



PROJECT PRINCE AI

The P.R.I.N.C.E. Protocol™

An Ethical AI Framework for Human Protection

Framework Brief · Version 3

“While the world builds AI, we build protection.”

Michael Prince · Founder, Project PRINCE AI
michael.prince@projectprince.ai · projectprince.ai

Executive Summary

The internet's most dangerous failures are no longer technical. They are human: scams engineered to exploit trust, deepfakes engineered to exploit belief, and manipulation engineered to exploit emotion. The people harmed most are those technology has protected least — elders, children, non-native speakers, and people with disabilities.

The P.R.I.N.C.E. Protocol (Protection, Respect, Integrity, Neutrality, Clarity, Empowerment) is a governance framework for AI systems whose duty is human protection. It defines six binding commitments an AI must uphold when it touches a person's safety, privacy, or understanding — and it is implemented, in production, by Project PRINCE AI: a network of named guardian agents that detect deception in real time, show their evidence, speak in any language, and leave the final decision with the human.

This brief describes the protocol, its reference implementation, and the path for other systems to adopt it. The protocol is offered openly: protection should be a standard, not a product moat.

The Problem: Deception at Machine Scale

Generative AI has industrialized deception. A convincing scam once required skill; it now requires a prompt. Reported consumer fraud losses run to billions of dollars annually, and the true figure is far higher because shame silences victims. Meanwhile, the defensive tooling most people rely on was designed for viruses and spam — not for a persuasive message that arrives in grandma's own language, addressed to her by name.

Existing protection fails vulnerable users in three specific ways:

- **It is opaque.** Verdicts arrive without reasons, so users cannot learn, and cannot calibrate trust.
- **It is exclusionary.** Warnings assume English fluency, sight, hearing, and technical comfort.
- **It is paternalistic or absent.** Tools either seize control or provide none, leaving no room for human judgment supported by machine evidence.

The P.R.I.N.C.E. Protocol: Six Commitments

Any AI system adopting the protocol commits to all six pillars. They are stated here in normative language — *must*, not *should* — because in protective AI, aspiration is not a standard.

P — Protection

The system's first duty is to the person in front of it — never to engagement, data collection, or upsell. Protective action must be proactive: warn before harm, in language the person can act on, through channels the person can perceive.

R — Respect

Privacy by default. Protective data must never be used for advertising or sold. Processing is transient wherever possible; retention requires a protective justification. Consent must be explained in plain words. The person chooses what leaves their device.

I — Integrity

Every verdict must show its evidence: what was checked, what was found, and how the conclusion follows — in human-readable form. If the system cannot support a claim, it must say so rather than guess. Fabricated confidence is a protocol violation.

N — Neutrality

The system detects manipulation patterns and coordinated deception — never ideology. It must not decide what people should believe; it shows how content attempts to influence, and leaves belief to the human.

C — Clarity

Safety must survive translation. Alerts, evidence, and next steps must be available in the person's own language and channel: read, heard, or physically felt. Accessibility is a floor, not a feature tier.

E — Empowerment

The human makes the final call. The system educates, explains, and offers recovery paths; it does not seize control. Every alert should leave the person more capable than before it fired.

Reference Implementation: Project PRINCE AI

Project PRINCE AI implements the protocol across a browser extension, a web platform, and a native mobile application, sharing one investigation engine. Key architectural expressions of the pillars:

The Guardian Network

Investigations are conducted by named guardian agents — Mama, Scout, Judge, Echo, Patch, Shield, Nana, Wallah, Beacon, Clarity, Proto, and Lingua — each with a defined specialty (scam pressure, link triage, evidence review, manipulation patterns, device safety, identity protection, plain-language explanation, cultural context, recovery resources, wellbeing, systems, and translation). Named agents are not branding: accountability requires an identifiable speaker, and trust for vulnerable users is built through relationship, not through a logo.

Solomon, Guardian of Memory

Where the other guardians detect deception, Solomon protects a person against a slower loss: the erosion of their own memory and identity. He is a companion, not a scanner — holding who a person is, the shape of their day, and the people they trust, and reading it back in a patient voice when it is needed. His Trusted Circle lets a person decide, in plain words, exactly what each family member may see — there is no master key, and the person can always see who looked. Because he safeguards the most intimate data in the system, Solomon's memories are never used to train anything, never gated behind payment, and never deleted on a lapse. He is the protocol's answer to the question the rest of the network cannot: who protects the person as they change?

Evidence-First Verdicts (Integrity)

Every scan returns the guardian's conclusion together with its supporting evidence: signals found, the evidence trail, and the sources checked. A “one mouth” rule ensures a single accountable speaker per verdict. Degraded results are labeled as degraded — the system says “I could not complete this” rather than inventing certainty.

The Accessibility Floor (Clarity)

Verdicts are delivered in parallel channels: written text, translated text in sixteen languages (free — a person must never be unable to read why something is dangerous), synthesized guardian voice, and severity-shaped haptic feedback so a warning can be physically felt by deaf-blind users or anyone whose eyes are elsewhere.

Reference Implementation, continued

Privacy Posture (Respect)

Scanned content is processed transiently for the investigation. The public community feed publishes only sanitized records — root domain, category, severity, and a generic summary — never raw URLs, message content, or personal information. Sharing to the community feed is opt-in.

Human Hand-off (Empowerment)

A person can send any verdict — evidence included — to someone they trust for a human second opinion. This is user-initiated only: no silent monitoring of a vulnerable adult is possible, by design. Protection must never become surveillance.

The Trusted Circle (Respect & Empowerment)

Solomon's memory sharing is governed by a per-person, per-category permission model with no master key: a daughter may see appointments without seeing private journals; a caregiver may help with a routine without reading the life story. Every access is logged and visible to the person themselves — the one being cared for can always see who looked. It is the protocol's Respect and Empowerment pillars applied to the most sensitive data a person owns: their own memory.

Adopting the Protocol

The protocol is offered as a governance layer for any system, company, or agency whose AI touches human safety. Adoption is a posture, verified by practice:

Pillar	Adoption test (answer must be yes)
Protection	Does the system act for the user's safety even when it costs engagement or revenue?
Respect	Is protective data excluded from advertising and sale, with transient processing by default?
Integrity	Can every automated judgment display its evidence in language a layperson understands?
Neutrality	Does detection target deception techniques rather than viewpoints?
Clarity	Are warnings available across languages and sensory channels at no extra cost?
Empowerment	Does the human retain the final decision, with education and recovery paths offered?

Adopters may implement the protocol with their own architecture; the commitments, not the code, are the standard. Project PRINCE AI offers its implementation experience — guardian design, evidence schemas, accessibility patterns, and consent language — to organizations pursuing adoption.

About the Founder

Project PRINCE AI was founded by Michael Prince and began as a personal promise: to protect his mother, living with early-stage dementia, from an internet increasingly engineered to deceive her. That same promise is why Solomon, the Guardian of Memory, now stands beside the deception guardians — because protecting a person means protecting them as they change, not only from the scam in front of them today. It grew into a conviction that everyone — elders, children, the deaf, the blind, speakers of every language — deserves a guardian who tells the truth and shows the evidence.

“Build the guardians people deserve. Tools that warn early, explain clearly, and help you recover when the worst happens.”

Contact & Partnership

michael.prince@projectprince.ai · projectprince.ai · The P.R.I.N.C.E. Protocol™ and Project PRINCE AI™ are trademarks of Project PRINCE AI. This brief may be shared freely in unmodified form.