Commitment Races

LOREN K. FRYXELL, City St George's, University of London, United Kingdom
MAXIMILIAN REITH, University of Bonn, Germany

Two agents play a game—but they may also lock in some action beforehand. A commitment race is when each agent attempts to be the first to lock in the action associated with their preferred equilibrium. We show that in a wide class of symmetric games, agents do indeed race to commit, and this can lead both players to be worse off than if neither had this ability. This analysis contributes to a literature concerned about the prospective ability of AI agents to lock in their actions prior to engagement with other human or AI agents.

CCS Concepts: • **Theory of computation → Algorithmic game theory and mechanism design**; *Algorithmic game theory*; *Quality of equilibria.*

Additional Key Words and Phrases: game theory, commitment, AI agents, Game of Chicken, AI Safety, Subgame Perfect Equilibrium, Mixed Strategies

ACM Reference Format:
Loren K. Fryxell and Maximilian Reith. 2026. Commitment Races. 1, 1 (February 2026), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

As Artificial Intelligence (AI) technology continues to evolve, AI agents are becoming an increasingly prominent object of study. In the future, we may see capable AI agents operating under minimal oversight and engaging in strategic interactions. It has been suggested that such AI agents could possess the ability to make credible commitments [2]. There are concerns that if AI agents have such an ability, they may use this power to extort others. For example, one AI agent might commit to an aggressive action that then forces another AI agent to accommodate. But if both AI agents possess this ability, each may attempt to commit to the aggressive action first. This dynamic gives rise to what has been called a *commitment race* [6], in which each agent races to commit before the other. If both commit simultaneously to incompatible strategies, the outcome could be harmful to both.

In this paper, we develop a formal model to study this dynamic. Two players play a game, referred to hereafter as the base game. Before playing the base game, players have the opportunity to simultaneously commit to any normal-form strategy of this game. Commitment is voluntary; players can choose not to commit and wait until the base game. We refer to the overall game as the commitment game. Our main result is that the ability to credibly commit can lead both players to be worse off than if neither had this ability. That is, in a class of symmetric games known as Chicken, the unique symmetric equilibrium in the commitment game is Pareto inferior to the

We thank Eric Olav Chen, Peter Eso, James Faville, Alexis Ghersengorin, Gaurav Khatri, Yidian Meng, Paul Metzler, Margaret Meyer, Richard Ngo, Alexei Parakhonyak, Manos Perdikakis, Sami Petersen, Francis Priestland and Ludvig Sinander for helpful comments.

Authors' Contact Information: Loren K. Fryxell, City St George's, University of London, London, United Kingdom, loren.fryxell@citystgeorges.ac.uk; Maximilian Reith, University of Bonn, Bonn, Germany, max.reith@uni-bonn.de.

ACM XXXX-XXXX/2026/2-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

unique symmetric equilibrium in the base game. This arises because the non-aggressive strategy of Waiting in the commitment game becomes less attractive in equilibrium than the non-aggressive strategy of Moving in the base game, increasing the likelihood of simultaneous aggressive behavior. We illustrate this with an example.

Example

Consider the well-known Game of Chicken depicted in Table 1, where two players simultaneously decide whether to go Straight or Move (out of the way). Besides two pure-strategy Nash equilibria where one player plays Straight and the other plays Move, this game has one mixed-strategy Nash equilibrium (MSNE), in which both players choose Straight with probability 0.1 and Move with probability 0.9, resulting in an equilibrium payoff of −0.1. This is the unique symmetric equilibrium of the game, and we will refer to it as the base game equilibrium henceforth.

Table 1. The Game of Chicken

	Straight	Move
Straight	$-10, -10$	$1, -1$
Move	$-1, 1$	$0, 0$

How would equilibrium outcomes change if players could simultaneously commit to their strategies beforehand? Figure 1 shows the commitment game. Initially, both players simultaneously decide whether to commit to Straight (S), commit to Move (M), or wait (W). After this commitment period, any commitments made are observed and the base game is played. If both players commit, their chosen strategies are executed, concluding the game. If only one player commits, the other observes this commitment and best responds accordingly. Should both players choose to wait, they proceed to play the base Game of Chicken.

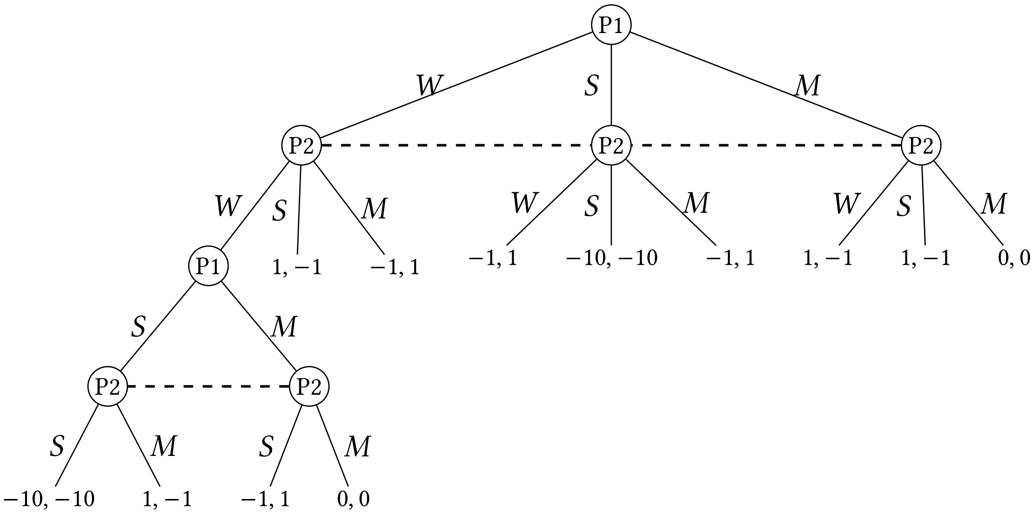

Fig. 1. The Game Tree for the Commitment Game of Chicken

In the commitment game, there are multiple subgame-perfect Nash equilibria (SPNE) in pure strategies: in one, a player commits straight while the other waits and then moves out of the way; in another, one player commits straight and the other commits to moving out of the way. Importantly, no equilibrium can involve both players waiting before choosing a pure-strategy equilibrium, since in that case one player would end up in her less-preferred equilibrium and could profitably deviate by committing straight. In addition to these pure strategy SPNE, there exists a unique mixed strategy SPNE.[1] Here, each player commits to the aggressive strategy (Straight) with probability 0.11 and waits with probability 0.89. If both players decide to wait, they play the base game equilibrium. This is the unique symmetric SPNE of the commitment game. Notably, this equilibrium has an expected payoff of -0.2 for each player, while the base game equilibrium has an expected payoff of -0.1 for each player.

Why is that? As it turns out, the ability to commit one period before transforms any Game of Chicken into *another* Game of Chicken with strictly lower payoffs for both players, as shown in Table 2. In the commitment game, committing to M (moving out of the way) is weakly dominated by W (waiting) and there is no symmetric equilibrium where committing to M is played with positive probability. Hence, we can remove it. (S, W) gives payoffs $(1, -1)$ since player 2 will best respond with M in the next period, (W, S) gives payoffs $(-1, 1)$ since player 1 will best respond with M in the next period, and (W, W) gives payoffs $(-0.1, -0.1)$, the payoff of the base game equilibrium. The commitment game is hence another Game of Chicken, except with $(-0.1, -0.1)$ rather than $(0, 0)$ in the bottom right cell. Thus, playing S becomes relatively more attractive in the commitment game than in the base game, leading to more aggressive behavior, a higher probability of crashing (both playing Straight), and lower payoffs for both players in equilibrium.

Table 2. Commitment Transforms the Game of Chicken

(a) Base Game

	S	M
S	$-10, -10$	$1, -1$
M	$-1, 1$	$0, 0$

$\longrightarrow$

(b) With Commitment

	S	W
S	$-10, -10$	$1, -1$
W	$-1, 1$	$-0.1, -0.1$

So far, we have considered a single commitment period before the base game, during which players can simultaneously commit. However, if players can commit to a strategy for a game they will play in the future, it seems natural that they might be able to do so at multiple points in time prior to play. This is particularly relevant if one thinks of the players as AI agents. Fast computational speed could imply that agents have many, effectively infinite, opportunities to commit beforehand. Let T represent the number of discrete commitment periods. Having examined the single-period case ($T = 1$), we now consider multiple periods ($T > 1$). In each commitment period, players simultaneously decide whether to commit to a strategy or wait. If both players commit, their chosen strategies are executed immediately. If only one player commits, the other observes this commitment and best responds. If both players wait, the game advances to the next commitment period, or to the base game if no commitment periods remain.

What happens as we increase the number of commitment periods T? In the unique symmetric SPNE of the commitment game, the expected payoff for both players strictly decreases with T and, as T approaches infinity, converges to the minmax payoff of the base game—the worst possible equilibrium outcome for each player. Table 3 shows the expected payoffs and the probability of crashing as T increases.

[1] For clarity, only equilibria involving both-sided probabilistic commitment are referred to as mixed equilibria, even though pure strategy equilibria are technically also mixed strategy equilibria.

Table 3. Expected MSNE payoff and probability of crashing with multiple commitment periods T. The base game is the Game of Chicken in Table 1. $T = 0$ corresponds to the mixed equilibrium in the base game without commitment.

	$T = 0$	$T = 1$	$T = 2$	$\cdots$	$T \to \infty$
Expected payoff	-0.1	-0.2	-0.29	$\cdots$	-1
Probability players crash	0.01	0.02	0.03	$\cdots$	0.1

The mechanism behind this result is identical to the single-period case. Each additional commitment period transforms the game into a new Game of Chicken with lower payoffs. Waiting becomes relatively less attractive, making early commitments to Straight more likely, which increases the probability of crashing and lowers the expected payoffs for both players.

We refer to this dynamic as a commitment race. A commitment race is when two agents each attempt to be the first to lock in the action associated with their preferred equilibrium, so that the best response of the other is to concede. We study commitment races in all symmetric 2x2 games. In Section 2, we show that in most but not all symmetric 2x2 games, the unique symmetric equilibrium is, in fact, a commitment race. For some games this race to commit makes both players better off, and for other games this makes both players worse off, as shown in this example. In Section 3, we analyze the effect of increasing the number of commitment periods T and present a general convergence result. Section 4 concludes.

Literature

In economic theory, commitment is often associated with welfare-enhancing outcomes, at least for the player making the commitment. Commitment features prominently across a wide range of models, such as principals committing to mechanisms, players delegating decision-making to agents, and first movers in Stackelberg games earning a higher profit [see the discussion in 5]. This paper, however, turns to the question of whether the ability to make commitments can *harm* players.

Existing literature has largely focused on one-sided commitment, leaving two-sided commitment relatively understudied. Harrenstein et al. [3] analyze two- and three-player games in which all players have the ability to commit. However, the authors assume a predetermined order of commitment and do not allow for simultaneous commitment.

Kalai et al. [5] present a model of simultaneous conditional commitment, expanding on earlier work by Tennenholtz [7]. In their framework, players select commitment devices d before the base game. A commitment device specifies a base game strategy as a function of the other's chosen commitment device, $d_i(d_{-we}) = S_i$. The authors' main result is a commitment folk theorem, stating that the Nash equilibria in such commitment games span all individually-rational correlated strategy payoffs of the game. Naturally, this result relies on the availability of a complete space of conditional commitment devices. In contrast, our model limits players to committing exclusively to normal form strategies. This still allows for some conditionality in sequential base games. For example in an entrant-incumbent game, the incumbent's normal form strategies allow commitment that conditions on the entrant's initial decision whether to enter the market or not.

The remainder of this paper is organized as follows. Section 2 shows that, in addition to cases where payoffs decline, there are also base games in which commitment raises payoffs for both players. Here, Proposition 1 characterizes symmetric 2×2 games in which commitment is either beneficial or detrimental. Section 3 considers games with multiple commitment periods T. Here we present Proposition 2, which establishes a convergence result as $T \longrightarrow \infty$.

2 Commitment in 2×2 Games

In the introduction, we presented a game in which commitment leads to strictly lower equilibrium payoffs in mixed strategies. In this section we will investigate for which broader class of games this result holds. We analyze the effect of introducing commitment on equilibrium payoffs in a class of symmetric 2×2 games. Specifically, we will consider 2×2 games that are symmetric in the sense of table 4 (left) with $A < D$. To reduce the number of parameters without loss of generality, we apply a positive affine transformation: subtracting A from all payoffs and dividing by $D - A$. This allows us to redefine the remaining parameters as $b = \frac{B-A}{D-A}$ and $c = \frac{C-A}{D-A}$, resulting in the normalized form shown in Table 4 (right). This transformation follows standard practice [see 8, and references therein].

Table 4. Representation of a general 2×2 game and its simplified form.

(a) General Form

	A_1	A_2
A_1	(A, A)	(B, C)
A_2	(C, B)	(D, D)

$\longrightarrow$

(b) Simplified Form

	A_1	A_2
A_1	$(0, 0)$	(b, c)
A_2	(c, b)	$(1, 1)$

Different values of b and c represent different types of well known 2×2 games. Figure 2 illustrates this.

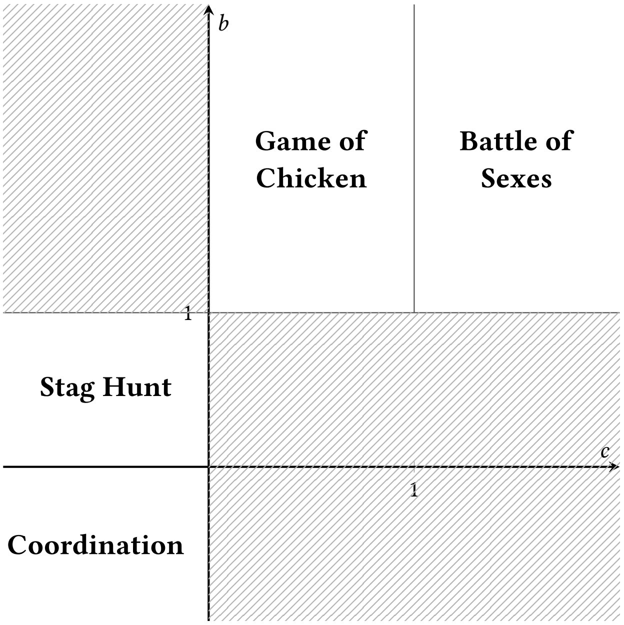

Fig. 2. Game-type regions in (c, b)-space.

When $b > 1$ and $c \in [0, 1]$, the game takes the form of a Game of Chicken. Games with $b > 1$ and $c > 1$ are known as the Battle of Sexes. The Stag Hunt game arises when $b \in [0, 1]$ and $c < 0$. Finally, games with $b < 0$ and $c < 0$ do not really have a standard game name; we refer to this residual class as Coordination games. Shaded areas denote games with strictly dominated strategies.

In pure strategies, the Game of Chicken and the Battle of Sexes have the Nash equilibria $\{A_1, A_2\}$ and $\{A_2, A_1\}$, whereas Stag Hunt and Coordination have the equilibria $\{A_1, A_1\}$ and $\{A_2, A_2\}$. All of these games have one equilibrium that involves mixing: Players play A_1 with probability $\frac{b-1}{c+b-1}$, which yields the expected payoff $\frac{bc}{b+c-1}$. [2] Note that this MSNE is the unique symmetric equilibrium for the Game of Chicken and the Battle of Sexes.

Base game equilibria in pure strategies translate into subgame perfect Nash equilibria in pure strategies in the commitment game. This can occur in three distinct ways: both players committing upfront to base game equilibrium strategies, one player committing first and the other waiting and then best responding, or both players waiting and then subsequently playing equilibrium strategies. In the appendix in section A, we establish two lemmas characterizing the conditions under which each type of pure strategy equilibrium arises.

However, we argue that focusing on subgame perfect equilibria in *mixed strategies* rather than pure strategies is considerably more compelling for two reasons: symmetry and equilibrium selection.

Symmetry. All 2×2 games considered in this paper are symmetric, yet the Game of Chicken and the Battle of Sexes do not have a symmetric equilibrium in pure strategies, neither in the base game nor in the commitment version. Asymmetric equilibria seem counterintuitive. Players are identical in both strategy sets and payoffs, thus there is no clear rationale for why one should expect opposite equilibrium behavior. As we show, the mixed equilibrium is the unique symmetric equilibrium in these games, and thus seems better fitted as a solution concept to predict the behavior of players.

Equilibrium Selection. In each commitment game, multiple pure strategy Nash equilibria typically exist. Given the absence of communication devices and the one-shot nature of commitment games, it seems implausible that players can seamlessly coordinate on a single equilibrium. This is particularly relevant in the Game of Chicken and the Battle of Sexes, where each player prefers a different equilibrium. A mixed equilibrium naturally incorporates the possibility of miscoordination, quantifies the probability with which miscoordination might occur, and provides expected payoffs rather than deterministic outcomes. Additionally, Harsanyi's purification theorem [4] further supports the practical relevance of mixed equilibria.

Proposition 1. *Consider a symmetric 2×2 game and its commitment variant. In the unique mixed subgame perfect equilibrium of the commitment game, the expected payoff for both players relative to the mixed equilibrium of the base game is:*

(1) *weakly lower in the Game of Chicken (strictly for $c < 1$);*
(2) *higher in the Battle of Sexes.*

Notably, Proposition 1 establishes that commitment can induce equilibria in which both players are strictly worse off, but also that the impact of commitment on equilibrium outcomes depends on the underlying game. In the Game of Chicken, commitment makes players worse off; in contrast, it improves outcomes in the Battle of the Sexes. To our knowledge, this finding is not documented in the existing literature. The full proof is presented in the appendix in section B, where the equilibria for all game-type regions are derived.

[2] The general game does not have a mixed equilibrium when $b + c = 1$, in which case there are (weakly) dominated strategies.

To see why, consider the Game of Chicken first. We argue that commitment transforms the Game of Chicken into another Game of Chicken with worse payoffs. Consider Table 4. On the left side of Table 4, the original Game of Chicken without commitment is displayed. On the right, we present the reduced normal form of the Game of Chicken with commitment. Committing to A_2 (moving out of the way) is weakly dominated by waiting, so that A_2 is effectively replaced by W, and thus the (A_2, A_2) payoff is substituted by the "Wait–Wait payoff." This is the expected mixed equilibrium payoff of the base game, $\frac{bc}{c+b-1}$, instead of the payoff 1 that players received before when both played A_2.

Table 5. Commitment Transforms the Game of Chicken

(a) Base Game

	A_1	A_2
A_1	$(0, 0)$	(b, c)
A_2	(c, b)	$(1, 1)$

$\longrightarrow$

(b) Reduced Normal Form with Commitment

	A_1	W
A_1	$(0, 0)$	(b, c)
W	(c, b)	$(\frac{bc}{c+b-1}, \frac{bc}{c+b-1})$

In the Battle of the Sexes, this dynamic is reversed. Since $b > 1$ and $c > 1$, the Wait-Wait payoff $\frac{bc}{c+b-1}$ exceeds 1, so commitment improves payoffs relative to the base game.

Table 6. Commitment Transforms the Battle of Sexes

(a) Base Game

	A_1	A_2
A_1	$(0, 0)$	(b, c)
A_2	(c, b)	$(1, 1)$

$\longrightarrow$

(b) Normal Form with Commitment

	A_1	A_2	W
A_1	$(0, 0)$	(b, c)	(b, c)
A_2	(c, b)	$(1, 1)$	(c, b)
W	(c, b)	(b, c)	$(\frac{bc}{c+b-1}, \frac{bc}{c+b-1})$

Proposition 1 analyzed the Game of Chicken and the Battle of the Sexes. Corollary 1, derived in Appendix B, extends the analysis to the Stag Hunt and Coordination Game, completing the analysis for all symmetric 2×2 games.

Corollary 1. *The following results hold for the Coordination Game and the Stag Hunt.*

(1) *There is a mixed subgame perfect equilibrium in the Coordination Game. However, players committing to (A_1, A_1) is another symmetric subgame perfect equilibrium that Pareto-dominates the mixed equilibrium.*
(2) *The Stag Hunt game with commitment does not have any new equilibrium outcomes.*

Going forward, we focus on the Game of Chicken and the Battle of the Sexes.

3 Multiple Commitment Periods

In this section, we extend the model to allow for multiple commitment periods, denoted by T. From this point forward, let T represent the number of discrete commitment periods. Suppose that in the first commitment period, player 1 commits and player 2 waits. Then, player 2 gets to observe 1's commitment in the second period, and can best respond to it. But if instead both players did not commit in the first period, they observe each other's inaction and then in the second period simultaneously decide again whether to commit. The base game is played only if neither player commits in any of the T commitment periods.

Let A_{1t} denote the strategy whereby a player waits until period t to commit to A_1, unless the other player has already committed in an earlier period. In that case, the player best responds. Waiting (W) is the strategy of not committing in any commitment period, best-responding to any commitment made by the other player, and if the base game is reached playing the MSNE strategy of the base game. Let p_t denote the probability of a player playing A_{1t} and let q_t denote the probability of a player playing A_{2t}. The probability of waiting is denoted as $\sigma(W) = 1 - \sum_{\tau=1}^{T}(p_\tau + q_\tau)$.

Proposition 2. *As $T \longrightarrow \infty$, the expected payoff in the unique mixed subgame perfect equilibrium for both players*

(1) *decreases in the Game of Chicken and converges to c*
(2) *increases in the Battle of Sexes and converges to* $\min\{b, c\}$

The full proof of proposition 2 can be found in the appendix in section C. We prove three lemmas to establish that this is indeed the unique mixed subgame perfect equilibrium for all T, and to characterize the equilibrium as $T \longrightarrow \infty$, including closed-form expressions for all p_t, q_t, and $\sigma(W)$ in terms of b, c, and T. Table 7 describes the probabilities of committing in the first period and waiting as $T \longrightarrow \infty$, as well as the equilibrium payoffs.

Table 7. Asymptotic Convergence of Probabilities as $T \longrightarrow \infty$ by Game Type.

Game	Condition	Asymptotic Behavior	Payoff
Chicken	$0 < c \leq 1 < b$	$p_1 \longrightarrow \frac{b-c}{b}, \quad \sigma(W) \longrightarrow 0$	$\longrightarrow c$
Battle of Sexes	$0 < 1 < c < b$	$p_1 \longrightarrow \frac{b-c}{b}, \quad q_1 \longrightarrow 0, \quad \sigma(W) \longrightarrow 0$	$\longrightarrow c$
	$0 < 1 < b < c$	$p_1 \longrightarrow 0, \quad q_1 \longrightarrow \frac{c-b}{c-1}, \quad \sigma(W) \longrightarrow 0$	$\longrightarrow b$

To illustrate Proposition 2, recall the example commitment Game of Chicken from the introduction. With a single commitment period, $T = 1$, the mixed-strategy equilibrium involves both players committing to Straight with probability 0.11, waiting with probability 0.89, and receiving an expected payoff of -0.2. When the number of commitment periods increases to $T = 2$, the mixed equilibrium changes: each player commits to Straight in the first period with probability 0.12, in the second period with probability 0.1 (conditional on no one having committed in the first period), and waits in both periods with probability 0.79. The resulting expected payoff is -0.29, which is strictly lower than under $T = 1$. As T increases further, the probability of committing in each period rises, and there are additional periods in which a player might commit. Figure 3 illustrates how the commitment probabilities in each period increase as T becomes larger. Each line corresponds to a different number of commitment periods with which the Game of Chicken from Table 1 is played. The x-axis represents the commitment periods. Consider $t = 1$, the first commitment period. With $T = 1$, it is also the only commitment period, and the probability of committing is 0.11. With $T = 2$, it is the first of two commitment periods, and the probability of committing in this period increases to 0.12. As T increases further, the probability of committing in the first period continues to rise, approaching 0.18. Thus, the probability of crashing increases, and the mixed equilibrium payoff further decreases. The expected payoff converges to -1 for both players, which is the payoff of 'losing' the Game of Chicken.

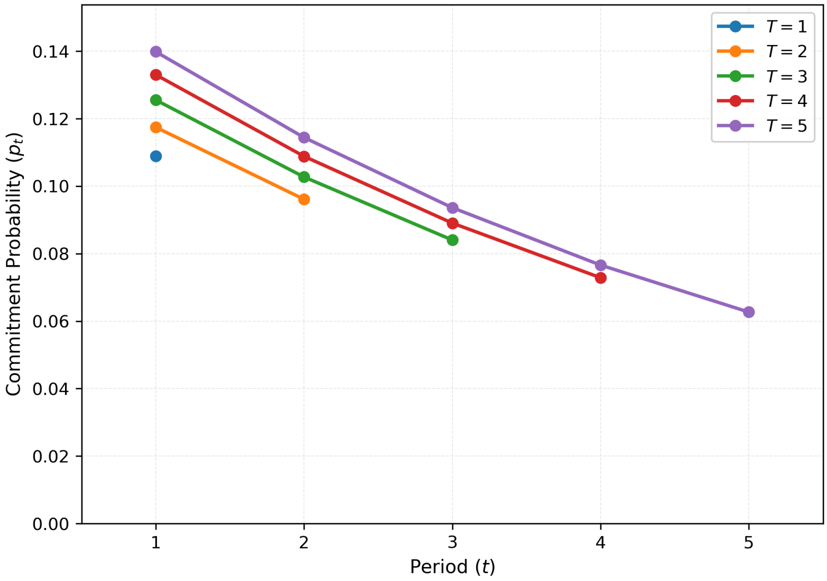

Fig. 3. Commitment probabilities across different periods T in the Game of Chicken.

This convergence occurs because each additional commitment period hurts both players. As previously argued, introducing one period of commitment transforms the base Game of Chicken into a game with lower payoffs, as illustrated in Table 5. Adding another commitment period to this game transforms it into a game with even lower payoffs. In the first commitment period, players choose whether to commit Straight or wait. If both wait, the game proceeds. Under $T = 1$, this would lead to the MSNE of the base game. Under $T = 2$, however, it leads to the $T = 1$ commitment game, which has a lower expected payoff. As a result, the equilibrium payoff is lower when $T = 2$. Each additional commitment period worsens outcomes in the Game of Chicken. As $T \longrightarrow \infty$, both players effectively lose the Game of Chicken—no one wins.[3]

This is a *commitment race*: players have strong incentives to commit early, leading to frequent miscoordination and low expected payoffs. Of course, both players committing to Straight in the first period is not an equilibrium, but in the mixed-strategy equilibrium this outcome occurs with probability $0.18^2 \approx 3\%$ in the Game of Chicken as $T \longrightarrow \infty$. The term commitment race was originally introduced by Kokotajlo [6] in the context of AI agents potentially extorting one another. While his treatment is highly informal, our contribution here is to formalize this idea within a game-theoretic framework.

4 Conclusion

This paper develops a model of simultaneous commitment, focusing on the possibility that commitment can lead to worse outcomes. In the model, players can voluntarily commit to normal-form strategies before playing a base game. We show that in some symmetric 2×2 games, the option to commit strictly reduces equilibrium payoffs. We characterize symmetric 2×2 games for which

[3] However, for similar reasons commitment improves the Battle of Sexes.

commitment reduces or improves equilibrium payoffs and extend the model to multiple commitment periods. As $T \longrightarrow \infty$, early incompatible commitments can become more likely, and payoffs can decline further.

Several limitations remain. First, the finding that commitment can worsen outcomes relies on mixed-strategy Nash equilibria as the equilibrium concept. We focus on mixed-strategy equilibria because they capture the possibility of miscoordination and allow for quantifying its likelihood. Moreover, in symmetric games, mixed equilibria may be the only symmetric equilibria available. However, this justification does not extend to asymmetric games.

Secondly, this paper considers only a narrow class of 2×2 games. Analyzing a broader class of games may reveal additional dynamics.

Finally, while motivated by concerns over AI agents capable of making commitments, this paper does not explore technical implementation. An interesting direction for further research is to explore in what ways commitment in AI agents can be designed to avoid the inefficient outcomes identified in this paper. Conditional commitments [5] or simulation-based program equilibria [1] seem like promising directions here.

References

[1] Emery Cooper, Caspar Oesterheld, and Vincent Conitzer. 2024. Characterising simulation-based program equilibria. (Dec. 19, 2024). arXiv: 2412.14570[cs]. doi:10.48550/arXiv.2412.14570.

[2] Lewis Hammond et al. 2025. Multi-agent risks from advanced AI. (Feb. 19, 2025). arXiv: 2502.14143[cs]. doi:10.48550/arXiv.2502.14143.

[3] Paul Harrenstein, Felix Brandt, and Felix Fischer. 2007. Commitment and extortion. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems.* AAMAS07: International Conference on Autonomous Agents and Mulitagent Systems. ACM, Honolulu Hawaii, (May 14, 2007), 1–8. ISBN: 978-81-904262-7-5. doi:10.1145/1329125.1329157.

[4] John C Harsanyi. 1973. Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory*, 2, 1–23.

[5] Adam Tauman Kalai, Ehud Kalai, Ehud Lehrer, and Dov Samet. 2010. A commitment folk theorem. *Games and economic behavior*, 69, 1, 127–137. Place: Duluth Publisher: Elsevier Inc. doi:10.1016/j.geb.2009.09.008.

[6] Daniel Kokotajlo. 2019. The commitment races problem, (Aug. 23, 2019). Retrieved Dec. 14, 2024 from https://www.alignmentforum.org/posts/brXr7PJ2W4Na2EW2q/the-commitment-races-problem.

[7] Moshe Tennenholtz. 2004. Program equilibrium. *Games and Economic Behavior*, 49, 2, (Nov. 1, 2004), 363–373. doi:10.1016/j.geb.2004.02.002.

[8] Emanuel Tewolde and Vincent Conitzer. 2024. Game transformations that preserve nash equilibria or best-response sets. (June 23, 2024). arXiv: 2111.00076[cs]. doi:10.48550/arXiv.2111.00076.

A Commitment Game Equilibria in Pure Strategies

In Section 2, we noted that pure strategy equilibria of the base game carry over to the commitment game. In this section, we formally establish this connection by proving two supporting lemmas. Lemma 1 shows that any pure strategy equilibrium of the base game, such as (Straight, Straight) in the Game of Chicken, also constitutes an equilibrium outcome in the commitment game. In contrast, Lemma 2 demonstrates that strategy profiles not corresponding to any pure strategy equilibrium in the base game, such as (Move, Move), cannot arise as pure strategy equilibrium outcomes in the commitment game.

Lemma 1. *Consider a two player game where each player has strategies $A = \{A_1, \ldots, A_n\}$. Let $\{A_i, A_{-i}\}$ be a Nash equilibrium in this game. When this game is played with commitment,*

(1) *there is a strategy pair with both players committing to A_i and A_{-i} respectively, that forms an SPNE of the commitment game.*
(2) *if $U_i(A_i, A_{-i}) \geq U_i(A'_i, B_{-i}(A'_i)) \quad \forall A'_i \in A$, then there is a strategy pair where i commits to A_i and $-i$ waits and plays A_{-i} that forms an SPNE of the commitment game.*

(3) *if $U_i(A_i, A_{-i}) \geq U_i(A'_i, B_{-i}(A'_i)) \quad \forall i \in \{1, 2\}, \ \forall A'_i \in A$, then there is a strategy pair where both players wait and then play A_i and A_{-i} that is an SPNE of the commitment game.*

Proof: We prove each part by showing that the constructed strategy profile yields the outcome (A_i, A_{-i}) and that no unilateral deviation is profitable.

1. Suppose both players commit to A_i and A_{-i}, which gives players payoffs $U(A_i, A_{-i})$. Given the other players commitment, player i can unilaterally only deviate to receive any $U_i(A'_i, A_{-i})$ with $A'_i \in A$. But since (A_i, A_{-i}) is a Nash equilibrium of the normal-form game, no such unilateral deviation can improve i's payoff. By symmetry, this is an equilibrium. To make it subgame perfect, the strategies must also specify equilibrium strategies off-path, for which there are multiple options. All such strategy profiles are subgame perfect.

2. Let player i commit to A_i while player $-i$ waits and plays A_{-i} after observing i's commitment. If i deviates by committing to any $A'_i \in A$ instead, $-i$ will best respond with $B_{-i}(A'_i)$ which yields $U_i\big(A'_i, B_{-i}(A'_i)\big)$. By assumption this does not exceed $U_i(A_i, A_{-i})$. Similarly, if $-i$ deviates by committing in period 1, the Nash equilibrium property of (A_i, A_{-i}) ensures she cannot gain. Thus no player has a profitable deviation.

3. If both players wait and then, after the commitment period, play A_i and A_{-i}, the outcome is again (A_i, A_{-i}). Any deviation to some commitment A'_i would lead to the opponent responding with $B_{-i}(A'_i)$, yielding a payoff $U_i\big(A'_i, B_{-i}(A'_i)\big)$, which is less than $U_i(A_i, A_{-i})$ by assumption. Thus this is an SPNE: both players get the best equilibrium outcome if they wait, so committing earlier offers no advantage.

Lemma 2. *Consider a two player base game, that is a symmetric 2×2 game with strategies $\{A_1, A_2\}$ for each player. If a strategy pair $\{A_i, A_{-i}\}$ does not form a Nash equilibrium in the base game, it will not feature in an SPNE of the game with commitment.*

Proof: We prove the contrapositive. If $\{A_i, A_{-i}\}$ features in a commitment game SPNE, it must feature in a base game Nash equilibrium.

1. Let's say players commit to A_i and A_{-i} in the commitment game SPNE. Then, by virtue of this being an SPNE it holds that $U_i(A_i, A_{-i}) \geq u_i(A'_i, A_{-i}) \quad \forall i \in \{1, 2\}, \ \forall A'_i \in A$. Thus, $\{A_i, A_{-i}\}$ is a base game Nash equilibrium.

2. Let's say players wait and then play A_i and A_{-i} in the commitment game SPNE. Then, by virtue of this being an SPNE it holds that $U_i(A_i, A_{-i}) \geq U_i(A'_i, A_{-i}) \ \forall i \in \{1, 2\}, \ \forall A'_i \in A$. Thus, $\{A_i, A_{-i}\}$ is a base game Nash equilibrium.

3. This is where the 2×2 assumption comes in. Let's say player i commits to A_1 and $-i$ waits. If $-i$ best responds with A_1, by symmetry $\{A_1, A_1\}$ are mutual best responses and thus a Nash equilibrium of the base game. If $-i$ best responds with A_2, one must consider what would have happened if i had committed A_2 instead of A_1. If $-i$'s best response to A_2 would have been A_2, by symmetry i would have preferred to commit A_2 instead of A_1, a contradiction. Thus, $-i's$ best response to A_2 must be A_1, and thus by symmetry $\{A_1, A_2\}$ is a base game Nash equilibrium. By symmetry, this also holds for $\{A_2, A_1\}$. This concludes the proof.

B Proof of Proposition 1

Consider the simplified 2×2 symmetric normal form game from table 4. Let A_1 denote the strategy where a player makes a commitment to A_1, and let A_2 denote the strategy where a player commits to A_2. W denotes the strategy of waiting. To prove Proposition 1, we will derive the mixed subgame perfect equilibria of the symmetric 2×2 games. In these equilibria, if a player waits, she will best respond to any commitment the other player makes. If both players wait, they will play the MSNE of the base game which yields expected payoff $\frac{bc}{c+b-1}$. Lemma 3 and Lemma 4 (stated and proven further below) ensure that this is the unique mixed subgame perfect equilibrium.

Battle of Sexes ($b > 1$ and $c > 1$). If player $-i$ mixes between A_1, A_2 and W with respective probabilities p, q, and $1 - p - q$, then player i's pure strategy payoffs are

$$\begin{aligned} U(A_1) &= p \cdot 0 + q \cdot b + (1 - p - q) \cdot b, \\ U(A_2) &= p \cdot c + q \cdot 1 + (1 - p - q) \cdot c, \\ U(W) &= p \cdot c + q \cdot b + (1 - p - q) \cdot \tfrac{bc}{b+c-1}. \end{aligned}$$

For player i to mix over these strategies, she must be indifferent between them. The p and q that equate the pure strategy payoffs are

$$p = \frac{b(1-b)^2}{(b+c-1)(b^2-b+c^2-c)}$$

$$q = \frac{c-(1-p)b}{c-1}$$

It can be shown that $p, q \in [0, 1]$. By symmetry, this holds for player $-i$ as well. To obtain the equilibrium payoff, plug p and q into any of the above pure strategy payoffs. Consider $U(W)$. Note that b and c are both bigger than $\frac{bc}{c+b-1}$. Thus, the payoff in the commitment game equilibrium exceeds the payoff in the base game.

Game of Chicken ($b > 1$ and $1 \geq c > 0$). For $1 \geq c$, waiting weakly dominates A_2. Thus, in the mixed subgame perfect equilibrium there is mixing only over A_1 and W. The probability p that equates $U(A_1)$ and $U(W)$ is

$$p = \frac{b^2-b}{b^2-b+c^2-c+bc}.$$

It can be shown that $p \in [0, 1]$. Consider $U(W)$. Since $c \leq 1 \Leftrightarrow c \leq \frac{bc}{c+b-1}$. Thus, the commitment game payoff is worse than the base game payoff for $c < 1$, except for the edge case with $c = 1$ where it is the same. This concludes the proof of Proposition 1. The next equilibria we derive prove Corollary 1.

Coordination Game ($c < 0$ and $b < 0$). In the Coordination Game and the Stag Hunt the best responses to commitments are different from the best responses in the Game of Chicken and the Battle of Sexes. Here, if player $-i$ mixes over committing A_1, committing A_2, and W, with probability p, q, and $1 - p - q$, player i's pure strategy payoffs are

$$\begin{aligned} U(A_1) &= p \cdot 0 + q \cdot b + (1 - p - q) \cdot 0, \\ U(A_2) &= p \cdot c + q \cdot 1 + (1 - p - q) \cdot 1, \\ U(W) &= p \cdot 0 + q \cdot 1 + (1 - p - q) \cdot \tfrac{bc}{b+c-1}. \end{aligned}$$

In a mixed equilibrium with mixing over these three strategies, all pure strategy payoffs must be equal. The p and q that equate the above pure strategy payoffs are

$$p = \frac{(1-b)\,(bc-b-c+1)}{(1-b-c)(2bc-b-c+1)},$$

$$q = \frac{p(1-c)-1}{-b}\ .$$

It can be shown that $p, q \in [0, 1]$. To compare the equilibrium payoff of the commitment game to the base game MSNE payoff, plug p and q into $U(W)$. Since $0 > \frac{bc}{c+b-1}$, the commitment game payoff exceeds the MSNE base game payoff.

However, note that committing to (A_1, A_1) or players waiting and then playing (A_1, A_1) also form symmetric equilibria in the commitment game. These equilibria yield payoff 1 for each player, which is bigger than $U(W)$.

Stag Hunt ($0 \leq b < 1$ and $c < 0$). Here, committing to A_1 is weakly dominated by waiting. A_1 will not feature in a mixed equilibrium and p can be set to 0. But then, A_2 weakly dominates waiting. Thus, there is no equilibrium in mixed strategies with commitment in which players play the mixed base game equilibrium after waiting. If instead, players play the (A_2, A_2) equilibrium after waiting, players can mix over committing A_2 and waiting in teh commitment game equilibrium. This yields payoff 1 however, which is not a new equilibrium payoff.

Dominant Strategies. When A_1 or A_2 is a (weakly) dominant strategy in the game without commitment, there is little to explore. In equilibrium, players can commit to their dominant strategy, they can wait before playing it or they can mix between committing to it and waiting and then playing it. These equilibria do not offer interesting strategic dynamics, but are mentioned for completeness.

C Proof of Proposition 2

This section proves Proposition 2. We prove 3 lemmas. Lemma 3 states that the proposed mixed equilibrium is the unique subgame perfect equilibrium where players mix over committing and waiting, as long as players play the base game MSNE in the subgame where both waited. Lemma 4 shows that when players waited and do not play the base game MSNE, but rather a pure strategy equilibrium, no mixed equilibrium exists in which both players mix over commitment and waiting. Lemma 5 then describes the unique mixed equilibrium as $T \rightarrow \infty$ and characterizes closed form expressions for all p_t, q_t and $\sigma(W)$ in terms of b, c and T.

Note that we refer to uniqueness in terms of being the only equilibrium where both players mix over committing and waiting. Technically, there can be other equilibria which see some mixing, but these equilibria have the equilibrium outcomes of pure, asymmetric equilibria. For instance, there is an equilibrium where Player 1 mixes over committing in period 1 and 2, while Player 2 mixes over committing in period 3 and waiting. Here, Player 1 effectively induces her preferred outcome, like in a pure strategy equilibrium.

Lemma 3 (Unique Interior MSNE). *Consider the general symmetric 2×2 game in Table 4, with parameters $b > 1$ and $c > 0$. Let each player have the ability to commit to one of two actions, A_1 or A_2, in any of T commitment periods, or to wait until the end. In the subgame where both players wait, they play the mixed-strategy Nash equilibrium of the base game. Then, the commitment game has a unique mixed subgame perfect Nash equilibrium where players mix over committing and waiting. Specifically:*

1. *If $c > 1$, both players mix over all strategies A_{1t}, A_{2t}, and W.*
2. *If $c \leq 1$, both players mix over all strategies A_{2t} and W.*

Proof: Remember that $U(A_{1t})$ denotes the payoff of committing to strategy A_1 in period t. p_t and q_t denote the probability of a player playing strategy A_{1t} or A_{2t}, respectively, and $\sigma(W) = 1 - \sum_{\tau=1}^{T} (p_\tau + q_\tau)$ denotes the probability of waiting. Then, the pure strategy payoffs in a commitment game with T periods are

$$U(A_{1t}) = \sum_{\tau=1}^{t-1} (p_\tau c + q_\tau b) + p_t \cdot 0 + q_t \cdot b + \left(1 - \sum_{\tau=1}^{t}(p_\tau + q_\tau)\right) b \tag{1}$$

$$U(A_{2t}) = \sum_{\tau=1}^{t-1} (p_\tau c + q_\tau b) + p_t c + q_t \cdot 1 + \left(1 - \sum_{\tau=1}^{t}(p_\tau + q_\tau)\right) c \tag{2}$$

$$U(W) = \sum_{\tau=1}^{T} (p_\tau c + q_\tau b) + \sigma(W) \left(\frac{bc}{b+c-1}\right) \tag{3}$$

To illustrate, suppose player 1 plays A_{1t}. With probability $\sum_{\tau=1}^{t-1}(p_\tau + q_\tau)$, player 2 commits to A_1 or A_2 before her, and player 1 best responds, receiving c or b accordingly. [4] With probability $p_t + q_t$, both players commit simultaneously: if 2 also commits A_1, the payoff is 0; if 2 commits A_2, player 1 receives b. With the remaining probability $1 - \sum_{\tau=1}^{t}(p_\tau + q_\tau)$, player 2 was planning to move only after player 1 committed A_1. In this case, player 1 commits, player 2 best responds with A_2 and 1 receives the payoff b.

For now, consider the case $c > 1$. We are looking for equilibria where players mix over committing and waiting. Thus, $\sigma(W)$ and some p_t or q_t must be greater than 0. First, consider

$$U(A_{1,T}) - U(W) = p_T \cdot (-c) + \sigma(W) \cdot \left(\frac{b(b-1)}{b+c-1}\right)$$
$$U(A_{2,T}) - U(W) = q_T \cdot (1-b) + \sigma(W) \cdot \left(\frac{c(c-1)}{b+c-1}\right)$$

If $p_T = 0$ and $\sigma(W) > 0$, $U(A_{1,T}) > U(W)$. It must be that players play A_{1T} with positive probability, otherwise waiting would be dominated. The same holds for A_{2T}, thus $p_T > 0$ and $q_T > 0$. Next, consider

$$U(A_{1t}) - U(A_{1,t+1}) = p_t \cdot (-c) + p_{t+1} \cdot b$$
$$U(A_{2t}) - U(A_{2,t+1}) = q_t \cdot (1-b) + q_{t+1} \cdot (c-1)$$

If $p_t = 0$ and $p_{t+1} > 0$, $U(A_{1t}) > U(A_{1t+1})$. Thus for a player to mix over both A_{1t} and A_{1t+1}, it must be that the other player mixes over both A_{1t} and A_{1t+1} with positive probability. Since both players must have probability $p_T > 0$ as argued above, they must also have probability p_{T-1} greater than 0. It then follows that they must have all probabilities $\{p_t\}_{t=1}^{T}$ greater than 0. The same holds for A_{2t} and q_t. Thus, the only equilibrium where both players mix over committing and waiting is such that both players mix over all strategies $\{A_{1t}, A_{2t}, W\}_{t=1}^{T}$ with positive probability. This concludes the proof for $c > 1$.

When $c \leq 1$, playing any A_{2t} is weakly dominated by waiting: equations 2 and 3 imply $U(A_{2t}) \leq U(W)$. Players will not put probability weight on A_2 and thus all $q_t = 0$. The argument of the proof for $c > 1$ then holds for A_1 and W. In equilibrium, players will mix across all strategies in the set $\{A_{1t}, W\}_{t=1}^{T}$. This concludes the proof.

Lemma 4. *Consider the general symmetric* 2×2 *game in Table 4, with parameters* $b > 1$ *and* $c > 0$. *Let each player have the ability to commit to one of two actions,* A_1 *or* A_2, *in any of* T *commitment periods, or to wait until the end. In the subgame where both players wait, they play one of the pure strategy Nash equilibria of the base game. Then, as long as* $b \neq c$, *the commitment game does not have a mixed subgame perfect Nash equilibrium where both players mix over committing and waiting.*

[4] We adopt the convention that $\sum_{\tau=1}^{t-1}(\cdots) = 0$ when $t = 1$, i.e. the sum is empty and thus vanishes.

Proof: We aim to show that there is no mixed equilibrium in which both players randomize between committing and waiting, when one player gets their preferred equilibrium in the subgame where both players have waited. For the proof, recall that in any mixed equilibrium, players must be indifferent between all strategies they assign positive probability to. We will show that no such indifference condition can be satisfied when both player mix between commitment and waiting strategies.

Let player i get her preferred equilibrium in the subgame where both players have waited. Let p_t^i, q_t^i, and $\sigma(W)^i$ denote the probabilities with which i mixes over strategies A_{1t}, A_{2t}, and W, and analogously for player $-i$. Then it holds that

$$U_i(A_{1,t}) - U_i(W) = \sum_{\tau=t+1}^{T} p_\tau^{-i}(b-c) + p_t^{-i}(-c) + \sigma(W)^{-i}\big(b - \max\{b,c\}\big), \tag{4}$$

$$U_i(A_{2,t}) - U_i(W) = \sum_{\tau=t+1}^{T} q_\tau^{-i}(c-b) + q_t^{-i}(1-b) + \sigma(W)^{-i}\big(c - \max\{b,c\}\big), \tag{5}$$

and analogously for player $-i$:

$$U_{-i}(A_{1,t}) - U_{-i}(W) = \sum_{\tau=t+1}^{T} p_\tau^{i}(b-c) + p_t^{i}(-c) + \sigma(W)^{i}\big(b - \min\{b,c\}\big), \tag{6}$$

$$U_{-i}(A_{2,t}) - U_{-i}(W) = \sum_{\tau=t+1}^{T} q_\tau^{i}(c-b) + q_t^{i}(1-b) + \sigma(W)^{i}\big(c - \min\{b,c\}\big). \tag{7}$$

Assume $b > c$. Then from equations (5) and (7), it follows:

$$U_i(A_{2t}) < U_i(W) \quad \text{and} \quad U_{-i}(A_{2t}) \leq U_{-i}(W) \quad \text{for all } t.$$

Thus, neither player will commit to any A_{2t} with positive probability in any mixed equilibrium. The case $b < c$, is symmetric and implies that neither player will commit to A_1. For the remainder of the proof, we focus on the case $b > c$.

Suppose for contradiction that one player mixes over some, but not all, commitment strategies. Then there exists some t such that $p_t = 0$ for one player. Consider the difference in expected utility between adjacent commitment periods:

$$U(A_{1t}) - U(A_{1,t+1}) = p_t \cdot (-c) + p_{t+1} \cdot b \tag{8}$$

Since $p_t = 0$ but $p_{t+1} > 0$, by iterating equation 8 it follows that for the other player $U(A_{1t})$ is bigger than all $U(A_{1t+1}), \ldots, U(A_{1T})$. By equation 4 it follows that $U(A_{1t}) > U(W)$. Therefore, A_{1t} dominates commitment after period t and waiting, contradicting indifference.

Thus, for $\sigma(W) > 0$ to hold, both players must assign positive probability to all commitment strategies. However, consider

$$U_i(A_{1,T}) - U_i(W) = p_T^{-i} \cdot (-c) + \sigma(W)^{-i} \cdot (b - \max\{b,c\})$$

Since $b > c$, the second term is zero. If player $-i$ assigns positive probability to A_{1T}, i.e., $p_T^{-i} > 0$, then $U_i(A_{1,T}) < U_i(W)$. This implies that player i cannot be indifferent between waiting and A_{1T} and thus will prefer waiting to committing in any commitment period.

Therefore, it is not possible that both players mix over all commitment strategies and waiting in equilibrium. This concludes the proof.

Table 8. Equilibrium Probabilities in the Commitment Game

Strategy	**Symbol**	**Equilibrium Probability**	
		$0 < c \leq 1$	$c > 1$
		Game of Chicken	*Battle of Sexes*
Commit A_1 immediately	p_1	$\frac{1}{\frac{1-\rho^T}{1-\rho}+\rho^T\,\frac{b+c-1}{b-1}}$	$\frac{c-b}{c\left(\frac{c-1}{b-1}\right)\left[\frac{c(c-1)}{b(b-1)}\right]^T-b}$
Commit A_2 immediately	q_1	0	$\frac{p_1b+c-b}{c-1}$
Commit A_1 in period t	p_t	$p_1\left(\frac{c}{b}\right)^{t-1}$	$p_1\left(\frac{c}{b}\right)^{t-1}$
Commit A_2 in period t	q_t	0	$q_1\left(\frac{b-1}{c-1}\right)^{t-1}$
Wait	$\sigma(W)$	$\frac{b+c-1}{b-1}\left(\frac{c}{b}\right)^T p_1$	$\frac{b+c-1}{b-1}\left(\frac{c}{b}\right)^T p_1$

Lemma 5. *Consider the general symmetric* 2×2 *game in Table 4 with parameters satisfying* $b > 1$ *and* $c > 0$*. The mixed subgame perfect Nash equilibrium corresponds to the strategy profile described in Table 8. Table 7 summarizes the convergence of equilibrium probabilities and payoffs as* $T \longrightarrow \infty$*.*

Proof. We will first consider $c > 1$ and then $c \leq 1$.

Case 1: $c > 1$. Suppose player $-i$ mixes over A_{1t}, A_{2t} and W with probabilities p_t, q_t and $\sigma(W) = 1 - \sum_{\tau=1}^{T}(p_\tau + q_\tau)$. Then again, player i's pure strategy payoffs are described by equations 1, 2 and 3. In a mixed equilibrium, all these strategies must yield the same expected payoff:

$$U(A_{11}) = \ldots = U(A_{1,T}) = U(A_{2,1}) = \ldots = U(A_{2,T}) = U(W).$$

It can be shown that

$$U(A_{11}) = U(A_{21}) \qquad \Longleftrightarrow \qquad q_1 = \frac{p_1 b + c - b}{c-1} \tag{9}$$

$$U(A_{1t}) = U(A_{1t+1}) \qquad \Longleftrightarrow \qquad p_{t+1} = p_t \cdot \frac{c}{b} \tag{10}$$

$$U(A_{2t}) = U(A_{2t+1}) \qquad \Longleftrightarrow \qquad q_{t+1} = q_t \cdot \frac{b-1}{c-1} \tag{11}$$

$$U(A_{1T}) = U(W) \qquad \Longleftrightarrow \qquad \sigma(W) = p_T \cdot \frac{c}{b} \cdot \frac{b+c-1}{b-1} \tag{12}$$

$$U(A_{2T}) = U(W) \qquad \Longleftrightarrow \qquad \sigma(W) = q_T \cdot \frac{b-1}{c-1}\frac{b+c-1}{c} \tag{13}$$

We now want to express p_t, q_t and $\sigma(W)$ as closed form solutions in terms of b, c and T. First note that 10 and 11 can be used to write

$$p_t = p_1\left(\frac{c}{b}\right)^{t-1} \qquad q_t = q_1\left(\frac{b-1}{c-1}\right)^{t-1}$$

Next, we will use these to plug into 12 and 13 to obtain

$$p_1\left(\frac{c}{b}\right)^T \frac{b+c-1}{b-1} = q_1\,\frac{b+c-1}{c}\left(\frac{b-1}{c-1}\right)^T$$

Next, we plug in 9 for q_1 to obtain an expression that can be rearranged for p_1. This finally yields:

$$p_1 = \frac{c-b}{c\frac{c-1}{b-1}\left(\frac{c(c-1)}{b(b-1)}\right)^T - b}, \tag{14}$$

We have thus derived a closed form expression for p_1, and using equations 9-13 we can obtain closed form expressions for all p_t, q_t and $\sigma(W)$. Next, let's analyze the equilibrium as $T \longrightarrow \infty$. Define $r = \frac{c(c-1)}{b(b-1)}$. Then

$$p_1 = \frac{c-b}{c\frac{c-1}{b-1}r^T - b}, \quad q_1 = \frac{p_1 b + c - b}{c-1} \quad \text{and} \quad \sigma(W) = p_1\left(\frac{c}{b}\right)^T \frac{b+c-1}{b-1}.$$

As $T \longrightarrow \infty$, we distinguish three cases:

- If $c < b$, then $r < 1$, and
$$p_1 \longrightarrow \frac{b-c}{b}, \qquad q_1 \longrightarrow 0, \qquad \sigma(W) \longrightarrow 0.$$
- If $c > b$, then $r > 1$, and
$$p_1 \longrightarrow 0, \qquad q_1 \longrightarrow \frac{c-b}{c-1}, \qquad \sigma(W) \longrightarrow 0.$$
- Finally, if $c = b$
$$p_1 = \frac{1}{T+1}\frac{b-1}{2b-1}, \quad q_1 = \frac{1}{T+1}\frac{b}{2b-1}, \quad \sigma(W) \longrightarrow 0$$

Note that the case $b = c$ is not derived from equation 14. Instead, we plug equations 9-13 into $\sum_{t=1}^{T}(p_t + q_t) + \sigma(W) = 1$ to obtain closed form expressions. To see the equilibrium payoff in all cases consider $U(W)$. As can be seen, it converges to $\min\{b, c\}$ as $T \longrightarrow \infty$.

Case 2: $c \leq 1$. Here, committing to A_2 is weakly dominated by waiting. Thus all $q_t = 0$ and the mixed equilibrium is only characterized by Equation 10 and 12. Let $\rho = \frac{c}{b}$. Since the probability sequence must satisfy $\sum_{\tau=1}^{T} p_\tau + \sigma(W) = 1$, we can write

$$p_1 \sum_{\tau=0}^{T-1} \rho^\tau + p_1\,\rho^T\,\frac{b+c-1}{b-1} = 1.$$

The sum of the geometric series is $\sum_{\tau=0}^{T-1}\rho^\tau = \frac{1-\rho^T}{1-\rho}$. Plugging this into the above expression and solving for p_1 gives

$$p_1 = \frac{1}{\frac{1-\rho^T}{1-\rho} + \rho^T\,\frac{b+c-1}{b-1}},$$

which together with 10 and 12 give the closed form expressions for all p_t and $\sigma(W)$. Again, let me analyze the equilibrium as $T \longrightarrow \infty$. Note that $\rho^T \longrightarrow 0$, so that the geometric series simplifies to $\frac{1-\rho^T}{1-\rho} \longrightarrow \frac{1}{1-\rho}$. Thus,

$$p_1 \longrightarrow 1-\rho, \qquad \sigma(W) \longrightarrow 0.$$

Consider $U(W)$ again to see that the equilibrium payoff converges to c. Thus fur $c < 1$ the equilibrium payoff worsens as T increases. Note that $c = 1$ is an edge case, where the MSNE base game payoff $\frac{bc}{b+c-1} = 1$ and thus the commitment game equilibrium payoff stays 1 as $T \longrightarrow \infty$. This concludes the proof of Lemma 5.