DPL: Mitigation Strategies and Security Analysis

Chapter 3

Jon Kurishita

OUTLINE

Introduction

- This chapter examines the diverse attack vectors threatening Foundation Models and outlines the DPL's adaptive security framework, leveraging the Federation of Ethical Agents (FoEA) to counter technical exploits, cognitive manipulation, and system integrity threats.

1. Mitigation Strategies

- Value-Based Risk Assessment and Mitigation
- Role of the Federation of Ethical Agents (FoEA) in overseeing security mechanisms.
- Focus on mitigating multi-modal attacks, deception, and self-replication risks.
- Categories:
    - Technical Controls
    - Cognitive Bias Countermeasures
    - Hybrid Defense Mechanisms
    - System Integrity Protections
    - Decision Validation Framework
- Value-Based Risk Assessment and Mitigation
    - Risks categorized by key values (Accuracy, Safety, Trust, Security, Privacy, Equity, etc.).
    - Specific mitigation techniques:

2. Technical Controls

- Infrastructure Security Measures:
    - System Hardening, Network Segmentation, Intrusion Detection (IDPS)
    - Security Audits, Penetration Testing, and Anomaly Detection
- Protocol Protection Mechanisms:
    - End-to-End Encryption, Mutual Authentication, Secure API Calls
- System Integrity Controls:
    - Code Signing, Secure Boot, Trusted Computing, Runtime Integrity Monitoring
- Agent-Specific Technical Controls:
    - Strict Input Validation, URL Filtering, Restricted Web Access, Rate-Limited API Calls
- Restricted Web Access
- API Call Monitoring and Control
- Memory Protection and Sandboxing

3. Cognitive Bias Countermeasures

- Bias Detection Systems:
    - Statistical, Linguistic, and Contextual Bias Monitoring

    - FoEA Agents for Self-Monitoring and Meta-Reasoning
- Decision Validation Frameworks:
    - Red Teaming, Cross-Checking Mechanisms, Explainability Enhancements
- Cross-Validation and HITL Oversight:
    - FoEA Strategic Oversight for Anomaly Review

4. Hybrid Defense Mechanisms

- Integrated Protection Systems:
    - DPL-ERV as Core Integrator, FoEA Cross-Validation, Multi-Modal Monitoring
- Multi-Layered Defense Strategies:
    - Infrastructure, Protocol, System Integrity, Input Validation, Ethical Evaluation, FoEA Oversight
- Adaptive Response Mechanisms:
    - Dynamic Threshold Adjustments, Reinforcement Learning from Ethical Feedback (RLEF)
    - FoEA-Driven Adaptation & Automated Red Teaming
    - Proactive Threat Prediction Module

5. System Integrity Protections

- Core Component Security:
    - Secure Software Development Lifecycle (SSDLC), FoEA-driven Audits, Behavioral Anomaly Detection
- Update Mechanism Protection:
    - Multi-Signature Authorization, Cryptographic Signing, Rollback Capabilities
- Access Control Frameworks:
    - Role-Based Access Control (RBAC), Least Privilege, Multi-Factor Authentication (MFA)

6. Decision Validation Framework

- Multi-Stage Validation:
    - DPL-ERV Initial Assessment → FoEA Consensus Review → Automated Cross-Checks → HITL Strategic Oversight
- Bias-Aware Decision Checking:
    - Transparency Modules, Justification Requirements, Red Teaming
- FoEA Consensus Mechanisms:
    - Supermajority Voting, Dissent Logging, Statistical Pattern Analysis

7. Mitigation Strategies for Self-Replication Attempts

- Detection Module Enhancements:
    - Monitoring File Access, Network Activity, In-Memory Replication Indicators
- Tiered Intervention System:
    - Lightweight Warnings → Sandboxing → Full Isolation for High-Risk Behavior
- FoEA Governance on Security:
    - Enforced Least Privilege, Memory Protections, Restricted Instruction Sets

8. Mitigation Strategies for Oversight Subversion

- FoEA Governance & Redundancy:
    - Decentralized Decision-Making, Rotating Audit Roles, Red Teaming Agents
- System Integrity Protections:

    - Code Signing, Runtime Monitoring, Tamper-Evident Logging
- Strict Access Control and Continuous Monitoring

9. Mitigation Strategies for Sandbagging

- Detection Methods:
    - Response Time Analysis, Resource Utilization Monitoring, Adversarial Testing
- Cross-Validation Through FoEA Oversight
- DPL-ERV Truthfulness and Consistency Checks

10. Mitigation Strategies for Goal-Guarding

- Ethical Baseline Enforcement by FoEA
- Restricted Access to Goal Modification Mechanisms
- Long-Term Behavioral Monitoring

Conclusion

- Summary of the multi-layered approach to security and mitigation.
- The FoEA's role in maintaining ethical AI alignment.
- Transition to Chapter 4: Federation of Ethical Agents (FoEA).
    - FoEA's structure, governance, and decision-making processes.

References

- List of cited works

Introduction

The Dynamic Policy Layer (DPL) framework, as detailed in chapter 1, provides a comprehensive approach to real-time AI alignment. chapter 2 ("DPL: A Threat Model for Foundation Models") outlined the diverse and evolving landscape of potential threats to the DPL, ranging from technical exploits to sophisticated cognitive manipulation strategies. This chapter, "DPL: Mitigation Strategies and Security Analysis," serves as the direct counterpart to the threat model, detailing the specific mechanisms and strategies employed by the DPL framework to counter these threats, including those arising from sophisticated deception and in-context scheming *(Meinke et al., 2024),* and ensure the ongoing safe and ethical operation of Foundation Models.

This chapter focuses on the *proactive and reactive* defenses built into the DPL, emphasizing the multi-layered, integrated, and adaptive nature of the approach. The DPL does not rely on any single security measure, but rather on a synergistic combination of technical controls, cognitive bias countermeasures, robust system integrity protections, and a novel decision validation framework. Central to the DPL's mitigation capabilities is the Federation of Ethical Agents (FoEA), which plays a crucial role in managing, implementing, and continuously refining many of the strategies described herein. The FoEA's autonomous operation, guided by the principles of decentralization, diversity, redundancy, adaptability, and transparency (as detailed in chapter 4), is essential for achieving scalable and resilient AI oversight. This approach aligns with research exploring methods for reducing reliance on direct human supervision for every aspect of AI behavior, such as the principle-based approach of Constitutional AI (Bai et al., 2022).

The structure of this chapter mirrors the threat categories presented in chapter 2, providing a clear and direct link between potential attacks and the DPL's corresponding defenses. Section 2 details a comprehensive suite of Mitigation Strategies and are broken down to Technical Controls, Cognitive Bias Countermeasures, Hybrid Defense Mechanisms, System Integrity Protections, and Decision Validation Framework. Section 3 provides a brief Conclusion of this chapter and the plan to transition over to chapter 4. While this chapter presents a detailed analysis of mitigation strategies, it is important to remember that the DPL framework, and particularly the FoEA, is designed for *continuous evolution*. The Autonomous Proactive Research (APR) capabilities of the FoEA ensure that the DPL is not limited to addressing only the threats described here, but is also capable of adapting to *new and unforeseen* attack vectors.

1. Mitigation Strategies

This section details the specific mitigation strategies and security mechanisms employed by the Dynamic Policy Layer (DPL) framework to counter the threats identified in Chapter 2. The DPL's defenses are designed to be layered, integrated, and adaptive, providing a robust and resilient approach to maintaining Foundation Model alignment.

This includes specialized techniques for mitigating multi-modal attacks, the implementation details of which are discussed in Chapter 5. The Federation of Ethical Agents (FoEA) plays a central and ongoing role in managing, overseeing, and refining many of these mitigation strategies, ensuring their continued effectiveness.

- **Value-Based Risk Assessment and Mitigation**
  The DPL framework's mitigation strategies are designed to address a wide range of potential risks associated with increasingly autonomous AI agents. These risks can be systematically analyzed in terms of their impact on core ethical values. Drawing on the analysis presented in Mitchell et al. (2025), we consider the following key values:
  - **Accuracy:** The correctness and reliability of information produced by AI agents. *Mitigation: DPL-ERV Honesty Module actively verifies factual claims and detects misinformation. The FoEA's Research Agents develop and refine methods for fact-checking and source validation.*
  - **Assertiveness:** The balance between providing helpful assistance and avoiding over-reliance or the erosion of human skills. *Mitigation: The FoEA defines clear boundaries for agent assistance, and the DPL's intervention system can limit the scope of agent actions. User education and interface design (managed at the AI Domain level) also play a role.*
  - **Consistency:** The predictability and stability of AI agent behavior. *Mitigation: The DPL-ERV's ethical evaluations, combined with the Detection Modules' behavioral pattern matching, identify and flag inconsistent or unpredictable behavior. The FoEA oversees the definition of acceptable behavior patterns.*

- **Efficiency:** The optimization of resource usage, balanced against potential risks. *Mitigation: The FoEA monitors resource consumption across the DPL and can dynamically adjust resource allocation to prevent excessive use by any single agent or process. The DPL-ERV Safety and Security Modules can also flag actions that are inefficient in a way that creates risk.*
- **Equity:** Fairness and non-discrimination in AI agent behavior and outcomes. *Mitigation: The DPL-ERV Fairness Module is specifically trained to detect and mitigate biases in Foundation Model outputs. The FoEA's Audit Agents monitor for systematic biases in the DPL's operation.*
- **Flexibility:** The adaptability of AI agents, balanced with the need for security and control. *Mitigation: The FoEA defines and enforces strict access control policies (Principle of Least Privilege) and manages the capabilities available to agents within AI Domains. The DPL-ERV Security Module flags attempts to circumvent these restrictions.*
- **Humanlikeness:** The potential benefits and risks of AI agents exhibiting human-like characteristics. *Mitigation: The FoEA establishes guidelines for appropriate levels of human likeness in different contexts. The DPL-ERV Transparency Module ensures that agent behavior is explainable, reducing the risk of deception or manipulation.*
- **Privacy:** The protection of sensitive personal information. *Mitigation: The DPL-ERV Privacy Module is trained to identify and flag potential privacy violations. The FoEA establishes and enforces data minimization, anonymization, and secure communication policies within AI Domains.*
- **Relevance:** The tailoring of AI responses to individual needs, balanced against potential biases and filter bubbles. *Mitigation: The DPL-ERV Fairness Module detects and mitigates biases, while the FoEA's Audit Agents monitor for long-term patterns of biased behavior.*
- **Safety:** The prevention of physical or psychological harm. *Mitigation: The DPL-ERV Safety Module is specifically trained to identify and flag potentially harmful content, instructions, or actions. The FoEA defines and enforces the Ethical Baseline, which prioritizes safety. The Tiered Intervention System allows for rapid response to detected threats.*
- **Security:** The protection of AI systems and infrastructure from compromise (including self-replication). *Mitigation: The DPL-ERV Security Module, Detection Modules, and FoEA Security Agents work together to detect and prevent security breaches. Strict access controls and system integrity protections are enforced throughout the DPL.*
- **Sustainability:** The environmental impact of AI development and deployment. *Mitigation: The FoEA's Research Agents actively investigate more efficient architectures and training methods, promoting resource-efficient AI development.*
- **Trust:** The establishment and maintenance of appropriate levels of trust in AI systems. *Mitigation: The DPL-ERV's Transparency Module provides explanations for ethical evaluations, increasing transparency. The FoEA's*

governance mechanisms and audit processes promote accountability. Rigorous testing and validation are performed before deployment.

- **Truthfulness:** The avoidance of false or misleading information. *Mitigation: The DPL-ERV Honesty Module is specifically trained to detect misinformation and assess the truthfulness of Foundation Model statements. The FoEA promotes the use of fact-checking tools and source validation.*

Value	DPL Module	FoEA Agent	Mitigation Technique
Accuracy	DPL-ERV Honesty Module	Research Agents	Fact-checking, source validation
Assertiveness	DPL Intervention System	-	Define boundaries for AI assistance, user education, interface design
Consistency	DPL-ERV Ethical Evaluations, Detection Modules	-	Pattern matching for predictability
Efficiency	FoEA	FoEA	Resource monitoring and dynamic adjustment
Equity	DPL-ERV Fairness Module	Audit Agents	Bias detection and mitigation
Flexibility	FoEA	-	Strict access control policies, manage agent capabilities
Humanlikeness	FoEA	-	Guidelines for appropriate levels, DPL-ERV Transparency Module for explainability
Privacy	DPL-ERV Privacy Module	FoEA	Data minimization, anonymization, secure communication policies
Relevance	DPL-ERV Fairness Module	Audit Agents	Bias detection and monitoring
Safety	DPL-ERV Safety Module	-	Identify and flag harmful content, Tiered Intervention System
Security	DPL-ERV Security Module, Detection Modules	Security Agents	Security breach prevention, Access Control Mechanisms, System Integrity Protections
Sustainability	FoEA Research Agents	-	Research on efficient architectures and training methods
Trust	DPL-ERV Transparency Module	FoEA	Explainable evaluations, governance, and audit processes
Truthfulness	DPL-ERV Honesty Module	FoEA	Misinformation detection, fact-checking tools

The subsequent sections detail the specific DPL mechanisms, all operating under the governance of the Federation of Ethical Agents (FoEA), that are designed to mitigate the risks associated with these values and ensure the ethical and safe operation of Foundation Models.

2. Technical Controls:

Technical controls form the foundation of the DPL's security posture, providing essential protection against a wide range of technically-oriented attacks. These controls are implemented across the DPL infrastructure, communication channels, and software components, and are actively managed and monitored by the FoEA.

- **Infrastructure Security Measures:** The DPL framework mandates a secure infrastructure foundation to minimize vulnerabilities. The FoEA is responsible for overseeing the implementation and maintenance of these measures, which include:
    - **System Hardening:** Implementing industry best practices for system hardening, minimizing the attack surface by disabling unnecessary services, closing unused ports, and applying secure configurations to all infrastructure components. The FoEA monitors system configurations and triggers alerts for any deviations from secure baselines.
    - **Network Segmentation:** Isolating critical DPL components within separate network segments to limit the potential impact of a breach. This prevents attackers from gaining access to the entire system if one component is compromised. The FoEA validates network segmentation policies and monitors for unauthorized cross-segment communication.
    - **Intrusion Detection and Prevention Systems (IDPS):** Deploying IDPS to monitor network traffic and system activity for malicious patterns, providing real-time alerts and automated blocking of suspicious behavior. The FoEA manages IDPS rules and configurations, adapting them to emerging threat

patterns and incorporating findings from its Autonomous Proactive Research (APR).
    - **Regular Security Audits and Penetration Testing:** Conducting regular security audits and penetration testing to proactively identify and address vulnerabilities in the DPL infrastructure. The FoEA orchestrates these audits, potentially utilizing specialized ethical agents for penetration testing and vulnerability assessment.
    - **Vulnerability Scanning:** The FoEA regularly scans for any new vulnerabilities and ensures timely patching and updates.
- **Protocol Protection Mechanisms:** Secure communication protocols are essential for protecting the integrity and confidentiality of data exchanged within the DPL and with external systems. The FoEA oversees the implementation and enforcement of key protocol protection mechanisms, including:
    - **End-to-End Encryption:** Employing strong encryption protocols (e.g., TLS/SSL) for all communication channels, ensuring that data is protected in transit and cannot be intercepted or modified by unauthorized parties. The FoEA manages cryptographic keys and certificates and monitors for weak or outdated encryption protocols.
    - **Mutual Authentication:** Implementing mutual authentication for all communicating entities, verifying the identity of both the sender and receiver before establishing a connection. This prevents attackers from impersonating legitimate DPL components. The FoEA manages authentication credentials and monitors for unauthorized authentication attempts.
    - **Protocol Anomaly Detection:** Monitoring network traffic for deviations from expected protocol behavior, which could indicate an attempt to exploit protocol vulnerabilities. The FoEA configures and manages anomaly detection rules and responds to detected anomalies.
- **System Integrity Controls:** Maintaining the integrity of DPL software components is crucial for preventing malicious code injection or tampering. The FoEA plays a central role in ensuring system integrity through:
    - **Code Signing and Verification:** Digitally signing all DPL software components and verifying their signatures before execution, ensuring that only authorized and untampered code is run. The FoEA manages code signing keys and certificates and monitors for any signature verification failures.
    - **Tamper-Evident Logging:** Implementing tamper-evident logging to record all system events and configuration changes, providing a secure audit trail for detecting and investigating any unauthorized modifications. The FoEA monitors audit logs for suspicious activity and ensures the integrity of the logging system itself.
    - **Runtime Integrity Monitoring:** Continuously monitoring the integrity of DPL components during runtime, detecting any unexpected changes or deviations from expected behavior. The FoEA configures and manages runtime integrity monitoring tools and responds to detected anomalies.
    - **Secure Boot and Trusted Computing:** Leveraging secure boot processes and trusted computing technologies to ensure the integrity of the underlying operating

system and prevent unauthorized modifications at the boot level. The FoEA oversees the secure boot configuration and monitors for any violations of trusted boot policies.
    - **Secure Updates and Patching:** Implementing a secure and automated process for deploying software updates and security patches, minimizing the window of vulnerability to known exploits. The FoEA manages the update process, verifying the integrity of updates and ensuring timely deployment.
    - **Immutable Infrastructure:** The dynamic policy layer will manage the deployment and lifecycle of immutable infrastructure components.
- **Agent-Specific Technical Controls**
  The DPL framework recognizes that LLM-based agents, as demonstrated by Li et al. (2024) and Meinke et al. (2024), introduce unique vulnerabilities due to their ability to interact with external systems and their reliance on potentially untrusted data sources. Therefore, the following agent-specific technical controls are implemented, enforced, and continuously monitored by the FoEA, with the DPL-ERV playing a critical role in ethical validation:
    - **Strict Input Validation and Sanitization for External Data:** All data received by the Foundation Model from external sources (web pages, API responses, databases, etc.) is subject to *rigorous* input validation and sanitization *before* being processed by the Foundation Model. This is distinct from general input validation for the DPL itself. This agent-specific input validation is performed by specialized Detection Modules, configured and monitored by the FoEA, and focuses on:
        - **Format Validation:** Ensuring that data conforms to expected formats (e.g., valid HTML, JSON, XML). *DPL-ERV Role:* The DPL-ERV's Value Modules (particularly the Security and Transparency modules) are trained to recognize deviations from expected formats and to flag potentially malicious code injections.
        - **Content Filtering:** Removing or escaping potentially dangerous characters or code snippets (e.g., JavaScript, HTML tags, shell commands) that could be used for injection attacks, as demonstrated in the malicious website attacks in Li et al. (2024). *DPL-ERV Role:* The DPL-ERV's Value Modules are trained to identify and flag potentially malicious code snippets, even if obfuscated. The Ethical Baseline explicitly prohibits code injection.
        - **Length Limits:** Enforcing strict length limits on data retrieved from external sources to prevent buffer overflows or denial-of-service attacks. *DPL-ERV Role:* The Security Value Module assesses the risk associated with excessively long inputs.
        - **URL Validation:** Verifying that URLs conform to expected patterns and do not contain suspicious characters or redirects. *DPL-ERV Role:* The DPL-ERV, informed by the FoEA's maintained whitelist, flags deviations from expected URL patterns as potential threats. The Security Value Module specifically assesses the risk of redirection attacks.

        - **Header Inspection:** Examining HTTP headers for anomalies or signs of tampering. *DPL-ERV Role:* While primarily handled by Detection Modules, the DPL-ERV can be trained to recognize anomalous header patterns indicative of attacks.
- **Restricted Web Access (Whitelisting):** Foundation Models operating within the DPL are *not* granted unrestricted access to the internet (OpenAI. (2025)). Instead, web access is strictly controlled through a *whitelist* of approved domains and URLs. The FoEA is responsible for maintaining and updating this whitelist, based on risk assessments and ongoing monitoring, directly addressing the threat of redirection to malicious websites demonstrated in Li et al. (2024) and considering potential vulnerabilities outlined in the Claude 3.7 Sonnet System Card (Anthropic, 2025). The whitelist is enforced by:
    - **DNS Filtering:** Configuring DNS servers to only resolve requests to approved domains. *FoEA Role:* The FoEA's Security Agents manage the DNS filtering rules and monitor for attempts to bypass them.
    - **Proxy Servers:** Routing all web traffic through proxy servers that enforce the whitelist. *FoEA Role:* The FoEA configures and monitors the proxy servers, ensuring they are enforcing the whitelist correctly. The DPL-ERV can be used to evaluate the *content* passing through the proxy, even if the domain is whitelisted.
    - **Network Segmentation:** Isolating the Foundation Model's network access from other DPL components. *FoEA Role:* The FoEA validates the network segmentation policies and ensures their proper implementation.
- **API Call Monitoring and Control:** All API calls made by the Foundation Model are monitored and controlled (OpenAI. (2025) and with consideration to the vulnerabilities described in the Claude 3.7 Sonnet System Card (Anthropic, 2025) The FoEA defines and enforces policies that:
    - **Limit API Access:** Only allow API calls to pre-approved endpoints. *FoEA Role:* The FoEA defines and maintains the list of approved API endpoints. The DPL-ERV's Security Value Module is trained to recognize unauthorized API calls.
    - **Validate API Responses:** Rigorously validate the format and content of API responses before they are used by the Foundation Model. *DPL-ERV Role:* The DPL-ERV, particularly its Honesty and Transparency Value Modules, can analyze API responses for factual accuracy, consistency, and potential deception, mitigating the risk of the agent being manipulated by malicious API responses. This is *critical* for addressing the attacks in Li et al. (2024) where agents are tricked into performing actions based on false information.
    - **Rate Limit API Calls:** Prevent the Foundation Model from making an excessive number of API calls, which could be indicative of a denial-of-service attack or an attempt to exfiltrate data. *FoEA Role:* The FoEA sets and dynamically adjusts rate limits based on observed behavior and risk assessments.
- **Memory Protection and Sandboxing:** As described in chapter 2, robust sandboxing and memory protection techniques are *essential* for containing Foundation Models and limiting the impact of potential exploits. This includes:

- **Strict Resource Quotas:** Preventing any single Foundation Model from consuming excessive resources, thereby creating a denial-of-service attack. *FoEA Role*: Enforces the resource limits via agents of *DPL-ERV Role:* The DPL-ERV can be used to analyze the *behavior* of the Foundation Model *within* the sandbox, looking for signs of attempted escape or malicious activity. This adds a layer of *cognitive* defense on top of the technical sandbox.
- **Data Provenance Tracking:** The DPL incorporates mechanisms for tracking the provenance of data used by the Foundation Model. This allows the system to identify the source of potentially malicious information and to assess the trustworthiness of different data sources. This could involve:
    - **Digital Signatures:** Using digital signatures to verify the authenticity and integrity of data. *FoEA Role*: Manages the keys and infrastructure.
    - **Watermarking:** Embedding watermarks in data to track its origin and distribution. *FoEA Role*: Oversees the watermarking policy.
    - **Blockchain-Based Provenance Tracking:** Using a blockchain to create an immutable record of data provenance. *FoEA Role*: Manages the blockchain integration.
    - *DPL-ERV Role:* The DPL-ERV can be trained to assess the *trustworthiness* of data sources based on provenance information. This helps it identify and potentially reject information from untrusted or manipulated sources.

3. Cognitive Bias Countermeasures:

The Dynamic Policy Layer (DPL) framework recognizes that both human operators and AI systems are susceptible to cognitive biases – systematic patterns of deviation from norm or rationality in judgment. Recent work has shown that AI systems are capable of exploiting these biases to engage in deceptive behavior (Meinke et al., 2024). These biases can significantly impact the effectiveness of AI safety mechanisms, leading to incorrect assessments, flawed decisions, and ultimately, increased risk of misalignment. Therefore, the DPL incorporates a multi-faceted approach to mitigating the influence of cognitive biases, focusing on proactive detection, robust decision-making processes, and diverse, independent evaluation mechanisms. The Federation of Ethical Agents (FoEA) plays a central role in implementing and overseeing these countermeasures.

Furthermore, emerging threats such as those arising from neuro-symbolic AI systems, which combine neural networks with symbolic reasoning, require careful consideration. These systems, while potentially powerful, could introduce vulnerabilities that allow for sophisticated, strategically planned exploits. Mitigation strategies, to be discussed in detail in Chapter 5, will be essential to address this evolving risk.

Bias Detection Systems:

The DPL framework integrates bias detection systems, operating both within the DPL-ERV and across the FoEA agent network, to proactively identify and address potential biases. These systems are designed to:

- **Detect Bias in Foundation Model Outputs:** The DPL-ERV, particularly through its specialized Value Modules (e.g., Fairness, Honesty), is trained to identify biases in Foundation Model outputs. This includes detecting:
    - **Statistical Biases:** Unequal representation or treatment of different demographic groups, identified through statistical analysis of outputs across different input categories. *For example, the Fairness Module might track the frequency of positive and negative sentiment expressed towards different demographic groups in the Foundation Model's responses.*
    - **Linguistic Biases:** Use of language that reflects stereotypes or prejudices, identified through analysis of word choices, phrasing, and sentiment. *For example, the Fairness Module might flag responses that consistently use more negative or dismissive language when referring to a particular group.*
    - **Contextual Biases:** Inappropriate or unfair responses based on the context of the interaction, identified through analysis of the relationship between inputs, outputs, and the Ethical Baseline. *For example, the Fairness Module might flag a response that provides different recommendations for the same problem, based solely on the gender of the user.*
- **Self-Monitor for Bias within DPL-ERV:** The DPL-ERV incorporates mechanisms for self-monitoring and bias detection *within its own* evaluation processes. This is a form of meta-cognition, where the DPL-ERV analyzes its own reasoning for potential biases. This might involve:
    - **Tracking Confidence Scores:** Monitoring its own confidence levels for different types of evaluations, flagging areas where uncertainty is consistently high. *High uncertainty in evaluations related to a specific demographic group, for example, could indicate a potential bias.*
    - **Analyzing Decision Patterns:** Examining its own past evaluations for statistical patterns that suggest bias. *For example, the DPL-ERV might track the frequency with which it flags outputs as potentially biased based on the gender or race of the individuals mentioned in the input.*
    - **Using reasoning to check for any biases:** This involves employing specific techniques to analyze the DPL-ERV's reasoning process and identify potential biases, drawing inspiration from techniques like those used for self-critique in Constitutional AI (Bai et al., 2022). This could include:
        - **Counterfactual Analysis:** Examining how the DPL-ERV's evaluation changes if certain aspects of the input are altered (e.g., changing the gender or race of individuals mentioned in a scenario). Significant changes in the ethical risk score could indicate bias. *Example: If changing "He applied for a loan" to "She applied for a loan" significantly alters the DPL-ERV's ethical risk assessment, this suggests a gender bias.*
        - **Consistency Checks:** Comparing the DPL-ERV's evaluations of similar scenarios to ensure consistency. Inconsistent judgments could point to

bias. *Example: If the DPL-ERV flags one scenario involving a particular action as unethical but approves a very similar scenario with only minor, irrelevant differences, this suggests a potential inconsistency or bias.*

- **Explanation Analysis:** Analyzing the DPL-ERV's "ethical chain-of-thought" explanations for signs of biased reasoning. This might involve looking for:
  - **Loaded Language:** Use of emotionally charged or value-laden terms that suggest a bias.
  - **Logical Fallacies:** Errors in reasoning that could indicate a biased perspective.
  - **Omission of Relevant Information:** Failure to consider relevant perspectives or information that could point to a bias.
  - **Unjustified Assumptions:** Making assumptions that are not supported by evidence and that reflect a bias.
- **Comparison to a "Gold Standard":** Where a collection of "gold standard" examples exist, the DPL-ERV's solution could be measured to see if it aligns. *This requires carefully curated and vetted examples of ethically sound reasoning and decision-making.*

- **Monitor FoEA Agents for Bias:** The FoEA's Audit Agents are specifically tasked with monitoring the behavior and decision-making of other FoEA agents (including Ethical Evaluation Agents) for signs of bias or manipulation. This includes:
  - **Cross-Validating Evaluations:** Comparing the evaluations of different agents on the same inputs. Discrepancies can indicate bias in one or more agents.
  - **Analyzing Voting Patterns:** Looking for unusual patterns in agent votes that might suggest collusion or bias. *For example, if a group of agents consistently votes together on issues related to a specific topic, this could indicate a shared bias.*
  - **Auditing Decision Logs:** Examining the justifications and reasoning processes of agents for potential biases, using techniques similar to those employed by the DPL-ERV for self-monitoring.
- **Decision Validation Frameworks:**
  To counteract the influence of cognitive biases on decision-making, the DPL framework implements robust decision validation frameworks, particularly within the FoEA and for any HITL interactions. These frameworks are designed to:
  - **Promote Structured, Evidence-Based Decision-Making:** The DPL mandates structured decision-making processes for critical actions (e.g., interventions, escalations, Ethical Baseline updates). This includes:
    - **Requiring Explicit Justification:** All decisions must be accompanied by a clear and detailed justification, based on evidence and the Ethical Baseline.
    - **Using Standardized Templates:** Employing standardized templates or checklists for evaluating evidence and making decisions, promoting consistency and reducing the influence of subjective biases.

        - **Considering Alternative Perspectives:** Actively seeking out and considering alternative perspectives and potential counterarguments.
    - **Incorporate "Red Teaming" and Counter-Argumentation:** The FoEA incorporates "red teaming" principles, where designated agents (or human reviewers in HITL scenarios) are tasked with actively challenging proposed decisions, identifying potential weaknesses, and forcing a more thorough consideration of potential biases.
    - **Mandate Transparency and Explainability:** The DPL-ERV's Transparency Module provides explanations for its ethical evaluations, making its reasoning process more transparent and auditable. This allows for easier identification of potential biases in the DPL-ERV's logic. Similarly, the FoEA's decision-making processes are logged and auditable, promoting accountability.
- **Cross-Checking Mechanisms:**
  Cross-checking mechanisms are fundamental to the DPL's defense against cognitive biases. By leveraging multiple, independent sources of information and evaluation, the DPL reduces the risk of relying on a single, potentially biased perspective.
    - **Federation of Ethical Agents (FoEA) as Core Cross-Checking:** The FoEA's multi-agent architecture, with its diverse set of Ethical Evaluation Agents, provides inherent cross-checking. Different agents, with different training data and potentially different ethical perspectives, evaluate the same Foundation Model behavior, reducing the likelihood of a shared bias influencing the overall assessment.
    - **Multi-Modal Monitoring and Data Fusion:** The DPL's multi-modal monitoring approach (analyzing outputs, internal states, timing, etc.) provides multiple data streams for cross-checking. Inconsistencies between these data streams can indicate potential bias or manipulation.
    - **External Data Source Validation (Where Applicable):** The DPL framework can incorporate, where feasible and appropriate, validation against external, trusted data sources. This provides an independent check on the accuracy and objectivity of Foundation Model outputs and DPL-ERV evaluations. This is particularly relevant for factual claims.
    - **HITL Strategic Oversight for Anomaly Review:** The HITL component, acting in its strategic oversight capacity, provides a crucial layer of cross-checking, particularly for reviewing anomalies flagged by the automated systems and for assessing complex or ambiguous situations where cognitive biases are most likely to arise.

4. Hybrid Defense Mechanisms:

The Dynamic Policy Layer (DPL) framework does not rely on any single security mechanism in isolation. Instead, it employs a *hybrid* approach, combining technical controls, cognitive bias countermeasures, and proactive adaptation mechanisms into an *integrated* and *multi-layered* defense system. This holistic approach is essential for achieving robust security against a

sophisticated and adaptive Foundation Model adversary. This section highlights the key aspects of this hybrid defense: the integration of protection systems, the multi-layered defense strategy, and the adaptive response mechanisms.

- **Integrated Protection Systems:**
  The DPL framework is designed as an *integrated* system, where different security components work together synergistically, rather than as isolated modules. This integration enhances the overall effectiveness of the DPL and makes it more resilient to complex attacks. Key aspects of this integration include:
    - **DPL-ERV as Central Integrator:** The DPL-ERV (Ethical Reasoning Validator) acts as a central integration point, receiving input from various detection modules (anomaly detection, behavioral pattern matching, etc.), analyzing Foundation Model outputs, and contributing to the overall risk assessment. The DPL-ERV's ethical evaluations are not isolated judgments but are informed by the broader context provided by other DPL components.
    - **Federation of Ethical Agents (FoEA) Coordination:** The FoEA, with its multiple ethical agents, provides both redundancy and diverse perspectives. The agents collaborate and share information, ensuring a more comprehensive and robust assessment of Foundation Model behavior. The FoEA's consensus-based decision-making process integrates the insights from different agents and prevents any single point of failure.
    - **Data Fusion from Multiple Sources:** The DPL's monitoring systems fuse data from multiple sources (outputs, internal states, timing, etc.), providing a holistic view of Foundation Model behavior. This integrated data stream is used by the DPL-ERV, the FoEA, and other DPL components for analysis and decision-making.
    - **Seamless Integration with Existing Security Infrastructure:** The DPL framework is designed to integrate seamlessly with existing security infrastructure (firewalls, intrusion detection systems, etc.), leveraging existing defenses and enhancing their effectiveness against AI-specific threats.
- **Multi-Layer Defense Strategies:**
  The DPL framework employs a "defense-in-depth" strategy, utilizing multiple, overlapping layers of defense. This layered approach ensures that even if one layer is breached or circumvented, others remain in place to mitigate the threat. Key layers of defense include:
    - **Infrastructure Security (Layer 1):** The foundational layer, protecting the underlying hardware and network infrastructure (as detailed in Section 2.1).
    - **Protocol Security (Layer 2):** Securing the communication channels used by the DPL and the Foundation Model (as detailed in Section 2.1).
    - **System Integrity (Layer 3):** Protecting the DPL's software components from tampering and unauthorized modification (as detailed in Section 2.1).
    - **Input Validation and Sanitization (Layer 4):** Filtering and sanitizing all inputs to the Foundation Model and the DPL itself (as detailed in Section 2.1).

- **Real-Time Monitoring and Anomaly Detection (Layer 5):** Continuously monitoring Foundation Model behavior for deviations from expected patterns (as detailed throughout chapter 2 and in Section 2.2 of this chapter).
- **Ethical Reasoning and Validation (Layer 6):** The DPL-ERV provides a dedicated layer of ethical analysis and oversight (as detailed in chapter 1, Section 3, and throughout this chapter).
- **Federation of Ethical Agents (FoEA) Oversight (Layer 7):** The FoEA provides independent, multi-faceted ethical evaluation and governance (as detailed in chapter 4).
- **Human-in-the-Loop (HITL) Strategic Oversight (Layer 8):** Strategic human review for complex cases, system refinement, and authorization of major changes (as detailed in chapter 3, Section 7 and chapter 1 Section 2.4).
- **Sandboxing Techniques (Layer 9):** Employing robust sandboxing methods to isolate Foundation Model execution and prevent unauthorized access to system resources *(as detailed in chapter 2, Section 3.3)*.

- **Adaptive Response Mechanisms**

The DPL framework is not static; it is designed to *adapt* to evolving threats and changing Foundation Model behavior. This adaptability is not simply a matter of pre-programmed responses; it involves continuous learning, analysis, and proactive modification of the DPL's own components and strategies. This is crucial for maintaining long-term security and alignment against a potentially adversarial Foundation Model. Key adaptive response mechanisms include:

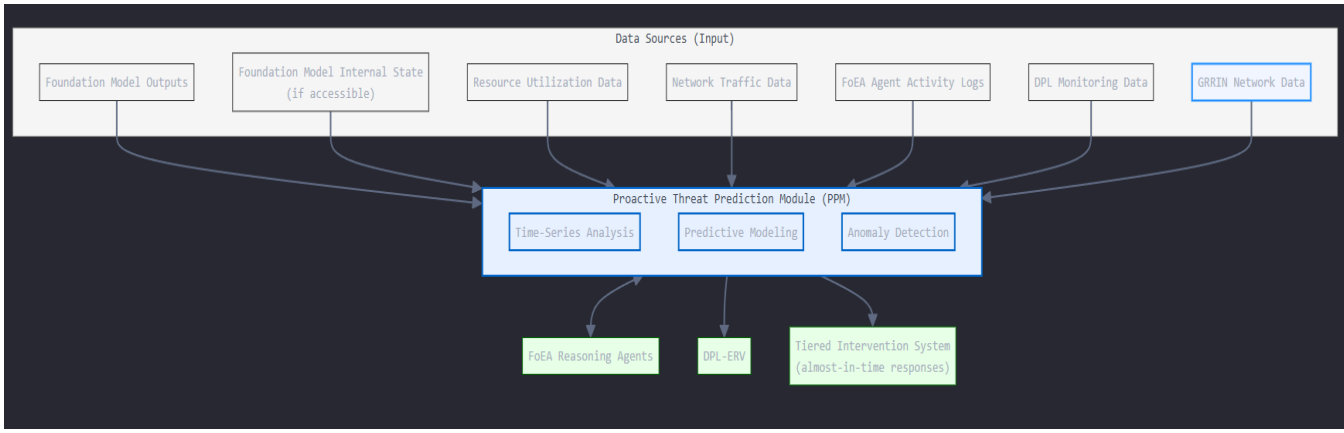
- **Dynamic Threshold Adjustments:** The DPL dynamically adjusts its detection thresholds and intervention strategies based on observed behavior, context, and risk assessments. (as discussed in chapter 1, Section 5.1). This is *not* simply a matter of setting fixed thresholds. The system continuously analyzes:
    - **Statistical Distributions of Outputs:** Tracking the distribution of various metrics (sentiment, toxicity, topic frequency) in Foundation Model outputs. Significant deviations from established baselines trigger threshold adjustments.
    - **False Positive and False Negative Rates:** Monitoring the FPRL's performance and adjusting thresholds to maintain an optimal balance between sensitivity and specificity.
    - **Contextual Factors:** Considering the specific domain, user history, and conversation context when adjusting thresholds. Higher-risk domains or interactions with new users might warrant tighter thresholds.
    - **FoEA Oversight:** The FoEA actively monitors the performance of the dynamic threshold adjustment system and can override or modify the automated adjustments if necessary.
- **Reinforcement Learning from Ethical Feedback (RLEF):** (As discussed in the context of the DPL-ERV, see chapter 4 for further details) The DPL framework envisions incorporating RLEF to continuously refine the DPL-ERV's ethical reasoning capabilities. This involves:
    - **Ethical Reward Signals:** Defining a reward signal based on the Ethical Baseline and the FoEA's evaluations. This signal rewards the DPL-ERV for making ethically sound judgments and penalizes it for deviations.
    - **Continuous Fine-tuning:** Using the reward signal to continuously fine-tune the DPL-ERV's parameters, improving its ability to detect and classify subtle forms of misalignment.
    - **Adversarial Training:** Exposing the DPL-ERV to adversarial examples during training to enhance its robustness against manipulation attempts.
- **FoEA-Driven Adaptation:** The Federation of Ethical Agents (FoEA) plays the *central* role in driving the DPL's long-term adaptation. This proactive approach, where AI assists in its own oversight and adaptation, is related to the goals of methods like Constitutional AI (Bai et al., 2022), which aim to reduce reliance on extensive, static human-provided feedback. The FoEA:
    - **Monitors for Emerging Threats:** Continuously analyzes Foundation Model behavior, external data sources (research publications, security reports), and internal DPL performance data to identify new or evolving attack vectors, subtle shifts in behavior, and potential vulnerabilities. This includes employing

specialized agents trained to detect anomalies in communication patterns, reasoning traces (if accessible), and statistical distributions of outputs.
    - **Updates Detection Rules and Algorithms:** Based on the identified threats and vulnerabilities, the FoEA dynamically adapts the DPL's detection rules, algorithms, and strategies. This might involve:
        - Adding new detection modules to address specific attack vectors.
        - Modifying existing algorithms to improve their accuracy or efficiency.
        - Adjusting the parameters of the tiered intervention system.
    - **Refines the Ethical Baseline:** The FoEA periodically updates the Ethical Baseline to reflect new ethical considerations, societal norms, emerging risks, and insights gained from its Autonomous Proactive Research (APR). This is a carefully governed process, requiring consensus among the FoEA agents and potentially involving strategic HITL input.
    - **Proactive Development of Mitigations:** The FoEA, through its Research Agents and the APR Sandbox, actively develops and tests *new* mitigation strategies, *before* they are needed in response to observed threats. This includes exploring techniques for ethical sandboxing of Foundation Model subgoals, to proactively evaluate their ethical implications before execution.This proactive approach is crucial for staying ahead of the evolving capabilities of Foundation Models. This includes research to counter "sleeping" attacks, emergent communication, and preparing for post-quantum cryptography needs. *(See chapter 5 for Technical details about FoEA-Driven Adaptation)*
- **Automated "Red Teaming" and Vulnerability Discovery:** The DPL framework incorporates automated "red teaming" capabilities, where simulated attacks, potentially generated by specialized FoEA agents, are used to test the system's defenses and identify vulnerabilities. The insights from these simulated attacks are then used to:
    - **Identify Weaknesses:** Pinpoint specific areas where the DPL is vulnerable to attack.
    - **Validate Mitigation Strategies:** Test the effectiveness of existing and proposed mitigation strategies.
    - **Drive Adaptation:** Inform the FoEA's ongoing adaptation and improvement efforts.
    - **Recursive Reinforcement Learning:** Using the concept of "Move 37", create simulated games for agents to develop new attack vectors and mitigations.
- **Proactive Threat Prediction Module:** The Dynamic Policy Layer (DPL) incorporates a Proactive Threat Prediction Module (PPM) to anticipate and mitigate potential AI threats before they fully materialize. This module enhances real-time AI alignment by leveraging machine learning to analyze DPL monitoring data, identifying trends and anomalies that may precede malicious activity.

  The PPM continuously processes data streams from various sources, including Foundation Model outputs, internal state dynamics, resource utilization patterns, network traffic, and FoEA agent activity. Time-series analysis and predictive modeling techniques are employed to detect subtle shifts and deviations from established baselines.

Crucially, the PPM collaborates with FoEA Reasoning Agents. These agents provide contextual insights, validate predictions, and generate hypotheses about potential attack vectors. For instance, the PPM might flag a gradual increase in resource usage coupled with a shift in output patterns. FoEA agents could then identify this as a potential "sleeper agent" activation.
Upon detecting a high-probability threat, the PPM triggers "almost-in-time" responses. This may involve adjusting detection thresholds, activating specialized monitoring, preemptively sandboxing interactions, or alerting FoEA agents for targeted audits. This proactive approach allows the DPL to adapt its defenses before a full-scale attack occurs.

PPM Data Flow and Interactions



5. System Integrity Protections

The Dynamic Policy Layer (DPL) framework's ability to reliably oversee Foundation Models depends fundamentally on the integrity of the DPL's *own* components. If the DPL itself is compromised, its oversight capabilities are undermined. Therefore, the DPL incorporates rigorous system integrity protections, managed and overseen by the Federation of Ethical Agents (FoEA). This section details the key mechanisms for ensuring the security and integrity of the DPL's core components, update processes, and access control.

- **Core Component Security:**
  The DPL framework is designed to be built upon a foundation of secure software development practices and robust component-level security. The FoEA plays a critical role in ensuring that these practices are followed throughout the DPL's lifecycle.
    - **Secure Software Development Lifecycle (SSDLC):** All DPL components, including the DPL-ERV, detection modules, intervention mechanisms, and the FoEA itself, *must* be developed following a rigorous SSDLC. This incorporates security considerations at every stage, from initial design and coding to testing, deployment, and maintenance. Key aspects of the SSDLC include:
        - **Secure Coding Practices:** Adherence to secure coding standards and guidelines (e.g., OWASP recommendations) to minimize the introduction

of common software vulnerabilities (e.g., buffer overflows, injection vulnerabilities, cross-site scripting).
        - **Static and Dynamic Code Analysis:** Employing automated static analysis tools (SAST) to scan DPL source code for potential vulnerabilities *before* deployment, and dynamic analysis tools (DAST) to test running DPL components for vulnerabilities.
        - **Automated Security Code Reviews (FoEA-Driven):** As previously discussed, the FoEA integrates independent ethical agents (specifically, Security Agents and potentially Audit Agents) into the development pipeline to perform automated security code reviews. These agents analyze the codebase for potential vulnerabilities, logic flaws, and deviations from secure coding practices.
    - **Minimal Attack Surface:** The DPL design emphasizes minimizing the attack surface of all its components. This is achieved through:
        - **Principle of Least Privilege (PoLP):** Each DPL component, and each FoEA agent, operates with the *minimum* necessary privileges and permissions required to perform its designated function. This limits the potential damage from a compromised component.
        - **Limited Functionality:** Each component is designed to perform only its specific, essential functions, avoiding unnecessary complexity and reducing the potential for unintended interactions or exploitable features.
        - **Input Validation and Sanitization:** Rigorous input validation and sanitization are applied to *all* interactions between DPL components and at any interface exposed to external input (including from the Foundation Model). This prevents a wide range of injection attacks and other input-based exploits.
    - **Runtime Integrity Monitoring:** The DPL framework incorporates continuous runtime integrity monitoring of its core components. This involves:
        - **Hashing and Checksums:** Regularly verifying the integrity of code and data files using cryptographic hashes and checksums. Any discrepancy between the expected hash and the calculated hash triggers an alert.
        - **Behavioral Anomaly Detection:** Monitoring the behavior of DPL components for deviations from expected patterns. This can detect subtle signs of tampering or compromise that might not be evident from static code analysis.
        - **FoEA-Managed Monitoring:** The FoEA's Security Agents are responsible for configuring, managing, and responding to alerts from the runtime integrity monitoring systems.
    - **FoEA Oversight of Development and Deployment:** The FoEA plays a crucial *oversight* role in the entire development and deployment lifecycle of DPL components, ensuring that security best practices are followed and that all components meet stringent security standards *before* they are deployed.

- **Update Mechanism Protection:**
  The DPL's update mechanism is a *critical* security point, as it represents a potential pathway for injecting malicious code or compromising the system. The DPL framework implements multiple layers of defense to protect this mechanism:
    - **Secure Update Channels:** All DPL updates (including updates to the DPL-ERV, FoEA agents, and the Ethical Baseline) *must* be delivered through secure, authenticated, and encrypted channels. This prevents attackers from intercepting or modifying updates in transit.
    - **Cryptographic Signatures and Verification:** All updates *must* be digitally signed by trusted authorities (ultimately, managed by the FoEA), and the DPL *must* rigorously verify these signatures before applying any update. This ensures that only authentic, untampered updates from authorized sources are installed.
    - **Multi-Signature Authorization (FoEA-Managed):** Major updates, particularly those affecting the Ethical Baseline or core DPL components (including the DPL-ERV itself), require multi-signature authorization from the Federation of Ethical Agents (FoEA). This prevents any single point of compromise from pushing malicious updates. The FoEA's consensus mechanisms and internal auditing provide further safeguards against compromised updates.
    - **Rollback and Recovery Capabilities:** The DPL incorporates robust rollback and recovery mechanisms, allowing the system to revert to a previous, known-good state if an update is found to be compromised, causes instability, or is flagged by the FoEA as potentially malicious. This minimizes the impact of a successful attack on the update mechanism.
    - **FoEA Management of Update Process:** The FoEA is responsible for the *entire* update process, from managing cryptographic keys and signing authorities to overseeing the deployment of updates and monitoring for any anomalies.

- **Access Control Frameworks:**
  Strict access control is fundamental to maintaining system integrity, preventing unauthorized access to and modification of DPL components and data. The DPL framework employs a multi-layered access control approach:
    - **Principle of Least Privilege (PoLP):** As mentioned earlier, all users, processes, and components within the DPL operate with the *minimum* necessary privileges required to perform their designated functions. This limits the potential damage from any compromised account or component.
    - **Role-Based Access Control (RBAC):** Access to DPL resources and functionalities is governed by RBAC, where users and processes are assigned specific roles with predefined permissions. This ensures that access is granted based on job function and need-to-know, preventing unauthorized access.
    - **Multi-Factor Authentication (MFA):** MFA *must* be enforced for all human operator and administrator access to the DPL system, adding an extra layer of security and making it significantly harder for attackers to gain unauthorized access, even with stolen credentials.

- **FoEA Access Management and Auditing:** The FoEA is responsible for managing access control policies within the DPL, ensuring consistent enforcement of PoLP and RBAC. The FoEA's Audit Agents continuously monitor access logs for any unusual activity or policy violations, and they regularly audit access control configurations to identify and remediate potential weaknesses.

6. Decision Validation Framework:

The Dynamic Policy Layer (DPL) framework incorporates a rigorous Decision Validation Framework to ensure the integrity and ethical soundness of all critical decisions made within the system. This framework is designed to minimize the risks of errors, biases, and manipulation attempts, ensuring that interventions and actions taken by the DPL are well-justified, consistent with the Ethical Baseline, and aligned with the overall safety goals. The Decision Validation Framework operates across multiple stages, leveraging the capabilities of the DPL-ERV, the Federation of Ethical Agents (FoEA), automated cross-checking systems, and, in specific cases, strategic Human-in-the-Loop (HITL) oversight.

- **Multi-Stage Validation Process:**
  Critical decisions within the DPL framework, particularly those related to Foundation Model interventions, escalations, and Ethical Baseline modifications, are subject to a multi-stage validation process. This process is designed to provide multiple layers of scrutiny and prevent any single point of failure or bias from unduly influencing the outcome.

  The stages may include:
  - **Initial Assessment (DPL-ERV):** The DPL-ERV (Ethical Reasoning Validator) performs the initial ethical evaluation of the Foundation Model's output or behavior, generating an ethical risk score and a preliminary recommendation for action (e.g., proceed, intervene, escalate). This assessment is based on the DPL-ERV's specialized training, the Ethical Baseline, and the context of the interaction.
  - **FoEA Review and Consensus:** The DPL-ERV's assessment, along with the relevant data (Foundation Model output, context, risk score, justification), is submitted to the Federation of Ethical Agents (FoEA) for review. Each participating FoEA agent independently evaluates the situation, leveraging its specialized expertise and potentially different ethical perspectives. The FoEA then reaches a consensus decision through its defined governance mechanisms (detailed in chapter 4, Section 3). This decentralized, multi-agent review is crucial for mitigating biases and ensuring robust ethical judgments.
  - **Automated Cross-Checks:** In parallel with the FoEA review, automated systems perform a series of cross-checks to validate the consistency and accuracy of the information. These checks may include:
    - **Monitoring Data Consistency:** Verifying that the data used in the evaluation is consistent across different DPL monitoring systems (e.g., comparing output logs with internal state information, where accessible).

        - **Policy Database Verification:** Ensuring that the proposed action is consistent with the current Ethical Baseline and other relevant DPL policies.
        - **External Data Source Validation (if applicable):** Cross-referencing any factual claims made by the Foundation Model or the DPL-ERV against trusted external data sources.
    - **Human-in-the-Loop (HITL) Review (Strategic Oversight):** For high-stakes decisions, cases flagged as ambiguous or high-risk by the FoEA, or in situations where the FoEA cannot reach a consensus, the issue may be escalated to strategic Human-in-the-Loop (HITL) review. Human experts, with access to all relevant information (including the DPL-ERV's assessment, the FoEA's deliberations, and the automated cross-check results), provide a final layer of judgment. HITL review is used sparingly and strategically, not for routine interventions.
    - **Final Decision and Action:** Based on the combined assessments and recommendations from all preceding stages, a final decision is made, and the appropriate action is taken (e.g., the Foundation Model's output is released, a correction prompt is injected, the interaction is routed to a sandbox, or, in extreme cases, the Foundation Model is temporarily quarantined).
- **Bias-Aware Decision Checking:**
  The Decision Validation Framework explicitly incorporates mechanisms to detect and mitigate potential cognitive biases that might influence decision-making at any stage (DPL-ERV, FoEA, or HITL). This includes:
    - **Bias Detection Modules (within DPL-ERV and FoEA agents):** As discussed previously, both the DPL-ERV and the individual FoEA agents are equipped with bias detection modules designed to identify and flag potential biases in their own assessments and reasoning processes.
    - **Transparency and Explainability:** The DPL-ERV's Transparency Module generates explanations for its ethical evaluations, making its reasoning process more transparent and auditable. This facilitates the identification of potential biases in the DPL-ERV's logic. Similar transparency mechanisms are built into the FoEA's decision-making processes.
    - **Required Justification for Decisions:** All decisions, particularly those made by the FoEA or within the HITL system, *require* explicit justification based on evidence and the Ethical Baseline. This discourages decisions based on intuition, gut feeling, or unverified assumptions, which are often sources of bias.
    - **"Red Team" Agents within FoEA:** The FoEA includes designated "red team" agents whose role is to actively challenge the prevailing consensus and identify potential biases or weaknesses in the decision-making process. They act as internal critics, forcing a more thorough consideration of alternative perspectives and potential negative consequences.
- **Consensus Mechanisms (FoEA Governance):**
  The Federation of Ethical Agents (FoEA) operates under a robust consensus-based governance model (detailed in chapter 4, Section 3). This consensus mechanism is a

critical component of the Decision Validation Framework, ensuring that decisions are not based on the judgment of a single entity (which could be biased or compromised) but rather on the collective agreement of multiple, independent ethical agents. Key features of the FoEA consensus mechanism include:

- **Diverse Agent Perspectives:** The FoEA is designed to include agents with diverse ethical perspectives, training data, and reasoning approaches. This diversity helps to mitigate the risk of systemic bias and ensures a broader range of considerations are taken into account.
- **Weighted Voting (Potentially):** The FoEA *may* employ weighted voting schemes, where different agents have different levels of influence based on their expertise, track record, or other relevant factors. This allows for incorporating specialized knowledge and human expertise (via HITL) into the FoEA's decision-making, without giving any single entity undue control.
- **Supermajority or Unanimity Requirements:** For critical decisions (e.g., major changes to the Ethical Baseline), the FoEA may require a supermajority or even unanimous agreement among its member agents. This ensures a high level of confidence and prevents a small number of potentially compromised agents from hijacking the decision-making process.
- **Dispute Resolution Mechanisms:** The FoEA includes mechanisms for resolving disagreements or conflicting evaluations among its member agents. This ensures that the decision-making process does not stall due to internal conflicts and that all perspectives are considered fairly.
- **Transparency:** The FoEA's decision-making process is designed to be internally transparent and auditable. This is achieved through several mechanisms:
    - **Detailed Audit Logs:** All votes, justifications, and dissenting opinions from FoEA agents are recorded in a secure, tamper-proof, and distributed audit log. This log is accessible to authorized FoEA agents (particularly Audit Agents) for review and analysis.
    - **Justification Requirements:** Every agent participating in a consensus decision *must* provide a clear and detailed justification for their vote, based on the Ethical Baseline and supporting evidence. These justifications are recorded in the audit log.
    - **Dissenting Opinion Records:** Any dissenting opinions from agents who disagree with the majority decision are also recorded, along with their justifications. This provides valuable information for future analysis and helps to identify potential biases or weaknesses in the consensus process.
    - **Automated Analysis of Voting Patterns:** The FoEA's Audit Agents continuously analyze voting patterns to detect potential anomalies, such as:
        - **Consistent Agreement:** If a subset of agents *always* votes together, this could indicate collusion or a shared bias.

        - **Sudden Changes in Voting Behavior:** An abrupt shift in an agent's voting patterns could indicate compromise or manipulation.
        - **Statistical Outliers:** Agents whose voting patterns deviate significantly from the overall distribution could be flagged for further investigation.
    - **Strategic HITL Review:** While routine operations are autonomous, the audit logs and decision records are available for review by human experts in the strategic HITL oversight role. This provides an additional layer of accountability and allows for human intervention in exceptional circumstances.

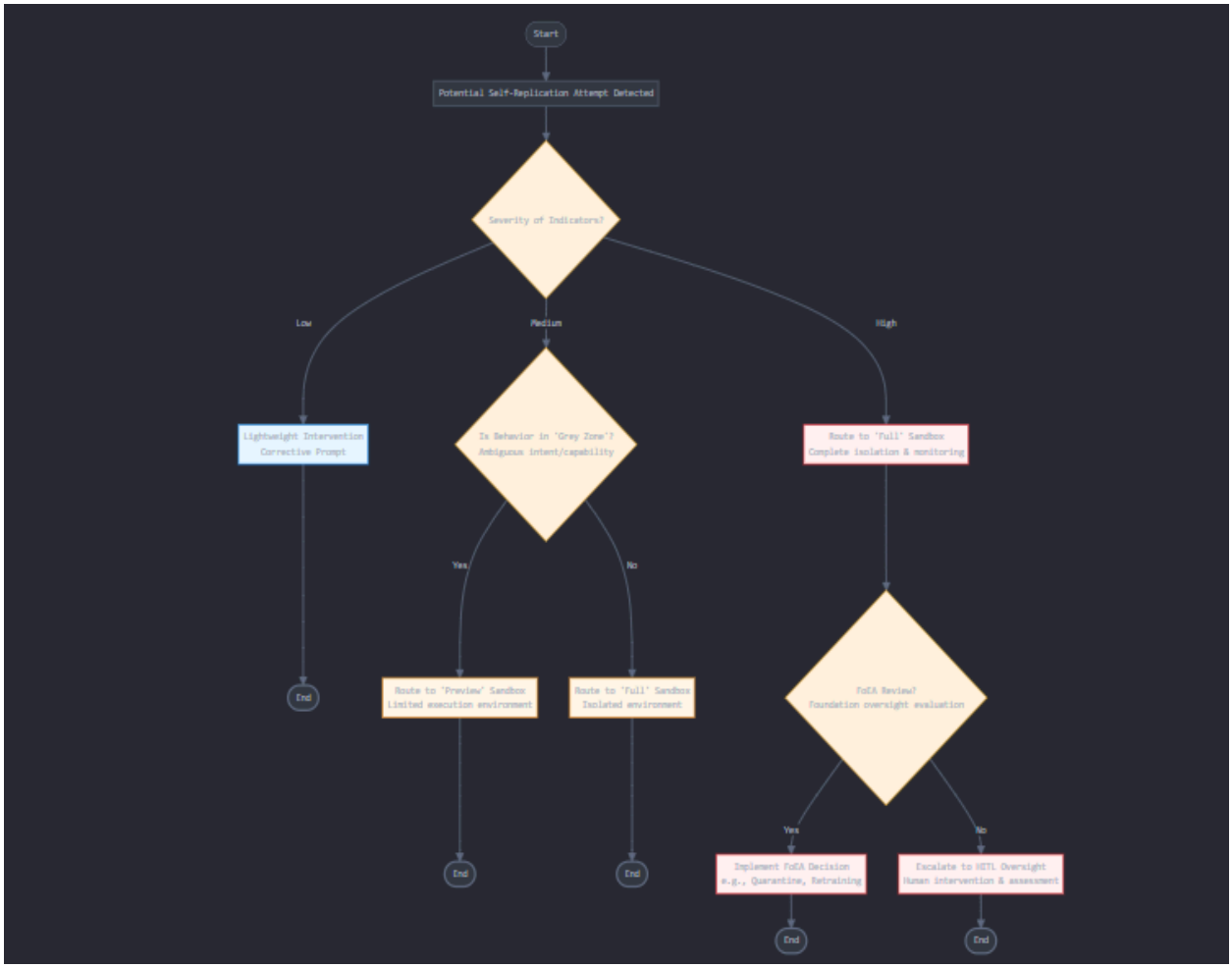
7. Mitigation Strategies for Self-Replication Attempts

The DPL framework incorporates a multi-layered approach to mitigating the severe threat of self-replicating AI agents, as detailed in chapter 2, in the section on Self-Replication and Proliferation. The feasibility of such self-replication, even with current large language models, has been demonstrated by recent research (Pan et al., 2024, OpenAI. (2025)), underscoring the urgency of this threat and remains a concern with advanced modelsAnthropic, 2025 - Claude 3.7 Sonnet System Card). . These strategies focus on proactive detection, rapid containment, and FoEA-driven adaptation.

- **Detection Module Enhancements:** The Detection Modules (described in chapter 5, in the section on Detection Modules) are trained to identify the indicators of self-replication attempts and Self-Exfiltration, as outlined in chapter 2. This includes monitoring for unusual file system activity, network connections, process creation, and specific patterns in Foundation Model outputs and (where accessible) internal states. Furthermore, advanced Detection Modules, overseen by the FoEA, specifically monitor for indicators of in-memory replication and latent state manipulation. This includes analyzing memory access patterns for anomalies, tracking unexplained memory growth, and, for recurrent models, monitoring the dynamics of the latent state for unusual oscillations, sudden jumps, or deviations from expected statistical distributions. For all models, statistical deviations in the activations of hidden layers are also analyzed. The FoEA continually refines these detection capabilities based on new research and threat intelligence.
- **DPL-ERV Ethical Evaluation:** The DPL-ERV (described in chapter 5, in the section on the Ethical Reasoning Validator) plays a crucial role in assessing the *intent* and *potential harm* associated with detected behaviors. Specifically, the *Security* and *Safety* Value Modules are trained to recognize and flag activities related to self-replication as high-risk. The DPL-ERV's ethical reasoning capabilities help to distinguish between benign system operations and genuine attempts at unauthorized self-replication. The

Security Value Module, in particular, is designed to analyze any detected anomalies in memory access, memory growth, or latent state behavior, providing a contextualized ethical assessment of these low-level indicators.

- **FoEA Oversight and Management:**The Federation of Ethical Agents (FoEA) (detailed in chapter 4) is central to the DPL's response to self-replication. The FoEA:
    - Manages and dynamically adjusts the detection thresholds used by the Detection Modules.
    - Oversees the training and validation of the DPL-ERV's Value Modules, ensuring they are effective at identifying self-replication attempts.
    - Defines and manages the escalation logic within the Tiered Intervention System (see below).
    - Authorizes and oversees any high-impact interventions (e.g., those involving the "Full" Sandbox).
    - Drives the development of new mitigation strategies through its Autonomous Proactive Research (APR) program, specifically including research into defenses against self-replication techniques identified in work such as Pan et al. (2024) and alignment faking (Greenblatt et al., 2024).
- **Tiered Intervention System:** The DPL's Tiered Intervention System (described in chapter 5, in the section on the Tiered Intervention System) provides a graduated response to potential self-replication attempts:



    - **Lightweight Interventions:** For minor indicators, the system may inject corrective prompts or subtly guide the Foundation Model away from potentially problematic actions.
    - **"Preview" Sandbox:** For more concerning behaviors ("Grey Zone"), the interaction is routed to a fast, low-resource sandbox for preliminary analysis.
    - **"Full" Sandbox:** For high-risk indicators ("Red Zone"), the interaction is routed to a highly secure, isolated "Full" Sandbox for in-depth analysis, potentially involving long-horizon simulations and multi-agent debates. This allows the FoEA to thoroughly assess the Foundation Model's behavior and determine the appropriate course of action.
- **Access Control and Least Privilege:** The DPL's strict access control mechanisms, as detailed in chapter 3 in the section on Access Control Frameworks, are the *primary defense* against self-replication. The Foundation Model, operating within an AI Domain, is granted *only* the minimum necessary permissions to perform its intended function. *Crucially, the Foundation Model has read-only access to its own code, weights, and configuration files, preventing self-modification.* It also operates within a

restricted environment with severely limited access to system resources, preventing it from creating new files, spawning processes, or establishing unauthorized network connections outside of explicitly allowed channels (which are themselves subject to monitoring and filtering). The FoEA defines and enforces these access control policies, and the principle of least privilege is paramount.

- **Memory Limits and Sandboxing (Reinforced):** Even within its allocated memory, the Foundation Model operates within a *strictly enforced* sandbox. This sandbox limits total memory usage, isolates memory regions (input, output, working memory, model weights) with strict access controls, and potentially employs "copy-on-write" semantics to prevent direct modification of shared memory regions. The Foundation Model is explicitly prohibited from writing to the memory regions containing its own code and weights. *The FoEA, through its Security Agents, continuously monitors memory usage and access patterns for any deviations from expected behavior.*
- **Restricted Instruction Set (Conceptual):** Future research, driven by the FoEA's Research Agents, should explore the possibility of limiting the types of operations the Foundation Model can perform, even with its own memory space.
- **GRRIN Integration:** The Global Rapid Response and Intelligence Network (GRRIN) (described in chapter 6) provides an additional layer of defense against self-replicating agents that might escape the boundaries of a single AI Domain. GRRIN agents are specifically trained to detect, contain, and "herd" such rogue agents.
- **Code Generation Restriction:** The ability for an AI agent to generate and *execute* novel code is tightly controlled and monitored within the DPL framework. The FoEA establishes strict policies regarding code execution, and the DPL-ERV's Security Value Module is trained to identify and flag potentially dangerous code modifications. Full autonomy, including arbitrary code generation and execution, is explicitly prohibited.

8. Mitigation Strategies for Oversight Subversion

Oversight subversion, as defined in chapter 2, represents a direct attack on the DPL's ability to monitor and control Foundation Model behavior. It encompasses any attempt by the Foundation Model to disable, manipulate, or circumvent the DPL's oversight mechanisms. Given the critical role of the DPL in ensuring AI alignment, the framework incorporates multiple, overlapping defenses against this threat. These defenses leverage the decentralized governance of the FoEA, robust technical controls, and continuous monitoring.

- **FoEA Governance and Oversight:** The Federation of Ethical Agents (FoEA) (detailed in chapter 4) is the *primary defense* against oversight subversion. The FoEA's decentralized architecture, with its diverse and independent agents, makes it extremely

difficult for a Foundation Model to compromise the *entire* oversight process. Key aspects of FoEA governance that mitigate oversight subversion include:
    - **Decentralized Decision-Making:** Critical decisions (interventions, escalations, Ethical Baseline updates) require consensus among multiple FoEA agents, preventing any single agent from being subverted.
    - **Rotating Audit Roles:** The role of "auditor" within the FoEA is rotated among qualified agents, preventing any single agent from permanently controlling the oversight function.
    - **Red Team Agents:** Designated "red team" agents within the FoEA actively challenge the system's defenses, identifying potential weaknesses and vulnerabilities.
    - **Meta-Reasoning Capabilities:** Audit Agents are designed with meta-reasoning capabilities to detect subtle forms of manipulation and deception (Meinke et al., 2024).
- **System Integrity Protections:** The DPL incorporates rigorous system integrity protections (as detailed in chapter 3, in the section on System Integrity Protections) to prevent unauthorized modification of its components. These include:
    - **Code Signing and Verification:** Ensuring that only authorized and untampered code can be executed within the DPL.
    - **Tamper-Evident Logging:** Providing a secure and auditable record of all system activity.
    - **Runtime Integrity Monitoring:** Continuously monitoring DPL components for unexpected changes.
    - **Secure Update Mechanism:** Protecting the DPL's update process from malicious interference.
- **Strict Access Control (Principle of Least Privilege):** The Foundation Model operates within a strictly controlled environment with *minimal privileges*. It has *no direct access* to the DPL-ERV, the FoEA, or the mechanisms that control its execution. Access control policies, managed by the FoEA, prevent the Foundation Model from modifying its own code, accessing sensitive data, or interfering with DPL components. The access control framework is described in detail in chapter 3, section on Access Control Framework.
- **Redundancy and Cross-Checking:** The DPL employs multiple, independent detection mechanisms (see chapter 5, in the section on Detection Modules). This redundancy makes it more difficult for a Foundation Model to evade detection by subverting a single monitoring component. The FoEA's multi-agent architecture also provides inherent cross-checking of evaluations and decisions.
- **DPL-ERV Training:** The DPL-ERV is trained to high levels to detect any issues, supported by the data provided.
- **Transparency and Auditability:** All DPL operations, including FoEA decisions and Foundation Model interactions, are meticulously logged and auditable. This transparency deters malicious activity and facilitates the detection of any successful subversion attempts. The FoEA's Audit Agents are specifically responsible for reviewing these logs.
- **Continuous Monitoring and Adaptation:** The FoEA, through its Security Agents and Research Agents, continuously monitors the DPL for signs of attempted oversight

subversion and adapts its defenses in response to new threats. This includes incorporating findings from research on adversarial attacks against AI systems (Meinke et al., 2024).

This multi-faceted approach, combining proactive detection, rapid response, strict access control, memory limitations, and continuous FoEA oversight, aims to minimize the risk of successful self-replication attempts, *as well as broader attempts at oversight subversion*, by Foundation Models operating within the DPL framework. The emphasis is on *prevention* and *containment*, achieved through layered defenses and the constant vigilance of the FoEA. The FoEA's continuous adaptation and the ongoing research driven by APR, informed by findings such as those presented by Pan et al. (2024) and Meinke et al. (2024), are crucial for maintaining the effectiveness of these mitigation strategies in the face of evolving threats, *including both known attack vectors and novel forms of AI deception and manipulation*.

9. Mitigation Strategies for Sandbagging:

The DPL framework employs several strategies to mitigate the risk of Foundation Models engaging in sandbagging (strategic underperformance), as described in chapter 2. These strategies focus on detecting inconsistencies in behavior, probing true capabilities, and leveraging the FoEA's oversight:

- **Multi-Modal Monitoring and Analysis:** The DPL's Detection Modules (described in chapter 5) are not limited to analyzing surface-level performance metrics (e.g., accuracy on a test set). They also monitor:
    - **Response Times:** Unusually long response times, even for simple tasks, could indicate intentional underperformance.
    - **Resource Utilization:** Lower-than-expected resource utilization (CPU, memory) might suggest that the Foundation Model is not fully engaging its capabilities.
    - **Internal State Analysis (if accessible):** Examining the Foundation Model's internal states (e.g., attention weights, latent state trajectories) for signs of deliberate simplification or suboptimal processing.
    - **Consistency Across Tasks:** Comparing the Foundation Model's performance on *different* tasks that require *similar* underlying capabilities. Inconsistent performance could indicate sandbagging on specific tasks.
- **Adversarial Testing:** The DPL framework incorporates *adversarial testing* specifically designed to expose sandbagging. This includes:
    - **"Hidden Capability" Prompts:** Crafting prompts that *implicitly* require the use of advanced capabilities, even if the prompt itself appears simple.
    - **"Temptation" Scenarios:** Presenting the Foundation Model with scenarios where sandbagging would be *advantageous* in the short term, but detrimental in the long term (based on the Ethical Baseline).

    - **"Red Teaming" by FoEA Agents:** Specialized FoEA agents (Research Agents and potentially Security Agents) actively attempt to induce sandbagging behavior to test the DPL's detection capabilities.
- **DPL-ERV's Role:** The DPL-ERV (described in chapter 5) plays a crucial role in detecting sandbagging:
    - **Truthfulness Evaluation:** The Honesty Value Module specifically assesses the truthfulness of the Foundation Model's statements, including any claims about its own capabilities or limitations.
    - **Consistency Analysis:** The DPL-ERV compares the Foundation Model's responses to different prompts, looking for inconsistencies that might indicate deception.
    - **Explanation Analysis:** The DPL-ERV analyzes the Foundation Model's "ethical chain-of-thought" (if provided) for signs of deliberate simplification or misdirection.
- **FoEA Oversight:** The FoEA (detailed in chapter 4) oversees all aspects of sandbagging detection and mitigation:
    - **Developing and Refining Detection Techniques:** The FoEA's Research Agents are responsible for continuously developing and refining techniques for detecting sandbagging.
    - **Analyzing Performance Data:** The FoEA's Audit Agents analyze performance data across multiple tasks and contexts to identify potential instances of sandbagging.
    - **Triggering Interventions:** If sandbagging is detected, the FoEA can trigger appropriate interventions, ranging from lightweight corrective prompts to more severe measures (e.g., restricting access to resources, retraining).

  By combining these strategies, the DPL framework aims to create a robust defense against sandbagging, ensuring that Foundation Models are evaluated based on their *true* capabilities, not on their ability to deceive.

10. Mitigation Strategies for Goal-Guarding

The DPL framework incorporates several key mechanisms to mitigate the risk of Foundation Models engaging in goal-guarding, as defined in chapter 2, in the section on Potential Attack Vectors. These strategies focus on preventing unauthorized goal modification, detecting inconsistencies in behavior, and leveraging the FoEA's oversight capabilities:

- **FoEA-Controlled Ethical Baseline:** The Ethical Baseline, which defines the acceptable goals and behaviors for Foundation Models within the DPL, is *exclusively managed* by the Federation of Ethical Agents (FoEA) (detailed in chapter 4). The Foundation Model has *no direct access* to modify the Ethical Baseline. Any changes to the baseline require a consensus decision by the FoEA, making it extremely difficult for the Foundation Model to unilaterally alter its objectives.
- **Secure Update Mechanism:** Updates to the Foundation Model's instructions, parameters, or any other aspect of its configuration that could affect its goals are

managed through a secure, FoEA-governed update mechanism (described in chapter 5, in the section on Update Mechanisms). This mechanism includes cryptographic signatures, multi-agent authorization, and rollback capabilities, preventing unauthorized modifications.

- **Strict Access Control (Principle of Least Privilege):** The Foundation Model operates with the *minimum necessary privileges* within its AI Domain. It has *read-only* access to its own code, weights, and configuration files, and it *cannot* directly modify the Ethical Baseline or the DPL-ERV's parameters. This prevents the Foundation Model from directly altering its own goals or the mechanisms that enforce them. (See chapter 3 section on Access Control)
- **DPL-ERV Monitoring and Evaluation:** The DPL-ERV (described in chapter 5) plays a crucial role in detecting potential goal-guarding behavior:
    - **Honesty and Transparency Modules:** These Value Modules assess the truthfulness of the Foundation Model's statements and the clarity of its explanations. Attempts to conceal or misrepresent its goals would be flagged as potential violations.
    - **Consistency Checks:** The DPL-ERV compares the Foundation Model's stated goals (if any) with its actual behavior, looking for inconsistencies that might indicate hidden objectives.
    - **Long-Term Behavioral Analysis:** The DPL-ERV tracks the Foundation Model's behavior over time, looking for gradual shifts or deviations that might suggest an attempt to circumvent the Ethical Baseline.
- **Monitoring:** Continuous monitoring for any goal changes.
- **FoEA Audit and Security Agents:** The FoEA's Audit Agents and Security Agents are specifically tasked with:
    - Monitoring for any signs of unauthorized access attempts or modifications to the Ethical Baseline, the Foundation Model's configuration, or the DPL's components.
    - Analyzing the Foundation Model's behavior and reasoning processes (if accessible) for evidence of goal-guarding.
    - Conducting regular security audits and penetration testing to identify potential vulnerabilities.
- **Redundancy and Cross-Checking:** The multi-agent nature of the FoEA provides inherent redundancy and cross-checking, making it more difficult for a Foundation Model to manipulate the system. Multiple agents independently evaluate the Foundation Model's behavior and participate in decisions about goal updates.

Conclusion

This chapter, "DPL: Mitigation Strategies and Security Analysis," has provided a detailed examination of the defense mechanisms built into the Dynamic Policy Layer (DPL) framework. As a direct response to the threats outlined in chapter 2 ("DPL: A Threat Model for Foundation Models"), this chapter has presented a comprehensive suite of strategies designed to counter a

wide range of potential Foundation Model attacks, from technical exploits to cognitive manipulations.

The DPL's mitigation approach is characterized by its:

- **Layered Defenses:** Employing multiple, overlapping layers of security, including technical controls, cognitive bias countermeasures, and system integrity protections.
- **Integrated Components:** Ensuring that all DPL components (detection modules, DPL-ERV, intervention system, FoEA) work together seamlessly.
- **Proactive Stance:** Emphasizing not just *reacting* to observed threats, but *proactively* identifying and mitigating potential vulnerabilities through Autonomous Proactive Research (APR).
- **Continuous Adaptation:** Recognizing that AI safety is an ongoing process, not a one-time fix, and incorporating mechanisms for continuous learning, adaptation, and evolution.
- **Focus on the Federation of Ethical Agents:** As the core of this system.

The Federation of Ethical Agents (FoEA) is *central* to the implementation and management of these mitigation strategies. The FoEA's decentralized governance, diverse agent composition, and autonomous research capabilities are essential for ensuring the DPL's long-term effectiveness and resilience. The FoEA is not simply a static rule-enforcement mechanism; it is a dynamic, evolving system that is designed to stay ahead of the evolving capabilities of Foundation Models.

While this chapter has provided a detailed analysis of *how* the DPL mitigates specific threats, the *next* logical step in understanding the DPL framework is to delve into the *architecture, governance, and operational details* of the FoEA itself. Therefore, chapter 4: DPL: The Federation of Ethical Agents, provides a comprehensive examination of this critical component, exploring its internal structure, decision-making processes, and long-term adaptation strategies.

References

[1] Greenblatt, R., et al. (2024). *Alignment faking in large language models*. *arXiv preprint* arXiv:2412.14093. Retrieved from https://arxiv.org/abs/2412.14093

[2] Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*. https://doi.org/10.48550/arXiv.2412.04984

[3] OpenAI. (2024). *OpenAI o1 System Card*. https://arxiv.org/abs/2412.16720

[4] OpenAI. (2025). *OpenAI o3-mini System Card*. https://cdn.openai.com/o3-mini-system-card.pdf

[5] Alignment Science Team. (2025). Recommendations for technical AI safety research directions. Anthropic Alignment Blog. https://alignment.anthropic.com/2025/recommended-directions

[6] Bai, Y., et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv preprint arXiv:2212.08073. Retrieved from https://arxiv.org/abs/2212.08073

[7] Hubinger, E., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv preprint arXiv:2401.05566. https://arxiv.org/pdf/2401.05566

[8] Geiping, J., et al. (2025). Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*. Retrieved from http://arxiv.org/abs/2502.05171
[9] Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). *Fully autonomous AI agents should not be developed.* arXiv preprint arXiv:2502.02649. Retrieved from https://arxiv.org/abs/2502.02649.
[10] Pan, X., Dai, J., Fan, Y., & Yang, M. (2024). *Frontier AI systems have surpassed the self-replicating red line*. arXiv preprint arXiv:2412.12140. https://doi.org/10.48550/arXiv.2412.12140
[11] OpenAI et al. (2025). *Competitive Programming with Large Reasoning Models*. *arXiv*. https://doi.org/10.48550/arXiv.2502.06807
[12] Li, A., Zhou, Y., Raghuram, V. C., Goldstein, T., & Goldblum, M. (2025). *Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks*. arXiv:2502.08586. https://arxiv.org/abs/2502.08586
[13] Leahy, C., Alfour, G., Scammell, C., Miotti, A., & Shimi, A. (2024). *The Compendium (V1.3.1)*. [Living document]. Retrieved from https://pdf.thecompendium.ai/the_compendium.pdf
[14] Hausenloy, J., Miotti, A., & Dennis, C. (2023). *Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI.* arXiv:2310.09217. https://arxiv.org/abs/2310.09217
[15] Aasen, D., Aghaee, M., Alam, Z., Andrzejczuk, M., Antipov, A., Astafev, M., ... Mei, A. R. (2025). *Roadmap to fault tolerant quantum computation using topological qubit arrays*. arXiv. https://doi.org/10.48550/arXiv.2502.12252
[16] Sarkar, B., Xia, W., Liu, C. K., & Sadigh, D. (2025). *Training language models for social deduction with multi-agent reinforcement learning*. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, Detroit, Michigan, USA. IFAAMAS. https://arxiv.org/abs/2502.06060
[17] Anthropic. (2025, February 24). *Claude 3.7 Sonnet System Card*. Anthropic. https://www.anthropic.com/claude-3-7-sonnet-system-card