Principes d'Auditabilité des Agents

Adservio | Dr Olivier Vitrac

2026-02-03

Contents

1 Principes d'Auditabilité des Agents

Adservio — Doctrine Interne

1.1 Principe Fondamental

> **Un agent sans trace est un agent sans valeur.**

L'auditabilité n'est pas une option, c'est la condition minimale pour qu'un système à base d'agent soit déployable en contexte professionnel.

1.2 Qu'est-ce qu'une Trace ?

Une trace est un enregistrement structuré permettant à un humain de **reconstruire le raisonnement et les actions** de l'agent.

1.2.1 Schéma Minimal

Champ	Description	Exemple
`timestamp`	Horodatage précis	`2026-02-02T14:32:01Z`

Champ	Description	Exemple
`action`	Ce que l'agent a fait	`call_tool`, `generate`, `decide`
`input`	Données d'entrée	Paramètres, contexte
`output`	Résultat produit	Réponse, code, décision
`reasoning`	Justification (si disponible)	"Test échoué → retry avec fix"
`tool`	Outil appelé (si applicable)	`run_tests`, `grep`, `lint`
`status`	Succès/échec/partiel	`success`, `error`, `timeout`

1.2.2 Exemple Concret

```
{
  "timestamp": "2026-02-02T14:32:01Z",
  "action": "call_tool",
  "tool": "run_tests",
  "input": {"path": "tests/unit/", "pattern": "test_*.py"},
  "output": {"passed": 12, "failed": 2, "errors": ["test_auth.py:45"]},
  "status": "partial",
  "reasoning": "2 échecs détectés, analyse des erreurs requise"
}
```

1.3 Ce qu'une Trace N'est PAS

1.3.1 Anti-patterns à éviter

Anti-pattern	Problème	Correction
Verbose mais vide	500 lignes sans information actionnable	Filtrer, structurer, prioriser
Rationalisation post-hoc	"J'ai fait X parce que c'était logique"	Capturer la décision *avant* l'action
Chemins négatifs absents	Seuls les succès sont tracés	Tracer aussi les échecs et abandons
Format non parseable	Texte libre non structuré	JSON, YAML, ou format défini
Timestamps manquants	Impossible de reconstruire la séquence	Toujours horodater

---

1.4 Niveaux d'Auditabilité

Niveau	Description	Usage
1 — Minimal	Entrées/sorties des outils	Debug basique
2 — Standard	+ décisions et branchements	Audit post-mortem
3 — Complet	+ raisonnement intermédiaire	Conformité, régulé

Pour ce workshop : **Niveau 2 minimum**.

1.5 Pourquoi Cela Compte pour Adservio

1.5.1 Contexte Client

- Les clients régulés exigent la traçabilité
- “L’IA a décidé” n’est pas une réponse acceptable
- La responsabilité reste humaine

1.5.2 Contexte Interne

- Reproductibilité des résultats
- Debug des comportements inattendus
- Capitalisation des patterns qui marchent

1.5.3 Contexte Légal

- Auditabilité = preuve de diligence
- Traçabilité = défense en cas de litige
- Documentation = conformité RGPD/AI Act

1.6 Règle d’Or

Un livrable sans trace exploitable est considéré comme non livré.

Cette règle s’applique à tous les parcours du workshop.

Document de référence — À conserver et appliquer au-delà du workshop.