Depth of Reasoning and Higher Order Beliefs

Tomasz Strzalecki*
Harvard University

Abstract

As demonstrated by the email game of Rubinstein (1989), the predictions of the standard equilibrium models of game theory are sensitive to assumptions about the fine details of the higher order beliefs. This paper shows that models of bounded depth of reasoning based on level-k thinking or cognitive hierarchy make predictions that are independent of the tail assumptions on the higher order beliefs. The framework developed here provides a language that makes it possible to identify general conditions on depth of reasoning, instead of committing to a particular model such as level-k thinking or cognitive hierarchy.

(JEL C72, D03)

*E-mail: tomasz_strzalecki@harvard.edu. Part of this research was done while I was visiting the Economic Theory Center at Princeton University to which I'm very grateful for its support and hospitality. I thank Roland Benabou, Colin Camerer, Sylvain Chassang, Kim-Sau Chung, Vincent Crawford, Drew Fudenberg, David Laibson, Stephen Morris, Wojciech Olszewski, Wolfgang Pesendorfer, Marcin Peski, Ariel Rubinstein, Itai Sher, Marciano Siniscalchi, Dale Stahl, Georg Weizsäcker, and Asher Wolinsky and two anonymous referees for helpful comments and Jeff Ely for encouragement. I also thank the audiences of the student seminar at Northwestern, behavioral economics seminar at Princeton, and the North American Winter Meeting of the Econometric Society in Atlanta. All errors are mine. This version: August 29, 2014.

1 Introduction

One of the assumptions maintained in the standard equilibrium analysis of game theory is that agents have unlimited reasoning ability—they are able to perform arbitrarily complicated iterative deductions in order to predict their opponent's behavior. For instance, the assumption of common knowledge of rationality entails everyone being rational, everyone knowing that their opponent is rational, everyone knowing that their opponent knows that they are rational, and so on ad infinitum. Such higher order beliefs assumptions may be useful approximations in certain common and simple strategic situations, but one would not expect them to hold uniformly in all interactions. Even simple games that we encounter in everyday economic interactions are complicated enough to suggest that people may not have enough cognitive ability to solve them inductively.

This observation that people do not, in reality, take the inductive reasoning to its logical conclusion has been widely recognized, one of the quintessential illustrations being the email game of Rubinstein (1989).[1] In this game, two generals are facing a common enemy. There are two possible informational scenarios. In the first one, it is common knowledge that the enemy is weak and the generals would like to coordinate an attack on him. In the second scenario, the enemy can be either strong or weak, and the generals want to attack only in the latter case. Only one general knows for sure if the enemy is weak and this knowledge is later shared between the two generals through a back and forth email exchange. At some finite point in time, an email gets lost and this leaves the generals in a situation of "almost common knowledge," but not "common knowledge", i.e., they both know that the enemy is weak, they both know that they both know this, and so on, but only finitely many times; see Section 2 for a formal description of the game.

Intuitively, the difference between the two scenarios should be small, especially if the number of emails is large. However, as Rubinstein shows, this difference is critical. In the second scenario, no matter how many email messages get exchanged, the generals will not be able to coordinate successfully: the only equilib-

[1] This game is related to the coordinated attack problem in computer science, see, e.g., Halpern (1986)

rium involves not attacking the enemy despite the fact that a successfully coordinated attack is an equilibrium in the game with common knowledge. Rubinstein finds it hard to believe that the generals will try to outguess each other and fail to coordinate even in cases when the number of messages they exchange is very large and intuitively close to the simple game of the first scenario. He finds this discontinuity of behavior with respect to higher order beliefs counterintuitive and writes:

> The sharp contrast between our intuition and the game-theoretic analysis is what makes this example paradoxical. This game joins a long list of games such as the finitely repeated Prisoner's Dilemma, the chain store paradox and Rosenthal's game, in which it seems that the source of discrepancy is rooted in the fact that in our formal analysis we use mathematical induction while human beings do not use mathematical induction when reasoning. Systematic explanation of our intuition [...] is definitely a most intriguing question.

This paper provides precisely such a systematic explanation. The model studied here uses the recent non-equilibrium approach to strategic thinking.[2] The premise of this approach is that each agent has bounded depth of reasoning, and that the actual bound depends on his "cognitive type." An agent with bound k can perform at most k "steps of reasoning," i.e., can iterate the best response correspondence at most k times. These "level-k" or "cognitive hierarchy" models have been successful at accounting for many of the systematic deviations from equilibrium behavior, such as coordination in market entry games, overbidding in auctions, deviations from the unique mixed strategy equilibrium, and other phenomena; however, there are also environments in which the predictive power of such models is low: Ivanov, Levin, and Niederle (2010) and Georganas, Healy, and Weber (2010). This paper provides a general framework within which such games can be analyzed.

The main result of this paper, Theorem 3, is that in the email game there exists a

[2]See Nagel (1995); Stahl and Wilson (1994, 1995); Ho, Camerer, and Weigelt (1998); Costa-Gomes, Crawford, and Broseta (2001); Camerer (2003); Camerer, Ho, and Chong (2004); Costa-Gomes and Crawford (2006); Crawford and Iriberri (2007a,b); Crawford, Gneezy, and Rottenstreich (2008); Crawford, Kugler, Neeman, and Pauzner (2009); Healy, Georganas, and Weber (2010).

finite number of messages such that coordination is possible among all "cognitive types," no matter how high their bound, provided that they receive at least that many messages. This result is not a simple consequence of the fact that players are bounded, but rather is an outcome of the strategic interaction between the players and in particular their beliefs about the boundedness of their opponents. The intuition for the result is that if a high level type believes mostly in lower levels who themselves believe in even lower levels, and so on, then the behavior of all types is largely determined by the actions of the least cognitively able type, making the coordination possible. In the extreme case, if everyone believes only in the lowest type, then coordination is immediate. The result holds under a fairly general assumption on those beliefs, and an extension shows that even if sophisticated types without a cognitive bound are allowed, the above conclusion still holds provided the fraction of such infinite types is not too large.

While capturing Rubinstein's intuition, this simple and rigorous argument makes predictions that are consistent with observed behavior. In an experimental study, Camerer (2003) shows that in a variant of the email game the subjects were able to successfully coordinate after receiving a certain finite number of messages, thereby behaving according to the strategy described in the above paragraph. Interestingly, after several repetitions of the game the number of messages required for a coordinated attack increased, eventually rendering the coordination impossible.[3] This behavior suggests that models of bounded depth of reasoning may be useful for capturing the initial behavior in games, whereas in more frequent interactions a larger role may be played by the standard equilibrium models of game theory and the higher order beliefs considerations.

The formal result of this paper sheds light on a large literature in game theory that has developed in response to Rubinstein's (1989) paper. The goal of one of the branches of this literature is to restore the continuity of the equilibrium behavior by redefining the notion of distance on higher order beliefs. According to this new notion of distance, receiving even a very large number of messages does not bring the

[3]There is evidence that in similar coordination games (Cabrales, Nagel, and Armenter, 2007; Heinemann, Nagel, and Ockenfels, 2004) experimental subjects behave differently; since this paper focuses on theoretical aspects, these differences will not be discussed here.

agents close to common knowledge and for this reason the difference in behavior between the two scenarios is no longer paradoxical. By contrast, this paper remains faithful to the intuitive notion of distance and changes the solution concept to the one that preserves continuity.[4]

Another branch of literature that ensued from Rubinstein's (1989) paper studies global games: games where the multiplicity of equilibria is eliminated by a "contagion" argument, very much like the one used in the email game.[5] In a common knowledge game with multiple equilibria, a particular equilibrium is selected based on the proximity of this game to a game of "almost common knowledge" where the equilibrium is unique. The tools developed in this paper may be useful for analyzing the extent to which such selection arguments rely on unbounded depth of reasoning; for a related paper see Kneeland (2012).

In the course of developing the argument of this paper, a general model of bounded depth of reasoning is constructed. This model nests the models existing in the literature as special cases identified with the specific assumptions they make about the relationship between the bound on an agent's reasoning and his belief about other agents' boundedness. The "universal" model of this paper provides the requisite formal language that makes it possible to vary independently the assumptions about these two components (bound and belief), making it easier to understand the dependence of the results on particular assumptions. The notion of cognitive type space developed here allows for studying richer forms of dependency between the bound and the belief, which offers a new direction for the analysis of models of bounded depth of reasoning and its applications to various economic settings.

[4]Formally, cognitive rationalizability is upper hemi-continuous in the product topology on higher order beliefs.

[5]Like the email game, global games are dominant-solvable, they take infinitely many rounds to solve, and the standard solution differs starkly between the complete and incomplete information versions of the game, see, e.g., Carlsson and van Damme (1993), Morris and Shin (1998) and Frankel, Morris, and Pauzner (2003), among others.

2 Email Game

Two generals at different locations face a common enemy whose strength is unknown. They have to independently decide whether to Attack (A) or Not Attack (N). There are two possibilities: the enemy is either strong (s) or weak (w); both players put a common prior probability $\frac{1}{2}$ on s and $\frac{1}{2}$ on w. Initially the information about the state of nature is known only to player 1. However, the following information communication protocol is at work: if player 1 learns that the enemy is weak (and only in this case) an email is sent from his computer to player 2's computer. Player 2's computer is programmed to send a confirmation message back to player 1. Player 1's computer automatically sends a confirmation to player 2 and so on. There is a probability $\varepsilon > 0$ that each message can get lost, in which case the communication stops; with probability 1 the process will stop after a finite number of messages. Players can not influence this protocol and they have to make their decisions only after the communication stops. The state space is equal to:

$$\Theta = \{(0,0),(1,0),(1,1),(2,1),(2,2),\ldots\}$$

where in state $(0,0)$ the enemy is strong and in all other states he is weak. In each state (i,j) the number of messages sent by player 1 is equal to i and the number of messages sent by player 2 is j. Each player knows only his number; they never know if the reason for no reply is that their message got lost or the reply to their message got lost. The partitions and posteriors of players are shown below in Table 1.

player 1	1	$\frac{1}{2-\varepsilon}$	$\frac{1-\varepsilon}{2-\varepsilon}$	$\frac{1}{2-\varepsilon}$	$\frac{1-\varepsilon}{2-\varepsilon}$	...
Θ	$(0,0)$	$(1,0)$	$(1,1)$	$(2,1)$	$(2,2)$	...
player 2	$\frac{1}{1+\varepsilon}$	$\frac{\varepsilon}{1+\varepsilon}$	$\frac{1}{2-\varepsilon}$	$\frac{1-\varepsilon}{2-\varepsilon}$	...	

Table 1: Partitions and posteriors in the email game.

Players' payoffs depend on the strength of the enemy: if the enemy is strong Not Attacking is a dominant action, if he is weak the game is one of coordination;

see Table 2.[6]

	Attack	*Not Attack*
Attack	$-2, -2$	$-2, 0$
Not Attack	$0, -2$	$0, 0$

(a) Enemy is strong

	Attack	*Not Attack*
Attack	$1, 1$	$-2, 0$
Not Attack	$0, -2$	$0, 0$

(b) Enemy is weak

Table 2: Payoffs in the email game.

If it was *common knowledge* that the enemy is weak, i.e., that the payoff matrix is the one on the right, then it would be possible for the two generals to coordinate and decide to Attack. However, with the information structure described above, w is never common knowledge. For example, if the state of nature is $(2, 1)$, then both players know w, 1 knows that 2 knows w, 2 knows that 1 knows w, but 2 *doesn't* know that 1 knows that 2 knows w. Nevertheless, if the number of messages sent by both players is high, then they have *almost common knowledge* of w in the sense that they know w, they know that they know w and so on, many many times. According to Rubinstein's intuition, both situations are almost the same in the minds of the players, so their behavior should not be different. Unfortunately, the following puzzling result obtains.

Theorem 1 (Rubinstein)**.** *The unique rationalizable strategy profile of the email game is that both players choose Not Attack, regardless of how many messages they got.*

[6]The same payoff matrix is used in Dekel, Fudenberg, and Morris (2006); the numerical values of payoffs do not play any role and can be replaced by any other numbers as in Rubinstein (1989). The game is slightly different from Rubinstein's (1989) original formulation in that when the enemy is strong Attacking is strictly dominated, whereas in Rubinstein's original game Attacking is a (Pareto-dominated) Nash equilibrium. This modification makes the analysis of the game simpler by making coordination even more difficult: it eliminates the motive for Player 1 to Attack when the enemy is strong, thus making the results of this paper even stronger.

3 Limited depth of reasoning

The literature on limited depth of reasoning postulates that each player has a bound k on reasoning, where $k \in \{0, 1, \ldots\}$. A player with bound $k = 0$ is a nonrational and nonstrategic type which is allowed to take any action; its behavior is used by other players to anchor their beliefs. All other players are rational, i.e., best respond to *some* belief about their opponents. A player with $k = 1$ best responds to the belief that his opponents are of type $k = 0$. Beliefs of players with $k > 1$ are defined according to some pre-specified rule: Some authors (e.g., Costa-Gomes and Crawford, 2006) assume that a player with bound k believes that his opponents' bound is $k - 1$, while other authors (e.g., Camerer et al., 2004) assume that a player with bound k has a nondegenerate belief on the set $\{0, 1, \ldots, k - 1\}$.

The approach of this paper is to use a general notion of a *cognitive type space*, which can accommodate any assumption about the beliefs of player k about his opponent. Furthermore, for the purposes of this paper it will be useful to uncouple the two pieces of the description above: the *cognitive type space*, i.e., the description of depth of reasoning and beliefs, and the *solution concept*, i.e., the description of the actions taken by each cognitive type (act irrationally, or best respond to a belief).

3.1 Notation

For any measurable space X with a σ-algebra Σ the set of all σ-additive probability measures on X is denoted $\Delta(X)$. We consider $\Delta(X)$ as a measurable space with the σ-algebra that is generated by all sets of the form $\{\mu \in \Delta(X) \mid \mu(E) \geq p\}$ for $E \in \Sigma$ and $p \in [0, 1]$. For $x \in X$ the Dirac measure on x is denoted δ_x.

3.2 Cognitive type spaces

The main notion introduced in this section generalizes various models studied in the literature while remaining faithful to one of their main assumptions, which is that agents are concerned only about opponents with bounds lower than themselves. There seem to be two main justifications for this assumption in the literature. The first one is to escape the fixed point logic of equilibrium, where a player

best responds to an opponent that best responds to that player. The second reason is that an agent with a low bound on reasoning simply cannot predict what an opponent with a high bound will do because that would require him use more steps of reasoning that are available to him. He is thus forced to *model* an opponent with a high bound as one with a lower bound that he is able to wrap his mind around.

The construction generalizes the existing models in the direction of allowing an agent with a given bound to have arbitrary beliefs about his opponents' bound (as long as it is below his own). This freedom offers a systematic way of nesting different modeling assumptions about beliefs inside one unified model and thereby examining which results are a simple consequence of the fact that the agents are bounded and which ones are more subtle and rely on the specific assumptions about the perception of the boundedness.

Similarly to the literature on the foundations of games of incomplete information there are two alternate approaches to the description of the beliefs of a player about the bounds of his opponents: hierarchies of beliefs and type spaces.[7]

The hierarchies of beliefs approach is an extensive description of the beliefs of an agent about the bound of his opponents, together with his beliefs about their beliefs about the bound of their opponents, and so on. Appendix A studies this approach in detail and its connection to the approach of cognitive type spaces.

The main focus of the paper is on the application of Harsanyi's idea of a type space in the context of depth of reasoning. A *cognitive type* of a player is a compact description of his hierarchy: it is composed of his own bound together with his belief about the cognitive types of his opponent. A *cognitive type space* is a collection of cognitive types of each player.

Definition 1 (Cognitive type space)**.** A *cognitive type space* C is a tuple $(C_i, k_i, \nu_i)_{i=1,\dots,I}$ such that C_i is a measurable space and the measurable mappings $k_i : C_i \to \mathbb{N}$ and $\nu_i : C_i \to \Delta(C_{-i})$ are such that

$$\nu_i(c_i)\left(\{c_{-i} \in C_{-i} \mid k_{-i}(c_{-i}) < k_i(c_i)\}\right) = 1 \text{ for all } c_i \text{ with } k_i(c_i) > 0. \tag{1}$$

[7] See, e.g., Harsanyi (1967); Mertens and Zamir (1985); Brandenburger and Dekel (1993); Heifetz and Samet (1998).

In this notation $c_i \in C_i$ denotes the cognitive type of player i, the number $k_i(c_i)$ is the cognitive ability of this type, i.e., his bound, and the distribution $\nu_i(c_i)$ is the belief that this type has about the cognitive types of his opponents. Equation (1) ensures that all players believe that their opponents' ability is below their own (the notation $k_{-i}(c_{-i}) < k_i(c_i)$ means that $k_j(c_j) < k_i(c_i)$ for all $j \neq i$).[8] [9]

This simple construction can capture many different assumptions about beliefs. For example, the collection of all possible cognitive hierarchies is a cognitive type space, which does not involve any assumptions about the beliefs.

Example 1 (Universal cognitive type space)**.** The universal cognitive type space $(C_i^*, k_i^*, \nu_i^*)_{i=1,\ldots,N}$ constructed in Appendix A is a cognitive type space, which captures all possible hierarchies of beliefs.

A feature of Example 1 is that there are multiple cognitive types with the same bound k that are distinguished by the beliefs they hold about the bound of their opponents. The cognitive type spaces used in the literature typically involve an assumption that rules this out: the level k of a type uniquely determines his beliefs.

Property 1 (Level determines beliefs)**.** For all i and all c_i the function $\nu_i(c_i)$ depends on c_i only through $k_i(c_i)$.

In such type spaces all agents of the same level have the same beliefs. This can be modeled as there being only one cognitive type of each level.[10] Any such type space is isomorphic to one with $C_i := \mathbb{N}$ and $k_i(k) := k$ for all $k \in \mathbb{N}$. The following two examples illustrate the most frequently used spaces.

Example 2 (Immediate-predecessor type space)**.** In this type space, all cognitive types believe with probability one that the opponent is their immediate predecessor in the cognitive hierarchy, i.e., $\nu_i^{\text{IP}}(k) = \delta_{k-1}$, a Dirac measure concentrated on $k-1$.

[8] The beliefs of the nonstrategic type 0 player don't matter because he is "just acting".

[9] Formally, the cognitive state space is a Harsanyi type space, where besides beliefs each player has an additional source of information: his cognitive level. The cognitive type space encodes all hierarchies of beliefs with the assumption of common belief that "your level is smaller than mine."

[10] As explained before, the cognitive type space is a description of the cognitive abilities of the agent, while the solution concept is a description of his actions. Thus, there being only one type of each level does not impose any restrictions on the number of possible actions taken. The multiplicity of actions for each cognitive type is modeled as part of a solution concept, described in Section 3.3.

In other words, a level 1 player believes that his opponent is of level 0, a level 2 player believes that his opponent is of level 1, etc. This type space was used for example by Costa-Gomes and Crawford (2006) and others and together with the solution concept, to be discussed in Section 3.3, it constitutes the well known "level-k model".

This type space has a property that, although as k grows types become smarter, their beliefs become farther away from any fixed distribution on the population of players. By contrast, in the following example a level k agent believes not only in level $k-1$ but also in lower levels, which results in the convergence of their beliefs as k grows.

Example 3 (Common-conditionals type space)**.** In this type space, beliefs are constructed as conditionals of a fixed, full support distribution $\lambda \in \Delta(\mathbb{N})$. Thus, $\nu_i^{\mathrm{CP}(\lambda)}(k)(n) = \frac{\lambda(n)}{\sum_{l=0}^{k-1}\lambda(l)}$ for $n < k$ and zero otherwise. In other words, a level 1 player believes that his opponent is of level 0, but a level 2 player believes that his opponent is of level 1 *or* level 0 and so on for higher levels. The role of λ resembles the role of the common conditionals in the Harsanyi type spaces. Such type spaces were used for example by Camerer et al. (2004), where λ was taken to be a Poisson distribution and together with the solution concept, to be discussed in Section 3.3, it constitutes the well known "cognitive hierarchy model".[11] In empirical estimation of this model, λ is taken to be the "objective distribution of types." Such an objective distribution is also necessary for estimation of the level-k model; however, there is no counterpart of it in the model itself.

Example 3 has the property that for any two cognitive levels k and n they share conditional beliefs on types lower than $\min\{k, n\}$. The following weaker condition does not restrict conditional beliefs, but requires that, as k grows to infinity, the beliefs converge to some distribution over levels (which doesn't have to be the common prior).

[11] See also Stahl (1993) and Stahl and Wilson (1995).

Property 2 (Convergent beliefs)**.**

$$\lim_{k\to\infty} \nu_i(k) \text{ exists.}^{12}$$

Property 2 is satisfied in Example 3 but violated by Example 2. The following even weaker condition, which is key to the results to follow, limits the extent to which the beliefs can diverge. It says that as k grows, agents put less and less weight on the types immediately below them: for any increasing sequence of levels there is a level in the sequence, such that all cognitive types of that level put at most probability κ on the opponent being above the previous level in the sequence.

Property 3 (Nondiverging beliefs)**.** There exists a constant $\kappa \in [0,1)$ such that for any strictly increasing sequence $\{k^n\}_{n=0}^{\infty}$ of natural numbers

$$\inf_n \sup_{c_i \in \{c_i \in C_i | k_i(c_i) = k^n\}} \nu_i(c_i)\left(\{c_{-i} \in C_{-i} \mid k_{-i}(c_{-i}) \geq k^{n-1}\}\right) < \kappa.$$

In the presence of Property 1, this condition boils down to the following simple one. There exists a constant $\kappa \in [0,1)$ such that for any strictly increasing sequence $\{k^n\}_{n=0}^{\infty}$ of natural numbers

$$\inf_n \nu_i(k^n)\left(\{k_{-i} \geq k^{n-1}\}\right) < \kappa.$$

It is immediate that the beliefs in the immediate-predecessor type space of Example 2 violate Property 3, while the beliefs in the common conditionals type space of Example 3 satisfy it with $\kappa = 0$. In general any beliefs that converge, i.e., satisfy Property 2, satisfy this property with $\kappa = 0$. As the following example shows, also beliefs that do not converge can satisfy Property 3.

Example 4 (A mixture of immediate-predecessor and common-conditionals type spaces)**.** Let $\alpha \in (0,1)$ and $\lambda \in \Delta(\mathbb{N})$ be a fixed, full support distribution. Define

[12] Formally, $\nu_i(k) \in \Delta\{0,\ldots,k-1\}$, so the limit is taken after imbedding all those measures in $\Delta(\mathbb{N})$. The topology on $\Delta(\mathbb{N})$ is the weak* topology.

$\nu_i(k) = \alpha\nu_i^{\text{IP}}(k) + (1-\alpha)\nu_i^{\text{CP}(\lambda)}(k)$. It is easy to verify that

$$\lim_n \nu_i\left(k^n\right)\left(\left\{k_{-i} \geq k^{n-1}\right\}\right) = \alpha,$$

that is, such beliefs satisfy Property 3 with $\kappa \in (\alpha, 1)$.

3.3 Solution concepts

This section defines the solution concept, which is a consistency condition on the actions taken by all cognitive types. A *game* is a tuple $G = (u_i, A_i)_{i=1,\ldots,I}$, where for each i, A_i is a set of *actions* and $u_i : A_1 \times \cdots \times A_I \to \mathbb{R}$ is a *payoff function*.[13] Given a game G and a cognitive type space C a strategy $\sigma_i : C_i \to A_i$ tells the agent what to do for each possible cognitive type that he might have. A strategy profile $(\sigma_i)_{i=1,\ldots,I}$ is a *cognitive equilibrium*[14] if and only if for all $i = 1, \ldots, I$, for all $c_i \in C_i$ with $k_i(c_i) > 0$, and all $a_i \in A_i$

$$\int u_i\left(\sigma_i(c_i), \sigma_{-i}(c_{-i})\right) \, \mathrm{d}\nu_i(c_i)(c_{-i}) \geq \int u_i\left(a_i, \sigma_{-i}(c_{-i})\right) \, \mathrm{d}\nu_i(c_i)(c_{-i}).$$

Formally, the notion of cognitive equilibrium can be seen as Bayesian Nash equilibrium on the cognitive space and letting every type with $k_i(c_i) = 0$ have a constant utility. Intuitively, cognitive equilibrium defines how beliefs about levels interact with beliefs about actions. It allows type $k = 0$ to take any action. Type $k = 1$ best responds to the action that type $k = 0$ is taking. Type $k = 2$ best responds to those two actions *and* his beliefs about the proportion of the types $k = 0$ and $k = 1$ of his opponent. Actions of higher types are determined similarly. A cognitive

[13] This formulation captures both normal-form games, as well as Bayesian games in their "type-agent representation," see, e.g., Myerson (1991). Note that this approach makes beliefs about the state of nature implicit in the payoffs (they are integrated out); thus, versions of a level-1 player who have different beliefs about their payoffs are modeled as distinct players with different (deterministic) payoff functions. For a model with explicit beliefs about the state (but implicit beliefs about cognitive ability), see Kets (2014) and Heifets and Kets (2013).

[14] "Cognitive equilibrium" is perhaps not the most fortunate term, but it retains the main idea of equilibrium, which is that players have correct beliefs about their opponents' strategies (while having possibly incorrect beliefs about their types or cognitive types). In particular, in case of a tie they have a correct belief about which of the actions will be taken by an opponent; the set-valued concept discussed below relaxes this assumption.

equilibrium can be thus computed in a simple iterative manner, given the action of type $k = 0$. The literature makes different assumptions about this action, sometime it is a uniform randomization over all the actions, sometime it is an action that is focal in the given game. The above solution concept can accommodate any such assumption.

Another possible solution concept is set-valued, analogously to rationalizability. Such *cognitive rationalizability* assigns to type $k = 0$ the set of all feasible actions, and then the iterative procedure described above determines the set of actions of each cognitive type. If C is the immediate predecessor type space, then this boils down precisely to iterative deletion of strictly dominated strategies, where the set of actions of each k shrinks as k increases, eventually converging to the set of rationalizable strategies. Under other assumptions on C, for example if it satisfies the nondiverging beliefs assumption, this convergence is slowed down and the limit set of actions is strictly larger than rationalizability.[15,16]

Finally, it is sometimes assumed that there are multiple instances of every type, each believing in a different action taken by the opponent. For example, Crawford and Iriberri (2007b) have two possible actions for $k = 0$ (uniform randomization and truthful reporting) and two possible actions for $k = 1$ (best response to randomization, best response to truthful reporting), and so on. Such multiplicity can be easily modeled as a refinement of the above cognitive rationalizability solution where the starting value for $k = 0$ is not the set of all actions, but some nonsingleton selection of them. The advantage of the conceptual separation of the cognitive type space and the solution concept that this paper adopts is that the assumptions about the belief structure can be discussed independently of the actions taken by the player of level $k = 0$.

[15] See Stahl (1993) who uses a similar solution concept with a common-conditionals type space.

[16] Another possible solution concept, considered by Ho et al. (1998) and Rogers, Palfrey, and Camerer (2009) is where each agent's action is a noisy best response to his beliefs.

4 The Analysis of the Email Game

This section discusses the predictions of the models of bounded depth of reasoning in the email game and compares various assumptions on beliefs, i.e., various cognitive type spaces.

The construction starts out by describing the behavior of the cognitive type $k = 0$. For both players, this strategy is simply to attack regardless of the number of messages received, see Table 3.[17]

player 1, $k = 0$	A	A		A		...
	$(0,0)$	$(1,0)$	$(1,1)$	$(2,1)$	$(2,2)$	...
player 2, $k = 0$	A		A		...	

Table 3: Level 0 actions in the email game.

Now, derive the behavior of types with $k = 1$. Player 1 in the information set $\{(0,0)\}$ will play N because it is a dominant action when the enemy is strong. In other information sets, his action will remain A because it is the best response to player 2 attacking and the enemy being weak. Similarly, for sufficiently small values of ε, Player 2 will choose N in the information set $\{(0,0),(1,0)\}$ and keep playing A in the remaining information sets; see Table 4.

player 1, $k = 1$	N	A		A		...
player 1, $k = 0$	A	A		A		...
	$(0,0)$	$(1,0)$	$(1,1)$	$(2,1)$	$(2,2)$	...
player 2, $k = 0$	A		A		...	
player 2, $k = 1$	N		A		...	

Table 4: Level 0 and 1 actions in the email game.

[17] This assumption is made here for simplicity of exposition. Alternative, perhaps more realistic, assumptions are discussed in Section 4.3.1.

A key point in the analysis involves a situation where player i of cognitive level k faces player j whose levels $l = 0, 1, \ldots, k-2$ chose A in both of the information sets considered by i to be possible, but type $l = k-1$ chooses N in one of these sets and A in the other. In such a situation player i will be referred to as being "on the fence."

The first example of the situation of being "on the fence" is the strategic choice of player 1 with $k = 2$. In the information set $\{(0,0)\}$ it is still dominant for him to choose N. However in the information set $\{(1,0),(1,1)\}$ his action will depend on his beliefs. If he puts a lot of weight on player 2 being of type $k = 1$ (who plays N in the information set $\{(0,0),(1,0)\}$ and A in the information set $\{(1,1),(2,1)\}$), he will choose the safe action N because the choice of A would involve a negative expected payoff. However, if he puts enough weight on player 2 being of type $k = 0$ (who plays A in both information sets that player 1 considers plausible), the expected payoff of playing A will become positive and player 1 will choose A himself; see Table 5.

player 1, $k = 2$	N	**A or N**	A		...
player 1, $k = 1$	N	A	A		...
player 1, $k = 0$	A	A	A		...
	$(0,0)$	$(1,0)$	$(1,1)$	$(2,1)$	$(2,2)$ ...
player 2, $k = 0$	A		A		...
player 2, $k = 1$	N		A		...

Table 5: Level 0, 1, and 2 actions in the email game.

Players of higher types also face the situation of being "on the fence": for example, if player 1 of type $k = 2$ chose N, then the behavior of player 2 of type $k = 3$ in the information set $\{(1,1),(2,1)\}$ will depend on his subjective probability of types $k = 2$ versus $k = 0, 1$ of player 1. For this reason, the further analysis of the game depends on the cognitive type space.

4.1 Lower Bound on Cooperation

The first step in the analysis considers a type space where player i who is "on the fence" always puts a high enough probability on player j being of a high cognitive type who plays N in one of the information sets considered by i. An extreme case of such a type space is the Immediate-predecessor type space of Example 2, where player i who is "on the fence" always puts probability 1 on j playing N in one of the information sets considered by i. Note, that this type space yields a lower bound on the occurrences of cooperation because player i of level k who is "on the fence" in a given information set always plays A in information sets to the right of that information set.

Theorem 2. *For any cognitive type space and any cognitive equilibrium of the email game where players of level $k = 0$ always choose A, player 1 of cognitive type k chooses A upon receiving more than $\frac{k}{2}$ messages and player 2 of cognitive type k chooses A upon receiving more than $\frac{k-1}{2}$ messages.*

Observe, that this lower bound on cooperation implies that upon receiving a large number of messages only the players of high cognitive types will choose the non-cooperative action N. This means that under a fixed probability distribution on types (which could, for example, describe the distribution of types from the point of view of the analyst) the probability of N being chosen converges to zero as the number of messages increases.

Corollary 1. *For any cognitive type space and any probability distribution on it, for any cognitive equilibrium of the email game where players of level $k = 0$ always choose A the probability that A is chosen upon receiving n messages converges to 1 as $n \to \infty$.*

Note, however, that although players choose A after receiving many messages, the number of required messages depends on the cognitive type of the player. Under the Immediate-predecessor type space the number of messages required for a player of level k to start playing A increases to infinity with k. Observe, that in the limit as $k \to \infty$, N is always chosen regardless of the number of messages, which is exactly the behavior coming from the iterative elimination of dominated strategies described in Theorem 1. In particular, there is no equilibrium in which a number of messages exists which makes all cognitive types play A.

4.2 Upper Bound on Cooperation

The second step of the analysis studies cognitive type spaces where players' beliefs about of the boundendess of their opponents slow down the iterative elimination process and create a uniform bound on the number of messages required for a player of any cognitive type to play A. In these cases there is at least one situation of player i of level k being "on the fence" and putting high enough a probability on player j being of level $l < k - 1$, so that he chooses A instead of N.

For example, in the first situation of being "on the fence" described above, if the subjective probability that $k = 2$ puts on $k = 0$ is high enough to convince him to play A, the unravelling will stop at $n = 1$ and all players with higher levels will play A after receiving one or more messages. For other type spaces the reasoning will have to continue; however, if in the limit the players "being on the fence" believe with sufficiently high probability that their opponent is below their immediate predecessor, the reasoning will have to stop at some number of messages n. The following theorem makes this idea formal.

Theorem 3. *For any cognitive type space satisfying Property 3 with a constant $\kappa = \frac{2-\varepsilon}{3}$ there exists a cognitive equilibrium of the email game and a number of messages n such that all cognitive types of both players choose to Attack if they receive n or more messages.*

Theorem 3 is not a simple consequence of the fact that players are bounded, but rather is an outcome of the strategic interaction between the players and in particular their beliefs about the boundedness of their opponents. The intuition for the result is that if a high level type believes mostly in lower levels who themselves believe in even lower levels, and so on, then the behavior of all types is largely determined by the actions of the least cognitively able type, making the coordination possible.

The content of Theorem 3 can be also expressed in the language of cognitive rationalizability. In that language, there exists n such that if players receive at least n messages (there is mutual knowledge of the state of order at least n), cognitive rationalizability predicts the same behavior as in the complete information game (common knowledge of the state). Therefore "almost common knowledge" and "common knowledge" result in the same predictions.

4.3 Robustness of the Results

4.3.1 Actions of level $k = 0$

Note that there exists the "bad" equilibrium where players of level $k = 0$ choose N independently of their information set and consequently all types of both players choose not to attack.[18] However, the "good" equilibrium does not rely on the level 0 agent attacking in all information sets. An alternative specification of the level 0 strategies involves playing N for the first m messages and A thereafter. Such a modeling choice results in the same conclusion as Theorems 2–3, no matter how high m is, and has the feature that all players, even the level $k = 0$ player, are behaving rationally, and in fact, for high values of m the players have "almost common knowledge of rationality". A third alternative for specifying the level 0 action is a mixed strategy choosing A with probability α and N with probability $1 - \alpha$. The payoffs of the game are chosen so that whenever an agent faces $\alpha = 0.5$, it is optimal for him to choose N, which is the reason why Theorem 1 holds for any $\varepsilon > 0$. However, any $\alpha > \frac{2}{3}$ guarantees that Theorems 2–3 hold.[19]

4.3.2 Boundedness of types

It is important to observe that Theorem 3 does not rely on everyone in the population being boundedly rational. In particular it could be assumed that there is a proportion δ of *sophisticated* players who are unboundedly rational and have a correct perception of δ. Formally, let $\lambda \in \Delta(\mathbb{N})$ and let ∞ denote the sophisticated type.[20] His belief $\nu_i^{\delta,\lambda}$ on $\mathbb{N} \cup \{\infty\}$ is defined by $\nu_i^{\delta,\lambda}(\infty) = \delta$ and $\nu_i^{\delta,\lambda}(k) = (1-\delta)\cdot\lambda(k)$. Let $\hat{C}_i^{\delta,\lambda} = C_i \cup (\{\infty\} \times \{\nu_i^{\delta,\lambda}\})$ and $\hat{C}^{\delta,\lambda} = \hat{C}_1^{\delta,\lambda} \times \hat{C}_2^{\delta,\lambda}$.

[18] The existence of this equilibrium is not troubling, as (N, N) is an equilibrium of the complete information game. It is rather the inexistence of the cooperative equilibrium in the standard setting (Theorem 1) that was the focus of Rubinstein's (1989) paper.

[19] Additionally, it can be shown that for any ε the payoff matrix of the game can be modified (by setting $-L$ for every occurrence of -2 in Table 2 and M for every occurrence of 1) so that the results of Theorems 2–3 obtain, while the only rationalizable outcome still involves not attacking (i.e., Theorem 1 still holds) as long as $(1 - \varepsilon)M < L < M$.

[20] Sophisticated types are similar to "worldly" types of Stahl and Wilson (1995), who with probability δ believe in the "naive Nash types". It follows from the proof of Theorem 4 that the same conclusion holds for "worldly" types.

Theorem 4. *For any cognitive type space C satisfying Property 3 with $\kappa = \frac{2-\varepsilon}{3}$, for any $\lambda \in \Delta(\mathbb{N})$, and for any $\delta < \frac{1}{3}$ there exists a cognitive equilibrium on $\hat{C}^{\delta,\lambda}$ with a number of messages n such that all cognitive types, including the sophisticated type, of both players choose to Attack if the number of their messages is bigger or equal than n.*

This theorem shows that the results of Theorem 3, i.e., the insensitivity to higher order beliefs, are not a simple consequence of the boundedness of the players, but rather of the assumption about the perception of the boundedness of the opponents. It is the strategic interaction of players under this assumption that delivers the result. Adding the unbounded players to the picture does not reverse the result because of the beliefs those unbounded players are endowed with: they put a sufficiently high weight ($1 - \delta > \frac{2}{3}$) on the bounded players. This insight is more general that the context of this theorem. The key role is played by the beliefs, rather than just the boundedness, because increasing the bound of a given agent without altering his beliefs leads to the same behavior, while keeping his bound constant but altering beliefs may lead to a change in behavior.

5 Discussion and relation to literature

5.1 Experimental work

There is an intriguing relationship between the results of this paper and the experimental findings reported in (Camerer, 2003, pp. 226–232). Experimental subjects were playing a version of the email game several times (with random repairings to avoid repeated games considerations) with strategies truncated at some number of messages.[21] In all repetitions of the game the fraction of subjects choosing to attack was increasing in the number of messages. In the early repetitions very few subjects chose N and the fraction of subjects choosing N after seeing six or more messages was zero. This is precisely the picture one would expect from Theorem 3.

Interestingly, after several repetitions of the game, the threshold number of messages after which the subjects switch to playing A started increasing, eventually

[21] The fixed upper bound on the number of messages makes this game formally different, especially in the light of the results of Binmore and Samuelson (2001).

surpassing the upper bound imposed on the number of messages in the experimental design. This suggests that the depth of reasoning considerations are especially important in "new" games, whereas in more frequent interactions a larger role may be played by the standard models of game theory and therefore the importance of higher order beliefs considerations may be higher.

One of the possible explanations for this increase in the threshold is that players updated their beliefs about the distribution of cognitive types in the player population (the matching was random and therefore precluded learning about any fixed player). Alternatively, the increased exposure to the game itself may have given the players more time to analyze the game and their own cognitive types increased.

5.2 Topologies on Higher Order Beliefs

As mentioned in the introduction, an important branch of game theory has developed in reaction to Rubinstein's observation, with the goal of identifying the notion of distance on higher order beliefs that would restore the continuity of the standard solution concepts. Under the natural notion of distance (product topology) the situation of "almost common knowledge of the game" (having received many messages) is close to the "common knowledge of the game," however the equilibrium behavior differs between the two situations. Monderer and Samet (1989) and most recently Dekel et al. (2006) and Chen, Di Tillio, Faingold, and Xiong (2009a,b) pin down a stronger topology on beliefs that makes behavior continuous in all games. Under this notion "almost common knowledge" is by definition far from "common knowledge," precisely because behavior differs between the two situations.

In contrast, the exercise in this paper is to fix the natural notion of distance (product topology) and find a solution concept (cognitive equilibrium or cognitive rationalizability) that is continuous in that topology. Under this interpretation, "almost common knowledge" is close to "common knowledge" and the behavior predicted in those two situations is close (in fact identical).

This interpretation of the findings is consistent with the recent contribution of Weinstein and Yildiz (2007), who show that no refinement of rationalizability that allows for two equilibria in the full information game can be continuous in the product topology. In order to obtain continuity, the solution concept has to be

changed to one that allows for strategies that are not rationalizable. Cognitive rationalizability has precisely this desired property: under the nondiverging beliefs assumption the solution concept is larger than rationalizability.

5.3 Mechanism Design

This section discusses the contrast between two notions of robustness in mechanism design. A recent literature (e.g., Bergemann and Morris, 2005), motivated by the Wilson (1987) doctrine of successive elimination of the common knowledge assumptions, remains ignorant about the higher order beliefs of the agents and is looking for mechanisms that do well under many different specifications of beliefs. Because in standard solution concepts behavior is discontinuous in beliefs, the designer has to be very conservative in his choice of the mechanism in order to ensure the desired outcome, especially if, following Weinstein and Yildiz (2007), he thinks that the product topology "captures a reasonable restriction on the researcher's ability to observe players' beliefs."

On the other hand, recent papers by Crawford and Iriberri (2007b) and Crawford et al. (2009) study design with agents who have a bounded depth of reasoning. With the "immediate-predecessor cognitive type space," Crawford and Iriberri (2007b) show that bounded depth of reasoning may lead to overbidding in auctions, which allows for an explanation of the winner's curse. Crawford et al. (2009) is a careful study of the optimal mechanism design problem taking into account the fact that players' cognitive abilities may be bounded. One of the key insightful observations is that in the setting of mechanism design, what is being designed is possibly a brand new game, which is likely to be played for the very first time. This makes the level-k models a relevant tool of analysis, but potentially introduces a concern about the robustness of the mechanism to the actions of lower types.

Because, as the results of this paper suggest, level-k behavior is not sensitive to the common knowledge assumptions, the first concern for robustness is attenuated. Those models are therefore a sensible alternate way of implementing the Wilson doctrine. However, they introduce another concern: instead of worrying about higher order beliefs, the designer has to worry about the lower order cognitive

types. The notion of the cognitive type space that this paper introduces makes it possible to explore this tradeoff between the two notions of robustness by varying the assumption about the beliefs about the players' depth of reasoning.

A Appendix: Universal Cognitive Type Space

This appendix defines the collection of all possible *cognitive hierarchies*, i.e., beliefs about the opponent's depth of reasoning and collects them in the *universal cognitive type space*.

A cognitive hierarchy of player i is a belief about the cognitive ability of $-i$, together with i's belief about $-i$'s belief about the cognitive ability of i, etc. The set Z_i^k denotes all possible cognitive hierarchies of a player with level k of cognitive ability and is defined recursively as follows.

$$\begin{aligned} Z_i^0 &:= \{0\} \\ Z_i^1 &:= \{1\} \times \Delta\left(Z_{-i}^0\right) \\ Z_i^2 &:= \{2\} \times \Delta\left(Z_{-i}^0 \cup Z_{-i}^1\right) \\ &\vdots \\ Z_i^k &:= \{k\} \times \Delta\left(\bigcup_{l=0}^{k-1} Z_{-i}^l\right) \end{aligned}$$

The set Z_i^0 is trivially a singleton, as the level 0 type does not act on any beliefs. Similarly, the set Z_i^1 is a singleton because the level 1 type believes in only one type of opponent: namely level 0. Beliefs of higher levels can be more complicated. For instance various types of level 2 of player i are distinguished by their beliefs in the proportion of levels 0 and 1 of players $-i$. Various types of level 3 of player i are distinguished by their beliefs in the proportion of levels 0, 1, and 2 of players $-i$ and their beliefs about the mix of various types of level 2 of players $-i$ (i.e. what mix of levels 0 and 1 of player i do the players $-i$ of level 2 believe in).

The set $\bigcup_{k=0}^{\infty} Z_i^k$ collects all possible cognitive hierarchies. Note that all hierarchies of beliefs are finite because of the assumption that all players have finite bounds and believe that their opponents have bounds strictly lower than themselves. We now show that the set $\bigcup_{k=0}^{\infty} Z_i^k$ with the natural belief mapping constitutes a type space. To be precise, for it to be a type space in the sense of Definition 1, the cognitive bound and belief mappings need to be measurable. Endow each set Z_{-i}^k with the product σ-algebra and each set $\bigcup_{l=0}^{k-1} Z_{-i}^l$ and $\bigcup_{l=0}^{\infty} Z_i^l$ with the direct

sum σ-algebra (see Fremlin, 2001, Section 214).

Definition 2 (Universal Cognitive Type Space). $C_i^* := \bigcup_{k=0}^{\infty} Z_i^k$; for any $c_i^* = (k_i, \nu_i) \in C_i^*$ let $k_i^*(c_i^*) := k_i$ and for any measurable $F_{-i} \subset C_{-i}^*$, $\nu_i^*(c_i^*)(F_{-i}) := \nu_i\left(F_{-i} \cap \bigcup_{n=0}^{k_i-1} Z_{-i}^n\right)$.

The following lemma ensures the correctness of this definition.

Lemma 1.

(i) For any measurable set $F_{-i} \subset C_{-i}^*$ and any $k \in N$ the set $F_{-i} \cap \bigcup_{n=0}^{k-1} Z_{-i}^n$ is measurable in $\bigcup_{n=0}^{k-1} Z_{-i}^n$.

(ii) The mapping $k_i^* : C_i^* \to \mathbb{N}$ is measurable.

(iii) The mapping $\nu_i^* : C_i^* \to \Delta(C_{-i}^*)$ is measurable.

Proof. Proof of (i): follows from the definition of the direct sum σ-algebra on sets $\bigcup_{l=0}^{\infty} Z_{-i}^l$ and $\bigcup_{l=0}^{k-1} Z_i^l$.
Proof of (ii): It suffices to observe that for each $n \in \mathbb{N}$ the set $(k_i^*)^{-1}(n) = \{c_i^* \in C_i^* | k_i(c_i^*) = n\} = Z_i^n$ is measurable in C_i^* by definition of the the direct sum σ-algebra.
Proof of (iii): By definition it suffices to show that for any measurable set $E \subseteq \Delta(C_{-i}^*)$ the set $(\nu_i^*)^{-1}(E)$ is measurable in C_i^*. It suffices to check this definition only for the generators of the σ-algebra on $\Delta(C_{-i}^*)$, i.e., sets of the form $E = \{\mu \in \Delta(C_{-i}^*) | \mu(F_{-i}) \geq p\}$ for some $p \in [0, 1]$ and some measurable set $F_{-i} \subseteq C_{-i}^*$. Thus, by the definition of the mapping ν_i^*, it suffices to show that the set $\{(k_i, \nu_i) \in C_i^* | \nu_i(F_{-i} \cap \bigcup_{n=0}^{k_i-1} Z_{-i}^n) \geq p\}$ is measurable in C_i^*. By definition of the the direct sum σ-algebra this set is measurable in C_i^* if and only if for each $k \in \mathbb{N}$ the set $\{\nu_i \in \Delta(\bigcup_{n=0}^{k-1} Z_{-i}^n) | \nu_i(F_{-i} \cap \bigcup_{n=0}^{k-1} Z_{-i}^n) \geq p\}$ is measurable in $\Delta(\bigcup_{n=0}^{k-1} Z_{-i}^n)$, but this is true because of (i) and how the σ-algebra on $\Delta(\bigcup_{n=0}^{k-1} Z_{-i}^n)$ is generated. □

The space C^* is universal in the sense that any hierarchy of beliefs that arises in some type space belongs to C. The following Definition and Theorem formalize this notion.

Definition 3. A collection of measurable mappings $h_i : C_i \to C_i^*$ for $i = 1, \ldots, I$, is *belief preserving* if $\nu_i^*(h_i(c_i))(E) = \nu_i(c_i)(h_{-i}^{-1}(E))$ for any measurable $E \subseteq C_{-i}^*$.

Theorem 5. *For any cognitive type space C there exists a collection of belief preserving mappings $h_i : C_i \to C_i^*$ for $i = 1, \ldots, I$.*

The proof is standard (Mertens and Zamir, 1985; Brandenburger and Dekel, 1993; Heifetz and Samet, 1998), except that the type space is a union of measurable spaces rather than a single measurable space.

Proof. For any cognitive type space C construct the mappings h_i by a recursion on k

$$h_i(c_i) := \begin{cases} 0 & \text{if } k_i(c_i) = 0, \\ (1, \delta_0) & \text{if } k_i(c_i) = 1, \\ \left(k_i(c_i), \nu_i(c_i) \circ h_{-i}^{-1}\right) & \text{if } k_i(c_i) \geq 2. \end{cases}$$

Note that this construction is correct since by definition if $(k, \nu_i^*) \in Z_i^k$, then the domain of ν_i^* is the collection of measurable subsets of $\bigcup_{l=0}^{k-1} Z_{-i}^l$ and the value of h_{-i} on those elements of C_{-i} which have $k_{-i} \leq k-1$ is already defined. Hence, h_{-i}^{-1} is well defined on $\bigcup_{l=0}^{k-1} Z_{-i}^l$. The fact that $h_{-i}^{-1}(G)$ is a measurable set for any measurable set $G \in C_{-i}^*$ follows from the measurability of the mappings h_i to be shown now.

To show that the mappings h_i are measurable note that for any measurable set $G \subseteq C_i^*$ we have $G = \bigcup_{k=0}^{\infty}(G \cap Z_i^k)$, so $h_i^{-1}(G) = \cup_{k=0}^{\infty} h_i^{-1}(G \cap Z_i^k)$; hence, it suffices to show that for any measurable $G_i^k \subseteq Z_i^k$ the set $h_i^{-1}(G_i^k)$ is measurable. To show this, proceed by induction. For $k = 0, 1$ the result is immediate and follows from the measurability of the mapping k_i. Assume now the result holds for $k' = 0, 1, \ldots, k$.

Note that any measurable $G_i^{k+1} \subseteq Z_i^{k+1}$ is of the form $\{k+1\} \times F$ for some measurable $F \subseteq \Delta\left(\bigcup_{l=0}^{k} Z_{-i}^l\right)$ and by definition the collection of measurable subsets of $\Delta\left(\bigcup_{l=0}^{k} Z_{-i}^l\right)$ is generated by the family of sets $\left\{\mu \in \Delta\left(\bigcup_{l=0}^{k} Z_{-i}^l\right) \mid \mu(E) \geq p\right\}$ for all measurable $E \subseteq \bigcup_{l=0}^{k} Z_{-i}^l$ and $p \in [0, 1]$. Thus, it suffices to show that the set $\{c_i \in C_i \mid k_i(c_i) = k+1\} \cap \{c_i \in C_i \mid \nu_i(c_i)\left(h_{-i}^{-1}(E)\right) \geq p\}$ is measurable for each measurable $E \subseteq \bigcup_{l=0}^{k} Z_{-i}^l$ and $p \in [0, 1]$. The measurability of the first component of the intersection follows from the fact that the mapping k_i is measurable for all i. From the inductive hypothesis it follows that the set $H := h_{-i}^{-1}(E)$ is a measurable subset of C_{-i}. Since the mapping ν_i is measurable for all i, it follows that the set

$\{c_i \in C_i \mid \nu_i(c_i) \in A\}$ is measurable for any measurable set $A \subseteq \Delta(C_{-i})$, in particular this is true for the set $A := \{\mu \in \Delta(C_{-i}) \mid \mu(H) \geq p\}$, which is measurable since the set H is measurable and by the definition of the σ-algebra on $\Delta(C_{-i})$.

To show that the mappings h_i are belief preserving note that for any measurable $E \subseteq C^*_{-i}$ by Definition 1 and the construction of the mappings h_i:

$$\begin{aligned}
\nu_i^*(h_i(c_i))(E) &= \nu_i(c_i) \circ h_{-i}^{-1}\left(E \cap \bigcup_{n=0}^{k_i(c_i)-1} Z_{-i}^n\right) \\
&= \nu_i(c_i)\left[h_{-i}^{-1}\left(E \cap \bigcup_{n=0}^{k_i(c_i)-1} Z_{-i}^n\right)\right] \\
&= \nu_i(c_i)\left[h_{-i}^{-1}(E) \cap h_{-i}^{-1}\left(\bigcup_{n=0}^{k_i(c_i)-1} Z_{-i}^n\right)\right]
\end{aligned}$$

$$\begin{aligned}
&= \nu_i(c_i)\left[h_{-i}^{-1}(E) \cap \bigcup_{n=0}^{k_i(c_i)-1} h_{-i}^{-1}(Z_{-i}^n)\right] \\
&= \nu_i(c_i)\left[h_{-i}^{-1}(E) \cap \{c_{-i} \in C_{-i} \mid k_{-i}(c_{-i}) < k_i(c_i)\}\right] \\
&= \nu_i(c_i)\left[h_{-i}^{-1}(E)\right],
\end{aligned}$$

where the last equality follows from Equation (1) in Definition 1. □

B Appendix: Proofs

B.1 Proof of Theorem 2

We will inductively prove the following four assertions. For any cognitive type space and any cognitive equilibrium of the email game where players of level $k = 0$ choose A

(i-k) Player 1 of level $2k$ chooses A upon receiving at least k messages,

(ii-k) Player 1 of level $2k + 1$ chooses A upon receiving at least k messages,

(iii-k) Player 2 of level $2k$ chooses A upon receiving at least $k - 1$ messages,

(iv-k) Player 2 of level $2k + 1$ chooses A upon receiving at least k messages.

Observe that if these assertions hold for any k, then the proof of 2 is complete. Also, observe that for both players the number of messages required to ensure his playing A is monotonically increasing in the his level.

First, note that assertions (i-0), (ii-0), (iii-0), and (iv-0) follow from the construction in Section 4. Second, suppose that for some k assertions (i-l), (ii-l), (iii-l), and (iv-l) hold for $l = 0, 1, \ldots, k$ and prove that they hold for $l = k + 1$.

To prove assertion (i-$k + 1$) suppose that player 1 of level $2(k + 1)$ receives n messages. This means that he thinks player 2 is of type $0, 1, \ldots, 2k, 2k + 1$ and received either $m = n - 1$ or $m = n$ messages. By the inductive hypothesis, player 2 chooses A if $m > k$. Thus, if $n - 1 > k$, or $n > k + 1$, then player 1 is sure that player 2 chooses A and chooses A himself.

The proofs of assertions (ii-$k + 1$), (iii-$k + 1$), and (iv-$k + 1$) are analogous. □

B.2 Proof of Theorem 1

Fix a cognitive type C space and a probability distribution $\lambda \in \Delta(C)$. Let $\lambda_i(k)$ be the induced probability that player i has level k. By Theorem 2, in any cognitive equilibrium of the email game where players of level $k = 0$ choose A, upon receiving n messages player 1 chooses A if he is of level $k < 2n$ and player 2 chooses A if

he is of level at least $2n+1$. This means, that the probability that player 1 chooses A upon receiving n messages is at least $\sum_{k=0}^{2n-1} \lambda_1(k)$ and the probability that player 1 chooses A upon receiving n messages is at least $\sum_{k=0}^{2n} \lambda_2(k)$. Since $\sum_{k=0}^{\infty} \lambda_1(k) = \sum_{k=0}^{\infty} \lambda_2(k) = 1$, it follows that $\lim_{n\to\infty} \sum_{k=0}^{2n-1} \lambda_1(k) = \sum_{k=0}^{2n} \lambda_2(k) = 1$. □

B.3 Proof of Theorem 3

Let $\gamma_i^{c_i} : \mathbb{N} \to \{A, N\}$ be the cognitive equilibrium strategies constructed iteratively in Section 4 starting from $\gamma_i^0(n) = A$ for all n.

Lemma 2. *For every i, for every n, and for every $c_i \in C_i$ if $\gamma_i^{c_i}(n+1) = N$ then $\gamma_i^{c_i}(n) = N$.*

Proof. Prove the contrapositive

$$\forall_i \forall_n \forall_{c_i \in C_i} \left[\gamma_i^{c_i}(n) = A\right] \Rightarrow \left[\gamma_i^{c_i}(n+1) = A\right]$$

by induction on $k_i(c_i)$. First observe that for all i and for all n it trivially follows that $[\gamma_i^{c_i}(n) = A] \Rightarrow [\gamma_i^{c_i}(n+1) = A]$ for all $c_i \in C_i$ with $k_i(c_i) = 0$.

Second, suppose that

$$\forall_i \forall_n \forall_{\substack{c_i \in C_i \\ k_i(c_i) \leq K}} \left[\gamma_i^{c_i}(n) = A\right] \Rightarrow \left[\gamma_i^{c_i}(n+1) = A\right] \tag{2}$$

and show that

$$\forall_i \forall_n \left[\gamma_i^{c_i}(n) = A\right] \Rightarrow \left[\gamma_i^{c_i}(n+1) = A\right]. \tag{3}$$

for all $c_i \in C_i$ with $k_i(c_i) = K+1$. Fix a probability measure $\lambda \in \Delta(C_{-i})$ and define

$$x_i^{\lambda}(n) := (-2) \cdot \lambda\left(\{c_i \in C_i \mid \gamma_i^{c_i}(n) = N\}\right) + 1 \cdot \lambda\left(\{c_i \in C_i \mid \gamma_i^{c_i}(n) = A\}\right).$$

Observe that by (2)

$$\forall_n \, x_i^{\lambda}(n) \leq x_i^{\lambda}(n+1). \tag{4}$$

First consider agent 1. Let λ be the belief of the cognitive type c_1 of agent 1 who got n messages and has $k_1(c_1) = K+1$. If $\gamma_1^{c_1}(n) = A$, then the payoff from choosing A

is greater than choosing N.

$$\frac{1}{2-\varepsilon}x_2^{\lambda}(n-1)+\frac{1-\varepsilon}{2-\varepsilon}x_2^{\lambda}(n)\geq 0.$$

By the independence assumption, λ is also the belief of the cognitive type c_1 of agent 1 who got $n+1$ messages; thus by (4)

$$\frac{1}{2-\varepsilon}x_2^{\lambda}(n)+\frac{1-\varepsilon}{2-\varepsilon}x_2^{\lambda}(n+1)\geq 0.$$

Hence, $\gamma_1^{c_1}(n+1)=A$. Second, consider agent 2. Let λ be the belief of the cognitive type c_2 of agent 2 who got n messages and has $k_2(c_2)=K+1$. If $\gamma_2^{c_2}(n)=A$, then the payoff from choosing A is greater than choosing N.

$$\frac{1}{2-\varepsilon}x_1^{\lambda}(n)+\frac{1-\varepsilon}{2-\varepsilon}x_1^{\lambda}(n+1)\geq 0.$$

By the independence assumption, λ is also the belief of the cognitive type c_2 of agent 2 who got $n+1$ messages; thus by (4)

$$\frac{1}{2-\varepsilon}x_1^{\lambda}(n+1)+\frac{1-\varepsilon}{2-\varepsilon}x_1^{\lambda}(n+2)\geq 0.$$

Hence, $\gamma_2^{c_2}(n+1)=A$. This proves (3) and concludes the proof. □

Proof of Theorem 3 Observe that by construction of the equilibrium for every k only finitely many actions can be N; thus,

$$\forall_i\forall_{c_i}\exists_n\forall_{m\geq n}\ \gamma_i^{c_i}(m)=A. \tag{5}$$

The goal is to show that

$$\exists_n\forall_i\forall_{c_i}\forall_{m\geq n}\ \gamma_i^{c_i}(m)=A. \tag{6}$$

Suppose (6) is false. Then

$$\forall_n\exists_i\exists_{c_i}\exists_{m\geq n}\ \gamma_i^{c_i}(m)=N.$$

It follows from Lemma 2 that

$$\forall_n \exists_i \exists_{c_i} \gamma_i^{c_i}(n) = N. \tag{7}$$

Claim:

$$\forall_i \forall_n \exists_{c_i} \gamma_i^{c_i}(n) = N. \tag{8}$$

Suppose not. There are two possible cases:

$$\exists_n \Big[\exists_{c_2} [\gamma_2^{c_2}(n) = N] \text{ and } \forall_{c_1} [\gamma_1^{c_1}(n) = A] \Big]. \tag{i}$$

and

$$\exists_n \Big[\exists_{c_1} [\gamma_1^{c_1}(n) = N] \text{ and } \forall_{c_2} [\gamma_2^{c_2}(n) = A] \Big]. \tag{ii}$$

But (i) cannot be true because by Lemma 2 $\gamma_1^{c_1}(n+1) = A$ for all c_1, which means that the cognitive type c_2 of player 2 is sure that player 1 chooses A, and should therefore choose A himself, i.e., $\gamma_2^{c_2}(n) = A$. Contradiction. Similarly, (ii) cannot be true. By Lemma 2 $\gamma_2^{c_2}(n+1) = A$ for all c_2. From (7) it follows that there exists i and c_i with $\gamma_i^{c_i}(n+1) = N$. The two last sentences imply that there exists c_1 with $\gamma_1^{c_1}(n+1) = N$. However, the cognitive type c_1 of player 1 is sure that player 2 chooses A, and should therefore choose A himself, i.e., $\gamma_1^{c_1}(n) = A$. Contradiction. This proves the claim, so (8) holds.

For any n define $k_i(n)$ to be the smallest $k_i(c_i)$ such that $\gamma_i^{c_i}(n) = N$.

Observe that

$$\forall_n \; k_2(n-1) < k_1(n) < k_2(n). \tag{9}$$

To verify the first inequality, suppose take any cognitive type c_1 of player 1 that has $k_1(c_1) = k_2(n-1)$. Then by Lemma 2 $k_1(c_1) \leq k_2(n)$ and player 1 of cognitive type c_1 who got n messages is sure that player 2 chooses A; hence $\gamma_1^{c_1}(n) = A$ and $k_1(n) > k_1(c_1) = k_2(n-1)$. To verify the second inequality, apply an analogous argument for player 2 of cognitive type c_2 with $k_2(c_2) = k_2(n)$.

To conclude the proof, consider player 2 of cognitive type c_2 with $k_2(c_2) = k_2(n)$ who received n messages. Under the independence assumption, player 1 plays N

with probability

$$\frac{1}{2-\varepsilon}\nu_2(c_2)\big(\{c_1 \in C_1 \mid k_1(c_1) \geq k_1(n)\}\big)+\frac{1-\varepsilon}{2-\varepsilon}\nu_2(c_2)\left(\{c_1 \in C_1 \mid k_1(c_1) \geq k_1(n+1)\}\right).$$

It follows from (9) that this probability is smaller than

$$\frac{1}{2-\varepsilon}\nu_2(c_2)\big(\{c_1 \in C_1 \mid k_1(c_1) \geq k_2(n-1)\}\big)+\frac{1-\varepsilon}{2-\varepsilon}\nu_2(c_2)\left(\{c_1 \in C_1 \mid k_1(c_1) \geq k_2(n)\}\right).$$

By assumption the second part is equal to zero, while by Property 3 with $\kappa = \frac{2-\varepsilon}{3}$

$$\inf_n \sup_{c_2 \in \{c_2 \in C_2 | k_2(c_2)=k_2(n)\}} \frac{1}{2-\varepsilon}\nu_2(c_2)\left(\{c_1 \in C_1 \mid k_1(c_1) \geq k_2(n-1)\}\right) < \frac{1}{3}$$

Thus, there exists n such that the payoff from playing A for any cognitive type c_2 of player 2 with $k_2(c_2) = k_2(n)$ who received n messages is larger than zero; hence, he should play A, rather than N implied by the definition of $k_2(n)$. Contradiction. □

B.4 Proof of Theorem 4

It follows from Theorem 3 that on C there exists an cognitive equilibrium with a number of messages n such that all cognitive types choose to attack having received n or more messages. The behavior of those types will be the same on $C^{\delta,\lambda}$, so only the behavior of type ∞ should be considered. Suppose that type ∞ of player i received $n+1$ or more messages. Playing N guarantees him a payoff of 0. The payoff of playing A will depend on what type ∞ of $-i$ is doing. The lowers possible payoff from choosing A is when that type is playing N and is equal to $\delta \cdot (-2) + (1-\delta) \cdot 1 \geq 0$; thus, no matter what type ∞ of player $-i$, type ∞ of player i has a strict incentive to play A as long as he receives $n+1$ or more messages. □

References

Bergemann, D. and S. Morris (2005): "Robust Mechanism Design," *Econometrica*, 1771–1813.

Binmore, K. and L. Samuelson (2001): "Coordinated action in the electronic mail game," *Games and Economic Behavior*, 35, 6–30.

Brandenburger, A. and E. Dekel (1993): "Hierarchies of beliefs and common knowledge," *Journal of Economic Theory*, 59, 10.

Cabrales, A., R. Nagel, and R. Armenter (2007): "Equilibrium selection through incomplete information in coordination games: an experimental study," *Experimental Economics*, 10, 221–234.

Camerer, C. F. (2003): *Behavioral Game Theory: Experiments in Strategic Interaction*.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A Cognitive Hierarchy Model of Games," *The Quarterly Journal of Economics*, 119, 861–898.

Carlsson, H. and E. van Damme (1993): "Global Games and Equilibrium Selection," *Econometrica*, 61, 989–1018.

Chen, Y.-C., A. Di Tillio, E. Faingold, and S. Xiong (2009a): "Common Belief, Rationalizability and Proximity of Types,," *mimeo*.

——— (2009b): "Uniform Topologies on Types," *mimeo*.

Costa-Gomes, M., V. P. Crawford, and B. Broseta (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69, 1193–1235.

Costa-Gomes, M. A. and V. P. Crawford (2006): "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*, 96, 1737–1768.

Crawford, V., U. Gneezy, and Y. Rottenstreich (2008): "The Power of Focal Points is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures," *American Economic Review*, 98, 1443–1458.

Crawford, V. and N. Iriberri (2007a): "Fatal Attraction: Salience, Naivete, and Sophistication in Experimental Hide-and-Seek Games," *American Economic Review*, 97, 1731–1750.

Crawford, V. P. and N. Iriberri (2007b): "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica*, 75, 1721–1770.

Crawford, V. P., T. Kugler, Z. Neeman, and A. Pauzner (2009): "Behaviorally Optimal Auction Design: An Example and Some Observations," *Journal of the European Economic Association*, 7.

Dekel, E., D. Fudenberg, and S. Morris (2006): "Topologies on Types," *Theoretical Economics*, 1.

Frankel, D. M., S. Morris, and A. Pauzner (2003): "Equilibrium selection in global games with strategic complementarities," *Journal of Economic Theory*, 108, 1–44.

Fremlin, D. H. (2001): *Measure Theory, Vol. 2: Broad Foundations*.

Georganas, S., P. J. Healy, and R. A. Weber (2010): "On the persistence of strategic sophistication," *Ohio State University Working Paper*.

Halpern, J. (1986): "Reasoning about knowledge: An overview," *Theoretical aspects of reasoning about knowledge*.

Harsanyi, J. C. (1967): "Games with Incomplete Information Played by "Bayesian" Players, I-III. Part I. The Basic Model," *Management Science*, 14, 159–182.

Healy, P. J., S. Georganas, and R. Weber (2010): "On the Persistence of Strategic Sophistication," *mimeo*.

Heifets, A. and W. Kets (2013): "Robust Multiplicity with a Grain of Naivete," Tech. rep.

Heifetz, A. and D. Samet (1998): "Topology-Free Typology of Beliefs," *Journal of Economic Theory*, 82, 324–341.

Heinemann, F., R. Nagel, and P. Ockenfels (2004): "The theory of global games on test: experimental analysis of coordination games with public and private information," *Econometrica*, 72, 1583–1599.

Ho, T.-H., C. Camerer, and K. Weigelt (1998): "Iterated Dominance and Iterated Best Response in Experimental "p-Beauty Contests"," *American Economic Review*, 88, 947–969.

Ivanov, A., D. Levin, and M. Niederle (2010): "Can relaxation of beliefs rationalize the winner's curse?: an experimental study," *Econometrica*, 78, 1435–1452.

Kets, W. (2014): "Finite Depth of Reasoning and Equilibrium Play in Games with Incomplete Information," Tech. rep.

Kneeland, T. (2012): "Coordination under limited depth of reasoning," *University of British Columbia Working Paper*.

Mertens, J. F. and S. Zamir (1985): "Formulation of Bayesian analysis for games with incomplete information," *International Journal of Game Theory*, 14, 1–29.

Monderer, D. and D. Samet (1989): "Approximating Common Knowledge with Common Beliefs," *Games and Economic Behavior*, 1, 170–190.

Morris, S. and H. S. Shin (1998): "A Theory of the Onset of Currency Attacks," CEPR Discussion Papers 2025, C.E.P.R. Discussion Papers.

Myerson, R. B. (1991): *Game theory*, Harvard university press.

Nagel, R. (1995): "Unraveling in Guessing Games: An Experimental Study," *The American Economic Review*, 85, 1313–1326.

Rogers, B. W., T. R. Palfrey, and C. F. Camerer (2009): "Heterogeneous quantal response equilibrium and cognitive hierarchies," *Journal of Economic Theory*, 144, 1440–1467.

Rubinstein, A. (1989): "The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge"," *The American Economic Review*, 79, 385–391.

Stahl, D. (1993): "Evolution of smart$_n$ players," *Games and Economic Behavior*, 5, 604–617.

Stahl, D. O. and P. W. Wilson (1994): "Experimental evidence on players' models of other players," *Journal of Economic Behavior and Organization*, 25, 309–327.

——— (1995): "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218–254.

Weinstein, J. and M. Yildiz (2007): "A structure theorem for rationalizability with application to robust predictions of refinements," *Econometrica*, 75, 365–400.

Wilson, R. (1987): "Game theoretic analyses of trading processes," in *Advances in Economic Theory: Fifth World Congress, Cambridge University Press, Cambridge, MA*, 33–70.