

Measuring Agency Degradation under Rational Interaction

A note on the possibility of stable improvement and risk-prevention of multiagent-system dynamics.

Abstract

This paper introduces a formal evaluation framework for assessing agency impact in rational multi-agent systems. The framework represents agents, dependency structures, and vulnerability using quantitative measures, and analyzes decision procedures to estimate how agency degradation propagates within a closed interaction domain. In contrast to assessment approaches that rely on subjective or value-based parameters, the proposed method emphasizes consistency, payoff information, and dependency impact to support comparative analysis of alternative decision profiles prior to implementation. The resulting outputs provide measurable and auditable indicators of agency impact, enabling the analysis of stability and coherence under rational multi-agent interaction.

1. Introduction

Evaluating the impact of decisions in systems composed of multiple interacting agents remains a persistent challenge across domains such as policy analysis, organizational governance, and multi-agent artificial intelligence. Existing assessment approaches often rely on qualitative judgments, context-specific value assumptions, or domain-dependent heuristics, which limits their comparability, auditability, and stability under changing conditions. As a result, decision-makers frequently lack consistent tools for assessing how local actions propagate systemic effects across interdependent agents.

In rational multi-agent environments, interactions are shaped not only by individual incentives but also by dependency structures that amplify or mitigate the consequences of decisions. Actions that appear locally advantageous may generate indirect costs through dependency chains, leading to instability, strategic retaliation, or long-term degradation of cooperative capacity. Despite extensive work in game theory on repeated interaction and incentive alignment, these insights are rarely integrated into practical evaluation frameworks capable of quantifying agency-related impacts prior to implementation.

This paper addresses this gap by introducing a formal evaluation framework designed to assess agency impact under rational multi-agent interaction. The framework models agents, vulnerability, and dependency relations using quantitative representations, and evaluates decision procedures by estimating how agency degradation propagates across a closed interaction domain. Rather than prescribing normative objectives, the framework focuses on internal consistency, payoff structure, and dependency propagation to enable comparative analysis of alternative decision profiles.

The framework is designed to be domain-agnostic, yet it finds natural applications in fields where multi-agent dependency and agency are critical, such as:

- Artificial Intelligence: assessing the impact of autonomous agent decisions in mixed human-AI systems;
- Organizational Governance: evaluating policy changes in hierarchical or networked organizations;
- Public Policy: analyzing social assistance programs or environmental regulations for systemic equity and stability;
- Supply Chain Management: modeling disruption propagation across interdependent actors.

In each domain, the framework supports structured comparison of decision alternatives without prescribing normative outcomes.

The proposed approach is intended to support analysis rather than prescription. It does not claim moral authority or completeness, nor does it assume universal value alignment among agents. Instead, it provides measurable and auditable indicators that allow decision-makers to examine the systemic implications of choices under rational interaction assumptions. By grounding evaluation in consistency and stability considerations, the framework offers a structured method for analyzing agency-related risks in complex systems.

The remainder of this paper is organized as follows. Section 2 reviews related work in computational ethics, game-theoretic interaction, and dependency modeling. Section 3 introduces the formal definitions used throughout the framework. Section 4 describes the evaluation framework in detail, including representation, propagation, and aggregation mechanisms. Section 5 discusses stability considerations under rational interaction. Section 6 presents simulated case studies, and Section 7 outlines the limitations of the proposed approach. Section 8 concludes with potential applications and directions for future work.

2. State of the Art - Multiagent System Interactions and their Assessment

2.1 Computational Ethics and Evaluation Frameworks

Computational ethics has emerged as an interdisciplinary field concerned with the formalization and automation of ethical reasoning in computational systems [1], [2]. Early approaches primarily focused on embedding normative theories—such as rule-based deontological constraints or utility-based optimization—directly into decision-making agents. While these methods provide clear prescriptions within bounded contexts, they often rely on explicit value assumptions that limit their applicability across heterogeneous domains and stakeholder environments.

More recent work has shifted toward evaluation-oriented frameworks that aim to assess ethical properties without directly prescribing actions [3]. These approaches typically generate scores, classifications, or qualitative assessments based on predefined criteria, enabling comparison across alternatives. Examples include risk-based assessment models, fairness metrics, and compliance-oriented evaluation systems. Although such frameworks improve transparency and auditability, they frequently depend on domain-specific heuristics or subjective weighting schemes that are difficult to generalize or justify across contexts.

A recurring challenge in computational ethics is the tension between normative expressiveness and formal consistency [4]. Highly expressive models can encode rich ethical considerations but often suffer from ambiguity, conflicting rules, or sensitivity to parameter selection. Conversely, more constrained models achieve formal clarity at the cost of excluding relevant aspects of agency, vulnerability, or indirect effects. As a result, many existing frameworks struggle to provide stable evaluations when applied to complex systems characterized by interdependence and feedback.

Another limitation common to existing approaches is the treatment of ethical impact as primarily local or static. Evaluations are often performed on isolated decisions or agents, without accounting for how effects propagate through dependency structures over time. In multi-agent settings, however, decisions frequently generate indirect consequences that alter incentives, redistribute constraints, or degrade the effective agency of other participants. Frameworks that fail to model such propagation risk underestimating systemic effects and mischaracterizing long-term stability.

These observations motivate the exploration of evaluation frameworks that emphasize internal consistency, dependency-aware impact analysis, and rational interaction dynamics. Rather than encoding prescriptive ethical rules, such frameworks aim to provide structured, quantitative assessments that support comparative analysis while remaining agnostic with respect to specific normative commitments.

2.2 Game-Theoretic Perspectives on Rational Interaction

Game theory provides a formal framework for analyzing strategic interaction among rational agents, particularly in settings where outcomes depend on the actions of multiple participants. Classical models focus on equilibrium concepts such as Nash equilibrium and incentive compatibility [5]. These tools have been widely applied to study cooperation, competition, and coordination in economic and multi-agent systems.

A central result in game-theoretic analysis is that short-term payoff maximization does not necessarily yield stable outcomes under repeated interaction. In iterated games, strategies that impose excessive costs on other agents—whether through defection, exploitation, or coercive constraints—often trigger retaliation, breakdown of cooperation, or inefficient equilibria. This observation has motivated extensive research on mechanisms that support cooperative behavior, including reputation systems, reciprocity, and conditional strategies.

From a systemic perspective, many game-theoretic models implicitly assume that payoffs capture all relevant consequences of interaction. However, in complex environments, decisions may alter the structure of interaction itself by introducing new constraints, dependencies, or asymmetries among agents. Such structural changes can affect future payoffs in ways that are not immediately reflected in local utility calculations. As a result, analyses that focus exclusively on immediate payoff outcomes may overlook longer-term effects on the agents' capacity to act effectively within the system.

Recent work has explored extensions of game-theoretic models to account for network effects, externalities, and inter-agent dependencies. These approaches highlight how the degradation of one agent's strategic options can propagate through interaction networks, reshaping incentives and equilibrium properties. Nevertheless, these models are typically developed for explanatory or predictive purposes rather than as general evaluation tools applicable across decision contexts.

2.3 Dependency and Impact Propagation Models

Modeling dependencies among agents and system components is a common approach in the analysis of complex systems, particularly in domains such as risk assessment, network

theory, and systems engineering. Dependency models are often used to represent how the failure, constraint, or modification of one element can affect others through direct or indirect relationships. Graph-based representations, in which nodes correspond to entities and edges encode dependency relations, are a standard tool for capturing such structures.

In risk analysis and systems safety, impact propagation models are employed to estimate how localized disruptions can cascade across interconnected components. These models typically assign weights or probabilities to dependency links, enabling the evaluation of systemic vulnerability and the identification of critical nodes. Similar techniques have been applied in financial networks, supply chains, and infrastructure systems to analyze contagion effects and systemic risk.

Within computational and organizational contexts, dependency modeling has also been used to study information flow, authority structures, and resource constraints. Such models provide valuable insights into how constraints imposed on one agent may indirectly limit the options available to others. In iterated interaction settings, cooperation and retaliation dynamics have been extensively studied [6]. However, these approaches are often domain-specific and tailored to particular forms of dependency, making them difficult to generalize across heterogeneous multi-agent environments.

A further limitation of many existing dependency-based models is their treatment of impact in isolation from rational interaction dynamics. While they can identify potential cascades or bottlenecks, they typically do not account for how agents adapt strategically in response to changing constraints. As a result, the interaction between dependency propagation and incentive-driven behavior remains underexplored in evaluative settings. Game-theoretic models traditionally assume that payoffs capture all relevant consequences of interaction [7].

These observations suggest the value of integrating dependency and impact propagation models with frameworks that explicitly consider rational agency and interaction stability. An evaluation approach that combines quantitative dependency modeling with rational interaction analysis can offer a more comprehensive view of how decisions affect agency across interconnected systems. Dependency relations are commonly represented using graph-based models in complex systems analysis [8]. Impact propagation models are widely used in systemic risk assessment [9]. Network-level disruptions can amplify localized effects through structural dependencies [10].

Dimension	Existing Frameworks (Computational Ethics, Game Theory)	Proposed Framework (Agency Impact Evaluation)
Normative Approach	Often prescriptive; relies on rules, duties, or utility maximization	Evaluative only; does not prescribe outcomes
Dependency Modeling	Limited or implicit; rarely formalized as weighted networks	Explicit, graph-based, with weighted directional dependencies
Impact Propagation	Usually local or static; indirect effects are not systematically modeled	Dynamic, dependency-aware, with iterative indirect propagation
Agency Representation	Rarely quantified; often treated as binary or qualitative	Multidimensional vector, normalized, comparable across agents
Auditability & Transparency	Often opaque; based on subjective weighting or black-box models	Fully explicit parameters; reproducible intermediate steps
Stability Considerations	Treated separately (e.g., equilibrium analysis in repeated games)	Integrated into impact evaluation via agency degradation metrics
Value Agnosticism	Usually assumes aligned values or ethical commitments	Makes no normative assumptions; focuses on structural consistency
Output Type	Often classificatory (e.g., ethical/unethical) or prescriptive	Comparative indicators (scalar/vector), enabling trade-off analysis
Domain Generality	Frequently context-specific or tied to particular ethical theories	Designed for cross-domain application with configurable dimensions

3. Formal Definitions

This section introduces the formal definitions used throughout the paper. The definitions are intended to be operational rather than ontological, providing a consistent basis for representation and evaluation within the proposed framework.

3.1 Agents

An agent is defined as an entity capable of pursuing objectives through actions within a shared interaction environment. Agents are assumed to exhibit rational behavior in the minimal sense that their actions are guided by incentive structures and expected outcomes. No assumptions are made regarding optimality, complete information, or uniform preferences across agents.

Let

$$A = \{a_1, a_2, \dots, a_n\}$$

denote the finite set of agents participating in the interaction domain.

3.2 Agency

Agency refers to an agent's effective capacity to pursue goals without externally imposed constraints that significantly limit available options. Within the framework, agency is treated as a measurable property that can vary across agents and domains.

Agency is represented as a vector:

$$\mathbf{g}_i = (g_{i1}, g_{i2}, \dots, g_{ik})$$

where each component corresponds to a domain-specific dimension of agency (e.g., decision autonomy, resource access, informational freedom). Components are normalized to a common scale to allow comparison and aggregation.

3.3 Vulnerability

Vulnerability captures the degree to which an agent's agency can be degraded by changes in its environment or by the actions of other agents. Vulnerability is modeled as a weighting factor that modulates the impact of dependencies and interactions.

For each agent $A(i)$, vulnerability is represented as:

$$\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{ik})$$

where higher values indicate greater susceptibility. Illustration, consider a simplified scenario

with two agents, a_1 and a_2 and two agency dimensions: *decisional autonomy* and *resource access*. Suppose:

$$\mathbf{v}_1 = (0.8, 0.3), \quad \mathbf{v}_2 = (0.2, 0.9)$$

F

Here, a_1 is highly vulnerable to changes in decisional autonomy ($v_{11} = 0.8$)

but less so in resource access. Conversely, a_2 is highly vulnerable in resource access

($v_{22} = 0.9$). If a decision reduces a_1 's autonomy, by 0.5 the effective

propagated impact to a_2 via dependency links will be modulated by these vulnerability weights, emphasizing domain-specific sensitivities.

3.4 Dependency Relations

A dependency relation exists when the agency of one agent is partially contingent on the actions, resources, or constraints imposed by another agent. Dependencies are represented using a directed weighted graph:

$$D = (A, E)$$

$$e_{ij} \in E$$

where each edge e_{ij} denotes dependency from A(i) to A(j).

$$w_{ij} \in [0, 1]$$

Each dependency is associated with a weight w_{ij} , representing the strength of the dependency and its potential to transmit impact.

3.5 Impact

Impact refers to a measurable change in an agent's agency vector resulting from a decision or interaction. Impact may be direct, affecting the agent immediately involved in a decision, or indirect, propagating through dependency relations to other agents.

For a given decision procedure p , the local impact on agent $a(i)$ is defined as:

$$\Delta \mathbf{g}_i(p) = \mathbf{g}_i^{\text{post}} - \mathbf{g}_i^{\text{pre}}$$

Indirect impact is computed by propagating local impacts through the dependency graph using weighted aggregation, as described in Section 4.

3.6 Interaction Domain

The interaction domain is defined as the closed set of agents and dependency relations under consideration. Closure implies that all relevant agents and dependencies affecting agency impact are included for the purpose of evaluation. This assumption enables bounded analysis while acknowledging that real-world systems may require iterative refinement of domain boundaries.

3.7 Decision Procedures

A decision procedure is any structured action, policy, or rule set whose implementation produces changes in agency across the interaction domain. The framework evaluates decision procedures based on their estimated impact profiles rather than their stated intentions or normative justifications.

3.8 Agency Degradation

Agency degradation occurs when a decision procedure produces a net negative change in one or more components of an agent's agency vector, either directly or through propagated effects. The framework does not assume that all agency degradation is undesirable; instead, degradation is treated as a measurable effect whose distribution and magnitude are subject to evaluation.

3.9 Evaluation Scope

All definitions in this section are intended to support comparative analysis within the framework. They do not imply moral judgments, prescribe acceptable outcomes, or assume universal applicability beyond the defined interaction domain.

4. The Evaluation Framework

This section describes the proposed evaluation framework in detail. The framework is designed to assess the impact of decision procedures on agency within a rational multi-agent interaction domain by combining quantitative representations, dependency-aware propagation, and aggregation mechanisms. The objective is to produce comparable and auditable indicators of agency impact without assuming normative preferences.

4.1 Representation of Agency States

Each agent

$$a_i \in A$$

is associated with an agency vector

$$\mathbf{g}_i$$

as defined in Section 3.2. The components of this vector represent domain-specific dimensions of agency and are normalized to a common scale to allow comparison across agents and domains.

The global agency state of the interaction domain is represented as the set:

$$G = \{\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n\}$$

Decision procedures are evaluated by estimating changes in these agency vectors before and after their hypothetical implementation.

4.2 Local Impact Estimation

For a given decision procedure ppp , the framework first estimates its **local impact** on directly affected agents. Local impact is defined as the change in an agent's agency vector resulting from the immediate effects of the procedure:

$$\Delta \mathbf{g}_i^{\text{local}}(p) = \mathbf{g}_i^{\text{post}} - \mathbf{g}_i^{\text{pre}}$$

Local impact estimation may be derived from empirical data, simulations, expert input, or policy specifications, depending on the application context. The framework remains agnostic regarding the source of these estimates, provided they can be expressed in quantitative form.

4.3 Dependency-Aware Impact Propagation

Local impacts may generate indirect effects on other agents through dependency relations. To capture this, the framework propagates local agency changes across the dependency graph D defined in Section 3.4.

For each dependency edge e_{ij} with weight w_{ij} a portion of the impact on agent $a(i)$ is transmitted to agent $a(j)$, modulated by both dependency strength and vulnerability factors. Indirect impact is computed iteratively according to:

$$\Delta \mathbf{g}_j^{\text{indirect}} = \sum_i w_{ij} \cdot \mathbf{v}_j \odot \Delta \mathbf{g}_i$$

where \odot denotes element-wise multiplication. Propagation may be bounded by depth limits, decay factors, or convergence thresholds to prevent unbounded amplification.

4.4 Impact Aggregation

For each agent, total impact is obtained by aggregating local and indirect components:

$$\Delta \mathbf{g}_i^{\text{total}} = \Delta \mathbf{g}_i^{\text{local}} + \Delta \mathbf{g}_i^{\text{indirect}}$$

To support comparative analysis, the framework aggregates agent-level impacts into a system-level indicator. One possible aggregation approach is:

$$S(p) = \sum_{i=1}^n \alpha_i \cdot \|\Delta \mathbf{g}_i^{\text{total}}\|$$

where α_i represents optional weighting factors reflecting agent relevance or exposure, and $\|\cdot\|$ denotes a chosen norm. Alternative aggregation schemes may be used depending on domain requirements.

4.5 Comparative Evaluation of Decision Procedures

Decision procedures are evaluated comparatively by computing their respective impact indicators $S(p)$ under identical interaction domain assumptions. Lower aggregate degradation values correspond to decision profiles that preserve agency more effectively under the framework's criteria.

Importantly, the framework does not label decisions as acceptable or unacceptable. Instead, it provides a structured basis for comparing potential outcomes in terms of agency impact, enabling informed analysis prior to implementation.

4.6 Auditability and Reproducibility

All components of the framework —agency vectors, dependency weights, vulnerability parameters, and aggregation functions —are explicitly specified and can be inspected, modified, and re-evaluated. This design supports auditability, reproducibility, and sensitivity analysis.

By separating representation, propagation, and aggregation stages, the framework allows analysts to examine how individual modeling choices influence final evaluations without obscuring intermediate results.

4.7 Implementation Considerations

The framework is implementation-agnostic and can be instantiated using standard numerical and graph-processing tools. Computational complexity depends primarily on the size of the agent set and the density of dependency relations. In practice, sparse dependency graphs and bounded propagation depth enable efficient evaluation in moderately large systems.

5. Stability Under Rational Interaction

This section examines how the proposed evaluation framework relates to stability considerations in rational multi-agent systems. Rather than introducing new equilibrium concepts, the analysis focuses on how agency impact profiles influence the persistence of interaction structures under repeated decision-making.

5.1 Agency Preservation and Interaction Stability

In rational multi-agent environments, agents adapt their behavior in response to perceived changes in incentives and constraints. Decision procedures that significantly degrade the agency of one or more agents may alter the strategic landscape by reducing available options, increasing asymmetries, or introducing coercive dependencies. Over repeated interactions, such changes can destabilize previously viable patterns of cooperation or coordination.

Within the proposed framework, agency degradation is treated as a measurable indicator of potential instability. Large or unevenly distributed negative impacts suggest an increased likelihood that affected agents will revise strategies in ways that disrupt existing interaction structures. Conversely, decision profiles that preserve agency across agents tend to support stable engagement by maintaining a broader set of feasible responses.

5.2 Repeated Interaction and Adaptive Response

Stability considerations become particularly salient in repeated interaction domains, where agents observe outcomes and update expectations over time. Even when a decision procedure yields short-term payoff advantages, its longer-term viability depends on how it reshapes agency and dependency relations.

Agency-preserving decisions reduce incentives for defensive adaptation, exit, or escalation. By contrast, procedures that consistently erode agency may provoke countermeasures, withdrawal, or structural reconfiguration of dependencies. The framework provides a means to anticipate such dynamics by evaluating impact profiles prior to implementation.

5.3 Dependency Structure and Systemic Effects

Dependency relations play a critical role in determining how agency degradation affects system stability. When dependencies are concentrated or highly asymmetric, localized impacts can propagate widely, amplifying their effects on interaction dynamics. In such cases, even moderate local degradation may produce significant systemic instability.

The framework's dependency-aware propagation mechanism highlights these effects by identifying decision procedures whose impacts disproportionately affect highly connected or

vulnerable agents. This supports the analysis of systemic risk without requiring explicit simulation of strategic behavior.

5.4 Consistency and Rational Adaptation

From a rational perspective, agents are expected to respond consistently to changes in their effective agency. While the framework does not assume equilibrium behavior, it aligns with the minimal expectation that agents adapt when constraints or opportunities shift.

By quantifying agency impact, the framework enables the comparison of decision procedures in terms of their compatibility with sustained rational interaction. Decisions that minimize disruptive agency shifts are more likely to be compatible with stable, repeatable interaction patterns.

5.5 Scope and Interpretation

The stability considerations discussed in this section are not intended to provide predictive guarantees. Rather, they offer an evaluative lens for assessing how decision procedures may influence the conditions under which rational interaction persists.

The framework does not claim that agency preservation is always optimal or sufficient for stability. Instead, it provides measurable indicators that can inform analysis in conjunction with domain-specific knowledge and empirical validation.

6. Case Studies

This section presents two illustrative case studies demonstrating the application of the proposed evaluation framework to corporate and social decision contexts. The cases are constructed using publicly observable characteristics and explicitly stated assumptions. The objective is to show how the framework enables structured comparison of decision procedures without prescribing normative conclusions.

6.1 Corporate Case Study: Environmental Intervention Models

6.1.1. Domain Setup

The interaction domain consists of the following agent categories:

- Local communities affected by environmental conditions
- An environmental intervention organization
- Corporate actors operating within the affected area
- Regulatory institutions overseeing compliance

Agency dimensions considered include environmental safety, resource access, economic stability, and decision autonomy. Dependency relations reflect reliance on environmental quality, regulatory approval, funding flows, and public trust.

6.1.2. Decision Procedures

Two alternative environmental intervention procedures are evaluated:

- **P₁ (Centralized Intervention):**
A large-scale cleanup program designed and executed by a single coordinating entity with limited local participation.
- **P₂ (Distributed Intervention):**
A decentralized cleanup model incorporating local community involvement in planning and execution.

6.1.3. Impact Assessment

Local impact estimates are derived from assumed differences in execution control, participation, and resource distribution. In P₁, environmental improvements are achieved

rapidly but agency gains are concentrated within the coordinating entity. In P_2 , improvements occur more gradually, but agency gains are distributed across community agents.

Dependency-aware propagation highlights that P_1 increases dependency concentration, particularly where communities rely on centralized decision authority for continued environmental maintenance. P_2 reduces such concentration by distributing operational responsibility, although it introduces coordination overhead.

6.1.4. Comparative Analysis

Aggregated impact indicators reveal distinct trade-offs. P_1 exhibits higher short-term efficiency with greater asymmetry in agency distribution. P_2 shows lower aggregate efficiency but greater preservation of agency across agents with high vulnerability.

These results illustrate how the framework makes visible structural differences in agency impact that may not be captured by cost or performance metrics alone.

6.2 Social Case Study: Policy-Level Assistance Mechanisms

6.2.1. Domain Setup

The interaction domain includes:

- Beneficiaries of social assistance
- Administrative institutions
- Service providers
- Funding contributors

Agency dimensions include access to services, autonomy of choice, administrative burden, and economic predictability. Dependencies arise from eligibility criteria, service provision constraints, and funding mechanisms.

6.2.2. Decision Procedures

Two policy-level decision procedures are considered:

- **P_3 (Conditional Assistance):**
A centralized assistance program with eligibility requirements and compliance conditions.
- **P_4 (Unconditional Distribution):**
A distributed assistance mechanism providing uniform access without conditional

constraints.

6.2.3. Impact Assessment

Local impact estimation indicates that P_3 increases access to targeted services while introducing administrative constraints that reduce beneficiary autonomy. P_4 improves autonomy and reduces administrative friction but introduces increased dependency between funding contributors and distribution mechanisms.

Propagation analysis shows that conditionality in P_3 amplifies agency degradation among vulnerable beneficiaries through compliance dependencies. In contrast, P_4 shifts agency pressure toward funding and administrative agents without directly constraining beneficiary choice.

6.2.4. Comparative Analysis

System-level aggregation reveals that P_3 produces more uneven agency impact profiles, with concentrated degradation among high-vulnerability agents. P_4 results in broader redistribution of agency effects, reducing extreme degradation at the cost of increased exposure elsewhere in the system.

The framework enables these distinctions to be articulated quantitatively without asserting the superiority of either policy design.

6.3 Summary of Case Study Insights

Across both case studies, the framework demonstrates its capacity to:

- Represent heterogeneous agents and dependencies
- Quantify and propagate agency impacts
- Support comparative analysis of decision procedures

These examples illustrate how the framework can be applied to diverse domains using publicly observable assumptions, supporting transparent and auditable evaluation of agency-related effects.

7. Limitations

While the proposed framework provides a structured approach for evaluating agency impact in multi-agent systems, several limitations should be acknowledged.

First, the framework relies on the specification of agency dimensions, vulnerability factors, and dependency weights. Although these elements are explicitly represented and auditable, their instantiation may involve modeling assumptions that vary across domains. Different analysts may produce different evaluations based on alternative but reasonable parameterizations. The framework does not eliminate such variability, but rather makes it transparent and subject to examination.

Second, impact estimation is inherently approximate. Local impact values may be derived from empirical data, simulations, or expert judgment, each of which introduces uncertainty. The framework does not claim predictive accuracy; instead, it supports comparative analysis under stated assumptions. As such, results should be interpreted as conditional on the chosen modeling inputs.

Third, the framework operates within a closed interaction domain. While this assumption enables bounded analysis, real-world systems often involve external agents and dynamic boundary conditions. Extensions to handle evolving agent sets and adaptive dependency structures remain an area for future work.

Fourth, the framework does not model strategic adaptation explicitly. Although stability considerations are discussed in relation to agency impact, the framework does not simulate equilibrium behavior or learning dynamics. Its outputs should therefore be viewed as evaluative indicators rather than predictions of agent behavior.

Finally, the framework is not intended to provide normative prescriptions. It does not define acceptable or unacceptable outcomes, nor does it encode ethical priorities. Its role is limited to offering a formal mechanism for assessing and comparing the agency-related impact of decision procedures.

8. Conclusion

This paper presented an evaluation framework for assessing the impact of decision procedures on agency within rational multi-agent systems. By combining quantitative representations of agency, dependency-aware impact propagation, and transparent aggregation mechanisms, the framework enables structured and auditable comparison of alternative decision profiles.

Unlike approaches that rely on explicit normative rules or subjective weighting schemes, the proposed framework emphasizes internal consistency, measurable impact, and rational interaction dynamics. It is designed to support analysis across diverse domains without prescribing values or outcomes.

Through illustrative corporate and social case studies, the paper demonstrated how the framework can reveal trade-offs and structural effects that are not readily captured by conventional cost- or outcome-based assessments. These examples highlight the framework's potential as a decision-support tool in complex environments where agency, dependency, and systemic stability are central concerns.

Future work may explore extensions to dynamic domains, integration with strategic simulation models, and empirical validation in applied settings. As complex multi-agent systems continue to shape technological, organizational, and social decision-making, frameworks that enable transparent evaluation of agency impact may contribute to more informed and resilient system design.

References

- [1] J. H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18–21, 2006.

[2] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.

[3] B. Mittelstadt et al., “The Ethics of Algorithms: Mapping the Debate,” *Big Data & Society*, vol. 3, no. 2, 2016.

[4] L. Floridi et al., “AI4People—An Ethical Framework for a Good AI Society,” *Minds and Machines*, vol. 28, pp. 689–707, 2018.

[5] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.

[6] R. Axelrod, *The Evolution of Cooperation*. Basic Books, 1984.

[7] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991.

[8] A.-L. Barabási, *Network Science*. Cambridge University Press, 2016.

[9] D. Acemoglu, A. Daron, A. Ozdaglar, and A. Tahbaz-Salehi, “Systemic Risk and Network Formation,” NBER Working Paper, 2015.

[10] D. Helbing, “Globally Networked Risks and How to Respond,” *Nature*, vol. 497, pp. 51–59, 2013.

[11] M. Wooldridge, *An Introduction to MultiAgent Systems*. Wiley, 2009.

[12] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

