Commitment and Credibility

Game Theory · Module 10 — Strategic Systems Integration

Node 10.4

Position in Knowledge Graph

System: Game Theory **Structural Domain:** Strategic Systems Integration **Node:** 10.4

A threat, a promise, a contract, or an institutional rule is only as powerful as its credibility. Throughout the course — from subgame perfection (3.3) to trigger strategies (5.3) to mechanism enforcement (6.1) — we have encountered the same constraint: a strategy that prescribes an action at some node is only part of an equilibrium if the player would actually carry out that action when the node is reached. This is the problem of **credibility**. The complementary problem is **commitment**: how to make a non-credible strategy credible by altering the structure of the game so that the player has no choice but to follow through. 10.4 unifies the credibility analysis scattered across earlier modules into a single formal treatment. It develops Schelling's theory of strategic commitment, Nash bargaining as a commitment baseline, hands-tying devices, strategic delegation, and the connection between commitment and institutional design (10.1). This node is the theoretical spine of Module 10: every institution (10.1), every mechanism layer (10.2), every contract (10.3), every collective-choice rule (10.5), and every ecosystem equilibrium (10.6) depends on whether the players' strategies are credible and whether they can commit.

Formal Definition

Let Γ be an extensive-form game (Module 3) with player set N and strategy profiles $s \in S$.

A **commitment device** for player i is a transformation of Γ into a new game Γ' in which player i's action set at some information set is restricted to a singleton (the player's choice is removed; they are forced to take a particular action). Formally, if h is an information set of player i and $A_i(h)$ is the action set at h, the commitment device replaces $A_i(h)$ with $\{a_i^*\} \subseteq A_i(h)$.

A strategy s_i is **credible** at information set h if $s_i(h) \in B_i(h, s_{-i})$, the best-response set at h given the opponents' strategies. A strategy profile s is **subgame-perfect** if every $s_i(h)$ is credible at every h — the definition of SPE from 3.3.

Schelling's commitment principle (1960): A player can improve their equilibrium payoff by *reducing* their own options — specifically, by committing to an action that would otherwise not be credible. The commitment is valuable precisely because it is irreversible: the opponent, observing the commitment, updates their best response and the resulting equilibrium shifts in the committer's favor.

Nash bargaining solution. Two players bargain over a surplus S with disagreement payoffs $d = (d_1, d_2)$. The Nash bargaining solution is

$$(x_1^*, x_2^*) = \arg\max_{(x_1,x_2)\geq d,\ x_1+x_2=S} (x_1 - d_1)(x_2 - d_2)$$

which yields $x_i^* = d_i + (S - d_1 - d_2)/2$. Each player gets their disagreement payoff plus half the surplus. The disagreement point is the **commitment baseline**: a player who can credibly threaten a worse disagreement outcome for the opponent improves their own bargaining position.

Strategic delegation. Player i appoints an agent (delegate) whose preferences differ from i's own, and the delegate plays the game on i's behalf. If the delegation is observable and irreversible (a commitment device), the opponent best-responds to the delegate's preferences, not to i's. The optimal delegate is chosen to maximize i's payoff given the opponent's best response to the delegate.

Intuition

Commitment is the act of voluntarily destroying your own future options. This sounds irrational — and in a single-agent optimization problem it is. But in a game, destroying options can be rational because it changes the opponents' beliefs about your future play, and hence their best responses.

Schelling's paradigmatic example: two cars heading toward each other on a one-lane road. If one driver visibly removes their steering wheel and throws it out the window, the other driver must swerve — the first driver has no option to swerve, and both drivers know this. The commitment (removing the steering wheel) converts a non-credible threat ("I will not swerve") into a credible one ("I cannot swerve"). The driver who commits first wins the game of chicken.

Three structural observations:

1. **Commitment requires observability.** If the opponent does not know you have committed, the commitment has no strategic effect. This is why commitment devices must be public: burning bridges must be visible to the enemy; contracts must be verifiable by both parties; institutional rules must be common knowledge.
2. **Commitment requires irreversibility.** If the commitment can be cheaply reversed, it is not credible. A player who "commits" to a tough bargaining position but can back down at any time has not committed. True commitment requires either physical irreversibility (the steering wheel is gone), legal enforcement (the contract is binding), or reputational cost (backing down destroys reputation, which is a repeated-game mechanism from 5.3).
3. **Too much commitment is costly.** A player who commits to a completely rigid strategy loses the ability to adapt to new information. The optimal commitment is selective: commit on the dimensions where commitment shifts the equilibrium in your favor, but retain flexibility on all other dimensions. This is why real-world commitment devices — contracts, constitutions, reputations — are always partial.

Hands-tying is the general term for commitment by restricting your own future options. The canonical forms are:

- **Contractual commitment:** signing a legally binding agreement (10.3).
- **Reputational commitment:** building a public track record that would be costly to violate (5.5).
- **Institutional commitment:** creating or entering an institution that restricts your future actions (10.1).
- **Physical commitment:** burning bridges, sinking costs, making irreversible investments.

Structural Role

10.4 is the theoretical core of Module 10. Every other node in the module presupposes a theory of commitment and credibility:

- **10.1 Institutions** are credible only if their enforcement mechanisms are self-enforcing (a credibility condition).
- **10.2 Layered mechanisms** require credible regulatory commitment — the regulator must credibly commit to enforcing the outer-layer rules.
- **10.3 Contracts** require credible enforcement by the legal system and credible effort by the agent.

- **10.5 Collective choice** requires credible commitment by voters and representatives to the outcomes of the voting process.
- **10.6 Strategic ecosystems** are stable only if the interaction norms within the ecosystem are credibly maintained.

Looking backward, 10.4 unifies: - **3.3 Subgame perfection:** the formal requirement of credibility in extensive-form games. - **5.3 Trigger strategies:** credibility of punishment as the binding constraint on cooperation. - **5.5 Reputation:** commitment through repeated-game reputation building. - **6.1 Mechanism design:** the designer's credible commitment to the mechanism's rules.

Mathematical Development

The value of commitment in sequential games

Consider a two-player sequential game where player 1 moves first and player 2 observes player 1's move. Let s_1^{NE} be player 1's Nash equilibrium action (without commitment) and s_1^{SPE} be their SPE action. Define the **commitment value** as:

$$\Delta_{\text{commit}} = u_1(s_1^C, B_2(s_1^C)) - u_1(s_1^{\text{SPE}}, B_2(s_1^{\text{SPE}}))$$

where s_1^C is the action player 1 would choose if they could commit (i.e., move first with the opponent best-responding), and B_2 is player 2's best response. In a simultaneous game, there is no first-mover advantage and $\Delta_{\text{commit}} = 0$. In a sequential game where commitment shifts the equilibrium, $\Delta_{\text{commit}} > 0$.

Example: Stackelberg competition. Two firms choose quantities. Simultaneous (Cournot) NE: $q_1 = q_2 = (a-c)/3$ with profit $(a-c)^2/9$ each. Sequential (Stackelberg) with firm 1 as leader: $q_1^S = (a-c)/2, q_2^S = (a-c)/4$, profit $(a-c)^2/8$ for the leader (higher than Cournot). The commitment value is:

$$\Delta_{\text{commit}} = \frac{(a-c)^2}{8} - \frac{(a-c)^2}{9} = \frac{(a-c)^2}{72} > 0$$

The leader's commitment to a large quantity is credible because the quantity is sunk (produced before the follower moves). Without sunk production, the threat to produce a large quantity is not credible (the leader would prefer to reduce output after seeing the follower's quantity).

Nash bargaining and outside options

The Nash bargaining solution $x_i^* = d_i + (S - d_1 - d_2)/2$ has two comparative statics relevant to commitment:

1. **Improving your disagreement payoff** $d_i \uparrow$ increases your bargaining share. Any action that credibly raises your outside option — threatening to walk away, developing a BATNA (best alternative to negotiated agreement) — improves your position.
2. **Worsening the opponent's disagreement payoff** $d_j \downarrow$ also increases your share. Any action that credibly worsens the opponent's outside option — making their alternatives worse, creating switching costs — improves your position.

The commitment dimension is that these threats must be credible: you must actually be willing to walk away, and the opponent must believe it.

Hands-tying: formal model

Player i can, before the game begins, choose a **commitment level** $k_i \in [0, 1]$, where $k_i = 0$ means full flexibility and $k_i = 1$ means complete rigidity. The commitment restricts i's action set to $A_i(k_i) \subseteq A_i(0)$ with $A_i(1) = \{a_i^*\}$.

Theorem (Value of Hands-Tying). In a two-player game where player 1 wants to credibly commit to an action a_1^* that is not a best response to player 2's equilibrium action in the uncommitted game, the optimal commitment level k_1^* is the minimum level that makes a_1^* the only available action, provided:

$$u_1(a_1^*, B_2(a_1^*)) > u_1(s_1^{\text{NE}}, s_2^{\text{NE}})$$

That is, the committed outcome is better for player 1 than the uncommitted equilibrium. If this condition fails, commitment has no value (committing to a threat that does not improve your equilibrium payoff is pointless).

Strategic delegation

Player i delegates to an agent with utility function $\tilde{u}_i \neq u_i$. The opponent observes the delegation and best-responds to $\tilde{u}_i$. Player i receives payoff $u_i(s_{\tilde{i}}, B_j(s_{\tilde{i}}))$ where $s_{\tilde{i}}$ is the delegate's equilibrium action under $\tilde{u}_i$.

Theorem (Optimal Delegation in Cournot Duopoly, Fershtman-Judd 1987). In a linear Cournot duopoly, the owner of firm i delegates to a manager with utility $\tilde{u}_i = \alpha u_i + (1 - \alpha) \cdot \text{revenue}$. The optimal delegation parameter is $\alpha^* < 1$: the manager is instructed to be "more aggressive" than profit-maximizing (partially maximizing revenue, which leads to higher output). This is credible because the manager's contract is observable. The result: the delegating firm produces more, the rival produces less, and the delegating firm earns higher profit than in the symmetric Cournot equilibrium — precisely the Stackelberg outcome achieved through delegation rather than sequential timing.

Credibility and time consistency

A commitment is **time-consistent** if the player, given the opportunity to revise their commitment at a later date, would choose to maintain it. Most interesting commitment devices are **time-inconsistent**: the player would prefer to deviate from the commitment after the opponent has already responded to it. This is why commitment must be physically, legally, or reputationally irreversible — if it were time-consistent, no commitment device would be needed.

Kydland-Prescott (1977): A government that commits to low inflation (to reduce inflationary expectations) has an incentive to deviate from this commitment once expectations are set (the Phillips curve tempts them to exploit low expectations with a surprise inflation). The commitment to low inflation is time-inconsistent. Central bank independence is the institutional commitment device that makes the low-inflation policy credible.

Worked Example

Commitment in a poker context: the image of unpredictability

Consider a regular at a weekly home game who announces: "I never fold top pair. If you bet into my top pair, I call. Always." This is a commitment strategy — by publicly restricting their action set (removing "fold" from the available actions when holding top pair), they change opponents' optimal play.

Without commitment. In the uncommitted equilibrium, opponents bluff into this player at the game-theoretically optimal frequency. The player folds top pair some fraction of the time (against large bets on scary boards). The equilibrium is mixed.

With commitment. If the player credibly commits to never folding top pair, opponents' optimal response is to never bluff into this player when the player could have top pair — the expected value of bluffing is negative against a player who always calls. The player gains: they capture all the value from opponents' bluffs (which now never come) but also from opponents' thin value bets (which must now beat top pair to be profitable).

Credibility analysis. Is this commitment credible? Only if: 1. **Observability:** Opponents know about the commitment (the player has announced it publicly and has a track record of following through). 2.

Irreversibility: The player actually follows through even when it is costly (e.g., top pair on a four-flush board against a large river bet). If the player folds top pair once, the commitment is broken and opponents return to bluffing. 3. **The repeated-game structure makes it self-enforcing:** the player maintains the commitment because the long-run value of the reputation (opponents never bluffing into them) exceeds the short-run cost of occasionally calling with a losing hand.

Formal model. Let v^{commit} be the per-session value of the commitment (opponents stop bluffing, player captures extra value) and c^{commit} be the per-session cost (occasionally calling with a losing hand). The commitment is self-enforcing if $v^{\text{commit}} > c^{\text{commit}}$. In a repeated game with discount factor δ, the commitment is SPE if:

$$\frac{v^{\text{commit}} - c^{\text{commit}}}{1-\delta} > \frac{v^{\text{deviate}}}{1-\delta} \cdot \delta + v^{\text{no-commit}} \cdot \frac{1}{1-\delta}$$

where v^{deviate} is the one-shot gain from deviating (folding top pair when you should call to maintain the commitment). This is exactly the trigger-strategy analysis of 5.3 applied to a commitment strategy.

Poker Anchor

Concept mapping

Formal object	Poker instantiation
Commitment device	Public statement: "I never fold top pair," or a pattern of play that functions as a commitment
Credibility	Track record of actually following through; reputation at the table
Hands-tying	Announcing a rigid strategy that restricts your own future options
Strategic delegation	Telling a friend at the table "if I try to leave before midnight, stop me" (committing to session length)
Nash bargaining	Negotiating a table change, a save in a tournament, or a chop deal
Disagreement point	What each player gets if the chop negotiation fails (continuing to play)
Time inconsistency	The temptation to abandon the commitment when it becomes costly

Concrete situation: final-table chop negotiation

Five players remain in a tournament. The payouts are \$10K, \$7K, \$5K, \$3.5K, \$2.5K. Current chip stacks: A has 45%, B has 25%, C has 15%, D has 10%, E has 5%. A deal is proposed to redistribute the remaining prize money.

Nash bargaining framework. The disagreement point for each player is their ICM equity (the expected payout if they play it out). The surplus from a deal is the risk reduction — each player avoids the variance of playing it out. The Nash bargaining solution splits the surplus equally (adjusted for the number of players):

$$x_i^* = d_i^{\text{ICM}} + \frac{S - \sum_j d_j^{\text{ICM}}}{5}$$

where S is the total remaining prize pool and d_i^{ICM} is player i's ICM equity.

Commitment in chop negotiations. Player A (chip leader) can improve their deal by committing to "I will not accept any deal that gives me less than $9K." If this is credible (A has a reputation for walking away from bad deals, or A genuinely prefers to play it out), opponents must accommodate. If it is not credible (A obviously wants to lock up money and everyone knows it), the threat is ignored.

Player E (short stack) has the weakest negotiating position: their disagreement payoff is lowest (high probability of busting in 5th). But E can improve their position by committing to "I will shove every hand if we don't deal" — making the game riskier for everyone. This worsens the disagreement payoff for all other players, which improves E's bargaining share. The commitment is credible if E's stack is short enough that shoving is nearly optimal anyway.

What this understanding buys you

Commitment is a strategic tool, not a character trait. When a player says "I never bluff" or "I always defend my blinds" or "I never fold to a 3-bet," they are deploying a commitment strategy — whether they know it or not. The game-theoretic question is whether the commitment is credible and whether it improves their equilibrium payoff. Some commitments are valuable (never folding top pair against known bluffers); others are costly (never folding when the opponent has the nuts). The theory tells you which is which.

Credibility is the binding constraint. Most poker "commitment strategies" fail not because the strategy is wrong but because the commitment is not credible. A player who says "I never fold top pair" but folds top pair when the pot is large has no commitment — opponents learn this and adjust. The player who actually never folds top pair, even when it costs them, has a genuine commitment device.

Chop negotiations are Nash bargaining. Every final-table deal is a Nash bargaining problem. Understanding the disagreement point (ICM equity), the surplus (risk reduction), and the role of commitment (credible threats to walk away or play aggressively) gives you a structural framework for these negotiations rather than relying on intuition or social pressure.

Cross-Links

- **3.3 Subgame perfection** — the formal credibility requirement for extensive-form strategies.
- **5.3 Trigger strategies** — commitment through repeated-game reputation.
- **5.5 Reputation and chain-store paradox** — reputation as a commitment device.
- **10.1 Games embedded in institutions** — institutions as credible commitment mechanisms.
- **10.3 Contracts and principal-agent** — contracts as legally enforced commitment.
- **10.5 Political economy** — constitutional commitment and credibility of collective-choice outcomes.
- **Economics / Schelling** — *The Strategy of Conflict* (1960): the original theory of strategic commitment.

Structural Compression

Invariant: A strategy is credible if and only if the player would actually execute it when the relevant decision node is reached; commitment is the act of altering the game's structure to make a non-credible strategy credible, either by removing alternative actions (hands-tying), delegating to an agent with different preferences, or building a reputation whose destruction is more costly than following through; the value of commitment equals the difference between the committed and uncommitted equilibrium payoffs.

Mechanism: Commitment transforms the game tree by restricting the committer's action set at future nodes; the opponent, observing the commitment, updates their best response; the resulting SPE of the modified game gives the committer a higher payoff than the SPE of the original game; credibility requires observability, irreversibility, and either physical constraint, legal enforcement, or reputational cost sufficient to deter deviation.

Cross-System Echo: Commitment and credibility in game theory are structurally identical to **precommitment in dynamical systems** — a controller that "commits" to a feedback policy by fixing certain gains (making them physically unmodifiable) achieves robustness at the cost of adaptability, exactly as a game-theoretic committer achieves favorable equilibrium outcomes at the cost of flexibility. The same tradeoff appears in **constitutional design** (rigidity vs. adaptability), **software architecture** (immutable interfaces vs. flexible implementations), and **biological development** (cell differentiation as irreversible commitment to a developmental pathway).

Exercises

1. In the game of Chicken with payoffs $(0, 0)$ for crash, $(2, 6)$ if player 1 swerves and player 2 goes straight, $(6, 2)$ if reversed, and $(4, 4)$ if both swerve: compute the mixed NE and show that player 1's expected payoff in the mixed NE is less than 6. Then show that if player 1 commits to going straight, the unique SPE gives player 1 payoff 6. Compute the commitment value.
2. In a Nash bargaining problem with $S = 100, d = (20, 10)$, compute the Nash bargaining solution. Then show that if player 1 can credibly raise their disagreement payoff to 40 (by developing an outside option), their bargaining share increases. Compute the increase.
3. In the Stackelberg duopoly with inverse demand $P = a - Q$ and constant marginal cost c, derive the leader's commitment value as a function of a and c. Show that it is positive for all $a > c$.
4. A player in a repeated game wants to commit to grim trigger. Show that the commitment is credible (self-enforcing) if and only if the punishment phase is itself an SPE continuation, which holds when the punishment is a stage Nash equilibrium. Relate this to 5.3.
5. In the Fershtman-Judd delegation model, the owner of a Cournot duopoly firm delegates to a manager who maximizes $\alpha \cdot \text{profit} + (1 - \alpha) \cdot \text{revenue}$. Derive the equilibrium quantities as a function of α, the optimal α^* for the delegating owner (when the rival owner does not delegate), and the equilibrium when both owners delegate.
6. A poker player commits to "always 3-betting with pocket aces preflop" (never slow-playing). Model the table's best response to this commitment. Under what conditions does this commitment increase the player's EV? Under what conditions does it decrease EV?
7. **Argue, structurally,** why the problem of credible commitment is fundamentally different from the problem of optimal strategy selection. Show that a player can have a strategy that is optimal-if-believed but not credible, and a strategy that is credible but not optimal-if-believed. Give one example of each from poker.

Game Theory · Module 10 — Strategic Systems Integration · Node 10.4
Concepts: nash-equilibrium, institutions, repeated-interaction
Prerequisites: 2.1, 3.3, 5.3, 10.1, 10.3
Enables: 10.5, 10.6