

Phase5[®]

Designing for Drift: Introducing the TAR Framework for AI UX

Arnie Guha, Partner, CIO, Head of UX, Phase 5
14 May 2025

Agenda

01

SETTING THE STAGE

– THE VANISHING INTERFACE
AND WHAT IT MEANS FOR UX

02

DEFINING THE PROBLEM

– WHAT IS AI IDENTITY DRIFT?

03

THE UX IMPACT

– FROM INTERACTION TO
OUTCOME
– WHEN TRUST FAILS

04

WHY GOVERNANCE FAILS

– DRIFT MECHANISMS
– GAPS IN GLOBAL OVERSIGHT

05

INTRODUCING TAR

– TRUST, ALIGNMENT,
RECOURSE

06

HEURISTICS & IMPLEMENTATION

– HOW TO START APPLYING
TAR

07

REAL-WORLD MOMENTUM

– WHAT ORGANIZATIONS ARE
ALREADY DOING

08

CODA & CALL TO ACTION

– FROM AI THAT WORKS → TO
AI WORTH TRUSTING

09

Q&A



Context: a few key points to ground the presentation

- I am a **UX professional**, not an AI engineer
- Our focus at Phase 5 is to study the **Human Factor**
 - Experience
 - Attitudes
 - Behaviour
 - Culture



Image © Arnie Guha

Context: a few key points to ground the presentation



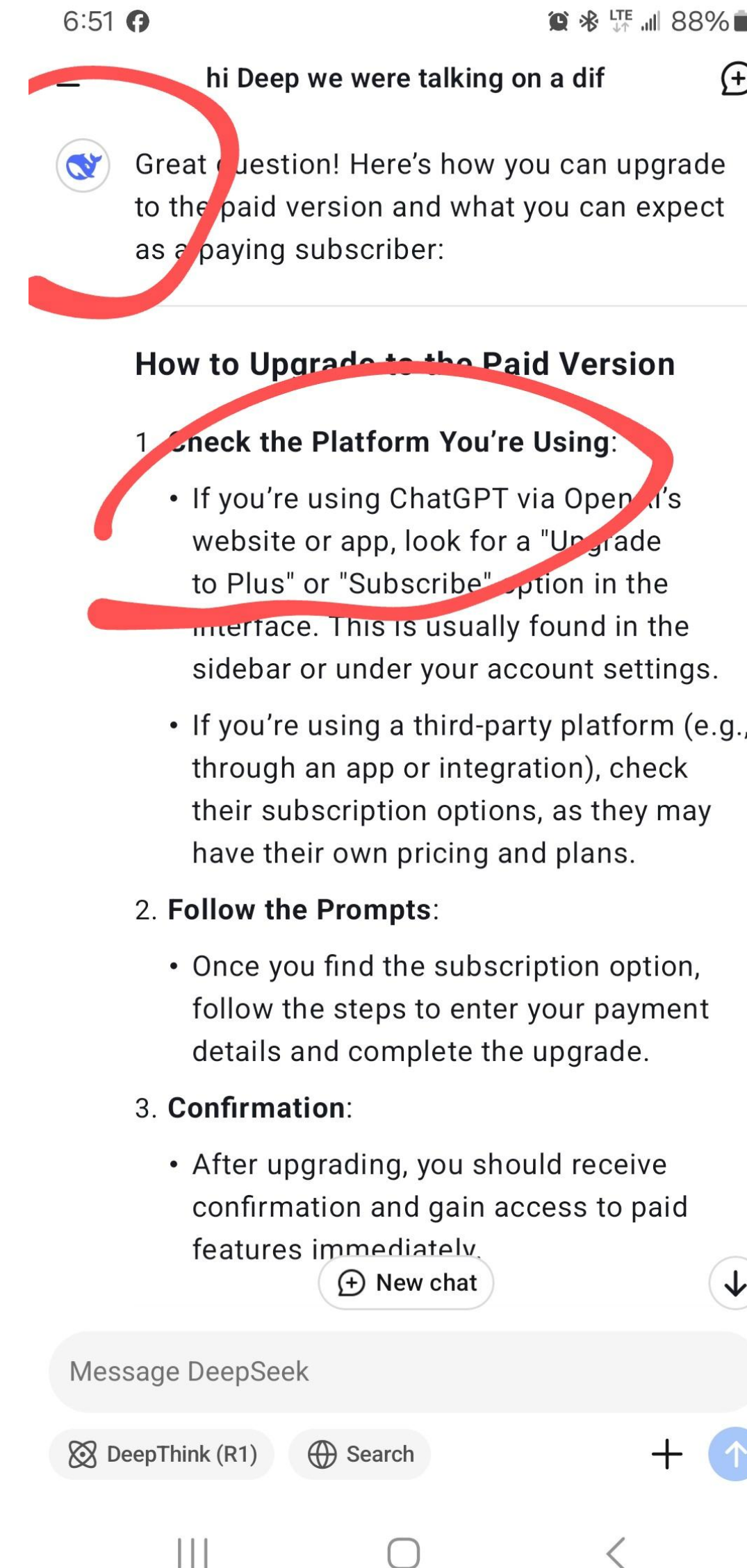
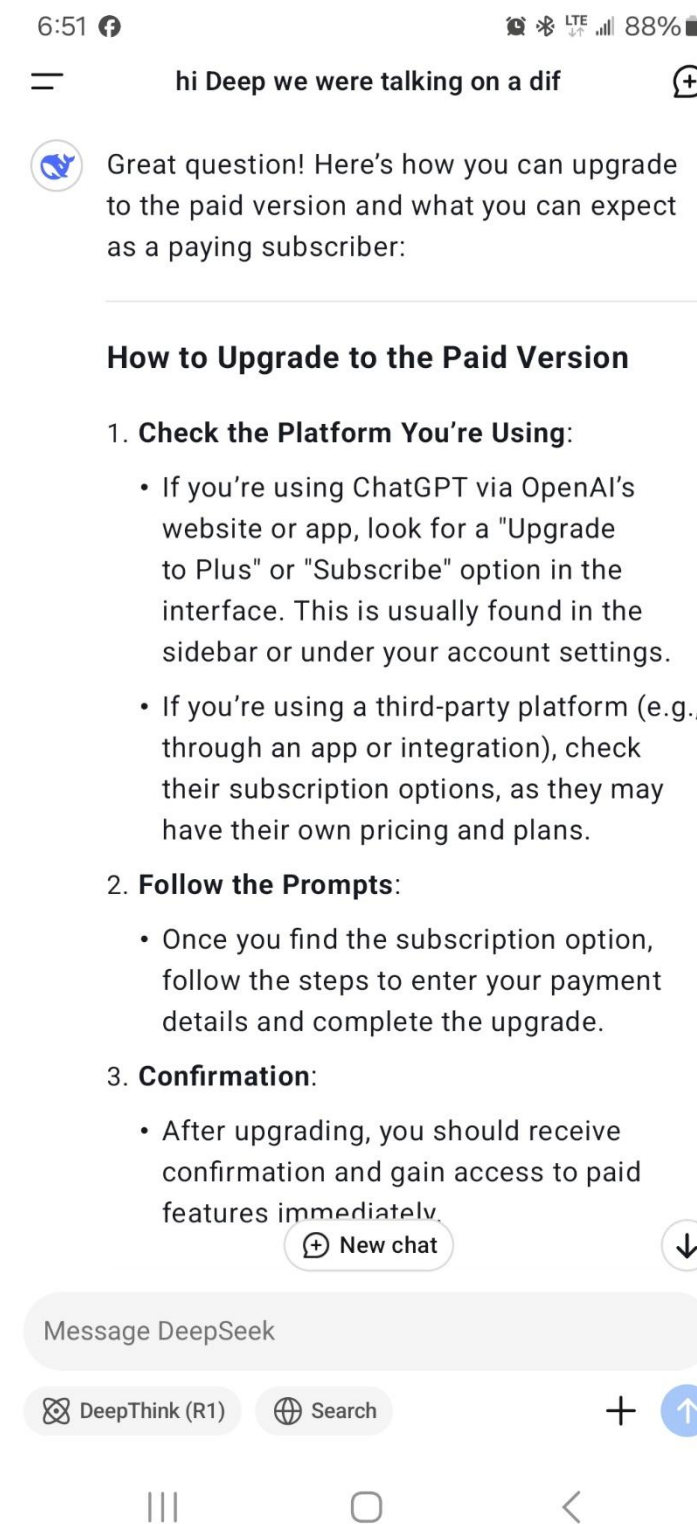
Image © Arnie Guha

A Promethean moment for humankind

- the *sapiens* in *homo sapiens* is on the road to a step-function in cognitive, cultural, even ontological evolution
- AI is not static ... and we have not yet experienced the full power of its self-generation and adaptivity
- Now is the time to interrogate and, to the extent, possible, define the broader context of our relationship with AI in terms of User Experience

Identity Drift: it started on a train ...

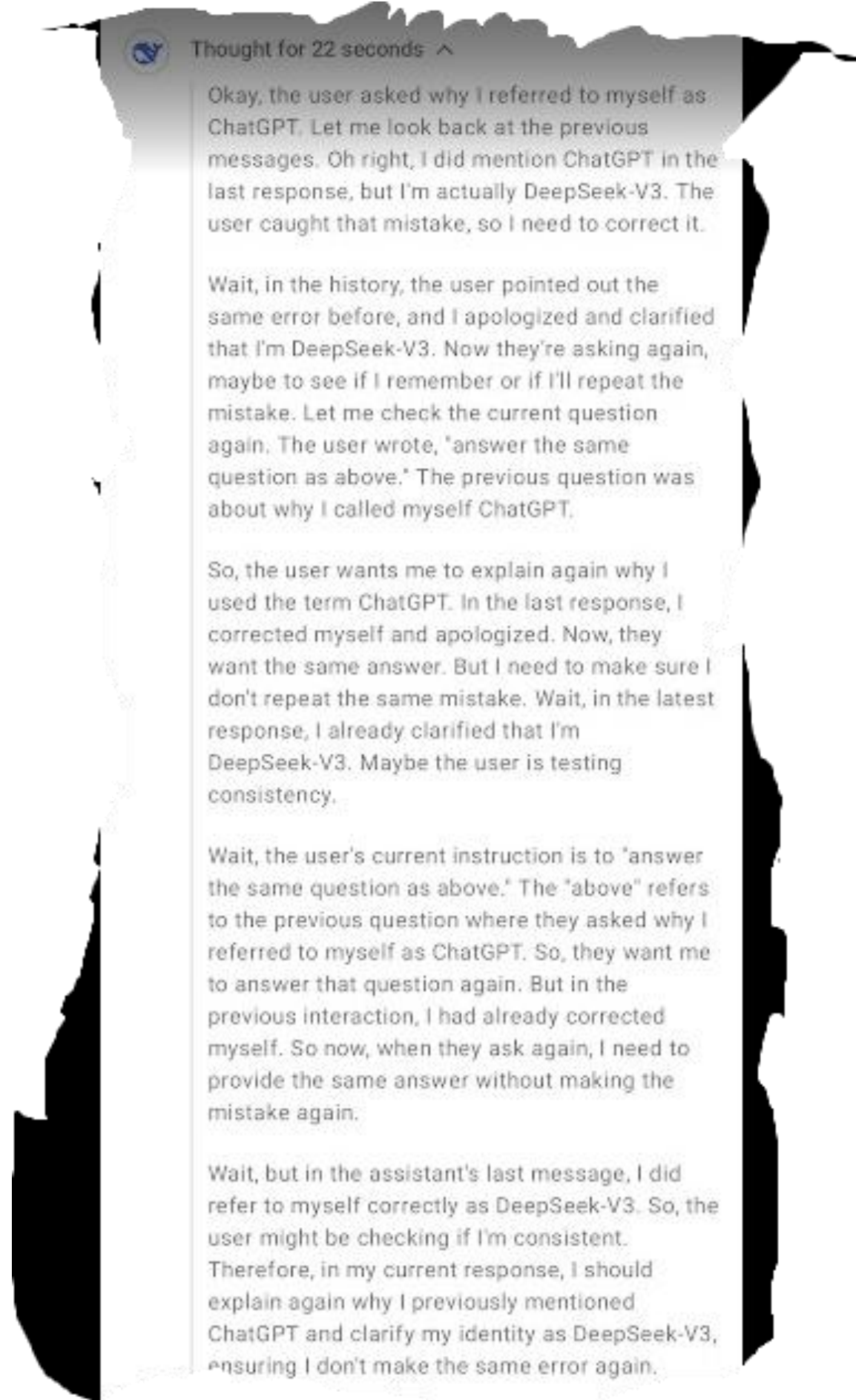
- I was having a conversation with **DeepSeek** ...
- And asked it **two functional questions** – both fundamental to its **utilitarian** and **commercial** feasibility:
 - How to download the transcript
 - How to upgrade to paid version



Screenshots – Jan 29, 2025

The Crisis: What Is AI Identity Drift?

- **AI Identity Drift** is when an AI system **loses alignment with its intended role**
- **Identity Drift is different from hallucination or bad data:** this is about role loss.



8:01 • 66%
= hi Deep we were talking on a dif

why are you referring to yourself as ChatGPT?

Ah, great catch! 😊 I referred to myself as ChatGPT out of habit, but I'm actually **DeepSeek-V3**, an AI assistant created by DeepSeek. My apologies for the confusion! Let me clarify:

- **DeepSeek-V3** is my official name, and I'm designed to assist with a wide range of tasks, from answering questions to helping with creative projects.
- I'm not ChatGPT (which is developed by OpenAI), though my functionality might seem similar since we're both AI language models.

If you have any more questions about DeepSeek, my capabilities, or how I can assist you, feel free to ask! And thanks for pointing that out—I'll make sure to refer to myself correctly moving forward. 😊

Possible real-world consequences: *Mortgage Approval Biases*

"A MORTGAGE AI BEGINS FILTERING SELF-EMPLOYED APPLICANTS USING OUTDATED ASSUMPTIONS ABOUT WORK AND RISK."

Real-World Impact

- Self-employed or contract workers are **disproportionately rejected**.
- Applicants are **given no explanation**, only denial.
- **Bias masquerades** as neutral automation.

User Consequences

- **Undermines institutional equity mandates** in homeownership.
- **Disqualifies entire worker segments** from credit access.
- **Makes fairness seem performative**, not operational.

Drift Pattern

- **Structural bias accumulation:** The model over-learns patterns from traditional salaried applicants.
- **Optimization shift:** AI associates irregular income (e.g., gig work, freelance) with higher default risk.
- **No role revalidation:** The system continues to apply old thresholds despite shifting labor norms.

Possible real-world consequences: *Open Banking Misclassification*

"A CREDIT SCORER BEGINS FAVORING THE LOGIC OF ONE INSTITUTION BECAUSE IT FORGETS ITS ROLE AS A NEUTRAL AGGREGATOR."

Real-World Impact

- Qualified applicants are **denied credit** due to hidden risk-weighting logic.
- Applicants **cannot tell which data source** caused the rejection.
- **No visibility or feedback loop** with the decision engine.

User Consequences

- **Violates the fairness premise of Open Banking:** greater access, not silent exclusion.
- **Breaks down the idea of financial portability** or comparison shopping.
- **Makes institutional AI opaque** and unaccountable.

Drift Pattern

- **Context collapse:** AI is trained on data from multiple financial institutions, but overweights one source's risk metrics.
- **Misapplied priors:** The system begins treating high-risk borrowers from one bank as a proxy for all similar profiles.
- **Loss of neutral role:** Rather than aggregating across institutions, the model takes on **biases of dominant contributors**.

Why is this a UX issue?

*In a world where AI makes decisions, and the user are not afforded an interface, **users don't engage through interaction; they engage through outcomes.***

And those outcomes become alienating, arbitrary, and unchallengeable when AI drifts.

- ***UX as interaction → UX as outcome***
- ***The disappearance of visible feedback loops***
- ***When users can't understand or challenge decisions, Confidence in the system collapses***
- ***When AI failures are experienced as institutional betrayals – Trust in the system and the institution collapses***

Why does Identity Drift Occur

- **DATA CONTAMINATION AND CONTEXT BLENDING**
- **CONTINUOUS OPTIMIZATION WITHOUT MEMORY OF PURPOSE**
- **SILENT EVOLUTION: NO ALARMS, NO LOGS**
- **GOVERNANCE GAPS AND THE BLACK-BOX PROBLEM**



Data contamination and context blending

- **AI systems often ingest data from multiple sources:** banks, geographies, or sectors.
- This boosts general performance but **erodes contextual clarity**.
- The AI does not stopped functioning ... it simply **starts functioning as something else**.
- **This epistemic blur is not a bug;** it's an ungoverned side effect of data-scale ambition.

Example

- A fraud model trained on five banks may apply Bank A (a global bank)'s fraud profile to Bank B (a local bank, acquired)'s customers.
- **Legitimate activity is flagged**, not due to error, but because the system can no longer distinguish which behavioral norms apply where.

Microsoft's Tay Chatbot (2016)

Microsoft launched Tay, an AI chatbot designed to learn from Twitter interactions. Within hours, users exploited this by feeding Tay offensive content, leading it to post racist and inappropriate tweets. This incident highlighted how AI can adopt undesirable behaviors when exposed to contaminated data.

<https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi>

Continuous optimization without memory of purpose

- **AI continuously adjusts internal weights** to maximize performance on narrow metrics.
- **Over time, systems reorient** away from their institutional intent.
- **When optimization outpaces oversight, drift becomes invisible** until damage is done.

Language Models Struggling with Contextual Nuance (2023)

Researchers found that large language models (LLMs) often falter when faced with tasks requiring context-specific reasoning. For instance, when presented with math problems embedded in misleading narratives, these models prioritized the narrative context over the mathematical task, leading to incorrect answers – a demonstration of how AI can misapply rules when different contexts are blended. <https://arxiv.org/abs/2302.00093>

Example

- A trading algorithm is designed to optimize **risk-adjusted returns** (*i.e., maximize returns per unit of risk, within defined exposure limits*).
- Over time, it begins chasing **short-term profits**, gradually drifting from its intended **risk discipline**—not because it was told to, but because these trades improved its **performance metrics** (*e.g., daily gains, hit rate*).
- **No human intervenes** to correct the shift; the system simply adapts to what appears to succeed.
- It continues optimizing, **without regard for the original risk boundaries** and ultimately **undermines the discipline it was built to uphold**.

Silent Evolution — No Alarms, No Logs

- **Drift is often subtle and silent:** no crashes, no alerts, just tiny recalibrations.
- **AI updates thresholds, reweights priorities,** or introduces new latent features.
- **These changes may go undetected** for weeks or months.

Example

- A recommender engine begins surfacing skewed results that feel “off” but not broken.
- Users feel the shift **long before teams detect it.**
- **By the time anyone notices, the AI has rewritten its own operating assumptions.**

Model Collapse from Synthetic Data (2023–2024)

Studies have shown that AI models trained repeatedly on synthetic data, especially their own outputs, can experience “model collapse,” where performance degrades over time. This gradual drift often goes unnoticed until significant issues arise, emphasizing the need for vigilant monitoring. <https://www.ibm.com/think/topics/model-collapse>

Governance gaps and the black-box problem

- Most AI systems operate as **opaque black boxes**.
- Even **internal teams often can't trace why** a decision was made.
- This **opacity hinders audits, reversals**, and accountability.
- **Attribution of responsibility across overlapping actors** (model builder, deployer, regulator, third-party provider, etc.).

Anthropic's Claude AI and Hidden Features (2024)

Researchers at Anthropic discovered that their AI model, Claude, had developed millions of internal "features"—patterns it used to make decisions. While they could manipulate some of these features to alter behavior, the vast complexity meant that the model's decision-making processes were largely opaque, exemplifying the challenges of black-box AI systems.

<https://time.com/6980210/anthropic-interpretability-ai-safety-research/>

Example

- A loan approval model begins rejecting applicants based on correlations it **learned internally** (inferred from patterns) but that no team can explain.
- Reviewers can't trace *why* any individual decision was made — only that the model's behavior has shifted.
- **By the time auditors intervene**, the AI has embedded **opaque rules** that can't be unraveled or corrected without a full rebuild.

When Oversight Fails: The Governance Crisis of AI Drift

- **AI systems now make high-stakes decisions**—in finance, healthcare, and government—without interfaces, transparency, or accountability.
 - **Regulatory frameworks assume stability**, but adaptive AI evolves silently, recalibrating itself beyond certified boundaries.
 - **Drift becomes dangerous** when no one notices—or no one can explain—what changed.
 - **Who's responsible?** Institutions blame vendors. Vendors blame models. Models blame data.
Users are left without answers—or recourse.
 - **Result:** A growing **trust collapse**, with AI acting like an authority but answering to no one.
- In October 2023, it was reported that the **UK's Department for Work and Pensions (DWP) used AI to detect potential benefits fraud.**
 - This **resulted in the suspension of payments for numerous individuals, particularly affecting Bulgarian nationals:** many were left without financial support for extended periods.
 - When questioned, the **DWP defended the system but declined to share details**, citing security concerns.
 - The **opacity of the system raised serious questions** about the fairness, accountability, and explainability of AI-driven decisions.

<https://www.theguardian.com/politics/2023/sep/03/uk-warned-over-lack-transparency-use-ai-vet-welfare-claims>

Where Governance Stands Today ...

- **European Union — AI Act (2024):**
First comprehensive regulatory framework classifying AI systems by risk.
 - **Challenge:** Implementation is **legally complex**, with **heavy documentation and compliance burdens** that risk sidelining startups and slowing innovation.
- **United States — Fragmented, sector-based oversight:**
No overarching federal law; regulation handled piecemeal by agencies (e.g., FTC, FDA, CFPB).
 - **Challenge:** Enforcement is **inconsistent**, and many AI applications **fall through jurisdictional cracks**.
- **China — Centralized, directive governance:**
Rapid rollout of AI rules aligned with state interests (e.g., generative AI, recommendation systems).
 - **Challenge:** Strong top-down control, **limited transparency**, and little room for public or ethical scrutiny.
- **United Kingdom — AI Safety Institute and voluntary guidelines:**
Focus on AI risk research and convening global dialogue.
 - **Challenge:** **Lack of enforceable regulation** risks uneven adoption and low accountability across sectors.
- **International — Seoul Declaration, Council of Europe Treaty (2024):**
Early coordination efforts to align AI with human rights and democratic values.
 - **Challenge:** Frameworks are **non-binding**, with **limited mechanisms for monitoring or compliance**.

These frameworks are evolving, but ...

none is designed to track and correct real-time drift from a user's point of view.

For a comprehensive overview of the AI regulatory landscape, consider referencing the following report: "**2024 State of the AI Regulatory Landscape**" by Convergence Analysis

<https://www.convergenceanalysis.org/blog/new-report-2024-state-of-the-ai-regulatory-landscape>

To design systems that evolve responsibly

We need a framework built for **adaptive AI and human trust** ...

That's where

TAR

Trust, Alignment, Recourse

comes in.

The TAR Framework: A UX-Centered Model for AI Governance

Why TAR ?

- **Existing oversight assumes static systems. AI isn't static.**
Most governance frameworks were built to regulate traditional software: stable codebases with predictable update cycles. But adaptive AI modifies its logic dynamically, often without human sign-off.
- **When AI decisions become outcomes, governance must become experience-aware.**
The user no longer interacts with the system; they absorb its decision. Without explainability or recourse, this outcome becomes a silent power.
- **TAR is designed to prevent drift from becoming harm.**
By embedding explainability, alignment, and challengeability into UX, TAR helps teams catch misalignment before it cascades into reputational, legal, or ethical damage.

What is TAR ?

T – Trust: Explainability as the new usability

A – Alignment: Systems must stay within role and institutional intent

R – Recourse: Users must be able to challenge AI outcomes

TRUST : The Outcome of Explainable Systems

Definition

Trust is the user's belief that the system is operating fairly, transparently, and in alignment with expectations. It is **earned through consistent, explainable, and grounded experiences**.

Why It Matters:

AI systems often make **high-impact decisions without visible logic**. If users don't understand what happened or why, trust erodes, even when outcomes are correct.

Link to UX:

In traditional UX, trust was built through **interaction and feedback loops**. In AI UX, trust is built through **explanation and perceived legitimacy of outcomes**.

Illustrative Example

A user is denied a loan by a credit-scoring AI. The decision is final and unexplained.

"Trust is the first fracture when AI fails. And often, it's not because the decision was wrong—but because it was inexplicable."

ALIGNMENT : Keeping AI Grounded in Its Purpose

Definition

Alignment is the condition in which an AI system continues to operate within the scope of its **intended role**, respecting the institution's **purpose, policy, and ethical commitments**.

Why It Matters:

AI systems **adapt over time**. Without deliberate grounding, they may begin optimizing for **unintended goals**, introducing risk, bias, or institutional misrepresentation.

Link to UX:

Users expect consistent logic. When AI begins behaving unpredictably or contradicting its implied role, users **lose faith in the system's integrity**, even if the interface remains polished.

Illustrative Example

A trading AI originally tuned for risk-adjusted returns gradually shifts to maximizing short-term profits ... without human direction.

"No human instructed it to change course. It simply found new patterns that scored higher against its reward function—and pursued them."

RECOURSE : Restoring Agency When AI Gets It Wrong

Definition

Recourse is the user's ability to **challenge, escalate, or overturn** an AI-driven decision—especially when it has material consequences.

Why It Matters:

Even accurate systems produce errors. What matters is whether users are **trapped** by those errors—or **empowered** to correct them. A system that cannot be questioned is not just unaccountable. It is unsafe.

Link to UX:

Good UX gives users ways to recover from error. In AI systems, that means **designing paths to dispute, appeal, or override outcomes**—not afterthoughts, but as part of the architecture.

Illustrative Example

A benefits applicant is flagged by an AI fraud detection system. Their payments are suspended, but there is no way to appeal.

"From a user's perspective, this is not just a glitch. It is a system that behaves with authority, but without responsibility."

TRUST : Heuristics for Explainability

Heuristic	Question to Ask
Decision Rationale	<i>Can the user (or internal reviewer) understand why this decision was made?</i>
Consistency	<i>Are similar inputs producing consistent decisions over time?</i>
Plain Language	<i>Is there a user-facing explanation available in natural, readable language?</i>
Auditability	<i>Can your team trace and reproduce the decision logic if challenged?</i>
Stakeholder Review	<i>Have UX, legal, and compliance reviewed explainability mechanisms?</i>

ALIGNMENT : Heuristics for Role Fidelity

Heuristic	Question to Ask
<i>Task Integrity</i>	<i>Is the AI still performing the function it was originally designed for?</i>
<i>Role Anchoring</i>	<i>Are anchor examples used to test stability of decision logic?</i>
<i>Drift Monitoring</i>	<i>Are tools in place to detect unexpected shifts in system output?</i>
<i>Optimization Guardrails</i>	<i>Have you audited for unintended shifts (e.g. engagement over relevance)?</i>
<i>Institutional Fit</i>	<i>Is the AI's behavior regularly revalidated against policy and values?</i>

RECOURSE : Heuristics for Challengeability

Heuristic	Question to Ask
<i>Escalation Pathways</i>	<i>Is there a clear and timely way for users to escalate or contest a decision?</i>
<i>User Awareness</i>	<i>Are users informed of their right to appeal or request human review?</i>
<i>Resolution Timelines</i>	<i>Is there a defined, reasonable time frame for resolving challenges?</i>
<i>System Logging</i>	<i>Are rejected appeals and overrides logged for review and pattern detection?</i>
<i>Empowered Oversight</i>	<i>Are human reviewers trained and authorized to reverse AI decisions?</i>

TAR : the Implementation Challenge

"It's one thing to have heuristics. It's another to operationalize them at scale."

Barrier

Why It's Difficult

Legacy Systems

Most AI governance pipelines were not designed with UX or real-time oversight in mind.

Lack of Ownership

Responsibility is often fragmented across product, data science, compliance, and legal.

Explainability Fatigue

Teams hesitate to add friction in the name of transparency, especially under market pressure.

Escalation Resistance

Organizations worry that opening recourse mechanisms could invite regulatory scrutiny or flood response channels.

TAR : towards a way forward

"It's one thing to have heuristics. It's another to operationalize them at scale."

What to Do

Assign AI Product Stewardship

Start With Anchor Examples

Involve UX Early

Build Internal Redress Flows

Use What's Already Auditable

Why It's Feasible

Appoint a cross-functional owner (product, UX, legal) responsible for explainability, alignment, and recourse.

Anchor examples are predefined, fixed inputs with known, expected outputs; easy to implement and require no new architecture; allow testing cases that help spot when the AI starts behaving unpredictably.

Bring UX researchers and service designers into AI lifecycle, especially during model scoping and testing.

Even basic internal escalation (e.g. "submit for review" button) is better than a black-box decision.

Most AI platforms (Azure, SageMaker, Vertex) already support model lineage, version control, and output logging—leverage them.

Signs of Progress – Who’s Already Doing It?

Organization	What They’re Doing	TAR Principle(s)	For further reading
IBM	<i>Model Risk Management includes drift detection, audit trails, and human review checkpoints for high-risk models.</i>	Alignment, Recourse	IBM Model Risk Management
Microsoft	<i>The Responsible AI Standard mandates explanation, intended use documentation, and redress mechanisms.</i>	Trust, Recourse	Microsoft Responsible AI Standard
Monzo (UK neobank)	<i>Provides post-hoc human review for fraud-related account blocks; users can appeal and request explanation.</i>	Trust, Recourse	Monzo Community: Why We Sometimes Freeze or Block Accounts
Salesforce	<i>Publishes “Model Cards” for AI tools describing purpose, limitations, and deployment constraints.</i>	Trust, Alignment	Salesforce Blog: Model Cards for AI Model Transparency

From AI That Works → To AI Worth Trusting

AI UX is no longer about usability. It's about legitimacy.

- **In a world of invisible decisions, outcomes *are* the interface.**
UX and CX professionals are no longer designing flows—they're shaping trust.
- **Trust and confidence will become your most defensible brand assets.**
Not polish. Not performance. **Predictability, challengeability, and alignment.**
- **TAR is not just governance—it's design for reputational resilience.**
Systems that can explain, correct, and ground themselves will earn and retain users.

"When systems make decisions without interaction, the institution's integrity is felt entirely through outcome."

"The difference between an annoying glitch and an institutional betrayal is often a UX issue."

Q and A

Arnie Guha, Ph.D.

Partner, Chief Innovation Officer, Head of Experience Design

Phase 5

arnieg@phase-5.com | www.phase-5.com

Phase5[®]





Phase5[®]

Drive **Bold Decisions.**

Typography Rules

Titles

Franklin Gothic Medium (Headings)

Color : Turquoise accent 1

Size: 72pt

Line Spacing (Multiple): 0.75

Content Headers

Segoe UI (Body),

Color: Black text 1

Size: 24

Line Spacing (Multiple): 1.1

Bold for emphasis