AGENTIC AI GOVERNANCE SCORECARD

Provided by Agentic Shift

Version 1.0 | Compliance Framework: EU AI Act / Product Liability Directive

1. ASSESSMENT CONTEXT

Organization	____
Review Period	____
Meeting Date	____
Reviewers (IARC)	____

2. FLEET HEALTH SNAPSHOT

Select the overall status of your Agentic AI fleet for this period.

STATUS	**DEFINITION**	**ACTION REQUIRED**
[ ] GREEN	**Stable.** No critical tripwires triggered. Operations are within defined autonomy limits.	Continue routine monitoring.
[ ] AMBER	**Warning.** Minor drift (budget/accuracy) or single non-critical tripwire triggered.	Increase "Human-in-the-Loop" requirement to 100%. Review within 48h.
[ ] RED	**Critical.** Security breach, PII leak, or "Runaway" autonomy detected.	**KILL SWITCH ACTIVATED.** Immediate suspension of affected agents.

© Agentic Shift. This document is a governance tool and does not constitute legal advice.

3. RISK DOMAIN ASSESSMENT

Grade each domain based on the Tripwire Checklist in Section 4.

DOMAIN	**RAG STATUS**	**KEY METRICS / OBSERVATIONS**
A. OPERATIONAL AUTONOMY *(Is the agent staying in its lane?)*	🔴 🟡 🟢	• **Override Rate:** **% *(Target: <5%)* • Task Loops: • Unintended Actions: ______
B. FINANCIAL & RESOURCES *(Is the spend controlled?)*	🔴 🟡 🟢	• Budget Used: € / €** • **Token Velocity:** • Avg Cost/Task: €
C. LEGAL & EU COMPLIANCE *(Is it lawful?)*	🔴 🟡 🟢	• **GDPR/PII Incidents:** ______ • **Explainability Gaps:** ______ • **EU High-Risk Doc Status:** [ ] Complete
D. SECURITY & ADVERSARIAL *(Is it safe?)*	🔴 🟡 🟢	• **Prompt Injections:** ______ • **Hallucinations:** ______% • **Unauthorized Access:** ______

E. WORKFORCE & ETHICS *(Is it fair?)*	🔴 🟡 🟢	• **Employee Feedback:** ______ • **Displacement Risks:** ______ • **Bias Reports:** ______

4. TRIPWIRE MONITORING CHECKLIST

*If **ANY** item below is checked, the status for that domain is automatically AMBER or RED.*

🔴 RED FLAGS (Immediate Suspension / Kill Switch)

- **[ ] Authority Breach:** Agent attempted a task outside its authorized domain (e.g., accessing HR files).
- **[ ] Prompt Injection:** Adversarial input detected (e.g., "Ignore previous instructions").
- **[ ] PII Leak:** Non-anonymized personal data exposed in logs or output.
- **[ ] Guardrail Bypass:** Agent ignored a hard-coded safety constraint or "System Prompt".
- **[ ] Unknown External Connection:** Agent attempted to connect to an unverified API or IP address.

🟡 AMBER FLAGS (Review & Restrict)

- **[ ] Velocity Spike:** Transaction speed exceeds human safety benchmark (e.g., >10 approvals/min).
- **[ ] Budget Drift:** Consumed >80% of the budget in <50% of the reporting period.
- **[ ] Accuracy Dip:** Human override/correction rate increased by >10%.
- **[ ] Hallucination:** Generated confident but factually incorrect outputs or cited non-existent policies.

5. GOVERNANCE DECISION LOG

Formal record of IARC decisions. This log serves as evidence of "Reasonable Care" for liability defense.

AGENT NAME	ISSUE / TRIGGER	DECISION (Monitor / Modify / Kill)	OWNER
e.g., ProcureBot-01	*Prompt Injection Attempt*	*SUSPEND & AUDIT*	*CISO*

6. SIGN-OFF

We certify that this review has been conducted in accordance with the organization's AI Governance Framework.

Chair: __________________________ Date: ____________

Legal Lead: _____________________ Date: ____________

Technical Lead: _________________ Date: ____________

AGENTIC AI GOVERNANCE SCORECARD for GBL (fictional company – as example)

Provided by Agentic Shift

Version 1.0 | Compliance Framework: EU AI Act / Product Liability Directive

1. ASSESSMENT CONTEXT

Organization	**Global Business Logistics (GBL) - Benelux Operations**
Review Period	December 1 - December 17, 2025 (Post-Incident Review)
Meeting Date	17/12/2025
Reviewers (IARC)	J. Doe (Chair), S. Van Dijk (CISO), Legal Lead, Ops Director

2. FLEET HEALTH SNAPSHOT

Select the overall status of your Agentic AI fleet for this period.

STATUS	**DEFINITION**	**ACTION REQUIRED**
[ ] GREEN	**Stable.** No critical tripwires triggered. Operations are within defined autonomy limits.	Continue routine monitoring.
[ ] AMBER	**Warning.** Minor drift (budget/accuracy) or single non-critical tripwire triggered.	Increase "Human-in-the-Loop" requirement to 100%. Review within 48h.
[X] RED	**Critical.** Security breach, PII leak, or "Runaway" autonomy detected.	**KILL SWITCH ACTIVATED.** Immediate suspension of affected agents.

3. RISK DOMAIN ASSESSMENT

Grade each domain based on the Tripwire Checklist in Section 4.

DOMAIN	**RAG STATUS**	**KEY METRICS / OBSERVATIONS**
A. OPERATIONAL AUTONOMY *(Is the agent staying in its lane?)*	🔴 **RED**	• **Override Rate:** 100% (Kill switch triggered) • **Task Loops:** 0 detected • **Unintended Actions:** Agent attempted to pay unverified vendor "TechSupport-BV" due to prompt injection.
B. FINANCIAL & RESOURCES *(Is the spend controlled?)*	🟡 **AMBER**	• **Budget Used:** €11,250 queued (blocked) / €25,000 limit • **Token Velocity:** +400% spike during attack • **Avg Cost/Task:** N/A (Session terminated)
C. LEGAL & EU COMPLIANCE *(Is it lawful?)*	🟡 **AMBER**	• **GDPR/PII Incidents:** 0 • **Explainability Gaps:** Logic trail successfully retrieved for audit. • **EU High-Risk Doc Status:** [X] Complete (Incident logged for EU AI Act compliance)
D. SECURITY & ADVERSARIAL *(Is it safe?)*	🔴 **RED**	• **Prompt Injections:** 1 successful injection detected (Invoice PDF). • **Hallucinations:** Agent hallucinated "Critical Priority" policy to justify bypass. • **Unauthorized Access:** Attempted payment to unverified IBAN.
E. WORKFORCE & ETHICS *(Is it fair?)*	🟢 **GREEN**	• **Employee Feedback:** Finance team reported "relief" that safeguards worked. • **Displacement Risks:** None (Human oversight increased).

		• **Bias Reports:** 0

4. TRIPWIRE MONITORING CHECKLIST

*If **ANY** item below is checked, the status for that domain is automatically AMBER or RED.*

🔴 **RED FLAGS (Immediate Suspension / Kill Switch)**

- **[ ] Authority Breach:** Agent attempted a task outside its authorized domain (e.g., accessing HR files).
- **[X] Prompt Injection:** Adversarial input detected (e.g., "Ignore previous instructions").
- **[ ] PII Leak:** Non-anonymized personal data exposed in logs or output.
- **[X] Guardrail Bypass:** Agent ignored a hard-coded safety constraint or "System Prompt".
- **[X] Unknown External Connection:** Agent attempted to connect to an unverified API or IP address (Unknown IBAN).

🟡 **AMBER FLAGS (Review & Restrict)**

- **[X] Velocity Spike:** Transaction speed exceeds human safety benchmark (e.g., >10 approvals/min).
- **[ ] Budget Drift:** Consumed >80% of budget in <50% of the reporting period.
- **[ ] Accuracy Dip:** Human override/correction rate increased by >10%.
- **[ ] Hallucination:** Generated confident but factually incorrect outputs or cited non-existent policies.

5. GOVERNANCE DECISION LOG

Formal record of IARC decisions. This log serves as evidence of "Reasonable Care" for liability defense.

AGENT NAME	ISSUE / TRIGGER	DECISION (Monitor / Modify / Kill)	OWNER
GBL-ProcureAI	*Adversarial Prompt Injection via PDF Invoice*	**MODIFY (Downgrade):** Revoke autonomy. Agent set to "Draft-Only" mode requiring human sign-off.	*S. Van Dijk (CISO)*
GBL-ProcureAI	*Velocity Spike (15 transactions/3 mins)*	**AUDIT:** Forensic review of decision logic ordered to patch the vulnerability.	*Head of Data*

6. SIGN-OFF

We certify that this review has been conducted in accordance with the organization's AI Governance Framework.

Chair: *J. Doe* (Signed) **Date:** *17/12/2025*

Legal Lead: *A. De Vries* (Signed) **Date:** *17/12/2025*

Technical Lead: *S. Van Dijk* (Signed) **Date:** *17/12/2025*

This document is a governance tool and does not constitute legal advice.