Institutionalising the Noble Person Test: Adversarial Debate as a Mechanism for Just Institutional Design

Shuhao Zhong *

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1980.v1

Keywords: institutional design; adversarial debate; red team; Noble Person Test; justice; democratic governance; sacrifice; burden of proof

 doi:10.20944/preprints202602.1980.v1

Article

Institutionalising the Noble Person Test: Adversarial Debate as a Mechanism for Just Institutional Design

Shuhao Zhong

Independent Researcher, China; persbective@163.com

Abstract

This paper develops the institutional implications of the Noble Person Test, a framework for evaluating justice proposed in [Shuhao Z., Beyond the Veil of Ignorance: The Noble Person Test as a Framework for Justice]. The Noble Person Test evaluates institutional arrangements by asking whether a hypothetical agent—default self-interested, intellectually honest, and persuadable under strict conditions—would accept the arrangement from every position within it. This paper argues that the test is best operationalised not as individual thought experiment but as **structured adversarial debate**: a red team representing those bearing the costs of an arrangement defaults to refusal, while a blue team representing those proposing the arrangement bears the burden of proving necessity and the absence of less costly alternatives. The paper derives four structural features that just institutions must possess, examines the relationship between the Noble Person Test and democratic governance, applies the framework to three domains of legal and policy controversy, and proposes concrete institutional mechanisms for implementing adversarial review. The paper draws on existing practices in military red-teaming, intelligence analysis, and judicial adversarial procedure to argue that the proposed mechanism is not utopian but an extension of proven institutional designs.

Keywords: institutional design; adversarial debate; red team; Noble Person Test; justice; democratic governance; sacrifice; burden of proof

---

1. Introduction

1.1. From Thought Experiment to Institution

Political philosophy has produced powerful thought experiments for evaluating justice—Rawls's original position, Scanlon's reasonable rejection test, Smith's impartial spectator—but has given comparatively little attention to the question of how these thought experiments should be *institutionalised*: embedded in actual decision-making processes so that they discipline real choices by real actors.

This gap matters. A thought experiment that lives only in the philosopher's study can criticise institutions but cannot improve them. Improvement requires a mechanism—a procedure that real institutions can adopt, that structures real debates, and that produces real decisions constrained by the thought experiment's logic.

In a companion paper [Author, companion paper], I proposed the Noble Person Test as a framework for evaluating institutional justice. The test asks: would a hypothetical agent—default self-interested, intellectually honest, and persuadable only when shown that not sacrificing produces worse long-term consequences and no less costly alternative exists—accept this arrangement from every position within it? The companion paper developed the concept, compared it with existing frameworks, and applied it to hard cases.

This paper takes the next step. It argues that the Noble Person Test is best operationalised as **structured adversarial debate**—not as a thought experiment conducted by a single person, but as an institutional mechanism in which opposing teams argue under fixed rules. It derives the institutional features that just arrangements must possess, examines the framework's relationship to democracy, applies it to specific legal and policy domains, and proposes concrete mechanisms for implementation.

 doi:10.20944/preprints202602.1980.v1

1.2. Why Adversarial Debate?

A single person applying the Noble Person Test faces a fundamental limitation: they cannot fully escape their own position, interests, and cognitive biases. Even with the best intentions, a policymaker imagining themselves in the least advantaged position will systematically underestimate the costs borne by that position and overestimate the necessity of the sacrifice they are proposing [5].

Adversarial debate addresses this limitation structurally. Instead of asking one person to simulate all positions, it assigns different positions to different teams:

- A **red team** represents those bearing the costs of the proposed arrangement. Their task is to refuse—to argue that the sacrifice is unnecessary, that less costly alternatives exist, or that the long-term consequences have been underestimated.
- A **blue team** represents those proposing the arrangement. Their task is to persuade—to demonstrate necessity, the absence of alternatives, and acceptable long-term consequences.

The debate proceeds under fixed rules derived from the Noble Person Test:

(i) The burden of proof lies with the blue team.
(ii) The default outcome is the red team's position—no sacrifice.
(iii) Both sides must rely on independently verifiable facts and logic.
(iv) Emotional manipulation, information suppression, and appeals to authority are prohibited.
(v) The scope of "long-term" is itself subject to argument.
(vi) Arrangements based on identity prejudice (characteristics irrelevant to the problem) are excluded from debate.
(vii) Conclusions are provisional—new evidence may reopen the debate.

This structure is not novel in practice. Military organisations have used red-teaming for decades to stress-test plans and assumptions [12]. Intelligence agencies employ structured adversarial analysis to reduce cognitive bias [4]. Judicial systems in common-law countries are built on adversarial procedure. What is novel is the application of adversarial structure to *institutional design decisions*, guided by a specific normative framework (the Noble Person Test) that determines the rules of engagement.

1.3. Plan of the Paper

Section 2 derives four structural features that just institutions must possess, using the Noble Person Test as the tool of derivation. Section 3 examines the relationship between the Noble Person Test and democratic governance. Section 4 applies the framework to three domains: self-defence law, capital punishment, and emergency powers. Section 5 proposes concrete institutional mechanisms for adversarial review. Section 6 addresses objections. Section 7 concludes.

2. Four Features of Just Institutions

The Noble Person Test, applied systematically across institutional positions, yields four features that any just institution must possess. Each is derived by the same method: place the Noble Person at a specific position, identify what they would refuse, and determine whether persuasion under the test's conditions could succeed. When persuasion cannot succeed—because a less costly alternative always exists—the arrangement fails, and the feature becomes a requirement.

2.1. Feature 1: Protection of Creative Incentives

Derivation. Place the Noble Person at the position of a creator—an entrepreneur, inventor, skilled worker, or artist—whose output is substantially redistributed.

The Noble Person defaults to refusal: "My output is being taken. I do not accept." Can they be persuaded? The blue team argues that redistribution funds public goods (security, infrastructure, education) without which the creator could not have produced their output. The Noble Person, being intellectually honest, acknowledges this.

 doi:10.20944/preprints202602.1980.v1

But the Noble Person also recognises a limit. A tax rate that leaves sufficient after-tax return preserves the incentive to create. A tax rate that approaches confiscation destroys it. Destroying creative incentives reduces total output, harms everyone including the least advantaged, and is therefore self-defeating in the long run. A less costly alternative—a rate high enough to fund public goods but low enough to preserve incentives—always exists.

Conclusion. Just institutions must allow creators to retain sufficient returns to sustain creative effort. The specific rate is within the acceptable range and is determined by empirical research and democratic deliberation, not by the theory.

This derivation does not rest on a rights-based argument ("creators have a right to their output") but on the Noble Person Test directly: the Noble Person at the creator's position cannot be persuaded to accept confiscatory redistribution because a less costly alternative exists.

2.2. Feature 2: Basic Security for All

Derivation. Place the Noble Person at the least advantaged position—unable to meet basic needs, with no prospect of improvement.

The Noble Person refuses: "I cannot survive in this arrangement. I do not accept." Can they be persuaded? The blue team would need to show that providing basic security is impossible or that its cost exceeds its benefit. But basic security systems are feasible—numerous societies have implemented them successfully. A less costly alternative (an arrangement with basic security) always exists.

Moreover, the Noble Person is intellectually honest: they know that anyone—including those currently advantaged—may fall into the least advantaged position through illness, accident, economic disruption, or technological change. Strength and weakness are *fluid conditions*, not fixed identities. Protecting the weakest is therefore rational self-insurance: it protects everyone against the possibility of becoming weak.

Conclusion. Just institutions must provide basic security for all members. This is not charity; it is rational institutional design derived from the Noble Person's self-interest and intellectual honesty.

2.3. Feature 3: Social Mobility

Derivation. Place the Noble Person at the position of someone born into the least advantaged circumstances.

The Noble Person refuses: "My birth completely determines my fate. I do not accept." Can they be persuaded? The blue team would need to show that preventing birth-determined outcomes is impossible or too costly. But public education, anti-discrimination law, and fair competition mechanisms are feasible and have been implemented in many societies. A less costly alternative always exists.

The Noble Person, being intellectually honest, also recognises that rigid stratification suppresses talent on a massive scale: people born into disadvantage who could have been innovators, leaders, or creators never get the chance. The long-term consequence is reduced collective capacity and increased instability as frustrated talent turns to resistance.

Conclusion. Just institutions must maintain meaningful social mobility. Differences in outcome due to differences in ability and effort are acceptable; differences in outcome determined entirely by birth are not.

2.4. Feature 4: Error Correction Capacity

Derivation. Place the Noble Person at any position within an institution that lacks mechanisms for discovering and correcting its own errors.

The Noble Person refuses: "If this institution makes a mistake—and it will, because all institutions do—there is no way to fix it. Errors accumulate until collapse. I do not accept." Can they be persuaded? The blue team cannot overcome this objection because an institution *with* error correction capacity is always a less costly alternative to one without it.

This derivation is particularly important because it is *internal* to the Noble Person Test. The Noble Person is intellectually honest: they know that anyone applying the Noble Person Test—including the

 doi:10.20944/preprints202602.1980.v1

present author—may be wrong. An institution without error correction is one in which mistakes made during the initial design, including mistakes in the application of the Noble Person Test itself, can never be rectified. The Noble Person's intellectual honesty compels them to reject such an institution.

Error correction requires specific institutional capacities:

(i) **Error detection**: freedom of speech, freedom of information, independent media, independent research.
(ii) **Error correction**: institutional flexibility, separation of powers, independent judiciary, regular policy review.
(iii) **Victim compensation**: targeted reparation for individuals harmed by erroneous policies.

These are simultaneously the preconditions for the adversarial debate mechanism to function. Without freedom of information, the red team cannot access the evidence it needs. Without institutional flexibility, debate conclusions cannot be implemented. Without victim compensation, the credibility of the entire system is undermined. The requirements of error correction and the requirements of adversarial debate converge on the same institutional capacities.

2.5. The Relationship Among the Four Features

Features 1–3 are *substantive requirements*: they specify what just institutions must provide. Feature 4 is a *capacity requirement*: it specifies what just institutions must be able to do.

Feature 4 is logically prior to Features 1–3. Without error correction, any initial design—however well it satisfies Features 1–3—will drift from justice as circumstances change and errors accumulate. Error correction is the mechanism by which institutions *maintain* justice over time, not merely achieve it at a single moment.

The four features jointly define a *dynamic equilibrium*: institutions must balance creative incentives against basic security and social mobility (Features 1–3), and must be able to adjust this balance as conditions change (Feature 4). Justice is not a fixed state but a continuously maintained condition, analogous to the dynamic equilibrium of a living organism.

3. The Noble Person Test and Democratic Governance

3.1. The Relationship: Floor, Not Replacement

The Noble Person Test does not replace democratic governance. It provides a **floor**—a set of arrangements below which no democratic decision can legitimately go.

Democracy determines specific policies: the precise tax rate, the retirement age, the level of social spending. These are choices within the acceptable range identified by the Noble Person Test. The test's role is to define the *boundaries* of that range: arrangements that the Noble Person at some position cannot be persuaded to accept are excluded regardless of how many people vote for them.

The relationship is analogous to that between a constitution and ordinary legislation. A constitution sets limits that legislation must respect. The Noble Person Test sets limits that democratic decisions must respect. Within those limits, the people decide.

3.2. Why a Floor Is Needed

Democratic majorities can produce unjust outcomes. A 60% majority voting to strip a minority of its property, liberty, or political participation is a familiar historical phenomenon [11]. Majority vote does not establish that the Noble Person at the minority's position could be persuaded: the fact that 60% of voters prefer an arrangement does not demonstrate that the arrangement is necessary, that its long-term consequences are acceptable, or that no less costly alternative exists.

The Noble Person Test provides a principled criterion for distinguishing legitimate democratic decisions from majority tyranny. A democratic decision is legitimate if it falls within the range that the Noble Person at every position would accept. It is illegitimate—regardless of the margin of victory—if the Noble Person at some position cannot be persuaded.

 doi:10.20944/preprints202602.1980.v1

3.3. Why the Test Does Not Become Antidemocratic

The concern that a normative test applied to democratic outcomes is inherently antidemocratic is understandable but misplaced.

First, the test only excludes arrangements below the floor—it does not select among arrangements above it. Above the floor, democratic choice operates freely. The test *expands* the domain of legitimate democratic choice by protecting the conditions under which meaningful democracy is possible: freedom of speech, access to information, protection of minorities, and institutional flexibility. Without these conditions, democracy degrades into a mechanism for ratifying the preferences of those who control information and coercion.

Second, the test itself can be applied democratically. The adversarial debate mechanism proposed in this paper is a *democratic* process: both sides argue publicly, the evidence is available to all, and the conclusion is subject to democratic review and revision. The test does not empower a philosopher-king; it structures public argument.

Third, existing democracies already apply normative tests to democratic outcomes. Constitutional judicial review in the United States, Germany, and many other countries involves unelected judges evaluating whether democratically enacted laws violate fundamental principles. The Noble Person Test provides a more principled and more transparent basis for such review than the often opaque reasoning of constitutional courts [2].

3.4. Imbalance in Either Direction Is Unjust

An important implication of the four features is that justice requires balance: institutions biased toward the wealthy fail the test at the least advantaged position (Feature 2); institutions biased toward redistribution fail at the creator's position (Feature 1). Neither extreme is just.

Bias toward the wealthy. Noble Person at the position of the poor: basic needs unmet, no opportunity for advancement, no hope. The Noble Person is intellectually honest: they recognise that mass despair produces social instability, which harms everyone in the long run. Less costly alternatives (basic security, social mobility mechanisms) exist. Persuasion fails.

Bias toward redistribution. Noble Person at the creator's position: output confiscated, incentive destroyed. The Noble Person is intellectually honest: they recognise that destroyed incentives reduce total output, harming everyone including the least advantaged in the long run. Less costly alternatives (moderate taxation that funds public goods while preserving incentives) exist. Persuasion fails.

The acceptable range lies between these extremes. The precise location within the range is a matter for democratic deliberation and empirical research, and it shifts over time—during economic prosperity, more redistribution is sustainable; during economic difficulty, the balance may need adjustment. Feature 4 (error correction capacity) is what makes this continuous adjustment possible.

4. Applications to Law and Policy

This section applies the adversarial debate framework to three domains of persistent legal and policy controversy. In each case, the analysis follows the same structure: identify the positions, assign red and blue teams, apply the debate rules, and derive conditional conclusions. The purpose is to demonstrate that a single framework, applied consistently, produces nuanced and defensible conclusions across diverse domains—without requiring domain-specific principles.

4.1. Self-Defence and the Limits of Proportionality

4.1.1. The Problem

Legal systems universally recognise some right of self-defence, but struggle with proportionality. When a person facing a lethal attack kills their attacker, courts must determine whether the defensive force was "proportionate." In practice, this requires the defender to have calibrated their response with precision in a moment of mortal terror—a requirement that many legal scholars and practitioners regard as unrealistic [3].

 doi:10.20944/preprints202602.1980.v1

4.1.2. Adversarial Analysis

Red team (defender's position). I am being attacked with a deadly weapon. I cannot calculate the precise force needed to stop the attack without killing the attacker. My default self-interest is survival. If the law requires me to calibrate force with precision during a life-threatening attack, the consequence is that all defenders must hesitate, and some will die because of that hesitation. The long-term institutional consequence is the erosion of the effective right to self-defence.

Blue team (attacker's position, attempting to argue for strict proportionality). The attacker's life also has value. Unlimited defensive force could license disproportionate violence.

Red team response. The Noble Person at the attacker's position is intellectually honest: they acknowledge that they initiated the lethal threat. The risk of disproportionate response is a natural and foreseeable consequence of their own action. A legal regime that prohibits effective self-defence to protect attackers from the consequences of their own attacks produces worse long-term outcomes than one that permits robust self-defence: it incentivises attack (attackers know defenders are legally constrained) and disincentivises resistance (defenders fear legal liability). No less costly alternative protects both the effective right of self-defence and the attacker's interest: the attacker's interest is adequately protected by the option of not attacking.

Conclusion. Defensive force causing death should be presumptively justified when the defender faced a credible lethal threat. The law should not require precise force calibration under conditions where such calibration is impossible. This conclusion is derived not from a "right to self-defence" but from the Noble Person Test: the Noble Person at the defender's position cannot be persuaded to accept a legal regime that effectively penalises survival, and the Noble Person at the attacker's position—being intellectually honest—cannot deny that they created the threat.

4.2. Capital Punishment

4.2.1. The Problem

Capital punishment is among the most contentious issues in legal philosophy. Abolitionists argue that the state should never kill; retentionists argue that some crimes deserve death; consequentialists argue about deterrent effects. The debate has persisted for centuries without resolution [9].

4.2.2. Adversarial Analysis

Red team (position of the person sentenced to death). I do not want to die. Default refusal.

Blue team. Must demonstrate two things: (i) that not executing this person produces worse long-term consequences, and (ii) that no less costly alternative (such as life imprisonment) exists.

On condition (i). In the vast majority of cases, life imprisonment is sufficient to protect society. The person is removed from the public; the threat is neutralised. Not executing them does not produce worse consequences.

However, a narrow class of cases exists in which life imprisonment may be insufficient: individuals who continue to pose lethal threats even within prison—killing other inmates, organising violence from within, or posing escape risks that cannot be managed. For such individuals, the blue team can argue that not executing them produces worse consequences (continued deaths within the prison system) and that life imprisonment is not a less costly alternative because it does not eliminate the threat.

On condition (ii). The red team counters with the problem of irreversibility and error. The Noble Person at the position of a potentially wrongly convicted defendant raises a decisive objection: death is irreversible, and judicial systems make errors. If the error rate is non-trivial, capital punishment will inevitably kill innocent people, and no subsequent error correction is possible.

Blue team response. The risk of error can be reduced, though not eliminated, by extreme procedural safeguards: the highest evidence standard, mandatory multi-level appellate review, independent forensic review, fully funded defence counsel, and a requirement that guilt be established beyond any reasonable doubt by unanimous judicial panels.

 doi:10.20944/preprints202602.1980.v1

Conclusion. The Noble Person Test produces a conditional and highly restrictive conclusion:

(i) Capital punishment is not justified in the general case. Life imprisonment is a less costly alternative for the overwhelming majority of offenders.
(ii) Capital punishment may be justified in the narrow case of individuals who demonstrably continue to pose lethal threats that life imprisonment cannot neutralise.
(iii) Even in the narrow case, extreme procedural safeguards are required to reduce the risk of executing innocent persons to the lowest achievable level.
(iv) The institution must provide for posthumous exoneration and compensation to dependents if error is later discovered.

This conclusion will satisfy neither committed abolitionists nor committed retentionists. That is a feature of the framework, not a bug: it reflects the genuine complexity of the issue and avoids the false simplicity of absolute positions. The Noble Person Test does not resolve the debate; it *structures* it by identifying precisely where the boundary lies and what conditions must be met on each side.

4.3. Emergency Powers

4.3.1. The Problem

Emergencies—wars, pandemics, natural disasters—create pressure to expand government powers: restrict movement, commandeer property, suspend ordinary legal protections. History shows that such expansions, once granted, are difficult to reverse, and that governments have strong incentives to declare emergencies for political advantage [1]. The challenge is to permit necessary emergency action while preventing abuse.

4.3.2. Adversarial Analysis

Red team (position of a citizen whose rights are being curtailed). I do not consent to the suspension of my rights. Default refusal.

Blue team (government claiming emergency powers). Must demonstrate three things:

(i) **The emergency is real and imminent.** The Noble Person is intellectually honest: they can evaluate evidence of the emergency independently. A fabricated or exaggerated emergency fails at this first step. The blue team must present evidence that the red team can independently verify—classified intelligence that cannot be shared fails the test.
(ii) **The power expansion has a definite time limit.** Open-ended emergency powers cannot satisfy the Noble Person: an arrangement that might never be reversed is one in which a less costly alternative (the same powers with a sunset clause) always exists.
(iii) **No less costly alternative exists.** Can the emergency be addressed without suspending rights? If targeted measures (quarantine of affected areas rather than national lockdown; enhanced surveillance of specific threats rather than mass surveillance) can achieve the same purpose, the broader measure is unjustified.

Red team counter-arguments. Even if all three conditions are met, the red team raises two further concerns:

- **Precedent effects.** Each grant of emergency powers makes the next grant easier. The long-term institutional consequence of normalising emergency powers is the erosion of rights during non-emergency periods. The blue team must address this by showing that the specific safeguards in place (sunset clause, judicial review, legislative override) are sufficient to prevent normalisation.
- **Compensation.** Those who bear the costs of emergency measures (business owners forced to close, individuals detained, communities displaced) must be compensated. Failure to compensate undermines the legitimacy of future emergency measures: if people observe that those who bear emergency costs receive nothing, they will resist future emergency measures even when genuinely necessary.

Conclusion. Emergency powers are justified if and only if:

 doi:10.20944/preprints202602.1980.v1

(i) The emergency is real, imminent, and independently verifiable.
(ii) The powers have a definite and short time limit, subject to renewal only through the same adversarial process.
(iii) No less intrusive alternative can address the emergency.
(iv) Those bearing the costs are compensated.
(v) Institutional safeguards against normalisation are in place.

This framework provides a principled and operational basis for judicial review of emergency declarations—a function that courts currently perform with little theoretical guidance.

4.4. Common Pattern Across Applications

All three applications follow the same structure:

(1) Identify the positions, especially the most burdened.
(2) Red team defaults to refusal.
(3) Blue team bears the burden of proving necessity and absence of alternatives.
(4) Both teams argue about long-term consequences, including precedent effects.
(5) Conclusions are conditional and include safeguards.
(6) Compensation for those bearing costs is independently required.

No domain-specific principles are needed. The same framework, the same rules, the same argumentative structure produce nuanced conclusions across self-defence law, criminal punishment, and emergency powers. This universality is a strength: it means that institutional actors do not need to learn different frameworks for different policy domains. One set of debate rules covers all cases.

5. Institutional Mechanisms for Adversarial Review

5.1. The Proposal

This paper proposes that major institutional decisions involving significant sacrifice—defined as decisions that impose substantial costs on identifiable groups—be subject to structured adversarial review before adoption. The review mechanism consists of:

(i) **Mandatory red team assignment.** For any proposed policy involving significant sacrifice, an independent team is assigned to represent the interests of those bearing the costs. The red team has full access to relevant information, adequate funding, and institutional protection against retaliation.
(ii) **Structured debate.** The blue team (proposing the policy) and the red team engage in formal adversarial debate under the rules specified in Section 1. The debate is public, recorded, and subject to external scrutiny.
(iii) **Burden of proof.** The blue team bears the burden of demonstrating that the sacrifice is necessary (not sacrificing produces worse long-term consequences) and that no less costly alternative exists. If the blue team fails to discharge this burden, the default outcome (no sacrifice) prevails.
(iv) **Provisional conclusions.** Debate conclusions include explicit conditions, time limits, and review dates. No conclusion is permanent.
(v) **Periodic re-evaluation.** At specified intervals, the debate is reopened. New evidence, changed circumstances, or previously unconsidered alternatives may alter the conclusion.
(vi) **Compensation mechanism.** Any policy adopted through adversarial review must include a specific, funded compensation plan for those bearing the costs.

 doi:10.20944/preprints202602.1980.v1

5.2. Existing Analogues

The proposal extends practices already in use in several institutional domains.

5.2.1. Military Red-Teaming

The United States military has institutionalised red-teaming since the early 2000s, establishing the University of Foreign Military and Cultural Studies ("Red Team University") at Fort Leavenworth [12]. Red teams are tasked with challenging operational plans, identifying vulnerabilities, and proposing alternatives. Their institutional independence is protected: they report outside the chain of command of the planners whose work they are reviewing.

The military experience demonstrates both the value and the preconditions of effective red-teaming. Value: red teams consistently identify flaws that planning teams miss. Preconditions: the red team must have genuine independence, adequate resources, access to information, and institutional protection. When these preconditions are not met—when red teams are understaffed, denied information, or punished for unwelcome conclusions—the mechanism degrades into a formality [12].

5.2.2. Intelligence Analysis

The intelligence community's use of structured analytic techniques—including Analysis of Competing Hypotheses, devil's advocacy, and Team A/Team B exercises—is motivated by the same insight: unchallenged analysis is systematically biased [4]. Tetlock & Gardner [10] demonstrated that structured adversarial reasoning significantly improves forecasting accuracy compared to individual judgment.

5.2.3. Judicial Adversarial Procedure

Common-law judicial systems are built on adversarial procedure: prosecution and defence present competing cases, subject to rules of evidence, before an impartial fact-finder. The analogy to the proposed mechanism is direct: the red team is the defence, the blue team is the prosecution, the debate rules correspond to rules of evidence, and the burden of proof is on the prosecution.

The judicial analogy also illustrates a crucial institutional requirement: the defence must be adequately resourced. A criminal trial in which the defendant has no lawyer is widely recognised as unjust, not because of an abstract right to counsel, but because the adversarial mechanism cannot function if one side lacks the capacity to argue effectively. The same logic applies to the proposed adversarial review: the red team must have resources comparable to the blue team's.

5.3. Where Should Adversarial Review Apply?

Not every institutional decision requires formal adversarial review. The mechanism is designed for decisions involving *significant sacrifice*—decisions that impose substantial costs on identifiable groups. Examples include:

- Major tax reform
- Decisions to go to war or use military force
- Declaration of emergency powers
- Criminal sentencing policy (especially capital punishment)
- Large-scale infrastructure projects requiring displacement
- Significant changes to social welfare systems
- Public health measures restricting liberty (quarantine, lockdown)

For routine decisions that do not involve significant sacrifice, the ordinary democratic process is sufficient. The adversarial mechanism supplements democratic governance for high-stakes decisions; it does not replace it for everyday policy-making.

 doi:10.20944/preprints202602.1980.v1

5.4. Institutional Independence of the Red Team

The single most important design requirement is the **independence of the red team**. If the red team is appointed by, funded by, or answerable to the same authority proposing the sacrifice, the adversarial mechanism is compromised.

Institutional independence requires:

(i) **Separate funding.** The red team's budget must not be controlled by the blue team or the decision-making authority.

(ii) **Separate appointment.** Red team members must be appointed through a process independent of the proposing authority—for example, by an independent oversight body, a judicial appointment committee, or a random selection mechanism.

(iii) **Protection against retaliation.** Red team members must be legally protected against retaliation for their conclusions, regardless of how unwelcome those conclusions may be.

(iv) **Access to information.** The red team must have the same access to relevant information as the blue team. Information asymmetry invalidates the debate (recall that the Noble Person's intellectual honesty requires independently verifiable facts).

(v) **Public transparency.** Debate proceedings and conclusions must be publicly available, with limited exceptions for genuinely sensitive national security information (and even these exceptions must be subject to independent judicial review).

These requirements are demanding but not unprecedented. Independent auditors, ombudsmen, inspectors general, and constitutional courts already operate with similar institutional protections in many democracies. The proposal extends this logic to a broader class of institutional decisions.

5.5. Failure Modes

The adversarial mechanism can fail in several ways:

(i) **Capture.** The red team is co-opted by the blue team or the proposing authority. *Mitigation*: separate appointment, funding, and legal protection.

(ii) **Formalism.** The debate is conducted as a formality with the conclusion predetermined. *Mitigation*: public transparency, external scrutiny, media coverage.

(iii) **Information asymmetry.** The blue team withholds relevant information. *Mitigation*: legal requirements for disclosure, penalties for withholding, independent verification mechanisms.

(iv) **Capacity imbalance.** The red team lacks the expertise or resources to mount effective arguments. *Mitigation*: adequate and protected funding, ability to retain independent experts.

(v) **Political override.** The decision-making authority ignores the debate's conclusion. *Mitigation*: legal weight attached to conclusions (e.g., requiring a supermajority to override a red team objection), judicial review.

No institutional mechanism is immune to all failure modes. The adversarial mechanism is not proposed as incorruptible but as *more resistant to failure than the alternative*—which is decision-making without structured adversarial challenge. The military and judicial experience confirms that adversarial mechanisms, despite their imperfections, systematically outperform unchallenged decision-making [10,12].

6. Objections and Replies

6.1. "This Would Paralyse Decision-Making"

The objection assumes that every decision would be subject to full adversarial review. It would not. The mechanism applies only to decisions involving *significant sacrifice*—a threshold that excludes routine governance. Within democracies, constitutional judicial review already imposes procedural requirements on legislation without paralysing the legislative process. The proposed mechanism is analogous: it adds a procedural requirement for high-stakes decisions, not for all decisions.

 doi:10.20944/preprints202602.1980.v1

Moreover, the adversarial process has a defined timeline. In genuine emergencies, expedited procedures can be employed—with the requirement that a full adversarial review follows within a specified period. This is analogous to the common legal practice of granting temporary injunctions pending full hearing.

6.2. "Who Decides What Counts as Significant Sacrifice?"

This is a genuine boundary problem. The paper proposes a functional criterion: a decision involves significant sacrifice if it imposes costs that the Noble Person at the cost-bearing position would, by default, refuse. In practice, this can be operationalised through threshold criteria: magnitude of cost (financial, liberty, or physical), number of people affected, and reversibility. The precise thresholds are themselves subject to democratic deliberation and periodic revision.

Boundary disputes are unavoidable in any institutional design. The relevant comparison is not between the proposed mechanism (which has boundary disputes) and an ideal mechanism (which does not), but between the proposed mechanism and the status quo (which has no principled threshold at all).

6.3. "The Red Team Cannot Truly Represent the Affected"

The red team represents the *Noble Person at the affected position*, not the actual affected individuals. This is a limitation: actual affected individuals may have concerns, knowledge, or perspectives that the red team does not capture.

The limitation is mitigated by several features. First, adversarial review is *supplementary* to democratic participation, not a substitute for it. Affected individuals can still vote, petition, organise, and protest. Second, the debate is public, allowing affected individuals to observe and challenge the red team's arguments. Third, the red team can and should consult affected communities in preparing its arguments.

But the objection has force, and the paper acknowledges it. The red team is a proxy, and proxies are imperfect. The adversarial mechanism is proposed as an *improvement* over unchallenged decision-making, not as a perfect representation of all affected interests.

6.4. "This Is Just Judicial Review Under a Different Name"

The mechanism shares features with judicial review but differs in important respects.

First, judicial review is *reactive*: it evaluates laws after they are enacted. The proposed mechanism is *proactive*: it evaluates policies before adoption.

Second, judicial review applies constitutional standards—often vague and subject to competing interpretations ("due process," "equal protection"). The proposed mechanism applies a specific and operational standard: can the Noble Person at the most burdened position be persuaded, given verifiable evidence, that not sacrificing produces worse long-term consequences and no less costly alternative exists?

Third, judicial review is conducted by judges with legal training. The proposed mechanism involves teams with diverse expertise—economic, social, technical—appropriate to the specific policy under review.

The mechanisms are complementary, not competing. Adversarial review before adoption and judicial review after enactment provide two layers of protection.

6.5. "In Non-Democratic Societies, This Mechanism Cannot Function"

This is correct, and the paper acknowledges it as a *scope condition* rather than a defect. The adversarial mechanism requires preconditions: freedom of speech, access to information, institutional independence, protection against retaliation. In societies where these conditions are absent, the mechanism cannot function effectively.

This scope condition is not unique to the proposed mechanism. *All* deliberative democratic procedures, including judicial review, parliamentary debate, and free elections, require similar pre-

 doi:10.20944/preprints202602.1980.v1

conditions. The Noble Person Test identifies these preconditions not as arbitrary requirements but as *logical consequences* of the framework itself: the Noble Person at any position within an institution lacking these conditions will refuse, because an institution with these conditions is always a less costly alternative (Feature 4, error correction capacity).

The implication for non-democratic societies is not that the framework is irrelevant but that the framework identifies what they lack: the conditions under which institutional justice is possible.

7. Conclusion

This paper has argued that the Noble Person Test—a framework for evaluating institutional justice proposed in a companion paper—is best operationalised as structured adversarial debate. The argument proceeds in four steps.

First, individual application of the Noble Person Test is subject to cognitive bias; adversarial debate mitigates this by assigning opposing positions to opposing teams and imposing fixed rules of engagement.

Second, the Noble Person Test, applied across institutional positions, yields four features that just institutions must possess: protection of creative incentives, basic security for all, social mobility, and error correction capacity. These features are derived from a single source—the inability to persuade the Noble Person at specific positions—and jointly define a dynamic equilibrium that institutions must continuously maintain.

Third, the adversarial framework produces nuanced, conditional, and operationally useful conclusions across diverse policy domains—self-defence law, capital punishment, emergency powers—without requiring domain-specific principles. The same rules, applied to different cases, yield appropriately different conclusions.

Fourth, the proposed mechanism is not utopian. It extends practices already proven in military red-teaming, intelligence analysis, and judicial adversarial procedure. Its institutional requirements—independence, funding, information access, transparency, protection against retaliation—are demanding but have precedent in existing democratic institutions.

The paper acknowledges significant limitations. The mechanism applies only to high-stakes decisions involving significant sacrifice; it supplements rather than replaces democratic governance; it requires preconditions (freedom of speech, institutional independence) that are absent in many societies; and the red team is an imperfect proxy for actual affected individuals. These limitations define the scope and the ambition of the proposal: not a complete theory of just governance, but one mechanism—grounded in a specific normative framework and amenable to institutional implementation—for improving the justice of high-stakes institutional decisions.

The relationship between justice and institutional design has received less attention in political philosophy than the relationship between justice and abstract principle. This paper is an attempt to redress that imbalance. If justice is not only a truth to be discovered but a structure to be built, then the tools of construction—procedures, mechanisms, debate rules—deserve the same philosophical scrutiny as the blueprints.

References

1. Agamben, G. (2005). *State of Exception*. Chicago: University of Chicago Press.
2. Estlund, D. (2008). *Democratic Authority: A Philosophical Framework*. Princeton: Princeton University Press.
3. Fletcher, G. P. (1998). *Basic Concepts of Criminal Law*. New York: Oxford University Press.
4. Heuer, R. J. (1999). *Psychology of Intelligence Analysis*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.
5. Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
6. Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.
7. Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.
8. Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

 doi:10.20944/preprints202602.1980.v1

9. Sunstein, C. R. & Vermeule, A. (2005). Is capital punishment morally required? Acts, omissions, and life-life tradeoffs. *Stanford Law Review*, 58(3), 703–750.
10. Tetlock, P. E. & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York: Crown.
11. Tocqueville, A. de. (1835). *Democracy in America* (Vol. 1). London: Saunders and Otley.
12. Zenko, M. (2015). *Red Team: How to Succeed by Thinking Like the Enemy*. New York: Basic Books.