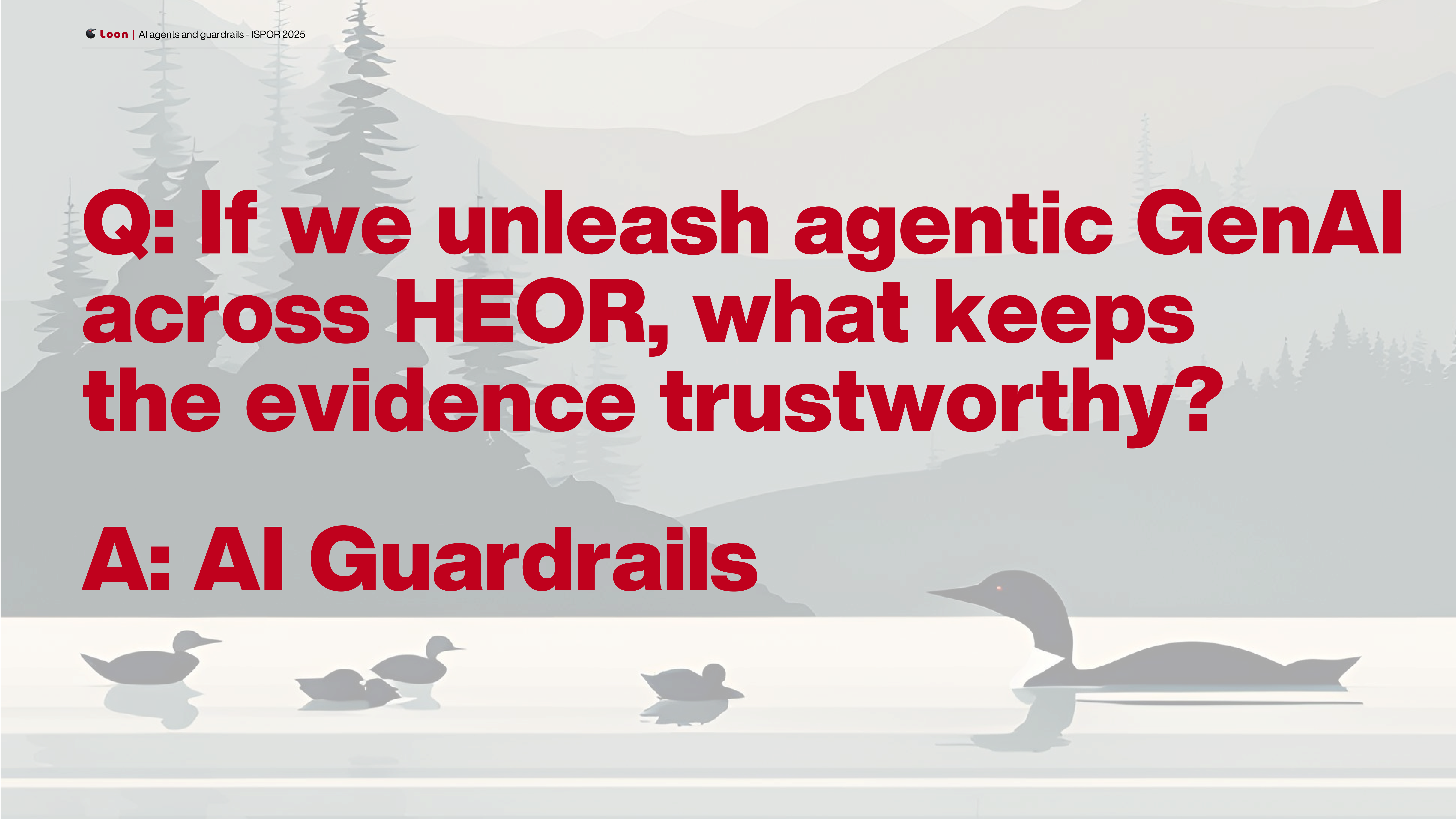


# Q: If we unleash agentic GenAI across HEOR, what keeps the evidence trustworthy?

# A: AI Guardrails

# *Conflict of Interest Statement/ My Biases:* Co-founder of Loon

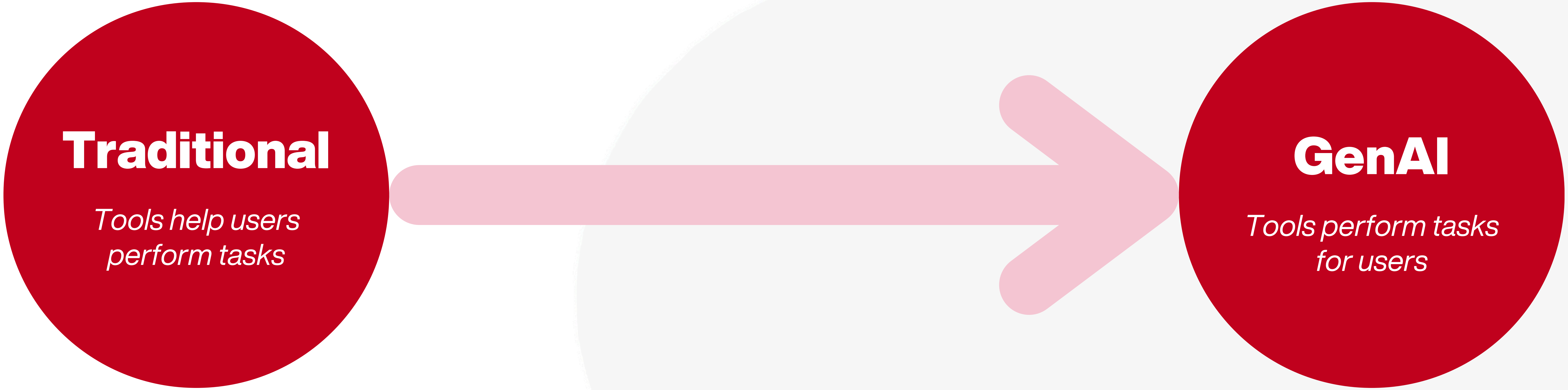
## Loon builds, validates, and utilizes agentic AI systems in HEOR.

*Loon, Ottawa, Ontario, Canada*

### Disclaimer

Loon has no financial interest in any tools, libraries, or frameworks mentioned in this presentation.

# GenAI software applications must have a *different thinking paradigm* than traditional software tools



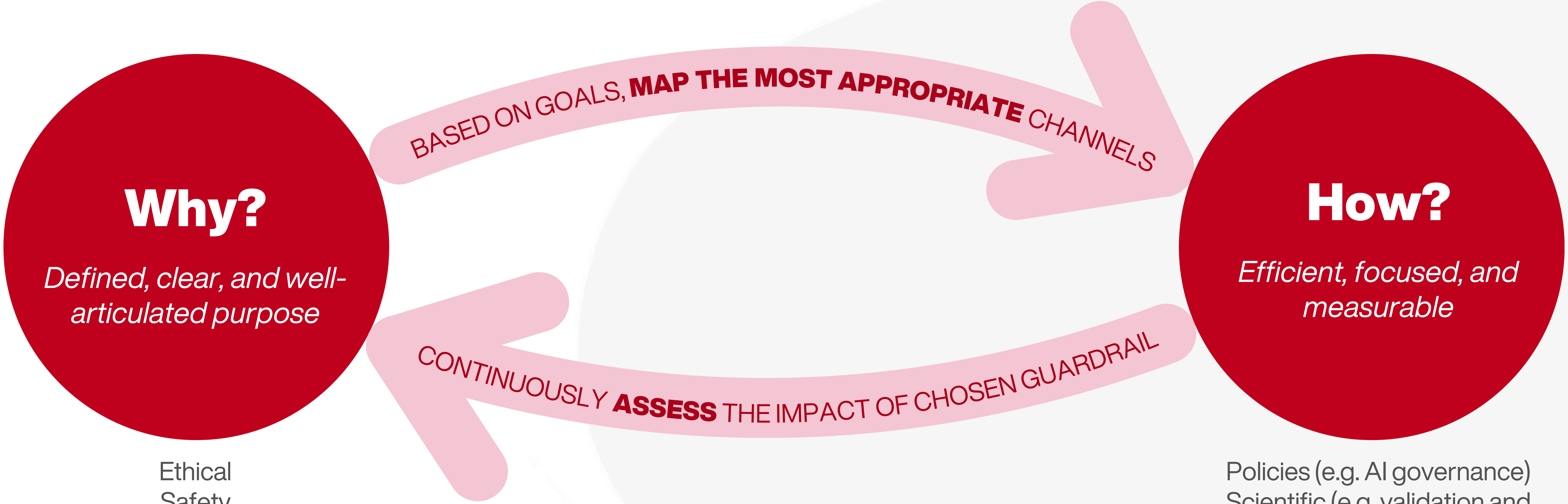
**Traditional**

*Tools help users perform tasks*

**GenAI**

*Tools perform tasks for users*

# *AI Guardrails:* Policies, constraints, and technical measures that keep AI behavior within safe and desired bounds.



Why?

*Defined, clear, and well-articulated purpose*

BASED ON GOALS, **MAP THE MOST APPROPRIATE** CHANNELS

How?

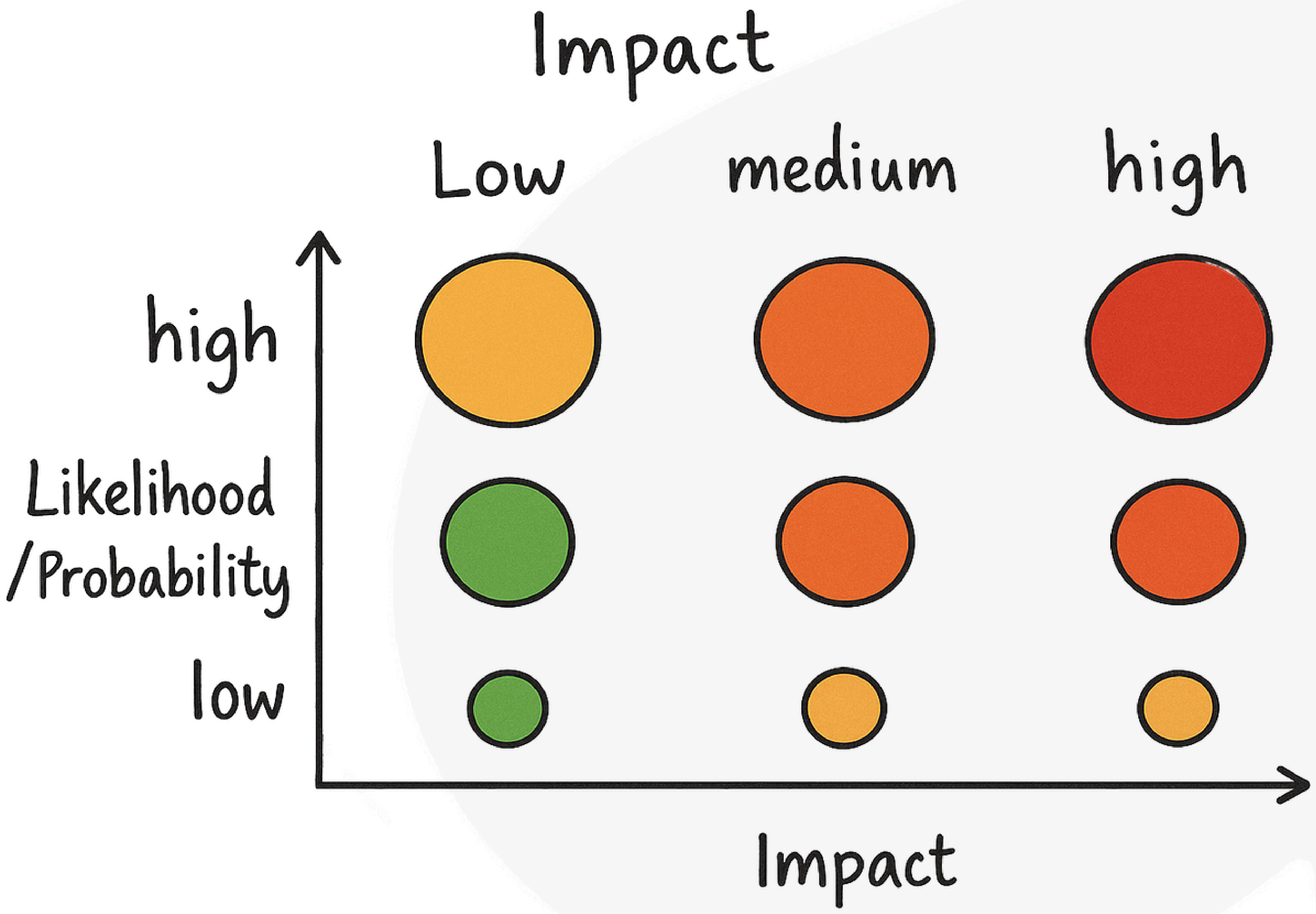
*Efficient, focused, and measurable*

CONTINUOUSLY **ASSESS** THE IMPACT OF CHOSEN GUARDRAIL

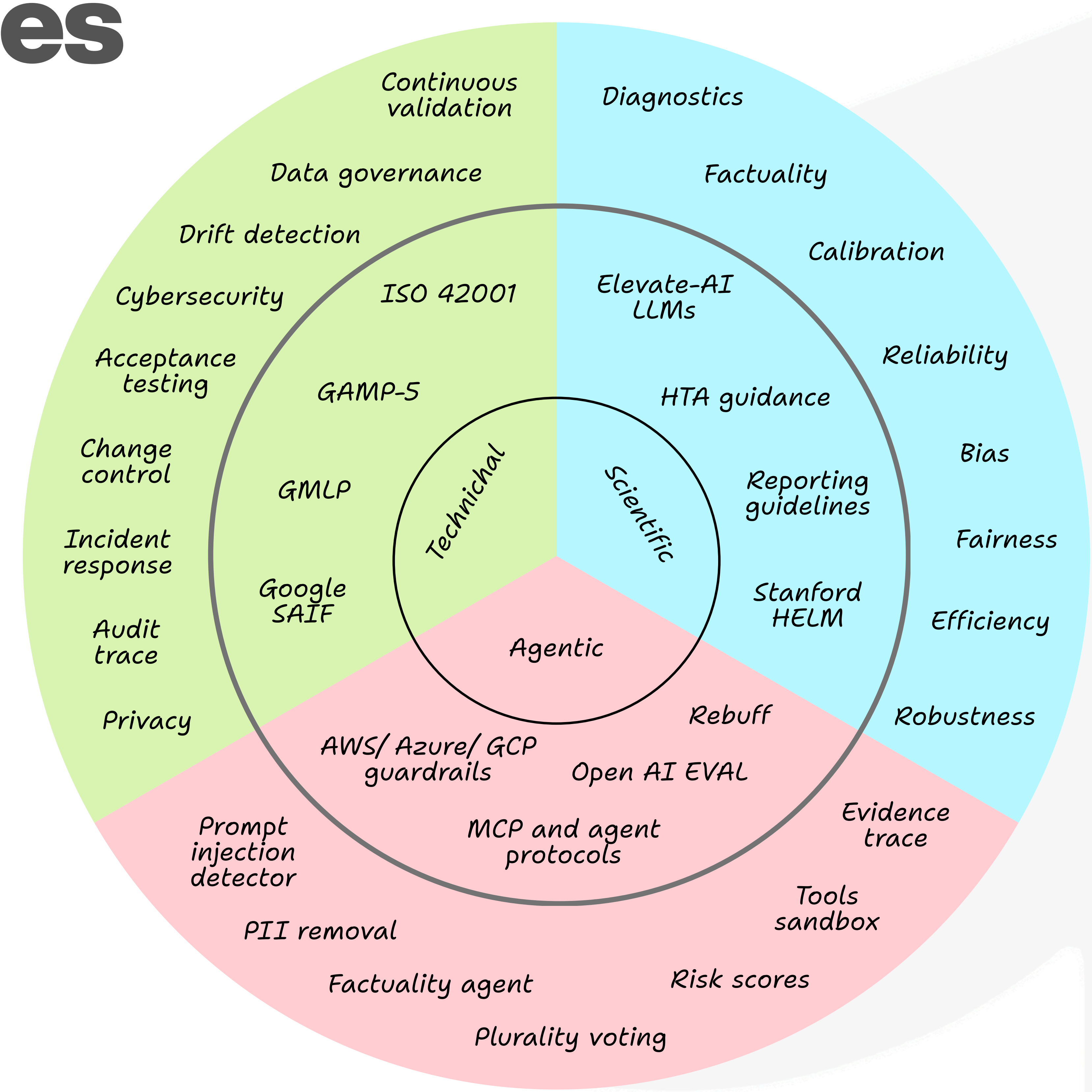
Ethical
Safety
Accuracy
Reliability
Transparency

Policies (e.g. AI governance)
Scientific (e.g. validation and calibration studies)
Technical (e.g. encryption, filters, factuality tests )

# AI Guardrails: Risk-based mindset



# AI Guardrails - The Lay of the Land: Three Lenses



# ***HTA Bodies:*** a Major stakeholder that will determine the pace of adoption

> *Concerns about the appropriateness, transparency and trustworthiness of AI do exist.*
>
> *NICE AI statement (also adopted by CDA-AMC)*

# *Scientific Lens:* Generating scientific evidence to support claims may be challenging, but it is essential

### Diagnostic studies

Assessment of ***accuracy, agreement, recall, F1 scores, and other relevant metrics.***

### Numerous variables

***Prompts, models, parameters, architecture, and many more*** variables need to be controlled and assessed.

### Calibration studies

***Assessment of the AI output confidence*** metric to determine when the AI output is suboptimal.

### No good gold standards

***No "ground truth"*** dataset for many HEOR tasks and none assessed for errors.

### Measurement studies

Prompts and agents assessment can borrow a lot from ***measurement science.***

### Resource intensive

Properly designed and conducted validation studies are ***expensive, complex, and time-consuming***.

# *Technical Lens:* Software development cycle with a flare for AI

## GMLP

*Principles adhered to by regulated devices*

ML-specific

## GAMP-5

*Industry standard for software in life sciences*

Software-specific

## ISO/IEC 42001:2023

*AI development and adoption management*

AI-specific

# *Agentic Lens:* Specific tools and libraries for good Agentic AI practices

## Cloud Provider Guardrails

*Tools for filtering prompts, denying topics, and more*

## OpenAI Eval

*A standard and tested template for evaluating the performance of each agent*

## MCP & Agent Protocols

*Rapidly evolving. Manges info control and flow*

# *Example of A Problem in Need of Guardrails:* Initial performance metrics may not reflect real-world use performance metrics

## Model Drift

*Traditional concerns of ML model drift apply to agentic AI systems*

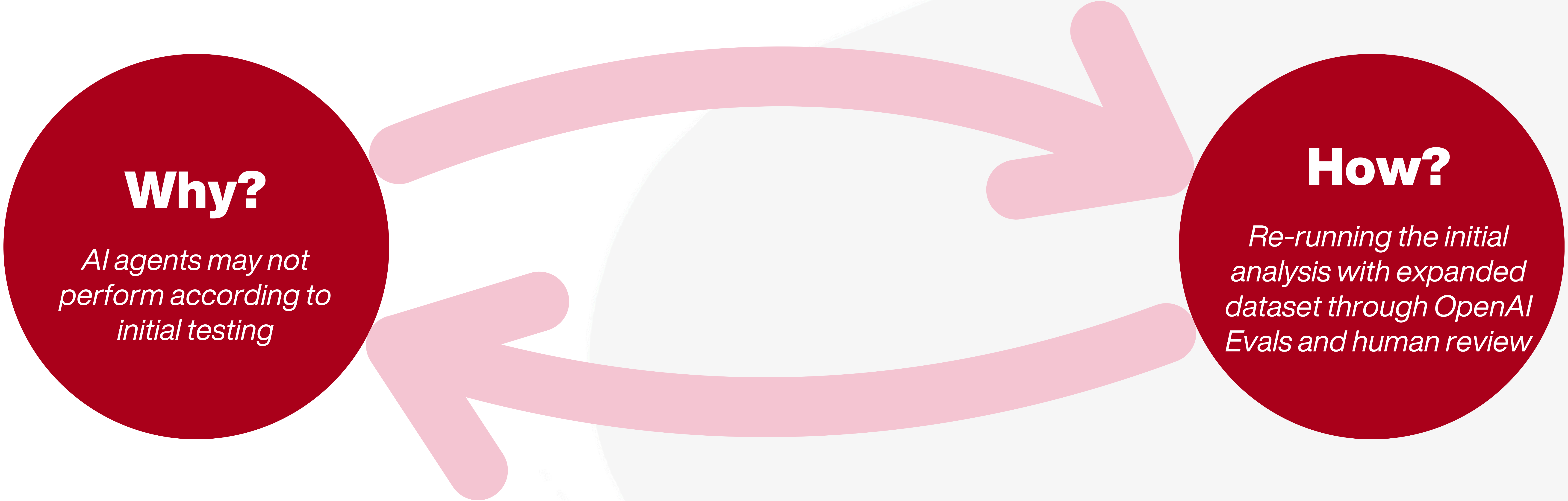
## Foundational Models

*Tweaks in third party foundational models can break validity*

## UX Effect

*Forcing certain inputs and outputs can affect performance*

# *Guardrail Example - Combining All Lenses:* Continuous validation using OpenAI Evals



**Why?**

*AI agents may not perform according to initial testing*

**How?**

*Re-running the initial analysis with expanded dataset through OpenAI Evals and human review*

# *As a tool user:*
# Minimum guardrails to look for

☑ **Diagnostic/performance measurement studies**

Comprehensive, transparent, and rigorous scientific work to scientifically test each tool claim. More proof, less brag.

☑ **Uncertainty/confidence calibration studies**

Ability to focus human-in-the-loop validation effort and the extent to which that method corresponds with true high-risk output .

☑ **Continuous validation studies**

Ensuring that there is a mechanism in place to deliver performance that matches, or ranges within within an established margin of error, to the original validation studies.

☑ **Change control mechanism**

Provide assurance that any change in the system will not lead to unexpected consequences.

# As a reviewer of AI-enabled research: Minimum guardrails to look for

☑ **Citing diagnostic/performance measurement studies in the methods**

Must be part of the protocol and treated as a tool used in the experiment/study.

☑ **Justification for human-in-the-loop validation method**

Calibration studies to support targeted validation or justification for other methods of validating of AI output.

☑ **Adherence to reporting guidelines**

For GenAI applications in HEOR, the ELEVATE-AI LLM framework is appropriate.

☑ **Complete transparency of AI output**

AI outputs and the resulting human decisions should be communicated clearly in a step-by-step fashion.

# *The Ultimate Guardrail in HEOR:* Can results be independently reproduced?

Same Methods

Same Data

Same Findings

# Thank you!