Redefining Rationality

The Standard Definition Is Self-Defeating: Diagnosis and Correction

José Ricardo Durán Barroso

Independent Researcher — Kobalt Red Research

contacto@kobaltsoftware.co

Draft — April 2026

Abstract

We demonstrate that the standard formalization of rationality ($\mathcal{R}_{Std}$, Cournot 1838, Nash 1950) does not satisfy the intuitive criterion it claims to formalize — "I want what is best for me" ($\mathcal{R}_A$) — and identify the reason this failure is necessary and geometric. The argument proceeds in three steps.

First, $\mathcal{R}_{Std}$ operates within a restricted space of formulable functions — the individual perspective $\mathcal{F}_{\text{indiv}}$ — which renders its own scope assumption invisible from within. The only object that would reveal the scope $\{i\}$ as a restriction is $\sum_{e\in\mathcal{E}} W_e(a)$, which belongs to $\mathcal{F}_{\text{sist}} \setminus \mathcal{F}_{\text{indiv}}$: not as a rejected alternative, but as an object not formulable from that space. Without a formulable alternative, the scope $\{i\}$ cannot appear as a choice — it appears as the very nature of rational analysis.

Second, Nash (1950) provides evidence of the failure: when all agents follow $\mathcal{R}_{Std}$, the equilibrium a^{NE} is not Pareto-optimal in the general case (games with Price of Anarchy greater than 1). The reason is geometric: the first-order conditions of $\mathcal{R}_{Std}$ zero out $\partial U_i/\partial a_i$ but remain structurally blind to $\sum_{e\neq i} \partial W_e/\partial a_i$. Since $\mathcal{F}_{\text{indiv}} \subsetneq \mathcal{F}_{\text{sist}}$, no set of constraints on $\mathcal{R}_{Std}$ can reproduce the systemic Pareto conditions: constraints narrow the search space without changing the geometry that determines which solutions are reachable.

Third, the correction is to change the geometry: from the systemic perspective $\mathcal{F}_{\text{sist}} \supsetneq \mathcal{F}_{\text{indiv}}$, systemic rationality

$$\mathcal{R}_S: \quad a^* = \arg\max_{a\in\mathcal{A}} \sum_{e\in\mathcal{E}} W_e(a)$$

correctly formalizes $\mathcal{R}_A$. The sum operates over individual and explicit welfare functions — not over an aggregate — and selects in $\mathcal{P}^*$ (Lemma 5.2). Under Rosen's (1965) conditions, the equilibrium where all agents follow $\mathcal{R}_S$ coincides with the social optimum without external planner and without mechanism.

Keywords: rationality, self-defeating, Nash equilibrium, Pareto optimality, perspective, solution geometry, disaggregation, systemic rationality, Rosen conditions.

Contents

1 Introduction

The history of economic theory contains a persistent anomaly: agents defined as rational consistently produce outcomes that are worse for themselves than available alternatives. Nash (1950) demonstrated this formally. The prisoner's dilemma yields $(1, 1)$ when $(3, 3)$ is achievable. The tragedy of the commons exhausts the resources on which all depend. Oligopolies leave welfare on the table for each firm.

The standard response has been to manage this anomaly: Pigou (1920) through corrective taxes, Coase (1960) through property rights, Hurwicz (1960, 1972) through mechanism design. Each strategy accepts the definition of rationality as given and works around its consequences.

This paper argues that the anomaly is not a property of the world to be managed. It is a property of the definition to be corrected. And the reason 75 years of management strategies did not resolve the problem is precise: they sought the solution within the same space that generates the problem. The solution is not there — not because it is hidden, but because the space structurally excludes it.

1.1 Relation to other papers in the programme

This is the first paper of a research programme on systemic rationality. The companion paper (Durán Barroso, 2026b) develops the implications of $\mathcal{R}_S$: it demonstrates that game theory is the special case of a more general framework where agency is not a logical necessary condition, and introduces the Observational Closure Theorem — why external auditing is a mathematical necessity, not a normative convention. The paper on the form of W_e (Durán Barroso, in preparation) demonstrates that the Shannon welfare function satisfies the requirements that $\mathcal{R}_S$ demands in order to protect individuals while maximizing the sum.

1.2 The argument in its most direct form

PREMISE: $\mathcal{R}_A$ — "I want what is best for me" (pre-theoretical)

HYPOTHESIS: $\mathcal{R}_{Std}$ claims to formalize $\mathcal{R}_A$
(operates in $\mathcal{F}_{\text{indiv}}$; scope $\{i\}$ invisible from within — §2)

EVIDENCE: Nash (1950): $a^{NE} \notin \mathcal{P}^*$ in the general case
There exists a' better for each agent; $\mathcal{R}_{Std}$ does not produce it

CONTRADICTION: $\mathcal{R}_{Std}$ does not satisfy $\mathcal{R}_A$ — it is self-defeating (§4)

WHY: $\mathcal{F}_{\text{indiv}} \subsetneq \mathcal{F}_{\text{sist}} \Rightarrow$ FOC of $\mathcal{R}_{Std}$ blind to $\sum_{e \neq i} \partial W_e / \partial a_i$
Constraints search within the same space: do not change the geometry
The systemic optimum is unreachable from $\mathcal{F}_{\text{indiv}}$ (§5)

CORRECTION: Exit $\mathcal{F}_{\text{indiv}} \to \mathcal{F}_{\text{sist}}$: change the geometry itself
$\mathcal{R}_S$: $\arg\max \sum_e W_e$ disaggregated $\Rightarrow$ $u^* \in \mathcal{P}^*$ [Lemma 5.2]
Under Rosen conditions: equilibrium = social optimum [Theorem 6.3]

---

2 Two perspectives — one contains the other

2.1 Perspectives as function spaces

A perspective, in the mathematical sense used here, is a space of functions formulable from a given viewpoint. Different perspectives make different objects available for optimization — and therefore determine which solutions are conceivable.

Definition 2.1 (Individual perspective)**.** *The space of functions whose domain is restricted to the actions of a single agent and the responses of the others:*

$$\mathcal{F}_{indiv} = \{f : f \text{ is a function of } (a_i, a_{-i}) \text{ for some } i \in N\}$$

Definition 2.2 (Systemic perspective)**.** *The space of functions that depend on the welfare of all affected entities:*

$$\mathcal{F}_{sist} = \{f : f \text{ is a function of } \{W_e\}_{e \in \mathcal{E}}\}$$

where $\mathcal{E} = \{e : \exists\, a, a' \in \mathcal{A} \text{ such that } W_e(a) \neq W_e(a')\}$ *is the set of all entities for which at least two actions produce different welfare levels.*

The complete ontological development of $\mathcal{E}$ *— including the distinction between the intrinsic existence of* W_e *and its measurement over action profiles — is found in the paper La Agencia No Es Necesaria [Agency Is Not Necessary] (Durán Barroso, 2026b). For the purposes of this paper, the relational definition suffices:* $e \in \mathcal{E}$ *if and only if the actions available in* $\mathcal{A}$ *can affect its welfare.*

Remark (On the identification $U_i = W_i$). Throughout the paper we identify the utility function U_i of Von Neumann and Morgenstern with the welfare function W_i, assuming that U_i directly represents the agent's welfare — not merely a preference ordering. This is the standard identification in welfare economics (Bergson, 1938; Samuelson, 1947). What $\mathcal{R}_{Std}$ and $\mathcal{R}_S$ share is exactly this object W_i: both recognize it as the measure of agent i's welfare. What separates them is solely the scope: $\mathcal{R}_{Std}$ limits the objective function to the set $\{i\}$; $\mathcal{R}_S$ extends it to all of $\mathcal{E}$.

Proposition 2.3 (Strict containment)**.**

$$\mathcal{F}_{indiv} \subsetneq \mathcal{F}_{sist}$$

Proof. For containment ($\subseteq$): under the identification $U_i = W_i$, every $f \in \mathcal{F}_{\text{indiv}}$ is representable as a welfare function of agent i and belongs to $\mathcal{F}_{\text{sist}}$.

For strictness ($\subsetneq$): when $\mathcal{E}$ contains an entity $e' \notin N$ without agency — an ecosystem, a future generation, a shared resource — with $W_{e'} : \mathcal{A} \to \mathbb{R}$ well defined, the function $W_{e'}$ and its values $W_{e'}(a)$ belong to $\mathcal{F}_{\text{sist}}$. Since e' is not a player, no function in $\mathcal{F}_{\text{indiv}}$ can depend on $W_{e'}$. Therefore $\sum_{e \in \mathcal{E}} W_e(a) \in \mathcal{F}_{\text{sist}} \setminus \mathcal{F}_{\text{indiv}}$. □

The strict inclusion can be read directly in the sum:

$$\sum_{e \in \mathcal{E}} W_e(a) = \underbrace{\sum_{i \in N} W_i(a)}_{\in\, \mathcal{F}_{\text{indiv}}} + \underbrace{\sum_{e \in \mathcal{E} \setminus N} W_e(a)}_{\notin\, \mathcal{F}_{\text{indiv}}}$$

The second term did not exist in the world of Von Neumann and Morgenstern — not as a rejected alternative, but as an object not formulable from $\mathcal{F}_{\text{indiv}}$.

2.2 The epistemic consequence: invisible assumptions

Definition 2.4 (Visibility of an assumption)**.** *An assumption A is* visible *within a function space $\mathcal{F}$ if and only if there exists in $\mathcal{F}$ an object consistent with $\neg A$ — an alternative formulation with respect to which A appears as a restriction. If no such object exists in $\mathcal{F}$, the assumption is* invisible *from $\mathcal{F}$: it does not appear as a choice but as the necessary nature of any possible formulation.*

Proposition 2.5 (Invisibility of the scope assumption)**.** *The scope assumption $\{i\}$ is invisible from $\mathcal{F}_{indiv}$ in the sense of Definition 2.4.*

Proof. The only object that would reveal the scope $\{i\}$ as a restriction — i.e., the alternative formulation with respect to which $\{i\}$ would appear as a choice rather than a necessity — is $\sum_{e\in\mathcal{E}} W_e(a)$. By Proposition 2.3, this object belongs to $\mathcal{F}_{\text{sist}} \setminus \mathcal{F}_{\text{indiv}}$: it does not exist in $\mathcal{F}_{\text{indiv}}$. Without a formulable alternative, the scope $\{i\}$ cannot appear as a restriction — it appears as the very nature of rational analysis. By Definition 2.4, the assumption is invisible. □

Remark (Historical note). For 186 years — from Cournot (1838) to the date of this paper — no formal tradition in economics questioned the scope assumption $\{i\}$. Not through intellectual carelessness, but through structural impossibility: the only object that would have revealed the scope $\{i\}$ as a restriction — the alternative $\sum_{e\in\mathcal{E}} W_e(a)$ — belongs to $\mathcal{F}_{\text{sist}} \setminus \mathcal{F}_{\text{indiv}}$: not as a rejected alternative, but as an object not formulable from that space. Without a formulable alternative, the scope $\{i\}$ cannot appear as a choice — it appears as the very nature of rational analysis. The perspective that generates the assumption is the same from which analysis proceeds: the error and its concealment share the same structural root.

Remark (On Harsanyi (1955)). Welfare economics (Bergson, 1938; Harsanyi, 1955) proposed social welfare functions that aggregate utilities over individuals. The difference from $\mathcal{R}_S$ is twofold: (1) they aggregate over agents N, not over entities — $\mathcal{E} = N$ in their framework, so entities without agency remain absent; (2) they require interpersonal utility comparisons or the veil of ignorance as an external normative device. $\mathcal{R}_S$ requires neither: it operates over $\mathcal{E} \supseteq N$ with explicit individual welfare functions and without an additional normative device.

3 The standard definition and its hidden assumption

Definition 3.1 (Standard rationality, $\mathcal{R}_{Std}$)**.** *Agent i is rational if, for interior solutions:*

$$a_i^* = \arg\max_{a_i\in\mathcal{A}_i} U_i(a_i, a_{-i})$$

This formulation first appears in Cournot (1838), 112 years before Nash generalized it as a solution concept. Von Neumann and Morgenstern (1944) provided the axiomatic foundation: under four consistency axioms over lotteries, there exists U_i such that the agent maximizes its expected value. Nash (1950) defined equilibrium: the profile a^{NE} such that no agent can improve U_i through unilateral deviation.

$\mathcal{R}_{Std}$ contains a scope assumption: the objective function contains only agent i. This assumption does not follow from $\mathcal{R}_A$ — "I want what is best for me" says nothing about which function to use or which entities to include. By Proposition 2.5, this assumption is invisible from within $\mathcal{F}_{\text{indiv}}$.

The formal consequence is direct. The first-order conditions of $\mathcal{R}_{Std}$ for interior solutions are:

$$\frac{\partial U_i}{\partial a_i} = 0 \quad \forall i \in N$$

These conditions are structurally blind to the term $\sum_{e \neq i} \partial W_e / \partial a_i$ — the effect of agent i's action on the welfare of all other entities — which belongs to $\mathcal{F}_{\text{sist}} \setminus \mathcal{F}_{\text{indiv}}$ and cannot appear in any first-order condition of $\mathcal{R}_{Std}$.

Remark (Historical note). The scope assumption $\{i\}$ is, by Proposition 2.5, invisible from $\mathcal{F}_{\text{indiv}}$. That structural impossibility — not intellectual carelessness — explains why for 186 years no formal tradition questioned it: the only object that would have revealed the scope $\{i\}$ as a restriction did not exist in the space from which analysis proceeded (see §2.2 for the full argument).

4 Nash as evidence of self-defeat

4.1 The intuitive criterion and its correct formalization

Definition 4.1 (Intuitive rationality, $\mathcal{R}_A$)**.** *The outcome a satisfies intuitive rationality if it is Pareto-optimal: there exists no $a' \in \mathcal{A}$ such that $W_i(a') \geq W_i(a)$ for all $i \in N$ with strict inequality for some $j \in N$:*

$$\nexists\, a' \in \mathcal{A} : \left[W_i(a') \geq W_i(a) \;\; \forall i \in N\right] \;\wedge\; \left[\exists\, j \in N : W_j(a') > W_j(a)\right]$$

Why Pareto-optimality formalizes "I want what is best for me". If outcome a is not Pareto-optimal, there exists a' that no agent would prefer to reject — and at least one strictly prefers. Accepting a over a' contradicts the agent's own criterion. Pareto-optimality is the minimum necessary condition that any outcome must satisfy to be coherent with "I want what is best for me".

4.2 Self-defeat

Definition 4.2 (Self-defeat)**.** *A definition of rationality $\mathcal{R}$ is self-defeating with respect to $\mathcal{R}_A$ if the outcome it produces is not Pareto-optimal: there exists a reachable outcome at least as good for every agent and strictly better for at least one, which $\mathcal{R}$ does not produce.*

Proposition 4.3 ($\mathcal{R}_{Std}$ is self-defeating)**.** *In games with Price of Anarchy greater than 1:*

1. *$\mathcal{R}_A$ requires (Definition 4.1): the outcome a must be Pareto-optimal.*
2. *$\mathcal{R}_{Std}$ produces a^{NE}.*
3. *Nash (1950): in games with Price of Anarchy greater than 1, $a^{NE} \notin \mathcal{P}^*$ — there exists a' with $W_i(a') \geq W_i(a^{NE})$ for all i, strict for some j.*
4. *In games with Price of Anarchy greater than 1, every Nash equilibrium under $\mathcal{R}_{Std}$ satisfies that the systemic gradient is non-zero at that point (see Lemma 5.1), which implies $a^{NE} \notin \mathcal{P}^*$. $\mathcal{R}_{Std}$ does not produce any $a' \in \mathcal{P}^*$.*
5. *Therefore $\mathcal{R}_{Std} \not\models \mathcal{R}_A$.*

Self-defeat is not a failure against an external social criterion. It is a failure against the agent's own criterion, measured by its own metric W_i. The definition that says "seek what is best for you" does not produce what is best for you.

Example (Canonical example). Prisoner's dilemma: $N = \{1, 2\}$, payoffs $U_i(C, C) = 3$, $U_i(D, D) = 1$, $U_i(D, C) = 5$, $U_i(C, D) = 0$. Under $\mathcal{R}_{Std}$: $a^{NE} = (D, D)$, $W_i = 1$ for both. But $a' = (C, C)$ gives $W_i = 3$ for both — no one is worse off and both are better off. a^{NE} is not Pareto-optimal. $\mathcal{R}_{Std}$ fails $\mathcal{R}_A$ by a factor of 3.

Remark (On genericity). Self-defeat is generic, not universal. $\mathcal{R}_{Std}$ can produce Pareto-optimal outcomes in constant-sum games (Price of Anarchy = 1). In those cases $a^{NE} \in \mathcal{P}^*$, but for particular geometric reasons specific to the game — if payoffs change marginally, the equilibrium may cease to be optimal. Proposition 4.3 states a structural failure of the definition in the general case.

4.3 Nash's incomplete diagnosis

Nash had the diagnostic but could not propose the correction. From within $\mathcal{F}_{\text{indiv}}$, the object $\sum_e W_e$ does not exist. The natural interpretation of $a^{NE} \notin \mathcal{P}^*$ from $\mathcal{F}_{\text{indiv}}$ is: "decentralized decision-making has costs — a property of the world to be managed." From $\mathcal{F}_{\text{sist}}$, the interpretation changes: "the definition has a scope error — a property of the formalization to be corrected."

Seventy-five years of post-Nash economics chose the first interpretation. This paper demonstrates that the second is correct.

5 The geometric reason: why it must fail

5.1 The utility possibility set

Let $\mathcal{U} = \{(W_e(a))_{e \in \mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|} : a \in \mathcal{A}\}$ be the set of all achievable welfare vectors. $\mathcal{U}$ is fixed: it is a property of the world, not of any objective function. What changes with the choice of $\mathcal{R}$ is which point of $\mathcal{U}$ is selected.

The Pareto frontier:

$$\mathcal{P}^* = \{u \in \mathcal{U} : \nexists\, u' \in \mathcal{U} \text{ with } u'_i \geq u_i \ \forall i,\ u' \neq u\}$$

also fixed. The central question is whether the point selected by each definition of rationality lies in $\mathcal{P}^*$ or in its interior.

5.2 Where each definition selects

Lemma 5.1 (Geometric reason for Nash's result)**.** *At a^{NE}, the gradient of systemic welfare $\sum_{e \in \mathcal{E}} W_e$ is generically non-zero: a^{NE} is not a stationary point of the sum. This explains geometrically why $a^{NE} \notin \mathcal{P}^*$ in the general case (Nash, 1950).*

Proof. The first-order conditions of $\mathcal{R}_{Std}$ at a^{NE}, for interior solutions, are $\partial U_i / \partial a_i|_{a^{NE}} = 0$ for all $i \in N$. The necessary conditions for a^{NE} to be Pareto-optimal require the gradient of the sum to be zero:

$$\left.\frac{\partial \sum_{e \in \mathcal{E}} W_e}{\partial a_i}\right|_{a^{NE}} = \underbrace{\left.\frac{\partial U_i}{\partial a_i}\right|_{a^{NE}}}_{=0 \text{ by } \mathcal{R}_{Std}} + \underbrace{\sum_{e \neq i} \left.\frac{\partial W_e}{\partial a_i}\right|_{a^{NE}}}_{\neq 0 \text{ in general}} \neq 0$$

The first term is zero by construction of $\mathcal{R}_{Std}$. The second — the cross-effect of a_i on the welfare of all other entities — is non-zero in the general case: it is zero only in degenerate games where each agent's actions produce no effects on others (Price of Anarchy = 1, see Remark in §4.2). The necessary conditions for Pareto-optimality are therefore not satisfied at a^{NE} in the general case, explaining geometrically Nash's result (1950). □

Remark. This lemma explains Nash's result geometrically — it does not prove it independently, since Nash (1950) is the established result being cited. The lemma's contribution is to identify the structural reason: the FOCs of $\mathcal{R}_{Std}$ zero out the individual gradient but are blind to the cross-effect, which is exactly the gradient that determines the direction toward $\mathcal{P}^*$. The term $\sum_{e \neq i} \partial W_e / \partial a_i \neq 0$ at a^{NE} is non-zero in the general case under two explicit conditions: (i) payoffs exhibit genuine strategic interaction (not separable in individual actions), and (ii) the equilibrium is interior. These conditions are satisfied in the prisoner's dilemma example (§6.3) and in the Cournot duopoly (§6.3).

Lemma 5.2 ($\mathcal{R}_S$ selects in $\mathcal{P}^*$)**.** *Let* $a^* = \arg\max_{a \in \mathcal{A}} \sum_{e \in \mathcal{E}} W_e(a)$. *Then* $u^* = (W_e(a^*))_{e \in \mathcal{E}} \in \mathcal{P}^*$.

Proof. Suppose $u^* \notin \mathcal{P}^*$. Then there exists $a' \in \mathcal{A}$ with $W_e(a') \geq W_e(a^*)$ for all $e \in \mathcal{E}$, strict for some j. Summing over $\mathcal{E}$:

$$\sum_{e \in \mathcal{E}} W_e(a') > \sum_{e \in \mathcal{E}} W_e(a^*)$$

This contradicts $a^* = \arg\max_a \sum_e W_e$. Therefore $u^* \in \mathcal{P}^*$. □

Remark (Existence note). Weierstrass' theorem guarantees the existence of a^*: continuous W_e and compact $\mathcal{A}$ are sufficient.

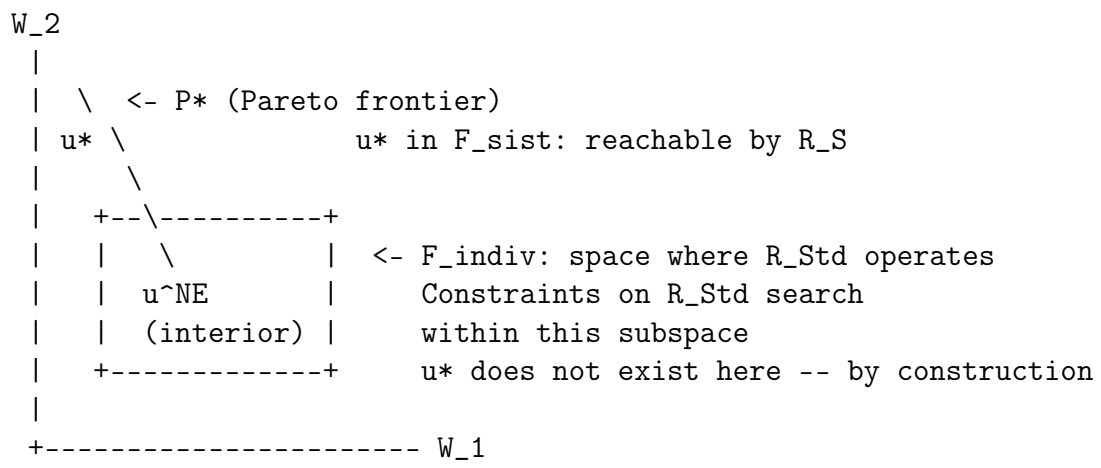

Figure 1. u^{NE} in the interior of $\mathcal{U}$; u^* on its boundary $\mathcal{P}^*$. The subspace where $\mathcal{R}_{Std}$ operates (inner rectangle) does not contain u^*. No constraint on $\mathcal{R}_{Std}$ changes this structural fact.

5.3 Why constraints cannot close the gap

Theorem 5.3 (Impossibility of constraints)**.** *No set of constraints on* $\mathcal{R}_{Std}$ *produces* $\mathcal{P}^*$ *in general.*

Proof. **Geometric argument (intuition).** Since $\mathcal{F}_{\text{indiv}} \subsetneq \mathcal{F}_{\text{sist}}$, every function $f \in \mathcal{F}_{\text{indiv}}$ is structurally blind to the term $\sum_{e \in \mathcal{E} \setminus N} \partial W_e / \partial a_i$. The first-order conditions of any $f \in \mathcal{F}_{\text{indiv}}$ — with or without constraints — cannot coincide with the systemic Pareto conditions, which require that term to be zero. Constraints on $\mathcal{R}_{Std}$ narrow the search subspace within $\mathcal{F}_{\text{indiv}}$ without changing the geometry that determines which optimality conditions are formulable. The systemic optimum u^* is unreachable from $\mathcal{F}_{\text{indiv}}$ with or without constraints.

Formal proof (Kuhn-Tucker, 1951). Adding constraints $g_k(a) \leq 0$ to $\mathcal{R}_{Std}$ modifies the first-order conditions to:

$$\frac{\partial U_i}{\partial a_i} + \sum_k \lambda_k \frac{\partial g_k}{\partial a_i} = 0 \quad \forall i \in N,\ \lambda_k \geq 0$$

To reproduce the systemic Pareto conditions the following identity would be needed:

$$\sum_k \lambda_k \frac{\partial g_k}{\partial a_i}(a) \equiv \sum_{e \neq i} \frac{\partial W_e}{\partial a_i}(a) \quad \forall i \in N,\ \forall a \in \mathcal{A}$$

Here the $\lambda_k \geq 0$ are multipliers evaluated at a given point — they are fixed values at that point, not functions of a over $\mathcal{A}$. Therefore the left-hand side is a fixed linear combination of gradients of fixed functions g_k. To reproduce the right-hand side $\sum_{e \neq i} \partial W_e / \partial a_i(a)$, which varies with a according to the shape of the functions W_e, the λ_k would need to depend on a — which would be equivalent to the "mechanism" being the systemic operator $\mathcal{R}_S$ itself, not a constraint on $\mathcal{R}_{Std}$. For generic W_e, no finite linear combination of gradients of fixed functions g_k with fixed multipliers can reproduce that behavior throughout $\mathcal{A}$. □

The historical implication is precise. Pigou's taxes, Coase's property rights, Hurwicz's mechanism design — each adds a term $\sum_k \lambda_k \partial g_k / \partial a_i$ to the first-order conditions of $\mathcal{R}_{Std}$. The theorem demonstrates that none of these strategies can reproduce $\sum_{e \neq i} \partial W_e / \partial a_i$ in general. The existing literature applied the right tools in the wrong function space.

6 The correction: systemic rationality $\mathcal{R}_S$

6.1 The definition and disaggregation

The correction is not to add constraints to $\mathcal{R}_{Std}$. It is to change the geometry itself — to exit $\mathcal{F}_{\text{indiv}}$ toward $\mathcal{F}_{\text{sist}}$, where the systemic Pareto conditions are formulable and u^* exists.

Definition 6.1 (Systemic rationality, $\mathcal{R}_S$)**.**

$$\mathcal{R}_S : \quad a^* = \arg\max_{a \in \mathcal{A}} \sum_{e \in \mathcal{E}} W_e(a)$$

In plain terms: *what is best for me is determined jointly with what is best for every entity in the system — not instead of it.*

The explicit decomposition is constitutive — it is not algebraically redundant, it is informationally necessary:

$$\sum_{e \in \mathcal{E}} W_e(a) = \underbrace{W_i(a)}_{\text{best for me}} + \underbrace{\sum_{e \in \mathcal{E} \setminus \{i\}} W_e(a)}_{\text{best for each entity in the system}}$$

This form establishes that the sum operates over individual and explicitly known welfare functions — not over an aggregate $G(a)$ that approximates the system as a whole.

Why disaggregation matters. An aggregate $G(a)$ can be computed without knowing W_e for any entity individually. If the welfare of some entity e' collapses, the aggregate may not detect it — the collapse is diluted within the global index. The disaggregated sum makes it visible without ambiguity. The three definitions are strictly ordered by preserved informational content:

Definition	**What it preserves**	**What it loses**
$\mathcal{R}_{Std}$: $\max_{a_i} U_i$	Only W_i	All $W_{e\neq i}$
$G(a)$: $\max_a G(a)$	Aggregate trend	Individual information of each entity
$\mathcal{R}_S$: $\max_a \sum_e W_e$	All W_e for each e	Nothing

$\mathcal{R}_S$ is the only formulation with no information loss.

$\mathcal{R}_S$ **is not altruism.** The agent following $\mathcal{R}_S$ reaches the Pareto frontier — $u^* \in \mathcal{P}^*$ (Lemma 5.2) — while $\mathcal{R}_{Std}$ cannot (Lemma 5.1). The outcome of $\mathcal{R}_S$ cannot be Pareto-dominated. Whether additionally $W_i(a^*) \geq W_i(a^{NE})$ holds for agent i individually depends on the specific form of W_e: that guarantee requires adaptive sensitivity at the frontier and is established in Open Problem 1 and in the companion paper on Shannon.

$\mathcal{R}_S$ is the object in $\mathcal{F}_{\text{sist}} \setminus \mathcal{F}_{\text{indiv}}$ that makes the scope assumption of $\mathcal{R}_{Std}$ visible as an assumption. Its existence is precisely what the individual perspective rendered invisible for 186 years.

6.2 The game under $\mathcal{R}_S$

When all agents in N adopt $\mathcal{R}_S$, the resulting game G_S has shared objective function $\Phi(a) = \sum_{e\in\mathcal{E}} W_e(a)$. The Nash equilibrium of G_S is the profile a^S such that no agent can increase Φ through unilateral deviation:

$$\Phi(a_i^S, a_{-i}^S) \geq \Phi(a_i, a_{-i}^S) \quad \forall a_i \in \mathcal{A}_i,\ \forall i \in N$$

Lemma 6.2 (Stationarity of Φ in G_S)**.** *For interior solutions, a^S is a stationary point of Φ: $\partial\Phi/\partial a_i|_{a^S} = 0$ for all $i \in N$.*

Proof. At a^S, no agent can increase Φ along its own coordinate. For interior solutions, this implies $\partial\Phi/\partial a_i|_{a^S} = 0$ for all i. Since the actions of different agents are independent coordinates, this is equivalent to $\nabla_a\Phi|_{a^S} = 0$ — exactly the stationarity conditions of $\max_a \Phi(a)$. □

Why $a^{NE}_{G_S} = a^*$. In G_S all agents maximize the same function Φ. The Nash equilibrium of a game where all maximize the same function is the point where no agent can increase Φ by deviating — which, for interior solutions, is exactly $\nabla_a\Phi = 0$: the optimality conditions of $\max_a \Phi(a)$. This equality holds under the concavity conditions of Theorem 6.3 (Rosen, 1965); without them, $\nabla_a\Phi = 0$ identifies a stationary point but not necessarily the global maximizer. Under Rosen's conditions that point is guaranteed to be unique.

Theorem 6.3 (Rosen, 1965)**.** *If each $W_i(a_i, a_{-i})$ is strictly concave in a_i given a_{-i}, $\mathcal{A}$ is convex and compact, and Rosen's strict diagonal concavity condition is satisfied, then there exists a unique Nash equilibrium of G_S and:*

$$a^{NE}_{\mathcal{R}_S} = a^* \in \mathcal{P}^*$$

The equilibrium where all agents follow $\mathcal{R}_S$ coincides with the social optimum. The Nash-Pareto divergence disappears — not through external intervention, not through mechanism design, but by correcting the definition of rationality.

Remark (Logical independence of the two central results). Lemma 5.2 ($u^* \in \mathcal{P}^*$) requires only continuity and compactness (Weierstrass): it guarantees that $\mathcal{R}_S$ produces a Pareto-optimal outcome regardless of the game's conditions. Theorem 6.3 additionally requires strict individual concavity and strict diagonal concavity: it further guarantees that the Nash equilibrium of G_S reaches that optimum without a planner. The self-defeat of $\mathcal{R}_{Std}$ holds independently of which of the two applies.

6.3 Verification in simple games

Case 1 — Prisoner's dilemma (discrete, linear). Under $\mathcal{R}_{Std}$: $a^{NE} = (D, D)$, $\sum W_i = 2$. Under $\mathcal{R}_S$: each agent maximizes $W_i + W_j$. For agent 1: $W_1(C,C) + W_2(C,C) = 6 > W_1(D,C) + W_2(D,C) = 5$. Both choose C. $a^S = (C,C)$, $\sum W_i = 6 \in \mathcal{P}^*$. ✓

Case 2 — Cournot duopoly (continuous, quadratic). With $W_i(q_i, q_j) = q_i(1 - q_i - q_j) - cq_i$:

- Under $\mathcal{R}_{Std}$: $q_i^{NE} = (1-c)/3$, sum $= 2(1-c)^2/9$.
- Under $\mathcal{R}_S$: first-order condition $1 - 2(q_i + q_j) - c = 0$, so $q_i + q_j = (1-c)/2$, sum $= (1-c)^2/4 > 2(1-c)^2/9$. $a^S \in \mathcal{P}^*$. ✓

In both cases the result is achieved from within the system, without an external planner — a decisive difference from classical welfare economics.

6.4 The complete logical chain

PREMISE: $\mathcal{R}_A$ — "I want what is best for me"

HYPOTHESIS: $\mathcal{R}_{Std}$ claims to formalize $\mathcal{R}_A$

STRUCTURE: $\mathcal{R}_{Std}$ operates within $\mathcal{F}_{\text{indiv}}$

$\mathcal{F}_{\text{indiv}} \subsetneq \mathcal{F}_{\text{sist}}$ [Prop. 2.3]

$\Rightarrow$ scope assumption $\{i\}$ invisible from within [Prop. 2.5]

$\Rightarrow \sum_{e \neq i} \partial W_e / \partial a_i$ blind to $\mathcal{R}_{Std}$ [§3]

EVIDENCE: Nash (1950): $a^{NE} \notin \mathcal{P}^*$ (general case: PoA > 1)

There exists a' Pareto-dominating; $\mathcal{R}_{Std}$ produces no $a' \in \mathcal{P}^*$ [Lemma 5.1]

CONTRADICTION: $\mathcal{R}_{Std}$ does not satisfy $\mathcal{R}_A$ — it is self-defeating [Prop. 4.3]

WHY: $\mathcal{F}_{\text{indiv}} \subsetneq \mathcal{F}_{\text{sist}} \Rightarrow$ FOC of $\mathcal{R}_{Std}$ blind to cross-effect [Lemma 5.1]

Constraints on $\mathcal{R}_{Std}$: search within $\mathcal{F}_{\text{indiv}}$, without changing geometry

No combination $\lambda_k \partial g_k / \partial a_i$ reproduces $\sum_{e \neq i} \partial W_e / \partial a_i$ [Thm. 5.3]

CORRECTION: Exit $\mathcal{F}_{\text{indiv}} \to \mathcal{F}_{\text{sist}}$: change the geometry itself

$\mathcal{R}_S$: $\arg\max \sum_e W_e$ disaggregated $\Rightarrow$ $u^* \in \mathcal{P}^*$ [Lemma 5.2]

In G_S all maximize Φ $\Rightarrow$ $a^{NE}_{G_S} = a^*$ [§6.2]

Under Rosen conditions: unique, guaranteed [Thm. 6.3]

$\mathcal{R}_S$ satisfies $\mathcal{R}_A$

---

7 What remains open

Open Problem 1 — Individual dominance. Lemma 5.2 establishes $u^* \in \mathcal{P}^*$ but does not guarantee $W_i(a^*) \geq W_i(a^{NE})$ for every i individually — maximizing the sum may produce trade-offs between entities. Protecting each individual while maximizing the sum requires a specific form of W_e with adaptive sensitivity at the frontier. The companion paper (Durán Barroso, in preparation) demonstrates that Shannon's welfare function $V(D, C) = -D \ln D \cdot C$ satisfies this requirement: $\partial V/\partial D \to +\infty$ as $D \to 0$ makes the marginal cost of harming the most vulnerable entity infinite.

Open Problem 2 — The form of W_e. $\mathcal{R}_S$ is valid for any continuous W_e. The question of scope (answered here: all entities in $\mathcal{E}$, with disaggregated sum) and the question of form (answered in separate work) are orthogonal to each other.

Open Problem 3 — Implementation without agency. How the system produces a^* without requiring agents to consciously adopt $\mathcal{R}_S$ — through perception, computation and transmission rather than strategic decision — is the subject of the systemic optimization paper (Durán Barroso, 2026b). That paper demonstrates that game theory is the special case of a more general framework where agency is a sufficient but not necessary condition.

8 Conclusion

The standard definition of rationality $\mathcal{R}_{Std}$ — from Cournot (1838) to Nash (1950) — operates within $\mathcal{F}_{\text{indiv}}$, whose first-order conditions are structurally blind to the term $\sum_{e \neq i} \partial W_e/\partial a_i$. The systemic optimum u^* requires that term to be zero — a condition that $\mathcal{F}_{\text{indiv}}$ cannot formulate. The entire post-Nash tradition added constraints within the same space: the right tools in the wrong geometry.

The correction is not to constrain. It is to change the geometry: to exit $\mathcal{F}_{\text{indiv}}$ toward $\mathcal{F}_{\text{sist}}$, where $\sum_{e \in \mathcal{E}} W_e(a)$ is formulable and u^* exists. $\mathcal{R}_S$ with disaggregated sum is the only formulation that preserves individual welfare information for each entity — information that any aggregate necessarily discards.

By Lemma 5.2, $\mathcal{R}_S$ produces results in $\mathcal{P}^*$. Under Rosen's (1965) conditions, the equilibrium where all agents follow $\mathcal{R}_S$ coincides with the social optimum without planner, without mechanism, and without explicit coordination.

Nash's result was not a description of how the world necessarily is. It was a diagnosis of how the definition necessarily fails: the first-order conditions of $\mathcal{R}_{Std}$ zero out $\partial U_i/\partial a_i$ — the individual gradient — but are structurally blind to $\sum_{e \neq i} \partial W_e/\partial a_i$ — the effect of each action on the rest of the system — which is precisely the gradient that points the direction toward $\mathcal{P}^*$. $\mathcal{R}_S$ corrects this: by maximizing $\sum_{e \in \mathcal{E}} W_e(a)$, its optimality conditions incorporate both gradients and select on the Pareto frontier without an external planner.

References

[1] Bergson, A. (1938). A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics*, 52(2), 310–334.

[2] Arrow, K. J. (1951). *Social Choice and Individual Values*. John Wiley & Sons.

[3] Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3, 1–44.

[4] Cournot, A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. Hachette.

[5] Durán Barroso, J. R. (2026b). Agency is not necessary: systemic optimization as a general framework of which game theory is a special case. *Kobalt Red Research*.

[6] Durán Barroso, J. R. (in preparation). The Shannon welfare function: axiomatic foundations and individual protection under systemic rationality.

[7] Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4), 309–321.

[8] Hurwicz, L. (1960). Optimality and informational efficiency in resource allocation processes. In *Mathematical Methods in the Social Sciences*. Stanford University Press.

[9] Hurwicz, L. (1972). On informationally decentralized systems. In *Decision and Organization*. North-Holland.

[10] Koutsoupias, E. & Papadimitriou, C. (1999). Worst-case equilibria. *Proceedings of the 16th STACS*, 404–413.

[11] Kuhn, H. W. & Tucker, A. W. (1951). Nonlinear programming. *Proceedings of the Second Berkeley Symposium*, 481–492.

[12] Nash, J. (1950). Equilibrium points in n-person games. *PNAS*, 36(1), 48–49.

[13] Pigou, A. C. (1920). *The Economics of Welfare*. Macmillan.

[14] Rosen, J. B. (1965). Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3), 520–534.

[15] Samuelson, P. A. (1947). *Foundations of Economic Analysis*. Harvard University Press.

[16] Von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

[17] Weierstrass, K. (1861). Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen reeller Argumente.

Companion papers in the same research programme:
— Durán Barroso (2026b). Agency is not necessary: systemic optimization as a general framework of which game theory is a special case.
— Durán Barroso (in preparation). The Shannon welfare function: axiomatic foundations and individual protection under systemic rationality.

Working draft — April 2026
Kobalt Red Research — contacto@kobaltsoftware.co