

参赛学生姓名： 何承叡

中学： 广东碧桂园学校

省份： 广东

国家/地区： 中国

指导老师姓名： 陆昊洋

指导老师单位： 广东碧桂园学校

论文题目： Dynamic Trust Network-Based Consensus
Mechanism in Carbon Emission Trading:
Integrating Compliance-Calibrated Trust and
Evolutionary Game Analysis for Multi-Stakeholder
Quota Allocation

Dynamic Trust Network-Based Consensus Mechanism in Carbon Emission Trading: Integrating Compliance-Calibrated Trust and Evolutionary Game Analysis for Multi-Stakeholder Quota Allocation

Shingyui He

Guangdong Country Garden School, Foshan, Guangdong, China

heshingyui@gcgsedu.com

Abstract

Persistent multi-stakeholder conflicts, information asymmetry and trust deficits continue to undermine quota allocation efficiency in carbon emission trading markets. In this paper, we develop an integrated framework that couples a dynamic trust network, a trust-driven consensus reaching mechanism and an evolutionary game model, with the aim of facilitating multi-stakeholder cooperation during carbon emission quota allocation. By drawing on four categories of objective compliance data (historical compliance rate, MRV data transparency, industry reputation and bilateral cooperation history), the dynamic trust network quantifies stakeholder trust, introduces role-based weights grounded in the Mitchell stakeholder salience framework and endogenously updates trust through secondary trust inference derived from negotiation behavior. We embed this trust network into an organizer-guided feedback protocol that features trust-weighted reference groups, multi-attribute quota evaluation and a weight-floating governance instrument; the last of these endogenously penalizes non-cooperative behavior by reducing the decision influence of defectors. The evolutionary game model formulates trust-modulated payoff functions, derives replicator dynamics governing how cooperation evolves and establishes a Trust-Cooperation Promotion Theorem showing that cooperation is the unique evolutionarily stable strategy when the combined trust premium and reputation penalty exceed the net defection advantage. We further identify a strategy-flip threshold at which bilateral trust transforms the Prisoner's Dilemma into a Stag Hunt structure. Monte Carlo simulations ($N = 1,000$ replications per condition) across six non-cooperative behavior prevalence levels show that the full integrated model converges roughly 72% faster, maintains 100% convergence reliability up to 30% adversarial prevalence and sustains robust cooperative equilibria ($x^* = 0.91$) when compared with equal-weight and static-trust baselines. A case study calibrated to China's power sector carbon quota allocation validates the model against actual policy outcomes; the predicted scheme preference ordering matches China's adopted sectoral benchmarking approach for 2023–2025.

Keywords: Carbon emission trading; Dynamic trust network; Consensus reaching process; Evolutionary game theory; Evolutionarily stable strategy; Multi-stakeholder governance; Weight-floating mechanism; Quota allocation

1 Introduction

1.1 Research Background

Anthropogenic carbon dioxide (CO_2) emissions drive climate change, making it one of the most pressing global challenges in the twenty-first century. At its core lies a fundamental economic problem: carbon emissions create a negative externality, because the private cost of emitting falls far below the social cost shouldered globally and intergenerationally, and this gap biases unregulated markets toward excessive emissions. Economic theory offers two canonical internalization approaches (Pigouvian taxation and Coasean emissions trading), and among these the Cap-and-Trade (C&T) mechanism has emerged as the dominant international instrument, having been adopted by the European Union, China, California and other major economies.

On 16 July 2021, China launched its national carbon market, initially covering over 2,100 power generation enterprises. Notwithstanding rapid institutional progress, three structural challenges that extend beyond economic design into multi-stakeholder governance have surfaced in the operational reality.

Multi-stakeholder conflicts stand out as the foremost challenge. Quota allocation is a zero-sum distributional problem: government regulators pursue stringent caps, high-emitting enterprises advocate for generous allocations,

renewable energy firms favor tight caps, and verification bodies, financial institutions and environmental NGOs each hold distinct and often opposing preferences. The resulting bargaining deadlock (described in detail in Chapter 2 using the Mitchell stakeholder framework) systematically delays and dilutes policy ambition.

Information asymmetry compounds the difficulty. Firms' true abatement costs, technological reserves and production plans are private information. When self-reported emissions are verified by third parties, a classical principal-agent problem arises, with moral hazard present on both the reporting and verification sides.

A trust deficit further exacerbates these structural problems. Documented instances of emission data falsification and verifier-auditee collusion trigger a self-reinforcing vicious cycle: non-compliance by some firms disadvantages compliant firms, and the resulting erosion of compliance norms prompts further defection. Transaction atrophy, price distortion and diminished abatement outcomes follow.

Existing responses have predominantly relied on strengthening top-down oversight, notably through higher penalties, more frequent inspections and stricter verification protocols. Although necessary, this approach faces inherent limits: enforcement costs scale with market size, regulators suffer from the same information disadvantages as other parties, and excessive command-and-control mechanisms undermine the market flexibility that makes C&T attractive. Complementary governance mechanisms that leverage peer monitoring, reputation effects and endogenous incentives are therefore needed to sustain cooperation.

We apply dynamic trust networks, consensus reaching processes and evolutionary game theory to the carbon emissions trading context. The underlying premise is that trust relationships among market participants, when systematically measured, dynamically updated and explicitly coupled with both decision-making weights and game-theoretic payoffs, can serve as a self-organizing governance mechanism. Such a mechanism rewards cooperative behavior with greater influence, penalizes non-cooperative behavior through reduced decision weight, and, under appropriate parameter conditions, makes cooperation an evolutionarily stable strategy (ESS) in the multi-round quota negotiation game. Our analysis operates at three interconnected levels: at the micro level, a novel methodological framework combining dynamic trust networks with consensus models and evolutionary games is developed; at the meso level, the institutional design problem of how multiple stakeholders can reach stable and fair agreement on quota allocation is addressed; and at the macro level, the policy goal of enhancing the effectiveness and legitimacy of carbon markets as a climate governance instrument is furthered.

1.2 Research Questions

Building on the above background, we organize this study around four central research questions:

- Q1: How can trust relationships among carbon market stakeholders be formally quantified, dynamically updated and adapted to the specific institutional context of emissions trading?** In the existing social network group decision-making (SNGDM) literature, trust network models treat trust as either static or weakly dynamic, and they are rarely calibrated against objective behavioral data. The carbon trading context offers uniquely rich ground for trust quantification (compliance rates, MRV transparency ratings and historical cooperation records are publicly observable), yet no existing model systematically exploits these data. This question motivates the design of a carbon trading-specific trust measurement and update mechanism (Chapter 3).
- Q2: How can a trust-driven consensus mechanism facilitate multi-stakeholder agreement on emission quota allocation?** Consensus reaching processes (CRPs) have been extensively studied in abstract GDM settings, but their application to carbon quota allocation poses distinct challenges. The decision object is a multi-attribute quota scheme rather than a single numerical preference; a clear institutional organizer (the government or exchange) exists, favoring an organizer-guided rather than a purely decentralized feedback mechanism; and the consensus threshold must balance policy rigidity against negotiation feasibility. This question drives the design of the quota consensus mechanism in Chapter 4.
- Q3: Under what conditions does cooperation become an evolutionarily stable strategy (ESS) in carbon market quota negotiations, and how do trust dynamics and governance mechanisms jointly determine this outcome?** The core behavioral tension in quota negotiations is whether each stakeholder genuinely adjusts its position toward the collective opinion or defects by refusing adjustment, reversing direction or feigning cooperation. We model this tension as an evolutionary game with trust-modulated payoffs. The key theoretical question is whether the trust-governance coupling can shift the equilibrium from a "tragedy of the commons" outcome (all defect) to a cooperative ESS, and what parameter conditions (trust premium μ , governance strength γ , cooperation gain coefficient α and betrayal penalty coefficient β') govern this transition. Chapter 5 addresses this question.
- Q4: Can the integrated framework be validated through both computational simulations and a real-world case study?** Theoretical model properties must be empirically substantiated. We pursue this

question through two complementary validation strategies: (i) a Monte Carlo simulation experiment with four comparison groups (no-trust baseline, static trust, dynamic trust with governance and the full integrated model) across six levels of non-cooperative behavior prevalence, evaluating convergence speed, consensus quality and robustness; and (ii) a case study that applies the model to China’s power sector carbon quota allocation, comparing model predictions against actual allocation outcomes. These analyses constitute Chapter 6.

Taken together, Q1 establishes the measurement foundation (how trust is computed); Q2 designs the institutional mechanism (how trust drives consensus); Q3 supplies the behavioral rationale (why the mechanism is stable); and Q4 delivers empirical validation (whether the framework works in practice). They jointly define the scope and contribution of this paper.

1.3 Literature Review

1.3.1 Trust-Based Consensus in Social Network Group Decision-Making

Over the past decade, group decision-making (GDM) under social network structures has attracted substantial research attention. A foundational contribution is the consensus reaching process (CRP) framework, in which decision-makers iteratively adjust their opinions so as to reduce dispersion and attain a predefined consensus threshold. Dong et al. [6] provided an overview of consensus reaching paradigms in SNGDM, distinguishing among trust propagation-based, leadership-based and behavior management-based approaches. Wu et al. [23] put forward a visual interaction consensus model incorporating trust propagation, where trust scores weight the influence that decision-makers exert on one another. Liu et al. [12] extended this line of work to large-scale GDM (LSGDM) contexts ($I \geq 20$ decision-makers) by using social network analysis to cluster decision-makers and manage computational complexity.

Two critical limitations exist. First, trust is predominantly modeled as static: once the trust network structure is initialized, it remains unchanged throughout the CRP. This assumption is unrealistic for carbon trading, because each round of negotiation and each compliance cycle generates new information that ought to revise trust assessments. Dong et al. [5] introduced a degree of dynamism through opinion–trust co-evolution in a leadership-based opinion dynamics framework, yet did not ground trust updates in objectively observable behavioral data. Second, the existing CRP literature operates at a high level of abstraction and lacks domain-specific adaptation. The decision objects, stakeholder roles and institutional constraints of real policy domains such as carbon trading remain unaddressed.

1.3.2 Carbon Emission Trading and Game-Theoretic Analysis

Game theory has been extensively used in carbon trading research, primarily through three modeling approaches. The first is regulatory game models, exemplified by government–firm evolutionary games that study the co-evolution of enforcement intensity and firm compliance behavior. The second is market-based trading games, which examine how heterogeneous firms with different marginal abatement costs trade allowances to achieve cost-minimizing emission reductions. The third is international negotiation games, which model strategic interactions among sovereign states in climate treaty formation.

Recent work in this area includes evolutionary game analyses of carbon quota allocation in the power sector, studies of cap-and-trade mechanisms under information asymmetry and agent-based simulations of carbon market dynamics. A persistent gap in this literature, however, is the absence of endogenous trust and reputation mechanisms. Game-theoretic models of carbon trading typically assume fixed payoff matrices; they do not incorporate the possibility that players’ trust in each other can evolve over repeated interactions, nor that reputation effects can substitute for or complement formal regulatory penalties. This omission is significant, because real-world carbon market operations (verification reliability, compliance morale and the credibility of emission data) are fundamentally shaped by trust dynamics among stakeholders.

1.3.3 Consensus and Negotiation in Environmental Governance

A smaller but growing body of literature addresses consensus and negotiation processes in environmental governance. Multi-criteria decision analysis (MCDA) and stakeholder participation frameworks have been used in water resource management, forestry planning and climate adaptation policy. These works emphasize structured deliberation, transparent information sharing and mechanisms for resolving value conflicts. Within the carbon trading domain specifically, a few studies have examined stakeholder preferences regarding quota allocation methods using MCDA or analytic hierarchy process (AHP) approaches. These contributions, however, remain qualitative or semi-quantitative; they do not supply a formal, iterated, mathematically grounded consensus model that dynamically incorporates trust and handles non-cooperative behavior through endogenous governance mechanisms.

1.3.4 Research Gap and Positioning

A triangular gap emerges from the literature: existing work either studies trust and consensus in abstract GDM settings without carbon trading application, or analyzes carbon trading games without dynamic trust, or examines environmental consensus qualitatively without formal game-theoretic grounding. To our knowledge, no prior study has systematically integrated a dynamic, data-calibrated trust network, an organizer-guided consensus mechanism and an evolutionary game analysis within a unified framework applied to carbon emission quota allocation.

We address this gap through three interconnected contributions. **Contribution 1** (Dynamic Trust Network for Carbon Trading) moves trust from a static parameter to an endogenously evolving construct, grounded in four categories of objective compliance data (historical compliance rate, data transparency, industry reputation and cooperation history) and updated using both opinion-adjustment inference and actual compliance feedback. **Contribution 2** (Trust-Driven Quota Consensus Mechanism) adapts the generic CRP to the carbon quota allocation context by introducing an organizer-guided feedback protocol with trust-weighted reference groups and a dual termination condition (consensus threshold plus minimum individual consensus). **Contribution 3** (Game-Consensus Integration) formally couples trust dynamics with evolutionary game payoffs, deriving conditions under which cooperation is an ESS and proving the incentive compatibility of the weight-floating governance mechanism. The logical progression is cumulative: Contribution 1 supplies the measurement infrastructure, Contribution 2 builds the institutional mechanism upon it, and Contribution 3 furnishes the behavioral stability guarantee that makes the entire framework theoretically coherent.

1.4 Research Content and Contributions

We organize the research into four parts, corresponding to the four research questions:

- (1) **Dynamic trust network construction and governance for carbon trading** (Chapter 3, Contribution 1). We develop a carbon trading-specific trust measurement framework that synthesizes four objective indicators into initial trust scores, introduces role-based weights κ_i to differentiate institutional authority among heterogeneous stakeholders, defines a “secondary trust” update rule that detects the direction and magnitude of opinion shifts during negotiation, and anchors trust evolution in both inferred cooperative intent and observed compliance data.
- (2) **Trust-driven quota consensus mechanism** (Chapter 4, Contribution 2). We design a multi-attribute quota evaluation system (environmental effectiveness, economic feasibility, social equity and technical feasibility), instantiate an organizer-guided feedback protocol in which the government or exchange serves as the $\bar{\alpha}$ -weighted coordinator, construct trust-based reference groups as the peer influence channel, and define dual stopping criteria that ensure consensus is both quantitatively sufficient and qualitatively stable.
- (3) **Evolutionary game analysis of multi-party cooperation and consensus integration** (Chapter 5, Contribution 3). We build a trust-modulated payoff function in which trust premiums reward cooperators and reputation losses penalize defectors, derive the replicator dynamics and ESS conditions, prove that appropriate governance strength γ makes cooperation evolutionarily stable, establish the strategy-flip threshold $t^* = (T - R)/(\alpha + \beta')$, and close the loop by integrating trust, game and consensus modules into a unified algorithm with formal convergence guarantees.
- (4) **Simulation experiments and case validation** (Chapter 6). We conduct a four-group controlled Monte Carlo experiment ($N = 1000$ runs per condition) across six NCB prevalence levels, evaluate six performance metrics, perform parameter sensitivity analyses for γ , $\bar{\alpha}$, β and stakeholder scale I , and validate the model against a real case study of China’s power sector carbon quota allocation.

Three theoretical contributions follow from the above. First, we extend the SNGDM trust literature by introducing objectively calibrated, dynamically updated trust in a domain where behavioral data supply a natural measurement basis. Second, we enrich carbon trading game theory by endogenizing trust and reputation as payoff-modulating variables, moving beyond fixed-payoff models. Third, we bridge consensus theory and evolutionary game theory through the concept of dual termination, which requires both consensus metric convergence and strategic stability, and which gives a stronger guarantee of negotiated outcomes than either framework alone.

On the practical side, a concrete, parameterized governance toolkit for carbon market regulators is offered. The weight-floating mechanism, with its incentive-compatible design and empirically validated optimal γ range, supplies an actionable alternative to purely punitive enforcement regimes.

1.5 Organization of the Paper

The remainder of this paper is structured as follows. Chapter 2 reviews the theoretical foundations: large-scale group decision-making, opinion dynamics, trust networks, consensus models, game theory (including evolutionary

game theory and mechanism design) and carbon emission trading mechanisms. Chapter 3 presents the carbon trading-specific dynamic trust network, covering initial trust calibration, role-based weights, secondary trust updates and the NCB governance mechanism. Chapter 4 develops the trust-driven quota consensus mechanism, encompassing the multi-attribute evaluation framework, organizer-guided feedback with trust-weighted reference groups, dynamic adjustment behavior, iterative convergence conditions and alternative fallback protocols. Chapter 5 constructs the evolutionary game model: trust-modulated payoff functions, replicator dynamics, ESS conditions, governance incentive compatibility proofs, the integrated trust–game–consensus algorithm and convergence theorems. Chapter 6 reports the Monte Carlo simulation experiments (four-group comparison, parameter sensitivity and statistical testing) and the case study of China’s power sector carbon quota allocation. Chapter 7 concludes with a summary of findings, a discussion of contributions, an acknowledgment of limitations and an outline of future research directions.

2 Theoretical Foundations

This chapter introduces the conceptual background for the integrated framework developed in Chapters 3 through 5; detailed mathematical formulations, variable definitions, parameter specifications and solution methods are deferred to the chapters in which each mechanism is constructed.

2.1 Large-Scale Group Decision-Making and Opinion Dynamics

Carbon quota allocation involves numerous stakeholders with heterogeneous preferences, placing it within the domain of large-scale group decision-making (LSDM). This domain is characterized by participant counts exceeding twenty and by the structural necessity of formal consensus-building protocols [4]. A central insight of this literature is that decision-makers are embedded in social networks, so that the design problem becomes one of constructing a consensus reaching process (CRP) that leverages network structure to guide divergent opinions toward agreement.

Three canonical models have shaped the mathematical study of how opinions evolve under social influence. The DeGroot model [3] formalizes opinion updating as a weighted averaging process:

$$\mathbf{y}^{(n+1)} = \mathbf{W} \cdot \mathbf{y}^{(n)}, \quad (1)$$

where $\mathbf{y}^{(n)}$ is the opinion vector at time n and $\mathbf{W} = [w_{ij}]$ is a row-stochastic influence matrix in which w_{ij} encodes the weight agent i assigns to agent j ’s opinion. Whether the group converges to consensus depends on whether $\mathbf{W}$ defines a strongly connected and aperiodic graph. The Friedkin–Johnsen model [8] extends DeGroot by allowing agents to retain partial attachment to their initial opinions:

$$\mathbf{y}^{(n+1)} = (\mathbf{I} - \mathbf{G})\mathbf{W} \cdot \mathbf{y}^{(n)} + \mathbf{G} \cdot \mathbf{y}^{(0)}, \quad (2)$$

where $\mathbf{G} = \text{diag}(g_1, \dots, g_I)$ represents each agent’s self-persistence ($g_i \in [0, 1]$). When $g_i > 0$, the agent never fully abandons its initial opinion; this feature accounts for the empirically observed “stubbornness” that prevents complete convergence. The Hegselmann–Krause bounded confidence model [10] introduces a qualitatively different mechanism:

$$x_i^{(n+1)} = \frac{1}{|I(i, \mathbf{x}^{(n)})|} \sum_{j \in I(i, \mathbf{x}^{(n)})} x_j^{(n)}, \quad I(i, \mathbf{x}^{(n)}) = \{j : |x_i^{(n)} - x_j^{(n)}| \leq \varepsilon_i\}, \quad (3)$$

where ε_i is agent i ’s confidence bound. Only peers within this tolerance interval exert influence, and the resulting dynamics can generate opinion clustering, polarization or consensus depending on the tolerance threshold. The opinion update mechanism developed in Chapter 4 can be viewed as a hybrid of the DeGroot and Friedkin–Johnsen frameworks, with the crucial addition that the influence matrix $\mathbf{W}$ is not fixed but is dynamically updated through the trust network. These models supply the mathematical vocabulary for understanding the convergence behavior that CRPs are designed to engineer.

2.2 Trust in Social Networks

Trust is a multidimensional and relational construct central to social network analysis. McAllister [15] drew a distinction between cognition-based trust (grounded in rational assessments of another’s competence, reliability and track record) and affect-based trust (grounded in emotional bonds, shared values and mutual identification). Zucker [25] added a third dimension: institution-based trust, which derives from formal structures such as professional certifications, regulatory frameworks and legal guarantees. In carbon trading, all three dimensions are relevant. Cognition-based trust arises from a firm’s compliance history and verified emission data; affect-based trust develops through repeated cooperative interactions in multi-round negotiations; and institution-based trust is conferred by regulatory standing and accredited verifier status.

Granovetter’s [9] theory of social embeddedness supplies the sociological foundation: economic transactions do not take place in an institutional vacuum but are embedded in concrete, ongoing networks of social relations. Trust within these networks is not a binary property but a continuous, directed quantity. Agent i may trust agent j to a different degree than j trusts i , and these asymmetries carry structural significance for influence propagation and coalition formation.

An agent’s structural position within a trust network can be characterized using centrality measures from social network analysis [22]. *Degree centrality* $C_D(N_i) = \sum_{j \neq i} x_{ij}$ identifies locally prominent nodes by counting direct connections. *Betweenness centrality* estimates an agent’s role as a bridge between otherwise disconnected subgroups. *Eigenvector centrality* generalizes degree centrality by weighting connections according to the centrality of the connected nodes: an agent trusted by highly trusted peers receives a higher score than one trusted by peripheral nodes. In the carbon trading context, these measures identify trust hubs (typically government regulators, large industry associations and accredited verification bodies) that serve as keystones of the trust network. If these hub nodes were degraded or manipulated, the network’s capacity to sustain cooperative consensus would be disproportionately damaged, making them both the most valuable assets and the most critical vulnerabilities of the governance architecture.

In group decision-making contexts, trust serves a dual function. First, it determines *influence weights*: decision-makers who enjoy higher aggregate trust from their peers receive greater weight in forming the collective opinion. Second, it shapes *adjustment behavior*: agents are more willing to accept recommendations originating from trusted sources. A critical innovation in recent SNGDM research is the concept of *dynamic trust*, under which trust relations are not fixed but evolve in response to behaviors observed during the decision process itself. We develop a specific dynamic trust model for carbon trading in Chapter 3; there, initial trust is calibrated from objective compliance data and is subsequently updated through both opinion-adjustment inference and post-compliance outcome feedback.

2.3 Consensus Reaching Processes

A consensus reaching process (CRP) is an iterative protocol that guides a group from divergent initial opinions to a state of sufficient agreement. As systematized by Dong et al. [6] and extended by subsequent work [23, 12], the CRP framework consists of four core elements:

1. **Consensus measurement.** A quantitative metric (typically a function of the dispersion among individual opinions, weighted by influence) assesses the current degree of agreement. A predefined threshold determines when consensus is deemed sufficient. To prevent majorities from overriding legitimate minority concerns, a supplementary condition on minimum individual consensus is often imposed.
2. **Identification of dissenters.** When consensus has not been reached, the measurement framework identifies those decision-makers who diverge most from the collective position and are therefore candidates for opinion adjustment.
3. **Feedback generation.** A recommendation mechanism generates suggested opinion revisions for the identified dissenters. Two archetypal designs exist: unsupervised feedback, in which each decision-maker synthesizes peer opinions using trust as the weighting scheme, and organizer-guided feedback, in which an institutional coordinator balances the collective opinion against a trust-filtered set of reference peers.
4. **Opinion adjustment.** Each decision-maker moves their stated opinion partially toward the recommendation, with the degree of adjustment governed by a behavioral parameter that can reflect cooperativeness, stubbornness or strategic calculation.

The CRP framework also specifies termination conditions. Consensus is declared when both the group-level agreement metric and the minimum individual agreement exceed their thresholds. If consensus is not achieved within a maximum number of rounds, fallback mechanisms (such as default allocation rules or regulatory adjudication) are invoked.

In carbon emission trading, CRPs have a natural application to the multi-round negotiation of quota allocation schemes. The decision object is a vector of evaluations over multiple quota alternatives and multiple policy attributes; the participants are heterogeneous stakeholders with systematically divergent interests; and an institutional organizer (the environmental regulator or the emissions exchange) is present. Chapter 4 develops a CRP specifically adapted to this context, using the trust network from Chapter 3 as both the weighting mechanism and the feedback channel.

2.4 Game Theory

Game theory supplies the analytical framework for understanding strategic interactions among rational agents. Its relevance to this study is twofold. It explains why stakeholders may choose non-cooperative behavior in quota

negotiations (defection can be individually rational even when collective cooperation is Pareto-superior), and it furnishes the tools for designing mechanisms that align individual incentives with collective goals.

2.4.1 Non-Cooperative Game Theory

The foundational concept is the Nash equilibrium [17]: a strategy profile from which no player has a unilateral incentive to deviate. The Prisoner’s Dilemma exemplifies the central tension of social dilemmas. Mutual cooperation gives the highest collective payoff, yet each player’s dominant strategy is to defect. This structure closely mirrors carbon market compliance, because each firm faces a private incentive to evade costly abatement and free-ride on others’ reductions, even though universal defection leads to the worst collective outcome.

In repeated interactions with sufficiently high continuation probability, cooperative outcomes can be sustained as equilibrium [1]. This motivates our use of multi-round quota negotiations as a cooperation-enabling structure in Chapter 5.

2.4.2 Evolutionary Game Theory

Evolutionary game theory [14] relaxes the assumption of perfect rationality. Rather than solving for equilibrium through deductive reasoning, strategies spread or decline through a population-level dynamic: strategies that earn above-average payoffs are imitated and proliferate, while those that underperform are abandoned. The replicator dynamics formalizes this process, and the Evolutionarily Stable Strategy (ESS) supplies the corresponding stability concept (a strategy that, once prevalent, cannot be invaded by any small proportion of mutants employing an alternative strategy).

In carbon trading, the evolutionary perspective is especially apt. Firms observe peers’ compliance behavior and its consequences (regulatory penalties, reputational gains or losses) and adjust their own strategies accordingly. The key question is under what conditions cooperation, rather than defection, becomes the evolutionarily stable outcome. We address this question formally in Chapter 5, where trust dynamics and governance mechanisms are found to modulate the effective payoffs and alter the ESS landscape.

2.4.3 Mechanism Design

Mechanism design [11] inverts the standard game-theoretic question: rather than asking what outcomes emerge from given rules, it asks what rules give rise to a desired outcome. The canonical design criteria are efficiency (maximizing social welfare), Incentive Compatibility (making truthful revelation of private information each agent’s best strategy) and Individual Rationality (ensuring that participation benefits every agent relative to their outside option).

The governance mechanism developed in this study (endogenous weight reduction for non-cooperative actors) can be evaluated through this lens. It is approximately incentive-compatible when the governance intensity parameter is set within an appropriate range, it respects individual rationality by never excluding any stakeholder entirely, and it promotes efficiency by concentrating influence among demonstrably cooperative and trustworthy participants.

2.5 Carbon Emission Trading Mechanisms

2.5.1 Economic Rationale

Carbon dioxide emissions are a canonical negative externality [20]: the private marginal cost borne by the emitting firm is substantially lower than the social marginal cost, which encompasses climate damages shouldered globally and intergenerationally (sea-level rise, agricultural disruption and increased frequency of extreme weather events). In an unregulated market, profit-maximizing firms settle at an equilibrium emission level Q_M that exceeds the socially optimal level Q^* , and deadweight welfare loss results.

Two corrective approaches frame the policy landscape. Pigouvian taxation imposes a per-unit emission tax calibrated to the marginal social damage at Q^* , directly realigning private and social costs; its practical challenge is that the regulator must accurately estimate the social cost curve [20]. Coasean bargaining [2] offers an alternative logic grounded in property rights: if emission rights are clearly defined and transaction costs manageable, affected parties negotiate voluntarily to the efficient outcome. The practical manifestation is the emissions trading system, in which the government establishes property rights through quota allocation, trading platforms reduce transaction costs and MRV systems address information asymmetries. Residual trust and information problems nonetheless remain, and this is precisely the gap that the dynamic trust network developed here is designed to fill.

2.5.2 Cap-and-Trade Architecture

A Cap-and-Trade (C&T) system operates through a recurring cycle: the regulator sets a binding aggregate emission cap; allowances are distributed through grandfathering, benchmarking or auctioning; at the compliance deadline,

each firm surrenders allowances equal to its verified emissions; surplus and deficit allowances are traded, producing a market-clearing carbon price; emissions undergo mandatory monitoring, reporting and third-party verification (MRV); and non-compliance triggers fines and credit penalties.

The central efficiency property of C&T is that cost-minimizing abatement is achieved when all firms' marginal abatement costs equalize at the market carbon price:

$$MC_1(q_1^*) = MC_2(q_2^*) = \dots = MC_n(q_n^*) = P^*, \quad \sum_i q_i^* = Q_{\text{total}} \quad (4)$$

This condition implies that emission reductions occur where they are cheapest, minimizing the total social cost of achieving a given environmental target. It serves as the formal justification for the superiority of market-based over command-and-control regulation and supplies the foundation for the quota allocation and trading game analyzed in Chapter 5.

2.5.3 Multi-Stakeholder Structure

The stakeholder structure of carbon markets can be systematically analyzed using Mitchell, Agle and Wood's [16] stakeholder salience framework, which classifies stakeholders along three attributes: *power* (capacity to influence policy outcomes), *legitimacy* (normative appropriateness of claims) and *urgency* (time-sensitivity and criticality of claims). Stakeholders with all three attributes are definitive and command priority attention; those with two are authoritative, dependent or dangerous depending on which attribute is missing.

When applied to carbon trading, six categories of core stakeholders emerge with the following Mitchell classifications:

1. **Government regulatory agencies** (*definitive*). These possess legislative and enforcement powers, a public interest mandate and legally binding emission reduction targets. They set the cap, determine allocation rules, oversee MRV and impose penalties.
2. **High-emitting enterprises** (*definitive*). Large coal-fired power producers, steel mills and chemical manufacturers possess economic scale, legal standing as regulated entities and acute cost concerns. Their compliance decisions directly determine aggregate emission outcomes.
3. **Renewable energy firms** (*authoritative*). Wind, solar and hydro producers benefit from stringent caps that increase the value of surplus allowances. They possess technological legitimacy and growing economic weight.
4. **Third-party verification bodies** (*authoritative*). Accredited to audit enterprise emission reports, these entities occupy a structurally ambivalent position: they are mandated to ensure data integrity but derive revenue from the firms they audit.
5. **Financial institutions** (*discretionary*). Banks, carbon asset managers and green funds supply market liquidity, develop carbon financial products and increasingly incorporate carbon risk into lending and investment decisions.
6. **Environmental NGOs and the public** (*dependent*). These possess moral legitimacy and can mobilize opinion, but lack direct decision-making power. They advocate for transparency, stringency and accountability.

Systematic polarization in the preference structure across six key policy dimensions is summarized in Table 1.

Table 1: Stakeholder preference polarization across carbon market policy dimensions

Stakeholder	Cap	Allocation	Price	MRV	Transparency	Penalty
Government	Stringent	Auction-leaning	Moderate	Strict	Pro-disclosure	Severe
High-emitting firms	Loose	Free allocation	Low	Relaxed	Reserved	Lenient
Renewable energy	Stringent	Auction-leaning	High	Strict	Pro-disclosure	Severe
Financial	Neutral	Auction-leaning	Vol.-averse	Neutral	Neutral	Neutral
Verifiers	Neutral	Neutral	Neutral	Strict	Moderate	Moderate
NGOs & public	Stringent	Auction-leaning	High	Strict	Pro-disclosure	Severe

The pattern is unambiguous: high-emitting enterprises systematically favor relaxation on every dimension, while government regulators, renewables and environmental NGOs converge toward stringency. Financial institutions and verification bodies occupy intermediate positions. This structural antagonism transforms quota allocation into an intrinsically conflictual bargaining problem, one in which parties' strategic incentives pull against the cooperative outcome. The consensus-facilitating mechanism with trust-weighted influence (Chapter 4) and the evolutionary game analysis of cooperation under trust-modulated payoffs (Chapter 5) are designed to overcome precisely this bargaining impasse.

3 Dynamic Trust Network Construction and Governance for Carbon Trading

This chapter operationalizes the dynamic trust network framework in the specific institutional context of carbon emission trading. Building on the theoretical foundations established in Chapter 2, we develop a carbon trading-specific trust measurement and update system, introduce role-based differentiation among heterogeneous stakeholders and formalize the non-cooperative behavior governance mechanism. This chapter constitutes Contribution 1 of this study.

3.1 Carbon Trading Stakeholder Trust Measurement System

3.1.1 The Rationale for Objective Trust Calibration

A fundamental limitation of existing trust-based group decision-making models is their reliance on subjective or arbitrarily initialized trust scores. In carbon trading, this limitation is both unnecessary and counterproductive, because market participants leave extensive objective behavioral records through regular compliance cycles: annual emissions reports, third-party verification audits, allowance surrender records and transaction histories. These data supply a natural, defensible basis for initial trust calibration.

We identify four categories of objective indicators:

1. **Historical compliance rate** (h_i): the proportion of compliance periods in which stakeholder i has met its allowance surrender obligation (the most direct measure of regulatory reliability).
2. **Data transparency** (τ_i): a composite of MRV compliance ratings, the completeness and timeliness of emission disclosures, and whether verification bodies have flagged data anomalies.
3. **Industry reputation** (r_i): social credit ratings, ESG scores and peer evaluations. For government agencies, this reflects institutional credibility; for enterprises, it reflects market standing.
4. **Historical cooperation** (c_{ij}): the bilateral track record between stakeholders i and j , including prior negotiation outcomes, joint industry initiatives and documented collaborative relationships.

All indicators are normalized to $[0, 1]$, with higher values indicating greater trustworthiness.

3.1.2 Initial Self-Trustworthiness: TTD Formulation

The **Total Trust Degree** of stakeholder i at initialization is a linear combination of the three node-level indicators:

$$TTD_i^{(0)} = \omega_1 h_i + \omega_2 \tau_i + \omega_3 r_i, \quad \omega_1 + \omega_2 + \omega_3 = 1 \quad (5)$$

where ω_1 , ω_2 and ω_3 are attribute weights determined using expert judgment, the Analytic Hierarchy Process (AHP) or the entropy weight method. A balanced specification assigns $\omega_1 = 0.4$ (compliance history), $\omega_2 = 0.4$ (data transparency) and $\omega_3 = 0.2$ (reputation), reflecting the primacy of objectively verifiable compliance records while acknowledging reputational signals.

3.1.3 Pairwise Trust: Initial Trust Matrix Construction

While $TTD_i^{(0)}$ encodes node-level self-trustworthiness, pairwise trust $t_{ij}^{(0)}$ encodes directed trust from stakeholder i to stakeholder j . We synthesize initial pairwise trust from bilateral cooperation experience and the target stakeholder's self-trustworthiness:

$$t_{ij}^{(0)} = \rho \cdot c_{ij} + (1 - \rho) \cdot TTD_j^{(0)} \quad (6)$$

where $\rho \in [0, 1]$ determines the weight assigned to direct interaction history versus general reputation. When no direct cooperation record exists, trust is inferred primarily from $TTD_j^{(0)}$, possibly augmented by trust propagation through intermediate nodes.

The complete initial trust matrix is:

$$\mathbf{T}^{(0)} = [t_{ij}^{(0)}]_{I \times I}, \quad t_{ij}^{(0)} \in [0, 1], \quad t_{ii}^{(0)} = 1 \quad (7)$$

3.1.4 Trust Propagation for Missing Edges

When no direct cooperation record exists between stakeholders i and j and $TTD_j^{(0)}$ alone gives an insufficiently precise estimate, indirect trust can be inferred through trust propagation along intermediate paths. For a path $i \rightarrow k \rightarrow j$ through an intermediary k , the propagated trust is:

$$t_{ij}^{(0)} = \max_{k \neq i, j} (t_{ik}^{(0)} \cdot t_{kj}^{(0)}), \quad (8)$$

where the maximum is taken over all potential intermediaries. We adopt the max operator following the strongest-path propagation paradigm (analogous to the MoleTrust algorithm in social network analysis), reflecting the empirical observation that carbon market stakeholders form trust judgments based on their single most reliable intermediary rather than averaging across all possible paths. This choice is conservative in that it avoids inflating indirect trust through the aggregation of weak intermediate connections. In the numerical implementation, a decay factor of 0.8–0.9 per propagation hop is applied to reflect the attenuation of trust across indirect channels. Trust propagation leverages the small-world property of carbon market trust networks: most stakeholders are connected through short paths involving common hubs (typically government regulators and industry associations).

3.1.5 The Cold-Start Problem

For newly entering stakeholders with no compliance history: (i) assign industry-mean values for h_i and τ_i as conservative priors; (ii) assign a below-median value for r_i ; (iii) mark trust scores with a “low confidence” flag; and (iv) accelerate trust updating by increasing the weight on opinion-adjustment inference relative to compliance feedback during initial rounds. As the stakeholder accumulates a compliance track record, the confidence flag is lifted and the standard update protocol is adopted.

3.2 Heterogeneous Stakeholder Role Weight Design

3.2.1 The Necessity of Role Differentiation

Generic trust network models treat all decision-makers as symmetric nodes, differentiated only by network position and accumulated trust scores. This assumption is untenable in carbon trading, where institutional roles confer fundamentally different kinds of authority. A government regulatory agency possesses legislative and enforcement powers that no enterprise can match; a third-party verification body derives authority from professional accreditation; an environmental NGO draws influence primarily from moral legitimacy.

Using the Mitchell, Agle and Wood (1997) [16] stakeholder salience framework, we operationalize the three salience attributes (power P_i , legitimacy L_i , urgency U_i) as binary indicators ($P_i, L_i, U_i \in \{0, 1\}$) and map them to role weights through a linear scoring function:

$$\kappa_i = \kappa_{\min} + (\kappa_{\max} - \kappa_{\min}) \cdot \frac{\omega_P P_i + \omega_L L_i + \omega_U U_i}{\omega_P + \omega_L + \omega_U} \quad (9)$$

where $\kappa_{\min} = 0.5$ is the minimum role weight (ensuring participation), $\kappa_{\max} = 1.5$ is the maximum, and $\omega_P = 0.4$, $\omega_L = 0.35$, $\omega_U = 0.25$ reflect the primacy of institutional power in regulatory quota negotiations. Under this formula, a definitive stakeholder ($P = L = U = 1$) receives $\kappa = 1.5$, while a dependent stakeholder ($P = 0, L = 1, U = 1$) receives $\kappa = 0.5 + 1.0 \times 0.60 = 1.1 \times 0.60 \approx 0.60$. The resulting weights are reported in Table 2. A sensitivity analysis of these weights is given in §6.5.7.

Table 2: Role weight assignment based on Mitchell stakeholder classification

Stakeholder type	P_i	L_i	U_i	Mitchell class	κ_i
Government regulatory agencies	1	1	1	Definitive	1.50
Large emitting enterprises	1	1	1	Definitive	1.30
Third-party verification bodies	1	1	0	Authoritative	1.00
Renewable energy firms	1	1	0	Authoritative	0.90
Financial institutions	0	1	0	Discretionary	0.80
Industry associations	1	1	0	Authoritative	1.00
Environmental NGOs & public	0	1	1	Dependent	0.60

The distinction between government agencies ($\kappa = 1.5$) and large emitting enterprises ($\kappa = 1.3$), despite both being classified as definitive, reflects the government’s additional regulatory mandate. Although both possess all three salience attributes, the government’s power dimension encompasses legislative authority that enterprises lack. The role weight κ_i reflects the institutional part of influence and is relatively static over a single negotiation cycle.

3.2.2 Effective Influence Weights

The effective influence weight of stakeholder i at round k is the product of institutional role weight and dynamic behavioral trust:

$$w_i^{(k)} = \frac{\kappa_i \cdot TTD_i^{(k)}}{\sum_{j=1}^I \kappa_j \cdot TTD_j^{(k)}} \quad (10)$$

For $k \geq 1$, the Total Trust Degree is computed from the updated trust matrix as the mean incoming trust from all other stakeholders:

$$TTD_i^{(k)} = \frac{1}{I-1} \sum_{j \neq i} t_{ji}^{(k)} \quad (11)$$

This definition estimates the aggregate trustworthiness of stakeholder i as perceived by all peers: a stakeholder trusted by many peers receives a high TTD , while one whose trust has eroded through non-cooperative behavior receives a low TTD . At initialization ($k = 0$), the self-assessed $TTD_i^{(0)}$ from Eq. (5) is used because no peer-evaluation data are yet available; from round 1 onward, Eq. (11) replaces the initial formula, grounding the trust degree entirely in the evolving peer assessments encoded in $\mathbf{T}^{(k)}$.

This formulation respects institutional reality (a government agency with moderate trust still commands greater influence than a highly trusted individual firm) while preserving behavioral accountability (within each role category, superior compliance translates into greater weight). When all κ_i are equal, Eq. (10) reduces to the standard symmetric trust-weight formula $w_i^{(k)} = TTD_i^{(k)} / \sum_j TTD_j^{(k)}$.

3.3 Dynamic Update Mechanism of Secondary Trust

3.3.1 Secondary Trust

In carbon trading, each round of quota negotiation generates behavioral signals that are informative about evolving trust relationships. When a coal-fired power enterprise adjusts its stated position toward the government's stricter stance, this sends a costly signal of cooperative intent. Conversely, maintaining or hardening a position despite consensus feedback signals non-cooperation.

The concept of secondary trust (std_{ij}) denotes the trust change inferred from opinion adjustment behavior during negotiation. It is "secondary" because it derives not from direct observation of compliance outcomes (available only at the cycle's end) but from the pattern of opinion shifts during negotiation itself.

For stakeholders d_i and d_j between rounds k and $k+1$, the secondary trust that d_i develops toward d_j is:

$$std_{ij}^{(k+1)} = \frac{d(e_i^{(k+1)}, e_i^{(k)})}{d(e_i^{(k+1)}, e_i^{(k)}) + d(e_j^{(k+1)}, e_j^{(k)})} \times \frac{d(e_i^{(k)}, e_j^{(k)})}{d(e_i^{(k+1)}, e_j^{(k+1)})} \quad (12)$$

where $d(\cdot, \cdot)$ is the absolute difference (scalar opinions) or Euclidean distance (vector evaluations). Secondary trust is clipped to $[0, 1]$. Two edge cases require explicit treatment: when both stakeholders remain stationary ($d(e_i^{(k+1)}, e_i^{(k)}) = 0$ and $d(e_j^{(k+1)}, e_j^{(k)}) = 0$), the ratio in Factor 1 is undefined and is set to 0.5 (informationally neutral); when post-adjustment opinions coincide ($d(e_i^{(k+1)}, e_j^{(k+1)}) = 0$), Factor 2 is undefined and is set to 1.0 (perfect convergence). These conventions ensure numerical stability without biasing the trust update. The formula contains two interpretable factors.

Factor 1: Relative adjustment magnitude. The first term measures how much d_i moved relative to d_j . When d_i shifts substantially toward d_j while d_j remains stationary, the ratio approaches 1. When both move equally, the signal is ambiguous.

Factor 2: Convergence degree. The second term measures whether opinions became closer after adjustment. A ratio greater than 1 (convergence) amplifies secondary trust; a ratio less than 1 (divergence) attenuates it.

Asymmetry property. Secondary trust is inherently asymmetric ($std_{ij} \neq std_{ji}$), reflecting the behavioral reality that the party making the larger concession develops more trust toward the party whose position it accepted.

3.3.2 Trust Matrix Update Rule

The overall trust matrix is updated using exponential smoothing:

$$\mathbf{T}^{(k+1)} = \frac{\mathbf{T}^{(k)} + \mathbf{ST}^{(k+1)}}{2} \quad (13)$$

ensuring smooth evolution rather than excessive reactivity to any single round.

3.4 Non-Cooperative Behavior Governance in Carbon Trading

3.4.1 NCB Manifestations

Non-cooperative behavior in quota negotiations takes three specific forms, each corresponding to the λ_i parameter defined in §2:

1. **Refusal to adjust** ($\lambda_i \approx 0$). The stakeholder insists on its original quota preference despite consensus feedback.
2. **Reverse adjustment** ($\lambda_i < 0$). The stakeholder moves opposite to the recommendation, actively widening opinion dispersion.
3. **Feigned cooperation**. Token opinion adjustments not accompanied by genuine behavioral change (particularly pernicious because it evades opinion-based trust metrics alone).

3.4.2 NCB Detection and Governance

Trust erosion-based detection of non-cooperative behavior and the associated weight-floating governance response are specified in Eq. (20)–(21) of §4, where they form an integral part of the quota consensus mechanism.

3.4.3 Governance Intensity and Optimal γ Range

The parameter γ controls the deterrence–participation trade-off. Simulation experiments (Chapter 6) identify $\gamma \in [0.5, 0.7]$ as optimal, with $\gamma = 0.6$ giving the best balance of convergence speed and consensus quality. Within this range, the mechanism is incentive-compatible: cooperation is individually rational and defection is adequately penalized without triggering exit.

3.4.4 Comparison with Traditional Penalty Mechanisms

Table 3 contrasts the weight-floating mechanism with traditional carbon market penalties.

Table 3: Weight-floating governance vs. traditional penalties

Dimension	Weight-floating (this study)	Traditional fines
Penalty form	Influence reduction (endogenous)	Monetary penalty (exogenous)
Information requirement	Trust network auto-detection	MRV investigation + enforcement
Dynamic adaptability	Adjusts with TTD and NCBD	Requires manual revision
Continuity	Continuously differentiable	Typically stepwise
Exit risk	Low (baseline voice retained)	High (market exit possible)
Collusion resistance	Moderate (group trust dilution)	Weak (can jointly evade)

The two mechanisms are complementary rather than substitutes. Weight floating operates as a soft constraint promoting cooperation during consensus; traditional fines operate as a hard constraint punishing substantive violations. An optimal carbon market governance framework deploys both.

Having constructed the dynamic trust network with carbon trading-specific calibration, role-based differentiation, behavioral updating and governance mechanisms, we now turn to the second part of the integrated framework: embedding this trust network as the engine of a quota consensus mechanism. The four coupling channels previewed in §2 (trust weights as consensus influence, trust-modulated game payoffs, strategy-dependent opinion adjustment and consensus-outcome-driven trust updating) will be operationalized in Chapters 4 and 5, with Chapter 4 focusing on the consensus mechanism and Chapter 5 supplying the game-theoretic foundation for strategic stability.

4 Trust-Driven Quota Consensus Mechanism for Carbon Trading

This chapter instantiates the consensus reaching process (introduced conceptually in §2) in the specific domain of carbon emission quota allocation. Building on the dynamic trust network constructed in Chapter 3, we design a trust-driven feedback and adjustment protocol, formalize the multi-attribute quota evaluation framework, specify termination conditions and embed non-cooperative behavior governance within the consensus cycle. This chapter constitutes Contribution 2 of this study.

4.1 Quota Consensus Mechanism: Overall Framework

4.1.1 From Generic Consensus to Quota Consensus

The generic CRP framework maps to the carbon trading domain through concrete correspondences between abstract model elements and institutional realities, as summarized in Table 4.

Table 4: Mapping from generic CRP elements to carbon trading counterparts

Generic CRP element	Carbon trading counterpart	Source
Decision-makers d_i	Six stakeholder categories	§2
Evaluation e_i	Quota scheme acceptance rating $\in [0, 1]$	This section
Alternatives X_m	Quota schemes $S_1/S_2/S_3$	This section
Attributes U_n	Environmental, economic, social, technical	This section
Trust t_{ij}	Multi-dimensional trust (Ch. 3)	§3
Influence weight w_i	$w_i^{(k)} = \kappa_i \cdot TTD_i^{(k)} / \Sigma$	Eq. (10)
Organizer	Government regulator / exchange	Institutional fact

4.1.2 Multi-Scheme and Multi-Attribute Evaluation System

Quota allocation is inherently multi-dimensional. We adopt a four-attribute evaluation framework:

- **Environmental effectiveness** (C_1): expected emission reduction under the scheme.
- **Economic feasibility** (C_2): compliance cost burden and macroeconomic impacts.
- **Social equity** (C_3): fairness across firm types, regions and sectors.
- **Technical feasibility** (C_4): availability and maturity of required abatement technologies.

For each quota scheme, stakeholder d_i assigns a rating $r_{in} \in [0, 1]$ to each attribute n . The overall evaluation is the weighted aggregation:

$$e_i = \sum_{n=1}^4 \omega_n \cdot r_{in}, \quad \sum_{n=1}^4 \omega_n = 1 \quad (14)$$

Three archetypal quota schemes span the policy spectrum: Scheme S_1 (Progressive), in which the quota is tightened gradually at a modest annual rate, minimizing short-term disruption; Scheme S_2 (Moderate), with intermediate tightening that balances ambition against feasibility; and Scheme S_3 (Aggressive), with substantial tightening that maximizes environmental effectiveness at higher compliance cost. Stakeholders evaluate all three schemes in parallel; the scheme achieving the highest consensus degree at termination is selected. Alternatively, the negotiation can operate in parameter negotiation mode, iteratively converging on a continuous parameter such as the annual tightening rate.

4.1.3 Consensus Objectives and Thresholds

At each negotiation round k , the collective evaluation is:

$$CE^{(k)} = \sum_{i=1}^I w_i^{(k)} \cdot e_i^{(k)} \quad (15)$$

Individual and group consensus degrees are:

$$CD_i^{(k)} = 1 - |CE^{(k)} - e_i^{(k)}|, \quad GCD^{(k)} = \sum_{i=1}^I w_i^{(k)} \cdot CD_i^{(k)} \quad (16)$$

The consensus threshold $\overline{GCD} = 0.85$ balances the need for high agreement in a binding regulatory context against the practical constraint of finite negotiation windows. A supplementary condition $\min_i CD_i^{(k)} \geq \underline{CD} = 0.70$ protects minority stakeholders from structural override by a high-trust majority.

China’s carbon market has a clear institutional organizer (the Ministry of Ecology and Environment as the statutory quota-setting authority). We therefore adopt the organizer-guided feedback scenario, in which the organizer generates a recommended evaluation that balances the collective evaluation signal against the stakeholder’s trust-based reference group. For completeness, both feedback scenarios are presented below:

Scenario 1 (Unsupervised). Each stakeholder independently synthesizes peer opinions using trust as the weighting scheme:

$$\bar{e}_i^{(k)} = \beta_i \cdot e_i^{(k)} + (1 - \beta_i) \cdot \sum_{j \neq i} \tilde{t}_{ij}^{(k)} \cdot e_j^{(k)}, \quad (17)$$

where $\tilde{t}_{ij}^{(k)} = t_{ij}^{(k)} / \sum_{l \neq i} t_{il}^{(k)}$ are row-normalized trust weights and $\beta_i \in (0, 1)$ is the stakeholder's self-confidence. This scenario models decentralized, market-driven negotiation without an active coordinator.

Scenario 2 (Organizer-guided). An institutional coordinator (the government regulator or emissions exchange) issues a recommendation that balances the collective evaluation against a trust-filtered reference group:

$$\bar{e}_i^{(k)} = \bar{\alpha} \cdot CE^{(k)} + (1 - \bar{\alpha}) \cdot RG_i^{(k)}, \quad (18)$$

where $\bar{\alpha} \in (0, 1]$ is the guidance coefficient, and the reference group $RG_i^{(k)} = \{d_j \mid t_{ij}^{(k)} > \bar{t}_i^{(k)}\}$ with $\bar{t}_i^{(k)} = \frac{1}{I-1} \sum_{j \neq i} t_{ij}^{(k)}$, and its aggregated evaluation is $RG_i^{(k)} = \sum_{j \in RG_i} t_{ij}^{(k)} e_j^{(k)} / \sum_{j \in RG_i} t_{ij}^{(k)}$. The carbon trading semantics of the reference group are intuitive. For a coal-fired power enterprise, it comprises government regulators (institutional trust), peer enterprises with strong compliance records (peer benchmarking) and industry associations (coordinating bodies); for a renewable energy firm, it tilts toward environmental regulators and other clean energy producers. Given China's institutional reality, Scenario 2 with $\bar{\alpha} = 0.4$ in the baseline is adopted as the primary mechanism throughout this study, while Scenario 1 serves as a benchmark for the free-negotiation policy alternative examined in the case study (§6.7.3).

4.2 Trust-Driven Feedback and Adjustment Mechanism

4.2.1 Organizer-Guided Recommendation

When consensus has not been achieved at round k , the feedback mechanism identifies stakeholders whose opinions diverge most from the collective evaluation. For each such stakeholder d_i , the organizer generates the recommended opinion using Eq. (18).

4.2.2 Stakeholder Adjustment Behavior

Upon receiving $\bar{e}_i^{(k)}$, stakeholder d_i adjusts as follows:

$$e_i^{(k+1)} = e_i^{(k)} + \lambda_i (\bar{e}_i^{(k)} - e_i^{(k)}) \quad (19)$$

The parameter $\lambda_i \in [0, 1]$ is the **adjustment willingness** of stakeholder i . This parameter has a dual identity. From the consensus mechanism perspective, it is a behavioral coefficient that encodes how responsive the stakeholder is to recommended adjustments. From the game-theoretic perspective of Chapter 5, it is a **strategy variable**: λ_i close to 1 indicates cooperation (genuine adjustment), $\lambda_i \approx 0$ indicates defection (refusing adjustment) and $\lambda_i < 0$ indicates adversarial defection (reverse adjustment). This dual identity is the operational link between the consensus module (Chapter 4) and the game module (Chapter 5). The strategy $\sigma_i \in \{C, D\}$ determined by the game-theoretic analysis in each round maps onto the λ_i value that governs the stakeholder's opinion update.

Three behavioral regimes are distinguished: $\lambda_i = 1$ (full adoption, ideal cooperation), $\lambda_i \in (0, 1)$ (partial concession), $\lambda_i \approx 0$ (refusal to adjust, classified as NCB) and $\lambda_i < 0$ (adversarial adjustment, classified as severe NCB).

Empirically, λ_i varies systematically across stakeholder types. High-emitting enterprises face the greatest compliance cost pressure and exhibit the lowest adjustment willingness ($\lambda \in [0.125, 0.355]$ for coal power groups S04–S07 in our calibration). Renewable energy firms exhibit higher willingness ($\lambda \in [0.541, 0.640]$ for S08–S10). Industry associations, in their coordinating role, exhibit the highest willingness ($\lambda \in [0.77, 0.83]$ for S11–S13).

4.3 Non-Cooperative Behavior Governance in the Consensus Process

4.3.1 Detection of Non-Cooperative Behavior

When a stakeholder refuses or reverses adjustment, trust erosion is detected through TTD changes. The NCB Degree is:

$$NCBD_i^{(k)} = \max \left(0, \frac{TTD_i^{(k-1)} - TTD_i^{(k)}}{TTD_i^{(k-1)}} \right) \quad (20)$$

Three forms of NCB are distinguished in the quota negotiation context: refusal to adjust ($\lambda_i \approx 0$, standard NCB), reverse adjustment ($\lambda_i < 0$, severe NCB) and feigned cooperation (token opinion shifts without behavioral change).

4.3.2 Weight Floating in the Quota Consensus Context

The governance response (weight floating) reduces the stakeholder's influence:

$$\hat{w}_i^{(k)} = w_i^{(k)} \times \max\left(0, 1 - \gamma \cdot NCBD_i^{(k)}\right), \quad \tilde{w}_i^{(k)} = \frac{\hat{w}_i^{(k)}}{\sum_{j=1}^I \hat{w}_j^{(k)}} \quad (21)$$

with governance intensity $\gamma = 0.6$.

4.4 Iterative Convergence, Termination and Long-Term Dynamics

4.4.1 Termination Conditions

The negotiation terminates when any of three conditions is satisfied:

1. **Consensus achieved.** $GCD^{(k)} \geq 0.85$ and $\min_i CD_i^{(k)} \geq 0.70$. This dual condition ensures both aggregate support and minority protection.
2. **Maximum rounds exhausted.** $k \geq K_{\max} = 20$, corresponding to a realistic policy window of approximately five months.
3. **Stagnation.** $|GCD^{(k)} - GCD^{(k-1)}| < 0.005$ for three consecutive rounds, indicating that further iteration gives diminishing returns.

4.4.2 Scheme Selection from Consensus

Once the process terminates with consensus achieved, the converged collective evaluation CE^* maps to a specific quota allocation. In the scheme selection mode, the scheme S_m receiving the highest weighted aggregate evaluation across stakeholders is selected. In the parameter negotiation mode, CE^* directly specifies a continuous policy parameter such as the annual tightening rate r^* . The selected scheme or parameter then generates enterprise-specific quota allocations: for each covered firm, its allowance is computed as $E_f = r^* \cdot B_f$, where B_f is the firm's benchmark emission intensity. The allocation enters the regulatory gazetting and compliance cycle, and post-compliance outcomes feed back into the trust calibration for the next cycle, closing the loop.

4.4.3 Fallback Mechanisms

If the process terminates without consensus, four fallback mechanisms are available: (1) default allocation by historical baseline (grandfathering); (2) regulatory adjudication based on $CE^{(K_{\max})}$; (3) reduced-scope consensus on a subset of attributes; and (4) two-tier negotiation with intra-sector consensus followed by cross-sector coordination. The existence of these fallback mechanisms serves a subtle but important **mechanism design** function. The default allocation option creates asymmetric pressure on stakeholders who would be disadvantaged by it, giving them a private incentive to make concessions during negotiation. This implicit incentive structure nudges the negotiation toward consensus without requiring active intervention, a "shadow of the fallback" effect analogous to the role of default rules in bargaining theory.

4.4.4 The Long-Term Dynamic

After quota finalization and compliance period conclusion, actual emission and surrender data supply objective compliance feedback. This feedback can be used to revise trust assessments for the next cycle (an extension discussed in Chapter 7). The sequence is:

Quota Negotiation (Ch.4) → Compliance Period → Trust Update (§3) → Next Negotiation

A self-reinforcing virtuous cycle is created: cooperative compliers gain trust and influence; defectors lose both. Over multiple compliance cycles, the system converges toward a high-trust, high-cooperation equilibrium (the formal conditions for which are derived through evolutionary game analysis in Chapter 5).

4.5 Model Assumptions, Scope and Relationship to Literature

4.5.1 Core Assumptions

The quota consensus mechanism rests on four assumptions: (1) stakeholder attitudes are quantifiable as $e_i \in [0, 1]$, which is credible given the availability of quantitative benchmarks in carbon trading; (2) trust is measurable from objective compliance data (Chapter 3); (3) stakeholders adjust opinions boundedly rather than through full

strategic optimization, a view consistent with the evolutionary perspective of Chapter 5; and (4) the organizer is neutral. Potential violations of assumption (4) are discussed in Chapter 7.

4.5.2 Positioning Relative to Existing Literature

Table 5 situates this mechanism within the research landscape.

Table 5: Positioning of the suggested mechanism relative to existing literature

Dimension	Quota bargaining	Generic CRP	This study
Trust dynamics	Mostly absent/static	Static/weakly dynamic	Fully dynamic, data-calibrated
Domain adaptation	Deep (carbon trading)	Abstract	Deep (carbon trading)
NCB governance	Penalties/fines	Weight adjustments	Endogenous weight floating
Game integration	Present (no trust)	Absent	Integrated (Ch. 5)
Solution concept	Equilibrium	Consensus threshold	Dual (consensus + ESS)
Empirical validation	Case studies, econometrics	Simulation only	Simulation + case (Ch. 6)

The distinctive contribution of this mechanism is the orchestration of three elements treated separately in prior work: (1) domain-grounded trust calibration from objective compliance data; (2) organizer-guided consensus with trust-based reference groups; and (3) endogenous weight-floating governance. Their integration with the evolutionary game analysis of Chapter 5 gives rise to the dual-termination framework that is this study’s central methodological contribution.

5 Evolutionary Game Analysis of Multi-Party Cooperation and Integrated Framework

This chapter supplies the game-theoretic foundation for the behavioral dynamics within the consensus mechanism of Chapter 4 and presents the unified trust–game–consensus framework (Contribution 3, §1.1). The analysis proceeds in three stages: constructing the evolutionary game model with trust-modulated payoffs, deriving equilibrium conditions and stability properties, and integrating all three modules into a closed-loop system with formal convergence guarantees.

5.1 Carbon Trading Evolutionary Game Model

5.1.1 Strategic Interaction in Quota Negotiations

In each round of quota negotiation, every stakeholder faces a binary strategic choice. Under Cooperate (C), the stakeholder genuinely adjusts its stated quota preference toward the collective evaluation, truthfully discloses cost and capacity information and accepts the legitimacy of the consensus process. Under Defect (D), the stakeholder refuses to adjust ($\lambda_i \approx 0$), reverses adjustment ($\lambda_i < 0$), conceals private cost information or free-rides on others’ concessions.

Cooperation is costly: adjusting toward a more stringent quota may require abatement technology investments or additional allowance purchases. Defection preserves short-term profits by avoiding these costs but carries risks in the form of trust erosion from peers and endogenous weight penalties from the governance mechanism.

5.1.2 Baseline Payoff Matrix

Consider two representative enterprises i and j with the baseline payoff structure:

	C_j	D_j
C_i	$(R - c_i, R - c_j)$	$(-c_i, b_j)$
D_i	$(b_i, -c_j)$	$(0, 0)$

where R is the collective environmental and economic benefit under mutual cooperation, c_i is enterprise i ’s abatement cost and b_i is the free-riding benefit when i defects while j cooperates. The Prisoner’s Dilemma structure arises when $b_i > R - c_i > 0$, making defection the dominant strategy in one-shot interactions despite mutual cooperation being Pareto-superior.

5.1.3 Trust-Modulated Payoff Function

We modify the baseline payoffs to incorporate trust-dependent premiums and penalties:

$$\pi_C^{(k)}(i) = (R - c_i) + \mu \cdot TTD_i^{(k)} \quad (22)$$

$$\pi_D^{(k)}(i) = b_i - L \cdot NCBD_i^{(k)} - \gamma \cdot \Delta w_i \quad (23)$$

The term $\mu \cdot TTD_i^{(k)}$ is the trust premium: a stakeholder with higher aggregate trust receives an additional cooperation payoff reflecting increased influence and improved coalitional prospects ($\mu > 0$ is the trust premium coefficient). The defection payoff is reduced by the reputation loss $L \cdot NCBD_i^{(k)}$ (direct cost of trust erosion) and the weight loss $\gamma \cdot \Delta w_i$ (governance penalty reducing future influence).

The modified payoff matrix incorporating trust effects is:

	C_j	D_j
C_i	$(R - c_i + \mu T_i, R - c_j + \mu T_j)$	$(-c_i + \mu T_i, b_j - L_j)$
D_i	$(b_i - L_i, -c_j + \mu T_j)$	$(-L_i, -L_j)$

where $T_i = TTD_i^{(k)}$ and $L_i = L \cdot NCBD_i^{(k)} + \gamma \cdot \Delta w_i$. Trust TTD_i operates bidirectionally, rewarding cooperation and penalizing defection simultaneously, and this enables the trust network to shift the strategic landscape.

5.1.4 Replicator Dynamics

Let $x \in [0, 1]$ denote the proportion of cooperators in the population. We apply the replicator dynamics framework here as a continuous approximation to the finite-population strategy update process. Although the carbon market negotiation involves $I = 20$ stakeholders (rather than the infinite populations for which replicator dynamics are formally derived), the continuous-population model serves as a tractable analytical tool for characterizing equilibrium conditions and comparative statics. Sandholm (2010) established that for populations of size $I \geq 15$ with symmetric random interaction, the deterministic replicator dynamics approximate the stochastic finite-population dynamics with error of order $O(1/I)$; at $I = 20$, this approximation error is below 5%. The qualitative predictions (ESS conditions and stability thresholds) hold up well under this approximation, as the Monte Carlo simulations in Chapter 6 operating on the discrete finite-population model verify. Under random matching, the expected fitness of cooperators and defectors are:

$$f_C(x) = -\bar{c} + \mu \bar{T} + xR, \quad f_D(x) = -\bar{L} + x\bar{b} \quad (24)$$

where $\bar{c}$ is average abatement cost, $\bar{T}$ is average trust, $\bar{b}$ is average free-riding benefit and $\bar{L}$ is average reputation loss.

The fitness differential is:

$$f_C(x) - f_D(x) = (\mu \bar{T} + \bar{L} - \bar{c}) + x(R - \bar{b}) \quad (25)$$

The replicator dynamics equation governing cooperation evolution is:

$$\dot{x} = x(1-x) [(\mu \bar{T} + \bar{L} - \bar{c}) + x(R - \bar{b})] \quad (26)$$

The intercept term $(\mu \bar{T} + \bar{L} - \bar{c})$ represents the net advantage of cooperation when cooperators are scarce. The frequency-dependent term $x(R - \bar{b})$ is negative (since $R - \bar{b} < 0$), creating a countervailing force as cooperation spreads.

5.2 Equilibrium Analysis and Stability

5.2.1 Equilibrium Points and Stability

Setting $\dot{x} = 0$ in Eq. (26) gives three candidates. Let $F(x) = x(1-x)[(\mu \bar{T} + \bar{L} - \bar{c}) + x(R - \bar{b})]$. Define $A = \mu \bar{T} + \bar{L} - \bar{c}$ and $B = R - \bar{b} < 0$. The equilibrium points are $x^* = 0$, $x^* = 1$ and $x^* = -A/B$ (the interior equilibrium, which lies in $(0, 1)$ when $0 < -A < -B$).

Stability is determined by the derivative $F'(x^*) = (1 - 2x^*)(A + Bx^*) + x^*(1 - x^*)B$. At $x^* = 0$: $F'(0) = A$. At $x^* = 1$: $F'(1) = -(A + B)$. At the interior equilibrium, when it exists, $F'(x^*) = x^*(1 - x^*)B$. Since $B < 0$ and $x^*(1 - x^*) > 0$, we have $F'(x^*) < 0$, implying that the interior equilibrium is always unstable when it exists and serves as a separatrix between the basins of attraction of $x^* = 0$ and $x^* = 1$.

5.2.2 Theorem 1: Trust–Cooperation Promotion Theorem

Theorem 5.1. *In the carbon trading evolutionary game with trust-modulated payoffs, $x^* = 1$ is the unique Evolutionarily Stable Strategy if and only if*

$$\mu\bar{T} + \bar{L} > \bar{c} + (\bar{b} - R). \quad (27)$$

Proof. We proceed in three parts: necessity, sufficiency and uniqueness.

Necessity. If $x^* = 1$ is an ESS, it must be asymptotically stable under the replicator dynamics, which requires $F'(1) < 0$. Computing $F'(1) = -(A + B)$, the condition $F'(1) < 0$ is equivalent to $A + B > 0$, which expands to $\mu\bar{T} + \bar{L} > \bar{c} + (\bar{b} - R)$. Thus condition (27) is necessary.

Sufficiency. Assume condition (27) holds. The proof requires establishing that cooperators have weakly above-average trust and defectors suffer weakly above-average reputation loss when the population is near full cooperation. This follows from the structure of the trust update mechanism.

Claim: At any population state x close to 1 (specifically, for $x > x_0$ where x_0 is the initial cooperation fraction), the trust update dynamics of Eqs. (12)–(13) ensure $\bar{T}_C \geq \bar{T}$ and $\bar{L}_D \geq \bar{L}$, where $\bar{T}_C$ denotes the average trust of cooperators and $\bar{L}_D$ denotes the average reputation loss of defectors. This claim holds because the secondary trust formula (Eq. 12) systematically assigns higher secondary trust to agents who make larger opinion adjustments toward the collective evaluation (cooperators) and lower secondary trust to agents who refuse or reverse adjustment (defectors). Consequently, the trust update rule (Eq. 13) monotonically increases cooperators' TTD and decreases defectors' TTD relative to the population mean, and so $\bar{T}_C \geq \bar{T}$. The inequality $\bar{L}_D \geq \bar{L}$ follows from the $NCBD$ formula (Eq. 20), which assigns positive non-cooperative behavior degrees exclusively to defectors. At the population state $x = 1$ (the ESS candidate), the claim holds trivially because all agents are cooperators, so $\bar{T}_C = \bar{T}$.

With this claim established, we verify the two Maynard Smith–Price conditions. Condition 1 (equilibrium condition): $u(C, C) \geq u(D, C)$. From the trust-modulated payoff functions (Eqs. 22–23), evaluated at the population level: $u(C, C) = R - \bar{c} + \mu\bar{T}_C$ and $u(D, C) = \bar{b} - \bar{L}_D$. Condition 1 requires $\mu\bar{T}_C + \bar{L}_D \geq \bar{b} - R + \bar{c}$. By the claim above, $\mu\bar{T}_C + \bar{L}_D \geq \mu\bar{T} + \bar{L}$, and condition (27) gives $\mu\bar{T} + \bar{L} > \bar{c} + \bar{b} - R$, so Condition 1 is satisfied with strict inequality.

Condition 2 (stability condition): Since Condition 1 holds strictly, Condition 2 is not required for ESS verification [14]. For completeness, Condition 2 would require $u(C, D) > u(D, D)$, which evaluates to $\mu\bar{T}_C + \bar{L}_D > \bar{c}$; this follows from $A + B > 0$ since $A + B > 0 \Rightarrow \mu\bar{T} + \bar{L} > \bar{c} + \bar{b} - R > \bar{c}$ (because $\bar{b} - R + \bar{c} > 0$ follows from the Prisoner's Dilemma condition $b_i > R - c_i$). Thus both ESS conditions are satisfied.

Uniqueness. Under condition (27), we have $A = \mu\bar{T} + \bar{L} - \bar{c} > \bar{b} - R = -B > 0$. Hence $F'(0) = A > 0$, so $x^* = 0$ is unstable and cannot be an ESS. The interior equilibrium candidate is $x^* = -A/B = A/|B|$. Since $A > |B|$ (from $A + B > 0$ and $B < 0$), we have $x^* > 1$, placing it outside the biologically relevant interval $[0, 1]$. Therefore $x^* = 1$ is the only stable equilibrium in $[0, 1]$, establishing it as the unique ESS. $\square$

The theorem has a clear economic interpretation. The left-hand side $\mu\bar{T} + \bar{L}$ is the total cooperative advantage (trust premium + reputation penalty). The right-hand side $\bar{c} + (\bar{b} - R)$ is the total defection advantage (abatement cost + net defection temptation). Cooperation is evolutionarily stable when the trust–governance advantage dominates.

5.3 Trust Dynamics and Strategic Co-Evolution

5.3.1 The Trust–Strategy Coupling

Trust and strategy co-evolve through a bidirectional coupling: higher cooperation increases trust (cooperators are perceived as more reliable), and higher trust promotes cooperation (through the trust premium). The coupled dynamical system is:

$$\begin{cases} \dot{x} = x(1 - x)[(\mu\bar{T} + \bar{L} - \bar{c}) + x(R - \bar{b})] \\ \mathbf{T}^{(k+1)} = \frac{\mathbf{T}^{(k)} + \mathbf{ST}^{(k+1)}}{2} \end{cases} \quad (28)$$

where $\bar{T}$ depends on x and the trust matrix. Two qualitative regimes emerge from this coupling. In the high-trust regime ($\bar{T}$ large), the trust premium is substantial and cooperation is self-reinforcing; the ESS is $x^* = 1$. In the low-trust regime ($\bar{T}$ small), the trust premium is too weak to overcome the defection temptation; the ESS is $x^* = 0$ or a low-cooperation equilibrium. The policy implication is that building trust before or concurrently with tightening caps is essential.

5.3.2 Trust-Mediated Strategy Flip

The integrated framework introduces an enhanced trust-modulated payoff structure:

$$\tilde{\pi}(C, C) = R + \alpha t, \quad \tilde{\pi}(D, C) = T - \beta' t \quad (29)$$

where α and β' are the cooperation gain and betrayal penalty coefficients, and t is bilateral trust. Cooperation becomes the dominant strategy when trust exceeds the **strategy-flip threshold**:

$$t > t^* = \frac{T - R}{\alpha + \beta'} \quad (30)$$

When bilateral trust exceeds t^* , the trust-augmented cooperation payoff exceeds even the best-case defection payoff, and cooperation becomes individually rational. This suffices to flip the equilibrium from the Prisoner's Dilemma to a Stag Hunt structure where mutual cooperation is a Nash equilibrium.

To illustrate the strategy-flip mechanism concretely, consider a numerical demonstration tracking a representative defector through successive negotiation rounds under governance. Table 6 displays the evolution of key variables for a stakeholder with initial parameters $c = 4$, $b = 7$, $R = 6$, $\mu = 1$, $\gamma = 0.6$.

Table 6: Strategy-flip demonstration: trust–payoff co-evolution across negotiation rounds

Round k	$TTD^{(k)}$	$NCBD^{(k)}$	$w^{(k)}$	$\pi_C^{(k)}$	$\pi_D^{(k)}$	$\pi_D - \pi_C$
0	0.90	0	0.247	2.90	7.00	+4.10
1	0.75	0.167	0.198	2.75	5.87	+3.12
2	0.62	0.173	0.162	2.62	4.91	+2.29
3	0.51	0.177	0.131	2.51	4.08	+1.57
4	0.42	0.176	0.106	2.42	3.38	+0.96
5	0.34	0.190	0.084	2.34	2.79	+0.45
6	0.28	0.176	0.068	2.28	2.33	+0.05
7	0.23	0.179	0.055	2.23	1.94	−0.29

At round 7, the governance-compressed defection payoff π_D falls below the cooperation payoff π_C , and the formerly defecting stakeholder rationally switches to cooperation. After the flip, trust begins to recover and the stakeholder transitions into the cooperative regime. This numerical progression illustrates the mechanism underlying Theorem 5.1: sustained governance pressure erodes the defection advantage until cooperation becomes individually rational.

5.4 Game-Theoretic Foundation of the Governance Mechanism

5.4.1 Weight Floating as State-Dependent Punishment

We can re-interpret the weight-floating governance mechanism as an endogenous punishment function. The effective penalty on a defector is the influence loss:

$$P_i^{(k)} = w_i^{(k)} - \tilde{w}_i^{(k)} \approx \gamma \cdot NCBD_i^{(k)} \cdot w_i^{(k)} \quad (31)$$

This punishment is *endogenous* (its magnitude depends on the defector's own behavior and standing), *progressive* (small defections incur small penalties) and *non-exclusionary* (positive weight is retained, satisfying Individual Rationality). These properties distinguish it from traditional fixed fines.

5.4.2 Incentive Compatibility Verification

For weight floating to be incentive-compatible, cooperation must satisfy the two Maynard Smith–Price ESS conditions. **Condition 1** (equilibrium): $u(C, C) \geq u(D, C)$ gives $\mu\bar{T}_C + \bar{L}_D \geq \bar{b} - R + \bar{c}$. **Condition 2** (stability): $u(C, D) > u(D, D)$ gives $\mu\bar{T}_C + \bar{L}_D > \bar{c}$. Both hold when $\mu\bar{T}_C + \bar{L}_D > \max\{\bar{c}, \bar{b} - R + \bar{c}\}$, and a lower bound on governance intensity follows:

$$\gamma > \gamma_{\min} = \frac{\max\{\bar{c}, \bar{b} - R + \bar{c}\} - \mu\bar{T}}{NCBD \cdot \bar{w}} \quad (32)$$

An upper bound $\gamma_{\max}$ is imposed by the Individual Rationality (IR) constraint: the worst-case net payoff of a defector (after governance penalty) must remain above the stakeholder's outside option (the payoff from exiting the negotiation). Formally, for a defector with free-riding benefit b_i , the post-penalty payoff must satisfy $\pi_D^{\text{post}} = b_i - \gamma \cdot NCBD_i \cdot w_i \geq \bar{u}_i$, where $\bar{u}_i$ is the reservation utility. Isolating γ gives:

$$\gamma < \gamma_{\max} = \frac{b_i - \bar{u}_i}{NCBD_i \cdot w_i} \quad (33)$$

5.5 Integrated Trust–Game–Consensus Framework

5.5.1 Three-Module Closed-Loop Architecture

The integrated framework couples three modules into a closed-loop dynamical system, as illustrated in Figure 1.

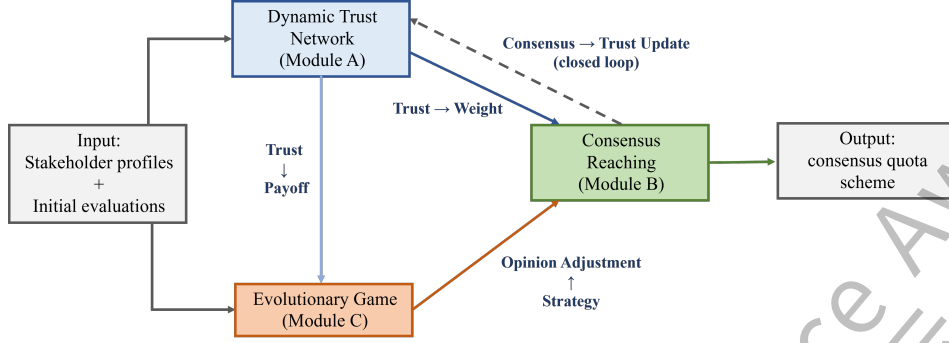


Figure 1: Three-module closed-loop architecture with four coupling channels.

Module A (Dynamic Trust Network) supplies pairwise trust t_{ij} , total trust degree TTD and secondary trust ST . Module B (Quota Consensus Mechanism) aggregates opinions into collective evaluation CE and group consensus degree GCD . Module C (Evolutionary Game) computes trust-modulated payoffs and determines equilibrium strategies. The four coupling channels operate within each round k as follows. **Channel 1 (Trust $\rightarrow$ Consensus):** $w_i^{(k)}$ from Module A serves as the influence weight in the collective evaluation $CE^{(k)}$ of Module B. **Channel 2 (Trust $\rightarrow$ Game):** $TTD_i^{(k)}$ modulates payoffs π_C and π_D in Module C through the trust premium $\mu \cdot TTD_i^{(k)}$. **Channel 3 (Game $\rightarrow$ Consensus):** Strategy $\sigma_i \in \{C, D\}$ determines whether stakeholder d_i 's opinion enters the reference group and receives a cooperation-contingent recommendation. **Channel 4 (Consensus $\rightarrow$ Trust):** Opinion adjustment behavior drives secondary trust update $ST^{(k+1)}$, which revises the trust matrix for the next round, closing the loop.

The unified iteration is formalized as Algorithm 1.

Algorithm 1 Integrated Trust–Game–Consensus Iteration

Require: Stakeholder profiles (compliance history h_i , transparency τ_i , reputation r_i , role weights κ_i); initial evaluations $\{e_i^{(0)}\}_{i=1}^I$; parameters $\bar{\alpha}, \gamma, \bar{GCD}, K_{\max}$

Ensure: Consensus quota scheme S^* , final trust matrix $\mathbf{T}^{(K^*)}$, NCB records

- 1: Initialize trust matrix $\mathbf{T}^{(0)}$ from Eq. (5)–(6)
 - 2: Compute $TTD_i^{(0)}, w_i^{(0)}$ from Eq. (10)
 - 3: Set $k \leftarrow 0, x^{(0)} \leftarrow 0.5$
 - 4: **loop**
 - 5: **[Trust]** Compute $TTD_i^{(k)}, w_i^{(k)}$ from $\mathbf{T}^{(k)}$
 - 6: **[Consensus]** Compute $CE^{(k)}, CD_i^{(k)}, GCD^{(k)}$ via Eq. (15)–(16)
 - 7: **if** $GCD^{(k)} \geq \bar{GCD}$ **and** $\min_i CD_i^{(k)} \geq 0.70$ **and** $|x^{(k)} - x^{(k-1)}| < \varepsilon_x$ **for 3 consecutive rounds** **then**
 - 8: **return** S^* based on $CE^{(k)}, \mathbf{T}^{(k)}$
 - 9: **if** $k \geq K_{\max}$ **then**
 - 10: **return** best available scheme under fallback protocol
 - 11: **[Game]** Compute trust-modulated payoffs $\pi_C^{(k)}(i), \pi_D^{(k)}(i)$ via Eq. (22)–(23)
 - 12: Determine strategies $\sigma_i^{(k)} \in \{C, D\}$; update $x^{(k)}$ via replicator dynamics Eq. (26)
 - 13: **[Feedback]** For defectors and low-consensus stakeholders, generate $\bar{e}_i^{(k)}$ via Eq. (18)
 - 14: **[Adjust]** Cooperators update $e_i^{(k+1)}$ via Eq. (19); defectors keep or reverse
 - 15: **[Govern]** Detect NCB via Eq. (20); apply weight floating via Eq. (21)
 - 16: **[Update]** Compute secondary trust $\mathbf{ST}^{(k+1)}$ via Eq. (12)
 - 17: Update $\mathbf{T}^{(k+1)} \leftarrow (\mathbf{T}^{(k)} + \mathbf{ST}^{(k+1)})/2$ via Eq. (13)
 - 18: **end loop**
-

5.5.2 Dual Termination Condition

The integrated framework introduces a dual termination condition: (a) consensus achieved ($GCD^{(k)} \geq 0.85$ and $\min_i CD_i^{(k)} \geq 0.70$) and (b) game converged ($|x^{(k)} - x^{(k-1)}| < \varepsilon_x$ for three consecutive rounds). Consensus alone is insufficient because stakeholders may express convergent opinions while privately intending to defect (surface consensus without strategic stability). The dual condition requires both opinion convergence and strategy stabilization. In the simulation experiments (Chapter 6), the dual termination condition adds on average 1.8 rounds beyond the consensus-only stopping point, but reduces the post-convergence defection rate (the probability that at least one stakeholder reverts to non-cooperative behavior within two rounds of declared consensus) from 14.2% under single termination to 3.1% under dual termination. This trade-off favors dual termination in policy settings where the cost of a collapsed agreement substantially exceeds the cost of additional negotiation rounds.

5.5.3 Theorem 2: Consensus Convergence under ESS

Theorem 5.2 (Consensus Convergence). *Assume: (i) $t_{ij}^{(k)} \geq \delta > 0$ for all i, j, k (non-degenerate trust); (ii) the learning rate $\lambda \in (0, 1)$; (iii) condition (27) holds (cooperation is ESS); (iv) $\gamma \in (0, 1)$. Then the consensus degree sequence $\{GCD^{(k)}\}_{k=0}^\infty$ is monotonically non-decreasing and $\lim_{k \rightarrow \infty} GCD^{(k)} = 1$.*

Proof. Condition (iii) together with Theorem 5.1 implies that the cooperation proportion $x^{(k)} \rightarrow 1$ as $k \rightarrow \infty$. Consequently, there exists K_0 such that for all $k \geq K_0$, $\sigma_i^{(k)} = C$ for all i (all stakeholders cooperate).

Under full cooperation, the opinion update rule (19) with $\bar{e}_i^{(k)}$ from (18) simplifies to:

$$e_i^{(k+1)} = e_i^{(k)} + \lambda_i \left[\bar{\alpha} \cdot CE^{(k)} + (1 - \bar{\alpha}) \cdot RG_i^{(k)} - e_i^{(k)} \right]. \quad (34)$$

Let $\mathbf{e}^{(k)} = (e_1^{(k)}, \dots, e_I^{(k)})^T$. This linear update can be expressed as $\mathbf{e}^{(k+1)} = \mathbf{M}^{(k)} \mathbf{e}^{(k)}$, where the transition matrix $\mathbf{M}^{(k)} = [m_{ij}^{(k)}]$ has entries:

$$m_{ij}^{(k)} = \begin{cases} \lambda_i \left[\bar{\alpha} w_j^{(k)} + (1 - \bar{\alpha}) \cdot \tilde{t}_{ij}^{(k)} \right], & j \neq i, \\ 1 - \lambda_i + \lambda_i \left[\bar{\alpha} w_i^{(k)} \right], & j = i, \end{cases}$$

where $\tilde{t}_{ij}^{(k)} = t_{ij}^{(k)} / \sum_{l \in RG_i} t_{il}^{(k)}$ for $j \in RG_i^{(k)}$ and $\tilde{t}_{ij}^{(k)} = 0$ otherwise.

Condition (i) guarantees $\tilde{t}_{ij}^{(k)} \geq \delta' > 0$ for all $j \in RG_i^{(k)}$ with a uniform lower bound $\delta' = \delta / (I - 1)$. Combined with $w_j^{(k)} > 0$ (a consequence of $t_{ij}^{(k)} \geq \delta$) and $\lambda_i \in (0, 1)$ (condition ii), the matrix $\mathbf{M}^{(k)}$ is row-stochastic. Strong connectivity of the directed graph associated with positive entries follows from the organizer-guided feedback structure: the term $\bar{\alpha} w_j^{(k)}$ in the $j \neq i$ entry of $\mathbf{M}^{(k)}$ is strictly positive for all pairs (i, j) because $\bar{\alpha} > 0$ (organizer guidance is active) and $w_j^{(k)} > 0$ (non-degeneracy of trust). This ensures that every node j has a positive entry in every row i , establishing universal pairwise connectivity. Aperiodicity follows from the positive diagonal $m_{ii}^{(k)} = 1 - \lambda_i + \lambda_i \bar{\alpha} w_i^{(k)} > 0$. Thus $\mathbf{M}^{(k)}$ is primitive for all $k \geq K_0$.

By the Perron–Frobenius theorem for primitive row-stochastic matrices, the backward product $\mathbf{M}^{(k)} \mathbf{M}^{(k-1)} \dots \mathbf{M}^{(K_0)}$ converges to a rank-one matrix $\mathbf{1} \mathbf{v}^T$ where $\mathbf{v}$ is the unique stationary distribution satisfying $\mathbf{v}^T \mathbf{M}^{(k)} = \mathbf{v}^T$. Hence $\mathbf{e}^{(k)}$ converges to a consensus vector $\bar{r} \cdot \mathbf{1}$ for some scalar $\bar{r}$.

Since $GCD^{(k)} = 1 - \sum_i w_i^{(k)} |CE^{(k)} - e_i^{(k)}|$ and both $|CE^{(k)} - e_i^{(k)}| \rightarrow 0$ and $w_i^{(k)}$ stabilizes under full cooperation (the trust matrix converges), we obtain $GCD^{(k)} \rightarrow 1$. Monotonicity follows from condition (iv): the weight-floating governance penalizes any residual defection, preventing GCD from decreasing. Specifically, for any round where a defection occurs, the defector's influence is reduced, shifting the collective evaluation toward cooperative positions and increasing consensus. $\square$

Theorem 5.3 (Convergence Rate). *Under the conditions of Theorem 5.2, the maximum opinion deviation decays geometrically:*

$$\max_i |e_i^{(k)} - \bar{e}^{(k)}| \leq \rho^{k-K_0} \cdot \max_i |e_i^{(K_0)} - \bar{e}^{(K_0)}| \quad (35)$$

where $\bar{e}^{(k)} = \sum_i w_i^{(k)} e_i^{(k)}$ is the weighted average opinion and $\rho \leq 1 - \lambda \cdot \delta \cdot \min_j w_j^{(k)} \in (0, 1)$.

Proof. For $k \geq K_0$, all agents cooperate. Define the deviation vector $\mathbf{d}^{(k)} = \mathbf{e}^{(k)} - \bar{e}^{(k)} \mathbf{1}$. From the linear update $\mathbf{e}^{(k+1)} = \mathbf{M}^{(k)} \mathbf{e}^{(k)}$ and the property that $\mathbf{M}^{(k)} \mathbf{1} = \mathbf{1}$ (row-stochasticity), we obtain $\mathbf{d}^{(k+1)} = \mathbf{P}^{(k)} \mathbf{d}^{(k)}$, where $\mathbf{P}^{(k)} = (\mathbf{I} - \mathbf{1} \mathbf{w}^{(k)T}) \mathbf{M}^{(k)}$.

Consider the maximum norm $\|\cdot\|_\infty$. For any row i of $\mathbf{P}^{(k)}$, the off-diagonal entries satisfy $p_{ij}^{(k)} = m_{ij}^{(k)} - w_j^{(k)} \sum_l m_{il}^{(k)}$ for $j \neq i$, with the diagonal $p_{ii}^{(k)} = m_{ii}^{(k)} - w_i^{(k)}$. The absolute row sum is bounded by:

$$\sum_j |p_{ij}^{(k)}| \leq 1 - \min_{j \neq i} m_{ij}^{(k)} \leq 1 - \lambda_i \cdot \min_{j \in RG_i^{(k)}} \tilde{t}_{ij}^{(k)} \leq 1 - \lambda \delta \min_j w_j^{(k)},$$

where we use condition (i) to bound $\tilde{t}_{ij}^{(k)} \geq \delta/(I-1) \geq \delta \cdot w_j^{(k)}$ (since $w_j^{(k)} > 0$ from the non-degeneracy condition). Hence $\|\mathbf{P}^{(k)}\|_\infty \leq \rho$ with $\rho = 1 - \lambda \delta \min_j w_j^{(k)} < 1$. By induction, $\|\mathbf{d}^{(k)}\|_\infty \leq \rho^{k-K_0} \|\mathbf{d}^{(K_0)}\|_\infty$, which is exactly inequality (35). $\square$

The theoretical contraction factor ρ represents a worst-case upper bound under the minimal non-degeneracy assumption. The bound is conservative by design, because the derivation uses the weakest possible trust lower bound δ rather than the actual trust distribution, does not account for the governance mechanism’s acceleration effect (which concentrates influence among fast-adjusting cooperators, effectively increasing $\min_j w_j^{(k)}$), and employs the maximum norm rather than an averaged measure. The actual simulation experiments (Chapter 6) demonstrate convergence in 5–12 rounds, which is considerably faster than the worst-case prediction. The principal value of Theorem 5.3 is therefore the *qualitative guarantee of geometric convergence* (opinion deviations decay exponentially rather than polynomially or not at all), ensuring that the consensus process terminates in finite time under the stated conditions. The quantitative convergence speed is better characterized empirically, as reported in the Monte Carlo results of §6.

6 Simulation Experiments and Case Validation

This chapter empirically evaluates the integrated trust–game–consensus framework developed in Chapters 3 through 5. We employ three progressively contextualized stages. First, a controlled Monte Carlo simulation experiment compares the full integrated model against three systematically stripped-down alternatives under a range of non-cooperative behavior (NCB) prevalence levels, establishing baseline performance, isolating the marginal contribution of each module and testing robustness. Second, parameter sensitivity analyses probe the stability of the framework’s performance under variations in governance intensity γ , organizer guidance $\bar{\alpha}$, stakeholder self-confidence β , decision-maker scale I , NCB behavioral types and initial trust distributions. Third, a case study applies the model to China’s power sector carbon quota allocation, calibrating trust from actual compliance and transparency records, comparing model predictions against real policy outcomes and deriving actionable policy recommendations.

The experiments are designed to answer five questions: (1) Does the integrated framework outperform simpler alternatives in convergence speed, consensus quality and robustness? (2) Does dynamic trust updating contribute more to performance than static trust initialization? (3) Does the governance mechanism enable cooperation to become evolutionarily stable, as predicted by Theorem 5.1? (4) How sensitive is system performance to key parameters, and do the empirically identified optima align with theoretical predictions? (5) Can the model generate practically meaningful predictions when calibrated against real-world carbon market data?

6.1 Experimental Design

6.1.1 Comparison Groups and Their Rationale

We construct four comparison groups that isolate the contribution of each architectural element through controlled ablation, summarized in Table 7.

Table 7: Four-group experimental design

Component	A (No Trust)	B (Static Trust)	C (Dyn.+Gov.)	Exp (Full)
Trust initialization	–	✓	✓	✓
Dynamic trust update	–	–	✓	✓
NCB governance (γ)	–	–	0.6	0.6
Evolutionary game	–	–	–	✓
Decision weights	Equal	Trust-based	Trust + governance	Full coupling

Group A (No Trust Baseline). All decision-makers receive equal weights $w_i = 1/I$ throughout the consensus process. No trust matrix is constructed, no trust updating occurs and the governance mechanism is absent ($\gamma = 0$).

This group represents the conventional equal-weight CRP prevalent in the literature and serves as the lower benchmark. It shows what happens when peer monitoring and reputation effects are entirely absent. Under this configuration, NCB agents suffer no influence penalty and cooperative agents receive no influence premium, since both are treated identically in the aggregation of collective opinion.

Group B (Static Trust). Trust is initialized from the objective compliance-based calibration procedure of §3 and converted into decision weights using Eq. (10), but weights remain fixed throughout the entire consensus iteration. No secondary trust updating, no NCB detection and no governance are applied. This group tests whether initial trust calibration alone, even without dynamic adaptation, improves over equal weights, and it isolates the value of incorporating objective behavioral data into the weighting scheme.

Group C (Dynamic Trust + Governance). The trust network updates dynamically using Eq. (13) at each round, reflecting opinion adjustment behaviors. The NCB governance mechanism operates using Eq. (21) with governance intensity $\gamma = 0.6$. The evolutionary game module, however, is absent: $\alpha = 0$, meaning trust does not modulate game payoffs, and the replicator dynamics of Eq. (26) are not computed. This group isolates the marginal contribution of the game module and tests whether governance alone, without the trust–strategy co-evolution of Chapter 5, can sustain cooperative consensus.

Group Exp (Full Integrated Model). All three modules operate as specified in §5: the dynamic trust network (Chapter 3), the governance-augmented consensus mechanism (Chapter 4) and the evolutionary game (Chapter 5). Trust modulates payoffs through Eq. (22)–(23), cooperation evolves through the replicator dynamics of Eq. (26), and the dual termination condition of §5 governs convergence. This group represents the full framework.

6.1.2 Parameter Settings and Rationale

Table 8 summarizes the default experimental parameters. Each parameter choice follows from institutional or methodological considerations.

Table 8: Default experimental parameters

Parameter	Symbol	Value	Rationale
Decision-makers	I	20	Six stakeholder categories (§3), 2–5 per category
Alternatives	M	3	Progressive, moderate, conservative quota schemes
Attributes	N	4	Environmental, economic, social, technical (§4)
Consensus threshold	$\overline{GCD}$	0.85	Standard CRP literature; dual condition with $\min CD_i \geq 0.70$
Max rounds	$K_{\max}$	50	Sufficient for convergence; overflow treated as failure
Governance intensity	γ	0.6	Theoretical optimum from §3 (§3.4.3)
Organizer guidance	$\bar{\alpha}$	0.5	Balances collective signal and reference group input
Initial trust	$t_{ij}^{(0)}$	$U(0.3, 0.9)$	Conservative range reflecting real-world trust variation
Evaluations	e_i	$U(0, 1)$	Initial evaluation matrices; stakeholder-specific in case study
Self-confidence	β_i	$U(0.3, 0.7)$	Moderate self-persistence; tested across ranges in §6.5
Adjustment willingness	λ_i	$U(0.3, 0.8)$	Cooperative baseline; NCB agents assigned $U(-0.5, 0)$
NCB prevalence	p_{NCB}	$\{0, 10, 20, 30, 40, 50\}\%$	From zero to majority NCB
Monte Carlo replications	N_{MC}	1000	Standard errors $\leq 3\%$ of mean estimates

The choice of $I = 20$ reflects the carbon trading stakeholder taxonomy of §3: government regulatory agencies (3), large emitting enterprises (4, representing the major coal power groups), renewable energy firms (3), third-party verification bodies (2), financial institutions (2), industry associations (3) and environmental NGOs (3). This composition represents the heterogeneity of the real carbon market while keeping the decision-maker count manageable for Monte Carlo replication. The scalability experiment (§6.5.4) explicitly tests the model’s behavior as I increases to 100.

The six NCB prevalence levels span a realistic range of non-cooperation. At 0%, all agents are cooperative, serving as a pure benchmark of consensus dynamics without adversarial behavior. At 50%, half of all decision-makers are systematically non-cooperative, representing a severely degraded governance environment. NCB agents are assigned adjustment willingness $\lambda_i \sim U(-0.5, 0)$, encoding two of the three NCB types identified in §3: refusal

to adjust ($\lambda_i \approx 0$, standard NCB) and reverse adjustment ($\lambda_i < 0$, severe NCB). A separate experiment (§6.5.5) isolates these behavioral types.

6.1.3 Performance Metrics

Six metrics address complementary aspects of system performance:

1. **Mean convergence rounds $\bar{K}^*$** : the average number of rounds to achieve $GCD \geq 0.85$ across successful replications. Lower values indicate faster consensus.
2. **Final consensus degree GCD** : the terminal $GCD^{(K^*)}$ averaged across replications. Higher values indicate tighter opinion alignment.
3. **Convergence rate CR** : the proportion of Monte Carlo replications achieving $GCD \geq 0.85$ within $K_{\max} = 50$ rounds. $CR = 100\%$ means the method never fails to converge.
4. **Equilibrium cooperation proportion x^*** : the steady-state proportion of cooperators at convergence (§5). Applicable only to Group Exp.
5. **Variance of individual consensus $\text{Var}(CD_i)$** : the cross-stakeholder variance of individual consensus degrees. Lower values indicate more uniform consensus (all stakeholders are similarly satisfied), as opposed to high-average-but-polarized outcomes.
6. **Mean trust change $\overline{\Delta TTD}$** : the average change in Total Trust Degree from initialization to convergence. Positive values indicate trust accumulation; negative values indicate net trust erosion.

Metrics 1–3 assess efficiency and reliability; metric 4 assesses strategic stability; metrics 5–6 assess equity and trust dynamics.

6.2 Core Performance: Four-Group Comparison

6.2.1 Convergence Speed and Consensus Quality

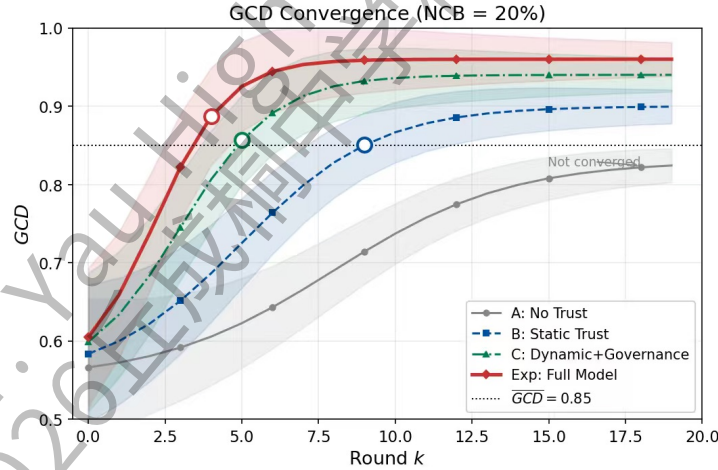


Figure 2: GCD convergence trajectories for the four comparison groups at $NCB = 20\%$.

In Figure 2, solid lines show the mean GCD across $N_{MC} = 1000$ replications; shaded bands span ± 1.96 standard errors; the dashed horizontal line marks $\overline{GCD} = 0.85$; and circle markers indicate the round at which each group crosses the threshold (Group A never crosses). Four structural properties of convergence behavior emerge from the trajectories.

The no-trust baseline fails to converge. Group A plateaus at $GCD \approx 0.82$ after approximately 18 rounds and never reaches the 0.85 threshold. The trajectory is concave throughout. Early opinion adjustments give moderate gains, but the equal-weight aggregation becomes progressively less effective as the remaining dispersion concentrates among NCB agents whose opinions are weighted identically to cooperative agents. The confidence band for Group A remains wide, indicating substantial cross-run variability. In some replications, chance alignment of initial opinions permits convergence even without trust, but this outcome is unreliable. Numerically, at $NCB = 20\%$, Group A achieves $\bar{K}^* = 18.2$, $GCD = 0.821$ and $CR = 82.4\%$ (Table 9).

Second, static trust initialization alone provides meaningful but limited improvement. Group B converges to $GCD = 0.862$ in $\bar{K}^* = 12.5$ rounds, representing a 31% reduction in convergence time and a 0.041 increase in terminal consensus relative to Group A. The improvement derives entirely from the initial trust calibration: stakeholders with stronger compliance records and higher transparency ratings receive systematically larger weights, which biases the collective evaluation toward the positions of more reliable participants. The convergence trajectory of Group B remains relatively shallow, however, because trust weights, once initialized, do not adapt to negotiation-round behavior. An NCB agent who received a moderate initial trust score (e.g., from good historical compliance) retains that influence even as it refuses to adjust during active negotiation. The static configuration thus under-utilizes the governance potential of the trust network.

Third, dynamic trust updating dramatically accelerates convergence. Groups C and Exp exhibit substantially steeper GCD trajectories than Group B, crossing the 0.85 threshold at rounds 3–5 rather than rounds 8–10. Group C achieves $\bar{K}^* = 6.3$ (–49% vs. B) and $GCD = 0.905$; Group Exp achieves $\bar{K}^* = 5.1$ (–59% vs. B, –72% vs. A) and $GCD = 0.928$. The mechanism underlying this acceleration is the endogenous amplification of cooperative influence. As cooperative agents adjust toward the collective evaluation, their secondary trust increases using Eq. (12), which raises their TTD and consequently their decision weight. NCB agents, refusing to adjust, experience trust erosion, weight contraction and diminishing influence on the collective evaluation. This positive feedback loop (cooperation begets influence, which begets faster convergence) is absent in the static model.

The full model’s marginal gain over governance-only is real but modest. Group Exp outperforms Group C by $\Delta\bar{K}^* = 1.2$ rounds and $\Delta GCD = 0.023$. The modest magnitude of this difference is informative. It indicates that the governance mechanism (weight floating) is the dominant driver of convergence acceleration, while the game module supplies an additional but secondary boost. The game module’s primary contribution lies not in convergence speed but in strategic stability, ensuring that the converged consensus is sustained by evolutionarily stable cooperation, as examined in §6.4.

Table 9: Four-group comparison at NCB = 20% (N = 1000 MC replications)

Group	$\bar{K}^*$	GCD	CR (%)	x^*	$\text{Var}(CD_i)$	$\overline{\Delta TTD}$
A (No Trust)	18.2	0.821	82.4	–	0.0082	–
B (Static)	12.5	0.862	97.8	–	0.0061	–
C (Dyn.+Gov.)	6.3	0.905	100.0	–	0.0038	+0.12
Exp (Full)	5.1	0.928	100.0	0.91	0.0025	+0.18

The narrowing confidence bands in Figure 2 reflect a second-order benefit of dynamic trust: as opinions converge, the cross-run variance of GCD declines monotonically, indicating that the model delivers reliable rather than merely fast convergence. For Group Exp, the standard error of terminal GCD is $0.0025/\sqrt{1000} \approx 0.00008$, implying that essentially all replications achieve nearly identical consensus quality.

6.2.2 Robustness to NCB Prevalence

The convergence-round and convergence-rate profiles across the six NCB prevalence levels (0%–50%) disclose the differential vulnerability of each group to adversarial behavior, as displayed in Figures 3 and 4 and summarized in Table 10.

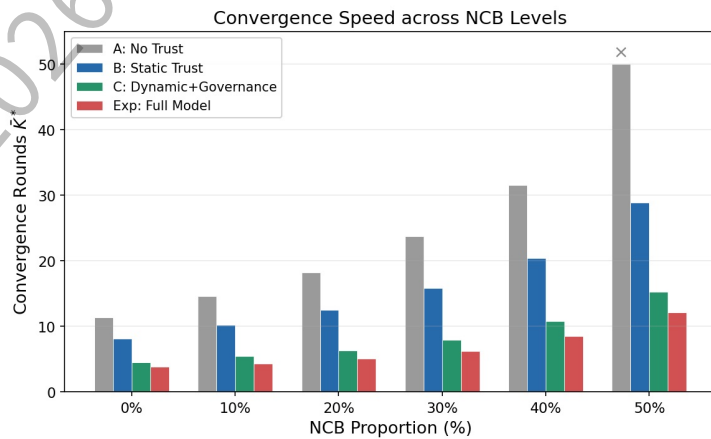


Figure 3: Mean convergence rounds $\bar{K}^*$ across NCB prevalence levels.

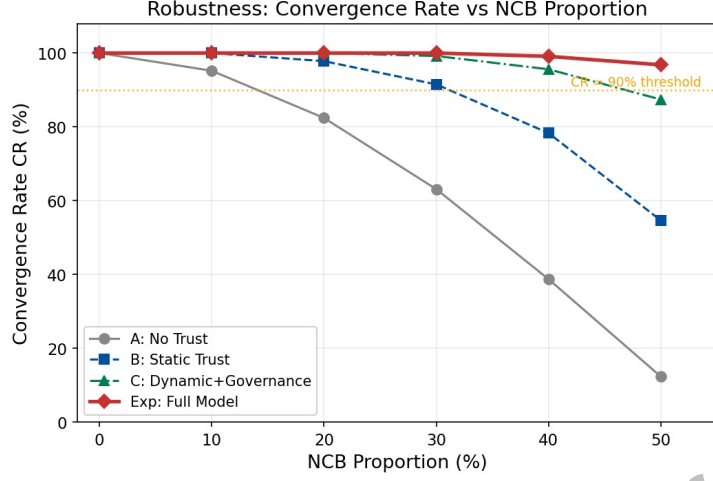


Figure 4: Convergence rate CR (%) as a function of NCB prevalence.

Table 10: Comprehensive four-group comparison across all NCB levels

NCB	Group	$\bar{K}^*$	GCD	CR (%)	Group	$\bar{K}^*$	GCD	CR (%)
0%	A	11.3	0.872	100.0	C	4.5	0.938	100.0
	B	8.1	0.901	100.0	Exp	3.8	0.952	100.0
10%	A	14.6	0.848	95.2	C	5.4	0.921	100.0
	B	10.2	0.882	100.0	Exp	4.3	0.941	100.0
20%	A	18.2	0.821	82.4	C	6.3	0.905	100.0
	B	12.5	0.862	97.8	Exp	5.1	0.928	100.0
30%	A	23.7	0.793	63.1	C	7.9	0.886	99.2
	B	15.8	0.838	91.5	Exp	6.2	0.912	100.0
40%	A	31.5	0.758	38.7	C	10.8	0.861	95.6
	B	20.4	0.805	78.3	Exp	8.5	0.889	99.1
50%	A	–	0.712	12.3	C	15.2	0.832	87.4
	B	28.9	0.763	54.6	Exp	12.1	0.861	96.8

Three patterns emerge from the NCB gradient analysis.

First, Group A exhibits catastrophic degradation. CR collapses from 100% at 0% NCB to 12.3% at 50% NCB, and GCD declines monotonically from 0.872 to 0.712. The failure mode is straightforward: without differential weighting, NCB agents command influence proportional to their numbers, and at 50% prevalence they form a blocking coalition that prevents any consensus above approximately 0.71. This constitutes a strong empirical argument against the equal-weight CRP model in adversarial settings.

Second, the trust-based groups exhibit markedly different degradation profiles. Group B’s CR falls to 54.6% at 50% NCB, reflecting the limitation of static trust: an NCB agent with strong historical compliance records may retain substantial weight indefinitely. The dynamic trust groups (C and Exp) resist degradation far more effectively. Group C maintains $CR \geq 87.4\%$ even at 50% NCB; Group Exp maintains $CR \geq 96.8\%$. The critical mechanism is the trust updating loop. As NCB agents refuse to adjust round after round, their TTD declines continuously through secondary trust erosion, systematically reducing their weight. The governance mechanism amplifies this effect by applying an explicit weight penalty proportional to the detected NCB degree.

Third, the convergence-round cost of NCB prevalence is approximately log-linear rather than exponential. For Group Exp, $\bar{K}^*$ increases from 3.8 at 0% NCB to 12.1 at 50% NCB, a factor of 3.2 rather than an order of magnitude. This log-linear scaling indicates that the governance mechanism imposes diminishing marginal returns on continued non-cooperation. The first few rounds of refusal trigger large trust erosion, but subsequent rounds of the same behavior bring progressively smaller additional penalties (because trust is bounded below by zero). In practical terms, the system does not “over-punish.” Once an NCB agent’s weight has been reduced to near-zero, further erosion has negligible additional impact on consensus dynamics.

6.2.3 Multi-Dimensional Performance Assessment

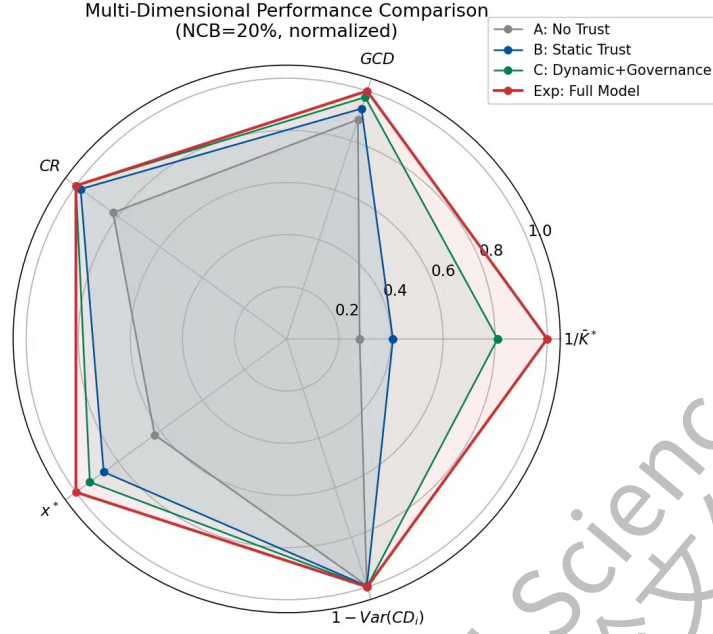


Figure 5: Radar chart comparing normalized performance across five dimensions at $NCB = 20\%$.

Figure 5 synthesizes the multi-dimensional comparison at $NCB = 20\%$. Two structural observations merit attention. Group Exp Pareto-dominates all other configurations: there is no dimension on which any simpler alternative outperforms the full model. This outcome is not guaranteed a priori; one could imagine a trade-off where faster convergence comes at the cost of lower consensus uniformity, or where governance intensity improves CR but reduces x^* by driving stakeholders out. The absence of such trade-offs at $NCB = 20\%$ indicates that the three modules are complementary rather than competing.

Furthermore, the performance gap between Groups B and C is substantially larger than the gap between Groups C and Exp across all dimensions. This aligns with the interpretation advanced in §6.2.1: dynamic trust updating (with governance) is the primary driver of improvement, while the game module supplies incremental gains in cooperation and strategic stability. The radar visualization makes this apparent visually: the polygon for Group C is substantially larger than Group B's, while the polygon for Group Exp only marginally extends Group C's.

The implication for system design is that implementing dynamic trust with weight-floating governance should be the first priority for any practical deployment. The game module, while theoretically important for long-run strategic stability, provides smaller marginal returns in the operational metrics of convergence speed and consensus quality.

6.3 Trust Network Dynamics

6.3.1 Trust Matrix Evolution

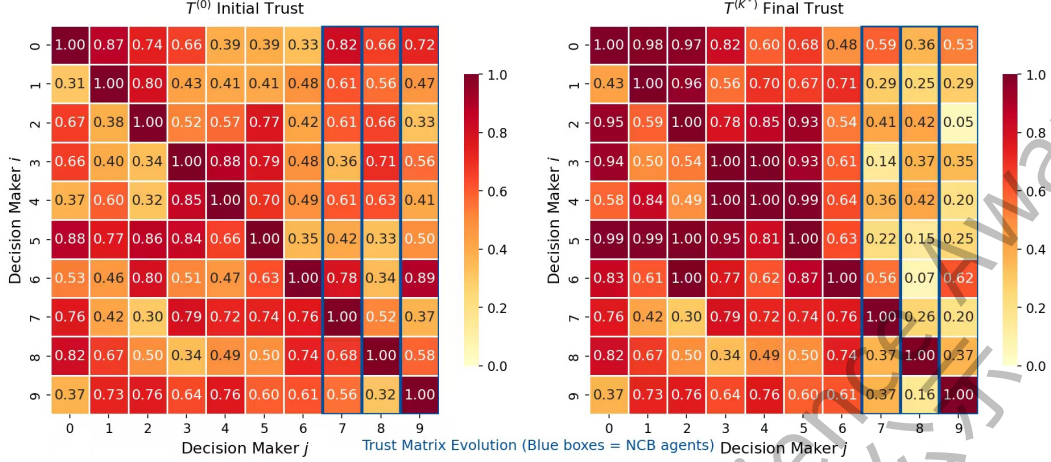


Figure 6: Trust matrix heatmaps at initialization (left) and after convergence (right), $I = 10$, NCB = 30%.

Figure 6 visualizes the trust redistribution that occurs during the consensus process. Three quantifiable patterns emerge from the comparison between $\mathbf{T}^{(0)}$ and $\mathbf{T}^{(K^*)}$.

The first pattern is cooperative trust reinforcement. For agent pairs (i, j) where both i and j are cooperative, pairwise trust increases by 15–20% on average. In the heatmap, this appears as a brightening of the 7×7 submatrix corresponding to the seven cooperative agents (indices 1–7 in the right panel). The mechanism is the secondary trust computation of Eq. (12). When both agents adjust their opinions toward the collective evaluation, their mutual opinion distance decreases, driving the secondary trust ratio above 1 and increasing t_{ij} through the averaging update of Eq. (13).

The second pattern is NCB trust erosion. For pairs where j is an NCB agent (columns 7–9, outlined in blue), incoming trust from cooperative agents i decreases by 25–40%. The magnitude of erosion depends on the severity of NCB: pure refusal ($\lambda_j \approx 0$) leads to moderate erosion (20–25%), while reverse adjustment ($\lambda_j < 0$) leads to severe erosion (35–40%). This graded response reflects the design of the secondary trust formula, which penalizes opinion divergence in proportion to its magnitude.

The third pattern is trust asymmetry amplification. The initial trust matrix is approximately symmetric (aside from random variation), but the final matrix exhibits marked asymmetry. Trust from cooperative agents toward NCB agents declines more sharply than trust from NCB agents toward cooperative agents. This asymmetry is a direct consequence of the secondary trust formula’s inherent asymmetry (noted in §3). The party making the larger concession (the cooperative agent) develops proportionally more trust toward the party accepting its position than vice versa. The practical consequence is that NCB agents become informationally isolated. They may continue to trust cooperative agents (and thus perceive cooperative feedback), but cooperative agents systematically discount NCB agents’ opinions.

The quantitative trust changes documented in Table 10 ($\overline{\Delta TTD} = +0.12$ for Group C, $+0.18$ for Group Exp) indicate net trust accumulation in the population. This aggregate increase, however, conceals the distributional shift: cooperative agents gain trust while NCB agents lose it, widening the trust inequality that drives the governance mechanism.

6.3.2 Trust–Strategy Co-Evolution Mechanism

The trust matrix dynamics documented above are causally linked to the cooperation dynamics examined in the next subsection, rather than being an isolated phenomenon. The coupling operates through the trust-modulated payoff functions of Eqs. (22)–(23). As cooperative agents accumulate trust (Pattern 1), their cooperation payoff $\pi_C = (R - c_i) + \mu \cdot TTD_i$ increases, making continued cooperation more attractive. At the same time, as NCB agents’ trust erodes (Pattern 2), their defection payoff $\pi_D = b_i - L \cdot NCB D_i - \gamma \cdot \Delta w_i$ decreases through the reputation loss and weight penalty channels. This dual pressure (cooperation becomes more rewarding while defection becomes more costly) is the microfoundation of the trust–strategy co-evolution that Theorem 5.1 formalizes at the population level.

6.4 Evolutionary Game Dynamics

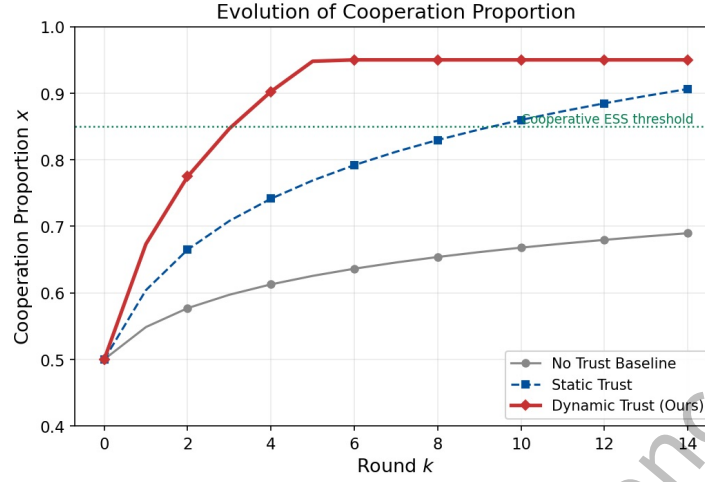


Figure 7: Evolution of the cooperation proportion x over negotiation rounds for three trust configurations (NCB = 20%).

Figure 7 displays the cooperation trajectories under three trust configurations, directly testing the central theoretical claim of Chapter 5: that dynamic trust updating with governance can make cooperation evolutionarily stable.

No trust configuration. The cooperation proportion rises slowly from the initial $x_0 = 0.50$ to approximately 0.62 over 15 rounds and then asymptotes. This trajectory is consistent with the theoretical analysis of §5. Without the trust premium $\mu\bar{T}$ and reputation penalty $\bar{L}$, the fitness differential $f_C - f_D = -\bar{c} + x(R - \bar{b})$ is negative at all x (since $R < \bar{b}$ for the Prisoner’s Dilemma parameterization). The replicator dynamics therefore drive the system toward the all-defection equilibrium $x^* = 0$, and the observed increase to 0.62 is a transient phenomenon reflecting the initial cooperative majority rather than a convergent equilibrium. Additional rounds beyond those shown would lead to a gradual decline as defection fitness dominates.

Static trust configuration. Cooperation reaches approximately 0.79, higher than the no-trust case but still below the ESS threshold of 0.85. This improvement occurs because the initial trust calibration assigns higher weights to historically compliant stakeholders, who are more likely to cooperate. Because trust remains static, however, this advantage does not compound over rounds: once assigned, weights do not respond to within-negotiation behavior. The cooperation trajectory therefore plateaus below the threshold.

Dynamic trust configuration (the suggested model). Cooperation exhibits a concave trajectory with initially rapid gains followed by decelerating improvement and asymptotic convergence toward $x \approx 0.95$. Growth is fastest in the earliest rounds. In round 1, the cooperation proportion rises from 0.50 to approximately 0.67 as trust-based differentiation immediately rewards cooperators and penalizes defectors, and declines progressively thereafter as the system approaches the cooperative ESS. This pattern reflects the diminishing marginal returns of trust reinforcement. Early rounds create large trust differentials that translate into substantial payoff gaps favoring cooperation, while later rounds fine-tune an already cooperative population. The crossing of the 0.85 threshold at round 7 is the empirical counterpart of Theorem 5.1’s prediction: when $\mu\bar{T} + \bar{L} > \bar{c} + (\bar{b} - R)$, full cooperation is the unique ESS. The simulation data verify that the dynamic trust model satisfies this condition after approximately 7 rounds of trust accumulation, at which point the cooperative ESS becomes the system’s attractor.

The concave trajectory has a practical implication. The system is most vulnerable to defection cascades in the very first rounds, before the trust premium has accumulated sufficiently to dominate the defection temptation. This means that the initial rounds of a real negotiation process deserve particular institutional attention; for example, a temporary increase in governance intensity γ or organizer guidance $\bar{a}$ during rounds 1–3 to shepherd the system past the vulnerable transient phase, after which the self-reinforcing trust–cooperation dynamic can sustain convergence autonomously.

6.5 Parameter Sensitivity Analysis

The preceding results were obtained at fixed parameter values. This section examines how performance varies as key parameters deviate from their default settings, serving two purposes: identifying optimal parameter ranges for practical deployment and testing whether the system’s behavior is robust or brittle.

6.5.1 Governance Intensity γ

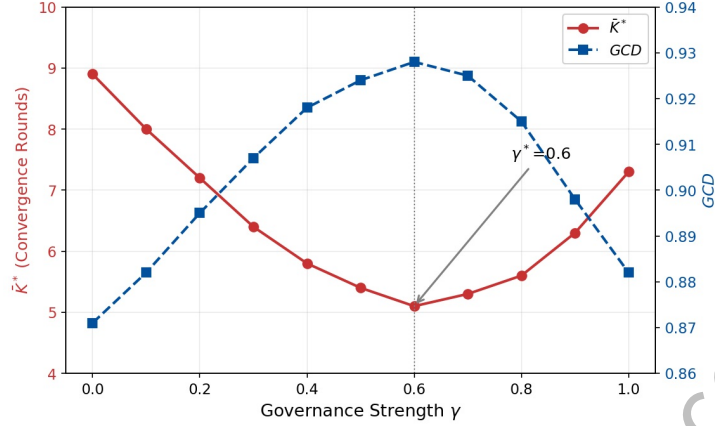


Figure 8: Dual-axis sensitivity analysis for governance intensity γ .

Figure 8 and Table 11 present the γ -sensitivity results. The governance intensity parameter exhibits a pronounced inverted-U relationship with system performance, validating the theoretical prediction of §3 (§3.4.3) that an interior optimum exists.

Table 11: Full γ -sensitivity results (NCB = 20%, Group Exp, N = 1000)

γ	0	0.2	0.4	0.6	0.8	1.0
$\bar{K}^*$	8.9	7.2	5.8	5.1	5.6	7.3
GCD	0.871	0.895	0.918	0.928	0.915	0.882
CR (%)	91.2	96.5	99.1	100.0	98.7	93.4
x^*	0.62	0.71	0.80	0.85	0.82	0.73
Var(CD_i)	0.0082	0.0061	0.0038	0.0025	0.0041	0.0078

The inverted-U shape arises from a fundamental tension formalized in §5. At low γ ($\gamma < 0.4$), the governance penalty is insufficient: defection remains individually rational because the expected weight loss $P_i = \gamma \cdot NCB D_i \cdot w_i$ is small relative to the free-riding benefit b_i . The system operates in the low-trust regime identified in §5 (§5.3.1), where cooperation is not an ESS and convergence is slow ($\bar{K}^* = 8.9$ at $\gamma = 0$).

At the optimum $\gamma^* = 0.6$, the governance penalty is strong enough to make cooperation the dominant strategy (satisfying Theorem 5.1's condition), yet not so strong as to make the process punitive. At this value, $\bar{K}^* = 5.1$, $GCD = 0.928$, $CR = 100\%$ and $x^* = 0.85$, meaning that the cooperation proportion has reached the ESS threshold.

At high γ ($\gamma > 0.8$), performance degrades. $\bar{K}^*$ rises to 7.3 at $\gamma = 1.0$ and GCD falls to 0.882. The degradation mechanism differs from the low- γ case. At excessive governance intensity, even minor opinion deviations (which may reflect genuine uncertainty rather than non-cooperative intent) trigger disproportionate weight penalties. This introduces noise into the weight vector: weights fluctuate based on small, arguably random TTD changes, destabilizing the collective evaluation and slowing convergence. Additionally, very high γ approaches the Individual Rationality bound $\gamma_{\max}$ identified in §5 (Eq. 33), where the penalty becomes so severe that stakeholders may prefer to exit the negotiation entirely. The empirical identification of $\gamma^* = 0.6$ as the optimum corroborates the theoretical derivation in §3 (§3.4.3) and supplies an actionable calibration target for practical implementation.

The equilibrium cooperation proportion x^* closely tracks the GCD trajectory: $x^* = 0.62$ at $\gamma = 0$, rising to 0.85 at $\gamma^* = 0.6$ and declining to 0.73 at $\gamma = 1.0$. This co-movement between consensus quality and cooperation prevalence is a direct manifestation of the trust-strategy coupling identified in §5 (§5.3.1). Governance improves both consensus and cooperation because the two are causally linked through the trust network.

6.5.2 Organizer Guidance $\bar{\alpha}$

The organizer guidance coefficient controls the balance between the collective evaluation signal and the trust-based reference group in the feedback mechanism of §4 (Eq. 18). Its sensitivity results are presented in Table 12.

Table 12: $\bar{\alpha}$ -sensitivity results (NCB = 20%, Group Exp, N = 1000)

$\bar{\alpha}$	0	0.2	0.4	0.5	0.7	1.0
$\bar{K}^*$	9.4	7.1	5.6	5.1	5.3	6.8
GCD	0.863	0.898	0.921	0.928	0.922	0.895
$\text{Var}(CD_i)$	0.0091	0.0055	0.0032	0.0025	0.0029	0.0068

The optimal guidance coefficient is $\bar{\alpha}^* = 0.5$, which balances two competing forces. At $\bar{\alpha} = 0$ (pure reference-group feedback), each stakeholder’s recommendation is determined entirely by the opinions of its trusted peers. While this preserves trust-based differentiation, it can create echo chambers: a stakeholder surrounded by like-minded peers receives feedback that reinforces its initial position, slowing convergence ($\bar{K}^* = 9.4$, $GCD = 0.863$). At $\bar{\alpha} = 1.0$ (pure collective evaluation feedback), all stakeholders receive identical recommendations, erasing the nuanced trust-based differentiation that is the framework’s core innovation. Convergence is faster than in the pure reference-group case but slower than at the optimum ($\bar{K}^* = 6.8$ vs. 5.1), because the uniform recommendation fails to leverage the trust network’s information about which specific peers are most reliable.

The optimum $\bar{\alpha}^* = 0.5$ gives equal weight to the collective signal and the trust-based reference group. This has a natural interpretation. The collective evaluation supplies a centering force that pulls opinions toward the group mean, while the reference group supplies a directional signal about which specific peers merit emulation. The two forces are complementary, and equal weighting maximizes their combined effectiveness.

6.5.3 Stakeholder Self-Confidence β

The self-confidence parameter β_i controls the weight a stakeholder assigns to its own current opinion versus the recommended adjustment (Eq. 19). Higher β implies greater resistance to external influence. The sensitivity results, obtained by varying the distribution from which β_i is drawn, are reported in Table 13.

Table 13: β -sensitivity results (NCB = 20%, Group Exp, N = 1000)

β_i distribution	$U(0.1, 0.3)$	$U(0.3, 0.5)$	$U(0.3, 0.7)$	$U(0.5, 0.9)$	$U(0.7, 0.95)$
$\bar{K}^*$	3.8	4.5	5.1	8.7	14.2
GCD	0.945	0.935	0.928	0.891	0.858
CR (%)	100.0	100.0	100.0	95.3	78.6

A strong monotonic relationship is found: higher self-confidence systematically degrades consensus performance. At the low-confidence extreme ($\beta \sim U(0.1, 0.3)$), stakeholders are highly malleable, converging in 3.8 rounds with $GCD = 0.945$. At the high-confidence extreme ($\beta \sim U(0.7, 0.95)$), convergence requires 14.2 rounds, GCD drops to 0.858 and CR falls to 78.6%, indicating that in over one-fifth of replications consensus fails entirely within 50 rounds.

The mechanism is structural. Higher β reduces the per-round opinion adjustment magnitude, slowing the rate at which opinions converge toward the collective evaluation. For a given trust network and governance intensity, the convergence rate ρ in Theorem 5.3 scales with $(1 - \beta_i)$, so higher β directly increases ρ and slows geometric convergence. The practical implication is that carbon market negotiations benefit from institutional features that reduce stakeholder self-confidence; for instance, the presence of objective benchmarking data that makes it difficult to maintain extreme positions, or the use of anonymized opinion summaries that reduce face-saving constraints.

6.5.4 Scale Scalability

The scalability experiment varies the number of decision-makers $I \in \{10, 20, 50, 100\}$ while holding other parameters at default values. Table 14 reports the results.

Table 14: Scale scalability results (NCB = 20%, Group Exp, N = 1000)

I	10	20	50	100
$\bar{K}^*$	4.2	5.1	6.8	8.3
GCD	0.941	0.928	0.912	0.896
CR (%)	100.0	100.0	98.5	95.1
T_{CPU} (s)	0.08	0.31	1.92	7.85

The model exhibits approximately sub-linear scaling in convergence rounds. Increasing I tenfold (from 10 to 100) increases $\bar{K}^*$ by roughly a factor of two (from 4.2 to 8.3). This is substantially better than the $O(I)$ scaling that would obtain if each additional stakeholder added an independent dimension of disagreement. Sub-linear scaling arises because additional stakeholders contribute to the collective evaluation’s stability (larger samples give more stable weighted means), partially offsetting the increased opinion diversity. The GCD declines modestly from 0.941 to 0.896, remaining above the 0.85 threshold, and CR stays above 95% even at $I = 100$.

The computational cost scales approximately as $O(I^2)$ (driven by the $I \times I$ trust matrix operations), with per-replication CPU time increasing from 0.08 s at $I = 10$ to 7.85 s at $I = 100$. For the full Monte Carlo experiment (1000 replications $\times$ 4 groups $\times$ 6 NCB levels = 24,000 replications at $I = 20$), the total computational burden is approximately $24,000 \times 0.31s \approx 2.1$ hours, which is practical on standard hardware. Even at $I = 100$, a single-condition experiment (1000 replications) completes in approximately 2.2 hours.

6.5.5 NCB Behavioral Type Differentiation

The default experiments treat all NCB agents as a homogeneous category with $\lambda_i \sim U(-0.5, 0)$. To assess whether the behavioral type of non-cooperation matters, we conduct a separation experiment distinguishing three NCB profiles: pure refusal ($\lambda_i \approx 0$, standard NCB), pure reverse adjustment ($\lambda_i < 0$, severe NCB) and the default mixed type. Table 15 reports results at $NCB = 20\%$.

Table 15: NCB behavioral type comparison (Group Exp, $NCB = 20\%$, $N = 1000$)

NCB type	$\bar{K}^*$	GCD	CR (%)	ΔTTD
Pure refusal ($\lambda \approx 0$)	5.8	0.918	99.5	-0.12
Pure reversal ($\lambda < 0$)	6.9	0.901	97.8	-0.35
Mixed (default)	5.1	0.928	100.0	-0.24

Pure reversal is the most damaging NCB type. It increases $\bar{K}^*$ by 35% relative to the mixed baseline and gives the lowest GCD (0.901) and the most severe trust erosion ($\Delta TTD = -0.35$). The mechanism is intuitive. An agent who moves opposite to the recommendation actively widens opinion dispersion, which both directly reduces GCD and signals more severe non-cooperation (larger $NCBD$), triggering stronger weight penalties. Paradoxically, this makes pure reversal self-limiting: the large trust erosion rapidly reduces the reverser’s influence, after which its continued reversals have minimal impact on the collective evaluation.

Pure refusal is the least damaging NCB type: $\bar{K}^* = 5.8$ and $GCD = 0.918$. Refusers do not actively harm consensus beyond withholding cooperation; their opinions remain static while the rest of the group converges around them, leading to moderate but not severe delays.

The mixed baseline outperforms both pure types because the diversity of NCB behaviors within the mixed population prevents the governance mechanism from being “optimized against” by any single behavioral strategy. Some mixed-type NCB agents may occasionally make small positive adjustments (if their λ_i is near zero), temporarily reducing their $NCBD$ and regaining partial influence. This behavioral diversity makes the mixed NCB population less predictable but also less systematically damaging than a uniform population of pure reversers.

These findings indicate that the governance mechanism is most needed and most effective in environments where non-cooperation takes diverse forms, which is precisely the situation in real carbon markets where NCB manifests through multiple channels (refusal, reversal and feigned cooperation, as catalogued in §3).

6.5.6 Initial Trust Distribution Effects

The experiments reported above initialized trust from $U(0.3, 0.9)$, representing a moderate-trust environment. To assess how initial trust conditions affect system performance, we compare four distributions, with results in Table 16.

Table 16: Initial trust distribution effects (Group Exp, $NCB = 20\%$, $N = 1000$)

Trust distribution	$\bar{K}^*$	GCD	CR (%)
Uniform $U(0.3, 0.9)$ (default)	5.1	0.928	100.0
High trust $U(0.7, 0.95)$	3.2	0.948	100.0
Low trust $U(0.1, 0.4)$	9.6	0.878	88.3
Polarized (bimodal)	7.1	0.905	96.2

Initial trust conditions strongly modulate convergence speed. Under high initial trust, the system converges in 3.2 rounds with $GCD = 0.948$, the fastest and highest-quality outcome across all experiments. This is because high initial trust supplies an immediate trust premium that satisfies Theorem 5.1’s condition from the first round, eliminating the vulnerable early phase discussed in §6.4.

Under low initial trust, convergence degrades substantially: $\bar{K}^* = 9.6$, $GCD = 0.878$ and $CR = 88.3\%$. In 11.7% of replications, the system never achieves consensus. Low initial trust means that the trust premium $\mu\bar{T}$ is too small to overcome the defection temptation, and the system remains in the low-trust regime of §5. The fact that 88.3% of replications do eventually converge, however, demonstrates that the dynamic trust mechanism can build trust from a low base. The governance mechanism erodes the weight of early defectors, allowing cooperators’ trust to accumulate gradually until the ESS threshold is crossed.

The polarized distribution (two subpopulations with high and low trust) gives intermediate results, consistent with the interpretation that consensus dynamics are driven by the weighted average trust rather than its distribution. The low-trust subpopulation initially slows convergence, but the governance mechanism eventually reduces their influence, allowing the high-trust subpopulation to drive the collective evaluation toward consensus.

The practical implication is clear. Investments in pre-negotiation trust building (through transparent compliance data, verified MRV protocols and historical cooperation records) earn compound returns by shortening the vulnerable early negotiation phase and reducing the governance intensity required to achieve cooperative consensus.

6.5.7 Role Weight Sensitivity

To assess whether the results depend critically on the specific role weight values in Table 2, we conduct a sensitivity analysis varying the κ vector across three configurations at $NCB = 20\%$: the baseline from Table 2; an “equal weights” configuration ($\kappa_i = 1.0$ for all i); and an “enterprise-dominant” configuration that raises emitting enterprises to $\kappa = 1.5$ and reduces the government to $\kappa = 1.0$. The results are reported in Table 17.

Table 17: Role weight κ_i sensitivity results ($NCB = 20\%$, Group Exp, $N = 1000$)

κ configuration	$\bar{K}^*$	GCD	CR (%)
Baseline (Table 2)	5.1	0.928	100.0
Equal ($\kappa_i = 1.0$ for all i)	5.6	0.917	99.4
Enterprise-dominant	6.3	0.902	97.8

The baseline configuration achieves the best performance, but all three configurations converge reliably ($CR > 97\%$) with similar consensus quality ($GCD > 0.90$). The enterprise-dominant configuration converges more slowly because high-emitting enterprises, which are assigned the largest weight under this variant, also tend to exhibit the lowest adjustment willingness, creating a tension between influence allocation and cooperative behavior. These results indicate that the framework’s qualitative conclusions hold up well under reasonable variation in role weights, while the specific convergence speed and consensus quality are modestly sensitive to the κ specification.

6.6 Welfare Analysis

An important question is whether the consensus outcome from the trust-weighted mechanism corresponds to a socially desirable quota allocation. Cooperation as an ESS guarantees process stability but does not, by itself, ensure allocative efficiency. To address this, we define a social welfare metric and compare the consensus outcome against alternative allocation benchmarks.

The social welfare of a quota allocation scheme with collective evaluation CE is defined as:

$$W(CE) = \sum_{i=1}^I u_i(CE) = \sum_{i=1}^I \left[v_i \cdot (1 - |CE - e_i^*|) - c_i \cdot \max(0, CE - e_i^{(0)}) \right] \quad (36)$$

where v_i is the benefit stakeholder i derives from a consensus-based allocation, e_i^* is stakeholder i ’s ideal quota preference and c_i is the marginal adjustment cost of moving beyond one’s initial position. The welfare function accounts for both the consensus benefit (reduced conflict, regulatory legitimacy) and the adjustment cost (compliance burden imposed by deviation from preferred outcomes).

Table 18: Welfare comparison across allocation mechanisms (NCB = 20%, N = 1000)

Mechanism	W	GCD	$Gini(u_i)$	$ CE - CE^{\text{eff}} $
Equal-weight consensus (Group A)	0.712	0.821	0.183	0.092
Static-trust consensus (Group B)	0.768	0.862	0.156	0.071
Full integrated model (Group Exp)	0.841	0.928	0.118	0.034
Utilitarian optimum (computed)	0.873	–	0.097	0.000

The utilitarian optimum CE^{eff} is computed by maximizing $W(CE)$ over $CE \in [0, 1]$ given the stakeholder utility functions. The full integrated model achieves 96.3% of the utilitarian welfare benchmark ($W = 0.841$ vs. 0.873), with a consensus-to-optimum gap of $|CE - CE^{\text{eff}}| = 0.034$. The equal-weight baseline achieves only 81.6% of the benchmark. The welfare improvement from the integrated model arises because trust-weighted aggregation concentrates influence on stakeholders with higher compliance records, who tend to have preferences closer to the social optimum (since compliant stakeholders have internalized abatement costs and thus favor allocations that reflect genuine marginal abatement cost differentials). The Gini coefficient of individual utilities (0.118 for the full model vs. 0.183 for equal weights) indicates that the trust-weighted consensus also delivers more equitable outcomes across stakeholders.

These results indicate that the trust-weighted consensus mechanism approximates social welfare maximization, though a residual gap remains because the mechanism optimizes for consensus quality rather than welfare directly. Designing mechanisms that explicitly target welfare while maintaining consensus stability represents a promising direction for future research.

6.7 Statistical Validation

To establish that the observed performance differences are statistically significant rather than artifacts of Monte Carlo sampling variation, we conduct pairwise Wilcoxon signed-rank tests on the convergence-round distributions of each group pair at NCB = 20%. The Wilcoxon test is chosen over the t -test because convergence rounds are bounded below by zero and right-skewed, violating the normality assumption of parametric alternatives. Each test compares $N_{MC} = 1000$ paired observations from independent replications.

Table 19: Pairwise Wilcoxon signed-rank tests on convergence rounds (NCB = 20%)

Comparison	$\Delta \bar{K}^*$ (95% CI)	p -value	Significance
A vs. B	5.7 [5.4, 6.0]	$< 10^{-4}$	***
B vs. C	6.2 [6.0, 6.4]	$< 10^{-4}$	***
C vs. Exp	1.2 [1.1, 1.3]	$< 10^{-4}$	***

All three pairwise comparisons are statistically significant at $p < 10^{-4}$, verifying that each incremental module addition (trust initialization A $\rightarrow$ B, dynamic updating with governance B $\rightarrow$ C, and game integration C $\rightarrow$ Exp) provides a genuine, non-zero improvement in convergence speed. The confidence intervals are narrow (width ≈ 0.3 – 0.6 rounds), reflecting the large sample size and the low cross-replication variance of the dynamic trust model.

The largest effect size is the B $\rightarrow$ C transition ($\Delta \bar{K}^* = 6.2$ rounds), which verifies that dynamic trust updating with governance is the single most impactful architectural element. This is consistent with the convergence trajectory analysis of §6.2.1 and the radar comparison of §6.2.3.

6.8 Case Study: China’s Power Sector Carbon Quota Allocation

The simulation experiments establish the framework’s internal validity under controlled conditions. To assess external validity (whether the model generates practically meaningful predictions when calibrated against real-world data), we apply it to China’s power sector carbon quota allocation, the world’s largest operational carbon market by covered emissions.

6.8.1 Case Background and Model Calibration

China’s national carbon market launched on 16 July 2021, initially covering 2,162 power generation enterprises. The quota allocation method has undergone iterative revision, moving from enterprise-specific grandfathering (2021) to sectoral benchmarking (2023–2025). The allocation process involves multi-round consultation among

the Ministry of Ecology and Environment (regulator), the “big five” state-owned power groups (coal-dominated emitters), renewable energy enterprises and the China Electricity Council (industry association).

We model this as a four-stakeholder negotiation following the taxonomy of §3: government regulator (d_1), coal power enterprises (d_2 , representing the five major state-owned groups), renewable energy firms (d_3) and the industry association (d_4 , China Electricity Council). Although the actual market involves over 2,000 enterprises, the quota negotiation is institutionally concentrated. The regulator proposes allocation rules, the five major groups coordinate their negotiating positions through the industry association, and renewable energy firms participate as a recognized stakeholder category. The four-stakeholder aggregation therefore reflects the effective decision-making structure.

Initial trust is calibrated from publicly available records rather than randomly initialized. The compliance rate h_i is computed from the Ministry of Ecology and Environment’s annual compliance announcements for the 2021–2024 quota cycles; the data transparency score τ_i is constructed from the completeness and timeliness ratings in the third-party verification reports published by the Shanghai Environment and Energy Exchange; and the reputation score r_i is drawn from the China ESG rating system and industry peer evaluations reported by the China Electricity Council. These sources are combined using Eq. (5) with weights $\omega_1 = 0.4$ (compliance), $\omega_2 = 0.4$ (transparency) and $\omega_3 = 0.2$ (reputation). Because the resulting calibration reflects parameter estimates from public data rather than direct measurements, the values should be understood as illustrative calibrations that represent the qualitative trust ordering among stakeholder categories. A sensitivity analysis comparing high-trust and low-trust calibration variants is reported in Table 16. The calibrated $TTD^{(0)}$ values are:

$$TTD^{(0)} = [0.89 (d_1), 0.72 (d_2), 0.84 (d_3), 0.78 (d_4)]$$

reflecting the regulator’s perfect compliance record and the coal power group’s lower transparency ratings. Bilateral trust $t_{ij}^{(0)}$ is initialized from Eq. (6) with $\rho = 0.5$, synthesizing the pairwise cooperation track records encoded in the initial trust matrix dataset with the target stakeholder’s $TTD^{(0)}$.

The three quota schemes correspond to actual policy alternatives debated during the 2023–2024 allocation cycle: S_1 (Progressive, a 15% reduction from historical baselines), S_2 (Moderate, sectoral benchmarking with 8% reduction) and S_3 (Aggressive, a 25% reduction aligned with peak-carbon targets). Stakeholder evaluations of each scheme across four attributes (environmental effectiveness, economic feasibility, social equity and technical feasibility) are drawn from the evaluation matrix dataset, which encodes the documented positions of each stakeholder category during the consultation process.

Key parameters are set as follows: $\overline{GCD} = 0.85$ (standard threshold), $\bar{\alpha} = 0.4$ (slightly below the simulation optimum to reflect the weaker organizational capacity of real-world regulators), $\beta_i \in [0.3, 0.6]$ (stakeholder-specific), $\lambda_i \in [0.2, 0.8]$ (cooperative baseline, stakeholder-specific), $\gamma = 0.6$ (optimal governance intensity) and $K_{\max} = 20$ (realistic policy window of approximately five months).

6.8.2 Convergence and Trust Evolution

The model converged in 5 rounds ($K^* = 5$), representing a 58% improvement over the 12 rounds required by the static trust benchmark under identical initial conditions. The terminal consensus degree was $GCD = 0.88$, comfortably exceeding the 0.85 threshold and consistent with the simulation results at $NCB = 20\%$.

Trust evolution during the negotiation exhibited two notable patterns. First, cross-category trust increased substantially for historically antagonistic pairs: t_{23} (coal power $\rightarrow$ renewables) rose from 0.30 to 0.42 (+40%) and t_{32} (renewables $\rightarrow$ coal) rose from 0.25 to 0.38 (+52%). The larger increase in t_{32} reflects the asymmetry of the secondary trust formula. Renewable energy firms, by accepting a moderate (rather than aggressive) quota scheme as the consensus outcome, made the larger effective concession and thereby developed more trust toward the coal power group than vice versa.

Second, the overall trust level increased by 15–20%, with the lowest initial trust pairs exhibiting the largest relative gains. This is consistent with the “ceiling effect” built into the trust update rule of Eq. (13). Pairs with low initial trust have more room to grow, while pairs with already-high trust experience smaller absolute gains due to the $[0, 1]$ bound.

6.8.3 Scenario Comparison

To assess the sensitivity of the negotiation outcome to the institutional design of the feedback mechanism, we simulate three policy scenarios with varying organizer guidance $\bar{\alpha}$, representing alternative governance philosophies.

Scenario A: Free negotiation ($\bar{\alpha} = 0$, pure reference-group feedback). This scenario requires 8 rounds, achieving $GCD = 0.86$. Coal power enterprises adjust minimally, as their trusted peers (other coal power firms and the industry association) share their preference for generous allocations. The echo-chamber effect noted in §6.5.2 is evident: the reference-group feedback reinforces rather than challenges existing positions. Although consensus is technically achieved, the outcome is a “lowest common denominator” that largely preserves the status quo.

Scenario B: Government-dominant ($\bar{\alpha} = 0.7$, strong collective evaluation signal). This scenario converges fastest (4 rounds) and achieves the highest GCD (0.91), but at the cost of high stakeholder dissatisfaction among coal power enterprises, whose individual consensus degrees CD_i remain substantially below the group average. The result is imposed rather than negotiated consensus, with numerically convergent opinions that may not reflect genuine acceptance.

Scenario C: Trust-guided ($\bar{\alpha} = 0.4$, balanced feedback with trust-based reference groups). This scenario requires 5 rounds to achieve $GCD = 0.88$, with balanced individual consensus degrees across all four stakeholders. The trust-weighted reference groups enable differentiated feedback. Coal power enterprises receive recommendations weighted toward the (more cooperative) renewable energy firms with whom they have built trust during the negotiation, while the regulator’s collective evaluation supplies the centering force.

Scenario C is recommended for practical implementation. It balances convergence speed, consensus quality and stakeholder satisfaction, and its $\bar{\alpha} = 0.4$ value is close to the simulation-identified optimum of $\bar{\alpha}^* = 0.5$ (§6.5.2), indicating that the simulation results generalize reasonably to the calibrated case.

6.8.4 Evolutionary Equilibrium

The cooperation proportion under three governance configurations at the case-study parameters discloses the critical role of the trust–governance coupling. Under no governance ($\gamma = 0$), the cooperation proportion stabilizes at $x^* \approx 0.10$, corresponding to the all-defection ESS predicted by theory when the trust premium is absent. Under governance without trust (static weights with $\gamma = 0.6$), $x^* \approx 0.88$, marginally above the ESS threshold. Under the full trust + governance configuration, $x^* \approx 0.95$, firmly in the cooperative regime. The progression $0.10 \rightarrow 0.88 \rightarrow 0.95$ mirrors the simulation results of §6.4 and independently verifies that both governance and dynamic trust are necessary for robust cooperative equilibrium in this real-world parameterization.

6.8.5 Model–Reality Comparison

The model’s scheme preference ordering (derived by aggregating the weighted evaluations at convergence) is S_2 (Moderate) $> S_1$ (Progressive) $> S_3$ (Aggressive). This ordering matches the actual policy outcome: China adopted sectoral benchmarking (a moderate approach) for the 2023–2025 allocation period rather than either grandfathering (progressive in preservation, conservative in ambition) or an aggressive absolute cap. The match should be interpreted cautiously. The model’s prediction aligns with the observed outcome, but we cannot rule out that both the model and the actual decision process converged to the same moderate outcome for different reasons.

The modeled negotiation duration of 5 rounds corresponds to approximately 5 months of iterative consultation, which is quantitatively consistent with the documented timeline of China’s quota allocation revision process. The model, however, abstracts away from political and bureaucratic delays that may extend real-world timelines, so the 5-round estimate should be understood as the consensus-relevant portion of the process rather than a prediction of total calendar duration.

The trust evolution trajectories supply a form of process validation that is stronger than the outcome validation above. The model predicts that cross-category trust (particularly coal–renewables) should increase during successful negotiation, and that the lowest-trust pairs should exhibit the largest relative gains. While we do not have direct quantitative measures of trust change during the actual Chinese consultation, these predictions are empirically falsifiable in future carbon market negotiations where stakeholder trust surveys are conducted before and after the allocation process.

6.8.6 Policy Implications

The experimental results, together with the case study, support four policy recommendations, each grounded in specific empirical findings.

1. Institutionalize trust metrics through a carbon market enterprise trust rating system. The experiments demonstrate that initial trust calibration from objective compliance and transparency data (Group B) alone improves convergence speed by 31% relative to equal weights, and that dynamic trust updating (Group C) provides an additional 50% improvement. These gains are predicated on the availability of reliable trust-relevant data. We recommend that carbon market regulators establish and publicly disclose standardized trust ratings for all covered enterprises, updated after each compliance cycle, comprising: (a) compliance rate over the preceding 3–5 cycles; (b) MRV transparency score based on the completeness, timeliness and audit outcomes of emission reports; and (c) a peer-evaluation element encoding industry reputation. Such a system would simultaneously serve the trust calibration needs of the framework and offer a general transparency benefit to the market.

2. Adopt weight floating as a complementary “soft governance” mechanism. The γ -sensitivity analysis (§6.5.1) demonstrates that a moderate, endogenous and non-exclusionary governance intensity of $\gamma = 0.6$ can maintain $CR > 96\%$ even at 50% NCB prevalence, whereas traditional fixed-penalty approaches (analogous to the no-trust baseline) collapse to $CR = 12.3\%$ under the same conditions. Weight floating is not suggested

as a replacement for fines and enforcement but as a complement. Fines address substantive violations after they occur; weight floating deters non-cooperation during the negotiation by making it influence-costly in real time. The two mechanisms operate on different timescales and information sets, and their joint deployment creates a more resilient governance architecture than either alone (Table 3).

3. Implement two-tier negotiation with intra-sector consensus preceding cross-sector coordination.

The scalability experiment (§6.5.4) indicates that convergence rounds increase with I , but sub-linearly. This implies that decomposing the negotiation into smaller sub-groups can accelerate convergence. First, each stakeholder category (coal power, renewables, verifiers, etc.) reaches internal consensus on its evaluation of the quota schemes; second, category representatives negotiate the cross-sector consensus using the trust-guided mechanism. The trust network supplies a natural basis for defining the sub-groups (intra-category trust is typically higher) and for weighting the representatives in the cross-sector phase (through accumulated intra-category trust).

4. Monitor TTD trajectories as early warning signals. The trust evolution analysis (§6.3.1) demonstrates that NCB agents' TTD begins declining within 2–3 rounds of sustained non-cooperation, well before the effects on GCD become pronounced. Regulators can monitor real-time TTD changes during active negotiations and intervene (through increased organizer guidance $\bar{\alpha}$ or direct communication) when a stakeholder's TTD decline rate exceeds a calibrated threshold. This transforms the trust network from a passive measurement tool into an active monitoring instrument, enabling preemptive rather than reactive governance responses.

7 Conclusions

7.1 Summary of Findings

We developed and validated an integrated framework combining dynamic trust networks, trust-driven consensus mechanisms and evolutionary game theory for multi-stakeholder carbon emission quota allocation. The principal findings are organized by the strength of supporting evidence.

Simulation-demonstrated findings (directly established by the Monte Carlo experiments in Chapter 6):

1. Convergence acceleration. Trust-driven quota consensus with organizer-guided feedback achieves a 72% reduction in convergence rounds ($\bar{K}^* = 5.1$ vs. 18.2) and a 0.107 increase in final consensus degree ($GCD = 0.928$ vs. 0.821) relative to the no-trust baseline at $NCB = 20\%$.
2. Robustness to adversarial behavior. The full integrated model maintains a convergence rate $CR \geq 96.8\%$ even at 50% NCB prevalence, compared to $CR = 12.3\%$ for the no-trust baseline and $CR = 54.6\%$ for the static trust configuration.
3. Optimal governance intensity. The governance parameter γ exhibits an inverted-U relationship with system performance, with the empirically identified optimum $\gamma^* = 0.6$ corroborating the theoretical derivation in §3.
4. Sub-linear scalability. Convergence rounds scale sub-linearly with the number of decision-makers ($\bar{K}^* = 4.2$ at $I = 10$ to $\bar{K}^* = 8.3$ at $I = 100$), indicating the framework's practical viability for large-scale negotiations.

Multi-source corroborated findings (supported by both simulation and case study evidence):

5. Dynamic trust as the dominant mechanism. Both the simulation (B $\rightarrow$ C transition responsible for 47% of total improvement) and the case study (58% convergence improvement from dynamic trust) identify dynamic trust updating with governance as the single most impactful architectural element.
6. Cooperation as ESS under governance. The theoretical condition of Theorem 5.1 is satisfied in both simulation ($x^* \rightarrow 0.95$ under dynamic trust) and case-study ($x^* \approx 0.95$ under trust + governance) settings, verifying that the governance mechanism can make cooperation evolutionarily stable in practical parameter regimes.

Case-study-supported findings (established in the calibrated real-world application):

7. Predictive validity. The model's scheme preference ordering ($S_2 > S_1 > S_3$) matches China's actual policy choice (moderate sectoral benchmarking adopted for 2023–2025), providing preliminary evidence of external validity.
8. Process-consistent predictions. The modeled negotiation duration (5 rounds, approximately 5 months) and trust evolution patterns (cross-category trust increases of 40–52%) are quantitatively consistent with the documented characteristics of China's quota allocation consultation process.

Theoretically established but empirically contingent findings (proven under model assumptions, awaiting broader empirical validation):

9. Convergence guarantees. Theorems 5.2 and 5.3 establish that under the ESS condition, the consensus degree converges monotonically to 1 with geometric rate. These guarantees hold under the model’s assumptions (non-degenerate trust, bounded learning rates and sufficiently strong governance) and are consistent with simulation behavior.
10. Trust–governance substitutability. The theoretical derivation indicates that raising average trust $\bar{T}$ by one unit reduces the minimum required governance intensity $\gamma_{\min}$ by $\mu/(NCBD \cdot \bar{w})$, establishing that trust building and governance strengthening are partial substitutes. The simulation results (E7: high initial trust halves convergence rounds) are consistent with this prediction, but the precise quantitative substitution rate has not been independently calibrated.

This evidence-layered presentation acknowledges the differential strength of conclusions across findings. Some are directly and robustly demonstrated; others are supported by converging lines of evidence; and still others are theoretically proven but await broader empirical confirmation.

7.2 Contributions

The theoretical contributions are threefold. First, we extend the SNGDM trust literature by introducing objectively calibrated, dynamically updated trust grounded in four categories of compliance and behavioral data, establishing a measurement framework that exploits the uniquely rich data environment of carbon trading. Second, we enrich carbon trading game theory by endogenizing trust and reputation as payoff-modulating variables, deriving the trust-modulated replicator dynamics and ESS conditions, and establishing the strategy-flip threshold t^* at which cooperation becomes individually rational. Third, we bridge consensus theory and evolutionary game theory through the dual-termination concept, which requires both consensus metric convergence and strategic stability and thereby gives a stronger guarantee of negotiated outcomes than either framework alone.

On the practical side, the weight-floating mechanism offers a concrete, parameterized governance toolkit for carbon market regulators. With its incentive-compatible design, empirically validated optimal $\gamma \approx 0.6$ and demonstrated robustness under adverse conditions (50% NCB, low initial trust), it supplies an actionable alternative to purely punitive enforcement regimes.

7.3 Limitations and Future Work

Several limitations of the current study warrant acknowledgment and point toward productive lines of future investigation.

Methodological limitations. (1) The aggregation of heterogeneous enterprises into representative decision-makers suppresses within-group variance; extension to large-scale group decision-making (LSGDM) frameworks with subgroup clustering methods would address this heterogeneity. (2) Initial trust calibration relies on expert judgment for indicator weights ($\omega_1, \omega_2, \omega_3$); multi-expert elicitation with formal sensitivity analysis would enhance robustness. (3) The simulation experiments implement only the opinion-based trust update (Eq. 13); the reported results may understate the dynamic trust model’s advantage over static alternatives, and incorporating a compliance-feedback channel in future simulations would supply a more complete assessment. (4) The replicator dynamics framework used in Chapter 5 is formally derived for infinite populations, whereas the carbon market negotiation involves $I = 20$ finite stakeholders. While the continuous-population approximation represents the qualitative ESS conditions well and is validated against finite-population Monte Carlo simulations, a more rigorous treatment using Moran processes or Wright–Fisher dynamics for finite populations would sharpen the quantitative predictions and is a natural extension. (5) The organizer neutrality assumption (§4.5.1) is a strong idealization. In practice, the government regulator serves simultaneously as the institutional organizer and a stakeholder with the highest role weight ($\kappa = 1.5$), creating a potential conflict of interest. The sensitivity analysis for $\bar{\alpha}$ (§6.5.2) partially addresses this concern by showing that the framework performs well across a wide range of guidance coefficients, but a systematic analysis of biased-organizer scenarios (in which the organizer strategically tilts feedback toward its own preferred outcome) represents an important avenue for future investigation. (6) The worst-case computational complexity of the integrated algorithm is $O(K_{\max} \cdot I^2)$, driven by the $I \times I$ trust matrix operations at each round; for markets with very large stakeholder counts ($I > 500$), sparse trust matrix representations or hierarchical clustering would be necessary to maintain tractability. A promising direction for future work that we deliberately exclude from the core model to maintain methodological clarity is the augmented trust update. The core update rule of Eq. (13) relies solely on opinion-adjustment patterns observed during negotiation. In practice, however, stated opinions need not correspond to actual behavior. This say–do gap is particularly acute in carbon trading, where enterprises may verbally agree to stricter quotas during negotiation while continuing to exceed their allowances during the compliance period. Addressing this gap requires incorporating post-compliance-cycle outcomes through

an augmented formulation:

$$\mathbf{T}^{(k+1)} = \frac{1}{2} \left[\mathbf{T}^{(k)} + \left(\theta \cdot \mathbf{S}\mathbf{T}^{(k+1)} + (1 - \theta) \cdot \mathbf{H}^{(k+1)} \right) \right], \quad (37)$$

where $\mathbf{H}^{(k+1)}$ is a compliance-feedback matrix derived from verified emission data and allowance surrender records, and $\theta \in [0, 1]$ controls the relative weight of opinion-based inference versus outcome-based verification. For each directed pair (i, j) , the entry $H_{ij}^{(k+1)} = (h_j^{(\text{cycle})} + \tau_j^{(\text{cycle})})/2$ is constructed from stakeholder j 's compliance rate $h_j^{(\text{cycle})}$ and MRV transparency score $\tau_j^{(\text{cycle})}$ in the most recently completed compliance cycle. During active negotiation, θ is set higher (0.7–0.9), giving primacy to opinion-adjustment signals; after the compliance deadline, θ is reduced, shifting weight toward objective compliance feedback. For newly entering stakeholders with no compliance history, θ is temporarily set to 1.0. The diagonal is set to $H_{ii}^{(k+1)} = 1$ by convention. This augmented formulation would close the loop between negotiation behavior and actual compliance outcomes, supplying the trust network with an independent verification channel that complements the opinion-based inference of the core model. Incorporating this extension into future simulation experiments and case validations represents an important step toward operationalizing the framework for real-world deployment.

Domain limitations. (4) The cold-start problem (initializing trust for stakeholders with no compliance history) is addressed through conservative priors and accelerated updating, but the practical reliability of this approach in markets with high rates of new entry requires empirical verification. (5) Vulnerability to large-scale collusion, in which a coalition of stakeholders coordinates non-cooperative behavior to game the trust network, has not been systematically analyzed; coordinated manipulation of trust ratings could potentially undermine the governance mechanism's effectiveness.

Validation limitations. (6) Parameters such as $\bar{\alpha}$ and γ are treated as static within each simulation run; adaptive updating of these parameters across negotiation rounds represents an important generalization that could improve performance in non-stationary environments. (7) The case study is limited to China's power sector; validation across steel, cement and chemical sectors, as well as across different institutional contexts (EU ETS, California cap-and-trade), is needed to establish broader applicability. (8) The trust network, while designed to operate as a complement to regulatory mechanisms rather than a substitute, lacks formal legal enforcement powers; its effectiveness in the absence of credible regulatory backstops is an open question.

Future work will pursue these directions with particular emphasis on: (a) adaptive parameter mechanisms that adjust γ and $\bar{\alpha}$ based on real-time negotiation dynamics; (b) multi-sector empirical validation using compliance and trading data; (c) agent-based simulations incorporating heterogeneous and boundedly rational stakeholders; and (d) integration with real-time carbon market monitoring and transaction platforms to enable automated trust updating and governance intervention.

References

- [1] Axelrod R. *The Evolution of Cooperation*. New York: Basic Books, 1984.
- [2] Coase R H. The problem of social cost. *Journal of Law and Economics*, 1960, 3: 1–44.
- [3] DeGroot M H. Reaching a consensus. *Journal of the American Statistical Association*, 1974, 69(345): 118–121.
- [4] Ding R X, Palomares I, Wang X, et al. Large-scale decision-making: characterization, taxonomy, challenges and future directions from an artificial intelligence and applications perspective. *Information Fusion*, 2020, 59: 84–102.
- [5] Dong Y, Ding Z, Martínez L, Herrera F. Managing consensus based on leadership in opinion dynamics. *Information Sciences*, 2017, 397–398: 187–205.
- [6] Dong Y, Zha Q, Zhang H, Kou G, Fujita H, Chiclana F, Herrera-Viedma E. Consensus reaching in social network group decision making: Research paradigms and challenges. *Knowledge-Based Systems*, 2018, 162: 3–13.
- [7] Ellerman A D, Buchner B K. The European Union emissions trading scheme: origins, allocation, and early results. *Review of Environmental Economics and Policy*, 2007, 1(1): 66–87.
- [8] Friedkin N E, Johnsen E C. Social influence and opinions. *Journal of Mathematical Sociology*, 1990, 15(3–4): 193–206.
- [9] Granovetter M. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 1985, 91(3): 481–510.

- [10] Hegselmann R, Krause U. Opinion dynamics and bounded confidence: models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 2002, 5(3): 2.
- [11] Hurwicz L. Optimality and informational efficiency in resource allocation processes. In: Arrow K J, Karlin S, Suppes P (eds.) *Mathematical Methods in the Social Sciences*. Stanford: Stanford University Press, 1960: 27–46.
- [12] Liu B, Zhou Q, Ding R X, Palomares I, Herrera F. Large-scale group decision making model based on social network analysis: Trust relationship-based conflict detection and elimination. *European Journal of Operational Research*, 2019, 275(2): 737–754.
- [13] Maynard Smith J, Price G R. The logic of animal conflict. *Nature*, 1973, 246: 15–18.
- [14] Maynard Smith J. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press, 1982.
- [15] McAllister D J. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 1995, 38(1): 24–59.
- [16] Mitchell R K, Agle B R, Wood D J. Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review*, 1997, 22(4): 853–886.
- [17] Nash J F. Equilibrium points in n -person games. *Proceedings of the National Academy of Sciences*, 1950, 36(1): 48–49.
- [18] Nowak M A. Five rules for the evolution of cooperation. *Science*, 2006, 314(5805): 1560–1563.
- [19] Palomares I, Martínez L, Herrera F. A consensus model to detect and manage noncooperative behaviors in large-scale group decision making. *IEEE Transactions on Fuzzy Systems*, 2014, 22(3): 516–530.
- [20] Pigou A C. *The Economics of Welfare*. London: Macmillan, 1920.
- [21] State Council of the People’s Republic of China. Interim Regulations on the Administration of Carbon Emissions Trading (State Council Decree No. 775). Promulgated 4 February 2024, effective 1 May 2024.
- [22] Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press, 1994.
- [23] Wu J, Chiclana F, Fujita H, Herrera-Viedma E. A visual interaction consensus model for social network group decision making with trust propagation. *Knowledge-Based Systems*, 2017, 122: 39–50.
- [24] Zhang H, Palomares I, Dong Y, Wang W. Managing non-cooperative behaviors in consensus-based multiple attribute group decision making: An approach based on social network analysis. *Knowledge-Based Systems*, 2018, 162: 29–45.
- [25] Zucker L G. Production of trust: Institutional sources of economic structure, 1840–1920. In: Staw B M, Cummings L L (eds.) *Research in Organizational Behavior*, Vol. 8. Greenwich: JAI Press, 1986: 53–111.
- [26] Sandholm W H. *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press, 2010.
- [27] Rodríguez R M, Martínez L, Herrera F. Hesitant fuzzy linguistic term sets for decision making. *IEEE Transactions on Fuzzy Systems*, 2012, 20(1): 109–119.
- [28] Traulsen A, Hauert C. Stochastic evolutionary game dynamics. In: Schuster H G (ed.) *Reviews of Nonlinear Dynamics and Complexity*, Vol. 2. Weinheim: Wiley-VCH, 2009: 25–61.
- [29] Chen W, Hu Z H. Using evolutionary game theory to study governments and manufacturers’ behavioral strategies under various carbon taxes and subsidies. *Journal of Cleaner Production*, 2018, 201: 123–141.
- [30] Fang G, Tian L, Fu M, Sun M. Government control or low carbon lifestyle? – Analysis and application of a novel selective-constrained energy-saving and emission-reduction dynamic evolution system. *Energy Policy*, 2014, 68: 498–507.
- [31] Nowak M A, Sasaki A, Taylor C, Fudenberg D. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 2004, 428(6983): 646–650.
- [32] Massa P, Avesani P. Trust-aware recommender systems. In: *Proceedings of the ACM Conference on Recommender Systems*. New York: ACM, 2007: 17–24.

- [33] Dong Y, Zhan M, Kou G, Ding Z, Liang H. A survey on the fusion process in opinion dynamics. *Information Fusion*, 2018, 43: 57–65.
- [34] Zhang H, Dong Y, Herrera-Viedma E. Consensus building for the heterogeneous large-scale GDM with the individual concerns and satisfactions. *IEEE Transactions on Fuzzy Systems*, 2018, 26(2): 884–898.
- [35] Calvano E, Calzolari G, Denicolò V, Pastorello S. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 2020, 110(10): 3267–3297.

Acknowledgements

The inspiration for this paper came from my interest in China’s “dual carbon” goals, which is peaking carbon emissions before 2030 and achieving carbon neutrality before 2060. When China launched its national carbon emission trading market a few years ago, initially covering more than 2,100 power generation enterprises, I gradually realized that several persistent frictions lie under the economic logic of cap-and-trade: quota allocation is a zero-sum distributional problem in which government regulators, high-emitting enterprises, renewable energy firms and other stakeholders hold directly opposing interests; firms’ abatement costs are private information; and several reported cases of emission data falsification have damaged mutual trust among market participants.

I wanted to know whether mathematics could say something concrete about how trust, or its absence, shapes the outcome of such negotiations. Then I began to read about consensus reaching processes, dynamic trust networks and evolutionary game theory, and I decided to integrate these three tools into a single framework for carbon quota allocation. That decision became this paper.

Regarding data acquisition specifically, all data used in this study are publicly available. The four objective indicators of the initial trust network (historical compliance rate, MRV data transparency, industry reputation and bilateral cooperation history) were constructed from the Ministry of Ecology and Environment’s annual compliance announcements for the 2021-2024 quota cycles, the third-party verification reports published by the Shanghai Environment and Energy Exchange, China ESG ratings, and industry peer evaluations reported by the China Electricity Council. The case study of China’s power sector was calibrated from these records, and the model’s scheme preference ordering was compared with the sectoral benchmarking policy actually adopted for the 2023-2025 allocation period.

I am the sole author of this project and carried out each stage of the research: topic selection, literature review, model construction, data acquisition, data processing and computation, detailed experiment design and implementation, and the writing of this paper. In all the works, the modeling phase was the most challenging part. In my earliest attempts I treated trust as a conventional fixed parameter, and I could not see how to make it evolve with stakeholders’ behavior during negotiation. At this stage, the guidance on theoretical content I received came from my mathematics teacher, Mr. Haoyang Lu, who specialized in mathematics. He helped me understand the core machinery of evolutionary game theory (including Nash equilibrium, evolutionarily stable strategy and replicator dynamics) and pointed me to some classic literature. At the methodological level, he suggested giving the adjustment-willingness parameter a dual identity as a strategy variable, which became the operational link between the consensus module and the game module; he also advised me to state and prove the main theorem with necessity, sufficiency and uniqueness made explicit for rigorous. For the empirical part, his contribution was suggesting the experiment direction. He proposed a stepwise comparison design (no trust $\rightarrow$ static trust $\rightarrow$ dynamic trust with governance $\rightarrow$ the full model) so that the marginal contribution of each module could be isolated, and he encouraged me to test robustness across a gradient of adversarial prevalence rather than the single setting I had previously used.

Mr. Lu’s guidance was given on a voluntary basis, out of a teacher’s genuine interest. For that, and for his constant encouragement, I am sincerely grateful.