Model Transparency Specification – SentinelX

# 1. Syfte

Detta dokument beskriver transparens, syfte och begränsningar för AI-modellerna som används inom SentinelX-plattformen. Specifikationen uppfyller kraven enligt Artikel 13–15 i EU:s AI-förordning.

# 2. Modellöversikt

- GroomingDetektor v0.9 (TFLite, mobilmodell)  
- VoicePatternFilter v1.2 (lokalt CLI-filter mot röstbedrägeri)  
- DNSPatternGuard v0.7 (Python-baserad DNS-anomalidetektor)  
- TextEvalCore v1.0 (svensk NLP, promptstyrd granskning)

# 3. Träningsdata

- Träningsdata består av syntetiskt genererat material och annoterade chattar från offentliga datakällor  
- Inga personuppgifter har använts under träning  
- DNS-detektion tränad på publikt phishing-dataset från 2023  
- Modellvalidering sker lokalt och manuellt

# 4. Begränsningar

- AI-modeller har svårigheter med slang, ironi och kontextberoende  
- Deepfake-detektion sker ej i realtid (endast varningsnivåer)  
- VoicePatternFilter fungerar bäst på svenskt eller engelskt tal  
- NLP-detektion är inte juridiskt bevis, endast varningssignal

# 5. Användarfeedback och förbättring

- Alla varningar loggas med möjlighet till manuell etikett av användare  
- Användarens återkoppling (rätt/fel) kan frivilligt bidra till förbättring av modell  
- Ingen data delas utan uttryckligt samtycke från användaren

# 6. Dokumentationstillgång

- Beskrivningar av modellernas logik, triggers och exempel finns tillgängliga i användarmanualen  
- CLI-kommandon dokumenteras i SentinelControlKit README  
- Guardian-agentens AI-modul dokumenteras i APK-guiden