Executive Briefing

Status: Januar 2026 | **Version:** 1.4 (Final Edit)

Thema: Revisionssichere Stabilitätsmessung & Human-AI-Coupling

---

RISK MANAGEMENT & GOVERNANCE

1. Das Problem: Die „Stille Instabilität“ von LLMs
Large Language Models (LLMs) sind statistisch offene Systeme. Ohne systemische Rückkopplung neigen sie zu „Drift“ (Halluzinationen, logischer Kollaps). Für regulierte Institute entsteht hier ein unkalkulierbares Haftungsrisiko, da herkömmliche Prüfverfahren nicht in Echtzeit skalieren.

2. Die Lösung: Der Waiva-Score (sReact™)
Waiva bietet eine kybernetische Sicherheitsschicht, die LLM-Outputs auf ihre strukturelle Stabilität prüft.

- **Transparenz:** Ein Score (0–90) macht Drift sofort sichtbar.
- **Accountability:** Ein systemischer Hardcap bei 90% sichert die menschliche Letztentscheidung.
- **Compliance:** Erfüllung der Überwachungspflichten nach **Art. 29 EU AI Act**.

3. Expertise: BaFin-erprobt & Standard-bewährt
Entwickelt von einem Experten für Banksteuerung (23 Jahre Banking, 10 Jahre Risk-Control):

- **Prüfungserfahrung:** Begleitung von zwei **§44 KWG Sonderprüfungen** der BaFin.
- **Benchmark-Referenz:** Entwickler eines Risikomessverfahrens für den **SVBW**, das von der Prüfungsstelle offiziell als **„Maßstab“** für alle Institute deklariert wurde.

SCIENCE, SECURITY & HUMAN OVERSIGHT

1. Wissenschaft: Das +1 Principle™
Waiva transformiert Ethik von einer moralischen Option in eine **mathematische Stabilitätsbedingung** (Negative Feedback).

- **Internationale Resonanz:** Fachlicher Austausch mit der **ETH Zürich**, **Cornell University** und dem **Mila Institute**.
- **Validierung:** Prof. James A. Yorke (Chaos-Theorie): „A valuable insight."

2. Active Governance: Human-AI-Coupling
Waiva befähigt den Nutzer zur aktiven Aufsicht (Art. 14 EU AI Act). Bei drohendem Drift (Score < 80) interveniert das System pädagogisch:

- **Expert-Nudging:** Interaktive „Drift-Ringe" erklären das Systemverhalten.
- **Korrektur-Impuls:** Konkrete Handlungsanweisungen via Tooltip: > *„+1 Tipp – Wenn die KI abdriftet, hilft ein gezielter Korrekturimpuls. Tippe hier für konkrete Vorschläge."*
- **Ergebnis:** Der Mitarbeiter wird zum qualifizierten Risiko-Manager in der KI-Interaktion.

3. Sicherheit: Privacy by Design

- **Hosting in Berlin (IONOS):** Volle digitale Souveränität, deutsches Recht.
- **Zero-Logging:** Transiente Analyse im RAM; keine dauerhafte Speicherung von Inhalten.
- **Doppelter Scrubber:** Mehrstufige Inhaltsreinigung zum Schutz von PII-Daten.

4. Das Waiva-Zertifikat (ab Plan L)
Bestätigt offiziell die Implementierung technischer Kontrollinstanzen gegenüber Wirtschaftsprüfern und der Aufsicht.

„Waiva weist nach, dass Verantwortung organisiert wurde – nicht, wie sich einzelne Personen verhalten haben."