UX DESIGN FOR TRUSTWORTHY AI: FIVE PRINCIPLES AND PRODUCT PATTERNS FOR TEAMS

***Gnatovych V.**, Senior Product Designer, Reddit*

Abstract. *This thesis explores UX design strategies for fostering trust in AI systems through real-world applications across healthcare, finance, and content moderation. It introduces a framework of five trust-enabling principles derived from cross-sector product implementations and academic literature (2018–2025). Designed from a UX and product lens, the framework helps teams operationalize explainability, user control, and system transparency in AI-powered experiences.*

The rise of artificial intelligence (AI) has redefined how users interact with digital products, from recommendation systems to moderation tools to clinical decision support. As designers and product teams increasingly build on AI-powered infrastructure, fostering user trust has become a critical responsibility. UX and product design must not only make AI intelligible and usable, but also visibly accountable and ethically aligned.

Designers must translate technical AI logic into user-facing explanations. Model Cards [1] offer one format, structured documentation describing a model's purpose, dataset, and performance. Interactive model cards [2] enable designers to embed modular explanations in interfaces tailored to user roles. In clinical tools, these have evolved into layered information systems: simple visuals for patients, data rationale for doctors, and audit logs for regulators, all within the same flow. In Google Health's AI-powered diagnostic tool, a model card appears as a side panel next to the output with expandable sections labeled "For Patients" and "For Clinicians», each showing tailored explanations and visual diagrams.

From onboarding moments to edge cases, users need to know what the AI is doing. Content moderation systems, for instance, now include "why was this flagged?" toggles and options to dispute decisions. Research [4] finds that visibility features, such as live updates, contextual tips, and system feedback, foster long-term trust, especially when they evolve with repeated interactions. Meta's content moderation UI shows a status history for each action (flagged, reviewed, reversed), with links to community guidelines and appeal forms.

Opt-in checkboxes are no longer sufficient in adaptive AI products. Product teams now use dynamic consent frameworks that let users adjust preferences on the fly. In financial or health applications, these systems include real-time permission editing, tiered data sharing, and blockchain-verified audit logs. Revolut's data sharing dashboard uses toggle-based controls, real-time tracking, and export logs so users can review their consent history.

Emerging interfaces like X-Selector [3] predict when users might want additional context, prompting them with options to pause, learn more, or decline specific outputs. This anticipates risk aversion and aligns with users' mental models. Designers increasingly build interfaces around Human-in-the-Loop (HITL) systems, which let users see where humans have corrected, verified, or reviewed AI decisions. Transparent feedback loops using Reinforcement Learning from Human Feedback (RLHF) are becoming standard. Salesforce's CRM platform flags predictions reviewed by sales managers, showing "human verified" badges next to forecast scores.

Just like users want to undo actions, they also want to know how the system made a choice. Provenance-enabled UX surfaces model lineage, data inputs, and decision pathways in product UIs. Klarna's credit decision engine displays a breadcrumb trail showing which factors contributed to the final output.

Designing for clinicians means clarity; designing for patients means reassurance. Visual explanations, step-based reasoning paths, and tiered explanations have helped systems support both. However, research [6] warns that overwhelming users can backfire; effective healthcare UX must balance transparency with information load. Financial AI requires traceable, regulation-compliant experiences. UX teams embed model explanations, opt-out features, and multi-tiered data logs directly into decision flows (e.g., loan application paths).

These principles were derived through a comparative analysis of over 20 AI-integrated product implementations, peer-reviewed UX and HCI research, and global regulatory frameworks published between 2018 and 2025. From this synthesis, we identify five core UX design principles that support user trust in AI systems (table 1).

Table 1 – Mapping Five UX Trust Principles to Real-World Design Patterns and Sector Applications

TRUST BUILDING UX STACK				
↓				
AI SYSTEM				
↓	↓	↓	↓	↓
1. LAYERED TRANSPARENC	2. CONTEXTUAL EXPLANATION	3. DYNAMIC CONSENT	4. PROVENANCE VISIBILITY	5. HUMAN TOUCHPOINTS
Offer multiple explanation depths and access points that fit user needs.	Embed just-in-time explanations at key decision points.	Let users modify data and model permissions fluidly.	Give users access to decision history and contributing factors.	Clearly indicate where human input has occurred or can be requested.

These principles are not features but foundations, building blocks for product teams designing responsible AI experiences. Trust is not a static state, it's earned, sustained, and tested through repeated interaction. For UX and product teams, this means designing systems that explain themselves, empower users, and acknowledge risk. The principles outlined here offer practical scaffolding to design AI-infused products that meet both user expectations and ethical standards.

Future work should explore how user trust evolves over time, how cultural context affects perception of algorithmic decisions, and how UX metrics for AI can align with global governance frameworks like the NIST AI RMF [5] or ISO 42001. Ultimately, designing for trust is designing for long-term relationships between people and intelligent systems.

References

1. Mitchell, M. et al. (2019). Model Cards for Model Reporting. FAccT. https://dl.acm.org/doi/10.1145/3287560.3287596.
2. Zhang, Y. et al. (2022). Interactive Model Cards for Explainable AI. arXiv:2205.02894. https://arxiv.org/abs/2205.02894.
3. Crisan, A. et al. (2024). X-Selector: Prediction-Aware Consent Design. arXiv:2404.03874. https://arxiv.org/abs/2404.03874.
4. UXMatters. (2024). Designing Visible AI Systems. https://www.uxmatters.com/mt/archives/2024/02/the-ai-advantage.php.
5. NIST. (2023). AI Risk Management Framework. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.
6. JAMA. (2024). Tiered Explanations Improve Patient Comprehension. https://jamanetwork.com/journals/jama/fullarticle/2798749.
7. European Commission. (2024). Digital Services Act. https://eur-lex.europa.eu/eli/reg/2022/2065/oj.