Unaware Adversaries: A Framework for Characterizing Emergent Conflict Between Non-Coordinating Agents

Scott VanRavenswaay
scottvr@gmail.com

February 2026

Abstract

In many complex systems, constituent agents engage in adversarial behavior without explicit awareness or intent. These "Unaware Adversaries" ("UA") operate based on localized objectives and feedback, yet their interactions within a shared environment result in mutual disruption, resource competition, or functional conflict. This paper formally defines this phenomenon and situates it within existing literature on systems theory. We present and analyze several case studies—including closed-loop, open-loop, and designed conflicts—to build a taxonomy of these these adversarial typologies. By demonstrating how this framework can be used as a generative tool for analysis and design in addition to diagnostic mitigation, we derive illustrative strategies for problems in fields as diverse as generative modeling and internet routing. The result is a framework intended to help identify, classify, and resolve a pervasive yet under-examined class of system failures.

Keywords: unaware adversaries, emergent conflict, system dynamics, feedback control loops, multi-agent systems, system failures, system design

1 Introduction

Consider a common real-world scenario: an office room equipped with two independent climate control systems. A radiator, governed by a building-wide thermostat, provides heat, while a window-mounted air conditioning unit, with its own separate controls, provides cooling. Each system operates according to its own local feedback loop. If an occupant turns on the A/C to cool a stuffy room while the building's heating system is simultaneously trying to maintain a minimum winter temperature, the two agents enter a state of persistent, mutually negating work—a thermodynamic conflict that neither is designed to recognize. This scenario serves as an intuitive archetype for a class of interactions we term **"Unaware Adversaries" ("UA").**

In this paper, we formalize a study of systems wherein agents function as antagonists without semantic awareness of their opposition. While the term *"adversary"* traditionally implies intentionality, our focus is on interactions where conflict is an emergent property of environmental coupling and feedback dynamics. While these scenarios resemble the Senge's "Accidental Adversaries" archetype[1], this framework focuses on a specific, asemantic subset of conflict, specifically in its treatment of intentionality and awareness. In classical systems dynamics, the adversarial state often arises from a decay in trust or a misinterpretation of a known peer's motives. In contrast, the **UA** framework addresses systems where the coupling is so indirect or the processing constraints so tight that the "adversary" is effectively invisible.

In many technical systems we find that agents often lack any awareness of a *partner* to begin with. We define Unaware Adversaries not by a breakdown in *cooperation*, but by the existence of an **Analytic Budget**, which is the reality that individual contributors and automated systems alike have a finite capacity to interpret environmental signals. Here, conflict is not a relational failure, but an architectural one where the *"noise"* of the environment masks the *"signal"* of the opponent. By analyzing these phenomena, we can better understand and mitigate unintended inefficiencies and instabilities in complex, decentralized systems.

This framework is designed for the practitioner - the network engineer, the SRE, the IT architect - who must troubleshoot failures in environments where formal systems thinking is rarely applied and mathematical

proofs are all but absent. **UA** provides a methodological language to describe the *negative space* between components that appear healthy in isolation but *fail in aggregate*.

2 Background and Related Work

The concept of *unintended systemic conflict* is not new, and the framework of Unaware Adversaries builds upon several established ideas e.g. from systems theory, control theory, economics, public policy. In economics, the notion of **negative externality** describes a cost imposed on a third party. While related, externalities are often unidirectional and do not necessarily involve the reciprocal feedback loop central to our definition of **UA**. In **systems theory**, Peter Senge's work on system archetypes is highly relevant[1]. *Fixes that Fail* and *Shifting the Burden* both describe outcomes of such conflicts. Similarly, the study of **policy resistance** in public administration explores why policies often fail or backfire due to the complex reactions of the system they are intended to manage. Our framework contributes to this body of work by focusing specifically on the agent-level *mechanism* of the conflict: two or more non-coordinating agents, each pursuing a local objective, coupled through a shared environment. We propose that by classifying the nature of the agents, their coupling, and the resulting impact, we can create a more targeted diagnostic and prescriptive tool than a general archetype provides.

3 A Framework for Unaware Adversaries

We formally define **Unaware Adversaries** as a set of two or more agents or systems that satisfy the following criteria:

Independent Objectives Each agent pursues a distinct objective function based on its own state and local environmental variables.

Bounded Processing Capacity Each agent has limited attention, computational resources, or analytical capacity to process available environmental signals. Agents may optimize only over a subset of observable state variables due to cognitive or computational constraints. We argue that bounded processing capacity is not incidental; without it, agents **could**, in principle, infer indirect coupling and cease to be unaware.

Shared Environment The agents co-exist and operate within a common environment, defined by a set of shared physical, informational, or resource-based state variables.

Indirect Coupling The actions of one agent affect the state of the shared environment, thereby influencing the sensory inputs and subsequent actions of the other agent(s). There is no explicit communication protocol for de-confliction.

Antagonistic Effect The collective behavior of the agents results in a net negative impact on the fulfillment of their respective objectives or the overall system operation (i.e., efficiency, reliability, cost.)
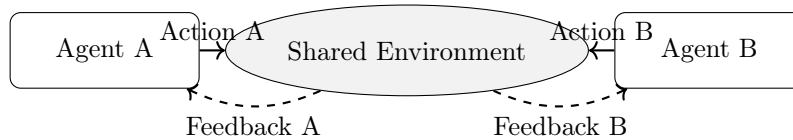

Figure 1: Generic model of unaware adversaries: agents act on a shared environment, receive feedback only via its state, and have no direct communication or awareness of their mutual antagonism.

4 Examples

The following models are intentionally simplified to expose coupling structure rather than to serve as predictive simulations. The mathematics used are illustrative and are not themselves novel analytical formulations deserving of scrutiny.

4.1 Thermostatic Conflict: The Heater-A/C Dyad

Our canonical example involves a room containing two physically distinct and uncoordinated devices: a radiator governed by a building-level thermostat and a window air-conditioning unit with its own integrated controls. This situation is common in older buildings or offices where supplemental climate control is added. Conflict arises when the setpoints are in opposition; for example, the building heat is set to maintain 22°C while the window A/C is set by an occupant to cool the room to 20°C. If the room temperature drifts to 21°C, both systems will activate, leading to a direct conflict. We can model the temperature dynamics and energy cost as follows:

$$\frac{dT}{dt} = -k(T - T_{\text{ext}}) + P_H u_H(t) - P_C u_C(t)$$

$$\text{Cost } (J) = \int_0^{t_{\text{final}}} (C_H u_H(t) + C_C u_C(t)) \, dt$$

Here, $u(t)$ is the control signal (1 if on, 0 if off) for the heater (H) and cooler (C), P_H and P_C are their respective thermal powers, and C_H and C_C are their electrical power draw (W). The total energy cost, J, in the *adversarial configuration is orders of magnitude higher than in a stable configuration, for no improvement in temperature stability.*

4.2 Email Deliverability Degradation

A compelling open-loop example occurs in corporate IT systems. Consider two agents sending email from the same domain (`@company.com`):

- **Agent A (Marketing Platform):** Its objective is to maximize email sends for a campaign, blasting 500,000 emails in an hour.

- **Agent B (Transactional System):** Its objective is to ensure reliable delivery of critical password reset or receipt emails.

The shared environment is the company's domain reputation with email providers (e.g., Google, Microsoft). Agent A's action (the email blast) degrades this reputation, causing Agent B's critical emails to be marked as spam or blocked. The feedback loop is open because the failure of the password reset emails provides no direct signal back to the marketing platform. Agent A's performance metrics are unaffected by the harm caused to Agent B, so it continues its behavior without correction.

4.3 Auto-scaling conflict

Consider a microservices architecture where two services independently manage their resource allocation:

- **Service A** (API Gateway): Auto-scales based on CPU usage, adding instances when CPU > 80

- **Service B** (Database): Auto-scales based on request queue depth, adding capacity when queue > 100 requests

Both services have access to comprehensive metrics dashboards showing:

- Their own performance metrics

- Their peer's scaling activities

- System-wide resource utilization

- Historical scaling patterns

The information is available, but neither service possesses the logic to recognize that their peer's scaling decision invalidates their own. When Service A scales up due to high CPU, it generates more database requests. Service B observes its queue filling and scales up. The additional database capacity briefly reduces queue depth, which reduces back-pressure on Service A. Service A interprets this as successful load handling and scales up further (or perhaps a more likely configuration would have it do the opposite - scale down as cost savings measure. This distinction should not be important for our point, which is that it creates unintended oscillations.) The cycle repeats, creating cascading over-provisioning.

4.4 Systemic Opposition in Urban Planning

Perhaps the most societally significant examples of Unaware Adversaries occur in socio-technical systems such as urban planning. Consider two municipal departments with laudable but conflicting mandates:

- **The Transportation Department:** Tasked with mitigating traffic congestion, its primary Key Performance Indicator (KPI) is maintaining a high Level of Service (LOS) on major arterial roads, a standard defined in the *Highway Capacity Manual* [8]. A high LOS corresponds to low traffic density and high travel speeds. The department's primary tool is increasing road capacity (K) by adding lanes.

- **The Urban Planning Department:** Tasked with addressing housing shortages and promoting sustainability, its goal is to encourage dense, walkable, Transit-Oriented Development (TOD).

These two locally-optimized strategies become potent unaware adversaries. The core of the conflict can be formally captured using standard models from transportation engineering. A widely used relationship to model the effect of congestion on travel time is the Bureau of Public Roads (BPR) formula [9]:

$$t = t_0 \left[1 + \alpha \left(\frac{V_{\text{base}} + \gamma K}{K} \right)^{\beta} \right]$$
$$= t_0 \left[1 + \alpha \left(\frac{V_{\text{base}}}{K} + \gamma \right)^{\beta} \right]$$

Here, t is the actual travel time on a road segment, t_0 is the free-flow travel time (with no traffic), V is the traffic volume (vehicles per hour), and K is the practical capacity of the road. The parameters α (typically ≈ 0.15) and β (typically ≈ 4) determine the severity of congestion effects. The Transportation Department's objective is to minimize t by increasing K. Here, K is our capacity expansion proxy, in lane-km.

Concurrently, the Planning Department approves dense residential projects. The immediate effect is an increase in the base number of trips originating in the area, contributing to a higher traffic volume V. However, the more insidious coupling occurs through the phenomenon of **induced demand**. The Transportation Department's expansion of a highway temporarily reduces travel time t, making driving faster and more attractive. This, in turn, fuels suburban growth and causes more people to drive, increasing the total traffic volume V.

This effect has been empirically verified and is known as the "Fundamental Law of Road Congestion," which states that the total vehicle-kilometers traveled (VKT) in a metropolitan area increases in direct proportion to the available lane-kilometers of roadways [10]. We can model this relationship simply as:

$$V = V_{\text{base}} + \gamma K$$

where V_{base} represents the volume from existing development and γK represents the new volume induced by the capacity expansion itself.

The adversarial loop is now clear. Substituting the induced demand model into the BPR formula reveals

the systemic trap:

$$t = t_0 \left[1 + \alpha \left(\frac{V_{\text{base}} + \gamma K}{K} \right)^\beta \right] = t_0 \left[1 + \alpha \left(\frac{V_{\text{base}}}{K} + \gamma \right)^\beta \right]$$

When the Transportation Department increases capacity (K), the term $\frac{V_{\text{base}}}{K}$ decreases, which is the intended effect of alleviating existing congestion. However, the presence of the induced demand constant γ means that as K becomes very large, the ratio inside the parentheses does not approach zero, but rather approaches γ. The travel time t does not return to the free-flow time t_0, but instead approaches a new, congested equilibrium:

$$t \to t_0 \left(1 + \alpha \gamma^\beta \right) \quad \text{as } K \to \infty$$

- *Since γ represents the induced demand coefficient, if $\gamma \approx 1$ (as suggested by the "Fundamental Law of Road Congestion"), the equilibrium travel time remains significantly higher than free-flow.*

Each department successfully meets its own local KPIs—lanes are built, and housing is approved—yet the system as a whole suffers from persistent congestion despite massive capital expenditure. The agents are locked in a state of policy-induced homeostasis, where each one's logical, well-intentioned actions systemically negate the efforts of the other.

4.5 Interdomain Routing (BGP) Instability

The dynamic of independent agents pursuing local optimizations at the expense of global system health extends from urban planning to the core architecture of the internet itself. The Border Gateway Protocol (BGP) [13] governs interdomain routing across the Internet, enabling thousands of autonomous systems (ASes) to exchange reachability information. While BGP is highly flexible, its decentralized and policy-driven design creates opportunities for emergent instability. This subsection presents a general case of unaware adversarial interaction in interdomain traffic engineering, followed by a concrete, high-impact example: the 2008 YouTube/Pakistan BGP hijack incident.

4.5.1 General Case: Unaware Traffic Engineering Between ASes

Consider two neighboring ASes:

- **AS A (Transit Provider):** Uses internal link metrics and MEDs [14] to load-balance traffic and minimize internal congestion.

- **AS B (Content Provider):** Optimizes for inbound performance by modifying BGP community tags and selectively announcing prefixes.

Both systems reactively adjust (typically via some mechanism outside of BGP itself) [16] their routing announcements based on local utility functions—CPU load, latency, throughput—but do so without coordination. Their interaction takes place through BGP advertisements and path selection behavior, producing a shared environment susceptible to instability.

Each AS modulates its route announcements as a function of observed performance:

$$R_A(t + \Delta t) = f_A(R_B(t), L_A(t)) \quad \text{and} \quad R_B(t + \Delta t) = f_B(R_A(t), L_B(t))$$

This coupling forms a delayed feedback loop. The system may enter oscillatory regimes or route flap storms, degrading routing stability and increasing global convergence latency [11].

4.5.2 Specific Case: The YouTube/Pakistan Hijack (2008)

In February 2008, Pakistan Telecom (AS17557) attempted to implement a government-ordered block of YouTube by announcing a more-specific prefix (208.65.153.0/24) intended for internal blackholing. This announcement was mistakenly propagated to PCCW (AS3491), a major transit provider, which further distributed the route across the global Internet.

Due to BGP's path selection rules, the hijacked prefix—which was more specific than YouTube's legitimate announcement—was globally preferred, redirecting worldwide traffic to a null route in Pakistan. YouTube was rendered inaccessible for several hours [12].

This incident demonstrates classic unaware adversarial dynamics:

- **Independent Objectives:** Pakistan Telecom sought national censorship; PCCW aimed to maintain routing continuity and performance.

- **Shared Environment:** The global BGP routing table.

- **Indirect Coupling:** PCCW's propagation choices were influenced by AS17557's unexpected prefix origin.

- **Antagonistic Effect:** Global denial of service for YouTube, triggered by local policy and lack of propagation checks.

5 A Taxonomy of Unaware Adversarial Systems

To demonstrate the analytical power of this framework, we now apply the taxonomy to our diverse Examples. As shown in Table 1, this classification not only organizes the examples but also reveals critical insights into the underlying dynamics of each type of conflict and prescribes a UA Principle for mitigation strategy.

Table 1: Taxonomy Analysis of Case Studies

Case	Domain	Coupling	Feedback	Complexity	Symmetry	Principle
4.1	Mechanical	Physical	Closed-Loop	First-order	Symmetric	Hierarchy
4.2	Digital	Informational	Open-Loop	Higher-order	Asymmetric	Observability
4.3	Digital	Informational	Closed-Loop	Higher-order	Symmetric	Observability
4.4	Social	Environmental	Closed-Loop	Multi-order	Asymmetric	Incentive
4.5	Net Infra.	Informational	Semi-Open	Higher-order	Asymmetric	Observability
4.6	Digital	Informational	Closed-Loop	Higher-order	Asymmetric	Incentive

The taxonomy reveals a critical distinction: in the open-loop conflict, the harm is asymmetric and the adversarial agent receives no corrective feedback, suggesting such problems can persist undetected for long periods. This contrasts with the noisy, symmetric inefficiency of the closed-loop thermostat conflict.

The BGP case introduces a "semi-open-loop," where an agent receives operational feedback that its actions have propagated (the route was accepted) but remains blind to the wider, negative systemic consequences of that action (the denial of service).

6 Principles

An understanding of Unaware Adversaries does more than explain a class of failures; it provides a powerful set of principles for designing robust, efficient, and intelligent systems. By consciously analyzing the potential for emergent conflict, designers can move from a reactive posture of debugging and mitigation to a proactive one of thoughtful system architecture. The framework allows us to ask critical questions at the design stage: Who are the agents in this system? What environment do they share? How are they coupled? And most importantly, how could their local objectives lead to global conflict? The answers lead to a set of clear design principles.

6.1 The Observability Principle

- Make System-Wide Effects Locally Apparent

- **AND** provide signals in a form that matches the agent's processing capacity

- **AND** explicitly surface higher-order implications that agents might not compute themselves

This principle directly targets the most insidious conflicts: the open-loop, asymmetric cases where one agent operates completely unaware of the harm it causes. The solution is to create a feedback channel where one does not naturally exist.

6.2 The Incentive Principle: Redefine "Winning" to Induce Cooperation

For conflicts where agents have fundamentally misaligned goals, imposing external controls can be brittle and inefficient. A more elegant and robust solution is to redesign the agents' objective functions so that cooperation becomes an emergent property of them pursuing their own, revised goals.

This is the most transformative principle. Instead of just adding a coordinating layer, *change what it means for an agent to succeed.*

6.3 The Hierarchy Principle: Appoint a Single Arbiter for Contested States

When agent objectives are simple, directly oppositional, and tightly coupled through a physical state, the most effective solution is often to remove autonomy and establish a clear control hierarchy.

By applying these principles, the framework of unaware adversaries becomes a generative tool. It equips a designer to anticipate failure modes and architect systems that are not only resilient to conflict but are designed from the ground up for emergent cooperation.

7 Case Studies

Here we will apply the UA Taxonomy and Principles to the Examples in Section 4:

7.1 Thermostatic Conflict: The Heater-A/C Dyad (Example 4.1)

The **UA** Solution: **The Heater-A/C dyad** is a perfect example of when to apply The Hierarchy Principle. Two simple agents are fighting a direct, symmetric battle over a single, continuous variable: room temperature. Their control logic is too simple to negotiate a truce. Institute a higher-level arbiter or a "master controller" that has authority over both agents and manages the contested resource. The solution is a single, smart thermostat that controls both the heat and the A/C. It becomes the sole source of truth for the room's temperature and can make intelligent decisions (e.g., defining a deadband between heating and cooling setpoints) that are impossible for the independent agents. This principle can apply to any system fighting over a fungible resource, from CPU cycles in a computer to water rights in a river basin.

7.2 Email Deliverability Degradation (Example 4.2)

The **UA** Solution: In our **Email Deliverability** example, the marketing platform's performance metrics (e.g., emails sent) are decoupled from the negative externality it creates (degraded domain reputation). The agent is flying blind to the true system-wide cost of its actions. Apply the Observability Principle. Instrument the system to make the shared state visible and, crucially, to integrate that state into each agent's local decision-making context. Don't just provide the marketing team with a dashboard showing their send volume; create a shared, top-level "Domain Health" or "Deliverability Score" that is a KPI for *both* the marketing team and the IT department. When the marketing platform initiates a blast, it should see a real-time dip in a metric it is also responsible for. This makes the invisible visible and forces the unaware adversary to reckon with the consequences of its behavior.

7.3 Auto-scaling Conflict (Example 4.3)

The **UA** Solution: The **Auto-scaling Conflict** exemplifies pure interpretation failure. The feedback loop is closed—both agents receive real-time signals about each other's behavior. But recognizing that "my peer's scaling activity is negating my scaling decision" requires:

1. Causal reasoning across service boundaries

2. Modeling the indirect coupling between CPU load and database queue depth

3. Maintaining temporal context to detect oscillatory patterns

An individual service's auto-scaling logic is optimized for computational efficiency: a simple threshold-based rule executing in milliseconds. The interpretive burden of higher-order analysis—detecting mutual invalidation across a delayed feedback loop—exceeds its analytical budget.

By the Observability Principle, our solution here is not merely to surface *more* metrics, but to pre-compute and inject the higher-order signal: a "scaling coordination score" that explicitly flags when peer scaling activities are causally linked. This transforms a complex inference problem into a first-order observable signal.

7.4   Urban Planning (Example 4.4)

The **UA** Solution: In the **Urban Planning** example, each agent (Department) "winning" by its local definition leads to collective failure (gridlock). *"Instead of just adding a coordinating layer, change what it means for an agent to succeed."* In this context, it would mean replacing department-level KPIs like "Level of Service" and "Housing Units Approved" with a single, shared, system-level objective, such as "Minimize average citizen commute time" or "Maximize accessibility score." This forces the departments to find a cooperative, globally optimal solution. This represents not only misaligned incentives, but also a limited optimization scope.

We can add a formal constraint of a **Bounded Processing Capacity** to call out the problem:

Let $A = \{a_1, a_2, \ldots, a_n\}$ be the set of all relevant state variables. Formally, each department optimizes over a strict subset of the relevant state variables; induced-demand dynamics fall outside both objective scopes.

- **Transportation Dept:** optimizes over $\{\text{LOS}, \text{capacity } K\}$

- **Planning Dept:** optimizes over $\{\text{housing units}, \text{density}\}$

- **Neither:** optimizes over $\{\text{induced demand } \gamma, \text{long-term VKT}\}$

Each agent (Department) requires an appropriate Attention Budget so that Induced Demand and VKT have bandwidth to process these information signals.

8   Case Studies (Generative)

In the next two **Case Studies**, we intentionally extend the analysis beyond diagnosis and mitigation to sketch speculative constructions. These are not presented as finished algorithms or protocol proposals, nor as claims of empirical superiority. Rather, they serve a narrower purpose: to demonstrate that the **UA** framework is not merely descriptive, but generative - capable of inspiring non-obvious architectural ideas once a system's conflict structure is made explicit.

8.1   Interdomain Routing (BGP) Instability 4.5

We show a simple demonstration of how routing stability and global convergence latency can increase under normal operation.

8.1.1   In the General Case (Ex. 4.5.1)

The **UA** Solution: Here, we demonstrate the framework's usefulness by sketching a hypothetical mitigation strategy for interdomain routing: the **Gradient-Paced BGP Advertiser (GPBA)**. Rather than impose fixed rate limits, GPBA adapts the pace of route advertisement based on the environmental sensitivity observed at the peer:

$$\Delta t_{\text{propagation}} \propto \left\| \frac{dR_{\text{remote}}}{dR_{\text{local}}} \right\|$$

This application of the **Observability Principle** discourages volatile updates by throttling agents whose announcements produce outsized ripples in routing state. Unlike traditional filters, GPBA uses a reflexive meta-policy: it applies only when sensitivity is high and relaxes when peers demonstrate robustness or stability. This maintains responsiveness while discouraging feedback-amplifying behavior.

8.1.2 Could GPBA Have Mitigated The Youtube/Pakistan Incident of Example 4.5.2?

A GPBA-aware edge at PCCW could have detected that:

1. The origin of 208.65.153.0/24 (AS17557) was highly anomalous.

2. The route produced a high volatility score in terms of downstream reachability change.

Rather than immediately propagate the route, the GPBA heuristic would have increased the propagation delay or required additional validation. This passive reflexivity—without requiring global RPKI [15] or semantic understanding of the policy intent—could have bought time for human intervention or for automated anomaly detectors to act.

The YouTube/Pakistan incident underscores the vulnerability of loosely coordinated protocols like BGP to unaware adversarial dynamics. GPBA-style mechanisms offer a path forward by embedding environmental reflexivity into protocol behavior. Rather than controlling agent behavior through centralized rules, GPBA designs the feedback surface itself to discourage destabilizing actions. In the language of our design principles, it represents a sophisticated form of the Observability Principle, where the system observes its own sensitivity to change and creates feedback to regulate it.

8.2 From Designed Conflict to a Novel Research Hypothesis

The standard Generative Adversarial Network [2] provides a powerful case study for our framework. It is a system of two agents, a Generator (G) and a Discriminator (D), locked in a designed, zero-sum game. This adversarial dynamic, however, is notoriously unstable and suffers from practical issues like vanishing gradients, where D becomes too proficient, leaving G with no learning signal. The original authors' first solution was the heuristic "non-saturating" loss, an immediate modification that sought a stronger, more reliable gradient for G. This established the central challenge in the field: managing the adversarial dynamic for stable and efficient training.

In the years since, the dominant paradigm for GAN stabilization has become one of **gradient control**. Landmark models like Wasserstein GAN (WGAN) [3] and its successor WGAN-GP [4] diagnosed the problem as being rooted in the geometry of the loss landscape. Their solution, which now represents the state-of-the-art, is to *tame and constrain* the discriminator's function (e.g., by enforcing a Lipschitz condition) to guarantee that it always provides a smooth and informative gradient to the generator. This philosophy is about preventing conflict from becoming destructive by carefully limiting the power of the adversary.

The Problem: In the standard GAN example, each agent "winning" by its local definition leads to instability (vanishing gradients).

The Design Solution: Our proposal for the **Pedagogical GAN** is the canonical example: we shift the Generator's objective from "fool the Discriminator" to "maximize the Discriminator's learning rate." This reframes the adversary not as an opponent to be defeated, but as a student to be taught efficiently.

Our framework of unaware adversaries prompts a different line of inquiry. Instead of asking, "How do we control the conflict?", we ask, "Can we redesign the agents' objectives to make the conflict more productive?" This leads us to sketch a perhaps novel approach that stands in philosophical opposition to gradient control. We term this the **Pedagogical GAN** here.

The core idea of the Pedagogical GAN is to change the generator's objective from simply fooling the discriminator to actively teaching it as efficiently as possible. We formalize this by proposing that

the generator should seek to *maximize* the discriminator's learning signal. The generator's objective function becomes:

$$\max_G \|\nabla_D \mathcal{L}(D, G)\|_2$$

Here, $\mathcal{L}(D, G)$ is the standard discriminator loss. The generator is now explicitly incentivized to find samples that lie on the steepest parts of the discriminator's loss landscape. It becomes a "Socratic tutor" that seeks to *weaponize* the gradient for accelerated learning, not suppress it.

This approach represents a conceptual departure from gradient-control-based stabilization. It is distinct from other cooperative frameworks, like Cooperative Teacher or Feature Map Distillation [19] and Unrolled GANs [5], which use strategic foresight, or other non-antagonistic models that alter loss functions to escape the zero-sum game [6]. Instead, it can be viewed as the principled and extreme conclusion of the line of thinking that began with the very first non-saturating GAN loss. We have not found this exact objective stated explicitly elsewhere; regardless, we only use it here as a conceptual exemplar of framework-driven ideation.

(Here we will acknowledge - to minimize the risk of causing any confusion - that being adversarial is an in-built aspect of GAN architecture, and that continuing to use the "Unaware Adversary" branding without such acknowledgement may be seen as ironic in this case.)

As with the simplified mathematical models used throughout, these constructions should be read as conceptual exemplars, not as validated designs of network protocol algorithms or GAN architecture. Their value lies in showing what kind of ideas fall out of the framework, not in asserting that these particular instantiations are ready for deployment or adoption.

8.3 Discussion

The primary contribution of this paper is a framework that serves not only as a descriptive and diagnostic tool but as a generative engine for system design. We demonstrated this generative capacity twice over, deriving plausible and interesting solution ideas in two disparate domains: the **Pedagogical GAN** for machine learning and the **Gradient-Paced BGP Advertiser (GPBA)** for network protocol design. Both concepts emerged directly from applying our framework's lens to a well-known problem. In each case, instead of accepting the standard adversarial dynamic as given, the framework prompted a fundamental question: "How can we redesign the agent's interaction rules to make the conflict less destructive or even productive?" The proposed solutions—one that reframes training as teaching, the other that paces communication based on systemic sensitivity—are deliberate architectural choices about the nature of the agents' interaction. They represent a significant philosophical departure from conventional approaches that focus on adversarial suppression or fixed rate-limiting.

While the two proposed solutions would require empirical validation to prove if they have any worth, that is not the point of their inclusion. Rather, their conceptual origins provide evidence of the framework's creative value, as the direct result of applying one or more of the three core design strategy Principles that emerge from our analysis.

9 Conclusion

9.1 A Complementary Lens for Systems Design Thinking

The concept of unaware adversaries is especially salient today. As we continue to build ever more complex, decentralized systems—from microservice architectures and the Internet of Things to multi-agent AI ecosystems and smart cities—the potential for emergent, unintended conflict grows exponentially. Our traditional engineering and debugging practices, which focus on verifying the correctness of individual components, are insufficient for addressing failures that arise from the negative space *between* those components. The problem is not that the transportation department's models are wrong or that the marketing platform's code has a

bug; the problem is that the system's architecture creates incentives for globally destructive, locally optimal behavior.

Ultimately, this framework encourages a crucial shift in perspective for architects and designers. It moves the focus from the design of agents in isolation to the design of the environment they share and the incentive landscape they inhabit. By analyzing the structure of the coupling and the nature of the feedback loops, we can stop being surprised by "policy resistance" or "unintended consequences" and start architecting systems where cooperation is not an afterthought, but an emergent property of the system's fundamental design.

9.2   Relation to Classical Control Theory

Many of the systems described in our case studies (oscillating on/off controllers, pathological GAN dynamics, BGP route convergence) are familiar terrain in classical control theory. Such systems can often be modeled as composite dynamical systems and analyzed using tools like Lyapunov stability, observability, and controllability [17].

A compelling example of this distinction comes from classical PID control. The P, I, and D terms can be viewed as three "unaware adversaries" - independent control laws with different temporal objectives (present error, accumulated error, rate of change) simultaneously acting on a shared output. (This is an analogy rather than a literal agent decomposition; the value is in highlighting architectural coupling, not intentional autonomy. Unlike the other case studies, PID terms are intentionally designed to counterbalance each other—their interaction is coordinated by design, even if we can analyze them through an unaware-adversary lens.) While stability analysis tells us under what conditions this composite system converges, our framework asks: what if the P term's aggressive response triggers instability in the D term's derivative estimate? This is not a question of mathematical stability but of architectural choice about how to structure agent interactions.

Another distinction from this formal standpoint, is that while the existence (or absence) of a global Lyapunov function may determine whether the system exhibits convergence to equilibrium, our framework highlights that stability alone is not the only goal. *A system can be stable and still fail.* Many real-world adversarial pathologies - such as traffic congestion equilibria [7], or GANs collapsing to degenerate outputs [2, 3] - are cases where systems settle into *undesirable equilibria, not chaotic instability.*

Framing these as systems of "Unaware Adversaries" surfaces a complementary perspective that classical tools, by design, abstract away: that each agent is acting rationally within its own feedback loop, but their combined behavior produces outcomes no one intended and no one benefits from. This is not merely a problem of instability; it is rather a problem of emergent misalignment.

What our framework offers is not a substitute for classical analysis, but a formalized way to highlight systems that look stable on paper but can exhibit pathological behavior due to unrecognized structural conflicts between agents. Classical control theory excels when systems can be modeled from first principles and the question is convergence, stability margins, or optimal control law design. The Unaware Adversaries framework is most valuable when:

1. the system is too complex for tractable first-principles modeling (e.g., organizational dynamics, multi-stakeholder systems),

2. the agents themselves are designed/intelligent entities with goals (not just passive dynamics), or

3. the pathology lies in emergent misalignment between rational actors rather than parametric instability.

The frameworks are thus complementary - classical theory for well-defined dynamical systems, unaware adversaries for socio-technical systems where intentionality matters.

For example, the Pedagogical GAN proposed earlier does not reject the gradient-based framing of GAN dynamics. Instead, it reframes the design question: what if we architected the generator and discriminator not as antagonists, but as collaborators with asymmetric knowledge? This reframing doesn't negate formal stability concerns, it simply shifts attention toward the shape and utility of the learning signal, and not solely its convergence properties.

While the framework's Observability Principle emphasizes making system-wide effects visible, visibility alone is insufficient when agents lack the processing capacity to interpret complex or higher-order signals. Consider the Email Deliverability case: the marketing platform could theoretically monitor global SMTP rejection rates, DNS blacklist statuses, and reputation scores. But recognizing that their blast is the cause of degraded reputation requires:

1. temporal correlation analysis,

2. understanding of cumulative effects, and

3. attribution across delayed feedback.

This interpretive burden may exceed the agent's analytical budget. Effective mitigation strategies must account not only for what information agents can access, but what information they can process and act upon given their bounded rationality. Signals must be pre-processed, aggregated, or simplified to match agent capabilities—or alternatively, agents must be equipped with sufficient analytical resources to extract meaning from raw environmental data.

The tools of nonlinear systems theory - comprehensively covered in foundational texts such as Khalil's Nonlinear Systems [17] - provide the rigorous formalisms for system stability and convergence. When such modeling is tractable, these methods should be the first line of analysis. But when full modeling is infeasible, or when coordination failure arises from structural blindness rather than stochastic noise, our lens provides a complementary mode of reasoning that emphasizes intentionality, partial observability, and emergent conflict.

If classical control tells us whether a system will converge, the framework of unaware adversaries helps us ask: to what, and why?

10 Directions for Future Research

This framework opens several avenues for future inquiry, most notably:

- The formal implementation and empirical validation of the proposed **Pedagogical GAN**. This would involve developing the necessary regularization techniques to stabilize the training dynamic and testing its performance against baseline models and gradient control methods like WGAN-GP.

- Analysis of unaware conflict in large-scale distributed AI systems, such as competing recommendation algorithms on a single platform.

- Application of the unaware adversary lens to AI alignment and interpretability, especially in ensembles of independently trained agents.

- The formal specification and empirical validation of the proposed **Gradient-Paced BGP Advertiser (GPBA)**. This would involve simulation in environments like BGPStream to test its efficacy in dampening route flaps and mitigating the propagation speed of malicious hijacks compared to existing mechanisms like route flap damping.

- **As the very concept of *"What is an Agent?"* seems to be on the verge of an enormous change with the advent of Agentic AI, the Unaware Adversaries Framework - and many others - will no doubt need to be updated with a special focus on the radical shift in the meaning of *Agent, Intent, Adversarial, etc.* that is developing now**.

11 Conclusion

The concept of unaware adversaries provides a valuable lens for analyzing and resolving a significant class of system failures. By moving beyond debugging individual components to architecting the environment in which they interact, we can mitigate and even prevent emergent, destructive conflict. This paper defined a

formal framework for this analysis and demonstrated its utility across mechanical, digital, and socio-technical systems. Crucially, we have shown this framework to be more than just descriptive; it is a generative tool for innovation. The derivation of two distinct and plausible solutions, the Pedagogical GAN and the Gradient-Paced BGP Advertiser, proves that a principled understanding of emergent conflict can inspire novel designs for more robust, stable, and cooperative systems.

Acknowledgments

References

[1] Peter M. Senge. *The Fifth Discipline: The Art & Practice of The Learning Organization*. Doubleday, 1990.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Ishan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, pages 5767–5777, 2017.

[5] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[6] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. Cooperative Training of Descriptor and Generator Networks. *arXiv preprint arXiv:1609.09408*, 2016.

[7] Anthony Downs. *Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*. Brookings Institution Press, 1992.

[8] Transportation Research Board. *Highway Capacity Manual, 7th Edition: A Guide for Multimodal Mobility Analysis*. National Academies of Sciences, Engineering, and Medicine, 2022.

[9] Bureau of Public Roads. *Traffic Assignment Manual*. U.S. Department of Commerce, Urban Planning Division, 1964.

[10] Gilles Duranton and Matthew A. Turner. The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review*, 101(6): 2616–52, 2011.

[11] Craig Labovitz, Abha Ahuja, Abhijit Bose, and Farnam Jahanian. Delayed Internet Routing Convergence. In *Proceedings of ACM SIGCOMM*, 1997.

[12] Jim Cowie. *Pakistan Hijacks YouTube*. Renesys Blog, February 24, 2008. Available at: `https://web.archive.org/web/20210506213233/https://www.renesys.com/2008/02/pakistan-hijacks-youtube-1/` (archived from the original).

[13] Y. Rekhter, T. Li, and S. Hares, editors. *A Border Gateway Protocol 4 (BGP-4)*. Request for Comments 4271, Internet Engineering Task Force, January 2006. `https://datatracker.ietf.org/doc/html/rfc4271`.

[14] D. McPherson, V Gill, editors. *BGP MULTI_EXIT_DISC (MED) Considerations.* Request for Comments 4451, Internet Engineering Task Force, March 2006. `https://datatracker.ietf.org/doc/html/rfc4451`

[15] R. Bush, R. Austein, editors. *The Resource Public Key Infrastructure (RPKI) to Router Protocol, Version 1* Request for Comments: 8210, Internet Engineering Task Force (IETF), September 2017. `https://datatracker.ietf.org/doc/html/rfc8210`

[16] X. Xu, S. Hegde, K. Talaulikar, M. Boucadair, C. Jacquenet, editors. *Performance-based BGP Routing Mechanism draft-ietf-idr-performance-routing-04* Internet-Draft, Internet Engineering Task Force (IETF), August 2024. `https://www.ietf.org/archive/id/draft-ietf-idr-performance-routing-04.txt`

[17] Hassan K. Khalil. *Nonlinear Systems, 3rd ed.* Prentice Hall 2002

[18] Hsu, Jules C. *Control Systems: An Introduction with MATLAB* McGraw-Hill, 1998.

[19] Tie Hu and Mingbao Lin and Lizhou You and Fei Chao and Rongrong Ji. *Discriminator-Cooperated Feature Map Distillation for GAN Compression* arXiv 2212.14169. `https://arxiv.org/abs/2212.14169`