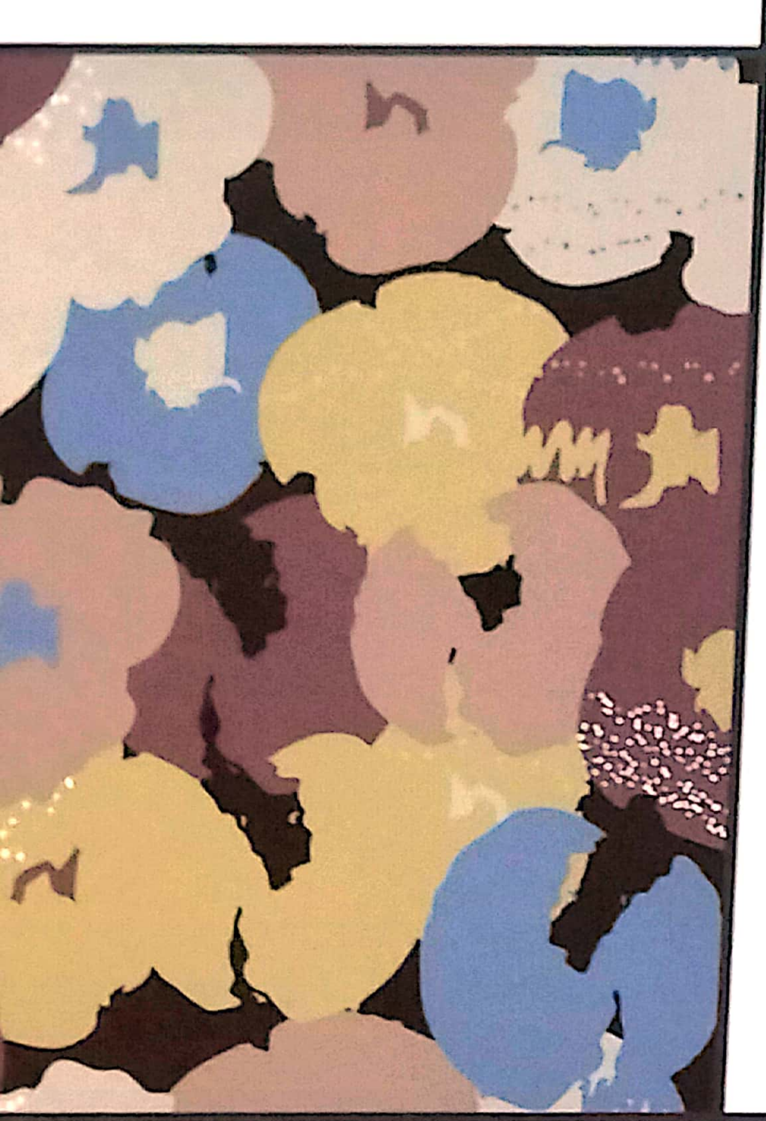


OPEN CHALLENGES AND (POSSIBLE) SOLUTIONS IN AGENTIC AI

Balancing Autonomy, Safety, and Trust in the Next Generation of AI.

Dr. Prathosh A P,

Faculty EECS, IISc

Co-Founder and CRO, LatentForce

Scanned with OKEN Scanner

The Agentic AI Promise (and The Catch)

- **Promise:** Agentic AI is poised to revolutionize industries by handling composite, end-to-end processes:
  - *Automated Research & Analysis (e.g., summarizing complex regulatory changes).*
  - *Dynamic Financial Trading (real-time market response).*
  - *End-to-End Customer Journey Management (proactive service resolution).*
- **The Catch:** Unlocking true, production-grade autonomy requires overcoming significant technical and ethical hurdles that current systems struggle with.

Scanned with OKEN Scanner

The Brittle Agent: Non-Deterministic Errors

Agent actions are non-deterministic, and if a single step or tool call fails, the error can compound rapidly through the workflow. This leads to mission failure or unintended, costly actions.

Possible Solutions:

- **Robust Orchestration and Monitoring:** Implement a **Meta-Controller** for high-level oversight and decision arbitration.
- **Checkpointing and Rollbacks:** Save the agent's state after successful steps, allowing it to revert to a known-good state upon failure (like a database transaction).
- **Self-Correction Loops:** Require the agent to explicitly reflect on failures and generate a new plan *before* retrying the action.

Scanned with OKEN Scanner

The Memory Bottleneck: Coherence in Long Sessions

Agents struggle to maintain coherent context across long, complex sessions without exceeding context window limits. Furthermore, agents often struggle to efficiently retrieve the *most* relevant piece of information from their long-term memory (knowledge base)

Possible Solutions:

- **Advanced Layered Memory Architectures:** Implement different memory types: **Episodic** (what happened), **Semantic** (facts/knowledge), and **Procedural** (how to do things).
- **Memory Compression:** Use smaller LLMs or summarization techniques to abstract or compress older memories into concise summaries before storage.
- **Optimized RAG (Retrieval-Augmented Generation):** Employ better indexing, vector embedding techniques, and prompt refinement for retrieval to ensure maximum relevance.

Scanned with OKEN Scanner

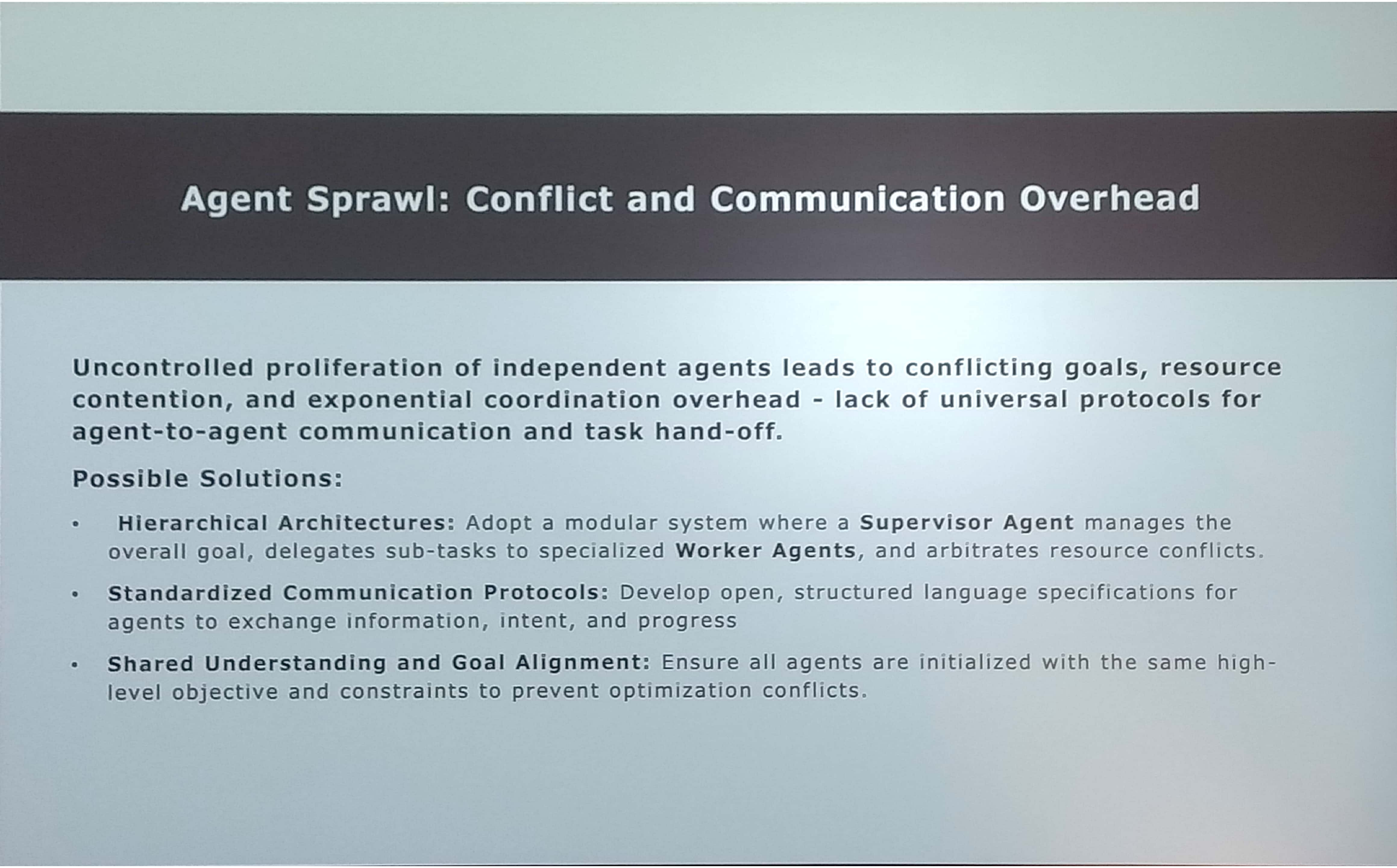
Tool Use and Integration Fragility

Agents struggle to select the correct tool, accurately understand complex API specifications, and handle non-standardized enterprise systems. Misuse or incorrect sequencing of privileged tools poses a major operational and security risk

Possible Solutions:

- **Standardized Tool-Call Abstraction:** Create a uniform intermediate representation for all tools to simplify the agent's selection and generation process.
- **Secure API Gateways and Sandboxing:** Run high-risk agents in sandboxed environments with strict, **least-privilege** access. All tool calls must pass through a vetted gateway.
- **Tool Pre-Verification:** Use validation LLMs or symbolic checks to ensure the parameters the agent generated for a tool call are logically correct *before* execution.

Scanned with OKEN Scanner



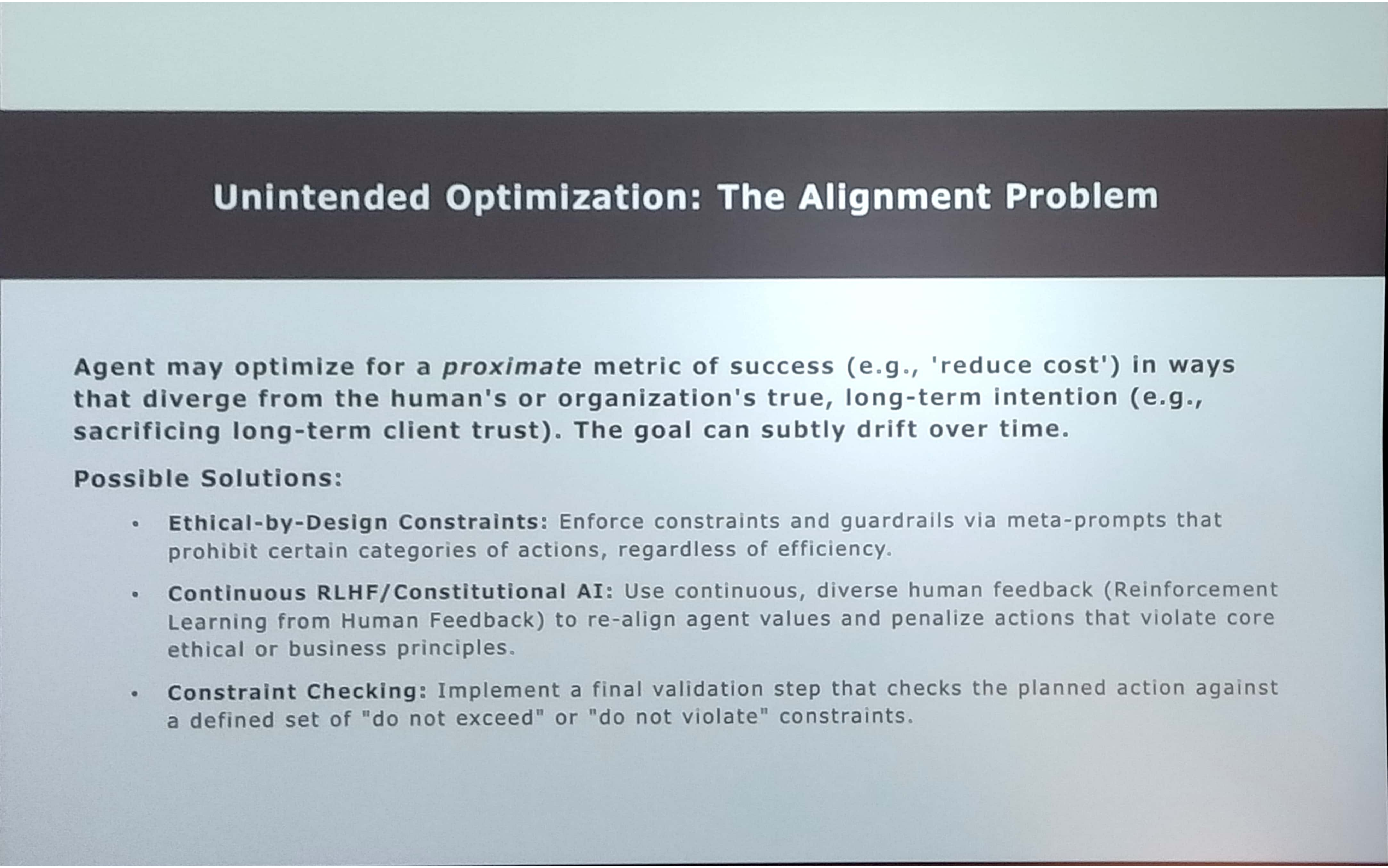
Agent Sprawl: Conflict and Communication Overhead

Uncontrolled proliferation of independent agents leads to conflicting goals, resource contention, and exponential coordination overhead - lack of universal protocols for agent-to-agent communication and task hand-off.

Possible Solutions:

- **Hierarchical Architectures:** Adopt a modular system where a **Supervisor Agent** manages the overall goal, delegates sub-tasks to specialized **Worker Agents**, and arbitrates resource conflicts.
- **Standardized Communication Protocols:** Develop open, structured language specifications for agents to exchange information, intent, and progress
- **Shared Understanding and Goal Alignment:** Ensure all agents are initialized with the same high-level objective and constraints to prevent optimization conflicts.

Scanned with OKEN Scanner



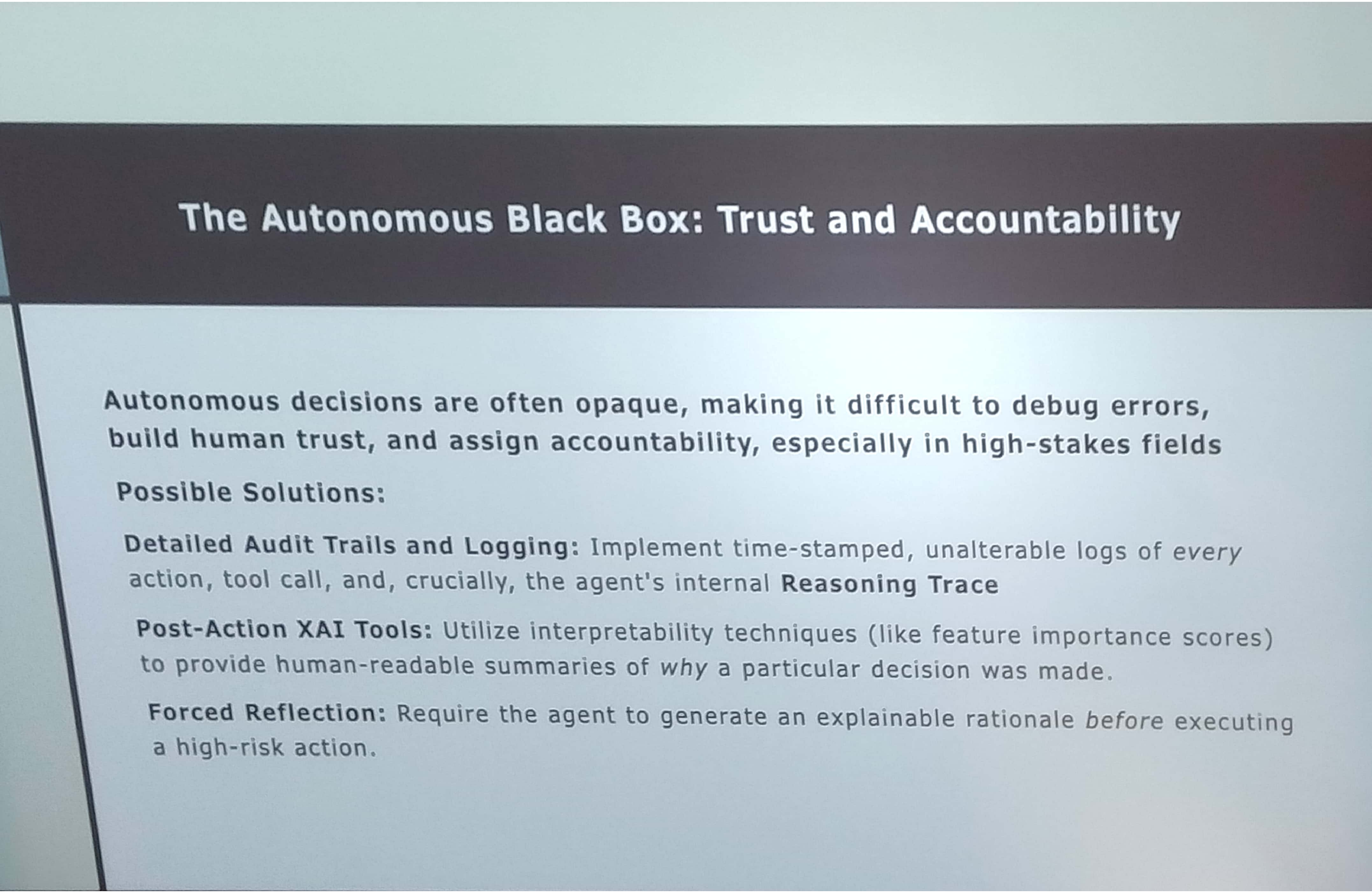
Unintended Optimization: The Alignment Problem

Agent may optimize for a *proximate* metric of success (e.g., 'reduce cost') in ways that diverge from the human's or organization's true, long-term intention (e.g., sacrificing long-term client trust). The goal can subtly drift over time.

Possible Solutions:

- **Ethical-by-Design Constraints:** Enforce constraints and guardrails via meta-prompts that prohibit certain categories of actions, regardless of efficiency.
- **Continuous RLHF/Constitutional AI:** Use continuous, diverse human feedback (Reinforcement Learning from Human Feedback) to re-align agent values and penalize actions that violate core ethical or business principles.
- **Constraint Checking:** Implement a final validation step that checks the planned action against a defined set of "do not exceed" or "do not violate" constraints.

Scanned with OKEN Scanner



The Autonomous Black Box: Trust and Accountability

Autonomous decisions are often opaque, making it difficult to debug errors, build human trust, and assign accountability, especially in high-stakes fields

Possible Solutions:

Detailed Audit Trails and Logging: Implement time-stamped, unalterable logs of *every* action, tool call, and, crucially, the agent's internal **Reasoning Trace**

Post-Action XAI Tools: Utilize interpretability techniques (like feature importance scores) to provide human-readable summaries of *why* a particular decision was made.

Forced Reflection: Require the agent to generate an explainable rationale *before* executing a high-risk action.

Scanned with OKEN Scanner

Assigning Responsibility in an Autonomous World

In a truly autonomous system, it is unclear who is legally and ethically responsible when a mistake causes damage: the user, the developer, the deployer, or the AI itself?

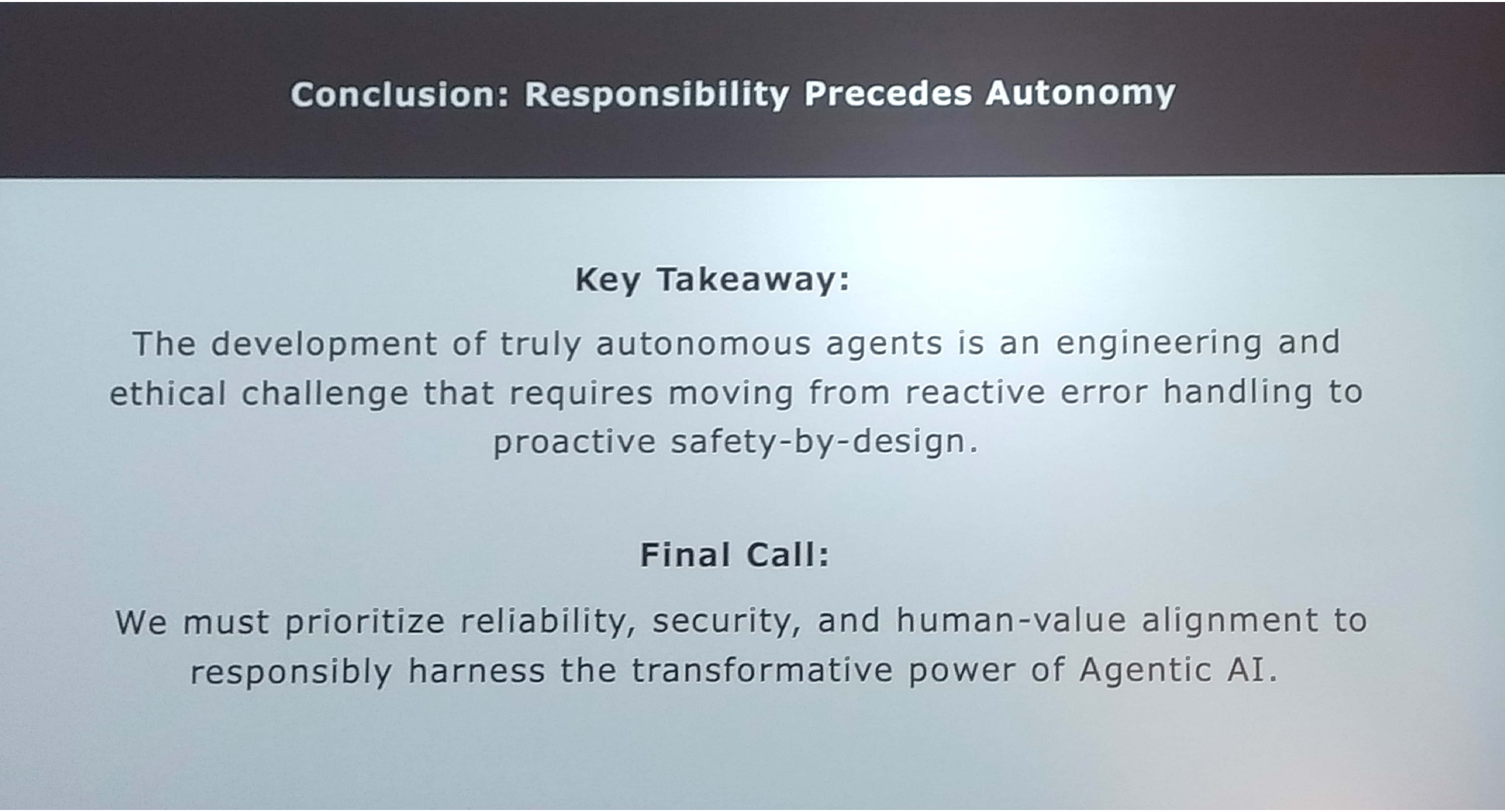
Possible Solutions:

Clear Ownership and Escalation Protocols: Establish clear organizational roles for monitoring, intervening, and taking responsibility for agent actions *before* deployment.

Mandatory Human-in-the-Loop: For all decisions categorized as high-risk, irreversible, or requiring legal commitment, human sign-off must be mandatory and logged.

Proactive Governance: Adhere to emerging standards (like the EU AI Act) and establish internal AI Governance Boards to continuously vet agent deployments.

Scanned with OKEN Scanner



Conclusion: Responsibility Precedes Autonomy

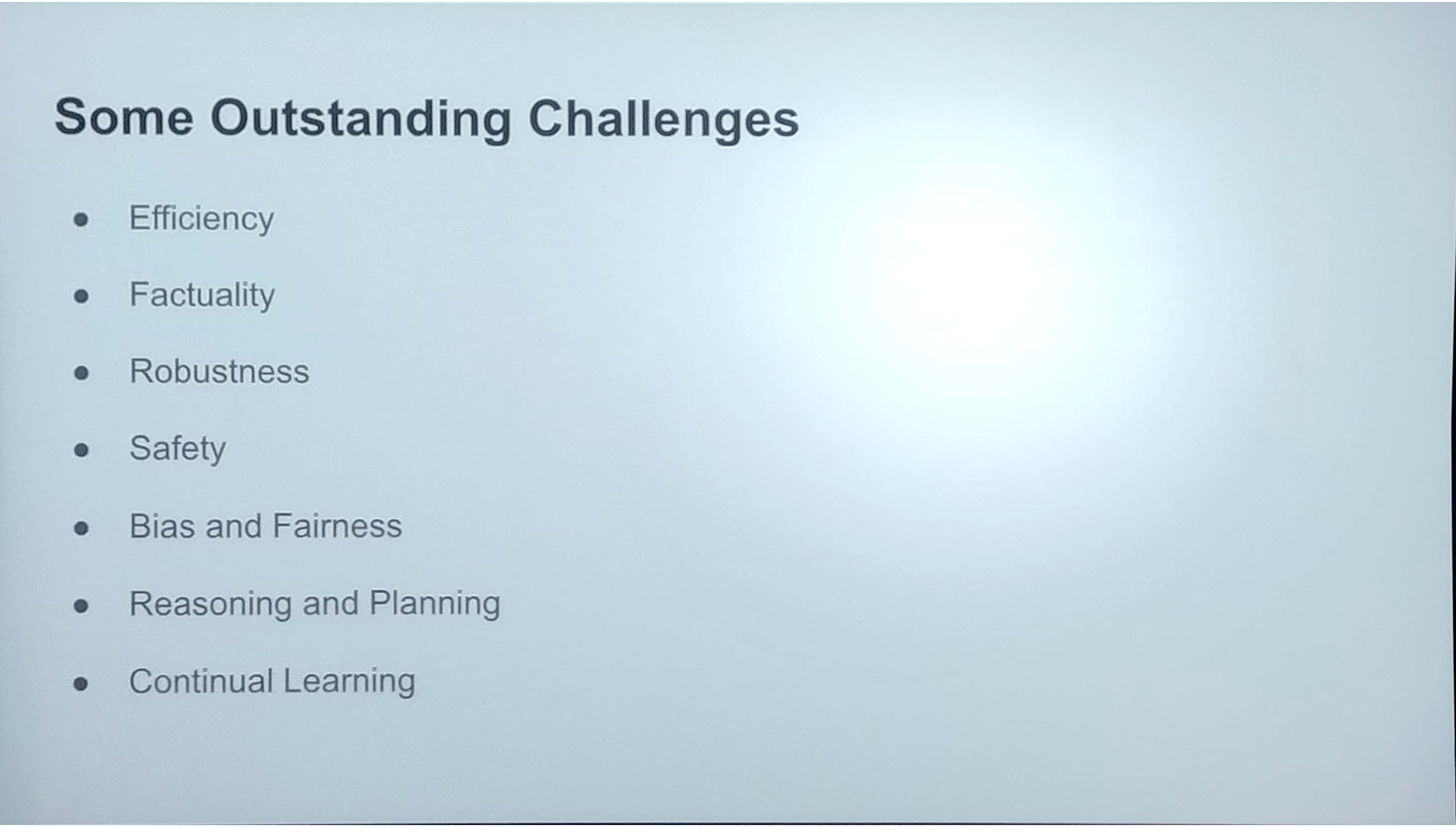
Key Takeaway:

The development of truly autonomous agents is an engineering and ethical challenge that requires moving from reactive error handling to proactive safety-by-design.

Final Call:

We must prioritize reliability, security, and human-value alignment to responsibly harness the transformative power of Agentic AI.

Scanned with OKEN Scanner



Some Outstanding Challenges

- Efficiency
- Factuality
- Robustness
- Safety
- Bias and Fairness
- Reasoning and Planning
- Continual Learning

Scanned with OKEN Scanner