Overmind Policy Gate & Trust Runtime

Security & Legal One-Pager (CISO/Legal)

1. Executive Summary

Overmind ermöglicht **Voice-gesteuerte Tool-Aktionen** (z. B. Ticketing, CRM, ERP, Knowledge) – **kontrolliert, regelbasiert und auditierbar**.

Kernmechanismus ist das **Policy Gate**: Es entscheidet pro Aktion anhand von **Rolle, Kontext, Risikostufe und Datenklasse**, ob die Aktion **automatisch**, **nur nach Bestätigung** oder **nur nach Freigabe (4-Augen)** ausgeführt wird. Jede Entscheidung erzeugt **Evidence** für Audit und Nachvollziehbarkeit.

2. Scope

In Scope

- Conversational Orchestration (Voice/Chat → Intent → Tool-Call)
- Policy Gate (RBAC/ABAC, Risk Matrix, Data Controls)
- Evidence/Audit Trail (tamper-aware, nachvollziehbar)
- Connector Layer (Tool-Scopes, Minimal Payload, Guardrails)

Out of Scope (typisch, je nach Projekt definieren)

- Vollständige Tool-Rechteverwaltung im Zielsystem (bleibt beim Kunden/IdP)
- DLP/SIEM-Betrieb (Integration wird unterstützt, Betrieb kundenseitig)

3. Control Objectives (CISO/Legal)

1. **Least Privilege & Zweckbindung** (nur notwendige Rechte & Daten)
2. **Deterministische Durchsetzung** (fail-closed, testbar, versioniert)
3. **Risk-Based Controls** (Grün/Gelb/Rot: Auto/Confirm/Approve)
4. **Data Protection by Design** (PII-Minimierung, Redaction, Retention)
5. **Auditability** (wer/was/warum/wann/mit welcher Freigabe)

4. Policy Gate – Wie Entscheidungen fallen

Inputs

- **Actor**: Identität, Rolle, Org/Team, Auth-Level (SSO/MFA), Delegation
- **Action**: normalisierte Aktion (z. B. ticket.close, erp.order.cancel)
- **Context**: Kanal, Workflow-State, Zeitfenster, Sicherheits-Signale
- **Data Class**: PII / Confidential / etc., Redaction-Status
- **Tool**: Zielsystem, Endpunkt-Scope, Effektklasse (read/write/external/destructive/financial)

Outputs (Entscheidungstypen)

- **ALLOW** (Auto-Execute)
- **REQUIRE_CONFIRMATION** (User-Bestätigung; Voice: explizite Zustimmung)
- **REQUIRE_APPROVAL** (4-Augen / Approver-Rolle)
- **NEED_MORE_CONTEXT** (Rückfragen erzwingen)
- **DENY** (Block + Policy-Reason)

Fail-Closed Prinzip

- Unbekannte Aktion / fehlender Scope / Sicherheits-Signal → **DENY** oder **NEED_MORE_CONTEXT** (keine Ausführung „auf Verdacht“)

5. Risikomatrix (Action × DataClass × Context → Control)

Ampel-Modell

- **Grün (Low)**: Lesen, Draft erstellen, nicht-kritische interne Aktionen → **Auto**
- **Gelb (Medium)**: Updates/Changes → **Confirm** (Parameter werden wiederholt)
- **Rot (High)**: irreversibel / extern / finanziell / Löschung → **Approve**

Beispielregeln (typisch)

- *External + PII* → **Confirm (explicit)** + **Redaction before log**
- *Order ≥ Schwellenwert* → **Approve** (Approver Role, strict evidence)
- *Publish Knowledge* → **Approve** (Knowledge Steward)
- *Delete Subject (DSAR-lite)* → **Approve** (Privacy Officer)

6. Data Protection & Privacy (GDPR-ready Controls)

Datenminimierung

- Minimal-Payload Templates pro Connector (nur notwendige Felder)
- Trennung: **LLM-Kontext** vs. **Tool-Payload** (PII bevorzugt im Connector, nicht im Prompt)

PII Handling

- PII Detection / Redaction vor Logs/Evidence
- Konfigurierbare Datenklassen je Use Case (PII/Confidential/Restricted)

Retention & DSAR-lite

- Konfigurierbare Aufbewahrung für Transkripte/Audio/Evidence
- Löschpfade (Subject-Delete) über definierte Prozesse/Rollen

7. Evidence & Audit Trail (Nachweisbarkeit)

Für jede relevante Aktion werden auditierbar erfasst:

- Policy-Version / Regel-ID (z. B. aus Policy-Repo)
- Canonical Request (redacted) + Decision + Reason
- Confirmation/Approval Event (wer, wann, wodurch)
- Tool-Execution Result (IDs, Status, Errors)
- Correlation IDs (End-to-End Trace)

Integrationspunkte

- Export/Forward in SIEM/Log-Pipeline möglich (kundenspezifisch)

8. Threat & Abuse Controls (Security)

- **Prompt-Injection/Tool-Abuse Guard**: Sicherheits-Signale führen zu deny/step-up
- **Rate Limits & Anomaly Signals**: Missbrauchserkennung (z. B. ungewöhnliche Häufung)
- **Defense in Depth**: Policy Gate + Connector Validations + Target-System Controls

9. Governance & Change Management

- **Policy-as-Code**: Versioniert, peer-reviewed, CI-Tests, staged rollout
- **Testbarkeit**: Unit-Tests pro kritischer Regel (e.g. „order_cancel requires approval“)
- **Rollenmodell (RACI)**
  - Policy Owner (Security/IT)
  - Approver Roles (Fachbereich/Legal je nach Aktion)
  - Evidence/Audit Owner (Compliance/Security)

10. What You Get (Audit-Ready Artefacts)

- Policy Catalogue (Actions, Controls, Thresholds)
- Risk Matrix (Ampel, Ausnahmen, Schwellenwerte)
- Evidence Schema & Sample Records (redacted)
- Connector Scope Matrix (endpoints, minimal payload)
- Test Suite (OPA/Rego) + Change Log

11. Offene Punkte zur finalen Abstimmung (Workshop-Checkliste)

- Datenklassen & „PII to LLM?“-Entscheidung je Use Case
- Schwellenwerte (financial, destructive), Confirmation-Text/UX
- Retention-Zeiten & DSAR-lite Prozess
- SIEM/Logging-Integration & Audit-Export
- Approver-Rollen und Eskalationspfade

Kontakt / Next Step

Kurz-Workshop (60–90 Min.) mit CISO/Legal/IT: *Scope & Risk Matrix festzurren, Approver-Rollen definieren, Evidence-Schema abnehmen.*

web@overmind.one