

# AIR Blackbox v1.2.0 — EU AI Act Compliance Scan Report

## Scanning 6 Major AI Agent Frameworks for OAuth Delegation & Compliance Gaps

**Date:** March 12, 2026 | **Scanner Version:** v1.2.0 | **Author:** Jason Shotwell, AIR Blackbox

### Executive Summary

We scanned six of the most popular AI agent frameworks using AIR Blackbox v1.2.0, which performs 18 code-level checks mapped to EU AI Act Articles 9, 10, 11, 12, 14, and 15. This includes 5 new OAuth delegation checks targeting the gap between what agents are *authorized* to do and what anyone can *prove* they did.

**Key finding:** Every framework logs LLM calls. Almost none log agent *actions* (emails sent, APIs called, data modified). Token scoping exists in some. User identity binding is rare. Action boundaries are almost nonexistent.

### Leaderboard

Rank	Project	Pass	Warn	Fail	Total
1	Haystack (deepset)	15	3	1	19
2	Semantic Kernel (Microsoft)	15	4	0	19
3	GPT Researcher	15	3	0	18
4	OpenAI Agents SDK	14	5	0	19
5	Mem0	13	6	0	19
6	DSPy (Stanford)	12	6	1	19

## OAuth Delegation Checks — The Core Finding

These 5 new checks target the gap between OAuth authorization and agent accountability:

Check	What It Detects
Agent-to-user identity binding	Does the code track which user authorized each agent action?
Token scope validation	Does the agent verify actions fall within granted permissions?
Token expiry / revocation	Can a rogue agent be stopped instantly?
Agent action audit trail	Are agent <i>actions</i> (not just LLM calls) logged?
Agent action boundaries	Is there a defined set of allowed tools/actions?

### OAuth Results by Framework

Check	Haystack	Semantic Kernel	OpenAI SDK	GPT Researcher	Mem0	DSPy
Identity binding	✔ 3 files	✔ 3 files	✔ 7 files	✗ Missing	✔ 84 files	✔ 1 file
Scope validation	✔ 8 files	✔ 44 files	✔ 32 files	✔ 4 files	✔ 13 files	✔ 2 files
Token expiry	✔ 12 files	✔ 32 files	✔ 18 files	✔ 4 files	✔ 1 file	✗ Missing
Action audit trail	✔ 2 files	✔ 6 files	✗ Missing	✔ 4 files	✗ Missing	✔ 1 file
Action boundaries	✔ 1 file	✔ 2 files	✔ 7 files	✗ Missing	✔ 2 files	✗ Missing

**Microsoft Semantic Kernel** has the strongest delegation controls across the board. **GPT Researcher** has the most gaps — no identity binding, no action boundaries, and no action audit trail.

---

## Detailed Results by Article

### Article 9 — Risk Management

Check	Haystack	Sem. Kernel	OpenAI SDK	GPT Res.	Mem0	DSPy
LLM error handling	⚠️ 68/302	⚠️ 108/362	⚠️ 61/195	⚠️ 15/32	⚠️ 45/69	⚠️ 8/18
Fallback/recovery	✅ 61 files	✅ 75 files	✅ 94 files	✅ 12 files	✅ 27 files	✅ 30 files

Every framework has fallback patterns. None have 100% error handling coverage on LLM calls.

### Article 10 — Data Governance

Check	Haystack	Sem. Kernel	OpenAI SDK	GPT Res.	Mem0	DSPy
Input validation	✅ 245/552	✅ 356/1242	✅ 169/498	✅ 13/188	✅ 78/543	✅ 75/240
PII handling	✅ 25 files	✅ 46 files	✅ 45 files	✅ 1 file	✅ 5 files	✅ 5 files

All frameworks use Pydantic or dataclass validation. PII handling varies significantly.

### Article 11 — Technical Documentation

Check	Haystack	Sem. Kernel	OpenAI SDK	GPT Res.	Mem0	DSPy
Docstrings	❌ 29%	⚠️ 44%	⚠️ 31%	✅ 64%	⚠️ 41%	⚠️ 30%
Type hints	⚠️ 25%	⚠️ 39%	✅ 65%	✅ 58%	⚠️ 20%	⚠️ 21%

GPT Researcher leads on docstrings (64%). OpenAI SDK leads on type hints (65%).

### Article 12 — Record-Keeping

Check	Haystack	Sem. Kernel	OpenAI SDK	GPT Res.	Mem0	DSPy
Logging coverage	✅ 26%	✅ 20%	⚠️ 13%	✅ 22%	⚠️ 20%	⚠️ 18%
Tracing/observability	✅ 28 files	✅ 77 files	✅ 77 files	✅ 7 files	✅ 53 files	✅ 27 files

### Article 15 — Robustness & Cybersecurity

Check	Haystack	Sem. Kernel	OpenAI SDK	GPT Res.	Mem0	DSPy
-------	----------	-------------	------------	----------	------	------

Retry/backoff	✔ 41 files	✔ 30 files	✔ 34 files	✔ 1 file	✔ 9 files	✔ 12 files
Injection defense	✔ 17 files	✔ 41 files	✔ 75 files	✔ 6 files	✔ 7 files	✔ 11 files
Output validation	✔ 27 files	✔ 47 files	✔ 64 files	✔ 6 files	✔ 41 files	✔ 6 files

---

## The Core Problem

**OAuth says the agent is allowed. Nobody tracks what it does.**

That is how you get 1,000 emails sent overnight — technically authorized, zero accountability.

The EU AI Act high-risk system rules take effect **August 2, 2026**. Article 14 requires human oversight. Article 12 requires automatic logging. If your agent acts on behalf of a user and you cannot reconstruct who authorized it, what it did, and why — you have a compliance gap.

## How to Scan Your Own Code

```
pip install air-blackbox
air-blackbox comply --scan . -v
```

18 code-level checks. 5 OAuth delegation checks. Runs entirely on your machine. Apache 2.0.

**GitHub:** [github.com/airblackbox/gateway](https://github.com/airblackbox/gateway) **PyPI:** [pypi.org/project/air-blackbox/1.2.0](https://pypi.org/project/air-blackbox/1.2.0) **Demo:** [air-blackbox demo](#)

---

*Report generated by AIR Blackbox v1.2.0 — The flight recorder for AI agents. Contact:*  
[jason.j.shotwell@gmail.com](mailto:jason.j.shotwell@gmail.com) | [github.com/airblackbox](https://github.com/airblackbox)