mathcal{C}$ for $(\mathcal{C}, (\mathcal{M}(C))_{C \in \mathcal{C}})$ on the understanding that each $\mathcal{M}(C)$ is defined in this way.

The purpose of working with this class of allowed priors is to ensure that there is some threshold in the number of observations at any given history, uniform across the population, such that all posterior point estimates of play at this history will be within any given tolerance of the sample average above this threshold (whenever the sample average is consistent with the support of their beliefs). That is, there are no 'unboundedly stubborn' priors, a possibility not ruled out by the weaker assumption of full support. The parameter η controls how stubborn agents are allowed to be in sticking with their prior in the face of their observations, with lower values of η allowing more stubbornness.

In what follows, I will say a class $(\mathcal{C}, (\mathcal{M}(C))_{C \in \mathcal{C}})$ satisfying all these conditions is a *candidate class*. That is, $(\mathcal{C}, (\mathcal{M}(C))_{C \in \mathcal{C}})$ is a candidate class if it satisfies 'same possible actions' and 'empiricism', all $C \in \mathcal{C}$ are closed and convex, and all sets of allowed priors $\mathcal{M}(C)$ are defined as above.

[7] The Euclidean measure of dimension appropriate to C, restricted to C and normalized to 1.

In the final subsection of the results, I will consider a class which is the union of several candidate classes but not a candidate class itself. This class will be constructed from the specific candidate classes identified in earlier subsections; as such, I do not attempt to define a larger collection of possibilities to which this case belongs in the way I have for candidate classes.

I will in what follows generally assume that η is small, and that N is large relative to η. That is, I consider cases where agents are allowed to be relatively stubborn – so that sufficiently small numbers of observations at a given history do not constrain the strategies they may choose in the subgame starting there – but where the sample size is also so large that only if a history is very rarely reached in the population overall will their observations at that history really be sufficiently small in this sense (excepting outliers with vanishing probability).

2.3.6 Interpretation

While the solution concept here is taken as primitive from the point of view of the formal analysis, rather than derived in a rigorous way from any larger process, we can have the following story in mind when interpreting it. There is some class of organizations within which personnel are often assigned to work on group projects with the structure of the game we consider. Each such organization has existed for a long time already, so that some convention or other is presumed to be present in it, but personnel come and go – if not from the organization as a whole, at least from whatever role within it makes one a candidate to play the game – too quickly to accumulate much personal experience from past play. They do, however, hear candid discussions of the outcomes of past games played by others in the organization, which they use to update their prior over which convention is present before playing themselves. A convention which persists across cohorts of agents who learn and behave in this way is a convention-affirming equilibrium.

A number of restrictive assumptions are present and should be highlighted. One, already mentioned, is that only the experience of others and not one's own past experience influences beliefs at the time of play. If one has many observations of the play of others, this only matters much if one plays unconventionally oneself (so that one accumulates more 'off-path' information than would be available otherwise); if the sample size is small, the fact that it grows over time might also matter in the alternative scenario where agents play more than once.

Additionally, we assume that agents are stuck with their partners – they cannot rematch and start over, say, or join a different group mid-game if they don't like what their opponent is doing. As we will see in what follows, agents who cheat in the first period will in some cases still have the ability to coordinate on the good outcome in the second period, since their opponent is stuck with them in what at that stage is a coordination game; the assumption of fixed partners is important for this. Because one's opponent is also playing only once, their past history cannot follow them into later games, or be observed in the present game; this also is relevant to the logic of the results for the present case below.

2.4 Results

In the next four subsections, I derive a locally complete CAE class for each unitary convention separately, and establish the conditions on payoff parameters for which each exists or fails to. In the final subsection, I address the (relatively modest) additional issues needed to derive a locally complete CAE class with components corresponding to all the unitary conventions that exist for given payoff parameters.

A general point about unitary conventions is worth noting at the outset. I will focus on cases where the probability of deviation from the convention is small enough that agents are significantly uncertain about play following a deviation, and have beliefs about it which are thus non-trivially prior-dependent (sets with probabilities larger than this cannot be CAE in any of the four cases). Since the sample size is assumed large relative to the bound on allowed priors, a frequency of deviation small enough to create such uncertainty is also small enough that the possibility of one's opponent deviating can be ignored in making (strict) payoff comparisons. Thus, an agent's choice to deviate or not depends only on their belief about their opponent's action at the second-stage history after they deviate and their opponent did not; their opponent's action at that history in turn depends only on their opponent's belief about the second-stage play of those rare agents who deviate in the first-round.

The LCCCs I derive for each of the four unitary conventions will also all have the property that all agents must always play the second-stage conventional action after the conventional first-stage action profile. Because the second stage is a coordination game and agents are certain the convention is usually followed, this is a necessary condition for best responding to any allowed belief. In considering PBSs 'close' to the LCCC, we technically need to check also PBSs where there is a slight probability of deviation after the first-stage conventional action profile. It is immediate from the support of their beliefs and the assumption that they stick with their priors after observing 'impossible' samples that such points are never affirmable. Having noted this here, I omit this step from the proofs below for simplicity.

2.4.1 The 'Collective Procrastination' Convention

I first consider the 'collective procrastination' convention. This convention prescribes the play of a stage-game dominant action in the first period followed by the best payoff for both players in the second; as such, playing along with it is a strict best response to any population where most other players also play according to it, no matter what beliefs one may hold about play off the conventional path. As this would suggest, it is particularly easy to construct a simple complete CAE set for this case.

Let $C^*_{LH} \subseteq \Sigma$ be the set consisting of all and only those σ with $\sigma(\emptyset) = 0$ and $\sigma(L_1, L_1) = 1$ (that is, the subset in which the entire population plays according to this convention). I will show that C^*_{LH} is a complete CAE set, from which it follows immediately that the singleton class $\mathcal{C}^*_{LH}$ consisting only of C^*_{LH} is a locally complete CAE class.

Proposition 13. *For all η sufficiently close to zero, C^*_{LH} is a complete CAE set.*

Proof. Observe first that agents whose beliefs are supported on C^*_{LH} must assign probability one to the event that their opponent follows the convention, because all $\sigma \in C^*_{LH}$ assign probability one to this event. Furthermore, for any sample observable at some $\sigma \in C^*_{LH}$, an agent's posterior belief will be the same as their prior; since all samples consist only of observations of the conventional terminal history, no agent has information that would cause them to update about play at those histories they can be uncertain about for beliefs supported on C^*_{LH}.

Accordingly, the question of whether a given σ is affirmable for C^*_{LH} reduces to the question of whether every strategy in the support of some strategy distribution consistent with it is a best response to the *prior* point estimate associated with some allowed prior over C^*_{LH} (that is, the posterior point estimate for the case where the posterior equals the prior).

I show first that no $\sigma \notin C^*_{LH}$ can be affirmable for C^*_{LH}. $\sigma \notin C^*_{LH}$ implies either $\sigma(\emptyset) > 0$ or $\sigma(L_1, L_1) < 1$. In the latter case, some agents who make low effort at first proceed to make low effort in the second round as well. Since they are certain given beliefs supported on C^*_{LH} that their opponent will make a high effort in such a case, they have an expected payoff of e from doing so, which is less than the expected payoff of V_H from making high effort in the second stage; this cannot be optimal, so such σ cannot be affirmable.

If $\sigma(\emptyset) > 0$, some agents are choosing to make a high effort in the first stage. For beliefs supported on C^*_{LH}, their expected payoff from doing so cannot exceed $-l + V_H$, while they have a larger expected payoff of V_H from following the convention. This cannot be optimal either. It follows that any σ affirmable for C^*_{LH} must be contained in C^*_{LH}.

It remains then to show that every $\sigma \in C^*_{LH}$ is in fact affirmable for C^*_{LH}; that is, we need to show that agents whose priors have support in C^*_{LH} could choose either action at the three histories never reached under C^*_{LH}. For the two histories believed unreachable given one's opponent's action – (L_1, H_1) and (H_1, H_1) – this is trivially true; since agents are certain these histories are never reached, play there does not enter their expected payoff calculation, so that any off-path play is consistent with best responding.[8]

For play at the history (H_1, L_1) following one's own deviation, note that as η approaches zero, it is possible to construct allowed priors which put almost all weight on $\sigma \in C^*_{LH}$ for which $\sigma(L_1, H_1)$ (the probability of my opponent playing H_2 if I am at (H_1, L_1), which is (L_1, H_1) from their perspective) is arbitrarily close to either zero or one. For example, one can construct priors which have the maximal density $\frac{1}{\eta}$ on some subset, and the minimal density η elsewhere. As η approaches zero, the first subset can be made arbitrarily small and thus contained within the subset of σ for which $\sigma(L_1, H_1)$ is within any tolerance of zero or one; since almost all weight in the prior is on this subset, and the posterior is the same as the prior, the posterior point estimate can also be placed within any tolerance of zero or one by this same scheme. Thus, either action can be played at (H_1, L_1) as well. This completes the proof. □

It is not difficult to see that playing along with the convention is also a strict best response for beliefs supported on any set in which the maximum probability of deviation is sufficiently small. Accordingly, no superset of C^*_{LH} which includes also small probabilities of deviation can be a CAE set – such distributions would not be playable for populations certain the convention was usually followed – and no class including such sets could thus be a locally complete CAE class, since including C^*_{LH} in addition to such sets would violate the 'same possible actions' condition, and no larger set is CAE itself.

[8]Note that this argument depends on the fact that we only require *ex ante* rationality, not sequential rationality. See Section 2.6.4 for discussion on this point.

There is a sense in which conventions which are strictly best to play along with knowing only the conventional path, regardless of beliefs about behavior at unconventional histories, are 'naturally' complete for this reason. The same will be true of the 'always low' convention considered next, for the parameter values under which it can occur.

Conventions for which the optimality of following the convention itself depends upon beliefs at unconventional histories, like the other two we will consider, are more delicate, as agents knowing only the convention may well choose to deviate from it. Non-degenerate locally complete CAE classes, rather than single complete CAE sets, arise naturally in such cases.

2.4.2 The 'Always Low' Convention

I now consider the 'always low' convention. There are two cases. When payoff parameters are such that even a guarantee of coordinating on high effort in the second round cannot compensate the agent for the cost of effort in the first period against an opponent who is almost certain to make a low effort, we can find a complete CAE set representing it.

When payoff parameters do not satisfy this condition, there is no locally complete CAE class for this convention. If it is possible for agents to respond to a deviator with high effort in the second period, choosing to deviate would then signal, by a forward induction-like argument, that the deviator expects to coordinate on high effort in the second period, thus incentivizing the other player to do the same and undermining the convention. If a candidate class of sets for this convention does not allow for this possibility, it cannot be locally complete.

Let $C^*_{LL} \subseteq \Sigma$ be the set of all and only those σ with $\sigma(\emptyset) = 0$ and $\sigma(L_1, L_1) = 0$ (again, the subset in which the entire population plays according to this convention). I will show first that C^*_{LL} is a complete CAE set, from which it again follows immediately that the singleton $\mathcal{C}^*_{LL}$ consisting only of C^*_{LL} is a locally complete CAE class, for a specific set of payoff parameters.

Proposition 14. *If $l > V_H - V_L$, then for all η sufficiently close to zero, C^*_{LL} is a complete CAE set.*

Proof. This proof parallels that for C^*_{LH}, with the same sequence of steps.

Agents assign probability one to other agents following the convention, and always have posteriors equal to their priors, given beliefs supported on C^*_{LL}. We can accordingly restrict attention to whether given strategies can be played for allowed prior point estimates in asking if a given σ is affirmable for C^*_{LL} or not.

Any σ outside C^*_{LL} either has a positive probability of high effort in the first period or of high effort after both players made low effort in the first period. The second is suboptimal for agents with beliefs supported on C^*_{LL} because it gives them 0 instead of V_L, the first is suboptimal because, by assumption, the first-period loss from making high effort when one's opponent doesn't, l, exceeds the difference between V_H and V_L, the highest possible gain from such a deviation in the second period (since agents following the convention get V_L in the second period, and V_H is the highest possible second-period payoff). Ruling out strategies which take either of these actions rules out also any σ outside C^*_{LL} being affirmable for C^*_{LL}.

To show any action can be played at the other three second-stage histories, and thus that all $\sigma \in C^*_{LL}$ are affirmable for C^*_{LL}, we note again that anything is a best response at the two believed unreachable given the opponent's strategies, and then construct beliefs at the remaining history (H_1, L_1) identical to those used for the same history in the proof of Proposition 13 $\square$

Say that a class of sets $\mathcal{C}_{LL}$ is an ϵ-component class for LL and $\epsilon \geq 0$ if it is a candidate class, and for all $C \in \mathcal{C}_{LL}$ and all $\sigma \in C$, $\sigma(\emptyset) \leq \epsilon$ and $\sigma(L_1, L_1) = 0$.

Proposition 15. *If $l < V_H - V_L$, then for all sufficiently small $\epsilon \geq 0$ and all η sufficiently close to zero, no ϵ-component class $\mathcal{C}_{LL}$ for LL is a locally complete CAE class.*

Proof. By 'same possible actions', since $\mathcal{C}_{LL}$ is a candidate class, either $\sigma(\emptyset) = 0$ for all $\sigma \in C$ and all $C \in \mathcal{C}_{LL}$ (the $\epsilon = 0$ case), or every $C \in \mathcal{C}_{LL}$ contains some σ with $\sigma(\emptyset) > 0$.

In the first case, all agents are certain that the convention will always be played, regardless of their priors and samples. By 'empiricism', we must thus consider in this case the singleton class consisting of only C^*_{LL} (since the convention is always played, agents cannot rule out any action probabilities off the conventional path).

If this class were to be a locally complete CAE class, it would have to be the case that all agents find it a strict best response to play according to the convention, for all possible off-path posterior point estimates; otherwise there would be PBSs with arbitrarily small positive probabilities of deviation which were affirmable for sets in the class but not included in a sufficiently small neighborhood of the class, violating local completeness.

But, as η approaches zero, posterior point estimates of $\sigma(L_1, H_1)$ arbitrarily close to 1 are possible for agents – σ with such values of $\sigma(L_1, H_1)$ are in the support of their beliefs, and their subjective probability cannot be reduced by observation when only the conventional path is observed. But, an agent who believed $\sigma(L_1, H_1)$ was close to 1 – that is, that an opponent who saw them make high initial effort would very likely make high second-stage effort – would instead want to deviate. So $\mathcal{C}_{LL}$ cannot be a locally complete CAE class for $\epsilon = 0$, as PBSs involving deviation would thus be affirmable for it.

I now consider the other, $\epsilon > 0$ case. By the 'same possible actions' condition all $C \in \mathcal{C}_{LL}$ must contain some σ with $\sigma(\emptyset) > 0$ if one of them does. Since all $C \in \mathcal{C}_{LL}$ are convex, the same must be true of some open subset of C, which must have positive probability under the prior. Thus all agents with beliefs supported on any $C \in \mathcal{C}_{LL}$ must believe deviations may occur with positive probability.

If there were to be some CAE set $C \in \mathcal{C}_{LL}$, both strategies which play along with the convention and those which deviate from it must be best responses to some possible posterior for agents with beliefs supported on C; if this were not true, σ with positive probabilities of both deviation and non-deviation – which we have established are contained in any $C \in \mathcal{C}_{LL}$ – could not be affirmable for C.

It is necessary for this, in turn, that C contain σ with $\sigma(L_1, H_1) > 0$ and σ with $\sigma(L_1, H_1) < 1$; that is, it is possible in C for a player whose opponent deviated to high effort in the first round to respond with high or low effort in the second round. If such a player could only respond with L_2, deviation could not be a best response to any belief, since a player deviating in such a case would get at most $-l + V_L$, while they get V_L from following the convention. If only H_2 was possible, non-deviation could not be a best response to any belief, since a deviator would then get $-l + V_H$ which for the payoff restriction we have assumed must exceed the payoff of V_L from following the convention. (Agents are not completely certain that their opponent will follow the convention, but the probability they will not is bounded by ϵ; since the inequality we have assumed is strict, for any values of the parameters satisfying it there is some choice of ϵ small enough to ensure that both the above payoff comparisons hold.)

Since choosing to deviate by playing H_1 in the first period is costly, it can only be rational for agents expecting a better payoff in the second round than the convention allows; this can only occur if the deviator plays H_2 in the second round (since V_L is the highest possible payoff if one plays L_2). If the payoff restriction assumed is satisfied, deviating and then playing H_2 is subjectively better than the convention if one assigns sufficiently high probability to the opponent playing H_2 (because $-l + V_H > V_L$, and sufficiently high probabilities of H_2 correspond to expected payoffs from such a deviation arbitrarily close to $-l + V_H$). Since deviating and then playing L_2 is worse than following the convention for any belief ($-l + V_L$, the maximal payoff from doing so, is less than the conventional payoff V_L), it is then also necessary, if C were to be a CAE set, that $\sigma(H_1, L_1) = 1$ for all $\sigma \in C$; that is, anyone who deviates plays H_2 subsequently.

But given this restriction, it cannot be rational for agents deviated on to play L_2, so it must also be necessary that $\sigma(L_1, H_1) = 1$ for all $\sigma \in C$. We have arrived at a contradiction. It is necessary that both actions be playable at this history and also that only H_2 is. We conclude that no ϵ-component class of the $\mathcal{C}_{LL}$ can be a locally complete CAE class, since no $C \in \mathcal{C}_{LL}$ can be a CAE set. □

Comparing these two results, there is a sense in which both this convention and the previous are robust with respect to beliefs: if agents are *a priori* certain these conventions are usually played by others, they always play according to them themselves, regardless of the other details of their priors. On the other hand, unlike collective procrastination, the always low convention is somewhat fragile with respect to payoff parameters – it becomes impossible to sustain as soon as there is any possibility of profiting by unilaterally deviating to high effort.

2.4.3 The 'Always High' Convention

I now consider the 'always high' convention. This differs from the previous two in that agents must take a stage-game dominated action in the first stage, which they can only do if they expect a sufficiently higher payoff in the second stage relative to the alternative. Somewhat counterintuitively, this cannot be supported by the expectation that most players will switch to low effort in the second period if their opponent makes low effort in the first. If the difference in value between the more and less ambitious completions in the second round is large enough that the threat of switching from the one to the other would deter low effort in the first round, it must also be the case that someone who deviates to low effort in the first round must be expecting to coordinate on the high outcome anyway, which thus also gives their opponent an incentive to choose high effort in the second round, undermining the convention. (Alternatively, if the possibility of responding to a deviation in this way is ruled out *a priori*, the resulting component cannot be locally complete.)

Instead, strategic uncertainty is needed to deter deviation. If it is possible for players who deviate to low effort to play either action in the second stage, the players deviated on will be unsure of which action to expect and thus also able to respond with either action, depending on the details of their priors. The possibility of both responses in turn also ensures it is possible in the first place for deviators to play either second-stage action, also depending on the details of their priors. This means that an agent thinking about whether to deviate faces a threat of miscoordinating in the second round if they do so. The threat is credible not because any player would wish to carry it out once they have arrived at the second stage, but because they lack the information they would need to avoid doing so.

If all agents believe miscoordination is too likely for deviation to be worthwhile, in expectation, the whole population can still play according to the convention in this case. But, there is no reason to expect all agents to have such beliefs; presumably most populations would have at least a few people who are subjectively confident, for whatever reason, that they can predict their opponent's response. In seeking LCCCs for this convention, we thus need to consider populations that may have small numbers of deviators. Agents whose priors strongly incline them to deviate will do so if they do not observe samples convincing them that deviation is not worthwhile. We thus need to allow for frequencies of deviation large enough to provide sufficient (discouraging) information about the outcome of deviation that no larger frequency of agents will want to deviate after observing the associated sample distribution (considering only frequencies smaller than this would allow PBSs with marginally higher frequencies of deviation to also be affirmable, violating part (ii) of the local completeness condition). In the CAE set within such a class that we construct, deviation will be rare, in the sense that it is observed seldom enough to allow agents to remain sufficiently uncertain about its outcome; the constructive part of the proof is thus close to that of the previous, with some modifications to account for the rare deviators and outlier samples.

Say that $\mathcal{C}_{HH}$ is an ϵ-component class for HH and $\epsilon \geq 0$ if it is a candidate class, and for all $C \in \mathcal{C}_{HH}$ and all $\sigma \in C$, $\sigma(\emptyset) \geq 1 - \epsilon$ and $\sigma(H_1, H_1) = 1$. The next three results show that some such $\mathcal{C}_{HH}$ is a locally complete CAE class for some payoff parameter values but not others.

Proposition 16. *If $g < V_H - V_L$, then for all sufficiently small $\epsilon \geq 0$ and all η sufficiently close to zero, no ϵ-component class $\mathcal{C}_{HH}$ for HH is a locally complete CAE class.*

Proof. The proof of this result mirrors that of the negative result for 'always low'.

If $\epsilon = 0$, 'empiricism' requires unrestricted play at unreached histories, and the absence of nearby affirmable PBSs requires a strict preference for playing according to the convention for all possible beliefs, and these contradict each other (since agents expecting the high outcome with high confidence in the second period even if they deviate will want to do so).

If $\epsilon > 0$, we need both actions to be possible at the history where one's opponent has deviated to make both deviating and non-deviating potential best responses, but deviation is possible only for agents who intend to play H_2 in this case (since deviating and then getting V_L is worse than playing according to the convention for these payoff parameters), so we also need H_2 to be the only possible response in order for those deviated on to be best responding. This too is a contradiction, and the two cases are exhaustive. □

For the next two results, we will need also conditions on the worst-case second-stage expected payoff, denoted $\underline{V} = \frac{V_H V_L}{V_H + V_L - e}$, which is achieved for the posterior point estimate making an agent exactly indifferent between the two second-stage actions. (Note that $\underline{V} < V_L$.)

The following is trivial from the fact that any strategy playing H_1 in the first round is strictly dominated under the condition stated (since the first-round gain from deviating, g, is assumed greater than the difference between the largest and smallest possible second-round payoffs, V_H and $\underline{V}$).

Proposition 17. *If $V_H - \underline{V} < g$, then for all sufficiently small $\epsilon \geq 0$ no ϵ-component class $\mathcal{C}_{HH}$ for HH is a locally complete CAE class.*

The remaining case is when g is between these two extremes.

Proposition 18. *If $V_H - \underline{V} > g > V_H - V_L$, then for some $\epsilon > 0$, for all η sufficiently close to zero, and for all N sufficiently large relative to ϵ and η, there exists an ϵ-component class $\mathcal{C}_{HH}$ for HH which is a locally complete CAE class.*

Proof. Consider the class $\mathcal{C}^{\epsilon}_{HH}$ consisting of the sets $C_{q,q'} = \{\sigma : \sigma(\emptyset) \geq 1-q, \sigma(H_1, H_1) = 1; \sigma(H_1, L_1), \sigma(L_1, H_1), \sigma(L_1, L_1) \in (q', 1-q')\}$, for $q \in (0, \epsilon], q' \in [0, \epsilon]$. I will show that $\mathcal{C}^{\epsilon}_{HH}$ is a locally complete CAE class for some $\epsilon > 0$. I first show that one of its elements (with q sufficiently close to zero, and q' sufficiently large relative to q) is a CAE set, and then show that there are no σ with $\sigma(\emptyset) \in [\epsilon, \epsilon + \epsilon')$ for sufficiently small ϵ' which are affirmable for any C_q, which is sufficient for $\mathcal{C}^{\epsilon}_{HH}$ as defined to satisfy condition (ii) of the definition of a locally complete CAE class.

I first show that for q sufficiently small, and q' large enough relative to q, $C_{q,q'}$ is a CAE set. This requires showing that, for a generic $\sigma \in C_{q,q'}$, σ is affirmable for $C_{q,q'}$. I will show this by showing the existence a collection of pure reaction functions which are best responses to certain kinds of allowed priors over $C_{q,q'}$, and showing that, for each σ, the convex hull of the strategy distributions $p(\sigma, r)$ induced by σ and each of these pure reaction functions r contains a strategy distribution which induces σ, which is sufficient for σ to be a fixed point for a mixed reaction function (with the same weights) supported on these r, and thus for σ to be affirmable.

The prevailing PBS is pinned down by the probabilities assigned by the prevailing strategy distribution to the 'reduced strategies' which specify an initial action, and specify an action at each of the two second-stage histories reachable given this initial action (these are technically equivalence classes of full strategies in the game). There are six such reduced strategies which can have positive probability within the component we consider – two which play H_1 initially, and then specify a choice of H_2 or of L_2 against an opponent who deviated (played L_1) initially, and four which play L_1 initially (that is, deviate), and then specify a choice of either H_2 or L_2 at each of the two second stage histories consistent with this. It will be convenient in what follows to make a unique choice of strategy distribution (over these reduced strategies) to associate with each PBS σ, from among the strategy distributions inducing it. Let this be the unique (among those that induce σ) distribution over reduced strategies for which the specified choices of actions at each history are statistically independent (that is, e.g. the frequency of agents who deviate and would play H_2 at either subsequent history is the product of the total frequency of deviation with the frequencies of H_2 at each of these histories separately; play of H_2 at one of them is not correlated with play of H_2 at the other, conditional on deviation).

I now show how to construct allowed priors for each of these six strategies in such a way that an agent holding such a prior would play the prescribed strategy 'almost always' in a population where observed deviations were sufficiently rare. Note that any prior which is the product of some density over $\sigma(\emptyset)$ with densities over the probability of H_2 at each of the three unconventional second-stage histories, $\sigma(H_1, L_1)$, $\sigma(L_1, H_1)$, and $\sigma(L_1, L_1)$ is an allowed prior if each of these densities is bounded between $\eta^{\frac{1}{4}}$ and $\frac{1}{\eta^{\frac{1}{4}}}$. Note also that, for each of the marginal densities on second-stage histories and for small η, there exist allowed priors putting the ceiling density $\frac{1}{\eta^{\frac{1}{4}}}$ on some interval with left endpoint $q^{'}$, some interval with right endpoint $1 - q^{'}$, and some interval centered on the worst-case probability associated with $\underline{V}$; and which place the floor density $\eta^{\frac{1}{4}}$ everywhere outside the given interval. The length of each such interval is uniquely determined by η, and approaches zero as η does. Since $q^{'}$ is assumed small, the intervals at the left and right ends of the range are in particular inside the range for which H_2 or L_2 is a strict best response for η small enough, and the interval centered on the $\underline{V}$ probability is strictly inside the range for which deviating initially is strictly suboptimal (in the case of $\sigma(H_1, L_1)$, specifically).

Thus, by considering priors which are products of such marginal densities we can make each of our six reduced strategies an *ex ante* best response, by choosing a marginal density for a second-stage history with very low or very high probability of H_2 if we wish an agent to play L_2 or H_2 at that history, respectively, and choosing a marginal density for $\sigma(H_1, L_1)$ which is centered on the worst-case miscoordination probability if we wish an agent not to deviate initially (for the extreme marginal densities at $\sigma(H_1, L_1)$, they will instead choose to deviate and then take the corresponding action).

Because the priors considered are product measures, only direct observations of play at each second-stage history can influence beliefs about it, and as the density for this history becomes arbitrarily concentrated on the interval chosen, a correspondingly large number of observations at this history are needed to prevent the posterior point estimate from specifying a probability making the prescribed action a best response.

Consider next the six pure reaction functions associated with some choice of a prior for each of our six reduced strategies. If all samples contained sufficiently few observations of each unconventional history, agents with each of these reaction functions would always play a single strategy, and we could simply choose weights over them corresponding to the weights on these strategies which generate the strategy distribution associated with σ. Because some samples will contain unusually large numbers of deviations, we also have to account for the small fraction of agents with each of these reaction functions who play a different strategy because they observe something unusual.

For any fixed N, as q is taken to zero, the frequency of samples with $n > 1$ or more observations at any unconventional history vanishes relative to the frequency of deviations, for any frequency of deviations less than or equal to q. In particular, the frequency of agents who observe more than one deviation can be made smaller than q'. By the same logic, the frequency of agents who observe more than two deviations can be made smaller than $(1 - \sigma(\emptyset)) \cdot (q')^2$. Since no agent with the priors chosen will, for η small enough, play a strategy different than that associated with their prior unless they observe more than one or two deviations, the strategy distribution $p(\sigma, r)$ induced by σ under the reaction function

r associated with each of these priors has an overall frequency of the strategy associated with this prior strictly higher than $1 - (1 - \sigma(\emptyset)) \cdot (q^{'})^2$, and thus a frequency of each other strategy less than $(1 - \sigma(\emptyset)) \cdot (q^{'})^2$.

Under any $\sigma \in C_{q,q'}$, by definition, the two non-deviating strategies must have frequency at least $q^{'}$ (so that each possible response to a deviator has probability at least $q^{'}$), and the four deviating strategies must have frequency at least $(1 - \sigma(\emptyset)) \cdot (q^{'})^2$ (the total frequency of deviation times the minimum frequency playing a given action at each of the two post-deviation histories conditional upon reaching them; recall that these last two are assumed independent under the strategy distribution associated with σ). Accordingly, 'outliers' alone, for agents who have any of our six pure reaction functions and observations from σ, can account only for a frequency of play of each of the six reduced strategies that is strictly less than that prescribed under the strategy distribution associated with σ.

It remains to show that each $\sigma \in C_{q,q'}$ is a convex combination of the strategy distributions $p(\sigma, r)$ generated by agents with each reaction function r who observe samples from σ. Establishing this will immediately imply that σ is a fixed point for the mixed reaction function whose support is these six pure reaction functions, with the same weights. Since this scheme will work for each such $\sigma \in C_{q,q'}$, this will establish that they are all affirmable for $C_{q,q'}$, and thus that $C_{q,q'}$ is CAE.

To begin, choose an arbitrary $\sigma \in C_{q,q'}$, fix any one of the six aggregate play distributions $p(\sigma, r)$, and fix any convex combination of the other five. By construction, $p(\sigma, r)$ has a greater frequency of its prescribed strategy than the strategy distribution associated with σ does, and any convex combination of the others has a smaller frequency of it. Thus, by the intermediate value theorem, there exists a unique convex combination of $p(\sigma, r)$ and the given convex combination of the others whose frequency of the prescribed strategy exactly matches that under the strategy distribution associated with σ.

Consider the self-map on the space of non-negative weights summing to one over the six $p(\sigma, r)$ (the 6-simplex) which assigns to each $p(\sigma, r)$ the unique weight whose existence was asserted in the previous paragraph, given the (renormalized to one) weights on the others under the input set of weights. It is easy to see that it is continuous. Since the 6-simplex is compact and convex, the Brouwer fixed point theorem applies. Since any fixed point of this map, by construction, gives the weights of a convex combination of the $p(\sigma, r)$ which is equal to the strategy distribution associated with σ, it follows that σ is affirmable for $C_{q,q'}$.

Thus, $C_{q,q'}$ is a CAE set, satisfying part (i) of the LCCC definition.

It remains to show that, for each $C_{q,q'}$, $A(C_{q,q'})$ is either contained in $C_{\epsilon,\epsilon}$ (which, given the nested structure, is equal to the union of all $C_{q,q'}$) or is distance at least some $\epsilon' > 0$ from $C_{\epsilon,\epsilon}$. Let ϵ be large enough that most agents have a sufficiently large subsample of actions at (H_1, L_1) to force their posterior point estimate of play at that history to be within some tolerance of the subsample average. This can be ensured to any tolerance with as small an ϵ as we like, by choosing N sufficiently large.

If ϵ is sufficiently small, the choice of whether to deviate or not reduces to a payoff comparison of the convention against the expected outcome when one deviates and one's opponent does not (that is, the possibility of one's opponent deviating is negligible). It is a best response to deviate and play L_2 (resp. H_2) if the probability of one's opponent playing L_2 (resp. H_2) is sufficiently high, and to not deviate when the probability of one's opponent playing H_2 is intermediate.

If one's posterior point estimate is constrained to be close to the subsample average, it can for a given subsample occupy at most two of these regions. In particular, no subsample of sufficient size is such that one agent can deviate to H_2 after seeing it and another deviate to L_2. Note in connection with this that for sufficiently large N, by the law of large numbers, most agents observe subsamples with average close to the true probabilities at the post-deviation history.

Assume for a contradiction that there were a σ with $\sigma(\emptyset)$ between ϵ and $\epsilon + \epsilon'$ which was affirmable for some $C_{q,q'}$. By the above reasoning, at least one of the two deviations (deviate and then H_2, deviate and then L_2) is such that almost no agents observing samples generated by σ will play it. If the other deviation is played with frequency at least ϵ, then for sufficiently large N almost all agents will expect the second-stage action associated with this deviation with high probability after observing L_1 in the first round, and match it, and almost all agents, including almost all deviators, will expect this as well. Accordingly, if σ were to be affirmable, almost all deviations would have to end in coordination on one of the two second-stage coordination outcomes. But if this were true, almost all agents would see it and thus would deviate, contradicting that the frequency of deviation under σ is no greater than $\epsilon + \epsilon^{'}$.

Thus we conclude no such σ are affirmable for any $C_{q,q'}$. $\square$

An important property of the LCCC described in Proposition 18, which supports the 'always high' convention, is that the conventional strategy and strategies that deviate from it correspond to different priors. A deviator is someone whose prior is sufficiently weighted toward one action or the other that they think they can predict their opponent's response with sufficient confidence to make deviating worthwhile in expectation. A non-deviator is instead someone who is sufficiently uncertain of the response to cheating that it is subjectively not worth taking the risk. Both types of beliefs are allowed over this CAE set, though a population which really does play the conventional path with high frequency must accordingly contain mostly agents who are too uncertain to risk cheating, through some combination of prior uncertainty and observed miscoordination.

2.4.4 The 'High, Then Low' Convention

Finally, I consider the 'high, then low' convention. For rather obvious reasons, this behavior cannot be enforced by any threat of coordinating on a different outcome in the second round. But can, nonetheless, be enforceable under a threat of miscoordination, albeit under payoff conditions in a sense more stringent than in the previous case: it is now the gap between the worst-case miscoordination payoff and the low effort coordination payoff, rather than the high effort payoff, that must exceed the gain from deviating.

Say $\mathcal{C}_{HL}$ is an ϵ-component class for HL and $\epsilon \geq 0$ if it is a candidate class, and for all $C \in \mathcal{C}_{HH}$ and all $\sigma \in C$, $\sigma(\emptyset) \geq 1 - \epsilon$ and $\sigma(H_1, H_1) = 0$. The proofs of the following are essentially identical to those of their counterparts for the 'always high' convention.

Proposition 19. *If $V_L - \underline{V} < g$, then for all sufficiently small $\epsilon \geq 0$ no ϵ-component class $\mathcal{C}_{HL}$ for HL is a locally complete CAE class.*

Proof. The maximal second-period loss in this case is less than the gain from deviating in the first period if one's opponent makes high effort there. Thus, if one's opponent is sufficiently likely to play H_1 (ϵ sufficiently small), any strategy prescribing play of H_1 oneself is strictly worse than some other strategy for any allowed prior, and thus no σ with positive probability of H_1 can be affirmable. □

Proposition 20. *If $V_L - \underline{V} > g$, then for some $\epsilon > 0$, for all η sufficiently close to zero, and for all N sufficiently large relative to ϵ and η, there exists an ϵ-component class $\mathcal{C}_{HL}$ for HL which is a locally complete CAE class.*

Proof. (sketch) This works for a class $\mathcal{C}^{\epsilon}_{HL}$ exactly paralleling $\mathcal{C}^{\epsilon}_{HH}$ in Proposition 18, above.

Allowed priors and pure reaction functions can be chosen exactly as before to justify each of the six possible strategies for 'uninformative' samples, and each σ in a sufficiently small set within the class can thus be shown to be affirmable.

For ϵ large enough, we can again show that almost all deviators will play one or the other of the second-round actions, and derive a contradiction, as before. □

2.4.5 All Unitary Conventions

The purpose of this final subsection is to prove that the results proved above for agents with beliefs over a single unitary convention are still valid if we consider instead agents with beliefs over all four unitary conventions. This is the main result of the paper, because it ties together all the previous cases and characterizes the play of agents who – realistically – consider multiple conventions possible until they learn from their sample which one they likely face.

For fixed values of payoff parameters, say an element of $\{LH, LL, HH, HL\}$ is a *live convention* if those parameter values are such that the payoff restrictions required for Propositions 13, 14, 18, and 20, respectively, are satisfied (Proposition 13 does not specify a payoff restriction, so LH – 'collective procrastination' – is always a live convention). Conventions which are not live are dead. Let $\mathcal{C}_{LH}$, $\mathcal{C}_{LL}$, $\mathcal{C}_{HH}$, and $\mathcal{C}_{HL}$ be the LCCCs identified in connection with those propositions. Call these the *components* associated with their respective conventions (note that two of these components depend on a parameter ϵ; I will assume it is the same ϵ in what follows). For $\Lambda \subseteq \{LH, LL, HH, HL\}$, let $\mathcal{C}^*(\Lambda)$ denote the collection of all $C \subseteq \Sigma$ which are unions of a collection of $|\Lambda|$ subsets of Σ, exactly one of which is in the component associated with each convention in Λ.

Theorem 3. *Fix values of the payoff parameters. Then for all sufficiently small $\epsilon > 0$, for all η sufficiently close to zero, and for all N sufficiently large relative to ϵ and η, $\mathcal{C}^*(\Lambda^*)$ is a LCCC, where Λ^* is the set of live conventions. Additionally, $\mathcal{C}^*(\Lambda^{'})$ is not a LCCC, if $\Lambda^{'}$ contains a dead convention.*

Proof. I will show that the analysis of each component separately does not, in this case, differ from the earlier one-component analysis in any way that could affect the results. The result then follows from Propositions 13-20.

Note first that no agent at a PBS in the HH component can assign positive probability to being in the HL component or vice versa, and the same is true of LH and LL (because unconventional second-stage actions are never played after the conventional first-stage profile in every case). Thus, agents at a PBS in the component for LH or LL can assign positive probability to being at HH or HL, and agents at HH or HL can assign positive probability to being at LH or LL (when these conventions are live).

All these connections are driven entirely by observation of the two samples associated with LH and LL, respectively – the samples in which all observations are of the LH path or the LL path – which can be extreme outliers for HH or HL. As N increases, the frequencies of such samples under any PBS in the component for HH or HL vanishes. Accordingly, any agent who observes them should believe they are at LH or LL, respectively, with probability that becomes arbitrarily high for any allowed belief.

Accordingly, the change in agents' beliefs when they have priors over $\mathcal{C}^*(\Lambda^*)$ or $\mathcal{C}^*(\Lambda^{'})$, instead of just having priors over the component they are in, is only this: If they are at HH or HL and observe one of the vanishingly rare samples which contains only observations of the LH path or the LL path, they believe, mistakenly but with overwhelming confidence, that they are in the LH component or the LL component (if the latter is a component they consider possible), instead of believing they are observing an extreme outlier. If instead they are at LH or LL, they assign some vanishingly small probability to being at HH or HL (if one or both of these is a component they consider possible), but still remain overwhelmingly convinced they are in the component they in fact are in.

The result then follows from noting that a change of posterior for a vanishing frequency of samples in the HH and HL cases, and a tiny sliver of doubt about the true component in the LH and LL cases, does not invalidate any of the previous arguments. To see this, it will help to group the above results into three categories.

Propositions 13, 14, 17, and 19 involve a sort of dominance argument – knowing the conventional path is usually followed is enough to make it a strict best response either to follow it or not to follow it. Changing the belief at the two extreme samples in question clearly cannot prevent most agents concluding that most others follow the true conventional path in any of these three cases, so these results all hold as before.

Propositions 15 and 16 are based on pseudo-forward induction reasoning about the thought process of a deviator. In the first case, almost all agents observing a deviator will still assign probability very close to one (reduced only slightly by the possibility they are seeing an outlier) to their intending subsequent high effort, and almost all prospective deviators will know they face a population of opponents of which this is true; this is sufficient for the argument to go through as before. In the second case, a vanishing fraction of agents will mistakenly believe they are in a different component than they are, but this will not significantly alter the calculations of a prospective deviator, who expects with very high confidence not to face such an opponent; thus again the argument goes through as before.

In the remaining cases, Propositions 18 and 20, I provided a method of constructing distributions of priors for which each strategy distribution in the set I showed was CAE could be a fixed point for a population where all agents best respond. This construction was such that the distributions of priors needed could be shown to exist without needing to know the strategies they lead to after 'unusual' samples. Accordingly, altering the strategy played at these two extreme outlier samples cannot invalidate the construction; it works as before, for the same choices of η, N, and ϵ. □

2.5 Comparisons With Other Concepts

In this section, I compare the results above to those obtainable with some natural alternative solution concepts, and explain the reasons for the differences. Since the results in this paper focus on cases where most agents play according to a single path, I shall restrict attention in the comparisons to such cases also. It is more natural for some concepts to compare the LCCC results to a 'pure' equilibrium for the alternative concept, where the conventional path is followed with probability one. I choose 'pure' or 'slightly mixed' cases for the concepts below according to what seems the most natural comparison. Where more than one version of an alternative concept is available, I focus on the version(s) that seem most directly comparable to the present work, also.

As in the results above, the decision to play according to the conventional path or not in the below cases usually hinges on a comparison between the payoff on the conventional path and that associated with a unilateral deviation. When the conventional path is known, the question of whether it can be sustained reduces to the question of which beliefs about post-deviation play in the second stage are permitted by a given concept, and whether this implies a different answer to whether/which deviations can be a best response than LCCC does.

It will again be useful to divide Propositions 13-20 into three categories for the purpose of comparison. Propositions 13, 14, 17, and 19 are results based on dominance; knowing the conventional path is usually followed is sufficient for following it (in the first two cases) or deviating from it (in the latter two) to be a strict best response, irrespective of beliefs about play at other histories. Accordingly, these results will continue to hold for any concept in which agents are rational and know the analogue of the 'conventional path' is usually followed. Rationality holds in all the alternative concepts I consider; knowledge of the conventional path fails only for 'unconventional' players in a self-confirming equilibrium.

The other two cases are Propositions 15 and 16, which are based on pseudo-forward induction reasoning, and Propositions 18 and 20, which involve the construction of a particular distribution over priors. I call these the 'pseudo-forward induction' cases and the 'constructive' cases, respectively. The extent of difference between these results and the analogous results for alternative concepts will largely hinge on whether the set of post-deviation beliefs which are possible for those concepts is different, and allows a different set of strategies.

2.5.1 Subgame-Perfect Equilibrium

I will focus on subgame-perfect equilibria (SPE) with a single, pure path of play. Thus, the four cases above in which most, but not necessarily all, agents play according to a given conventional path are compared to candidate SPE strategy distributions in which literally all agents play according to the same path. Since agents know the strategies of others in a

SPE (or act as if they do), the notion, natural for CAE, that agents are uncertain of play at rarely reached histories does not translate well to the SPE setting.

The subgame following each of the four first-stage histories is the same game in every case, and has three Nash equilibria – two pure and one mixed – whose payoffs to both players are V_H, V_L, and $\underline{V}$, respectively. Asking whether a given pure path can be sustained in SPE is thus equivalent to asking whether it has a higher overall payoff than first-stage deviation from it followed by one of these three payoffs (since we can choose any of the three to be the post-deviation outcome). Since both the highest (V_H) and lowest ($\underline{V}$) possible payoffs in the second stage are among the three that can be chosen as the (SPE candidate) outcome following a deviation, a given conventional path can be supported as a SPE so long as there is *any* belief about post-deviation play that makes playing along with it rational. It is immediate from this that LCCC refines SPE in this context, since play of the conventional path in a LCCC must be a best response to some belief(s) about post-deviation play.

The refinement is in fact strict. Specifically, SPE allows the outcomes ruled out by the pseudo-forward induction cases of Propositions 15 and 16. We are free in constructing SPEs to mandate that the V_L outcome (or even the $\underline{V}$ one) be played after a deviation, making deviating to high or low effort, respectively, in these two cases 'not worth it'. Doing so creates a completely arbitrary expectation which is self-enforcing (and thus remains purely counterfactual) once it has somehow been placed in the minds of all agents. This is ruled out in LCCC because it violates the 'local completeness' condition – there is no reason agents should be certain all other agents have such an expectation in the absence of any decisive empirical or *a priori* reason to be certain of this – and the presence of even a sliver of doubt about the expectations others may have leads to an actual deviator being interpreted as expecting high effort, and thus to an incentive to deviate.

The ‘constructive’ cases of Propositions 18 and 20 are also supportable as SPEs, though the reasoning behind them is different. Since there are no actual deviators in a pure SPE, we need only find beliefs justifying adherence to the path, and assigning the mixed equilibrium after a deviation – for which the worst-case miscoordination payoff $\underline{V}$ is achieved – gives more or less the same belief in SPE that motivated adherence to the path in the LCCCs constructed in these two cases.[9] In these SPE cases, we take agents to know for sure that this miscoordination danger is present, so that it could not be rational for them to deviate, given the equilibrium strategies. In the LCCC cases, agents were not certain that deviation was not worthwhile; most just had beliefs (potentially as a result of both their priors and some limited empirical information) making them too nervous to try it.

There is an argument against the plausibility of the mixed Nash of the second-stage subgame in a SPE, based on its ‘fragility’. How compelling this argument is in relation to the comparisons made here depends on the extent to which the analogous belief distributions in the LCCCs of Propositions 18 and 20 are ‘less fragile’. This in turn depends on not yet fully answered questions about which distribution-of-priors/strategy distribution pairs induce the CAE outcomes in these LCCCs. It is also not clear e.g. how ‘dynamically stable’ these LCCCs outcomes would be under various ways of operationalizing that question. But, to the extent one accepts on intuition that the mixed expectation is reasonable in LCCC but not SPE, we instead have incomparability: LCCC rules out ‘always low’ and ‘always high’ for the parameter values in the pseudo-forward induction cases, while SPE does not; and SPE rules out ‘always high’ for the parameter values of Proposition 18 and rules out ‘high, then low’ altogether, while LCCC allows both these cases.

[9]Though note that the beliefs assigned to deviators in these LCCCs could not coexist with the beliefs given to non-deviators within the same SPE, since the disagreement among such agents about the true strategy profile violates the condition that they know the true strategy profile in a SPE. The assumption that there are no deviators in the SPE analogues for these cases is necessary here.

To summarize: SPE with the second-stage mixed Nash allowed differs from LCCC in allowing 'always low' for all parameter values, not just $l > V_H - V_L$, and in allowing 'always high' for $g < V_H - V_L$. SPE without the second-stage mixed Nash also differs from LCCC in not allowing 'high, then low', and in not allowing 'always high' for parameter values $V_H - V_L < g < V_H - \underline{V}$.

2.5.2 Non-Strategic Bayesian Inference ('Sampling Equilibrium')

Next, I consider a case identical to the current model, except that agents have priors over all of Σ, with densities bounded between multiples η and $\frac{1}{\eta}$ as before. We will take our predictions in this case to be just the set of strategy distributions affirmable for Σ and these allowed priors. This is, in essence, a modified version of the current model in which strategic reasoning – embodied in the requirement that priors be supported on CAEs – has been stripped out. It is more or less identical in spirit, if not exactly the same formally, to the Bayesian inference variant of sampling equilibrium in Salant and Cherry (2020).

In studying this case, I consider the parameter limit where η approaches zero (the bounds on priors become very loose) and N becomes large relative to η, and I focus on strategy distributions in which all but at most fraction ϵ of agents play according to a given 'conventional path', where ϵ is small relative to N. This is the same parameter limit and class of strategy distributions considered in the results about possible LCCCs, above. The key observation to make about this case is that agents with such beliefs and samples facing such a strategy distribution will have very high confidence that their opponent will in fact play according to the given conventional path (because N is large relative to η), but have more or less unrestricted posteriors about off-path play (ϵ being sufficiently small relative to N implies their subsample of play at all off-path histories is also small, and η is assumed to be small, so that their posterior point estimates need not be significantly restricted by small subsamples).

The four cases involving only beliefs about the conventional path (Propositions 13, 14, 17, and 19) are thus the same in this case as they were for LCCC and SPE; under all three concepts agents know the conventional path is usually followed and that is decisive by itself in all these cases.

The pseudo-forward induction cases are again not ruled out in this case, however. Since beliefs off-path are unrestricted in this case, it is possible in particular for all agents to expect payoff V_L in the second stage after a deviation, and if all agents held such beliefs they could play according to the 'always low' or 'always high' path for the parameter values of Proposition 15 and Proposition 16, respectively.

The two 'constructive' cases of Propositions 18 and 20 are again supportable here, albeit for a different reason than in either of the previous two cases. Since off-path beliefs can be completely arbitrary in this case, we can in particular assign beliefs to all agents inducing the expected payoff $\underline{V}$ after a deviation, as in the proofs of the above propositions. Unlike in the SPE case, we can also assign a small number of them beliefs motivating deviation as well; indeed, we will need to if strategy distributions involving a positive frequency of deviators are to be fixed points.

To summarize, this 'sampling equilibrium' case differs from LCCC in allowing 'always low' for all parameter values, not just $l > V_H - V_L$, and in allowing 'always high' for $g < V_H - V_L$.

2.5.3 Self-Confirming Equilibrium

In a self-confirming equilibrium (Fudenberg and Levine, 1993), all agents have correct beliefs about the play they will face at histories they can reach when playing their own strategy against the prevailing strategy distribution in the population. In particular, agents who play in a way compatible with a given 'conventional path' will know that most other agents also do so, and will also know the true distribution over the second-round actions of rare deviators, if first-round deviation occurs with positive probability. Rare deviators from a given conventional path will know the first-round action distribution and the true distribution over responses to a first-round deviation, but may not know the true conventional path (i.e. the second round action distribution after both agents play 'conventionally' in the first round). Deviators may not know the behavior of other deviators, and non-deviators may not know the response of other non-deviators to deviators. Since all agents best-respond to beliefs that are correct about the distribution of play given their own strategy, play at every second-stage history reached with positive probability must correspond to a Nash equilibrium – if it did not, some of the agents who reach that history and thus know the distribution over their opponents' play there would not be best responding – but this restriction need not be reflected in the beliefs of agents.

There is a stronger argument against the mixed Nash as a second-stage outcome in the case of SCE, since SCE represents the limiting outcome of a process of learning about the consequences of one's own strategy and the mixed Nash of a 2×2 coordination game tends to be dynamically unstable under the sorts of learning processes one might imagine this outcome being derived from. Because of this, and for simplicity, I will focus on cases where pure Nash outcomes are played at reached second-stage histories. Note that this does not prevent agents from having mixed beliefs about histories they do not reach themselves. It is natural, in comparing SCE to LCCC, to consider both SCE where literally all agents follow the conventional path and SCE where a small fraction do not. I consider these cases in turn.

In the case where literally all agents follow a given conventional path, they know that all agents do so, and have beliefs which are otherwise unconstrained. Accordingly, the beliefs they can hold are precisely the same as in the preceding case, of 'Bayesian sampling equilibrium' in the parameter limit paralleling the analysis in this paper. The differences with LCCC are thus the same as well: 'always high' and 'always low' are allowed for the parameter values of Propositions 15 and 16 under 'pure' SCE but not LCCC.

In the other case, where there are some deviators, the allowed beliefs for non-deviators are not importantly different – they still believe with near-certainty their opponent will follow the conventional path if they do, and still do not know the distribution of play they would face if they deviated, and so it is possible for them to follow the path in a SCE under the same parameter conditions as before. As noted above, the play after a deviation will correspond to one of the two pure second-stage Nash equilibria. So, what we need to determine is in which cases a small fraction of deviators can find it optimal, for beliefs correct about the consequences of their own strategy, to deviate in the first round and then play according to one of the pure second-stage Nash equilibria.

The two 'dominance' results from the LCCC case which ruled out play of a given conventional path – Propositions 17 and 19 – will still hold in this case, since non-deviators know the conventional path and would thus know that they are getting a payoff worse than any possible result of deviation. But since the deviators in this case do not know the conventional path, it is not guaranteed that the two 'dominance' results ensuring all agents play a given conventional path – Propositions 13 and 14 – hold in this case. Since a deviating agent has correct beliefs about the consequences of deviating, they may deviate in a case where agents knowing the true conventional path cannot rationally deviate only if their belief about the value of following the conventional path is incorrect in a pessimistic direction. Since their belief about play after the conventional first-stage history is unrestricted, it is without loss, from the point of view of whether deviation from a given conventional path is supportable, to set the subjective value they attach to it at $\underline{V}$ in every case. Since their payoff after a deviation is maximized if the population coordinates on high effort after one player deviates, it is also without loss to assume this is true. If deviation from a given conventional path can be supportable in SCE at all, it will be supportable for this belief about the conventional path and this objective post-deviation play. From this, it is immediate that we may have SCE in which most, but not all, agents play the 'collective procrastination' path if $l < V_H - \underline{V}$ and in which most, but not all, agents play the 'always low' path if $l < V_L - \underline{V}$ (the interpretation in both cases being that their stage-game loss from deviating is less than their imagined second-stage gain from deviating). This is the only case in which deviation from 'collective procrastination' is possible.

In the psuedo-forward induction cases, any actual deviators must make high effort. Since the actual play of deviators must be known in SCE to non-deviators, actual play after a deviation must also be coordination on high effort in SCE. Even so, non-deviators may still believe that they would receive the miscoordination payoff $\underline{V}$ from deviating, so we cannot rule out play of the 'always low' and 'always high' conventional paths for the parameter values of Propositions 15 and 16, even if deviations happen with positive probability. This is different from LCCC, but matches the case of SCE without deviators.

Finally, in the constructive cases of Propositions 18 and 20, we can support the 'always high' and 'high, then low' paths as before when there is a positive fraction of deviators, but with the new restriction that play after a deviation always coordinates on either low or high effort. Thus, there cannot be actual miscoordination, but non-deviators may still fear miscoordination because they are uninformed about how other non-deviators respond to deviation; the presence of an objectively superior deviation, whose actual payoff is known to the deviators, need not induce them to switch to deviation as well.

The key property of SCE involved in much of the foregoing is that an agent who chooses a rare deviant strategy can remain wholly ignorant about what happens to agents who choose a conventional strategy; somehow they never manage to learn about the outcomes experienced by the vast majority of other agents around them. This is, frankly, hard to motivate in a setting like the informal 'many people work on joint projects within a large organization' story I have told about the game in this paper, if we imagine these agents never hear reports about what happens to others; it may be easier to motivate if we add features to the story that make these reports subjectively unreliable, because of e.g. a lack of trust. The condition that non-deviators exactly know the behavior of deviators even when deviation is very rare also requires these agents to have played a very large number of times if it is to be a good approximation to what they have learned. On the other hand, in fairness, the ignorance about one's own strategy inherent in CAE may be implausible in a setting where agents repeatedly play an unconventional strategy even a modest number of times and learn from their own experience. And there are clearly settings where a concept in which the information available to agents takes a form which is somewhere between these two extremes would be plausible; such an approach is left to future work.

In summary: SCE with zero deviations exactly matches the preceding 'Bayesian sampling equilibrium' case. SCE with positive probability deviations differs from LCCC in allowing small numbers of deviators from the 'collective procrastination' and 'always low' paths in some cases, in allowing the 'always low' and 'always high' paths to be followed for the parameter values of Propositions 15 and 16, respectively, and in requiring post-deviation play to involve objective coordination on one of the pure second-stage Nash outcomes in the 'constructive' cases of Propositions 18 and 20.

2.5.4 Rationality and Common Strong Belief in Rationality

In considering the outcomes that would be predicted by rationality and common strong belief in rationality, or RCSBR (Battigalli and Siniscalchi, 2002), it is natural to restrict attention to cases in which all agents believe a given conventional path is followed at least most of the time, as the analogue of agents having observed this in the LCCC case. Following Battigalli and Friedenberg (2012), the outcomes consistent with RCSBR and this additional restriction are given by the Δ-rationalizability procedure of Battigalli and Siniscalchi (2003), for a subset of beliefs about the opponent's strategy that place initial probability at most ϵ on strategies not consistent with a given conventional path (where 'beliefs' in this case are conditional probability systems ala Battigalli and Siniscalchi (2002)). Starting from the set of all strategies, the algorithm iteratively deletes strategies which are not a sequential best response to some belief satisfying both the given restriction and strong belief in the subsets of strategies at each previous stage of the algorithm.

In the cases of Propositions 13 and 14, where knowing most others play according to the conventional path – or, translated to this case, knowing one's opponent plays according to it with initial probability close to one – is sufficient to make doing so oneself a strict best response, the first stage of the algorithm clearly eliminates all and only those strategies which do not play according to the given conventional path. Since this same collection of strategies can all be best responses to beliefs assigning probability one to remaining on the conventional path whenever play has not yet gone off it (this is strong belief in the strategy subset surviving the first round, which implies and is thus consistent with the initial belief restriction), the algorithm terminates after one round, and we are left with a non-empty predicted set of strategies any pair of which play the conventional path when facing each other. This, translated to the language of this concept, 'matches' the LCCC result: most agents playing these paths is possible for the given parameter values, and agents must do so with probability one if most do so.

In the cases of Propositions 17 and 19, where playing according to the conventional path is worse than any possible outcome of not doing so, the first round of the algorithm eliminates all strategies that can play according to the conventional path. Strong belief in the complement of the set of strategies that do so contradicts the initial belief restriction, so we are left with an empty set of strategies. This, also, matches the LCCC result, translated to the RCSBR context – such paths cannot be played by most agents for the parameter values in question.

In the pseudo-forward induction cases, we, perhaps not surprisingly, derive an actual forward induction argument along the same lines in the RCSBR case. The order of moves is the same as in the LCCC case. Given the initial belief restriction, we eliminate strategies that deviate and then make low effort in the first round, while strategies that deviate and make high effort survive. This leads to strategies that respond to deviation with low effort being eliminated in the second round (they cannot be best responses to agents who strongly believe in the first-round set), and thence to elimination of strategies that follow the conventional path in the third. Agents who strongly believe the latter restriction – no possible strategy follows the conventional path – cannot satisfy the initial belief restriction, so we are again left with an empty set of strategies in the fourth round. This again matches the LCCC case: these paths cannot be followed for the parameter values of Propositions 15 and 16.

Finally, in the 'constructive' cases of Propositions 18 and 20, the first round eliminates strategies that do not play according to the conventional path in the second stage conditional on both players doing so in the first stage, but nothing else – we can choose beliefs supporting each other strategy in essentially the same way we did in the proofs of those propositions, all of which satisfied the initial restriction imposed here.[10] Since all of these beliefs placed probability one on the second stage of the conventional path being followed if the first was, they are also consistent with strong belief in this first round restriction. Thus, the algorithm terminates after the first round. This allows the outcomes of Propositions 18 and 20, with the same set of deviations allowed to occur with positive probability. But it is weaker than these results for the LCCC case, in that it does not require deviation to be rare. Thus, RCSBR differs from LCCC in this case by being more permissive, because the requirement that the play of agents who believe in the initial belief restriction reproduce the kind of observations which generated the initial belief restriction (this is, essentially, 'local completeness' – the second part of the LCCC definition) is absent.

[10]Note also that all of them placed positive probability on every first-stage outcome, so in focusing on these beliefs we need not worry about the general differences between standard Bayesian priors and conditional probability systems.

In summary, RCSBR differs from LCCC only in allowing a wider range of outcomes in the 'constructive' cases of Propositions 18 and 20. This difference in outcomes, and the difference in interpretation of other cases, is related to the fact that the LCCC framework provides an explicit account of what agents observe, while the framework of the analogous RCSBR case – even with an initial belief restriction inspired by what is observed in the corresponding LCCC cases – does not.

2.5.5 Distinguishing Between Concepts Empirically

The differences between the predictions of the various concepts discussed above and those of LCCC can also in principle form the basis for testing between these concepts empirically. The most natural setting would be to somehow get data on behavior within an actual large organization in which many small-scale projects of the same kind are regularly undertaken, and where this has been going on for long enough for behavior in any given project to be influenced by expectations derived from the sorts of behavior which have characterized work on past projects, which people within the organization presumably hear about. The crucial requirement is that there be a way to measure the level and trajectory of effort (i.e. which of the pure paths in the game we should interpret a given outcome as) and a way to measure the payoff characteristics of the game, so we know which of the key parameter inequalities applies (one could look, for example, at opportunity costs associated with higher effort, if both an individual's effort on the project in question and their performance across other job tasks were observable). For given payoff conditions, some observed outcomes will falsify some concepts and support others (for example, seeing low effort throughout when $l < V_H - V_L$ is evidence against LCCC, but consistent with SPE).

A complication is that some of these concepts are tied to specific notions about where agents get their information, which may or may not be observable. LCCC and 'sampling equilibrium' both presuppose hearing about what happened to others but limited personal experience, while SCE presupposes extensive personal experience but little or no learning from hearsay. SPE and RCSBR are more opaque, though motivating stories could be found. Accordingly, if one knew the amount and nature of information about one another individuals in a given organization had, one might restrict attention to testing the concepts intended to explain behavior for such information structures, to see whether their predictions coincide with actual behavior of agents who get their information in such a way. If the information is not known, the exercise becomes instead a joint test of whether the concept is good and whether the information structure in the data is appropriate to the model.

Beyond these specific comparisons, there is a distinctive reasoning process inherent in CAE and LCCC which one might attempt to test in general. I briefly discuss three possible methods for doing so – belief elicitation, experimental replication of the 'entering a world in equilibrium and wondering which it is' environment, and 'story elicitation'.

In the first case – belief elicitation – one could approach agents in some real-world setting whose features fit the CAE story (a large number of agents engage in some common sort of strategic interaction with other agents, they engage in such interactions only a few times themselves but are well-informed about what happened to others in the same situation, and such interactions have been happening regularly for a long time), and measure by some reasonable method what probabilities they assign to various actions of their opponent (according to some way of operationalizing the available 'actions' in a real-world 'game'). To the extent one fixes a game to describe the situation and an LCCC consistent with the observed distribution of play in the game, there will be definite predictions about these beliefs that are often stronger than could be obtained by merely inverting the best response correspondence (in 'collective procrastination', for example, agents should believe literally everyone will follow the convention, which is stronger than needed for it to be one's own best response).

The background story of CAE according to which a game has been played by many people, and newcomers have heard and formed expectations based on reports of past play, over a long period of time, cannot be directly mapped into a single experiment, where time and scale are necessarily limited. But, there are some imperfect workarounds. One might recruit from some existing population already believed to play according to some potentially CAE outcome 'in the wild' and have them play the same game in the laboratory, maybe stressing at the beginning that it is the same game they play in other settings and that their partners are from the same population they encounter in those settings. This is, arguably, a direct sample of play from the existing possibly-CAE outcome, and one can assess the likelihood it really came from a strategy distribution which is CAE. Or, in the case of games which have already been played in many previous experiments, one might inform agents at the outset of the distribution of outcomes in past iterations of the same experiment they are about to participate in (or maybe about outcomes for subjects 'similar to you' – the 'same population', in a more abstract sense), and see if agents react to this in a way consistent with the reactions agents in a CAE might have to a comparable sample. Or, one might randomly match agents in an artifically enlarged population containing the subjects and a much larger number of computerized opponents pre-programmed to play along with a given CAE, and show agents a sample of the play of the computer agents against each other before the agents play, explaining that they will typically encounter such opponents but may occasionally encounter each other (and won't observe whether their opponent is human or not). For many CAE outcomes, this infusion of new agents with a potentially different prior distribution need not preserve the equilibrium, but the direction of departure should be predictable. Alternatively, one might elicit beliefs at the start of the experiment and then ensure the CAE played by the computers is a fixed point for the distribution of priors among the subjects (so that the CAE ought to reproduce itself even with the new entrants, up to some noise associated with the finite population size).

Finally, one might take a more qualitative approach, asking agents to tell stories about the kinds of behavior they expect others to engage in in a given environment, and about what agents who act in a given way are likely to be thinking to motivate them to do so, and see whether the descriptions are consistent with CAE. One can ask this about a real world setting the agents have experience with, or a hypothetical setting; the latter has the disadvantage of being more artificial but the advantage of being more controllable. In the hypothetical case, one may also ask for an *ex ante* or *ex post* answer – with or without being given a 'sample' (or just a description) of behavior in the population being considered. Conformance with CAE/LCCC only really requires having posterior beliefs of the kind predicted – a population of agents who arrive at such posteriors would always behave in the ways predicted, even if their mental states were totally different than the model describes before they observe their sample, a point of particular importance with respect to LCCCs in games where the number of possible 'conventional paths' is very large – but the *ex ante* answer would also be interesting. In addition to revealing whether subjects get the 'right answer', this exercise could be illuminating on a number of other points. Do subjects typically try to reason in terms of other people's observations and beliefs about others' strategies, even if they reach different conclusions? If their descriptions fit with some CAE set, is it generally 'complete' (i.e. LCCC) or does it often end up describing a smaller CAE set? If so, which one(s), and is this an oversight or something subjects stick to even when prompted to consider the 'missing' possibilities? Are some of the factors relevant to CAE/LCCC – thinking about people's observations, beliefs about others' beliefs and strategies, payoffs/optimization, or the possibility of sustaining both deviation and non-deviation when deviations are common enough to be often observed – more often considered than others, and are factors irrelevant to CAE/LCCC sometimes considered? And how does all this depend on the characteristics of the subject, the game, and the setting? This sort of qualitative investigation would compliment the more quantitative approaches, and provide guidance about when the kind of reasoning postulated in a CAE is a good match to how agents actually reason. In the event that the stories told about hypotheticals were less in accordance with CAE than stories about situations agents have real-world experience with, we would also have evidence that such reasoning emerges in equilibrium more readily than in the mind of any given agent in the abstract (perhaps because speculation about the range of possible beliefs others might have and the strategies that might result gets around through the same channels as observations in the samples do; this sort of crowdsourced strategic reasoning would seem plausible on grounds of casual empiricism).

2.6 Discussion

2.6.1 Payoff Parameters and Live Conventions

The results above constitute a set of predictions about which of our four unitary conventions might arise for certain values of the parameters of the game. Let us now take stock and interpret these results in light of our original story.

The collective procrastination convention is, as we have seen, the most robust, being a live convention for all payoff parameters and being consistent with all allowed prior beliefs. We predict that, in any organization where this pattern of first low effort and then high effort has become widespread – and thus widely observed – all agents will fall into conformity with it, regardless of the details of the game (the exact nature of the 'joint project') and of their beliefs, including what other conventions they might have come to conform to upon entering an organization with a different history.

The 'always low' convention is similarly robust with respect to beliefs if the cost of initial effort when one's opponent slacks initially is higher than any possible gain. As soon as such a gain is possible, however – which can only mean the deviator hopes to coordinate on high second-stage effort and thus to make a high effort in both rounds – the convention becomes untenable. The threshold between these cases involves a comparison between the loss l from making high initial effort when one's opponent makes low effort (the amount by which the cost of effort exceeds the value added of only one player making an initial effort), to the difference in value between the two second-stage coordination outcomes. When the former exceeds the latter – when the cost of first-round effort is relatively large compared both to the value added from one-player first round effort (the other factor determining l) and to the difference between high- and low-effort outcomes in the second round – the convention can arise, and indeed will be played by any population which observes it is already widespread.

We can think of the case where the 'always low' convention is live as representing a case where effort in the project in question is ultimately not very important to the outcome. It needs to be done, but no one will be all that much better off because it is also done well. In such a case, the direct effect of exerting high initial effort is mostly downside – more work for oneself, with the improvements ultimately being not so important – and the enticement of a higher quality completion of the project is also not very enticing. Otherwise, we have relatively higher gains from effort – a lower net loss from unilateral initial effort and/or a greater difference in second-stage outcomes – which would motivate agents to find a way out of this convention if they believed themselves to be stuck in it. This convention is thus live in cases where effort is relatively less important. In this sense, it occurs in cases where it is less harmful for it to occur, though it is still inefficient.

The 'always high' convention splits into three cases – one where it is live and two where it isn't, for different reasons. These cases correspond to different values of the parameter g, interpreted as the gain from making low effort in the first period if one's opponent makes high effort (the amount by which the averted cost of effort exceeds the gain from both players making high effort rather than just one). When g is less than the difference between the high and low second-stage coordination payoffs, deviation 'signals' intended high effort, as in the analogous case for 'always low'. When g is higher than the difference between the high second-round coordination payoff and the worst-case miscoordination payoff, making high initial effort is strictly dominated.

We can think of the former case – g lower than the difference between the second-stage coordination outcomes – as representing a case where the value added of high effort (in both periods) is relatively important compared to the cost of first-period effort. The only difference is that in this case the comparison involves the value of first period effort when one's opponent also exerts effort, rather than when they do not. Intuitively, 'always high' is dead in this case because the value of the better second-stage coordination outcome is 'too high to risk'. One must accordingly conclude that if one's partner slacks off at first, they do not subjectively perceive themselves to be risking it – they still expect to coordinate on the high second-stage outcome, and will act accordingly. Such a partner may be irritating, but the only best response is to play along with what they are doing – the loss to both parties of failing to do so acts as a kind of blackmail to prevent a different response. Since both partners can anticipate this response, the convention is undermined.

The other case where 'always high' is dead – when g is too high – involves the risk not of the worse coordination outcome but of miscoordination in the second round. This has a different interpretation. While it is costly to cut corners in an organized way that makes the outcome less good than it could have been, it is more costly to end up working on incompatible things that leave the project unfinished or in shambles at the end. If the gain from avoiding first period effort is so high that even the risk of this loss cannot outweigh it, there is a natural sense in which the temptation to avoid it is too overwhelming to admit any other action.

The case where 'always high' is live puts g in between these two extremes. It is not worth exerting high initial effort just to avoid cutting corners in the second stage, but it is worth doing so if a complete breakdown of the project is too likely otherwise. For agents working on a project of this type, in an organization where making high effort throughout is typical, a partner who slacks off at first is hard to read, because they may be thinking two different things. They may believe their lack of initial effort will just be ignored, so that they can get away with it and coordinate on the high second-stage outcome anyway. They may also believe it will convince their partner they are 'unserious' about the project and trigger a low-effort regime in the second stage, also. If such deviations are unusual, one will generally not have much information about the relative frequency of partners who think in these two different ways. It is for this reason that choosing to deviate in the first place is risky – one's partner may misread one's own intentions, leading to chaos.

Whether 'high, then low' is live also depends on the risk of miscoordination. It differs from the 'always high' case in that it is the comparison between the worst miscoordination expected payoff and the low-effort coordination outcome in the second stage that matters. Avoiding effort in the first round is obviously optimal if one expects effort in the second round to remain reliably low. But, as in the 'always high' case, a partner who deviates can be thinking two different ways, and if such behavior is unusual the difficulty of knowing what they are thinking, and the associated risk of deviating in the first place, can maintain the convention if agents are subjectively unsure what to expect.

'High, then low' ceases to be live if the gain from deviating becomes large enough that it necessarily outweighs the risk of miscoordination in the second round. This happens for a lower value of g than under 'always high', since the loss is relative to the worse coordination outcome in this case. There is no lower bound on g necessary for this convention to be live, however, because there is no case in which certainty of the better coordination outcome can justify deviation but certainty of the worse one cannot. This has an interesting consequence: When g – the gain from initially slacking on a partner making high effort – is sufficiently low, the 'always high' convention cannot occur, but the 'high, then low' convention can. Intuitively, agents who have less to lose in trading the conventional second-round outcome for a different coordination outcome also have more to fear regarding their partner's inability to predict their intentions. This can be effort-promoting, when the inherent cost of effort is relatively minor.

Finally, let us take stock of which *sets* of live conventions are allowed for given parameter values – which possible ways of handling a project with given payoff properties might emerge in different organizations in which projects of this kind are worked on. 'Collective procrastination' is always an option, so the question can be framed in terms of which combinations of the other three can be live or dead.

'Always high' and 'high, then low' both depend on g but not l. As we have seen, 'always high' is live for an intermediate interval of g, while 'high, then low' is live for all g below some lower bound. This lower bound is necessarily lower than the high endpoint of the interval where 'always high' is live; thus there are always some values of g which are too high for 'high, then low' to be live, but where 'always high' is still live. Conversely, as noted above, sufficiently small g will make 'always high' dead but 'high, then low' live. The comparison between the lower endpoint for 'always high' ($V_H - V_L$) and the cutoff for 'high, then low' ($V_L - \underline{V}$) is indeterminate. When the former is higher – this happens when the difference in value of the second stage coordination outcomes is large and the cost e of higher second-stage effort (which reduces the net gain from the high outcome) is lower – only one of 'always high' and 'high, then low' can be possible for any project, with the former requiring strictly higher values of g than the latter. In the reverse case, there is a range of g for which both conventions are possible, consisting of the lower part of the range for 'always high' and the higher part of the range for 'high, then low'. Intuitively, when the loss of moving from the high to the low second-stage coordination outcome is large relative to the additional loss from miscoordination, agents will be willing to accept the latter for lower gains from initial slacking than would cause them to (possibly) accept the former. The interval of values supporting 'always high' is also narrower in this case; there is a smaller range of g for which agents accept the worse coordination loss but not the miscoordination loss. In the converse case, the risk of miscoordination is relatively more important, and there are accordingly cases where it can motivate agents to stick to both conventions.

The other convention, 'always low', is the only one whose live or dead status depends on l (and not on g). As such, it can be live or dead independently of the others. We may have only 'collective procrastination', only 'collective procrastination' and 'always low', or either of these subsets together with 'always high', 'high, then low', or both. (in particular, all four conventions are live when g is intermediate and l is large). When multiple conventions are live, each live convention will tend to be reproduced by agents in an organization where they observe it is already widely followed.

There is a parallel, however, between the condition on l making 'always low' dead and the condition on g making 'always high' dead on the grounds that deviation 'signals' high effort – both require that $V_H - V_L$ is large relative to the first-stage parameter l or g. Recall that $l \geq g$ corresponds to a 'convex' map from first-round effort to value added, and $g \geq l$ to a 'concave' one. In the convex case, mutual effort more than doubles the value added of one-player effort, so it is is less costly (and thus more beneficial) to slack when one's opponent is also doing so. This causes 'always high' to be ruled out on 'signaling' grounds for more values of $V_H - V_L$ than 'always low' is. It is inherently more tempting to slack off when the marginal contribution of one's own effort is lower, and the marginal contribution of effort when one deviates in 'always low' is less here, while the loss from deviating from 'always high' is greater (thus undermining the risk of miscoordination for more parameters, since a larger gain is required to be willing to accept the lower coordination outcome in exchange for slacking initially). In the concave case, the marginal contribution of one player exerting effort exceeds that of the other player also doing so. This makes the inherent temptation to deviate from 'always high' greater (and thus creates a risk of miscoordination for more values), while also increasing the temptation to deviate from 'always low'.

Finally, thinking about equilibrium beliefs as well as payoffs, we should note that there is no case in which high initial effort is the most ‘robust’ outcome. It always requires special beliefs that are not guaranteed by knowledge of the convention alone. But, high initial effort is possible in a convention-affirming equilibrium for some payoff values, and when it does occur the presence of the right kind of beliefs in the population is an important part of the explanation for why it has in fact occurred.

2.6.2 Cooperation

In comparing the present work to the previous work on the emergence of cooperation, we should note at the beginning that this literature has tended to focus on indefinitely repeated games, while I focus on a simpler, two-stage game. As such, the conclusions about e.g. which payoff parameters support which outcomes need to be interpreted in light of this difference between games, as well as between solution concepts. The ability to compare a fully known conventional path to an unknown (but one-dimensional and straightforward) post-deviation outcome in the second stage – central to the logic of all the results in the present paper – depends on the fact that play after the initial effort choices consists only of a one-shot coordination game.

If the second stage were to be decomposed further into another extensive game (thus moving in the direction of indefinite repetition, though not necessarily all the way to it), (a) the full distribution of play following an initial deviation would be harder to estimate based on observed deviations, (b) what a deviation ‘signals’ would not be as clear (since it could become ‘I will trick you initially and cheat you in the future’ instead of guaranteeing ‘I hope to coordinate on the high outcome’), and thus (c) whether a cooperative outcome emerges or not is likely to be more prior-dependent, since the conventional path observed is less informative about the full strategy distribution and agents’ deductions from their strategic reasoning about how deviations are likely to play out are less powerful. There is no way to fully work out how much and in what way these differences between games affect the comparisons of results, but they should be noted up front, so that too much is not read into the difference in solution concept alone.

The 'evolution of cooperation' literature is one natural source of comparisons with the present work. It has generally focused on cases where inefficient outcomes will be undermined by mutations (e.g. Binmore and Samuelson (1992), Fudenberg and Maskin (1990)), though there are cases justifying inefficiency (e.g. Volij (2002)) or arguing for continual long-run change in strategies/outcomes (e.g. Imhof et al. (2005), Garcia and van Veelen (2016)) as well. In the present work, there is no comparable case to be made for the 'robustness' of efficiency; 'collective procrastination', with its only partial cooperation, is the 'robust' outcome here, though this is probably partly an artifact of the different game considered. There is an informal sense in which the forward induction-like argument ruling out 'always low' for certain parameters is reminiscent of the 'secret handshake' argument of Robson (1990) for efficiency-promoting mutants – a deviation 'signals' something allowing it to prosper and take over – though the fully efficient 'always high' can be undermined in the same way in the present context.

There is much evidence in experiments and adjacent theoretical work for the importance of the parameters l and g, in particular the importance for cooperation of both being small. Subjects can struggle in experimental settings to find their way to cooperation when they are not (dal Bo and Frechette, 2018). The parameter l can be understood to determine the 'riskiness' of cooperation if the opponent's strategy is uncertain (Blonski and Spagnolo, 2015), which matters in such settings over and above the short-term value of defecting from a cooperative outcome given by g. One key difference between the present setting and the 'cold start' natural in thinking about experiments is that the ability to observe the existing conventional path – and the presumption that one enters a world in equilibrium – removes strategic risk from the conventional path, so that the fear of being defected on if one follows a cooperative convention is no longer a significant consideration. This likely explains why l plays such a limited role in the results of the present paper: Lowering it sufficiently rules out 'always low', but raising it never makes it harder to sustain high initial effort; this is because the high-initial-effort convention is observed, and taken for granted by agents.

In the present context, the fully cooperative outcome 'always high' also requires g to be of moderate size, rather than small. The intuition – that projects which are 'too important to be completed badly' must require high effort from all parties at the end no matter what, and that only when the second round difference is not too important is it possible one's opponent will just 'give up' if you don't play along with the convention – is tied to the different game considered here, which has only a one-shot coordination game after the initial effort choices. It is also tied to the observability of the conventional path: Knowing what one's opponent is risking by deviating, and thus which second-stage actions a belief leading to deviation is compatible with, requires knowing the conventional path they are deviating from.

Finally, we can consider the relative magnitude of payoffs and payoff differences in the first vs. the second stage, which is a proxy for the size of the discount factor in indefinitely repeated games. 'Scaling up' the second stage can only move 'always low' from live to dead, and can only move 'high, then low' from dead to live; both these results are consistent with the notion that greater weight on the future promotes initial cooperation. The 'always high' case is more complicated, with an increasingly important second stage moving it from dead to live, then live to dead again; the intuition in this latter case is just a restatement of that for g, about the 'signalling' properties of cases where the second stage effort is too important not to coordinate on.

2.6.3 Forward Induction

There is in much of the foregoing a kind of forward induction-like reasoning involved in the results, where agents see a deviation by their opponent from the conventional path and infer from this something definite about the action this opponent must intend to take in the second round. This effect is not true forward induction in the sense the literature has considered (see e.g. Kohlberg and Mertens (1986), Stalnaker (1998), Battigalli and Siniscalchi (2002), Battigalli and Friedenberg (2012)), because it does not involve changing beliefs about an opponent after observing a zero-probability event. Rather, agents have ex ante uncertainty over a collection of possible equilibria, some of these have a low but positive probability of deviation, and the set of equilibria considered possible may impose restrictions on which second-stage action the deviators will play (and may need to impose such restrictions in order to be a CAE set).

In a CAE set, all agents deviating in these equilibria are behaving rationally given their beliefs because this is part of the definition of a CAE set, and the interpretation of their behavior by agents with beliefs over some CAE set follows from this. A convention which is to correspond to a CAE set must be playable even for agents who make such inferences, and if it cannot be, it is ruled out as CAE.

There is no case studied in this paper in which agents encounter 'impossible' events which may force them to revise their beliefs about their opponent in the course of play. This is not to say that the analysis could not be extended to examine such cases or that it might not be interesting, only that I do not consider such possibilities here, and it is not necessary to consider them to obtain the pseudo-forward induction effects which arise in the present framework.

2.6.4 The 'Empiricism' Condition

In Propositions 15 and 16, where I ruled out LCCCs for 'always low' and 'always high' when the difference in second-stage payoffs was too large, LCCCs with a positive frequency of deviation were ruled out by pseudo-forward induction arguments. The argument against an LCCC with literally no deviations instead leaned on the 'empiricism' condition. This condition requires that any and all off-path behavior be considered possible if a history is literally off-path (reached with probability zero) and it was then easy to show that such unrestricted off-path beliefs can justify deviations, undermining local completeness. This second step was fairly simple; the deeper issues at work here concern the meaning and justification of the 'empiricism' condition itself.

As an example of what this rules out, consider the singleton distribution in which 'always high' is always played and any deviation is followed by coordination on 'always low' (this is basically 'grim trigger', translated to the present context). It is CAE, as any singleton consisting of a Nash equilibrium would be.

The intuitive problem with such cases is that they involve a sort of prior coordination on certainty of particular strategies derivable neither from observation nor knowledge of the process. In interesting, nontrivial CAE sets, restrictions on what kind of play can occur should be derived from observations and best responses to reasoning about observations – not necessarily about one's own observations, but at least about general facts connecting who observes what, and who thinks whom else might have observed what,...to who can play which strategy.

A case in which a history is *a priori* certain to be literally never reached is a case in which there are no such observations to reason about, and thus no basis for having restricted beliefs about it. If it might be reached, there is some empirical information about it somewhere in the population, so we allow for a much more permissive range of cases, though as we have seen all of these may be ruled out for other reasons.

A final point should be noted in connection with this. The assertion that 'all possible play is allowed' at zero probability histories is unproblematic here because the subgame starting at all unreached histories is a coordination game (so that all second-stage actions are stage-game rationalizable). In a game where there were dominated actions at later stages, we might want to employ a somewhat different assumption to rule out beliefs that assign positive probability to those dominated actions.

2.7 Other Related Literature

The CAE concept in general is related to the concept of sampling equilibrium (Osborne and Rubinstein (2003), Salant and Cherry (2020)), in that both define equilibrium as a strategy distribution that reproduces itself through the behavior of agents who observe some sample from it. CAE differs in studying agents who engage in sophisticated strategic reasoning about each other and about the larger process they are embedded in, rather than the behavior of agents who employ simpler decision rules under various statistical estimation procedures.

CAE is also related to self-confirming equilibrium (Fudenberg and Levine, 1993), especially the 'rationalizable' variants (Dekel et al. (1999), Fudenberg and Kamada (2015), Fudenberg and Kamada (2018)), in that it defines a notion of equilibrium in which agents are partially uncertain of the strategy distribution. It differs in that agents are uncertain of the outcomes of strategies rarely played by *others*, not the strategies they themselves don't play. It differs also from the 'rationalizable' variants in imposing a rationalizability-like restriction on the larger process the game is embedded in, not strategies in the game itself.

Further connections between the literature and CAE in general are discussed in Hudson (2023).

Chapter 3

The Confidence Game: Equilibrium Bargaining under Strategic Uncertainty

Abstract

I study a simple bargaining game, in which populations of firms and workers bargain over whether they will split output according to a high or low wage, and study it from the perspective of convention-affirming equilibrium, a notion of equilibrium for agents who are strategically sophisticated but do not know each other's strategies ex ante. I characterize a natural solution set in which all matches result in immediate agreement on either the high wage or the low wage. In the high-wage case, all agents are able to infer each other's equilibrium strategies from the available information. In the low-wage case, their equilibrium behavior instead reflects their subjective confidence in the demands they can successfully make, which is prior-dependent due to limited information. The results thus highlight how both strong strategic inferences and subjective confidence (or lack thereof) in the absence of such inferences can support convention-affirming equilibrium bargaining outcomes.

3.1 Introduction

Many real-world bargaining situations involve agents who must come to some agreement relatively quickly, and whose bargaining outcomes are thus shaped by their perceptions of their near-term outcomes rather than some (potentially) exhaustive process of exploration. Workers may need to have some job in order to continue to pay their bills; firms may need to maintain a sufficient number of employees in order to continue operations. The assumption of a short, hard time-limit is plausibly a good approximation to reality in many such cases.[1]

If agents inform themselves about what to expect in their own situation by observing or hearing about what happened in the bargaining of others, and if it is typical for those around them to reach immediate agreement when they bargain, they may remain uncertain what would happen if they tried to demand a more favorable agreement for themselves than what is typically agreed upon. Thus, in the absence of some prior understanding of the full strategies of others – which is not easy to motivate in a setting where only initial agreement is observed – it is natural for strategic uncertainty to inform their own choice of bargaining strategy.

This paper is an attempt to understand the equilibrium outcomes which would arise in a market occupied by agents who must make some agreement quickly,[2] and who observe only the typical outcomes around them and thus face strategic uncertainty. Such outcomes can depend on agents' subjective 'confidence' – the probability they assign to a demand for a better-than-usual outcome being accepted, when their observations are consistent with a range of values of this probability. They can also be restricted by 'objective' deductions agents make about their opponent – if one's opponent cannot optimally pursue a given strat-

[1]Hurkens and Vulkan (2015) also consider bargaining in a population of agents with hard deadlines. They consider a more general class of possibilities, from the more traditional point of view of subgame perfection.

[2]It is unclear whether the inability to make a larger number of offers and counteroffers, or to rematch repeatedly, is truly necessary for the kind of effects I consider and results I derive. But the intuition is clearest in this case, and it is the case I consider in this paper.

egy given the typical behavior one observes, and believes they also observe, one may deduce from their rationality that they will not, and act accordingly. Both of these possibilities will be important to the results.

I study a simple alternating-move bargaining game played by agents who are randomly matched from two continuum populations (who for concreteness I call *firms* and *workers*). Within each match first the firm and then the worker offer either the high wage or the low wage. If their proposals are the same, the match ends with surplus between them being divided according to that wage. If not, first the firm has a single chance to accept or reject the worker's offer. If the firm rejects, then the worker accepts or rejects the firm's offer. Acceptance in either case also constitutes agreement on the division associated with that wage. Mutual rejection ends the match with no deal. Workers may rematch once in their lifetimes, and care about the (discounted) value of whichever wage they agree to. Firms may not rematch, and thus always prefer some deal to no deal.

I study the outcomes of this game predicted by convention-affirming equilibrium. A convention-affirming equilibrium of the game is a distribution over strategies, and hence over outcomes, with the property that it is playable by a population of agents who believe they are in some convention-affirming equilibrium or other, and make inferences and choices rationally on the basis of such beliefs. There is an obvious circularity in this notion; it is dealt with by taking sets of convention-affirming equilibria (or CAE sets) to be the primitive solution concept. A set of possible strategy distributions is a CAE set if any distribution in it could be the aggregate play of agents who all take this set to be the set of all strategy distributions that are possible. Intuitively, the CAE concept predicts the possible behavior of agents who rationally respond to a correct understanding of the process they are acting within, and believe other agents rationally respond to the same correct understanding, and so on, but do not know exactly which strategy distribution they face; they learn the latter only partially by observing a sample of the outcomes of others' games.

Among CAE sets, I focus on those that are, in a certain sense, maximal. If it were possible to expand a CAE set slightly, by adding more strategy distributions at the 'margin', and arrive at something that is also a CAE set, there would be a sense in which the original was 'unnaturally small' – restricting agents to believe in a smaller set of similar possibilities than they reasonably might. I address this by constructing what I call a *locally complete CAE class* (LCCC for short). A LCCC is a *collection* of (generally 'similarly shaped') sets with the property that one of them is CAE and all of them are such that a population of agents who believe they are the set of possible strategy distributions cannot itself play a strategy distribution which is close to, but not within, the union of sets in the class.

The goal is to establish properties of CAE sets which are the 'largest', and hence the most interesting, within a given region. Since characterizing the largest sets themselves would involve prohibitive technical difficulties, I instead look at classes of sets constructed so as to include all CAE sets in a given region, along with some other sets that are not CAE. In proving that some property holds for all sets in the class, or for all sets in the class that are CAE, I also prove that this property holds for the 'largest' CAE sets within the class, without first needing to establish which exactly these are. The predictions of interest are still these largest CAE sets; the LCCC construction is just a device for indirectly proving things about them that would be very hard to prove directly.

The LCCC concept in general is very permissive; there can be LCCCs including many sets that are very different from any CAE set they contain. It is accordingly reasonable to state general criteria by which to judge whether a LCCC constitutes an 'interesting' or 'reasonable' prediction – whether it is 'tight', in a qualitative sense at least, around the largest CAE sets it contains. For LCCCs whose sets are all contained in some small neighborhood of strategy distribution space – those that all agree some pure path of play is almost always played, for example – a reasonable criterion is that all sets in the LCCC should agree on which actions, at histories reached with positive probability under strategy distributions in sets in the class, are played commonly, which rarely, and which never (i.e. each strategy distribution in each set in the LCCC should designate common, rare, and never played status to the same actions, and this designation should also be the same across all sets in the LCCC). The main result will feature an LCCC whose elements are unions of sets from two such neighborhoods; the sets in each of these neighborhoods separately do in fact exhibit such agreement about the 'qualitative frequency' of actions.

In the LCCC I construct, all convention-affirming equilibria will be such that most of the population agrees immediately, through their initial offers, on the same wage. That is, there are convention-affirming equilibria where most matches end in immediate agreement on the low wage and – as it turns out – a single convention-affirming equilibrium where all agents agree immediately on the high wage, but there are no cases where both wages are agreed on with substantial probability within the same population. What is interesting in this result is not just what happens but why it happens, and how the 'why' differs across the two cases. Widespread initial agreement on either wage precludes reaching, and thus observing, what would have happened after an initial disagreement; thus, agents have little empirical information about the consequences of playing differently. Nevertheless, there is an important sense in which the high-wage equilibrium is based on 'objective' thinking, while the low-wage cases depend on 'subjective' factors.

When there is immediate agreement on the high wage, all agents observe this and believe they are in a case where this happens. Since all matches end in agreement, all workers are in their first match, and could rematch if their current match ended with no deal. Accordingly, since they believe they are certain to agree on the high wage in the rematch, they will reject a lower offer. Firms anticipate this, and thus offer a high wage initially, since neither rejection nor delaying agreement is in their interest. Thus, even though later play is not observed, a unique prediction about it is derivable from the agents' interactive reasoning about each other. Agents' posteriors will concentrate on this unique prediction after observing that immediate agreement on the high wage is typical, so there is thus a sense in which they come to 'know the equilibrium strategy profile', despite not being informed of it *a priori*.

In the low-wage case, by contrast, there is nothing inherent in the set of possible equilibria which requires a worker to match a low initial offer by the firm, or, if they make a high counteroffer, to accept or reject the firm's offer if the counteroffer is rejected. Similarly, there is no inherent restriction on whether the firm must accept or reject a high counteroffer. The intuition for such CAEs is as follows. A worker who makes a high counteroffer and then rejects the firm's initial offer if it is rejected may be acting rationally if they believe most other firms would accept the high counteroffer and thus, in particular, that the firm they rematch with is likely to. Since workers may hold such beliefs, a firm who sees a high counteroffer may believe the worker they face does, and thus be motivated to accept the counteroffer rather than risk rejection. But because of this, it may be rational for a worker who is not so confident as the first type of worker to 'bluff' by making a high counteroffer, but then accept the low offer if it is rejected. Accordingly, firms may also rationally 'call' a potential bluff. The accept and reject strategies for firms and for workers thus form a self-consistent set of possibilities, each of which can be motivated by some belief over the others.

A worker who thinks the firm is unlikely to accept a high counteroffer will instead agree to the low offer initially. In any low-wage equilibrium, by definition, most workers do this. But it is somewhat subjective, requiring a particular conjecture about counterfactual play, when other conjectures are also possible. This subjective lack of confidence can be driven by prior beliefs, observations, or some combination. But it must be prior-dependent in the sense that observed deviations are not frequent enough to swamp the prior about them completely; we will see in the main result that a strategy distribution allowing for precise estimation of the post-deviation play could not be a convention-affirming equilibrium.

The rest of this section discusses the related literature. Section 3.2 defines the model and Section 3.3 the solution concept. Section 3.4 contains the results. Section 3.5 compares the results to those induced by alternative solution concepts. Section 3.6 discusses some further issues. Section 3.7 concludes.

3.1.1 Related Literature

The most directly related work is the literature on decentralized matching and bargaining in markets following Rubinstein and Wolinsky (1985), which embeds the alternating-offer bargaining game of Rubinstein (1982) in a larger dynamic market and studies its steady-state equilibria. Much of this literature has studied the extent to which the outcomes in such a 'decentralized' market can approximate classical Walrasian outcomes, and how it depends on the details of the model and definitions (e.g. Rubinstein and Wolinsky (1990), Gale (1986a), Gale (1986b), Gale (1987), Binmore and Herrero (1988), McLennan and Sonnenschein (1991), Lauermann (2013)).

The present work differs from this literature in developing a model of equilibrium bargaining under strategic uncertainty, while this previous work studies perfect equilibrium outcomes in which the strategies of others are known. Persistent strategic uncertainty is natural in a market where most bargains end in immediate agreement – it is not clear how one would know the strategies of others if they are never observed – and developing a model of equilibrium bargaining that allows for this is thus one of the main contributions of the present paper. The game I study is in many ways also simpler than the (potentially) infinitely repeated alternating-offer game of Rubinstein (1982) which much of this literature is based on. This simplicity is a benefit of the present approach; allowing for strategic uncertainty opens up new directions for inquiry which may be of interest even in fairly simple games.

Friedenberg (2019) also considers bargaining under strategic uncertainty, with a known equilibrium path in the spirit of self-confirming equilibrium. This paper differs from the present one in considering an alternating bargaining game in isolation, without the larger matching and rematching structure considered here, and in applying the rationality and common strong belief in rationality condition of Battigalli and Siniscalchi (2002), which shares with the present approach the assumption that agents engage in strategic reasoning, but situates it within a given play of the game itself rather than the larger sampling-and-matching process considered in the present work.

Also related is the evolutionary bargaining model of Young (1993b), and the subsequent work based on it (e.g. Agastya (1997), Agastya (1999), which study the evolution of coalitions, and Saez-Marti and Weibull (1999), which considers a version with 'level 1'-reasoning agents; see also the different evolutionary approach of Ellingsen (1997)). It is similar to the present framework in involving a population of agents whose expectations are determined by observing the divisions of surplus which are already typical – 'conventional' – in the population they are in. It differs in studying the Nash demand game (Nash, 1953), rather than the simple alternating-offer game studied here. It also differs in studying the long-run average outcomes of agents who best respond to the sample average they see and sometimes make mistakes, rather than a notion of equilibrium among agents who engage in strategic reasoning about each other and the process as in the present model.

The concept of convention affirming equilibrium in general is related to sampling equilibrium (Osborne and Rubinstein (2003), Salant and Cherry (2020)) in that it involves a notion of equilibrium in which agents observing a sample from the population around them act in a way that preserves the prevailing strategy distribution. It differs from this literature in studying agents who engage in strategic reasoning about one another and the process they are embedded in, rather than the impact of different statistical inference procedures on the behavior of agents who do not engage in strategic reasoning.

The convention-affirming equilibrium concept also has general connections with the concept of self-confirming equilibrium (Fudenberg and Levine, 1993), especially the 'rationalizable' variants (Dekel et al. (1999), Fudenberg and Kamada (2015), Fudenberg and Kamada (2018)), in that both involve agents who do not observe and are unsure about part of the strategy distribution they face. Convention-affirming equilibrium differs in that agents are unsure about the outcomes of strategies not often tried by *others*, not the outcomes of strategies they themselves don't play. It differs additionally from the rationalizable variants of SCE in imposing a rationalizability-like restriction on the larger population-sampling-and-inference model the game is embedded in, not on strategies within the game itself.

Further connections between the literature and CAE in general are discussed in Hudson (2023).

3.2 Model

I consider a setting with two continuum populations which I will call for concreteness *firms* (i) and *workers* (j). Successive generations of each are matched at random to play a simple bargaining game which determines how the value of output is split between them. Workers in each generation may rematch at most once if they walk away from a deal in their first period of life, but firms may only match with workers in the period they appear (so that ending with no deal is always a loss for them). Although the story and motivation are dynamic, I will define the formal solution concept in such a way that (convention-affirming) equilibrium outcomes can be studied in a static way.

3.2.1 The Game

I first describe the game played between a matched firm and worker. In the following section I will describe the larger matching process in which the game is embedded. I will assume two wage levels are possible – a low wage and a high wage – and seek to understand if and under what conditions workers will be paid the low wage, the high wage, or some mixture over the two.

When a firm and worker are first matched, the firm makes a wage offer from the set of possible wages $W = \{w_L, w_H\}$, with $0 < w_L < w_H < 1$. The worker then makes a counteroffer, also from W. If these two offers are the same, the game ends, with the worker getting payoff w and the firm payoff $1 - w$, where w is the wage offer made by both parties.

If the two offers are different, the firm chooses whether to accept the counteroffer or reject it. In the latter case the worker decides whether to accept the original offer or reject it. If the firm accepts the counteroffer, or if the worker accepts the original offer after the firm rejects the counteroffer, the firm and worker get payoffs $1 - w - c$ and $w - c$, respectively, where w is the wage thus accepted and $c > 0$ is a small cost borne by both parties as a result of not agreeing initially.[3] I assume in particular that $c < w_l$ and $c < 1 - w_H$ so that neither party ever feels compelled not to push for a better deal due to this cost alone.

If both parties reject after failing to match initially, the firm gets zero, and the worker gets either zero (if they are in their second and last match), or the payoff in their second match, discounted by the discount factor $\beta \in (0, 1)$.

I maintain two additional assumptions on the payoff parameters of the game, beyond the above. These are

$$c < (1 - \beta) w_L \tag{3.1}$$

and

[3]We can think of this as representing, say, awkwardness or bruised feelings in the working relationship due to its more adversarial beginning. This interpretation makes more sense than would treating it as cost of delay, since the worker can get out of this cost again by rematching.

$$(1-\beta)(w_L - c) < \beta(w_H - w_L) \quad (3.2)$$

Inequality (1) says that the loss c to a worker from accepting the low wage at a later stage in their first round rather than accepting it initially is less than the loss due to discounting of accepting the low wage at a given stage in their first round rather than accepting it at the same stage of their second round. That is, the cost of going to the second round exceeds that of going to a later stage of the first round, at least for the low wage. Inequality (2) says that the loss due to discounting of accepting a low wage at the end of the second round, rather than the end of the first, is less than the discounted gain of getting the high wage instead of the low wage.

The first inequality ensures that the worker will not wish to move to the second round just to avoid the cost of initial disagreement c, without expecting any other gain. The second ensures that the worker will in fact wish to move to the second round if they think they are sufficiently likely to get a high rather than a low wage by doing so. Assuming both inequalities thus allows us to focus on the interesting case: workers whose decision to 'walk' or not reflects their subjective assessment of the options awaiting them elsewhere in the market.

3.2.2 The Matching Structure

There are two continuum populations, of firms (i) and workers (j). Firms are matched with some number of workers in a way that is random and independent of any characteristics of individuals. Since the payoff of their game with one worker does not depend on the outcome of games with any others, it will not matter for firms' strategies how many workers they are matched with, or what the further details of this matching process are. Implicitly, I assume a linear technology where each worker produces an output with value 1, independent of the number of other workers employed.

I assume that firms are only concerned with the outcome of their interactions with the workers they are presently matched with (say, because they are run by managers with short tenures who care only about profit under their own management), and that all interactions are resolved simultaneously, so that firms face no dynamic considerations, neither forward-looking incentives nor learning from their own past experience.

Workers are matched with exactly one firm at a time, and may rematch at most once in their lives. If their first match ends in mutual rejection, they rematch at the same time the next cohort makes their first matches. If this match also ends in mutual rejection, they get nothing (so that optimally they will always accept at the end of the second match).

The overall payoff of workers is the payoff of their first match, if it ends in matching offers or acceptance, or the payoff of their second match, discounted by $\beta \in (0, 1)$, if they match twice. Their incentive to reject at the end of the first match thus depends on their subjected expected outcome in the second match. Workers know whether they are in their second match or not but firms do not observe this, so that a firm will play the same strategy against all workers, but workers may condition their strategy on their match number.

Although the story here is implicitly dynamic – we have successive generations matching, with some workers potentially playing in two periods – the formal model and solution concept I consider here will be static. Given the assumption that play is constant across periods and that all agents believe this, to be made below, we can study the conditions for such a constant distribution of play to be a convention-affirming equilibrium or not. In particular, since the fraction rejecting in the first match is fixed in a given equilibrium, the total population of workers playing in each period will be all the first-round workers plus this fraction of rematchers, with the latter group playing the conditional second-round strategy distribution associated with those whose first round outcomes ended in rejection. All this will be made precise in the next section.

3.2.3 (Simplified) Strategy Spaces

I denote firms by i and workers by j, and their strategy spaces by S_i and S_j. For simplicity, I will assume that firms and workers each choose from only 3 strategies, so that $S_i = \{s_i^M, s_i^A, s_i^R\}$ and $S_j = \{s_j^M, s_j^A, s_j^R\}$ (mnemonic for *M*atch, *A*ccept, *R*eject).[4] Technically, these will actually be equivalence classes of strategies which play identically against each of the three allowed equivalence classes for the other player role; since they are outcome-equivalent in this way I will just refer to them as strategies in what follows. It would not be difficult to show that firms and workers optimally must play strategies in one of these three classes, starting from the set of all their extensive-game strategies, but it will be simpler for our purposes to just treat this strategy restriction as an assumption.

All worker strategies immediately accept the high wage. That is, if the firm's initial offer is w_H, they all make counteroffer w_H, ending the game immediately with w_H as the realized wage. They differ in their play after the firm initially offers w_L. The 'match' strategy s_j^M makes counteroffer w_L after initial offer w_L, ending the game immediately in this case as well (actions at histories after a mismatch are not reachable given s_j^M and are thus not specified). The 'accept' strategy s_j^A makes a high counteroffer, but then accepts the initial offer if the counteroffer is rejected. The 'reject' strategy s_j^R makes a high counteroffer and rejects the initial offer if the counteroffer is rejected.

The 'match' strategy for the firm, s_i^M, makes initial offer w_H (thus ending the game with immediate agreement on w_H for any possible worker strategy). The 'accept' strategy s_j^A makes a low initial offer but then accepts a high counteroffer if one is made. The 'reject' strategy s_j^R makes a low initial offer and rejects a high counteroffer.

[4] S_j is the space of stage-game strategies for workers. They will choose an element of S_j to play in their first match, and another, possibly different element if they reach their second match. The connection between these two stage-game strategy choices will be explained a little later.

I also label the terminal histories associated with the various pairings of these strategies, as follows. Let z_H and z_L denote the terminal histories in which the high wage and the low wage, respectively, have been agreed upon immediately. The other three terminal histories begin with a low initial offer and a high counteroffer: z_A denotes the case where the firm accepts the counteroffer, z_{RA} the case where the firm rejects it but the worker accepts the initial offer, and z_{RR} the case where both parties reject, ending the game with no deal. Let $Z = \{z_H, z_L, z_A, z_{RA}, z_{RR}\}$ denote the set of all these terminal histories.

The relationship between strategy pairs and terminal histories is summarized in Figure 3.1.

	s_j^M	s_j^A	s_j^R
s_i^M	z_H	z_H	z_H
s_i^A	z_L	z_A	z_A
s_i^R	z_L	z_{RA}	z_{RR}

Figure 3.1: Strategies and Terminal Histories

3.2.4 Strategy Distributions

A strategy distribution for the population of firms is an element $\sigma_i \in \Delta S_i$ and a strategy distribution for the population of workers is an element $\sigma_j = (\sigma_j^1, \sigma_j^2) \in (\Delta S_j)^2$, the first coordinate being the distribution among the workers in their first match, and the second the distribution prevailing among the workers *who reach their second match* when they are in it. A strategy distribution overall is then denoted $\sigma = (\sigma_i, \sigma_j) \in \Delta S_i \times (\Delta S_j)^2$. Let Σ denote the space of strategy distributions, and Σ_i and Σ_j the spaces of strategy distributions for firms and workers, respectively.

Let $\lambda_\sigma \in [0,1]$ be the frequency with which first matches end in mutual rejection under σ_i and σ_j^1. Since workers only rematch after reaching the terminal history z_{RR}, which is reached only under the strategy pair (s_i^R, s_j^R), we have $\lambda_\sigma = \sigma_i(s_i^R)\sigma_j^1(s_j^R)$. The fraction of first-round matchers in the worker population is thus $\frac{1}{1+\lambda_\sigma}$ and that of second-round matchers is $\frac{\lambda_\sigma}{1+\lambda_\sigma}$. Let $\bar{\sigma}_j = \frac{1}{1+\lambda_\sigma}\sigma_j^1 + \frac{\lambda_\sigma}{1+\lambda_\sigma}\sigma_j^2$ then denote the overall strategy distribution among workers of both vintages.

Let $S_i(z)$ and $S_j(z)$ denote the strategies for firms and workers, respectively, that are capable of reaching terminal history z. Each strategy distribution σ induces a unique probability $\sigma(z)$ of each terminal history z, according to the formula $\sigma(z) = \sigma_i(\{s_i : s_i \in S_i(z)\})\bar{\sigma}_j(\{s_j \in S_j(z)\})$. I will in what follows sometimes abuse notation and write $\sigma(s_i)$ for $\sigma_i(s_i)$ and $\sigma(s_j)$ for $\bar{\sigma}_j(s_j)$.

3.2.5 Samples and Sample Distributions

Before being matched (in the case of workers, before the first match), all agents in both populations observe a random sample of the terminal histories reached in the matches of others under the prevailing strategy distribution (in observing others, neither workers nor firms observe whether the worker was on their first or second match). We can imagine that prospective workers or managers intending to enter a given profession or industry talk to others having done the same, but that their experiences are iid across the population and thus informative about the distribution but not the particular firms or workers who our soon-to-be entrants will encounter in the future. Agents will use their sample to update their beliefs about the prevailing strategy distribution in the market before choosing their own strategies.

Formally, fix an integer N, assumed to be a commonly known parameter of the model. Each agent (firm and worker) will observe a random sample of the terminal histories z reached in N matches. A *sample* is a tuple $x = (x_z)_{z \in Z}$, with $x_z \in \{0, 1, ...N\}$ for each $z \in Z$ and $\sum_{z \in Z} x_z = N$. Let X denote the set of all possible samples.

The prevailing strategy distribution σ induces a probability of each terminal history $\sigma(z)$ as described above. The distribution over samples under σ is thus multinomial, with probability mass function

$$f_\sigma(x) = \frac{N!}{\prod_{z \in Z} x_z!} \prod_{z \in Z} \sigma(z)^{x_z}$$

3.2.6 Reaction Functions and Aggregate Play

A pure reaction function for firms is a map $r_i : X \to S_i$ which assigns a choice of firm strategy after every possible sample. A pure reaction function for workers is a map $r_j : X \to S_j^2$ which assigns a pair of worker strategies to each possible sample – one to be played in the first match, one in the second. Write r_j^1 and r_j^2 for the projections of r_j onto the first and second match coordinates, respectively (e.g. $r_j^1(x)$ is the first round strategy played after observing sample x). Note that only workers whose first match strategies are consistent with reaching the mutual rejection terminal history z_{RR} play in the second round, so the reaction function for other workers contains some irrelevant information.

Write R_i and R_j for the spaces of pure reaction functions for firms and workers, respectively. A mixed reaction function for firms is a probability measure $\rho_i \in \Delta R_i$ and a mixed reaction for workers is a probability measure $\rho_j \in \Delta R_j$. Note that since R_i and R_j are finite, these are probability distributions over finite sets. A pair $\rho = (\rho_i, \rho_j)$ provides a complete description of how a population will behave in aggregate under any possible distribution over samples. As we will see below, mixed reaction functions correspond to distributions over prior beliefs; we can accordingly treat as somewhat interchangeable the distribution over priors prevailing in a population and the mixed reaction functions describing that population.

The aggregate play induced by σ and ρ, denoted by $\alpha(\sigma, \rho)$, is the strategy distribution which would prevail in a population with mixed reaction functions ρ and observations generated by σ. It is pinned down by its components for firms, first-match workers, and second match workers as $\alpha(\sigma, \rho) = (\alpha_i(\sigma, \rho), \alpha_j^1(\sigma, \rho), \alpha_j^2(\sigma, \rho))$. The components are defined as follows.

$$\alpha_i(\sigma, \rho_i)(s_i) = \sum_{x \in X} f_\sigma(x) \rho_i(\{r_i : r_i(x) = s_i\})$$

$$\alpha_j^1(\sigma, \rho_j)(s_j) = \sum_{x \in X} f_\sigma(x) \rho_j(\{r_j : r_j^1(x) = s_j\})$$

$$\alpha_j^2(\sigma, \rho_j)(s_j) = \sum_{x \in X} f_\sigma(x) \rho_j(\{r_j : r_j^1(x) \in S_j(z_{RR}), r_j^2(x) = s_j\})$$

where $S_j(z_{RR})$ is the subset of worker strategies consistent with reaching the mutual rejection terminal history z_{RR}. $\lambda_{\alpha(\sigma,\rho)}$, $\alpha_j(\sigma, \rho)$, and $\bar{\alpha}_j(\sigma, \rho)$ are then defined analogously to λ_σ, σ_j, and $\bar{\sigma}_j$, above.

Definition 16. *A population behavior strategy σ^* is a fixed point for mixed reaction function ρ if $\sigma^* = \alpha(\sigma^*, \rho)$.*

Note in particular that since $\alpha(\sigma, \rho)$ defines a continuous self-map on Σ for each ρ, each mixed reaction function has at least one fixed point, by the Brouwer fixed point theorem. Note also that $\alpha_i(\sigma, \rho_i)$ is the weighted sum over $\alpha_i(\sigma, r_i)$ for r_i in the support of ρ_i, with the weights induced by ρ_i. In particular, the subset of Σ_i that can be $\alpha_i(\sigma, \rho_i)$ for some ρ_i supported on a given subset of R_i is the convex hull of the $\alpha_i(\sigma, r_i)$ for r_i drawn from this subset. The analogous statements are true for $\alpha_j^1(\sigma, \rho_j)$ and $\alpha_j^2(\sigma, \rho_j)$.

3.3 Convention-Affirming Equilibrium

3.3.1 Priors and Updating

Let C be a closed subset of Σ, understood as the subset of strategy distributions an agent considers possible, and let $\mu \in \Delta C$ be an agent's prior. Agents update their prior using their sample and Bayes' rule. For each sample x with positive probability under their prior, their posterior belief μ_x after seeing sample x assigns to each Borel subset B of C the probability

$$\mu_x(B) = \frac{\int_B f_\sigma(x) d\mu}{\int_C f_\sigma(x) d\mu}$$

I will assume in what follows that an agent's prior and their sample are statistically independent. Thus, the distribution of posteriors under a given sample distribution and a given distribution of priors will be that resulting from independent draws from each followed by application of the above formula.

3.3.2 Best Responses

After updating in light of their sample, agents choose strategies by forming a posterior point estimate of the distribution of play they face, and then choosing a best response to this point estimate (slightly modified to weed out dominated strategies, in a way to be described below). An agent's reaction function is said to be a best response to their prior if their strategy choice after every sample is a best response to their posterior point estimate in this way.

Definition 17. *The posterior point estimate (PPE) of σ under posterior μ_x, denoted $\hat{\sigma}(\mu_x)$, is*

$$\hat{\sigma}(\mu_x) = \int_C \sigma d\mu_x$$

Let $\sigma_i^e \in \Delta S_i$ and $\sigma_j^e \in \Delta S_j$ denote the distributions over firm and worker strategies, respectively, that assign equal probability to all three strategies. I will define best responses as the limit of best responses to a mixture of this belief with the agent's PPE, as the weight on the former is taken to zero. This is for technical reasons, to rule out play of dominated strategies at histories that will not be reached in some of the CAE we study later.

Definition 18. *A pure reaction function for firms r_i is a best response to a prior μ if for each sample x,*

$$r_i(x) \in \lim_{\epsilon \to 0} \arg\max_{s_i \in S_i} v_i(s_i, (1-\epsilon)\bar{\hat{\sigma}}_j(\mu_x) + \epsilon \sigma_j^e)$$

where $v_i(s_i, \bar{\sigma}_j)$ is the expected payoff for a firm of strategy s_i playing against the worker strategy distribution $\bar{\sigma}_j$.

Definition 19. *A pure reaction function for workers r_j is a best response to a prior μ if for each sample x,*

$$r_j^1(x) \in \lim_{\epsilon \to 0} \arg\max_{s_j \in S_j} v_j^1(s_j, (1-\epsilon)\hat{\sigma}_i(\mu_x) + \epsilon \sigma_i^e)$$

and

$$r_j^2(x) \in \lim_{\epsilon \to 0} \arg\max_{s_j \in S_j} v_j^2(s_j, (1-\epsilon)\hat{\sigma}_i(\mu_{x'}) + \epsilon \sigma_i^e)$$

where x' is the original sample x with an additional observation of z_{RR} added, and v_j^1 and v_j^2 are the expected payoff functions for first and second match workers, respectively.

Since a worker only reaches the second round if their first match ends in mutual rejection, they will play in any case where they reach the second round with one more data point in which a firm strategy consistent with mutual rejection was encountered; their second round strategy must thus be a best response to a posterior that incorporates this. The difference in their first and second round expected payoffs is that they get zero from mutual rejection in the second round, but the discounted expected second round payoff in the first round. The subgame after an initial mismatch is depicted in Figure 3.2, below. $V(\mu, x)$ is zero for second-round workers, and for first-round workers it is the limiting maximized value of v_j^2, which in particular anticipates the slightly more pessimistic posterior an agent will have upon reaching the second round.

		Worker A	Worker R
Firm	A	$1 - w_H - c, w_H - c$	$1 - w_H - c, w_H - c$
	R	$1 - w_L - c, w_L - c$	$0, \beta V(\mu, x)$

Figure 3.2: The Post-Mismatch Subgame

3.3.3 Convention-Affirming Equilibrium

Given a closed subset $C \subseteq \Sigma$, let $\mathcal{M}(C) \subseteq \Delta C$ be a set of priors over C, called the *allowed priors* over C. A convention-affirming equilibrium set will be a pair $(C, \mathcal{M}(C))$ with the property that each $\sigma \in C$ could be the aggregate play of a population of agents who all hold allowed priors $\mu \in \mathcal{M}(C)$.

Which strategy distributions could be the aggregate play for a given allowed set is formalized as follows.

Definition 20. *A strategy distribution σ is affirmable for $(C, \mathcal{M}(C))$ if there exists a mixed reaction function ρ such that σ is a fixed point for ρ and each pure reaction function in the support of ρ is a best response to some prior $\mu \in \mathcal{M}(C)$. Write $A(C, \mathcal{M}(C))$ for the set of σ which are affirmable for $(C, \mathcal{M}(C))$.*

Definition 21. *$(C, \mathcal{M}(C))$ is a convention-affirming equilibrium set if $C \subseteq A(C, \mathcal{M}(C))$.*

This definition leaves open the possibility that there may be additional affirmable strategy distributions which are not included in the CAE set. If this is the case, there is a danger of arbitrariness in the use of CAE as a prediction, since some outcomes ruled out seem reasonable in exactly the same sense those predicted do. One solution is to focus on CAE sets that contain all affirmable strategy distributions; I call these CAE sets *complete.*

Definition 22. $(C, \mathcal{M}(C))$ *is a complete convention-affirming equilibrium set if* $C = A(C, \mathcal{M}(C))$.

In a convention-affirming equilibrium, agents understand the convention-forming process, and have correct beliefs about which conventions can actually occur, but are uncertain which convention actually holds. They are uncertain also about the distribution of priors over the true convention held by other agents, which may differ from their own and one another's, and possible hypotheses about the true convention are tied to hypotheses about the distributions of priors for which it can occur. A complete CAE set does not restrict these hypotheses beyond requiring that they be possible for some populations holding allowed beliefs over the set, while an incomplete CAE set incorporates additional restrictions imposed by the analyst on which conventions, and hence belief distributions, agents consider possible, and also the restriction that one of these is the convention which actually occurs.

It is obviously attractive to work with complete CAE sets if possible, since the set of outcomes they predict is transparently self-consistent, with no additional restrictions. The collection of all CAE sets includes also many which are too small to be reasonable – e.g. subsets of larger CAE sets with some affirmable strategy distributions arbitrarily deleted, or singleton sets which are best responses to themselves only because they exclude most off-path possibilities.

But working with complete CAE sets is not always possible. The most fundamental problem is that, for many sets of allowed priors, complete CAE sets may fail to exist. The reason is that the set of affirmable strategy distributions $A(C, \mathcal{M}(C))$ is tied to the subset of pure reaction functions which can be best responses for agents with beliefs in $\mathcal{M}(C)$. Since R is finite, and thus has finitely many subsets, there are accordingly also finitely many sets which can be $A(C, \mathcal{M}(C))$ for some C and $\mathcal{M}(C)$, and in general none of them may coincide with any of the C for which this is true. Even in cases where a complete CAE set does exist, it may be very difficult to characterizes, for technical reasons of limited economic interest.

In the following section, I will accordingly define a modified solution concept which coincides with completeness when a single complete set exists, but allows us to focus on a class of 'similar' sets rather than a single set in other cases. This sidesteps most of the thorniest technical problems associated with completeness. If this class is well-chosen, the interpretation of showing that it is a solution to this modified concept will have a similar economic interpretation to showing a single CAE set is complete.

3.3.4 Locally Complete CAE Classes

I will say a collection of subsets of Σ is a *locally complete CAE class* if it contains a CAE set, and if each affirmable strategy distribution for each set in the class is either contained in some other set in the class, or is separated by some minimal distance from all sets in the class.

Definition 23. *Let $\mathcal{C}$ be a class of closed subsets $C \subseteq \Sigma$ and let $\mathcal{M}(C)$ denote the set of allowed priors over each $C \in \mathcal{C}$. Say that $(\mathcal{C}, (\mathcal{M}(C))_{C \in \mathcal{C}})$ is a locally complete CAE class if*

(i) There exists $C \in \mathcal{C}$ s.t. $(C, \mathcal{M}(C))$ is a CAE set;

(ii) There exists some $\epsilon > 0$ s.t. for each $C \in \mathcal{C}$, and each $\sigma \in A(C, \mathcal{M}(C))$ which assigns positive probability only to samples with positive probability under some $\sigma' \in C' \in \mathcal{C}$, either $\sigma \in \bigcup_{C \in \mathcal{C}} C$, or $\sigma \notin B_\epsilon(\bigcup_{C \in \mathcal{C}} C)$.[5]

Note that a singleton class consisting of a complete CAE set is necessarily a locally complete CAE class, since by definition it is CAE and contains all points affirmable for itself. General locally complete CAE classes differ from this special case in two ways. First, I

[5] Where $B_\epsilon(\bigcup_{C \in \mathcal{C}} C)$ is the ball of radius ϵ around the union, according to the supremum metric – the set of all σ whose coordinates are all within ϵ of those of some σ' in some $C \in \mathcal{C}$.

require that any strategy distribution affirmable for a set in the class is *either* contained in some set in the class or is not 'close' to the class. This is the 'local' part; there is no 'global' requirement that all affirmable points are in the class. Second, the 'completeness' criterion applies within the class, rather than within a single set – nearby affirmable points for one set in the class must be contained in another, but not necessarily in the first set itself. The restriction to samples possible within the class avoids cases where a strategy distribution is affirmable only because agents have arbitrary posteriors after a zero-probability event.

Locally complete CAE classes, in general, do not inherently impose non-trivial restrictions. For instance, the class of all closed subsets of Σ, each endowed with the set of all possible priors over them, is always trivially a locally complete CAE class.[6] Rather, the quality of the prediction is tied to the choice of class, particularly the extent to which all sets in the class occupy a 'similar' region of Σ (so that the weakening of the concept relative to that of a single complete CAE set is mild).

3.4 Results

In the main result to follow, I will construct a LCCC consisting of two possible outcomes – immediate agreement on the high wage and immediate agreement on the low wage. In the high-wage case, all firms will play s_i^M and all workers will play s_j^R, so that this bargaining outcome is literally reached in every match. In the low-wage case, firms will play either s_i^A or s_i^R (both compatible with immediate agreement on the low wage), and most workers will play s_j^M (compatible with the same), but some may also play s_j^A or s_j^R.

Informally, the logic of the LCCC is as follows. Firms and workers enter the world with prior beliefs about both the relative likelihood of different CAE within each of these two

[6] There is always at least one CAE set because any singleton consisting of a Nash equilibrium is a CAE set, and the game has a Nash equilibrium.

cases, and about the relative likelihood of these two cases overall. But since no sample agents might observe is consistent with both cases – firms only make high offers in the first case and only make low offers in the second case – agents will be certain which case they are in once they observe their sample. Accordingly, only the distribution over conditional priors within each case will influence which strategy distributions consistent with each case are affirmable for a given population.

In the high wage case, since only one strategy is ever played by each side of the population, affirmability (of this single high-wage CAE) requires only that these strategies are strict best responses to each other. Since agents observing samples from this distribution are certain it is the true one, this implies affirmability (and the absence of other affirmable points, in which other strategies are played with positive probability.)

In the low wage case, all agents are convinced that all firms make low offers and that high counteroffers are rare, but workers are unsure whether firms would accept or reject a high counteroffer, and firms are unsure whether a worker who made a high counteroffer would accept or reject the firm's original offer if the counteroffer were rejected. There are many such distributions that agents consider possible, that differ on how rare the rare high counteroffers are, and also differ on the conditional distribution of play following a high counteroffer. Showing that all strategy distributions in such a collection are affirmable, and thus part of a CAE set, requires showing that we can assign agents prior beliefs over the collection in a way that makes each strategy distribution within the collection consistent with subjectively optimal behavior for all agents. Roughly speaking, this involves varying the priors of firms to make the subjective probability that a high-counteroffer worker is 'bluffing' high or low, and varying the priors of workers to make the subjective probability of a firm 'calling a bluff' high or low. A CAE distribution where all these strategies are played with nontrivial frequency will involve agents with posteriors that differ along these dimensions, so that they view different strategies as best responses to the environment they face. Such disagreement in equilibrium is possible so long as play after a high counteroffer is rarely observed. As observations of high counteroffers become more common, there comes a point where posteriors can no longer differ to the necessary degree; this defines the boundary of the low-wage component of the LCCC.

3.4.1 Firm and Worker Best Responses

In this first subsection, I explain how agents' beliefs influence their choice of strategy within the set of strategies playable in the low-wage case. This will help to clarify what is going on in the proof below, and why.

The Firm's Problem

For a firm choosing between s_i^A and s_i^R, the only relevant issue is the probability with which a worker who makes a high counteroffer (a worker playing either s_j^A or s_j^R) will accept the firm's initial (low) offer if the firm rejects their counteroffer. Denote this probability by q_j. It is the ratio, under the firm's posterior point estimate $\hat{\sigma}(\mu_x)$, of the probability of s_j^A to the total probability of both s_j^A and s_j^R.

Conditional on not initially agreeing (under initial agreement these strategies generate identical payoffs), the firm's expected payoff from s_i^A is $(1 - w_H - c)$ while their expected payoff from s_i^R is $q_j(1 - w_L - c)$. Thus, s_i^A is weakly better if $q_j \leq \frac{1-w_H-c}{1-w_L-c}$. This threshold value for the subjective probability q_j completely determines the firm's choice between these two strategies.

The Worker's Problem

The worker's problem is more complicated, because they are deciding among three s-trategies and considering outcomes across (possibly) two matches. Workers, for the possible strategies we consider here, believe any firm will play either s_i^A or s_j^R, so their own choice of strategy must hinge on their posterior point estimate of the probability of s_i^A. Denote their first-round posterior point estimate of it by q_i and their second-round posterior point estimate of it by $q_i^{'}$. These are different, because in the second round they have an additional observation. Since the second round is only reached if they encounter a firm playing s_j^R, thus reaching the terminal history z_{RR}, $q_i^{'}$ is unique given their posterior belief in the first round and in particular is foreseeable from the first round, though $q_i^{'} < q_i$.

The strategy s_j^M, which agrees immediately to the low wage, yields a certain payoff w_L. The expected payoff of s_j^A is $q_i(w_H - c) + (1 - q_i)(w_L - c)$ in the first round and $q_i^{'}(w_H - c) + (1 - q_i^{'})(w_L - c)$ in the second round. Rearranging, s_j^M is weakly preferred to s_j^A in the first round if $q_i \leq \frac{c}{w_H - w_L}$ and in the second if $q_i^{'} \leq \frac{c}{w_H - w_L}$.

The strategy s_j^R can only be a best response in the first round, since in the second it is strictly worse than s_j^A whenever the firm might play s_i^R. In the first round, since they lead to different outcomes only when playing against a s_i^R firm, the payoff comparison between the two hinges on the relative values of $w_L - c$ (the payoff of accepting the firm's initial offer if the counteroffer is rejected) and $\beta V(\mu, x)$, the discounted expected value of the second round, conditional on continuing to it. Under the maintained assumption on payoff parameters that $c < (1 - \beta)w_L$, the worker cannot reject if they intend to play s_i^M in the second round. The payoff from the latter would be βw_L, while that from accepting at the end of the first round would be $w_L - c$, and it is easy to see the maintained assumption implies the latter exceeds the former.

Thus, any worker playing s_j^R in the first round will play s_j^A in the second, and we calculate $V(\mu, x)$ accordingly. Given $q_i^{'}$ (itself implicitly a function of μ and x), $V(\mu, x) = q_i^{'}(w_H - c) + (1 - q_i^{'})(w_L - c)$, the expected payoff of s_j^A in the second round. Playing s_j^A in the first round is thus subjectively better than playing s_j^R if $\beta V(\mu, x)$ is less than $w_L - c$, which reduces to the condition $q_i^{'} \leq \frac{1-\beta}{\beta} \frac{w_L - c}{w_H - w_L}$.

Note that the choice between s_j^A and s_j^R in the first round thus depends on the estimate of the firm acceptance probability the worker would have in the *second round*. Note also that, under the assumed condition $c < (1-\beta)w_L$, we also have $0 < \frac{c}{w_H - w_L} < \frac{1-\beta}{\beta}\frac{w_L - c}{w_H - w_L} < 1$, where the last inequality follows from $(1-\beta)(w_L - c) < \beta(w_H - w_L)$, the other maintained assumption on parameters.

Thus, we can characterize the conditions for each first-round worker strategy to be a best response as follows. When $q_i \leq \frac{c}{w_H - w_L}$, s_j^M is a best response. When $q_i^{'} \geq \frac{1-\beta}{\beta}\frac{w_L - c}{w_H - w_L}$, s_j^R is a best response. And when either $\frac{c}{w_H - w_L} \leq q_i \leq \frac{1-\beta}{\beta}\frac{w_L - c}{w_H - w_L}$ or $\frac{c}{w_H - w_L} \leq q_i^{'} \leq \frac{1-\beta}{\beta}\frac{w_L - c}{w_H - w_L}$, s_j^A is a best response. (Since $q_i^{'} < q_i$, either of these inequalities implies the negation of the previous two.) Thus, in qualitative terms, first-round workers immediately agree to the low wage if they think the frequency of firms that would accept a high counteroffer is low, make a high counteroffer and reject at the end of the first round if they think this frequency is high, and make a high counteroffer but accept at the end of the first round if they believe it is intermediate.

The Frequency of Second-Round Deviators

Let q_j^1 denote the probability a *first round* worker plays s_j^A, conditional on their playing either s_j^A or s_j^R. Since all workers who play s_j^R in the first round play s_j^A in the second, and since the fraction of these who make it to the second round is determined by the probability $1 - q_i$ of firm rejection, the conditional probability of acceptance among all workers, q_j, can be written in terms of q_j^1 and q_i as

$$q_j = \frac{q_j^1 + (1 - q_j^1)(1 - q_i)}{1 + (1 - q_j^1)(1 - q_i)} \tag{3.3}$$

Note that this is the conditional frequency of s_j^A among the workers who deviate (play some strategy other than s_j^M) not the overall frequency.

When $q_i = 0$ (all firms reject), this ranges from $\frac{1}{2}$ to 1 as q_j^1 ranges from 0 to 1. When $q_i = 1$ (all firms accept), it ranges from 0 to 1. There is thus a dependency of the range of possible values of q_j on the value of q_i. This is, in general, a reason why firms might care about observations of what other firms are doing, though it will not much affect matters in the result I prove below.

Let $\Sigma^* \subseteq \Sigma$ denote the set of all $\sigma \in \Sigma$ with the property that $\sigma_j^2(s_j^A) = 1$. These are the strategy distributions in which all second-round workers play s_j^A, which as we have seen are the only cases consistent with worker optimization (against the possible firm strategies I will consider). Clearly, equation (3) and the associated restrictions on $\bar{\sigma}_j$ will be satisfied by any $\sigma \in \Sigma^*$.

3.4.2 The Match-High/Match-Low LCCC

Let $\mathcal{C}_H$ be the singleton class consisting of the set $C_H \subset \Sigma^*$ which is itself a singleton consisting of the unique strategy distribution σ_H which satisfies $\sigma(s_i^M) = \sigma(s_j^R) = 1$.

Let $\mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ be the class of $C_L \subset \Sigma^*$ containing all and only those σ satisfying $\sigma(s_i^M) = 0$, $\sigma(s_j^M) \geq 1 - \epsilon$ for some $\epsilon \in (0, \bar{\epsilon})$, and $\frac{\sigma(s_j^A)}{\sigma(s_j^A)+\sigma(s_j^R)}, \frac{\sigma(s_j^R)}{\sigma(s_j^A)+\sigma(s_j^R)}, \sigma(s_i^A), \sigma(s_i^R) \in (\epsilon, 1-\epsilon)$ for some $\epsilon \in [0, \underline{\epsilon})$. That is, $\mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ is a class of sets indexed by the maximum frequency $\bar{\epsilon}$ of workers who can 'deviate' from the low-wage regime by making a high counteroffer and the minimum fraction $\underline{\epsilon}$ of agents who deviate or respond to a deviation in any particular way conditional on being in a match with deviation, and we consider different possible upper bounds on these parameters by writing the class as a function of $\bar{\epsilon}$ and $\underline{\epsilon}$.

Let $\mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$ be the class consisting of all unions of C_H with some $C_L \in \mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$. I will say the two classes $\mathcal{C}_H$ and $\mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ are the *components* of $\mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$, and similarly that C_H and C_L are the components of $C = C_H \bigcup C_L \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$.

Say that a strategy is *live* for a component of C if it has positive probability under some strategy distribution in the component. Otherwise it is *dead*. The strategies S_i^M and S_j^R are the only live strategies for C_H, while all strategies except S_i^M are live for each C_L.

Fix a parameter $\eta \in (0,1)$, which is common knowledge. For $C = C_H \bigcup C_L \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$, let $\mathcal{M}(C)$ be defined as follows. Consider the collection $\mathcal{M}_L(C_L)$ of priors over C_L which have densities bounded by lower and upper multiples of the uniform density on C_L, with bounds η and $\frac{1}{\eta}$. $\mathcal{M}(C)$ contains all priors which are mixtures of some prior in $\mathcal{M}_L(C_L)$ with the point mass on σ_H, with strictly positive weights on each.

I show that $(\mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon}), \{\mathcal{M}(C)\}_{C \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})})$ is a LCCC, for some choices of $\bar{\epsilon}, \underline{\epsilon}$.

Theorem 4. *For all $\bar{\epsilon}$ sufficiently close to zero, for all η sufficiently close to zero, for all N sufficiently large given η and $\bar{\epsilon}$, there exists $\underline{\epsilon} > 0$ such that $(\mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon}), \{\mathcal{M}(C)\}_{C \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})})$ is a locally complete CAE class.*

Proof. We need to show that $(\mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon}), \{\mathcal{M}(C)\}_{C \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})})$ satisfies conditions (i) and (ii) of the LCCC definition.

Note first that the set of terminal histories which have positive probability under some $\sigma \in C_L \in \mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ and the set with positive probability under σ_H are disjoint. Thus, for agents observing a sample drawn from some element of either component, all strategy distributions in the other component will have zero posterior probability, and the agent's posterior given their sample will be the same as the posterior of an agent whose prior is equal to the conditional prior over the component they are in.

In looking for mixed reaction functions that make a strategy distribution in one component a fixed point, we can thus focus on conditional priors over that component and corresponding strategies that can be played after samples possible under that component. In particular, in finding a CAE set in the class to satisfy part (i) of the definition, we can check the condition for σ in each component separately.

Any $C \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$ has the singleton consisting of σ_H as its C_H component, and as noted agents making observations at σ_H will attach probability one to the event that they are in fact at σ_H, and their posterior will thus be trivial: concentrated on one point. Thus, it suffices to show that playing along with σ_H – playing s_i^M for firms and s_j^R for workers – is a best response for agents with posterior certainty of σ_H.

For firms facing s_j^R for sure, s_i^M is a strict best response. Playing s_i^R instead would result in a zero payoff, and s_i^A would result in a payoff of $1 - w_H - c$, which is worse than the payoff $1 - w_H$ from playing s_i^M.

All worker strategies match the high offer. For workers who were to somehow encounter a firm making a low offer, it is a strict best response to make a high counteroffer and reject the low offer, since the worker expects immediate agreement on the high wage with probability one in their second match, and thus cannot optimally choose to end their first match with a strictly lower payoff than this. (Note that the definition of best response as a limit of mixtures of the PPE with a small full support hypothesis is invoked here, as are the assumed conditions on payoff parameters.)

To find a CAE set $C \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$, it thus suffices to find some $C_L \in \mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ with the property that all $\sigma \in C_L$ can be the aggregate play of a population with some allowed distribution of posterior beliefs over C_L; we can then use $C_L \bigcup \{\sigma_H\}$ as our CAE set satisfying (i).

There are two live firm strategies (s_i^A and s_i^R) and three live worker strategies (s_j^M, s_j^A, and s_j^R) for $\mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$. I will show there exist choices of one allowed prior for each of these five strategies – chosen so that agents holding such a prior will for a high-probability subset of samples always choose that strategy – and then show that each $\sigma \in \mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ is a fixed point for some mixed reaction function supported on the pure reaction functions induced by these priors. Since the aggregate play of any such mixed reaction function at σ is just the weighted sum of the aggregate play induced by each of the pure reaction functions, $\alpha_i(\sigma, r_i)$ and $\alpha_j^1(\sigma, r_j)$,[7] weighted by their probabilities under the mixture, it suffices to show that

[7]As explained in the previous section, σ_j^2 is necessarily a point mass on s_j^A whenever workers best respond,

some such mixture can be found for each $\sigma \in C_L$.

For the firm strategies s_i^A and s_i^R, I choose priors whose weight is concentrated on σ with values of $\frac{\sigma(s_j^A)}{\sigma(s_j^A)+\sigma(s_j^R)}$ close to 0 and 1, respectively. That is, firms with the prior for s_i^A will be very certain that most deviating workers reject and vice versa. Formally, note that given the definition of the set of allowed beliefs $\mathcal{M}_L(C_L)$, there is a unique cutoff value ψ such that there exists an allowed prior assigning the ceiling density ($\frac{1}{\eta}$ times the uniform density) to all $\sigma \in C_L$ with $\frac{\sigma(s_j^A)}{\sigma(s_j^A)+\sigma(s_j^R)} \leq \psi$ and the floor density (η times the uniform density) to all other σ, and that ψ approaches zero as η does.

Let the prior μ_i^A associated with s_i^A be the unique prior defined in this way. Let the prior μ_i^R associated with s_i^R then be defined analogously to place the ceiling density on σ with $\frac{\sigma(s_j^A)}{\sigma(s_j^A)+\sigma(s_j^R)} \geq 1-\psi$ (where the ψ in this case may be different, even for the same η, but still goes to zero as η does).

The priors for the worker accept and reject strategies, μ_j^A and μ_j^R are then defined identically in terms of the unique floor/ceiling cutoff in the probability $\sigma(s_i^A)$ of firm acceptance. Let the prior μ_j^M for s_j^M be defined instead to place the ceiling density on the interval of values of $\sigma(s_i^A)$ of uniquely determined (and vanishing in η) length ψ, which is centered on the value $\frac{1}{2}(\frac{c}{w_H-w_L} + \frac{1-\beta}{\beta}\frac{w_L-c}{w_H-w_L})$ (i.e. the midpoint of the range of $\sigma(s_i^A)$ making s_j^M a best response).

The set C_L that we are proving is part of a CAE set will be such that its maximum frequency of deviation $\epsilon_1 \in (0, \bar{\epsilon})$ is very small, and the minimum relative frequencies of the accept/reject strategies for both players $\epsilon_2 \in (0, \underline{\epsilon})$ is large relative to ϵ_1. In particular, for ϵ_1 chosen small enough for any fixed N, deviation will be rare enough that the total frequency of agents who make more than one observation other than immediate agreement on the low wage (z_L) is strictly less than any positive multiple of ϵ_2.

It is easy to see that, for agents with any of the five priors defined above and η small

and we have built this into the definition of C_L by assuming it is a subset of Σ^*. It thus suffices to show the mixed reaction function reproduces the strategy distributions for firms and for first-round workers.

enough, they cannot optimally choose to play a strategy other than that associated with the prior unless they observe more than one terminal history which is not z_L. Observing z_L does not inform them about the relative probabilities of acceptance and rejection they care about, and as η approaches zero the concentration of the prior on values making the prescribed strategy a strict best response must dominate the impact of one observation.

Thus, from the above conditions on ϵ_1 and ϵ_2, the aggregate play $\alpha_i(\sigma, r_i)$ for firms with a reaction function r_i which is a best response to μ_i^A or μ_i^R must for any $\sigma \in C_L$ assign probability strictly less than ϵ_2 to their not playing the associated strategy. Thus, $\sigma_i(s_i^A)$ lies in between the probabilities of s_i^A under the two $\alpha_i(\sigma, r_i)$, for any two such r_i, so that by the intermediate value theorem there exists a convex combination of the two $\alpha_i(\sigma, r_i)$ whose probability of s_i^A exactly matches $\sigma_i(s_i^A)$. In this case, the probability of s_i^R must obviously match too, so the mixed reaction function with corresponding weights on the two r_i is consistent with σ being a fixed point, for each $\sigma \in C_L$.

It thus suffices to show that there is a mixed reaction function for first-round workers that reproduces each σ_j^1. Since there are three worker strategies in play, the argument is a little more complicated, but in the same spirit. Note first, by the same logic and under the same limit conditions as above, that the frequency of agents observing more than two deviations is less than any positive multiple of $\epsilon' \epsilon_2$ (i.e. the minimum frequency of each deviating worker strategy, if the minimum frequency *conditional on deviation* is ϵ_2 and the total frequency of deviating workers is $\epsilon' \in (0, \epsilon_1)$). More than two observed deviations are needed to swamp a stubborn prior for the same reason more than one is, so we conclude also that for each $\sigma \in C_L$, with associated total frequency of deviation ϵ', the aggregate play $\alpha_j^1(\sigma, r_j^1)$ associated with some reaction function r_j^1 which is a best response to any of the three chosen worker priors assigns probability less than $\epsilon' \epsilon_2$ to play of any strategy other than the prescribed one, and thus in particular to each of the non-prescribed strategies individually.

It remains to show that for each $\sigma \in C_L$, σ_j^1 is a convex combination of the $\alpha_j^1(\sigma, r_j^1)$

for some choice of three r_j^1 that are best responses to μ_j^M, μ_j^A, μ_j^R, respectively. From this it immediately follows that a mixed reaction for first-round workers with the same weights is consistent with each $\sigma \in C_L$ being a fixed point. Putting this together with the mixed reaction function for firms constructed above for each $\sigma \in C_L$, we then conclude that each is in fact a fixed point, and that $C_L \bigcup \{\sigma_H\}$ is thus a CAE set.

To begin with, observe that for each of the three $\alpha_j^1(\sigma, r_j^1)$ and for each convex combination over the other two, there exists a unique convex combination of $\alpha_j^1(\sigma, r_j^1)$ and the given convex combination over the others such that the resulting distribution assigns the same probability to the strategy associated with the given r_j^1 that σ does. This is again guaranteed by the intermediate value theorem.

Next, consider the self-map on the set of weights for all possible convex combinations over the three $\alpha_j^1(\sigma, r_j^1)$ (the 3-simplex), which assigns to each of the $\alpha_j^1(\sigma, r_j^1)$ the unique weight whose existence was asserted in the previous paragraph, given the weights on the other two (normalized so that they sum to one). It is easy to see that it is continuous. Since the 3-simplex is compact and convex, the Brouwer Fixed Point Theorem applies. By construction, any fixed point of this map gives the weights of a convex combination of the $\alpha_j^1(\sigma, r_j^1)$ which is equal to σ_j^1.

This completes the proof of part (i).

It remains to show part (ii). We need only check σ supported on the live strategies for one of the two components, as any other σ will assign positive probability to some samples with zero probability under all $\sigma^{'} \in C$, for all $C \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$. Trivially, since there are no strategy distributions other than σ_H supported on the live strategies for $\mathcal{C}_H$, there are no such strategy distributions which are affirmable for any $C \in \mathcal{C}^*(\bar{\epsilon}, \underline{\epsilon})$.

Consider some strategy distribution σ supported on the live strategies for $\mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ which is δ-close to the union over $\mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$ for some $\delta > 0$, but not contained in any $C_L \in \mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$. It must then satisfy $\sigma(s_i^M) = 0$ and $\sigma(s_j^M) \in (1 - \bar{\epsilon}, 1 - \bar{\epsilon} - \delta]$. As N becomes large relative to η, almost all agents at σ must have samples with arbitrarily large numbers of observations of play after (w_L, w_H), and thus in particular almost all workers must have posterior estimates of the probability a firm accepts a high counteroffer (the relative frequency of s_i^A among those playing either s_i^A or s_i^R) which are arbitrarily close to the true value.

The best response of workers depends only on their estimate of this probability. First-round workers optimally play s_j^M for sufficiently low estimates of the firm acceptance rate, s_j^R for sufficiently high estimates, and s_j^A for those in the middle (and almost all workers are first round under any $\sigma \in C_L \in \mathcal{C}_L(\bar{\epsilon}, \underline{\epsilon})$). In particular, since this middle region is of fixed length, as the number of observed deviations increases and most workers come, by the Law of Large Numbers, to have PPEs close to the true frequency of firm acceptance, there cannot be both a nontrivial fraction of both s_j^M and s_j^R played. Since any point close to $\mathcal{C}_L$ prescribes a high frequency of s_j^M, we conclude that almost all deviant workers must accept in such a case, which is rational only if a sufficient number of firms reject. If they did, then most firms would have samples that convince them most workers accept, so almost all firms should reject. But if this were the case, almost all workers – in particular, more than $1 - \bar{\epsilon}$ – should play s_j^M, which contradicts the presumed frequency of s_j^M for the σ we consider. We thus conclude no such σ can be affirmable. □

Remark 3. *The LCCC identified here is arguably the 'largest', in a qualitative sense, among those that do not consider outcomes in which multiple paths in the game are played with substantial probability. (If there were a single path of play always played other than immediate agreement, it would be worse for someone than immediate agreement, and thus not CAE.) While we could consider such mixed cases, they would more or less by definition be knife-edge cases and thus not necessarily interesting. If we leave aside such cases, we can consider the LCCC of Theorem 4 a more or less a complete description of convention-affirming predictions for this game.*

3.5 Comparisons With Other Concepts

In this section, I compare the LCCC derived in Theorem 4 to the predictions of a number of alternative solution concepts which could be applied to the same setting. For the sake of apples-to-apples comparison, I shall restrict attention to predictions of these other concepts in which there is immediate agreement on either the high wage or the low wage. Where multiple versions of these concepts are available, I try to focus on the one that offers the closest possible comparison.

3.5.1 Subgame-Perfect Equilibrium

I take a subgame-perfect equilibrium for the game in this paper to be a case where the (pure or mixed) strategy distribution played by each of the three player roles – firm, first-round worker, second-round worker – is commonly known among all players, and where these strategies are objectively optimal in the subgames starting at all non-terminal histories (including the optimal choice of a first-round worker to reject at the end of their first match or not, given their objective continuation value if they do). One difficulty in translating to the SPE context is that the 'matching' strategies s_i^M and s_j^M do not specify behavior at all histories (since they guarantee these histories are not reached, given their opponents' possible strategies). I get around this by also specifying in the SPE context which of the other two strategies (A or R) a M agent would deviate to, if they were to deviate. It is natural in considering SPE analogues of the above results to focus on cases where the paths of play we focus on (immediate agreement on either the high or low wage) are followed with probability one; since the true strategy profile is known, the diverse beliefs that could motivate both adherence to and deviation from the low wage LCCC could not arise in SPE.

There is always immediate agreement on the high wage just in case firms always choose the strategy s_i^M (since all worker strategies immediately accept a high offer). Thus, any SPE supporting this outcome will assign firms this pure strategy and assign some mixed strategy to each vintage of workers. As noted in the LCCC case as well, a worker who expects to receive the high wage with certainty upon rematching (in the SPE case, because they know the firm's strategy *a priori*) cannot optimally accept a low wage at any point in their first match. This implies the first-round worker must play S_i^R with probability one. Since second-round workers have zero probability in this case, a firm who deviates from s_i^M must deviate to s_i^A; otherwise they are not playing optimally after a high counteroffer, since the worker is certain to reject. Given this, second-round workers must play s_j^A, since doing so leads to the counteroffer being accepted. Since this whole derivation was from necessary conditions, this is the only SPE with immediate agreement on the high wage. It is the same outcome as the high-wage part of the LCCC of Theorem 4, except that strategies are fully specified even at histories they make unreachable, and are known *a priori* rather than derived in part from observations and reasoning about observations.

There is always immediate agreement on the low wage if firms mix between s_i^A and s_i^R and workers all play s_j^M. From the calculations in Section 3.4.1, s_j^M is optimal only when the worker expects the relative frequency of s_i^A in this mixture to be sufficiently low. In SPE then, it is necessary that the true weight on s_i^A in the firm's mixture in fact be sufficiently low. From the same calculations, we know that s_j^M being optimal for first-round workers implies s_j^A being better than s_j^R for first-round workers; thus, a first-round worker would deviate to s_j^A if they were to deviate from s_j^M when s_j^M is optimal. Since all workers are first-round workers in this case also, this implies all firms must play s_i^R (since it is a strict best response to a s_j^A worker in the subgame after deviation to a high counteroffer). This satisfies the condition on worker payoffs for first-round workers to optimally play s_j^M and deviate to s_j^A if they deviate, and implies the same optimal play for second-round workers. This is again the unique SPE for this case, as we have derived everything from necessary conditions.

The key difference in outcomes between SPE and LCCC is that the LCCC beliefs are 'objective' only in the high wage case, while the SPE beliefs are objective in both cases, and also specify counterfactual behavior to a greater degree than is necessary in LCCC. This difference arises because knowledge of strategies in a LCCC must be derived from agent observations and strategic reasoning if it is to be present, and full knowledge is not always derivable in this way. In a SPE, full knowledge is assumed *ex ante*, having already been arrived at by an unspecified process, so that the two cases cannot differ on this point.

To summarize: Both LCCC and SPE predict probability-one adherence to immediate agreement in the high-wage case, and sufficient knowledge of strategies to anticipate one's opponent's response to a deviation, though SPE specifies one's own optimal post-deviation behavior as well. In the low-wage case, LCCC predicts that positive probability deviations may arise that may or may not be profitable, and that whether they are profitable is unknown to agents, whereas in SPE the value of deviation is known and worse than the on-path payoff, as all firms reject high counteroffers and all agents know this.

3.5.2 Non-Strategic Bayesian Inference ('Sampling Equilibrium')

Next, I consider a case identical to the current model, except that agents have priors over all of Σ, with densities bounded between multiples η and $\frac{1}{\eta}$ as before. We will take our predictions in this case to be just the set of strategy distributions affirmable for Σ and these allowed priors. This is, in essence, a modified version of the current model in which strategic reasoning – embodied in the requirement that priors be supported on CAEs – has been stripped out. It is more or less identical in spirit, if not exactly the same formally, to the Bayesian inference variant of sampling equilibrium in Salant and Cherry (2020).

In studying this case, I consider the parameter limit where η approaches zero (the bounds on priors become very loose) and N becomes large relative to η, and I focus on strategy distributions in which all but at most fraction ϵ of agents agree immediately on either the high or the low wage, where ϵ is small relative to N. This is the same parameter limit and class of strategy distributions considered in the LCCC of Theorem 4, except that I allow for the possibility of positive-probability deviation from the high wage case. The key observation to make about this case is that agents with such beliefs and samples facing such a strategy distribution will have very high confidence that immediate agreement will occur if they play along with it (because N is large relative to η), but have more or less unrestricted posteriors about play if they don't (ϵ being sufficiently small relative to N implies their subsample of play at all unusual histories is also small, and η is assumed to be small, so that their posterior point estimates need not be significantly restricted by small subsamples).

It is easy to see that this case allows immediate agreement by all agents on either the high or low wage. They could all have priors that are very concentrated on near-certain rejection if they don't, for example. The key difference in this case, relative to SPE and LCCC, is that positive probability deviation is possible for both cases. Just as we constructed beliefs in the low-wage part of the LCCC for which deviators expected their counteroffer to be accepted and non-deviators expected it not to be, we can construct beliefs for firms in this case for which workers would almost always accept a low offer if one were made. Firms with such beliefs would in fact deviate in this setting, because they do not think strategically about what the workers observe, and thus do not incorporate the restriction that workers who know high firm offers are very common must optimally reject low offers. This is still true of workers who observe near-universal agreement on the high wage, however, so it is still the case that deviant firms in the high-wage case must be rejected, while high counteroffers in the low wage case might be accepted or rejected.

To summarize: This case differs from LCCC in allowing firms to deviate to a low offer in the high-wage case, although such offers must still be rejected. It allows for the same range of behavior as LCCC in the low-wage case.

3.5.3 Self-Confirming Equilibrium

In a self-confirming equilibrium (Fudenberg and Levine, 1993), all agents have correct beliefs about the play they will face at histories they can reach when playing their own strategy against the prevailing strategy distribution in the population. It is natural to consider comparisons with both SCE where literally all agents agree immediately on either the high or the low wage, and SCE in which some small fraction of agents may not do so.

In the first case – immediate agreement is literally always reached – the beliefs of agents about what would happen following a deviation to a low initial offer (in the high wage case) or a high counteroffer (in the low wage case) are completely unrestricted (deviation to a low counteroffer in the high wage case or a high initial offer in the low wage case is strictly dominated given universal immediate agreement). Thus, it is possible to assign priors to agents that expect rejection after any deviation, and thus support both outcomes. So, there are SCE where immediate agreement on each wage happens with probability one. It remains to check which cases with immediate agreement by most but not all can also arise in SCE.

If there is almost always immediate agreement on the high wage, all workers will immediately accept the high wage in every match where they are offered it and will thus know this almost always happens. They will thus reject any low offer in their first round, for the same reasons as before. Since most workers in such a case are first round and all first-round workers behave this way, any firm which offered a low wage initially would experience rejection in almost every case. Such a firm would not know the outcome of offering a high wage, but the lowest possible payoff of doing so not much less than the payoff they are already receiving given their current strategy. There is thus a sense in which there cannot be positive probability deviations from the high-wage case in a SCE with 'reasonable beliefs'. The only beliefs supporting this outcome are those for which workers are expected to refuse the high wage with near-certainty, with the degree of certainty needed approaching one as deviations by firms become rare. This would be ruled out by, for example, beliefs about off-path play bounded by some η as in the allowed class considered in this paper, if the frequency of deviation were sufficiently small.

If there is almost always agreement on the low wage, all workers know the low wage is almost always offered, and in particular believe this is also true in the second round. Workers who accept the low wage do not know what would happen if they made a high counteroffer instead, and thus may believe that it would almost certainly be rejected, so that accepting the low wage is a subjective best response. A worker who makes a high counteroffer does better than accepting the low wage (an outcome they know they can guarantee if there is a low initial offer) only if sufficiently many firms objectively accept a high counteroffer, and workers who do make high counteroffers know whether this is true or not. Accordingly, a necessary condition for there to be high counteroffers in a mostly-low-wage SCE is that firm strategies do in fact lean toward acceptance in this way. A firm who plays s_i^A (make a low initial offer, but accept high counteroffers) never learns whether the worker was 'bluffing' or not, so there are always firm beliefs that allow firms to play s_i^A. Firms who play s_i^R can be present too so long as most workers who deviate are playing s_j^A, so that rejection is not objectively suboptimal. Thus, we can have low-wage SCE with positive probability deviation so long as both sides of the market lean sufficiently toward acceptance after a deviation. This is more restrictive than the LCCC case, because of the requirement that deviators know the consequences of their own deviation.

In summary, SCE (with 'reasonable beliefs') agrees with LCCC that the high wage outcome will involve immediate agreement with probability one, but for a different reason – empirical knowledge of the deviators about the true consequences of deviating, rather than deduction about one's opponent's strategy from knowledge of their payoffs and observations. SCE also agrees with LCCC that there can be positive probability deviations from the low wage outcome, but these must take a more restricted form, again because of the requirement that deviators have learned the true consequences of deviating empirically.

3.6 Discussion

3.6.1 Simpler Versions of the Game

Since the game considered here has up to four moves – with agreement or disagreement reachable, in a sense, at three of them – one might instead consider truncated versions with the last one or two moves deleted. One might also consider a case where no rematching is allowed. In this section, I provide a brief, informal sketch of what might change in such cases.

Consider first a case in which the possibility of worker rematches is as before, but the bargaining game has fewer moves. If there are only the first two moves, with the match ending immediately with no deal if the two proposals are not the same, then under large sample sizes workers have a very accurate estimate of their expected payoff from rematching (i.e. the probability they will be offered the high wage), and firms have a very accurate estimate of what most workers' estimates will be. Thus, there is no significant room for uncertainty or disagreement about which offers workers will accept. We would accordingly expect CAE in this case to coincide with Nash equilibria.

If, instead, the third move – the firm accepting or rejecting the counteroffer – were included but the game ended with no deal following a firm rejection, the firm would never reject, since doing so gives them the worst outcome. Were we to give firms rather than workers the ability to rematch, so that rejection could conceivably benefit them, we might instead have a case much like that studied in the present paper. If firms and workers typically observe a large subsample of worker responses to an initially low offer, then both sides have a largely accurate estimate of the firm's expected payoff from making a low offer and from rematching. If only high offers are typically made, there may instead be uncertainty and potential disagreement about both these expected payoffs. The LCCC for such a case would plausibly look much like that constructed in the present work, with the roles of firms and workers (and high and low wages) reversed. The game studied in this paper is arguably more realistic as a description of at least some wage bargaining situations – where the firm's offer, once made, stays on the table for a predetermined length of time whether or not a counteroffer is made or rejected within this time window.

If there are no rematches, both parties understand that the final mover will always want to accept. This would seem to induce a backward induction logic which also eliminates the possibility of equilibria with lingering strategic uncertainty.

3.7 Conclusion

In this paper, I have characterized a LCCC solution to a simple bargaining game for agents in two randomly matched populations. The result highlights the importance of subjective confidence in supporting some convention-affirming equilibria in such a setting, and provides proof-of-concept for the general value of studying this notion of equilibrium in bargaining situations.

There are a number of directions in which future work might extend these results to more complicated games and settings. Some natural possibilities include games with larger numbers of moves or potential rematches, games with incomplete information about which offers a player may rationally accept (i.e. that are better than no deal), settings in which the set of individually rational divisions for each side of the market varies over time (so that there can be changes such that e.g. firms which could afford to pay a high wage before no longer can, due to, say, changes in technology which only they observe directly), and settings in which the bargaining and matching structure is explicitly embedded in a larger market model (with, e.g. explicitly modeled demand, production technology, and/or entry/exit).

There is also a rationale for considering the effect of hypothetical policy changes which somehow shift the distribution of prior beliefs on one or both sides of the market (i.e. some kind of intervention making firms or workers more or less confident about what they can demand through some change exogenous to the game and matching/sampling process itself). This would be a 'comparative dynamics' exercise – comparing the equilibria in the population before and after the change, and the adjustment process between them – and as such would require appending an explicit dynamic process to the model from which adjustment dynamics were derived.

The contribution of this paper thus hopefully consists both of the specific results derived therein, and in opening up opportunities for these various further directions to be explored.

Conclusion

In this work, I have introduced the concept of convention-affirming equilibrium, and employed it to study a few specific games. This work should hopefully illustrate the value of the concept; that it is intuitive, relatively easy to use, and capable of generating interesting and novel outcomes even within fairly simple and standard games. Future work should both broaden and deepen our understanding of the concept itself – its foundations, its general properties, and general techniques which are useful in employing it – and seek to apply it to a wider set of contexts; both more elaborate models of the same economic phenomena studied here, and other economic applications of interest.

Bibliography

Agastya, Murali (1997), "Adaptive play in multiplayer bargaining situations." *Review of Economic Studies*, 64.

Agastya, Murali (1999), "Perturbed adaptive dynamics in coalition form games." *Journal of Economic Theory*, 89.

Alesina, Alberto and Paola Giuliano (2015), "Culture and institutions." *Journal of Economic Literature*, 53.

Banerjee, Abhijit V. (1992), "A simple model of herd behavior." *The Quarterly Journal of Economics*, 107.

Battigalli, Pierpaolo and Amanda Friedenberg (2012), "Forward induction reasoning revisited." *Theoretical Economics*, 7.

Battigalli, Pierpaolo and Marciano Siniscalchi (2002), "Strong belief and forward induction reasoning." *Journal of Economic Theory*, 106.

Battigalli, Pierpaolo and Marciano Siniscalchi (2003), "Rationalization and incomplete information." *Advances in Theoretical Economics*, 3.

Bednar, Jenna and Scott Page (2007), "Can game(s) theory explain culture?" *Rationality and Society*, 19.

Bergin, James and Barton L. Lipman (1996), "Evolution with state-dependent mutations." *Econometrica*, 64.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992), "A theory of fads, fashion, custom, and cultural change as informational cascades." *Journal of Political Economy*, 100.

Binmore, Ken and Larry Samuelson (1997), "Muddling through: Noisy equilibrium selection." *Journal of Economic Theory*, 74.

Binmore, Kenneth G. and Larry Samuelson (1992), "Evolutionary stability in repeated games played by finite automata." *Journal of Economic Theory*, 57.

Binmore, K.G. and M.J. Herrero (1988), "Matching and bargaining in dynamic markets." *Review of Economic Studies*.

Blonski, Matthias and Giancarlo Spagnolo (2015), "Prisoners' other dilemma." *International Journal of Game Theory*, 44.

dal Bo, Pedro and Guillaume R. Frechette (2018), "On the determinants of cooperation in infinitely repeated games: A survey." *Journal of Economic Literature*, 56.

Dekel, Eddie, Drew Fudenberg, and David K. Levine (1999), "Payoff information and self-confirming equilibrium." *Journal of Economic Theory*, 89.

Ellingsen, Tore (1997), "The evolution of bargaining behavior." *Quarterly Journal of Economics*.

Ellison, Glenn (1993), "Learning, local interaction, and coordination." *Econometrica*, 61.

Ellison, Glenn and Drew Fudenberg (1993), "Rules of thumb for social learning." *Journal of Political Economy*, 101.

Ellison, Glenn and Drew Fudenberg (1995), "Word-of-mouth communication and social learning." *The Quarterly Journal of Economics*.

Fearon, James D. and David D. Laitin (1996), "Explaining interethnic cooperation." *American Political Science Review*, 90.

Friedenberg, Amanda (2019), "Bargaining under strategic uncertainty: The role of second-order optimism." *Econometrica*, 87.

Fudenberg, Drew and Yuichiro Kamada (2015), "Rationalizable partition-confirmed equilibrium." *Theoretical Economics*, 10.

Fudenberg, Drew and Yuichiro Kamada (2018), "Rationalizable partition-confirmed equilibrium with heterogeneous beliefs." *Games and Economic Behavior*, 109.

Fudenberg, Drew and David K. Levine (1993), "Self-confirming equilibrium." *Econometrica*, 61.

Fudenberg, Drew and David K. Levine (1998), *The Theory of Learning in Games*. MIT Press.

Fudenberg, Drew and Eric Maskin (1990), "Evolution and cooperation in noisy repeated games." *American Economic Review*, 80.

Gale, Douglas (1986a), "Bargaining and competition part i: Characterization." *Econometrica*, 54.

Gale, Douglas (1986b), "Bargaining and competition part ii: Existence." *Econometrica*, 54.

Gale, Douglas (1987), "Limit theorems for markets with sequential bargaining." *Journal of Economic Theory*, 43.

Garcia, Julian and Matthijs van Veelen (2016), "In and out of equilibrium i: Evolution of strategies in repeated games with discounting." *Journal of Economic Theory*, 161.

Greif, Avner (1994), "Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies." *Journal of Political Economy*, 102.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath (2001), "In search of homo economicus: Behavioral experiments in 15 small-scale societies." *AEA Papers and Proceedings*.

Hudson, Aristos A. (2023), "Convention-affirming equilibrium." *Working paper*.

Hurkens, Sjaak and Nir Vulkan (2015), "Dynamic matching and bargaining with heterogeneous deadlines." *International Journal of Game Theory*, 44.

Imhof, Lorens A., Drew Fudenberg, and Martin A. Nowak (2005), "Evolutionary cycles of cooperation and defection." *Proceedings of the National Academy of Sciences, USA*, 102.

Kandori, Michihiro, George J. Mailath, and Rafael Rob (1993), "Learning, mutation, and long run equilibria in games." *Econometrica*, 61.

Kohlberg, Elon and Jean-Francois Mertens (1986), "On the strategic stability of equilibria." *Econometrica*, 54.

Krishnan, Venkatarama (2006), *Probability and Random Processes*. Wiley.

Lauermann, Stephan (2013), "Dynamic matching and bargaining games: A general approach." *American Economic Review*, 103.

Lewis, David (1969), *Convention: A Philosophical Study*. Blackwell.

McLennan, Andrew and Hugo Sonnenschein (1991), "Sequential bargaining as a noncooperative foundation for walrasian equilibrium." *Econometrica*, 59.

Milgrom, Paul R. (1981), "Good news and bad news: Representation theorems and applications." *The Bell Journal of Economics*, 12.

Nash, John (1953), "Two-person cooperative games." *Econometrica*, 21.

Osborne, Martin J. and Ariel Rubinstein (2003), "Sampling equilibrium, with an application to strategic voting." *Games and Economic Behavior*, 45.

Rescorla, Michael (2019), "Convention." In *The Stanford Encyclopedia of Philosophy* (Edward N. Zalta, ed.), summer 2019 edition, Metaphysics Research Lab, Stanford University.

Robson, Arthur J. (1990), "Efficiency in evolutionary games: Darwin, nash and the secret handshake." *Journal of Theoretical Biology*, 144.

Rubinstein, Ariel (1982), "Perfect equilibrium in a bargaining model." *Econometrica*, 50.

Rubinstein, Ariel and Asher Wolinsky (1985), "Equilibrium in a market with sequential bargaining." *Econometrica*, 53.

Rubinstein, Ariel and Asher Wolinsky (1990), "Decentralized trading, strategic behavior and the walrasian outcome." *Review of Economic Studies*, 57.

Saez-Marti, Maria and Jorgen W. Weibull (1999), "Clever agents in young's evolutionary bargaining model." *Journal of Economic Theory*, 86.

Salant, Yuval and Josh Cherry (2020), "Statistical inference in games." *Econometrica*, 88.

Stalnaker, Robert (1998), "Belief revision in games: Forward and backward induction." *Mathematical Social Sciences*, 36.

Vanderschraaf, Peter (2001), *Learning and Coordination: Inductive Deliberation, Equilibrium and Convention*. Routledge.

Volij, Oscar (2002), "In defense of defect." *Games and Economic Behavior*, 39.

Young, H. Peyton (1993a), "The evolution of conventions." *Econometrica*, 61.

Young, H.P. (1993b), "An evolutionary model of bargaining." *Journal of Economic Theory*, 59.