

AI-native Delivery Experience

Evidential Independence and Enforceable Human Oversight

[Brief for open publishing. This is not the full presentation.]

Spyros Ktenas
Ionian University

Prof. Panagiotis Vlamos
Ionian University

IntDEx

- Agentic LLMs tools fundamentally transform how digital solutions (software) are conceived, designed, and delivered.
- Human generated deterministic content vs fast non-deterministic, emergent behavior driven by intents and interactions between prompts, models, and data.
- Author – tool separation vs Tools and Authors mix

[Brief for open publishing. This is not the full presentation.]

IntDEx – Key Elements

- Principles/Characteristics
- Delivery Flow
- AI-native team roles and responsibilities
- Expanding on
 - Risk controls
 - Autonomy and response
 - Cost control
 - Governance
 - Ethics
 - HITL
 - Automatic continuous behaviour validation
 - Quality vs throughput maximalisation

[Brief for open publishing. This is not the full presentation.]

| The risk of AI checking itself

One model can write every part

The requirements

The code

The tests

The final report

The same mistake can appear in every part.

A complete record alone cannot prove that the result is correct.

The paper proposes independent evidence and human approval that can block work as an essential part of the AI delivery engine.

[Brief for open publishing. This is not the full presentation.]

Evidence must come from another source

The paper ranks evidence from strongest to weakest

Rank (Strength)	Evidence source	Simple example
4	System execution	A test report or a recorded system response
3	Named human	A recorded review of the actual work
2	Another AI model	A review by a different provider model family
1	Earlier work record	A comparison with a previous version
0	The authoring model	Its own claim that the work is correct

Self-reports stay unverified. Evidence must also be strong enough for the risk.

[Brief for open publishing. This is not the full presentation.]

Human review must be able to stop work

A real approval record

A named human makes the decision.
The record explains the reason.
It names the files the human checked.
It includes evidence and a time.

**AI can create a checkpoint.
Only a human can approve it.**

Oversight debt

Pending major decisions
add to the debt.

**Work stops when debt
exceeds the set limit.**

Review covers changes to security,
data, behaviour, scope and release.

[Brief for open publishing. This is not the full presentation.]

Indicative additional chapters:

- ☐ Seven rules make the process checkable
- ☐ The study used one public-health platform (+data)
- ☐ Human review introduced changes on a number of checkpoints (+data)
- ☐ A second model helped, but raised false alarms (+data)
- ☐ The rebuild worked, but security became weaker (+data)
- ☐ Positioning against established approaches
- ☐ Artefact specification
- ☐ Templates/Examples (HITL checkpoint/XMODEL response)
- ☐ Evidence vs Severity Matrix
- ☐ ...

These checks test the records. They cannot prove that a reviewer is right.

[Brief for open publishing. This is not the full presentation.]

Thank you!



[Brief for open publishing. This is not the full presentation.]