Fraudfinder LTD
71-75 Shelton St
London
WC2H 9JQ
United Kingdom

Email: alexander@fraudfinderai.com

Fraudfinder AI Policy

1. Purpose and Commitment

This policy outlines Fraudfinder's principles and operational standards for the responsible development, deployment, and use of Artificial Intelligence (AI), including GenAI, across our product suite. Our mission is to replace manual fraud review with **accurate, fast, and secure AI** systems that enhance human decision-making while preventing abuse.

---

2. Scope

Applies to all Fraudfinder AI systems, models, and tools used in document fraud detection (e.g., bank statements, payslips, utility bills), including white-labeled deployments and integrations.

---

3. Principles

1. Accuracy & Reliability

- We aim for >99% fraud detection accuracy, measured continuously with real-world and synthetic test datasets.
- All model performance is monitored against **True Positive Rate (TPR)** and **True Negative Rate (TNR)** KPIs.

2. Transparency

- We do not use "black-box" AI. All models provide a **decision rationale** (e.g., fraud indicators per document section).
- Audit logs track decisions at file level, including metadata, confidence scores, and versioned model.

3. Human Oversight

- AI outputs are presented as **recommendations, not determinations**. Clients retain full control and can override or escalate any flagged files via their internal review processes.
- We maintain an **escalation protocol** for edge cases, anomalies, and low-confidence predictions.
- Policy-based guardrails automatically detect and block disallowed content or actions (e.g., privacy violations, harassment, or other policy non-compliance) before any AI output reaches human reviewers.

4. Privacy & Data Protection

- All documents are processed in compliance with **GDPR, the UK Data Protection Act 2018, and relevant AML regulations**.
- Personally identifiable information (PII) is **encrypted in transit and at rest**, and retained only as necessary to deliver services.
- **By default, client data is used in pseudonymised form to improve model performance** (e.g., through continuous learning and evaluation).
- Clients who wish to **opt out of contributing their data to model training** or require **full data deletion** may do so via our **Enterprise tier**.
- Clients can request full **audit logs** at any time.

5. Anti-Abuse & GenAI Misuse Detection

- We actively detect GenAI-generated or tampered documents, using watermark checks, pixel-level anomaly detection, and NLP-based context validation.
- We **do not use GenAI to generate, manipulate, or simulate documents**.
- Internal use of GenAI tools (e.g., for content, code, ops) is governed by a **restricted access policy** and reviewed quarterly.

6. Bias & Fairness

- All models are evaluated to minimise **disparate impact**, particularly for protected characteristics (e.g., names, addresses).
- We avoid using demographic or sensitive variables in training or inference.

---

4. Model Governance

Area	Practice
Model Development	Peer-reviewed codebase; reproducible pipelines
Model Training	Performed on controlled, labelled datasets; no personal data used
Model Versioning	All model versions are stored, tagged, and backward-compatible
Testing & Validation	Conducted on separate **test and validation** datasets
Incident Response	Any performance degradation triggers review within 24 hours

Third-party Models	Must pass internal review; no uncontrolled APIs allowed

---

5. Internal Use of AI Tools

Employees may use AI tools (e.g., ChatGPT, Copilot) under the following conditions:

- No client data or confidential IP is input.
- Output is reviewed before use in any production-facing content.
- All usage must comply with our **Acceptable Use and Data Security Policies**.

---

6. Client Transparency

- Clients are informed that AI is used during onboarding and contractually.
- Clients can request an explanation for any decision made by the AI.
- White-label clients are responsible for end-user disclosures, though Fraudfinder provides templated guidance.

---

7. Review & Updates

This policy is reviewed every 12 months or after any major product release involving AI.

Approved by:
Alexander Siedes, Chief Executive Officer

Effective Date: 06 July, 2025
Next Review Date: 05 July, 2026