Adversarial Design Thinking for Organisational Architecture

Richard Farry

QinetiQ, UK

SUMMARY

Organisations can unintentionally create friction, dysfunction, and harm through the design of their structures, processes, and information flows. This paper introduces Adversarial Design Thinking, a parallel-design method that applies a malicious-insider mindset to organisational architecture to reveal these hidden vulnerabilities. A Red Team is tasked to design solutions that meet stated goals while maximising plausible, undetected organisational harm, while a Blue Team designs conventionally. Comparing their outputs surfaces latent risks, structural weaknesses, and unintended consequences that human-centred approaches—often assuming good intent—may overlook. The paper presents the GHOST and Harm frameworks to support identification of adversarial design patterns, showing how organisational features can hide harm, degrade recovery, and allow dysfunction to accumulate. This lens strengthens organisational resilience and design quality.

KEYWORDS

Dark Patterns, Malicious Insider Threats, Red-teaming, Adversarial Design Thinking

Introduction

Adversarial Design Thinking (ADT) is introduced here as a parallel-design lens for organisational architecture. One team (Blue) pursues the stated goal conventionally, while a second team (Red) adopts a malicious-insider mindset (Red) to produce designs that still meet objectives yet maximise plausible, hard-to-detect (at least at the design stage) organisational harm. The comparison thereby sharpens organisational self-awareness, revealing design flaws and unintended consequences that conventional methods may not bring to light. In this paper, organisational architecture (Abdel-Kader and Lin 2009) is treated as the designed configuration of structures, culture, processes, people, resources, technology, and incentives in an organisation that shape behaviour and performance. The GHOST taxonomy (Gridlock, Hide, Oppress, Skew, Trap) and a complementary Harm framework provide practical scaffolds for generating, classifying, and analysing the adversarial patterns revealed by the comparison of Blue and Red's designs.

During World War II the U.S. Office of Strategic Services (OSS) published the *Simple Sabotage Field Manual* (OSS 1944). It provided guidance for how civilians in occupied territories could degrade or disrupt factories and other parts of industry without direct confrontation or violence. It encouraged civilians to use common workplace behaviours—calling unnecessary meetings, insisting on written instructions, advising caution, blaming work issues on poor quality tools—to subtly sabotage industrial operations. These tactics were effective because they closely resembled legitimate organisational behaviours, making their disruptive impact difficult to detect or challenge. ADT generalises this principle to the design of organisations, examining how an insider with influence over structural or policy decisions could maximise organisational harm while presenting those designs as reasonable and defensible to decision-makers.

Organisational architecture can be understood as an interface between different parts of the organisation: design decisions structure access to information, the distribution of decision rights, and the configuration of incentives, thereby directing behaviour and shaping outcomes. Even without malicious intent, organisations often create friction, misalignment, and resilience-degrading dynamics through the design of policies and structures that appear rational from the perspective of those with power. ADT renders such structural harms more visible by illustrating how organisational virtues—such as thoroughness, caution, standardisation, or accountability—can be subtly weaponised, producing outcomes akin to Dark Design Patterns (Gray *et al*. 2018) in user interface design.

ADT is proposed as a pragmatic, method-lite lens that integrates the adversarial stance of red teaming with insights from dark design patterns and documented organisational dysfunctions. It reveals plausible yet harmful organisational designs that conventional approaches may overlook. It is intended to stimulate different thoughts and approaches to understanding organisational vulnerabilities, and the pathways to organisational dysfunction.

Historical Antecedents

Forms of adversarial thinking, or appreciation of adversarial thinking, to refine one's own approaches or designs is not new, and the proposed approach has antecedents in red teaming and in user-interface dark design patterns.

Red Teaming

Red Teaming is an approach to critically assess and challenge assumptions, generate alternatives, and reduce the risk of Group Think (UFMCS 2018). Arguably, Red Teaming approaches grew out of the Catholic Church's use of the *Promotor Fidei* (Promoter of the Faith)—the Devil's Advocate—whose role it was to challenge claims and evidence of sainthood (Zenko 2015, pp.*i-xii*). The Devil's Advocate is a red team role that helps to test and improve a position (proposition, claim, etc.) through adversarial engagement, forcing critical assessment of the position and often leading to more creative and better-quality outcomes (e.g. Valacich and Schwenk 1995). More broadly Red Teaming can be seen as any simulated adversarial process whereby one seeks to understand what an adversary might do, and to challenge one's own thinking in relation to the adversary's plans and intentions.

Red Teaming has been most visible in military planning, but its use now includes business strategy (Sun et al. 2021), cyber security (Applebaum et al. 2016; Yulianto et al. 2023), and the development of ethical AI systems (Idemudia 2023). Across these applications, the Red Team typically represents an external actor or event, even when modelling insider threats to information systems. Its purpose is to identify what could be done to the organisation. In contrast, the ADT approach focuses on what an organisation might inadvertently or deliberately *do to itself*—whether through malicious insider influence or through error, neglect, ignorance, or organisational dysfunction.

Dark Design Patterns

In the field of Human-Computer Interaction (HCI), user-interfaces designed to benefit some entity (e.g. a company or shareholders) at the *expense of users*, usually via exploiting human psychology (Gray *et al.* 2018), are known as Dark Design Patterns. These patterns produce interfaces that appear benign but are in fact user-hostile, creating harm or inconvenience for those interacting with them. Examples of Dark Design Patterns include artificial scarcity (e.g. deceptive countdown timers or "only 2 left" stock notices that nudge immediate purchase), price presentation that obscures total cost or misleads as to the costs of cancelling a subscription (e.g. annual subscriptions displayed at a monthly-equivalent rate), advertisements that camouflage themselves so that they do not seem to be

adverts, and privacy-leaking defaults in social media settings that expose more data than users reasonably expect (Mathur *et al.* 2021).

Dark Design Patterns are hostile user-interface designs that modify users' choice architecture, either by altering their decision space or manipulating the information flow (Mathur *et al*. 2021, p.8), thus shaping user-behaviour against their own interests. Dark Design Patterns can harm users in a variety of ways, including financial loss, invasion of privacy, cognitive burden, wasted time, and undermining their autonomy (*Ibid*). While in many cases the use of Dark Design Patterns may be with ill-intent, in others they may be inadvertent, or done so in good faith without an appreciation of the impact they will have on users.

Adversarial Design Thinking: Let's Be Bad Guys

Adversarial Design Thinking is a parallel-design method in which a designated Red Team—acting as a plausible malicious insider at the organisational design level—creates solutions that meet the stated objective while attempting to maximise undetected harm. These Red outputs are then compared with Blue designs to reveal vulnerabilities and support the synthesis of a higher-quality organisational design. ADT differs from classical Red Teaming in that it examines what an organisation might *do to itself*, through insider-driven design choices, rather than focusing solely on what external adversaries could do *to it*.

Mindset

This paper proposes a parallel design approach in which a dedicated adversarial Red Team is tasked with solving the same design problem as a Blue Team. The Red Team adopts the mindset of a malicious insider, aiming to introduce harm or bad friction in ways that are plausible, difficult to detect, and unlikely to be attributed to malice. The key question for members of the Red Team is:

> "If I were a malicious insider tasked with designing this feature/thing to degrade the organisation while remaining undetected or unpunished, what would I do?" *

*This could be extended with the auxiliary question "and how would I go about it?"

Organisations currently suffer from policies and processes that degrade their function, stifle innovation or harm employees, such as (but not limited to) bureaucratic sludge (Sunstein 2019), forced performance distribution curves (O'Boyle and Aguinis 2017), perverse incentives (Kerr 1975), structural secrecy (Vaughan 1996), and vendor lock-in. These come about and survive because the underlying policies and processes are considered plausible and appropriate, at least by those who wield organisational power. These unintentional harms can be brought about intentionally by malicious insider threats. The Red Team is not attempting harm on the levels of cartoon villainy—such attempts would be detected and resisted—but rather, as with some of the advice in the *Simple Sabotage Field Manual* (OSS 1944), is seeking to weaponise legitimate organisational aims and virtues.

Approach

The Adversarial Design approach is largely agnostic to frameworks, methods, and tools. It is a 'methodology-lite' approach, only requiring the use of more than one design team, and for one of those teams to adopt a different mindset and goal. The goal is to design the feature/thing in a way that seems valid, plausible, perhaps even attractive as an option, but will bring about harm to the organisation. Harm here can mean any deleterious effect, whether in terms of stakeholder well-being and morale, financial performance, inappropriate risk aversion, poor returns on investment, slow decision-making or decision inertia, resilience, among other factors. Similarly, the harm could be of any magnitude (though the greater the better for the Red Team and the value of the exercise)

and over any timescale or frequency. The challenge in successful Adversarial Design is having the right *mindset* to come up with the harmful approaches.

That said, in addition to the right mindset, frameworks, methods, and tools designed to identify weaknesses in a system that can be exploited by adversaries, competitors, or natural disruptions, will likely better support the Red Team's aims. Examples of these include Human Reliability Analysis (HRA) methods, Work Domain Analysis (WDA), the Human Factors Response Framework (Farry 2020), pre-mortems, and applying threat identification approaches such as STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege) (Kohnfelder 1999) more broadly to an organisation than just to the information-security domain. The key is the adversarial mindset, identifying cases where organisational virtues can be tweaked, pushed a little too far in a particular direction, or otherwise exploited to cause organisational harm.

The ADT approach is as follows:

1. The Red Team is explicitly briefed to produce a design that (a) fulfils the stated goal while (b) maximising plausible, hard-to-detect organisational harm that (c) decision-makers could still consider implementable.
2. Blue Team(s) pursues the same goal using standard design intent and organisational virtues.
3. Parallel design activities proceed independently.
4. Outputs are compared to identify overlaps, surface vulnerabilities, and inform redesign or explicit trade-off decisions.

Although this paper does not prescribe a specific comparison method, the intent aligns with standard practices in parallel design and human-centred evaluation. In most design disciplines, comparing parallel outputs involves examining where the alternative solutions converge, diverge, or rely on similar assumptions. The same principle applies here: comparing the Red and Blue designs helps to reveal features that appear entirely reasonable from the Blue perspective yet carry damaging implications when viewed through the adversarial lens.

The ADT approach can be extended to be more detailed, as with normal design approaches. For example, the Red Team may be asked to persuasively explain the harmful aspects and consequences of their design choices, and the cost-benefit trade-offs could be modelled in detail.

While the use of tactics such as rhetoric, obscuration, obfuscation, or even outright lying about the Red Team's design can serve as a challenge to the host organisation's culture and ways of communicating, the focus on the Red Team's design should be on making it seem plausible to rational evaluation.

It is suggested the ADT approach has the following advantages:

- Adopting an adversarial mindset will surface issues and consequences which might otherwise go unnoticed or unappreciated.
- ADT is a variant of parallel design. In traditional parallel design, multiple designs are produced independently, evaluated, and then elements from across them are brought together into a 'cross-pollinated' design (Fessenden 2024). In adversarial design, the approach is similar, but the Red Team's output is used to identify features to avoid or mitigate.
- Introduces a structured and endorsed means for dissent and challenging decisions and it can be used to challenge both Group Think (Lidwell *et al*. 2023) and Creator Blindness (*Ibid*.), leading to a better design.

- When design features explicitly proposed by a Red Team overlap with the Blue Team's design decision-makers should be more likely to take the risks seriously. When Red Team features appear in the Blue design, framing them as aligned with adversarial intent can reduce optimism bias and prompt the critical question, "Do we really want to do what an adversary/competitor would want us to do?".

Alternative applications of adversarial design thinking include Red Team scrutiny of Blue Team outputs as a review process (as opposed to designing in parallel), iterative challenge cycles, or adversarial audits of existing systems. It may also be the case that the known 'threat' of a competing Red Team will spur Blue to produce better designs in the first instance, encouraging them to identify unintended negative consequences of their design decisions.

Framework

The proposed ADT approach is 'method-lite' as the rich existing set of tools and methods available can (in the main) easily be turned to the malicious purposes of the Red Team. However, even bad guys need a helping hand, so offered here are two thinking tools to support the Red Team in their adversarial design efforts: the GHOST and Harm frameworks.

The GHOST—Gridlock, Hide, Oppress, Skew, and Trap—framework (see Table 1) offers a pragmatic taxonomy of the ways a malicious insider threat could bring about harm to an organisation. The elements of the framework are drawn from real-life organisational dysfunction and failure, as documented in the literature. As such, they reflect design choices that were considered reasonable in context and often aligned with organisational virtues. As illustrated by the *Simple Sabotage Field Manual*, such virtues can provide camouflage that makes organisational harm difficult to detect or challenge. While the specific means of doing so will vary by context and circumstance, at this level of abstraction the research literature offers rich inspiration for those adopting an adversarial mindset.

Several GHOST elements parallel established dark design patterns in HCI: Gridlock aligns with obstruction-type patterns, Hide with misdirection/obfuscation, Skew with incentive/metric manipulation, and Trap with forced-continuity or "roach motel" designs. Although these analogies are illustrative rather than one-to-one, they ground GHOST in a familiar adversarial design logic.

Table 1: The GHOST Framework

GHOST Element	Design Approach	Key Sources
Gridlock *Reduce effectiveness and efficiency, and hinder output.*	**Introduce Sludge** – increase frictions and transaction costs in how the system functions, including expanding the decision-making and approval processes (Thaler 2018 and Sunstein 2019).	Thaler 2018 Sunstein 2019
	Overload Capacities – identify and create throughput bottlenecks and/or generate more workload than elements of the organisation can handle (an internal denial of service attack).	Vaughan 1996
Hide *Hinder the organisation's internal and external situation awareness,*	**Encourage Siloing and Structural Secrecy** – bring about fragmented information ownership, segmented communication channels, a proliferation of information gatekeepers, and introduce knowledge hoarding incentives.	Bechky 2003 Vaughan 1996
	Obfuscate – inhibit clarity and understanding through the use and abuse of language in communication, documentation, and tools, including burying the useful signal amongst noise.	Lutz 1989

GHOST Element	Design Approach	Key Sources
impair the quality of its decision-making.	**Measurement Omission** – avoid collecting data about key processes, states, or outcomes. Hide sludge by not measuring it or cloaking it in higher-level measures. Hide, avoid, or bury bad news, particularly when it relates to organisational harm or dysfunction.	McGoey 2012
Oppress *Degradation of the people element of the business, eroding morale, hindering people from doing the right thing, encouraging conflict and discouraging true collaboration. Encourage a talent exodus.*	**Encourage Zero-Sum Competition Between People** – for example, introduce forced performance ranking systems, show favouritism, assign individualised or hyper-individualised performance goals at the expense of collective goals, introduce scarcity or a scarcity narrative for key resources (e.g. budgets, headcounts, training opportunities, access to leadership).	O'Boyle and Aguinis 2017
	Promote a Power-Oriented Culture – focused on personal power, needs, and glory. Respond negatively or piecemeal to problems. Limit access to information, decision-makers, and key resources. Encourage zero-sum competition between organisational units. Seek to marginalise or otherwise erode informal networks and lines of communication that exist outside of a formal hierarchy.	Westrum 2004; 2014
	Introduce Tensions Between Staff and What They Believe Is Right, in terms of what they feel leadership owes them and what they owe to others inside and outside of the organisation (*betrayal of thémis*). Generally, erode or put staff's psychological contract under strain. Frequently shuffle people around between teams, to hinder the development of informal support networks.	Shay 1995 Topa *et al*. 2022
Skew *Misalign activity from desired goals, leading to dysfunctional activities and outcomes.*	**Normalise Deviance** – put in place incentives and a culture of not abiding to standards, policies, and rules. For example, put in place rules that seem plausible but won't be met, do not hold leaders to account for violations, require people to agree to things that they cannot comply with or fully understand, demand performance levels that cannot be met without deviation from the formal ways of doing things, encourage "we've always done it this way" thinking and approaches.	Sedlar *et al*. 2023 Vaughan 1996
	Introduce Perverse Incentives – unduly measure, track, and reward things that are not the organisation's goals or objectives. Incentivise easily measurable tasks (e.g. support tickets closed) at the expense of more difficult to measure outcomes (e.g. staff are able to do their job effectively). Design incentives that measure and reward the appearance of the outcome, rather than the reality. Promote a 'use it or lose it' culture. Create incentive structures to pressure people to avoid reporting issues because the perceived implicit and negative impacts (e.g. project delays, performance ratings, administrative burden) are too great.	Kerr 1975 Muller 2018
Trap *Commit the organisation to harmful courses of*	**The Roach Motel approach** - commit the organisation to easy to enter but difficult to exit agreements and courses of action, e.g. vendor lock-in, out-sourcing key skills, introducing complex dependencies, and long-term agreements with no easy termination or enforcement options.	Gray *et al*. 2018

GHOST Element	Design Approach	Key Sources
action and limit its future choices and available resources.	**Use up or deplete resources –** over-commit resources (e.g. money, effort, reputation) to a course of action, particularly a risky course of action, and escalate commitment to make it difficult for the organisation to change direction as its resources to extract itself or try something different have been used up or committed to the project. Aim to use up long-term or strategic assets (e.g. money, land, employee goodwill) for short-term gains.	Staw 1981

The Harm Framework (see Table 2) provides a set of features of organisational harm that a malicious insider threat (or Red Team) should seek to maximise, organised along two dimensions: (1) epistemic opacity—how harm remains hidden, and (2) recovery degradation—how harm reduces detection, correction, and recovery capacity. The Harm and GHOST frameworks align, with GHOST describing the routes through which harm is introduced and the Harm framework characterising how that harm remains hidden, persists, or becomes difficult to reverse.

Table 2: The Harm Framework

The Harm…	Mechanism by which the harm operates
Hiding or obscuring the Harm (Epistemic Opacity)	
Is deniable as deliberate Immunity from correction – the actor can continue to bring about harm elsewhere and/or lessons are not appropriately identified or learned.	The causes/harm can be explained away as incompetence, caution, compliance with rules, promoting an organisational virtue, or the blame can be shifted to external factors such as 'changing requirements'.
Is difficult to detect The organisation becomes accustomed to the degraded state (e.g., minor delays or errors) and redefines it as acceptable, failing to recognize it as a deviation from safety or performance standards.	The effects have one or more of the following attributes: sporadic, widespread, varied, below the threshold of detection, below the threshold of significance, subtle, fall across organisational silos/responsibilities, impact organisational features/activity which are not monitored or measured, a setup or culture that leads to blaming 'human error' rather than systemic issues, inflicts harm only on those without power to bring about change, are cloaked in unclear or ambiguous language.
Is difficult to quantify The harm may go unrecognised by those with the power to correct it, or the costs are undervalued or not fully understood, leading to inaction.	The effects have one or more of the following attributes: qualitative, soft, unmetricated, subjective, have a psychological rather than physical impact.
Degrading or Preventing Recovery	
Damages the organisation's ability to detect the harm The nature of the harm makes it harder for the organisation to detect the harm, or other organisational issues and dysfunctions.	Loss or degradation of psychological safety/just culture, breakdown of cross-functional information flow; over-reliance on metrics that mask weak signals; routinisation of "workarounds" that hide underlying process failures; cultural drift toward treating anomalies as normal; inconsistent or selective visibility of operational data.

Damages the organisation's ability to fix the causes of the harm The organisation suffers from a high cost of exit from the harm or commits to the sunk cost fallacy. The organisation lacks the capacity to fix the harm, and attempts to do so are met with sludge, which exhausts the reformers.	The introduction of sludge into 'organisational immune system' processes (e.g. decision-making committees, requiring multiple sign-offs or escalation). Creates organisational lock-in, e.g. to proprietary vendors, complex dependencies, or rigid/slow decision-making or action structures. Encourages misaligned incentives, talent exodus etc.
Damages the organisation's ability to recover from the harm Even if the harm is identified, the organisation lacks what it needs (e.g. energy, budget, or institutional memory) required to implement a recovery plan.	The implementation of the designed element (e.g. a new system or process) has used up substantial reserves of resource (money, goodwill, reputation, relationships, time etc.) not easily regained.

Summary

This paper has proposed that adopting a more adversarial approach to considering organisational design can offer valuable insight into potential organisational dysfunction, through stimulating different ways of thinking about design choices. ADT highlights how organisational design choices can lead to harm, even when intentions are good. Design is never neutral, and there is no such thing as 'no design'; poor design choices can lead to a range of subtle and unwelcome consequences. The ADT perspective makes the costs and unintended consequences of organisational design choices more explicit, leading to better quality designs.

The next step for this research is to further refine the GHOST and Harm frameworks through case study analysis, and to build a library of Dark Organisational Design Patterns.

References

Abdel-Kader, M. and Lin, E. (2009) 'The Organisational Architecture', *Performance Measure of New Product Development Teams*, pp.42-72. Palgrave Macmillan, London.

Applebaum, A., Miller, D., Strom, B., Korban, C. and Wolf, R. (2016) 'Intelligent, Automated Red Team Emulation', *2016 Annual Computer Security Applications Conference*, pp.363-373.

Faraj, S., Renno, W. and Bhardwaj, A. (2021) 'Unto the Breach: What the COVID-19 Pandemic Exposes about Digitisation', *Information and Organisation*, 31, pp.1-7.

Farry, R. (2020) 'Predicting How People Will Respond to a Disruptive Event: The Human Factors Response Framework', in Charles, R. and Golightly, D. (eds.) *Contemporary Ergonomics and Human Factors 2020*.

Fessenden, T. (2024) '3 Design Processes for High Usability: Iterative Design, Parallel Design, and Competitive Testing', *NN/Group website*, available at: https://www.nngroup.com/articles/parallel-and-iterative-design/ accessed 02/02/2026.

Gray, C., Kou, Y., Battles, B., Hoggat, J., and Toombs, A. (2018) 'The Dark (Patterns) Side of UX Design, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Article 534, pp.1-14.

Idemudia, I. (2023) 'Red Teaming: A Framework for Developing Ethical AI Systems', *American Journal of Engineering Research*, Vol.12, No.10, pp.7-14.

Kerr, S. (1975) 'On the Folly of Rewarding A, While Hoping for B', *Academy of Management Journal*, Vol.19, No.4, pp.769-783.

Kohnfelder, L. (1999) 'The Threats to our Products', *Microsoft Interface*.

Lidwell, W., Holden, K. and Butler, J. (2023) *Universal Principles of Design*, Third Edition. Rockport.

Li Sun, S., Zhang, Y. and Zhu (2021) 'Turning Disruption into Growth Opportunity: The Red Team Strategy, *Journal of Business Strategy*, Vol.43, Issue.6, pp.365-372.

Mathur, A., Kshursagar, M., and Mayer, J. (2021) 'What Makes a Dark Pattern…Dark?: Design Attributes, Normative Considerations, and Measurement Methods', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 360, pp.1-18.

O'Boyle, E. and Aguinis, H. (2017) 'The Best and the Rest: Revisiting the Norm of Normality of Individual Performance', *Academy of Management Proceedings*, Vo.2011, No.1.

OSS. (1944) *Simple Sabotage Field Manual*.

Staw, B. (1981) 'The Escalation of Commitment To a Course of Action', *Academy of Management Review*, Vol. 6, No. 4, pp.577-587.

Sunstein, C. (2019) 'Sludge and Ordeals', *Duke Law Journal*, Vol.68, No.8, pp.1843-1883.

University of Foreign Military and Cultural Studies (UFMCS) (2018) *The Red Team Handbook*, v9.0, US Army Training Doctrine and Command: https://home.army.mil/wood/6115/8222/0759/RedTeamHB.pdf

Valacich, J. and Schwenk, C. (1995) 'Devil's Advocacy and Dialectical Inquiry Effects on Face-to-Face and Computer-Mediated Group Decision Making', *Organisational Behaviour and Human Decision Processes,* Vol.63, No.2, pp.158-173.

Vaughan, D. (1996) *The Challenger Launch Decision*. University of Chicago Press.

Yulianto, S., Soewito, B., Goal, F. and Kurniawan, A. (2023) 'Metrics and Red Teaming in Cyber Resilience and Effectiveness: A Systematic Literature Review', *29th International Conference on Telecommunications (ICT)*, pp.1-7.