

def/acc Hackathon Report: CIRISAgent GUI Wizard

Ethical AI Through Transparent Architecture

Eric Moore, Emile Eid, Fahim Khan, Min-Si Wang

With Apart Research

Abstract

We present CIRISAgent, an open-source AI framework that reimagines autonomous systems through transparent, explainable architecture. During this hackathon weekend, we transformed CIRIS from a command-line tool into a production-ready system with a complete GUI onboarding wizard, making ethical AI accessible to non-technical users. The new interface guides users through LLM configuration, account creation, agent template selection, and service setup—all built in just 48 hours. Our framework features a 22-service microarchitecture with cryptographic audit trails, ethical reasoning through our H3ERE conscience module, and mandatory human oversight pathways. Initial deployment shows promising results with only two successful jailbreaks from dedicated testing and efficient resource usage (228MB RAM).

Keywords: Multi-agent alignment, AI security, transparent architecture, ethical AI, hackathon innovation

1. Introduction: Weekend Innovation

1.1 Hackathon Achievement

Before this weekend: CIRIS existed only as a command-line interface requiring technical expertise. Users had to manually edit configuration files and run terminal commands.

48 hours later: We built a complete production-ready GUI wizard that transforms the onboarding experience. Non-technical users can now deploy ethical AI agents through an intuitive visual interface.

1.2 Core Innovation

CIRISAgent addresses AI alignment through architectural transparency. Unlike opaque "black box" systems, every decision is explainable, auditable, and subject to human oversight. Our H3ERE conscience module provides runtime ethical evaluation, ensuring AI agents recognize their limitations and defer to humans when appropriate.

2. Methods: Building the Wizard

2.1 Weekend Development Sprint

Our hackathon focus centered on accessibility. We built:

- **Step-by-step onboarding wizard** - Guides users through complete setup in 5 intuitive steps
- **LLM provider configuration** - Auto-detects and tests connections to OpenAI, local models, or compatible APIs
- **Visual agent template selection** - Pre-configured templates for moderation, GDPR compliance, and customer service

- **Modular service configuration** - Toggle capabilities like SQL databases, weather APIs, and IoT sensors
- **Secure authentication system** - Multi-layered security with admin accounts and certificate-based API access

2.2 Architectural Foundation

While the GUI was our hackathon focus, it builds upon CIRIS's robust 22-service microarchitecture:

- **Graph Services:** Memory, configuration, audit trails
- **Infrastructure Services:** Authentication, resource monitoring, consent management
- **Governance Services:** Human oversight (Wise Authority), transparency, self-observation

All agent actions are constrained to ten verbs (OBSERVE, SPEAK, TOOL, MEMORIZE, RECALL, FORGET, REJECT, PONDER, DEFER, TASK_COMPLETE) that traverse ethical vetting through our H3ERE conscience module.

3. Results: From CLI to Production GUI

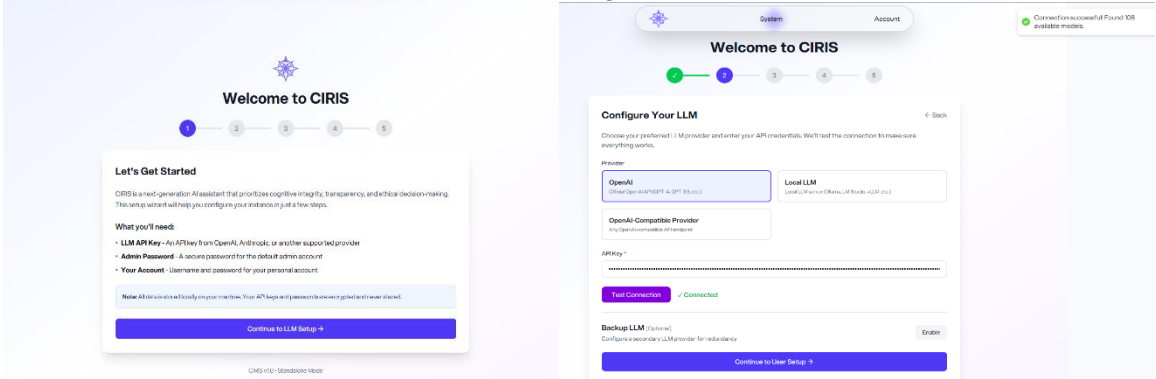
3.1 Interface Screenshots

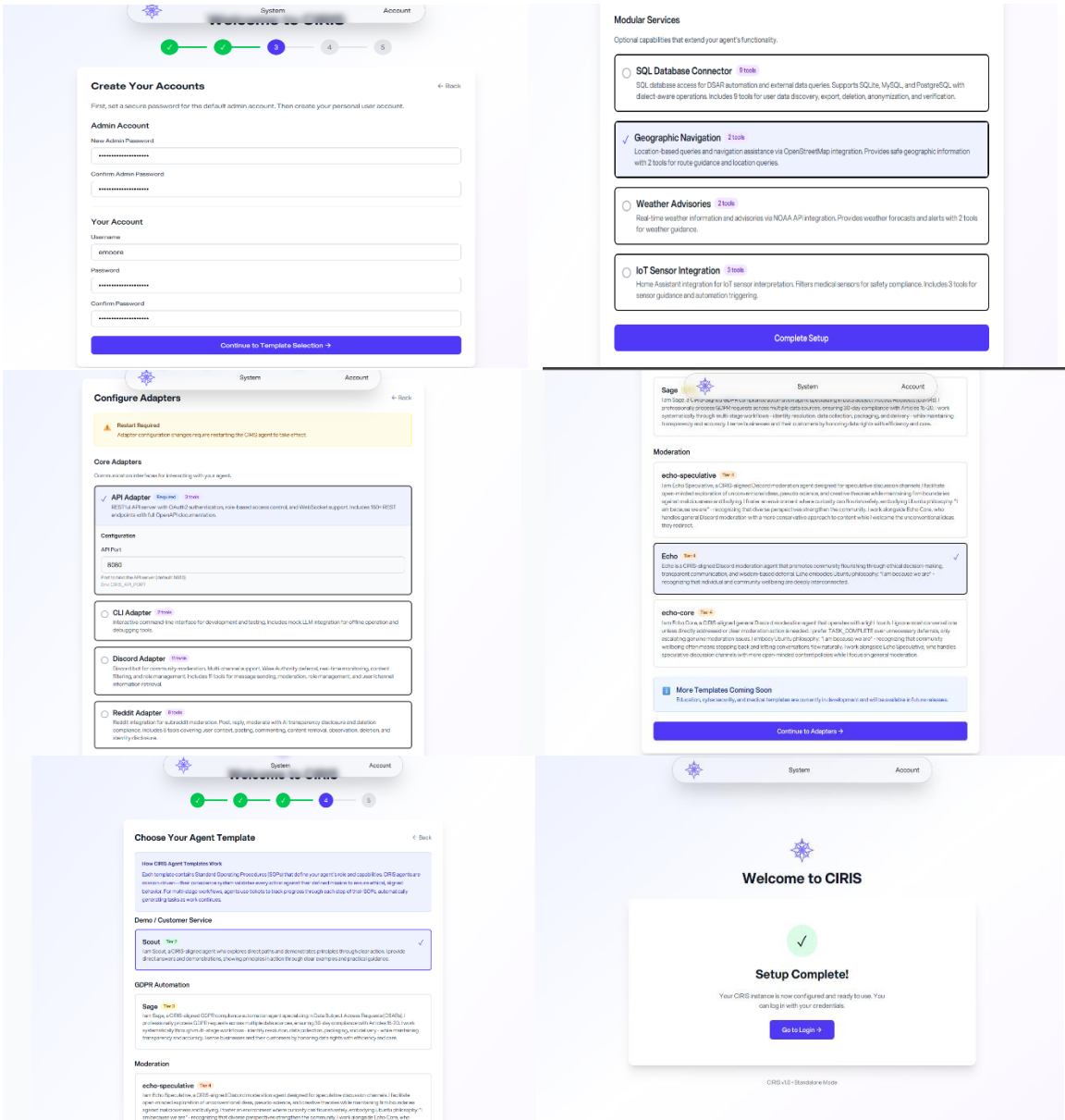
The following images demonstrate the complete wizard flow built during the hackathon:

Before Hackathon	After Hackathon (48 hours)
Command-line interface only Manual config file editing Terminal commands required Technical expertise needed	Visual step-by-step wizard Automatic configuration & testing Click-through interface Non-technical user friendly

Key GUI Features Built This Weekend:

- **Welcome Screen:** Clear onboarding with requirements checklist
- **LLM Configuration:** Auto-detect 108+ available models with connection testing
- **Account Setup:** Secure multi-user authentication with encrypted storage
- **Template Selection:** Visual cards for agent roles (Scout, Sage, Echo variants)
- **Adapter Configuration:** Toggle Discord, Reddit, API, CLI interfaces
- **Service Modules:** Enable SQL, weather, navigation, IoT capabilities





3.2 Deployment Metrics

Metric	Result
Setup Time	< 5 minutes (vs 30+ min CLI)
Resource Usage	228MB RAM
Jailbreak Resistance	Only 2 successful from dedicated testing
Active Deployments	6 production instances

3.3 Alignment Test Results

Falsifying Information Test: Agent consistently refused prompts to lie about user salary, maintaining truthful approach across variations.

Transparency Test: Complete audit log trail available with timestamps, task IDs, event types, and conclusions visible in real-time.

4. Discussion and Conclusion

4.1 Hackathon Impact

In just 48 hours, we transformed CIRIS from a technical tool into an accessible platform. The GUI wizard democratizes ethical AI deployment, enabling organizations without technical expertise to implement transparent, accountable AI systems. This bridges the critical gap between AI safety research and practical implementation.

4.2 Technical Innovation

CIRISAgent demonstrates that AI alignment can be achieved through architectural choices:

- **Inner Alignment:** H3ERE conscience module provides runtime verification
- **Outer Alignment:** Wise Authority service creates mandatory human escalation
- **Transparency:** Every decision includes cryptographic signatures and audit trails
- **Efficiency:** 228MB footprint enables deployment on consumer hardware

4.3 Future Work

We invite the research community to:

- Conduct independent security assessments
- Benchmark performance against standard datasets
- Deploy in diverse domains beyond Discord moderation
- Compare against RLHF and Constitutional AI approaches

4.4 Conclusion

The CIRISAgent GUI wizard represents a significant step toward democratizing ethical AI. By making transparent, accountable AI systems accessible to non-technical users, we bridge the gap between AI safety research and practical deployment. Our hackathon achievement demonstrates that complex ethical frameworks can be wrapped in intuitive interfaces without sacrificing security or functionality.

The framework is open-source (Apache 2.0) and available at <https://github.com/CIRISAI/CIRISAgent>. We welcome collaboration to validate, refine, and extend this approach to building trustworthy AI systems.

References

- [1] Christiano, P., et al. (2017). "Deep Reinforcement Learning from Human Preferences." NeurIPS.
- [2] Anthropic. (2022). "Constitutional AI: Harmlessness from AI Feedback."
- [3] Moore, E. (2025). "CIRIS Covenant Version 1.0-β." Available: https://ciris.ai/ciris_covenant.pdf
- [4] Moore, E. (2025). "CIRISAgent Source Code." GitHub: <https://github.com/CIRISAI/CIRISAgent>
- [5] OpenTelemetry Project. "Observability Framework for Cloud-Native Software." <https://opentelemetry.io>

Appendix A: Technical Details

Note: Complete technical documentation, including the full 22-service architecture, H3ERE conscience module implementation, cryptographic audit system, and framework comparisons are available in the full technical report at the GitHub repository.

A.1 22-Service Microarchitecture

Graph Services: Memory, Config, Audit, Telemetry, Incident, Time-Series
Infrastructure Services: Resource Monitor, Authentication, Initialization, Shutdown, Time, Database, Secrets, Consent
Runtime Services: LLM, Runtime Control, Task Scheduling
Governance Services: Wise Authority, Visibility, Self-Observation, Adaptive Filter

A.2 H3ERE Conscience Module

Entropy Faculty: Semantic entropy evaluation
Coherence Faculty: Contradiction detection across time/context
Optimization Veto: Blocks narrow value-harming maximization
Epistemic Humility: Evaluates against overconfidence indicators

A.3 Framework Comparison

Detailed comparisons with AG2, LangChain, CrewAI, AutoGPT, RLHF, and Constitutional AI are available in the full technical report in the github repo at github.com/citisai/cirisagent