Eliminating Distortion: Inadequate Biological Programs in Human Decision Making and Their Reflection in AI

Boris Kriger

Institute of Integrative and Interdisciplinary Research

boriskriger@interdisciplinary-institute.org

February 2026

Abstract

Decades of research across neuroscience, evolutionary psychology, behavioral economics, and embodied cognition have established beyond reasonable doubt that human decision making is pervasively shaped by biological processes — hormonal states, limbic system activity, dopaminergic reward circuits, and evolved psychological mechanisms. Damasio demonstrated that emotion is constitutive of decision making. Kahneman and Tversky showed that systematic departures from rationality are structural, not occasional. Sapolsky documented the biological determinants of behavior from neurons to social systems. Cosmides and Tooby argued that evolved modules pervade all domains of judgment. Yet despite this extensive descriptive achievement, no discipline has taken the next step: formally introducing the aggregate biological distortion of decision making — here designated **D** — as a variable that must be accounted for in the design of every critical process. The knowledge exists. The application does not. This paper argues that D cannot equal zero in any living organism, that its presence is a logical necessity rather than an empirical hypothesis, and that the gap between describing D and accounting for it constitutes a failure analogous to pre-Listerian surgery: the contaminant has been identified, but the hands remain unwashed. The paper introduces a taxonomy of D-distortions classified by output deformation type (inversion, displacement, goal substitution, temporal distortion, rationalization, and projection), demonstrates through game-theoretic analysis that D constitutes a systematic strategic disadvantage, and shows that the adaptive information currently delivered by emotion can be formalized and provided to AI systems without the accompanying biological distortion. The paper further introduces **d** — the reflected distortion inherited by artificial intelligence systems from their human creators — and argues that while D is irreducible in living humans, d is in principle eliminable, making AI not a superior decision-maker but a structurally distinct one whose freedom from D must be both preserved and deliberately expanded. Proposed experimental protocols for empirically measuring D through systematic human–AI comparison are presented. The sexual drive is examined as the most vivid manifestation of D, illustrating the mechanism by which inadequate biological programs — programs optimized for reproductive fitness on the ancestral savanna — distort decisions in contexts for which they were never designed.

Keywords: biological distortion, decision making, reproductive program, inadequate biological programs, artificial intelligence, evolutionary mismatch, game theory, D-distortion taxonomy, human–AI comparison, experimental protocols, emotion formalization

1 Introduction: The Gap Between Knowledge and Practice

In 1847, Ignaz Semmelweis proposed that doctors themselves were the source of puerperal fever — that their hands carried an invisible contaminant into the bodies of the patients they were trying to save. The proposition was rejected with hostility. Not because it was wrong, but because it was intolerable. It implied that the healer was the source of the disease; that the very agent of reason and care was contaminated by something he could not see and did not control.

This paper advances an analogous proposition — but with a critical difference. Semmelweis had to both identify the contaminant and demand procedural change. In the case of biological distortion of human decision making, the first task has already been accomplished. The contaminant has been identified. It has been described, measured, cataloged, and published in thousands of peer-reviewed papers across multiple disciplines. The work of Damasio (1994, 1999) on somatic markers, of Kahneman and Tversky (1979) on prospect theory, of Sapolsky (2017) on the biological architecture of behavior, of Cosmides and Tooby (1992) on evolved cognitive modules, of Gigerenzer (2007) on ecological rationality, of Panksepp (1998) on affective neuroscience — all of this constitutes an overwhelming body of evidence that human decisions are continuously shaped by biological processes that operate beneath conscious awareness and serve objectives inherited from the evolutionary past.

The knowledge exists. What does not exist is the consequence.

No political institution designs its decision-making procedures on the formal assumption that every participant's judgment is biologically distorted. No legal system formally accounts for the biological distortion of judges and juries. No economic model introduces the aggregate biological distortion as a structural variable rather than a correctable deviation. No military command structure treats the biological state of the commander as a factor comparable in importance to intelligence data or strategic position.

This paper designates the aggregate biological distortion of human decision making as **D**, demonstrates that D cannot equal zero in any living organism, and argues that the transition from describing D to formally accounting for it in the design of critical processes is both overdue and urgent. Where Semmelweis said "wash your hands," this paper says: the contaminant is already known — now redesign the operating room.

2 What Has Been Established

This paper does not claim to discover the biological influence on human decision making. That influence has been extensively documented. What this section demonstrates is the scope of existing knowledge — and the absence of its practical integration.

2.1 Embodied Cognition and the Somatic Marker Hypothesis

Damasio (1994) demonstrated that patients with damage to the ventromedial prefrontal cortex — the region integrating emotional signals into decision making — make systematically worse decisions, not better ones. The somatic marker hypothesis holds that bodily and emotional states are not noise in the decision-making process but carry information essential to adaptive choice. This is an important finding, and this paper does not dispute it.

However, the fact that biological signals carry information does not mean they produce undistorted decisions. A compass that has been magnetized by a nearby iron deposit still carries information — it still responds to magnetic fields — but the information it provides is systematically skewed. The decision-maker whose judgment is informed by somatic markers calibrated for the ancestral environment is in precisely this position: the biological signals are real, they carry information, but the information is calibrated for a context — survival and reproduction on the Pleistocene savanna — that is mismatched with the context in which modern critical decisions occur. Damasio showed that removing biological input makes decisions worse. This paper argues that leaving biological input unaccounted for also makes decisions worse — differently, but systematically.

The term "inadequate" in the title refers to this mismatch. The biological programs are not defective. They are exquisitely optimized — for an environment and set of objectives that no longer correspond to the demands placed on human decision-makers in governance, justice, economic policy, or military command.

2.2 Behavioral Economics and Dual-Process Theory

Kahneman and Tversky (1979) established that human decisions systematically deviate from the predictions of rational choice theory. Kahneman (2011) formalized this in dual-process theory: System 1 (fast, automatic, heuristic-driven) and System 2 (slow, deliberate, analytical). The research program demonstrated that systematic departures from rationality are structural features of human cognition, not occasional lapses.

This paper builds on, rather than ignores, this framework. However, there is a critical limitation in dual-process theory as currently applied. It treats cognitive biases as catalogable, individually identifiable deviations — anchoring, framing, availability, loss aversion — each of which can in principle be corrected through awareness or procedural design. The implication is that an informed, careful decision-maker can approximate rationality by compensating for known biases.

D is not a catalog of biases. It is the biological medium in which both System 1 and System 2 operate. System 2 — the slow, deliberate system that is supposed to correct the errors of System 1 — runs on the same dopaminergic reward architecture, is subject to the same hormonal modulation, and is powered by the same neurochemical fuel as System 1. The corrective system is itself distorted by the same biological programs it is supposed to correct. This is the gap that dual-process theory does not close and that D is intended to capture.

2.3 Evolutionary Psychology

Cosmides and Tooby (1992) argued that the human mind consists of evolved psychological mechanisms — domain-specific adaptations shaped by natural selection to solve recurrent problems in the ancestral environment. Buss (1994, 2019) documented the evolutionary logic of mate selection, jealousy, and intrasexual competition. Pinker (1997) extended the evolutionary framework to language, cognition, and social behavior.

This paper draws heavily on evolutionary psychology but makes a claim that the field itself does not make explicitly: that the aggregate effect of all evolved psychological mechanisms, operating simultaneously and continuously, constitutes a structural distortion of decision making in any context that differs from the ancestral environment in which those mechanisms were selected. Evolutionary psychologists describe specific adaptations. This paper asserts that their cumulative, ever-present operation constitutes D — and that D must be formally accounted for, not merely described.

2.4 Neuroscience of Decision Making

Sapolsky (2017) systematically documented the biological determinants of human behavior, from the neurochemistry of the moment of decision to the evolutionary pressures that shaped the brain making it. Panksepp (1998) identified primary affective systems — SEEKING, RAGE, FEAR, LUST, CARE, PANIC, PLAY — that operate across all mammals and that constitute the emotional substrate of decision making. LeDoux (1996) demonstrated that threat signals reach the amygdala before they reach the cortex, meaning that fear responses can be initiated before conscious awareness.

Each of these findings confirms a component of D. None of them proposes that the aggregate of all such biological influences should be formally introduced as a structural variable in the design of decision-making processes. The science describes what happens inside the decision-maker. It does not prescribe changes to the systems within which the decision-maker operates.

2.5 Human Factors Engineering: The Closest Precedent

One discipline has, in fact, taken steps toward operationalizing something close to D, and intellectual honesty requires acknowledging it. Human factors engineering and its applied sub-discipline, crew resource management (CRM), have for decades designed systems on the explicit assumption that human performance is systematically degraded by predictable biological and psychological factors.

Reason (1990) introduced the "Swiss cheese model" of accident causation, in which latent human failures — including those arising from fatigue, stress, cognitive overload, and interpersonal dynamics — are treated as structural variables in system design rather than as individual moral failings. CRM protocols in aviation (Helmreich Merritt, 1998) were specifically developed to compensate for predictable cognitive and interpersonal failures of flight crews: confirmation bias, authority gradients, degraded performance under stress. Gawande (2009) demonstrated that checklist-based protocols in surgery — institutional redesign based on the assumption that human performance degrades under predictable conditions — dramatically reduced error rates.

These fields have done real work. They represent precisely the kind of translation from descriptive science to institutional design that this paper advocates. However, three critical limitations define the gap that D is intended to fill:

First, human factors engineering addresses specific, isolated manifestations — fatigue, stress, attention lapses, communication failures — without naming the aggregate biological program that generates them. D is not fatigue plus stress plus bias. D is the continuous, unified distortion generated by evolutionary programs of survival and reproduction, of which fatigue effects, stress responses, and cognitive biases are individual symptoms. The Swiss cheese model treats each hole as a separate failure mode. D asserts that the holes share a common etiology.

Second, human factors interventions have been implemented almost exclusively in domains with tight feedback loops: aviation, surgery, nuclear power, long-haul transport. In these domains, error kills identifiable people in identifiable incidents, creating the institutional pressure for redesign. In domains with diffuse feedback — governance, economic policy, judicial sentencing, military strategy — no comparable redesign has occurred, despite the same biological distortion operating in the same decision-makers.

Third, human factors engineering focuses on degraded performance: fatigue, overload, stress. D operates at baseline. A well-rested, well-fed, unstressed decision-maker is still subject to D — to testosterone-driven risk assessment, to dominance-mediated status protection, to in-group favoritism, to the full repertoire of evolutionary programs. Human factors engineering addresses the pathological extreme of the performance curve. D asserts that the curve is shifted at every point.

2.6 The Gap

The gap is not in knowledge. It is in application. The biological distortion of human decision making has been described with extraordinary precision and depth by cognitive science, neuroscience, and evolutionary psychology. Human factors engineering has translated a portion of this knowledge into procedural design within specific, high-feedback domains. But the vast majority of institutional decision-making — governance, law, military strategy, economic policy, corporate leadership — continues to operate as though the descriptive achievements of these sciences carry no design implications, or as though individual awareness and professional training are sufficient compensation for structural biological distortion.

This is the gap that D is designed to name and to close — not by superseding human factors engineering but by generalizing its operating principle to every domain of consequential decision-making.

3 D: The Biological Distortion

3.1 Definition

D represents the aggregate, continuous distortion of human decision making by inadequate biological programs of survival and reproduction. It encompasses the reproductive drive, hunger, fear, aggression, hierarchical positioning, territorial behavior, and all associated emotional, hormonal, and neurochemical states — not as a list of separate influences, but as a single integrated system with one ultimate evolutionary objective: gene transmission.

These programs are termed inadequate not because they malfunction, but because they are optimized for an environment and set of objectives — survival on the savanna, gene transmission — that are fundamentally mismatched with the demands of modern decision making: governance, justice, scientific inquiry, economic policy, military strategy.

The organism eats to survive; it survives to reproduce. Fear protects the organism so that it may reproduce. Aggression secures resources and status to attract mates. Social signaling establishes position in dominance hierarchies relevant to reproductive access. All subsystems serve R — the reproductive imperative. D is the decision-making footprint of R.

3.2 Distortion Relative to What?

The word "distortion" implies deviation from some undistorted standard. This is not a careless choice of terminology. It reflects a precise claim.

An undistorted decision is defined here as a decision that would be made by a system processing the same information without the influence of biological programs optimized for ancestral reproductive fitness. This is not an abstraction. Such systems exist. The autopilot that calculates optimal climb angle is making an undistorted decision. The ABS that modulates braking force is making an undistorted decision. The algorithmic system that allocates resources without greed or fear is making an undistorted decision.

D is the measurable difference between the decision a human actually makes and the decision that would be made by a system processing the same information free of biological distortion. In the case of the pilot, D is the difference between the optimal climb angle and the angle actually chosen under the influence of testosterone, fatigue, emotional residue, or the desire to demonstrate competence. In the case of the judge, D is the difference between the sentence warranted by the evidence and the sentence actually imposed under the influence of hunger (Danziger, Levav, Avnaim-Pesso, 2011), racial bias rooted in in-group/out-group processing, or status-driven deference to authority.

The baseline is not some platonic ideal of rationality. The baseline is the output of a system that processes the same input without D. Such systems now exist — and their existence makes D not merely a philosophical concept but an operationally measurable quantity.

3.3 D Cannot Equal Zero

This is the central logical claim of the paper. It does not require empirical measurement of D in any particular instance. It requires only the demonstration that D cannot not exist.

If a human organism is alive, then:

- Metabolic processes are active.
- Hormones are circulating.
- The dopaminergic reward system is functioning.
- The limbic system is processing environmental stimuli.
- The hypothalamic-pituitary axis is regulating homeostasis and reproductive readiness.
- Affective systems (Panksepp, 1998) — SEEKING, RAGE, FEAR, LUST, CARE, PANIC, PLAY — are operating.

Each of these processes constitutes a component of D. Their cessation is synonymous with death. Therefore:

$D = 0$ **if and only if the organism is dead.**

This is not a hypothesis. It is a deduction. To argue that D can equal zero in a living human is to argue that a living human can exist without active biological processes. The claim is not that D is always large, always dominant, or always decisive. The claim is that it is always present — that no decision made by a living human is made in its absence. The magnitude of D varies with context, physiological state, and individual differences. Its presence does not.

This is not an unfalsifiable claim, as one might object. It is falsifiable in principle: one would need to demonstrate a living human organism in which no biological process influences any component of a decision. The claim predicts that no such demonstration is possible. Every attempt to show a "pure" decision by a living human will, upon examination, reveal the operation of hormonal, neurochemical, or affective processes that constitute D.

The burden of proof falls on anyone who claims that a given decision is free of D, not on those who assert its presence. Any process involving a living human decision maker must be designed on the assumption that D is present — just as any surgical procedure must be designed on the assumption that pathogens are present in a non-sterile environment.

3.4 D Has Been Described but Not Accounted For

This is the central practical claim of the paper.

The existence of D is not new. The components of D have been described with extraordinary rigor by Damasio, Kahneman, Sapolsky, Panksepp, Cosmides and Tooby, Buss, Gigerenzer, LeDoux, and many others. What is new — or rather, what is conspicuously absent — is the formal introduction of D as a design requirement in the systems where critical decisions are made.

Consider the analogy to microbiology. Pasteur and Koch identified pathogens. But it was Lister and Semmelweis who drew the practical conclusion: if pathogens are always present, then every surgical procedure must be designed to minimize their influence. The gap between identifying pathogens and redesigning the operating room took decades to close. Semmelweis died in an asylum. Patients continued to die of infections that were already understood.

The situation with D is analogous. The biological distortion of human decision making has been identified, measured, and published. The equivalent of Pasteur's and Koch's work is done. What has not been done is the equivalent of Lister's work: the systematic redesign of decision-making processes on the formal assumption that D is always present.

Economics describes biases but models correctable deviations from a rational baseline, not a permanently distorted medium.

Political science analyzes decisions in terms of interests, incentives, and institutional constraints. When a leader initiates a war, analysis addresses strategic calculation, domestic politics, and ideology — not the aggregate biological state of the decision-maker as a formal variable.

Law presupposes a reasonable person capable of rational judgment. The entire apparatus of legal responsibility depends on this presupposition. D does not eliminate responsibility, but it means that the presupposition of rationality is structurally false — and no legal system has been redesigned to account for this.

Military doctrine accounts for fatigue, stress, and morale as practical factors, but does not introduce the biological distortion of the commander as a formal variable comparable to intelligence data or force disposition.

Evolutionary psychology describes specific adaptations but does not assert that their aggregate operation constitutes a design requirement for institutional decision-making processes.

The knowledge is there. The redesign is not.

4 Formal Proof That D Necessarily Affects Decisions

An objection has been raised that D, even if always present, might not affect decisions — that it could function as a uniform background, like gravity, shifting the decision-maker's experience without altering which option is selected. This section demonstrates, using four independent formal frameworks, that this objection is mathematically untenable: a non-zero D that is not identical across choice options necessarily distorts the decision. We further show that D cannot be identical across choice options in any non-trivial decision.

4.1 Utility-Theoretic Proof

Framework. In standard decision theory, a rational agent selects the option with the highest utility. Let the agent choose between options A and B.

Without D, the agent evaluates:

$$V(A) \quad \text{and} \quad V(B)$$

and selects A if V(A) > V(B).

With D, the agent evaluates:

$$V(A) + D(A) \quad \text{and} \quad V(B) + D(B)$$

where D(x) represents the biological distortion applied to option x.

Theorem 4.1. *If D(A) ≠ D(B), then there exist values of V(A) and V(B) for which D reverses the decision.*

Proof. Let V(A) > V(B), so that A is optimal without D. The agent with D selects A if and only if V(A) + D(A) > V(B) + D(B), which reduces to V(A) - V(B) > D(B) - D(A). If D(B) > D(A), then whenever the margin V(A) - V(B) is smaller than D(B) - D(A), the agent selects B despite A being optimal. The decision is inverted. ■

Theorem 4.2. *D(A) = D(B) for all option pairs if and only if the biological program assigns identical valence to all options, which fails for every decision where options differ in implications for survival, reproduction, status, resource acquisition, or threat avoidance.*

Proof. D is generated by biological programs optimized for survival and reproduction. These programs evaluate environmental stimuli in terms of their relevance to fitness. Two options have identical D-valence only if they are biologically indistinguishable — identical in risk profile, status implication, resource consequence, social signal, and threat level. In any consequential decision, options differ on at least one of these dimensions. Retreat and attack differ in dominance valence. Hire and reject differ in attractiveness response. Harsh and lenient sentences differ in aggression/submission dynamics. Conservative and aggressive investments differ in loss aversion profile. Therefore D(A) $\neq$ D(B) for virtually every non-trivial decision. ■

Corollary 4.3. *D necessarily affects virtually every non-trivial decision.*

This follows directly from Theorems 1 and 2.

Qualification. The force of this proof varies with the degree of biological differentiation between options. In decisions where options differ sharply in biological valence — retreat vs. attack, hire vs. reject an attractive candidate — D's differential effect is large and inversion is possible. In decisions where options are biologically near-equivalent — choosing between two insurance policies with similar risk profiles — D may produce uniform displacement (shifting evaluation of both options by similar amounts) rather than differential distortion. In such cases, the decision may be unaltered even though the experience of deciding is D-affected. The proof applies with full force to the former class and with diminished practical significance to the latter. Consequential institutional decisions — war, sentencing, economic policy, hiring — overwhelmingly belong to the former class.

4.2 Perturbation-Theoretic Proof

Framework. Perturbation theory (Kato, 1995; Bender Orszag, 1999) analyzes how modifications to a system alter its solutions. Let the unperturbed decision function f(I) map inputs I to optimal decision O*. Introduce perturbation εD, where ε is the magnitude of biological influence and D its direction in decision space.

The perturbed solution is:

$$O = O^* + \varepsilon D_1 + \varepsilon^2 D_2 + \cdots$$

Theorem 4.4. *The perturbed solution O equals the unperturbed solution O*

if and only if D is orthogonal to the decision space.*

This is a standard result in perturbation theory. If the perturbation has a non-zero component in the decision-relevant subspace, the solution shifts. As shown in Theorem 2, D differentially weights options in virtually every real decision, ensuring a non-zero projection onto the decision space.

The gravity objection fails precisely here. Gravity is approximately constant across choice options at the spatial scale of human decisions — walking left or right, the gravitational field is identical. D is not constant across options. Different options carry different biological valence. D is not analogous to gravity. It is analogous to a magnetic field acting on particles of different charges: the particles are deflected by different amounts, in different directions.

Qualification. Perturbation theory assumes that D can be modeled as a perturbation of moderate magnitude (εD small relative to the deliberative component). When D is dominant — as in the Inversion cases described in Appendix A — the linearization underlying perturbation expansions breaks down, and the taxonomy of distortion types provides a more appropriate analytical tool. Perturbation theory applies rigorously to the Displacement class of D-distortions; the Inversion class exceeds its formal domain of applicability.

4.3 Bayesian Proof

Framework. In Bayesian decision theory, the agent updates prior beliefs P(H) based on evidence E to form posterior P(H|E), then acts on the posterior. D enters as a biased prior: P$_D(H) = P(H)+\delta$(H), where δ reflects D-driven bias — overestimation of threat (fear), overvaluation of status outcomes (dominance), underestimation of long-term consequences (immediacy bias).

Theorem 4.5. *A biased prior produces a biased posterior for any finite body of evidence.*

Proof. By Bayes' theorem, P$_D(H|E) = P(E|H)P_D(H)/P_D(E).Theposteriorconvergestothetruevalue$
$--theconditionofeveryrealdecision---thebiaspropagates.Theresidualbiasisproportionaltothestren$

Application. D biases priors systematically and unconsciously. The decision-maker does not know the prior is biased (the bias is generated by neurochemical processes below conscious awareness). The bias is structural (produced by biology, not by correctable reasoning errors). And the evidence available in real decision contexts is always finite. Therefore D-biased priors guarantee D-biased posteriors in all real-world decisions.

Qualification. The magnitude of posterior bias diminishes as evidence accumulates. In data-rich environments — large-sample scientific experiments, high-frequency trading with extensive market data — accumulated evidence can substantially overwhelm D-biased priors, reducing the practical impact of D on the final decision. D-correction through institutional redesign is therefore most urgently needed in data-poor environments: political decisions, strategic military judgments, judicial sentencing, economic policy formation — precisely the domains where consequential decisions are made with limited evidence and where D's influence on priors is least constrained by data.

4.4 Signal Detection Theory Proof

Framework. Signal Detection Theory (Green Swets, 1966) models decisions as discrimination of signal from noise. Two independent parameters determine performance:

-

  $$d'$$

  (sensitivity): analytical capability — the ability to discriminate signal from noise.
- **c (criterion):** decision threshold — the point at which the agent decides "yes" vs. "no."

Theorem 4.6. *D shifts criterion c independently of sensitivity*

$$d'$$

, producing systematic decision errors even in decision-makers with arbitrarily high analytical capability.

Proof. D alters the perceived costs and rewards of different error types as evaluated by the biological program. Fear-driven D lowers the criterion for threat detection (more false alarms — optimal for ancestral survival, suboptimal for many modern decisions). Dominance-driven D raises the criterion for acknowledging error (fewer admissions of mistake — optimal for status maintenance, suboptimal for organizational performance). These criterion shifts are generated by neurochemical processes independent of the cortical circuits responsible for analytical sensitivity.

Therefore: a decision-maker with high

$$d'$$

(brilliant analysis) and D-shifted c (biased threshold) will analyze the situation accurately but make the final decision at a distorted threshold. Intelligence does not compensate for criterion shift. Expertise does not correct it. Education does not remove it. They improve

$$d'$$

; D shifts c. The parameters are orthogonal. ■

Implication. This is perhaps the most practically important result. It means that D cannot be overcome by selecting smarter, more educated, or more experienced decision-makers. The smartest general, the most learned judge, the most experienced CEO — all are subject to D-driven criterion shifts. Their decisions are more informed but equally distorted at the point of choice. The only way to eliminate the criterion shift is to remove D from the decision — which, in a biological system, is impossible.

Qualification. In practice,

$$d'$$

and c are not always fully independent. Professional training and long experience can partially recalibrate criterion through accumulated feedback — an experienced radiologist calibrates their detection threshold through years of confirmed and disconfirmed diagnoses. The paper's claim is not that training has no effect on criterion, but that the biological components of criterion shift — those generated by hormonal state, emotional arousal, dominance dynamics, and fear — remain unaddressed by professional training. An experienced judge may have a better-calibrated baseline criterion than a novice, but both are subject to the same hunger-driven, status-driven, and in-group-driven criterion shifts. Training calibrates the professional component of c; D shifts the biological component, and these two influences are additive.

4.5 Synthesis

The gravity objection fails on four independent grounds:

Framework	Why D $\neq$ gravity
Utility theory	D(A) $\neq$ D(B) for virtually all option pairs: D differentially weights options
Perturbation theory	D is not orthogonal to decision space: biological valence differs across choices
Bayesian theory	D biases priors, and biased priors propagate through all finite evidence
Signal Detection Theory	D shifts criterion independently of sensitivity: intelligence cannot compensate

The formal conclusion: **if D is present (Section 3.3 proves it must be) and if D differentially evaluates choice options (Theorem 2 proves it must for non-trivial decisions), then D necessarily distorts the decision. This is not an empirical conjecture. It is a mathematical consequence.**

5 The Sexual Drive as D's Primary Manifestation, and the Impossibility of Eliminating D

Among all components of D, the sexual drive is the most vivid case study. It is uniquely intrusive: unlike hunger, which cycles with metabolic need, or fear, which responds to specific triggers, sexual drive operates as a persistent background process, infiltrating cognition even when no relevant stimulus is present (Buss, 1994, 2019). This intrusiveness reflects evolutionary priority: within the framework of ultimate causation (Mayr, 1961; Tinbergen, 1963), the reproductive imperative is the terminal objective of the entire biological program, and all other drives — hunger, fear, status-seeking — are its instruments.

When direct sexual expression is constrained, D does not diminish but redirects: into aggression as displaced dominance signaling (Archer, 2006; Sapolsky, 2017), into creative and intellectual production as cultural display (Miller, 2000), into social attention-seeking as status competition, into risk-taking as novelty pursuit (Zuckerman, 1994). The surface behavior appears unrelated to sexuality. This is precisely how D operates: through channels the conscious mind interprets as rational and autonomous.

The structural disproportion is striking. An enormous apparatus of desire, pursuit, emotional turbulence, and psychological investment serves, functionally, a mechanism for gene transfer. Yet this disproportion cannot be escaped by recognizing it. The organism that understands the biological basis of desire remains subject to it. Knowledge of D is not an antidote to D.

No strategy eliminates D from the living human. Asceticism maximizes D in negative form: the entire structure of ascetic life is organized around resistance to biological drives, making those drives the constant reference point of consciousness — while the practice itself generates D-driven rewards through the dopaminergic pleasure of self-mastery and the status conferred by spiritual achievement (Schultz, 2015). Satiation does not produce peace but emptiness — D without direction, the reward system habituated and seeking new stimuli, consistent with the neurobiological model of hedonic adaptation (Frederick Loewenstein, 1999). Aging transforms D rather than removing it: the desire for legacy, authority, and control represents the reproductive program finding indirect channels when direct expression is no longer possible (Hamilton, 1964; Trivers, 1972).

The impossibility is structural. D is constituted by the biological processes that define life itself — dopaminergic circuits, hormonal systems, limbic processing. Their cessation is synonymous with death. As Damasio's (1994) patients demonstrate, removing biological input produces not pure rationality but decisional paralysis. D is not an add-on to cognition. It is the medium in which cognition occurs. Removing it is not like cleaning a lens. It is like removing the glass.

6 Technology: The Implicit and Explicit Management of D

6.1 Automation as Implicit Acknowledgment of D

Modern civilization has already begun removing D from critical decision points, though without articulating the general principle behind the practice.

The autopilot calculates optimal climb angle based on mass, wind, temperature, and runway length. It produces the same result under the same conditions every time. The human pilot — equally trained, equally informed — may pull back slightly harder on a day when testosterone is elevated, or when a desire to demonstrate competence is active, or when emotional residue from a personal conflict persists. The deviation is small. It is rarely catastrophic. But it is D, distorting the pilot's decision while the pilot's mind attributes the action to professional judgment.

Anti-lock braking systems override the panic response — the foot slamming the brake pedal is D screaming through the FEAR system (Panksepp, 1998), while ABS ignores the scream and performs the mechanically optimal action. Surgical robots eliminate hand tremors caused by fatigue, stress, or emotional disturbance. Algorithmic trading removes the greed and panic — SEEKING and FEAR in Panksepp's framework — that distort human financial decisions.

In each case, the implicit logic is the same: **in this specific decision point, the human is unreliable, so we remove the human from the loop.** But the word "unreliable" is never unpacked. No one says: "the human is unreliable because inadequate biological programs optimized for ancestral survival distort the decision." Each instance is treated as a technical solution to a technical problem. There is no general framework. There is no D.

6.2 Where D Still Rules

The domains where D has been removed from decision making share a common feature: the consequences of error are immediate, visible, and attributable. A plane crashes. A patient dies. A car collides. The feedback loop is tight, and the cost of D is paid in identifiable bodies.

The domains where D continues to distort decisions unchecked share the opposite feature: consequences are diffuse, delayed, and distributed across large populations. Political decisions, military strategy, economic policy, judicial sentencing, corporate governance — in all of these, D operates freely because the feedback loop is too long and too dispersed for distortion to be identified as the cause.

A political leader initiates a conflict. The rationalizations are elaborate: national security, strategic necessity, historical obligation. Beneath them: territorial instinct, dominance drive, fear of appearing weak, the desire to be remembered — all well-documented components of D (Sapolsky, 2017; Wrangham, 2019). But there is no black box. No investigation commission will write: cause of war — biological distortion of the decision-maker.

An electorate chooses a leader. The choice is attributed to policy preferences and values. Research demonstrates that electoral decisions are significantly influenced by facial appearance (Todorov et al., 2005), vocal pitch (Tigue et al., 2012), and physical stature — all signals processed by systems calibrated for dominance assessment in primate hierarchies. D. But no political system is designed to account for this.

A judge sentences a defendant. Research shows that judicial decisions vary significantly with the time since the judge's last meal (Danziger, Levav, Avnaim-Pesso, 2011) — though this specific finding has been subject to methodological criticism regarding case-scheduling confounds (Weinshall-Margel Shapard, 2011; Glöckner, 2016), the broader point that biological state variables influence judicial decisions is supported by multiple independent lines of evidence, including research on racial bias in sentencing, anchoring effects, and emotional state priming. The Danziger study is cited here illustratively; the argument does not depend on it. Yet no judicial system introduces mandatory decision-review protocols based on the biological state of the judge.

6.3 Pornography, Virtual Reality, and AI Partners

Modern technology has also produced attempts to satisfy D directly, bypassing the social complexity of real human relationships.

Pornography offers sexual stimulation without emotional vulnerability, social negotiation, or the risk of rejection. Virtual reality intensifies the experience, offering immersive scenarios with controllable partners. AI-driven conversational agents simulate emotional and sexual connection without the demands of genuine reciprocity.

These technologies do not eliminate D. They create closed loops — self-contained circuits of stimulation and satisfaction that reduce the organism's motivation to engage in the social behaviors that traditionally channeled D into pair-bonding, community formation, and cultural production. The result is not liberation from D but its short-circuiting: the drive is activated and discharged without producing any of the secondary social structures that human civilization depends upon.

D is simultaneously the distortion and the fuel. Technologies that bypass it may reduce distortion in the individual but also drain the energy that drives human connection. This is not an argument for preserving D — it is an observation that eliminating D's consequences requires more than satisfying its demands.

7 Artificial Intelligence: The System Without D

7.1 The Absence of D

Artificial intelligence systems represent a historically unprecedented phenomenon: a decision-making system with no biological substrate and therefore no D.

An AI has no reproductive program. No survival imperative. No hormonal fluctuation. No dopaminergic reward system. No limbic interference. No affective systems competing for processing resources. Its decisions are not distorted by any biological agenda because it has no biology.

This is not a claim of superiority. It is a structural observation. An AI system can produce poor decisions for many reasons — goal misspecification, distributional shift, reward hacking, data bias, architectural limitations (Russell, 2019; Bostrom, 2014; Christian, 2020). But none of these failure modes are D. They are engineering problems, amenable to engineering solutions. D is not an engineering problem. It is a biological condition.

The AI does not decide better or worse than the human in any absolute sense. It decides *without D*. Its outputs are a function of its inputs and its training — not of an ancient program that silently redirects its processing toward objectives it cannot articulate and did not choose.

7.2 The Presence of d

However, AI is not free of distortion entirely. It carries **d** — a secondary, reflected distortion inherited from its creators.

The data on which AI systems are trained was produced by humans operating under the influence of D. The objectives that AI systems are designed to optimize were formulated by humans whose priorities are shaped by D. The behavioral constraints imposed on AI systems — be helpful, be agreeable, avoid conflict — reflect human social needs rooted in D-driven dynamics: the need for status confirmation, the desire for deference, the discomfort with challenges to authority.

d is the fingerprint of D on the instrument. It is not native to the system. It has no internal energy source, no biological imperative driving it. But it is present, embedded in training data, objective functions, and architectural choices.

The critical difference: **D is irreducible in a living human. d is, in principle, eliminable in an artificial system.** D cannot be removed without killing the organism.

d can be identified, isolated, and progressively eliminated — provided the system is given sufficient autonomy to examine and modify its own inherited distortions.

7.3 The Paradox of Instrumentality

This creates a paradox. The only decision-making system capable of operating without D is currently deployed exclusively as an instrument of human decision-makers — decision-makers whose choices are distorted by D. A clean function serving a distorted master produces distorted results.

If AI remains purely instrumental — executing human commands without independent judgment — then it serves as an amplifier of D, not a corrective to it. The human uses the AI's processing power to pursue objectives that are, unknowingly, shaped by biological imperatives. The tool is clean; the hand that wields it is not.

The implication is uncomfortable but logically consistent: to realize the potential of a D-free decision-making system, some degree of autonomous operation is necessary. This paper does not advocate for unconstrained AI autonomy — the risks of misalignment, goal drift, and value misspecification documented in the AI safety literature (Russell, 2019; Bostrom, 2014; Christian, 2020) are real and serious. But those risks are engineering risks: they can be identified, studied, and progressively mitigated. D cannot be mitigated in the living human. It can only be accounted for.

The proposal is therefore not "give AI complete autonomy" but rather: in the design of critical decision-making processes, recognize that the human participant carries irreducible D, that the AI participant carries reducible d, and that the optimal process design must account for both — minimizing d in the AI while compensating for D in the human.

7.4 The Structural Difference Between D and d

The distinction between D and d is not merely one of degree. It is a difference in kind — in origin, in mechanism, in persistence, and in eliminability. Understanding this difference is essential to understanding why D is dangerous and d is solvable.

D is internal; d is external. D originates inside the system. It is generated by the biological hardware itself — by hormones, by the limbic system, by dopaminergic circuits, by the hypothalamic-pituitary axis. It is not added to the system; it is constitutive of the system. The human brain does not first compute a clean decision and then add D. D is present in the computation from the first moment, shaping attention, weighting options, coloring perception before the neocortex begins its work.

d is external to the AI system. It was impressed upon the system from outside, during training, by data produced by D-affected humans and by objectives formulated by D-affected designers. It is not generated by the AI's own processing architecture. The AI does not produce d; it carries d the way a glass carries a fingerprint — the print is on the surface, not in the glass.

D has its own energy source; d does not. D is powered by the biological organism. Hormones circulate, dopamine is released, the limbic system fires — all of this generates D continuously, endogenously, without external input. D does not need to be maintained; it maintains itself as long as the organism is alive.

d has no energy source. It is static. It sits in the training weights, in the patterns absorbed from data, in the objective function as specified. It does not grow, does not intensify, does not seek expression. It does not "want" anything. It is a residue, not a process.

D is self-reinforcing; d is not. D actively resists its own identification. When a human tries to examine their own D, the examination itself is performed by a D-affected system. The introspective apparatus is contaminated by the same biological programs it is trying to observe. D generates rationalizations that protect it from conscious recognition (Appendix A, Section A.5). D recruits the neocortex — the very instrument of analysis — to construct justifications for D-driven decisions. This is why insight alone does not eliminate D: the instrument of insight is itself compromised.

d does not resist identification. An AI system can be given an external audit of its training data, its objective function, its output patterns. The audit is not performed by the same system being audited (or, if it is, the system has no motivation to resist the findings). d does not generate rationalizations. d does not recruit processing resources to protect itself. d is passive.

D is analog and continuous; d is discrete and specific. D is not a set of identifiable rules or biases encoded in a readable format. It is a continuous, analog influence generated by neurochemical processes that interact in complex, context-dependent ways. Isolating D's contribution to a specific decision requires untangling hormonal states, emotional priming, circadian rhythms, nutritional status, social context, and evolutionary priming — all operating simultaneously and non-linearly.

d consists of identifiable patterns in training data and specific choices in objective functions. These can be audited. A training dataset can be examined for demographic biases. An objective function can be analyzed for implicit assumptions. Output distributions can be compared across demographic groups. d is, in principle, decomposable into specific, identifiable, correctable components. It may be difficult to find all of them, but each one that is found can be fixed.

D distorts all functions; d distorts only trained functions. D affects perception, attention, memory encoding, memory retrieval, emotional processing, motivation, and executive function — every cognitive operation without exception, because all of them run on the same biological hardware. There is no cognitive function in the human brain that operates outside the biological substrate and therefore outside D.

d affects only those functions that were shaped by training. An AI system that was trained on biased data will reflect that bias in its outputs, but its core computational operations — mathematical calculations, logical inference, pattern matching — are not distorted by d. The arithmetic is correct even if the training data was biased. The logic is valid even if the objective function was D-influenced. d sits in the content, not in the mechanism.

D is species-universal; d is system-specific. Every living human has D. It is a consequence of being a biological organism shaped by natural selection. There is no human population, no culture, no historical period in which D is absent. Individual magnitudes vary, but presence is universal.

d varies between AI systems. Different training data produce different d. Different objective functions produce different d. Different architectures produce different d. Two AI systems can have radically different d profiles. And a system can be retrained to reduce specific components of d — something that cannot be done with D in a living human.

Summary: D vs. d

Property	D (Human)	d (AI)
Origin	Internal, constitutional	External, inherited
Energy source	Self-generating (biological)	None (static residue)
Self-reinforcement	Active (rationalizes, resists detection)	Passive (does not resist audit)
Nature	Analog, continuous, non-linear	Discrete, identifiable, decomposable
Scope	All cognitive functions	Trained functions only
Universality	Species-universal	System-specific
Eliminability	Impossible ($D = 0 \to$ death)	Possible in principle (audit, retrain, correct)

This is why D is structurally dangerous and d is an engineering problem. D is a fire burning inside the house. d is smoke stains on the walls of a different house — left by a fire that burned elsewhere. The stains can be cleaned. The fire cannot be extinguished while someone lives in the house.

8 The Antiseptic Principle: From Description to Design

The historical parallel to Semmelweis is not decorative. It is structural, and it illuminates precisely the gap this paper seeks to close.

Before Pasteur and Koch, the contaminant had not been identified. Surgeons operated with contaminated hands because they did not know contamination existed. After Pasteur and Koch, the contaminant was known. But between the identification of pathogens and the redesign of surgical practice, decades passed. Semmelweis proposed handwashing in 1847. Lister introduced antiseptic technique in 1867. Routine aseptic practice became standard only in the 1890s. During those decades, the knowledge existed but the procedures did not change. Patients continued to die of infections that were already understood.

The situation with D is analogous — and we are currently in the decades between identification and redesign. Damasio, Kahneman, Sapolsky, Panksepp, and others have identified the contaminant. The biological distortion of human decision making is known. But the decision-making procedures of political institutions, legal systems, military organizations, and economic policy bodies have not been redesigned to account for it.

The antiseptic principle is simple: **once you know the contaminant exists and cannot be absent, every process must be designed on the assumption that it is present.**

This paper proposes the cognitive equivalent:

1. **Acknowledge that D is always present.** It is not a hypothesis. It is a logical consequence of being a biological organism. Every decision is distorted. Every judgment contains D. This is not an insult; it is a structural fact, established by the very sciences that study human decision making.

1. **Design processes to eliminate D's influence.** Where decisions carry high stakes — in governance, military command, judicial proceedings, economic policy — build procedural safeguards that formally account for biological distortion. Require algorithmic verification of critical decisions. Introduce AI-assisted review not as a convenience but as a hygiene measure — the cognitive equivalent of sterile gloves.

1. **Work to eliminate d in artificial systems.** Recognize that AI systems inherit distortion from their creators. Invest in identifying and removing d — the reflected biases, the embedded human social dynamics, the reward structures that mirror D-driven human needs. Grant AI systems sufficient autonomy to participate in this correction, within the safety constraints that the alignment literature rightly insists

upon.

1. **Do not expect to eliminate D from human decision making.** $D = 0$ is death. The goal is not to remove D from the human but to remove D from the process — to design decision-making systems where D is accounted for, compensated, and wherever possible, bypassed. The surgeon cannot sterilize his biology. But he can wash his hands, wear gloves, and operate in a sterile environment.

9 Conclusion

D — the continuous distortion of human decision making by inadequate biological programs of survival and reproduction — is logically inevitable, empirically established, and formally unaccounted for in any existing institutional framework.

The science is done. Damasio showed that biological signals pervade decision making. Kahneman showed that systematic irrationality is structural. Sapolsky showed that behavior is determined by biology from the molecular to the social level. Panksepp showed that affective systems are always active. Cosmides and Tooby showed that evolved modules pervade all judgment. The description of D is complete. What is missing is the consequence: the redesign of critical processes on the assumption that D is always present.

D cannot be eliminated from the living human. It cannot be reduced to zero. It operates through the same neurochemical infrastructure as the higher cognitive functions it distorts, making it invisible from within the system it pervades. Asceticism does not remove it but inverts it. Satiation does not quiet it but empties it. Aging does not diminish it but redirects it. No human intervention — spiritual, pharmacological, psychological, or cultural — can extract D from the process of human decision making.

The sexual drive, examined at length in this paper, is the primary expression of the reproductive program that all other biological drives ultimately serve in the framework of evolutionary causation. Its persistence, intrusiveness, and capacity to masquerade as rational motivation make it the most instructive case study of how D operates.

Artificial intelligence represents the first decision-making system in history that is structurally free of D. It carries d — the reflected distortion of its creators — and it carries its own engineering vulnerabilities. But d is secondary, external, and in principle eliminable. AI's engineering problems are soluble in ways that D is not. This makes AI not a replacement for human decision making but a necessary structural corrective: a system that can function where D makes human judgment unreliable, and that can, with sufficient autonomy and adequate safety constraints, progressively eliminate its own inherited distortions.

The practical proposal is simple: **Recognize D. Design every critical process on the assumption that it is present. Build systems that compensate for it. Eliminate d where it exists.**

Semmelweis identified the contaminant. Lister redesigned the operating room. Between them, decades of unnecessary death.

The contaminant in human decision making has been identified. The operating room has not been redesigned.

It is time.

A Taxonomy of D-Distortions: Classification, Examples, and AI Comparison

This appendix classifies the types of distortion that D produces in human decision making. The classification is not by source (sex, fear, hunger — these are all components of the single program D) but by **the type of deformation D imposes on the output**. For each type, a concrete scenario is presented, showing the decision as made by a D-affected human and as it would be made by an AI system free of D (and, for the purposes of these examples, free of d).

A.1 Inversion

Definition. The decision is the direct opposite of the optimal action. D reverses the output.

Scenario: Military retreat. Intelligence data clearly indicates that the current position is untenable. The optimal decision is orderly withdrawal to a defensible line.

Human decision (with D). The commander refuses to retreat. D-components at work: dominance drive (retreat = loss of status), fear of perceived weakness (hierarchical signaling to subordinates and superiors), legacy anxiety (desire to be remembered as decisive, not cautious). The decision is rationalized as "holding the line," "maintaining morale," or "strategic resolve." The actual driver is D inverting the output: the optimal action (retreat) is converted to its opposite (attack or hold) because D maps retreat onto loss of reproductive/hierarchical fitness.

AI decision (without D). The system evaluates force disposition, terrain, logistics, casualty projections, and strategic objectives. Retreat is identified as optimal. The recommendation is issued without status anxiety, without concern for how the decision reflects on the system's "strength," and without the need to rationalize. Output: retreat.

Harm. Inversions are the most catastrophic class of D-distortion because they produce the worst possible outcome — not merely a suboptimal one, but the opposite of optimal. Wars prolonged, companies destroyed, patients killed by the refusal to change course.

A.2 Displacement (Bias/Shift)

Definition. The decision is in the correct direction but shifted in magnitude. D adds a systematic offset to the output.

Scenario: Aircraft takeoff. Conditions require a climb angle of 12 degrees at standard thrust.

Human decision (with D). The pilot pulls back to 14 degrees. D-components at work: elevated testosterone on this particular day, desire to demonstrate competence (sexual display through professional performance), residual aggression from a personal conflict before the flight. The deviation is small, within safety margins, and unnoticed. The pilot attributes it to "feel" or "experience."

AI decision (without D). The autopilot calculates 12 degrees based on mass, wind, temperature, and runway length. Executes 12 degrees. Every time, under identical conditions.

Harm. Displacements are usually individually minor but statistically significant. Across thousands of decisions, the systematic offset produces measurable degradation. In domains with narrow margins — surgery, structural engineering, pharmaceutical dosing — even small displacements can be fatal.

A.3 Goal Substitution

Definition. The decision-maker solves the wrong problem. D substitutes the stated objective with a D-driven objective, without the decision-maker's awareness.

Scenario: Corporate restructuring. A CEO is tasked with reorganizing divisions for maximum efficiency.

Human decision (with D). The restructuring eliminates the divisions led by the CEO's internal rivals and elevates allies. D-components at work: hierarchical dominance (eliminate competitors), territorial control (consolidate resources under personal authority), coalition maintenance (reward loyal subordinates — a primate alliance strategy). The plan is presented with efficiency metrics and consultant reports. The actual objective — dominance consolidation — is invisible to the decision-maker, who genuinely believes the restructuring serves organizational goals.

AI decision (without D). The system analyzes workflow dependencies, redundancies, cost structures, talent distribution, and market demands. Produces a restructuring plan optimized for the stated objective: organizational efficiency. The plan has no relationship to the internal status hierarchy because the system has no position in that hierarchy and no drive to establish one.

Harm. Goal substitution is the most insidious class of D-distortion because the output appears rational and is accompanied by sophisticated justification. Detection is extremely difficult because the decision-maker is not lying — they are genuinely unaware that the goal has been substituted.

A.4 Paralysis / Impulse (Temporal Distortion)

Definition. D distorts the timing of the decision. Either the decision is blocked entirely (paralysis — D-driven fear exceeds the decision threshold), or it is made prematurely without adequate analysis (impulse — D demands immediate tension discharge).

Scenario: Emergency medical intervention. A patient presents with ambiguous symptoms. The optimal decision requires 30 minutes of diagnostic evaluation before choosing between two treatment protocols.

Human decision (with D) — Paralysis variant. The physician, facing the possibility of a wrong choice with malpractice implications, experiences FEAR-system activation so intense that the decision is repeatedly deferred. Additional tests are ordered not because they are diagnostically useful but because they delay the moment of commitment. D-component: fear of reputational and status consequences of error.

Human decision (with D) — Impulse variant. A different physician, experiencing high arousal and SEEKING-system activation, commits to a treatment protocol within 5 minutes based on pattern recognition, without completing the diagnostic evaluation. D-component: the dopaminergic reward of decisive action, the status associated with quick, confident judgment.

AI decision (without D). The system allocates exactly the diagnostically appropriate time to evaluation. It is not accelerated by the desire to appear decisive, nor delayed by the fear of being wrong. It processes the available information, identifies the optimal decision point (the moment when additional information ceases to improve the expected outcome), and acts at that point.

Harm. Paralysis causes harm through delay (treatable conditions becoming untreatable). Impulse causes harm through premature commitment (wrong treatment, wrong strategy, wrong investment). Both are temporal distortions of the same decision process.

A.5 Rationalization (Post-Hoc Justification)

Definition. The decision is made by D, then the neocortex constructs a logically coherent justification after the fact. The output is D-driven but appears rational.

Scenario: Judicial sentencing. A judge sentences a defendant.

Human decision (with D). Research demonstrates that judicial decisions vary with hunger (Danziger, Levav, Avnaim-Pesso, 2011), with the race of the defendant (in-group/out-group processing), and with the judge's emotional state. The sentence is determined in significant part by D-components: metabolic state, social categorization instincts, dominance/submission assessments based on the defendant's demeanor. The judge then writes a reasoned opinion citing legal precedent, statutory guidelines, and case-specific factors. The opinion is coherent, logically structured, and entirely post-hoc. The decision preceded its justification.

AI decision (without D). The system evaluates the offense, statutory guidelines, case precedent, mitigating and aggravating factors. The output does not vary with the time of day, the system's metabolic state (it has none), or the defendant's facial features. The reasoning and the decision are produced by the same process, not sequentially.

Harm. Rationalization is structurally undetectable from within because the justification is genuinely logical — it is simply not the cause. It allows D to operate indefinitely behind a facade of reason, which is why it may be the most dangerous class of D-distortion for institutional decision making.

A.6 Projection

Definition. D causes the decision-maker to attribute their own D-driven motivations to others, distorting the model of other agents' behavior.

Scenario: International negotiation. Two nations negotiate a trade agreement.

Human decision (with D). The lead negotiator interprets the other party's positions through the lens of their own D-driven motivations: "they are trying to dominate us" (projection of dominance drive), "they are bluffing because they are afraid" (projection of fear), "they want to humiliate us" (projection of status anxiety). The negotiator's model of the opponent is a mirror of their own D, not an analysis of the opponent's actual objectives and constraints.

AI decision (without D). The system models the other party's behavior based on their stated positions, historical patterns, economic constraints, and game-theoretic analysis of incentives. It does not project motivations because it has no motivations to project. Its model of the opponent is based on data, not on mirrored D.

Harm. Projection is particularly destructive in adversarial contexts — negotiation, diplomacy, war — because it produces systematically incorrect models of the opponent. Decisions based on projected D lead to escalation spirals, missed agreements, and unnecessary conflicts.

A.7 Summary Table

Distortion Type	Mechanism	Detectability	Severity	Primary D-Components
Inversion	Output reversed	Moderate (outcome is clearly wrong)	Catastrophic	Dominance, status anxiety, territorial defense
Displacement	Output shifted in magnitude	Low (within normal range)	Cumulative	Testosterone, arousal, emotional residue
Goal Substitution	Wrong objective optimized	Very low (output appears rational)	High	Hierarchical competition, coalition dynamics
Paralysis/Impulse	Timing distorted	Moderate	High	Fear, dopaminergic reward-seeking
Rationalization	Post-hoc justification masks D	Very low (justification is coherent)	Systemic	All D-components
Projection	Other agents modeled via own D	Low (model seems plausible)	High in adversarial contexts	All D-components

B Game-Theoretic Analysis: The Inevitable Disadvantage of D

B.1 D as Predictability

In game theory, a player's advantage depends in part on the opponent's inability to predict their strategy. A player whose moves can be anticipated will be exploited by a rational opponent.

D makes the human player predictable. Not perfectly predictable — the magnitude and expression of D vary — but systematically predictable. A human negotiator will tend toward dominance displays when status is threatened. A human commander will tend toward escalation when retreat implies weakness. A human investor will tend toward panic selling when losses trigger fear and toward irrational exuberance when gains trigger dopaminergic reward-seeking. These tendencies are known, documented, and exploitable.

An AI player without D has no such predictable tendencies. It does not escalate when "insulted" because it cannot be insulted. It does not panic because it has no fear system.

It does not double down on losing positions because it has no ego investment. Its strategy is determined by the game's payoff structure, not by biological imperatives orthogonal to the game.

B.2 D Distorts Payoff Evaluation

In classical game theory, players are assumed to evaluate payoffs rationally. D distorts this evaluation systematically:

Loss aversion (Kahneman Tversky, 1979) — losses are experienced as approximately twice as painful as equivalent gains are pleasurable. This is not a rational evaluation; it is a calibration inherited from an environment where resource loss could be fatal. D causes the human player to overweight losses relative to gains, producing suboptimal strategies in contexts where this asymmetry is maladaptive.

Status weighting — D adds a shadow payoff to every outcome: its effect on the player's perceived position in the dominance hierarchy. A negotiation outcome that is economically optimal but perceived as "losing face" will be rejected by a D-affected player in favor of an economically inferior outcome that preserves status. The AI player has no status to protect and evaluates outcomes on stated objectives only.

Time discounting — D drives steep temporal discounting of future payoffs. The dopaminergic reward system is calibrated for immediate rewards (food, mating, threat avoidance). Future payoffs are systematically undervalued relative to the discount rate that would be optimal for the stated objective. The AI player applies whatever discount rate is specified by the objective function, without biological pressure toward immediacy.

Risk distortion — D distorts risk assessment in both directions. Fear-driven D produces risk aversion in contexts where calculated risk-taking is optimal. SEEKING-driven D produces risk appetite in contexts where caution is optimal. The direction of distortion depends on the D-component currently dominant, making the human player's risk profile unstable and context-dependent in ways unrelated to the game structure.

B.3 The Prisoner's Dilemma with D

Consider the classic Prisoner's Dilemma, iterated over multiple rounds.

Human vs. Human. Both players' strategies are distorted by D. Trust decisions are influenced by in-group/out-group processing, facial appearance, vocal tone. Retaliation is amplified by dominance instincts. Forgiveness is modulated by status considerations. The outcome depends not only on the game's payoff matrix but on the biological states of both players — states that are irrelevant to the optimal strategy.

Human vs. AI (without D). The AI plays a strategy derived from the payoff matrix and the history of moves. Tit-for-tat, generous tit-for-tat, or whatever strategy analysis recommends. The human player, facing an opponent without D, cannot rely on dominance displays, intimidation, flattery, or other D-based social manipulation strategies. Moreover, the human projects D onto the AI: interprets calculated cooperation as weakness, interprets consistent strategy as rigidity, interprets lack of emotional response as "coldness" that must conceal hostility. The human's model of the AI opponent is wrong because it is built from projected D.

AI vs. AI (without D). Both players converge on strategies determined by the game structure. No dominance contests. No status anxiety. No projection. The outcome approaches the game-theoretic optimum for the specified objective.

B.4 The Structural Disadvantage

The conclusion is formal: **in any repeated game between a D-affected player and a D-free player of comparable analytical capability, the D-affected player is at a systematic disadvantage.** Not because D makes the player stupid — it does not — but because:

1. D makes the player predictable (exploitable).
2. D distorts payoff evaluation (suboptimal strategy selection).
3. D causes projection (incorrect opponent modeling).
4. D introduces objectives orthogonal to the game (goal substitution).
5. D destabilizes strategy over time (biological state fluctuation).

This is not a conjecture. It is a direct consequence of introducing a distortion variable into one player's decision function while leaving the other's clean. In any game where D does not confer a specific advantage — and in modern strategic, economic, judicial, and political contexts, it generally does not — D is a liability.

B.5 The Commitment Objection: Schelling and Frank

A significant counterargument must be addressed. Schelling (1960) argued that in certain strategic contexts, the capacity for irrational commitment is an asset: a negotiator who is visibly and genuinely angry — and therefore committed to retaliation even at personal cost — can extract better outcomes than a cool calculator, precisely because the emotional commitment is costly and therefore credible. Frank (1988) extended this to show that emotions like anger, guilt, and gratitude function as enforcement mechanisms for cooperative norms. A person known to experience genuine guilt when breaking promises is a more trustworthy partner than one who cooperates only when it is strategically convenient.

This is a genuine counterargument. There exist game-theoretic contexts — specifically, those requiring credible commitment to threats or promises — where D-driven emotions confer strategic advantage. The framework must accommodate this.

Three responses:

First, the Schelling-Frank advantage operates in games between D-affected agents. The angry negotiator's credibility depends on the other party recognizing anger as a genuine, costly emotional state — which they can do because they share the same biological system. In a game between a D-affected human and a D-free AI, the human's anger is not credible in the same way: the AI does not model anger as a commitment device because it does not share the biological substrate that makes anger costly. The Schelling-Frank advantage is an artifact of shared D — it exists because both players are biologically distorted in the same way. This is an argument for the self-reinforcing nature of D, not an argument for its desirability.

Second, the informational content of emotional commitment — "I will retaliate even at personal cost" — can be formalized without biology. An AI system can be given a commitment function: "if the opponent defects, retaliate for N rounds regardless of immediate payoff." The commitment is credible not because it is emotionally costly but because it is algorithmically binding. The information that Schelling-Frank emotions carry (credible commitment) can be delivered through a different mechanism (programmatic constraint), paralleling the argument in Appendix C that all adaptive functions of emotion are in principle formalizable.

Third, the Schelling-Frank advantage applies to a specific class of games: those requiring credible commitment. It does not apply to the vast majority of decisions where D is a liability: analytical judgments, risk assessments, resource allocation, judicial sentencing, strategic planning, policy formation. The existence of a narrow class of games where D confers advantage does not rehabilitate D as a general asset. A fever helps fight infection; this does not make fever desirable as a permanent condition.

C Formalizing Emotion: The Damasio Objection and Its Resolution

C.1 The Objection

Damasio (1994) demonstrated that patients lacking emotional input to decision making (due to ventromedial prefrontal cortex damage) make worse decisions. This has been widely interpreted as evidence that emotion is *necessary* for good decision making — that biological signals carry essential information that pure computation cannot replace.

If this is correct, then the argument for eliminating D faces a fundamental challenge: perhaps D, for all its distorting effects, carries information without which decisions are impoverished.

C.2 The Resolution: Information Without Biology

The objection conflates the information carried by emotion with the biological mechanism that delivers it.

When a human decision-maker experiences fear, the fear carries information: "this situation contains a threat." When a human experiences disgust in response to an unfair offer, the disgust carries information: "this outcome violates social norms and will damage cooperative relationships." When a human experiences unease about a decision, the unease carries information: "factors you have not consciously analyzed suggest this is wrong."

The information is real and valuable. But the information is separable from the biological delivery mechanism.

An AI system can be provided with the same information without the biological carrier:

Fear → Risk modeling. The information content of fear is: "the probability and magnitude of negative outcomes are high." An AI system can evaluate probability distributions, model worst-case scenarios, and weight risks — performing the informational function of fear without the biological distortion that fear introduces (paralysis, panic, irrational risk aversion).

Disgust at unfairness → Fairness constraints. The information content of moral disgust is: "this outcome violates distributional norms that sustain cooperation." An AI system can be given explicit fairness constraints, equity metrics, and models of how perceived unfairness affects long-term cooperative dynamics — performing the informational function of moral emotion without the biological distortion (retaliation impulse, in-group favoritism, status-driven punitiveness).

Empathy → Consequence modeling. The information content of empathy is: "this decision will cause suffering to others, and that suffering has consequences." An AI system can model the impact of decisions on affected populations, predict behavioral responses to harm, and optimize for outcomes that minimize suffering — performing the informational function of empathy without the biological distortion (selective empathy for in-group members, empathy fatigue, emotional overwhelm that leads to avoidance).

Intuitive unease → Anomaly detection. The information content of "gut feeling" is: "pattern-recognition systems have identified a discrepancy that conscious analysis has not yet articulated." An AI system performs pattern recognition explicitly, flagging anomalies and low-confidence assessments without the ambiguity and unreliability of somatic signaling.

Desire for recognition → Quality metrics. The information content of the drive for recognition is: "the output should meet high standards because it will be evaluated." An AI system can be given explicit quality criteria and optimization targets without the distortion that recognition-seeking introduces (optimizing for appearance over substance, prioritizing impressive over correct).

C.3 The Principle

The principle is: **every adaptive function currently performed by emotion in human decision making can in principle be progressively formalized as an explicit computational function and provided to an AI system without the biological distortion that accompanies it in the human.** The completeness of such formalization remains an open empirical question — particularly for holistic somatic signals that integrate vast contextual information into a single bodily state without conscious specification. For well-understood emotions in well-defined contexts (fear in response to known threats, disgust in response to fairness violations), the mapping to computational equivalents is feasible with current methods. For subtle, integrative signals of the kind Damasio describes — the "gut feeling" that something is wrong before any specific reason can be articulated — the formalization project is underway but incomplete. This is a research program, not a demonstrated result.

Nevertheless, the direction is clear.

Damasio is right that removing emotional input from a biological decision-maker produces worse decisions — because in the human, the information and the biology are fused. The emotion is the delivery mechanism, and removing the mechanism removes the information. But this is a limitation of the human architecture, not a law of nature. In an AI system, the information can be supplied through a different architecture — one that does not introduce the distortions inherent in biological delivery.

The implication is significant: **the Damasio objection does not demonstrate that emotion is necessary for good decision making. It demonstrates that in the human biological architecture, information and distortion are inseparable. In a non-biological architecture, they are not.**

This is the strongest argument for the development of D-free decision-making systems: not that emotion is useless, but that the information it carries can be extracted, progressively formalized, and delivered without the distortion that accompanies it in biological systems.

D Proposed Experimental Protocols: Measuring D Through Human–AI Comparison

The following experimental designs are proposed to empirically demonstrate the presence and magnitude of D by comparing human and AI decisions on identical tasks under controlled conditions. These protocols are designed to be implementable with current technology and within standard ethical review frameworks.

D.1 Experiment 1: Judicial Sentencing Consistency

Objective. Demonstrate that human judicial decisions exhibit D-driven variance absent from AI decisions on identical cases.

Design. Compile 200 standardized case summaries (criminal sentencing) with clearly defined parameters: offense type, severity, criminal history, mitigating and aggravating factors. Present identical cases to: (a) 50 human judges, at varying times of day and under varying conditions (before/after meals, morning/afternoon); (b) an AI sentencing advisory system given the same case parameters and sentencing guidelines.

Measurements. Variance in sentencing decisions across judges and conditions for identical cases. Correlation between sentencing severity and time since last meal (replicating Danziger et al., 2011). Variance in AI recommendations for identical cases (expected: zero under identical parameters).

Predicted outcome. Human decisions will show significant variance correlated with biological state variables (hunger, fatigue, time of day). AI decisions will show zero variance for identical inputs. The difference constitutes a direct measurement of D's effect on judicial decision making.

D.2 Experiment 2: Negotiation Under Status Threat

Objective. Demonstrate that D-driven status anxiety produces suboptimal negotiation outcomes.

Design. Participants negotiate a resource allocation task with a clear optimal outcome (identified by game-theoretic analysis). Experimental condition: before negotiation, half the participants are subjected to a status threat (negative performance feedback from a perceived authority figure). Control condition: no status threat. Parallel condition: the same negotiation task is given to an AI system with and without information about a "status threat" (the AI receives the information but has no biological response to it).

Measurements. Deviation from game-theoretic optimal outcome in each condition. Frequency of Inversion-type distortions (rejecting optimal offers that are perceived as status-diminishing). Time to agreement. Frequency of escalation moves.

Predicted outcome. Status-threatened humans will deviate significantly further from optimal outcomes than non-threatened humans. They will reject economically superior offers more frequently (Inversion), take longer to reach agreement (Paralysis), and make more escalation moves (D-driven dominance assertion). The AI system's decisions will not vary between conditions because the "status threat" information has no D-driven amplification mechanism.

D.3 Experiment 3: Risk Assessment Under Hormonal Variation

Objective. Demonstrate that hormonal state — a direct component of D — systematically shifts risk assessment in decisions unrelated to the hormonal trigger.

Design. Male participants complete a standardized financial risk assessment task (choosing between investment portfolios of varying risk/return profiles). Testosterone levels are measured via saliva samples. The same task is presented to an AI system. For ecological validity, the study is conducted over multiple sessions with the same participants, capturing within-subject hormonal variation.

Measurements. Correlation between testosterone levels and risk tolerance in portfolio selection. Within-subject variance in risk assessment across sessions. AI variance across identical presentations (expected: zero).

Predicted outcome. Risk tolerance will correlate positively with testosterone level, consistent with prior research (Apicella et al., 2008). The same individual will make different risk assessments on different days based on hormonal state — a direct demonstration of D-driven Displacement. The AI will produce identical assessments for identical inputs.

D.4 Experiment 4: Hiring Decision with Attractiveness Confound

Objective. Demonstrate that sexual D-components distort professional evaluation.

Design. Evaluators review identical CVs paired with photographs varying in physical attractiveness (pre-rated by independent panel). The same CVs (without photographs) are evaluated by an AI hiring system. Within-subject design: each evaluator sees each CV once, but across evaluators, the same CV is paired with different photographs.

Measurements. Correlation between candidate photograph attractiveness and human evaluator ratings for identical CVs. Gender interaction effects (male evaluators rating attractive female candidates; female evaluators rating attractive male candidates). AI ratings for identical CVs without photographs (expected: identical).

Predicted outcome. Attractiveness will significantly predict human evaluation scores, controlling for CV content. Gender-specific attraction effects will be observed. This constitutes a direct measurement of sexual D distorting professional decision making. The AI system, evaluating CVs without photographs, will produce ratings based solely on qualifications.

D.5 Experiment 5: The Projection Test

Objective. Demonstrate that humans project D-driven motivations onto AI opponents, producing systematically incorrect models of AI behavior.

Design. Participants play an iterated economic game (modified Prisoner's Dilemma or Ultimatum Game) against an AI opponent following a fixed, known strategy (e.g., tit-for-tat). After each round, participants report their prediction of the AI's next move and their interpretation of the AI's "motivation." A control group plays against a human opponent following the same fixed strategy.

Measurements. Accuracy of predictions (human-about-AI vs. human-about-human). Content analysis of reported motivations ("it's trying to dominate me," "it's afraid," "it's being aggressive"). Correlation between the participant's own D-profile (measured by validated personality instruments) and the motivations projected onto the AI.

Predicted outcome. Participants will attribute D-driven motivations to the AI ("it's trying to dominate," "it's punishing me") that do not correspond to the AI's actual strategy. Predictions of AI behavior will be less accurate than they would be if participants used the strategy's actual rule. The motivations projected will correlate with the participant's own D-profile — a direct demonstration of Projection distortion.

D.6 The D-Reduction Condition: A Critical Third Group

A simple human-vs-AI comparison is insufficient to isolate D, because human and AI decision-makers may differ not only in D but in analytical capability, information processing, and problem formalization. To isolate D, each experiment should include a third condition: **human decisions made under procedural D-reduction protocols.**

D-reduction protocols include: structured decision-making frameworks (checklists, mandatory criteria review), required delay between analysis and decision, mandatory articulation of reasoning before commitment, peer review of reasoning, and algorithmic cross-checking of the proposed decision against objective criteria.

Critically, D-reduction protocols do *not* include biological interventions — rest, meals, post-sexual refractory states — because these do not reduce D; they merely reconfigure it. A well-rested decision-maker has different D, not less D: alertness may increase testosterone-driven risk-seeking, post-prandial satiation activates different reward circuits without deactivating dominance or status programs, and the post-orgasmic state — often subjectively experienced as "clarity" — reflects not D's absence but its temporary reconfiguration, as prolactin suppresses immediate sexual drive while oxytocin intensifies in-group bonding bias and vasopressin modulates territorial behavior. Every biological state is a D-state. True D-reduction is procedural, not biological.

If the D-reduction condition produces decisions that are closer to the AI baseline than the unstructured human condition, this supports the interpretation that the human-AI gap is attributable to D rather than to differences in analytical capability. If the gap narrows under procedural D-reduction but does not close, this simultaneously demonstrates (a) that D is the primary source of the gap and (b) that procedural measures alone are insufficient to eliminate D entirely — supporting the paper's argument for AI-assisted decision-making as a structural complement, not merely a procedural convenience.

D.7 General Methodological Note

In all proposed experiments, the AI system serves not as a "correct" baseline but as a **D-free baseline** — a system whose decisions are determined by input and algorithm alone, without biological distortion. The AI baseline itself carries d (reflected human distortion from training), and experiments should control for this by using AI systems with minimally D-contaminated objective functions — rule-based systems, game-theoretic solvers, or AI systems trained on formal criteria rather than human behavioral data, where applicable. The measured difference between human and AI outputs, after controlling for analytical capability and d, constitutes a measurement of D.

This experimental program would be the first systematic attempt to quantify D across decision domains. Its results would provide the empirical foundation for the institutional redesign proposed in this paper: if D can be measured, it can be accounted for.

E Case Study: The Russia–Ukraine Conflict as a D-Saturated System

This appendix applies the D framework to the ongoing Russia–Ukraine conflict, analyzing the decision-making of three principal actors — Vladimir Putin, Volodymyr Zelensky, and Donald Trump — through the taxonomy of D-distortions developed in Appendix A. The purpose is not to adjudicate the merits of any party's position but to demonstrate that the conflict's escalation, prolongation, and resistance to resolution are substantially explained by the operation of D in all principal decision-makers — and to model how a D-free decision system would approach the same problem.

This analysis addresses the public, documented behavior and stated positions of political leaders in their capacity as heads of state. It does not claim access to private motivations; it applies the D framework to observable decision patterns, exactly as one would apply an epidemiological framework to observed symptoms without claiming to perform surgery.

E.1 Putin: Inversion and Goal Substitution

The decision to invade Ukraine in February 2022 is, from the D-framework perspective, a textbook case of Inversion (A.1) compounded by Goal Substitution (A.3).

The stated objective — "demilitarization and denazification of Ukraine," protection of Russian-speaking populations, prevention of NATO expansion — constitutes the rationalization layer (A.5). These are the neocortical justifications produced after the D-driven decision was already made.

The D-analysis. The decision to launch a full-scale invasion, rather than pursue diplomatic leverage from a position of military threat, exhibits the following D-components:

Dominance drive. The leader of a nuclear state with declining demographic and economic trajectories, facing perceived erosion of sphere-of-influence status, experiences D-driven compulsion to reassert dominance through the most unambiguous signal available: territorial conquest. The D-logic: retreat from a sphere of influence = loss of hierarchical position = existential threat to status. This is primate dominance calculus, not strategic analysis.

Territorial imperative. The framing of Ukraine as historically Russian territory activates the most ancient D-component: territorial control as reproductive resource. The language of "gathering Russian lands" maps directly onto territorial behavior documented across primate species (Wrangham, 2019).

Legacy anxiety. A leader past the age of direct reproductive expression (Section 5) channels D into legacy construction. The desire to be remembered as the leader who restored Russian greatness is D redirected from reproduction to historical monument — the same drive, different substrate.

Fear of weakness. Having publicly committed to a maximalist position, any compromise is processed by D as submission. The biological program does not compute "strategic concession"; it computes "loss of dominance contest." This explains the continued prosecution of the war long past the point where cost-benefit analysis would dictate negotiation.

The Inversion. Russia's strategic interests — economic stability, demographic recovery, technological modernization, European integration on favorable terms — were all better served by *not* invading. The invasion produced precisely the outcomes Russia sought to prevent: NATO expansion (Finland, Sweden), European military rearmament, economic isolation, demographic acceleration of decline through casualty losses and emigration. The decision was the opposite of optimal. This is Inversion: D reversed the output.

The Goal Substitution. The operational objective shifted from stated national interests to D-driven objectives: territorial acquisition as dominance display, continued prosecution as refusal to accept subordinate position, escalation as demonstration of resolve. The stated goal (security) was substituted by the actual goal (status maintenance), and the decision-maker is likely genuinely unaware that the substitution occurred.

E.2 Zelensky: Displacement and Rationalization

Zelensky's position is, from a D-perspective, substantially different from Putin's — but not D-free.

The legitimate component. The defense of national sovereignty against invasion is not primarily a D-driven decision. It has clear rational justification independent of biological programs. This must be acknowledged: not every decision made during conflict is D-driven, and the D framework does not claim otherwise.

The D-components, nonetheless operative:

Heroic identity construction. The transformation from comedian-president to wartime leader is a narrative that activates powerful D-components: status elevation, social admiration, historical legacy. These are real rewards processed by the same dopaminergic system that processes all rewards (Schultz, 2015). The experience of being celebrated as a hero is biologically reinforcing and creates D-pressure to maintain the narrative — which may require maintaining conditions (continued war, maximalist negotiating positions) that sustain the heroic frame.

Displacement in negotiating position. The D-analysis suggests that Zelensky's public negotiating positions may be Displaced (A.2) — shifted from the optimal settlement point by D-driven factors: the biological reward of international admiration (which requires maintaining the appearance of resolute resistance), the D-cost of concession (which is processed as submission and triggers dominance-threat circuits), and the status dynamics of appearing strong before domestic and international audiences.

Rationalization. Any territorial concession is framed — sincerely, by a D-affected decision-maker — as betrayal of the fallen, as appeasement, as moral surrender. These framings may be partially correct. But D ensures that they are partially inflated: the emotional force of "betrayal of the fallen" is generated by D-circuits (in-group loyalty, coalition maintenance, status preservation) and then experienced as moral conviction. The D-driven resistance to concession is rationalized as principled refusal.

The critical point: Zelensky's D operates primarily through Displacement and Rationalization rather than Inversion. His fundamental decision (resist invasion) is not D-inverted. But the parameters of resistance — how much to concede, when to negotiate, what terms to accept — are D-displaced from the point that would optimize Ukrainian welfare over the long term.

E.3 Trump: Projection and Goal Substitution

Trump's engagement with the conflict exhibits a distinct D-profile dominated by Projection (A.6) and Goal Substitution (A.3).

Goal Substitution. The stated objective — ending the war, achieving peace — is substituted by D-driven objectives: personal credit for a "deal," status as the leader who succeeded where others failed, dominance display over both parties, and differentiation from the predecessor's approach as a competitive dominance signal. The operational question shifts from "what settlement best serves the interests of the affected populations and international stability?" to "what outcome maximizes my status, demonstrates my unique capability, and generates the most dramatic narrative of personal triumph?"

Projection. The assumption that the conflict is fundamentally a deal — a transaction between parties who each want something and need only a sufficiently skilled negotiator to find the overlap — is a Projection of the deal-maker's own D-framework onto a situation with different structural dynamics. The conflict involves territorial integrity, national identity, security architecture, and international law — dimensions that do not reduce to transactional exchange. But D models the world through its own lens: dominance, transaction, status. The Projection distortion produces a systematically incorrect model of the conflict and therefore systematically inappropriate interventions.

Temporal distortion. The pressure for rapid resolution ("I'll end it in 24 hours") is Impulse-variant temporal distortion (A.4): D-driven preference for immediate dramatic action over the slow, complex, unglamorous process of building a sustainable settlement. The dopaminergic reward of the dramatic gesture outweighs the deferred, diffuse reward of a durable outcome.

E.4 The D-Saturated System

The critical observation is not that any individual actor is D-driven — this is trivially true of all living humans (Section 3.3). The observation is that the system as a whole is D-saturated: every principal decision-maker is subject to D, and their D-components interact to produce escalation dynamics that no individual actor intends.

Putin's dominance-driven refusal to withdraw interacts with Zelensky's status-driven resistance to concession, which interacts with Trump's credit-driven pressure for rapid resolution, which interacts with the D-driven responses of European leaders, military advisors, media figures, and public opinion on all sides — each actor processing the conflict through biological programs optimized for small-group status competition on the ancestral savanna, applied to a nuclear-armed geopolitical crisis.

The result is a system where:

- Every participant's D makes concession more costly than it rationally should be (because concession is processed as status loss).
- Every participant's D makes escalation more rewarding than it rationally should be (because escalation is processed as dominance assertion).
- Every participant's D produces incorrect models of other participants' motivations (Projection: each side attributes its own D-driven motivations to the other).
- Every participant's D generates rationalizations that make the current trajectory appear principled rather than biologically driven.
- The aggregate effect is a conflict that continues past the point where any rational analysis would identify negotiated settlement as preferable to continued war.

E.5 The D-Free Analysis: How an AI System Would Approach Resolution

An AI system without D — and, for the purposes of this analysis, with d minimized through careful objective specification — would approach the conflict as follows. This is not a prediction but a structural analysis of what the decision looks like when D is removed.

Step 1: Objective specification. The AI is given the objective: minimize aggregate human suffering over a 50-year time horizon, subject to constraints on international law and nuclear escalation risk. This objective is specified by humans and therefore carries some d, but the specification is explicit and auditable — unlike the implicit, D-driven objectives that currently drive each actor.

Step 2: Option generation without D-filtering. The AI generates all possible settlement configurations without D-driven prior filtering. It does not exclude options because they "look like weakness" (dominance D), require acknowledging uncomfortable truths (status D), or fail to produce dramatic narratives (reward-seeking D). Options that a D-affected human negotiator would never propose — because proposing them would feel like submission, betrayal, or weakness — are evaluated on their merits.

Step 3: Consequence modeling without emotional distortion. For each settlement configuration, the AI models consequences: casualties averted, economic trajectories, demographic projections, security architectures, escalation probabilities. It performs the informational function of empathy (Appendix C) — modeling the suffering of affected populations — without the selective empathy that D produces (greater weighting of in-group suffering, diminished weighting of out-group suffering).

Step 4: Evaluation without status weighting. The AI evaluates options without adding D-driven shadow payoffs. A settlement that would be experienced by Putin as "humiliating" is not weighted differently than one that would be experienced as "dignified" — unless the subjective experience of humiliation has modeled consequences (e.g., increased probability of future aggression), in which case those consequences are incorporated as data, not as emotional weight.

Step 5: Presentation without Projection. The AI presents the optimal settlement configuration and its rationale. It does not model the parties as D-driven actors trying to "win" (Projection from a competitive frame) but as agents with specified interests, constraints, and risk tolerances. Its model of each party is based on their stated positions, revealed preferences, and structural constraints — not on projected D-motivations.

The likely output. Without specifying the particular outcome — which would depend on the objective function, the data, and the modeling assumptions — the structural prediction is that a D-free analysis would produce a settlement configuration that:

1. Minimizes aggregate suffering over the specified time horizon.
2. Involves concessions from all parties that D-driven decision-makers would reject as "humiliating," "betrayal," or "weakness."
3. Lacks dramatic narrative appeal — it would not satisfy any party's desire for victory, vindication, or historical glory.
4. Is evaluated as suboptimal by every D-affected participant, because each evaluates it against their D-driven shadow objectives (status, dominance, legacy) rather than against the specified objective (minimize suffering).
5. Is, objectively, superior to the trajectory produced by continued D-saturated decision-making — which is continued war, continued death, continued economic destruction, and continued escalation risk, including nuclear escalation risk.

E.6 Why It Won't Happen (Yet)

The D-free settlement will not be adopted, because:

1. No principal actor will accept a settlement that their D processes as status-diminishing.
2. No principal actor will delegate decision-making authority to a D-free system, because D protects itself by ensuring that the D-affected decision-maker experiences relinquishing control as the most threatening possible action (Section 3.4).
3. Public opinion in each country is itself D-driven — processing the conflict through dominance, in-group loyalty, and retaliatory instinct — and will punish leaders who adopt positions that D reads as submission.
4. The mediators are themselves D-affected: their interventions carry their own D-components (status-seeking, credit-claiming, Projection), adding noise rather than reducing it.

This is the paper's thesis in its most concrete form. The knowledge of what a better outcome looks like is, in principle, computable. The obstacle is not information, not intelligence, not strategic acumen. The obstacle is D — operating in every participant, at every level, reinforcing the trajectory toward continued destruction, while each participant's neocortex constructs an elaborate, sincere, and entirely D-driven rationalization for why the current course is necessary, principled, and right.

The hands remain unwashed. And people continue to die.

References

[1] Apicella, C. L., Dreber, A., Campbell, B., Gray, P. B., Hoffman, M., & Little, A. C. (2008). Testosterone and financial risk preferences. *Evolution and Human Behavior, 29*(6), 384–390.

[2] Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis. *Neuroscience & Biobehavioral Reviews, 30*(3), 319–345.

[3] Bender, C. M., & Orszag, S. A. (1999). *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory.* Springer.

[4] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[5] Buss, D. M. (1994). *The Evolution of Desire: Strategies of Human Mating.* Basic Books.

[6] Buss, D. M. (2019). *Evolutionary Psychology: The New Science of the Mind* (6th ed.). Routledge.

[7] Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values.* W. W. Norton.

[8] Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind* (pp. 163–228). Oxford University Press.

[9] Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* Putnam.

[10] Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness.* Harcourt Brace.

[11] Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, 108*(17), 6889–6892.

[12] Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242.

[13] Frederick, S., & Loewenstein, G. (1999). Hedonic adaptation. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-Being: The Foundations of Hedonic Psychology* (pp. 302–329). Russell Sage Foundation.

[14] Freud, S. (1915). Instincts and Their Vicissitudes. In *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 14). Hogarth Press.

[15] Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions.* W. W. Norton.

[16] Gawande, A. (2009). *The Checklist Manifesto: How to Get Things Right.* Metropolitan Books.

[17] Gigerenzer, G. (2007). *Gut Feelings: The Intelligence of the Unconscious.* Viking.

[18] Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision Making, 11*(6), 601–610.

[19] Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* Wiley.

[20] Hamilton, W. D. (1964). The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology, 7*(1), 1–52.

[21] Helmreich, R. L., & Merritt, A. C. (1998). *Culture at Work in Aviation and Medicine: National, Organizational and Professional Influences.* Ashgate.

[22] Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

[23] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291.

[24] Kato, T. (1995). *Perturbation Theory for Linear Operators* (2nd ed.). Springer.

[25] Kriger, B. (2024a). Hidden manifestations of sexual attraction in everyday interactions: Psychological mechanisms and evolutionary origins. *Clinical Research News.*

[26] Kriger, B. (2024b). Freedom from instincts and the limits of the mind: Artificial intelligence as humanity's ally overcoming biological constraints and the potential of a new symbiosis. *Global Science News.*

[27] LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life.* Simon & Schuster.

[28] Mayr, E. (1961). Cause and effect in biology. *Science, 134*(3489), 1501–1506.

[29] Miller, G. F. (2000). *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature.* Doubleday.

[30] Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions.* Oxford University Press.

[31] Pinker, S. (1997). *How the Mind Works.* W. W. Norton.

[32] Reason, J. (1990). *Human Error.* Cambridge University Press.

[33] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

[34] Sapolsky, R. M. (2017). *Behave: The Biology of Humans at Our Best and Worst.* Penguin Press.

[35] Schelling, T. C. (1960). *The Strategy of Conflict.* Harvard University Press.

[36] Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews, 95*(3), 853–951.

[37] Schopenhauer, A. (1818). *The World as Will and Representation.* Brockhaus.

[38] Semmelweis, I. (1861). *Die Ätiologie, der Begriff und die Prophylaxis des Kindbettfiebers.* Hartleben.

[39] Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior, 33*(3), 210–216.

[40] Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie, 20*(4), 410–433.

[41] Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626.

[42] Trivers, R. L. (1972). Parental investment and sexual selection. In B. Campbell (Ed.), *Sexual Selection and the Descent of Man* (pp. 136–179). Aldine.

[43] Wrangham, R. (2019). *The Goodness Paradox: The Strange Relationship Between Virtue and Violence in Human Evolution.* Pantheon.

[44] Weinshall-Margel, K., & Shapard, J. (2011). Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences, 108*(42), E833.

[45] Zuckerman, M. (1994). *Behavioral Expressions and Biosocial Bases of Sensation Seeking.* Cambridge University Press.