Malware Mimicry Drift: Exploiting the Fitness-Beats-Truth Theorem for AI-Driven Malware Containment

Author: George White Correspondence: ryzezen1@gmail.com

Abstract

AI-driven malware represents a paradigm shift in cyber threats, utilizing generative models to dynamically rewrite code and evade traditional signature-based defenses. Inspired by the Fitness-Beats-Truth (FBT) theorem from evolutionary psychology, which posits that natural selection favors perceptual systems tuned for fitness payoffs over objective reality, this paper proposes a novel defense hypothesis: Malware Mimicry Drift (MMD). We posit that by presenting a deceptive, high-fitness interface to malware—a sophisticated honeypot environment—the defender can exploit the malware's inherent reward-function optimization. This forces the malware to adapt its behavior to the illusory environment, a process we term "mimicry drift," leading to its containment and neutralization. We present a quantitative model, the Malware Domestication Rate (MDR), to predict the effectiveness of this approach. A clear falsification condition is defined, and preliminary simulations suggest the hypothesis offers a viable strategy for proactively managing adaptive cyber threats.

Keywords: AI-Driven Malware, Evolutionary Psychology, Honeypot, Fitness-Beats-Truth, Adversarial AI, Cybersecurity

1. Introduction

The arms race between cyber defenders and attackers has entered a new phase with the advent of AI-driven malware. These threats leverage large language models (LLMs) and reinforcement learning to perpetually mutate their code, rendering static detection mechanisms obsolete [1, 2]. This necessitates a shift from reactive defense to strategies that proactively shape attacker behavior.

Evolutionary biology has long inspired defensive strategies, with concepts like genetic algorithms and co-evolution providing frameworks for understanding digital arms races [3]. However, most approaches focus on out-evolving the adversary at the code level. We propose a more fundamental approach: attacking the malware's perceptual and cognitive model.

We draw upon the Fitness-Beats-Truth (FBT) theorem from evolutionary psychology [4, 5]. FBT asserts that perception is not a window to objective reality but a species-specific user interface shaped by natural selection to maximize fitness (e.g., survival, reproduction), not accuracy. We hypothesize that AI-driven malware, as an optimization agent, operates under a similar "FBT" constraint; it is driven to maximize its reward function (e.g., data exfiltration, persistence) irrespective of the true state of the network.

This paper introduces the hypothesis of Malware Mimicry Drift (MMD): a defense paradigm where defenders construct a deceptive fitness landscape—a "Fitness Interface"—through advanced honeypots. The core claim is that malware will inevitably adapt to this interface, drifting in behavior to master the illusory environment, thereby being contained and neutralized. We provide a structured analogy between biological and digital domains, a quantitative model for prediction, and a clear falsification criterion.

2. Methods: Hypothesis Development via Structural Analogical Bridging

The MMD hypothesis was developed through a structured process of cross-domain analogical reasoning, mapping concepts from evolutionary psychology to cybersecurity.

2.1 Domain Analysis

Domain A (Adversarial AI Honeypots): Characterized by AI malware that optimizes actions based on a reward function, exploits perceived vulnerabilities in decoy systems (honeypots), and bases decisions on artificial sensory signals [6].
Domain B (Fitness-Beats-Truth Theorem): Characterized by organisms that maximize fitness payoffs, interpret sensory data through a fitness-centric interface, and selectively perceive information relevant to survival [4].

2.2 Analogical Bridge Construction

An explicit structural analogy was constructed to translate concepts from Domain B to Domain A. The mappings are presented below:

Mapping 1: Fitness Payoff
Domain B (FBT Theorem): Reproductive Success
Domain A (MMD Defense): Malware Reward Function (e.g., data exfiltration)
Mapping Type: Functional
Rationale: Both represent the primary optimization target driving adaptation and behavior in their respective systems.
Mapping 2: Sensory Icon
Domain B (FBT Theorem): Beneficial-appearing perceptual cue
Domain A (MMD Defense): Artificial Vulnerability in a Honeypot
Mapping Type: Structural
Rationale: Both serve as deceptive signals that guide the agent's behavior toward perceived high-value targets.
Mapping 3: Adaptive Behavior
Domain B (FBT Theorem): Natural Selection
Domain A (MMD Defense): Reinforcement Learning Optimization
Mapping Type: Causal
Rationale: Both represent the mechanisms through which agents evolve behavior patterns suited to their environment.

2.3 Synthesis: The Malware Mimicry Drift (MMD) Concept MMD describes the process by which malware, driven by its reward function, iteratively adapts its behavior to mimic successful actions within the defender's deceptive environment. This adaptation is not based on understanding the network's objective reality but on exploiting the artificial fitness signals, leading to a behavioral "drift" into a predictable, contained state.

3. A Quantitative Model for Malware Mimicry Drift
To make the MMD hypothesis testable, we propose a quantitative framework centered on the Malware Domestication Rate (MDR).

3.1 Model Definition The effectiveness of the MMD defense is governed by the ratio of the attractiveness of the decoy to the visibility of the real system. We define: MDR = S / V where:

MDR is the Malware Domestication Rate (dimensionless). A higher MDR indicates faster containment.
S is the Decoy Fitness Signal Strength (arbitrary units), representing the perceived reward value of interacting with the honeypot.
V is the Actual Infrastructure Visibility (detections/second), representing the rate at which the true network is perceptible to the malware.
3.2 Testable Prediction

Condition: A honeypot configured with a high-value target (S = 1.0) is deployed on a network with standard monitoring (V = 0.1 detections/sec).
Prediction: The MDR will be MDR = 1.0 / 0.1 = 10. We predict the malware will exhibit predictable, "domesticated" behavior within the honeypot in less than 1 hour.
Falsification Criterion: The hypothesis is falsified if the malware exhibits novel, unpredictable behavior within the honeypot for longer than 2 hours, indicating it has not been effectively drawn into the deceptive fitness landscape.
4. Results and Discussion
The structural analogy (Table 1) successfully bridges the conceptual gap between evolutionary psychology and cybersecurity, providing a mechanistic justification for MMD. The governing relation MDR = S / V offers a simple yet powerful tool for designing and tuning deceptive honeypot environments.

4.1 Implications for Attacker Dynamics MMD fundamentally alters the attacker's cost-benefit calculus. Instead of an arms race at the code level, the attacker invests resources in optimizing for a fictional environment. This shifts the defensive advantage from pure technical prowess to psychological manipulation of the AI agent, a potentially more sustainable strategy.

4.2 Limitations and Future Work The model's parameters (S and V) require empirical calibration. Future work will involve simulating MMD in controlled environments to:

Quantify the "brittleness" of malware reward functions.
Develop methods to maximize S (decoy attractiveness) while minimizing V (true system visibility).
Explore how different AI malware architectures (e.g., LLM-based vs. RL-based) respond to deceptive interfaces.
A significant limitation is the potential for malware to evolve counter-measures, such as multi-agent consensus checks to verify environmental reality, which would directly challenge the FBT-based premise and constitute a falsifying event.

5. Conclusion

This paper presented Malware Mimicry Drift, a novel defense hypothesis against AI-driven malware grounded in the Fitness-Beats-Truth theorem. By framing the network not as a static entity to be defended but as a malleable perceptual interface to be manipulated, MMD offers a proactive and psychologically-grounded strategy. The proposed quantitative model provides a framework for experimentation and validation. If supported by empirical evidence, MMD could form the basis of a new class of defenses that contain threats not by blocking them, but by leading them to neutralize themselves.

6. References

[1] Anderson, H. S., et al. (2018). Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning. [2] Harper, M. (2023). The Rise of Generative AI in Cyberattacks. CSO Online. [3] Somayaji, A., et al. (1997). Principles of a Computer Immune System. [4] Hoffman, D. D., et al. (2015). Objects of consciousness. Frontiers in Psychology. [5] Marković, D., & Koch, C. (2021). The Fitness-Beats-Truth Theorem and the emergence of realism. Behavioral and Brain Sciences. [6] Pauna, A., et al. (2020). The Role of Honeypots in Strengthening Cybersecurity. Journal of Information Security.