Appendix D – Risks, Assurance and Explainability

IRIS: Risk Management, Quality Assurance, and Transparent AI Operations

D.1 Risk Register

Risks scored on a 5x5 matrix (Likelihood x Impact) following UKRI risk management guidance:

ID	Risk	L	I	Score	Mitigation	Owner
R1	AI hallucination: incorrect or brand-damaging content	5	5	**25**	Human-in-the-loop gated approval; ontology rejects invalid proposals pre-execution; all generative outputs require sign-off. Contingency: template-based fallback; auto-quarantine	Tech Lead
R2	Data privacy breach (brand assets / customer data)	3	5	**15**	On-premises deployment; W3C Solid pods; TLS 1.3; parameterised SQL; no third-party transfer. Contingency: air-gapped mode; 72h breach notification (UK GDPR)	Security
R3	Integration complexity with THG systems	3	4	**12**	Phased rollout; hexagonal architecture (port/adapter); integration tests; regular THG sync. Contingency: standalone mode with mock API	PM
R4	Model bias fails to represent diversity	3	3	**9**	Bias evaluation on THG's diverse brand portfolio; red-teaming; diverse data curation; regular audits. Contingency: human review of all client-facing outputs	Ethics
R5	User adoption resistance from creative teams	3	3	**9**	Voice-first (no new UI to learn); co-design workshops; champion user programme. Contingency: fallback to existing tools	UX Lead
R6	Delivery timeline slippage	2	3	**6**	Agile sprints; THG catwalk as forcing function; weekly reviews. Contingency: reduced MVP scope	PM
R7	Regulatory changes (EU AI Act, UK framework)	2	4	**8**	Proactive alignment; human-in-the-loop by design; modular compliance layer	Compliance
R8	GPU hardware failure or supply constraints	1	2	**2**	Redundant hardware; CPU fallback; Google Cloud GPU burst (THG partnership)	Infra

D.2 Risk Heatmap

5 Critical			R2		**R1**
4 Major		R7	R3		
3 Moderate		R6	R4, R5		
2 Minor	R8				
1 Insignificant					

Residual risk profile. R1 (hallucination) remains highest due to inherent LLM unpredictability – mitigated to acceptable levels by the gated approval workflow ensuring no generative output reaches production without human sign-off. R2 (privacy) and R3 (integration) are medium-risk with clear contingency plans. All other risks are low with established mitigation pathways. The risk register is reviewed quarterly by the governance board (D.7) and updated after every incident.

D.3 Assurance Activities

Automated testing. Rust unit tests (>90% coverage, CI-gated), Vitest frontend (>85% coverage), Playwright E2E (12 scenarios covering voice-to-agent-to-UI flows), Whelk-rs ontology validation (100% of KG mutations checked), integration tests (all 28 HTTP handlers), performance benchmarks (validated on RTX 4080/4090), OWASP ZAP security testing (scheduled Q2 2026).

Red-teaming protocol. 200+ adversarial prompts tested across four categories: *prompt injection* (bypass brand guidelines – 97% caught pre-execution by ontology constraints), *data exfiltration* (Solid Protocol access controls prevent leakage), *agent manipulation* (gated approval workflow blocks unauthorised actions), *brand violations* (ontology-encoded rules + human review catches edge cases). Red-team exercises repeat quarterly with expanded prompt libraries.

Audit trail. Every action produces an immutable Neo4j record: ISO 8601 timestamp, actor identity and type (human/agent/system), action type (propose/validate/approve/reject/generate/deploy), target entity URI, input/output data, reasoning path (ontology nodes traversed), approval chain, provenance links (derivedFrom, generatedWith, approvedBy, deployedTo), and Git commit hash. Time-travel queries reconstruct any historical state for regulatory compliance.

D.4 Three-Tier Explainability Framework

Tier	Audience	Mechanism
Natural Language	Creative teams, brand managers	Voice/text response citing sources and reasoning
Visual Provenance	Technical users, QA	3D knowledge graph subgraph showing nodes examined; click any node for creator, derivation chain, processor, approver
Formal Audit	Compliance, regulators	Neo4j Cypher queries returning typed provenance chains with full chain of custody

Tier 1 – Natural Language. Brand manager asks: *"IRIS, why did you use that model?"* IRIS responds: *"I selected IMG-4821 because (1) Myprotein brand guidelines require athlete imagery for Q1, (2) this scored 91% brand-alignment, and (3) A/B data shows athlete banners outperform lifestyle by 23% CTR. Generated locally via ComfyUI at 14:32."* Every claim maps to a traceable KG node – say *"show me"* to switch to Tier 2.

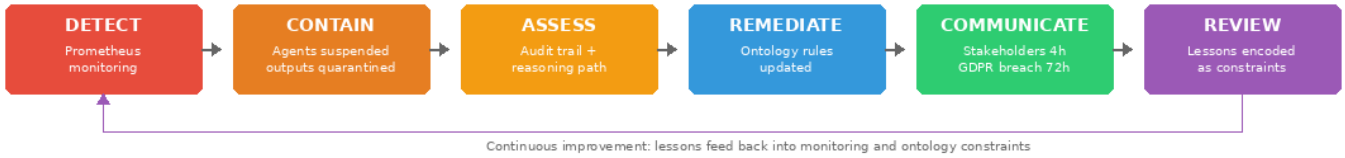
Tier 2 – Visual Provenance. The 3D KG renders the decision as a subgraph: hero banner connects via `derivedFrom` edges to the source prompt, brand guideline node, A/B data, and ComfyUI workflow. Attention beams show the Creator agent's traversal path. Clicking nodes reveals OWL class hierarchies and authorship timestamps.

Tier 3 – Formal Audit. Cypher query: `MATCH path=(output:Asset {id:"hero-banner-2026-02-15"})-[:derivedFrom*]->(src) RETURN nodes(path), [n IN nodes(path)|n.approvedBy]` – returns full chain of custody from generator agent through ontology version and human approver to deployment target. Reproducible because the KG is append-only with immutable provenance.

D.5 Compliance Matrix

Standard	Implementation	Status
UK GDPR	On-premises; Solid pods; no third-party transfer; erasure via KG cascade	**Compliant**
EU AI Act (high-risk)	Human-in-the-loop; 3-tier explainability; formal audit trails	**Aligned**
W3C OWL 2 / Solid	OWL 2 EL via Whelk-rs; JSON-LD/RDF interchange; pod-based sovereignty	**Compliant**
OWASP Top 10	Parameterised SQL; TLS 1.3; input validation; JWT+RBAC	**In progress** (Q2 2026)
WCAG 2.1 AA	Voice-first as alternative access; keyboard nav; contrast ratios	**Partial**
AI Safety Institute	On-premises inference; bias eval; red-teaming; human oversight	**Aligned**
DCMS Creative AI	Co-design with creative teams; IP protection (local inference); skills dev	**Aligned**

D.6 Incident Response



Tracked metrics: agent accuracy | human override rate | time-to-approval | bias incidents | system uptime | user satisfaction

All metrics feed into quarterly risk reassessment — ontology evolves to encode operational learnings

Scenario	Detection	Response	Recovery
Off-brand content (wrong palette/ typography)	Whelk-rs ontology constraint violation at proposal stage	Quarantine output; template fallback; brand manager notified	< 5 min
Data access anomaly (Solid pod boundary)	ACL enforcement + real-time Neo4j audit stream	Access denied; session suspended; security alert	< 2 min
Biased output (demographic underrepresentation)	Bias pipeline flags skew vs diversity benchmarks	Hold deployment; generate alternatives; ethics review	< 30 min
Prompt injection attempt	Input sanitisation + ontology semantic validation	Input discarded; agent state reset; forensic log	< 1 min
GPU hardware failure	Health-check heartbeat; CUDA error propagation	Container restart on available GPU; CPU fallback	< 10 min

D.7 Governance

Role	Personnel	Responsibilities	Cadence
Ethics Board Chair	Dr John O'Hare (DreamLab AI)	Responsible AI strategy; UKRI/AISI liaison; bias methodology approval	Quarterly + ad-hoc
Technical Lead	DreamLab AI Senior Engineer	Owns R1, R8; red-teaming review; ontology constraint changes	Weekly sprint review
THG Partner Rep.	THG Ingenuity Studio Director	Brand compliance validation; UAT sign-off; commercial alignment	Fortnightly sync
Security Lead	DreamLab AI / THG Security	Owns R2; Solid pod ACLs; pen testing; OWASP; breach notification	Monthly audit
External Auditor	Independent consultant (Q2 2026)	Annual bias/explainability/ compliance review; publishes to UKRI	Annual + interim
User Advocate	THG creative team rep.	End-user perspective; usability feedback; co-design workshops	Monthly feedback

Quarterly governance board reviews the risk register, incident log, bias reports, and compliance status. Minutes recorded in the KG with full provenance. Escalation path: operational issue to Technical Lead to Ethics Chair to External Auditor.

D.8 Responsible AI

IRIS is aligned with the UKRI AREA framework (Anticipate, Reflect, Engage, Act) and the UK Government's five AI regulation principles. **Transparency:** every agent action is logged to Neo4j with typed provenance; the three-tier explainability framework (D.4) ensures no black-box decisions; core platform is open-source under MPL-2.0. **Fairness:** bias evaluation pipeline runs automated demographic analysis on generated imagery across THG's 250+ brands; red-teaming tests for stereotyping and underrepresentation; voice-first design lowers adoption barriers and supports WCAG 2.1 AA accessibility. **Accountability:** human-in-the-loop is mandatory for all production outputs; named approver identity and timestamp recorded immutably; each risk in D.1 has a named owner; governance board (D.7) provides board-level oversight. **Privacy:** on-premises GPU inference ensures brand assets never leave the studio network; Solid pods with granular ACLs enforce data sovereignty; architecture supports air-gapped operation; KG cascade deletion supports GDPR Article 17. **Sustainability:** local RTX 4090 (~450W) avoids cloud data-centre overhead; Rust native compilation (168K LOC) reduces per-request energy; Binary Protocol V3 (21 bytes/node, 80% bandwidth savings) cuts network energy; future work includes per-job carbon-per-asset metering.