



Beyond the Model: Towards a safety ecosystem for AIES

Michael Thate and Zoe Szajnfarder

The George Washington University

Funding acknowledgement: (1) NSF NRT-AI-FW-HTF: Co-Design of Trustworthy AI in Systems (DTAIS) and (2) the NIST-NSF Institute for Trustworthy AI in Law and Society (TRAILS)

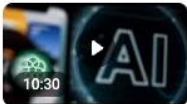


Motivation: why we need to look “beyond the model”

Lots of attention on AI safety...

Reuters
https://www.reuters.com › business › openai-faces-senat...
US lawmakers call for new AI rules after Anthropic ...
60 minutes ago — OpenAI said on Wednesday it was pushing for mandatory national AI safety requirements in the United States on concern the technology could ... [Read more](#)

Videos :

 **AI safety fears grow as experts warn technology could spiral out ...**
YouTube · FOX 32 Chicago · 1 hour ago

 **AI researcher warns companies are ignoring catastrophic risks**
YouTube · PBS NewsHour · 18 hours ago

 **Should we be worried after Anthropic researcher's warning of ...**
YouTube · Sky News · 19 hours ago

Most solutions point to better...

...*guardrails*

... *alignment*

... *testing*

Model level interventions

Missing the larger eco-system interventions

This is **not** how safety works in mature industries. Are there lessons for the AI space?

Approach

Mapped “safety ecosystems” from high-risk, low-incident contexts, looked at incidents in each to see how threats are mitigated and drew out common governance principles



Specific incident:
Engine fire



Pig gets sick



Drug recall

Illustrative Example

What are the mechanisms/controls in place that so that **this type of event is rare?**

What are the mechanisms/controls in place that so **if it happens, harm is limited?**

Fractured material,
misaligned seal,
Over-fueled
Overtaxed engine
Electrical storm
Bird flies in...

Manufacturing guidelines,
inspection and maintenance,
crew training ...

**Adverse event
(engine catches fire)**

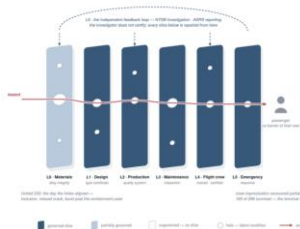


Crew training, emergency
procedures, back-up engine

+ ground crew,
emergency response



Commercial Aviation



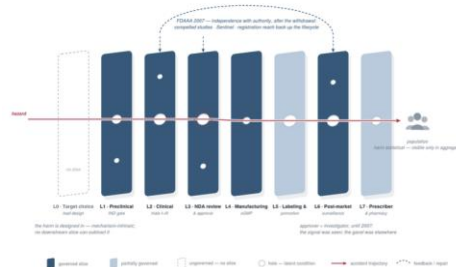
The process continues - transfer (20% after completion - transfer the report, and steps the form is signed)

time

1. Generate ideas
2. Develop goals
3. Transport it along
4. Assemble the team
5. Strengthen it along
6. HINDP & validation
7. Cash chain

customer satisfaction at the end

Pharmaceuticals



Tight-knit farming communities follow more stringent, generational knowledge about what types of cleanliness matter.

Government bears cost of certain types of testing and monitoring

Parents teach their kids about food safety best practices (beyond best before dates)

Cross-case takeaways

- All three were governed by a complex network of control and mitigation mechanisms,
 - targeted at different **actors** (e.g., pilots vs. mechanics, farmers vs. grocery stores)
 - over different **phases of the lifecycle** (e.g., manufacturing vs. operations).
- We identified a mix of **de jure** (rules) and **de facto** (norms) in each setting.
 - In some cases, norms were the main piece of the layer (e.g., farming practices inherited over generations).
 - The rules came from many **different agencies and entities**.
- While the specifics of the layers were tailored to the context, they all leveraged **multiple** barriers, balancing mechanisms to **limit likelihood** (in advance) and strategies to **mitigate harms** (after an incident occurred).
- All had established strong feedback loops to promote **continuous learning**.

Important caveat: these ecosystems grew up slowly over decades

Implications for “AI” (beyond the model)

- It matters what kind of “AI” and how humans can and do interact with “it,” and when

ChatBot



A Drone (and AIES)



Individuals

Data curators, model architects, safety/alignment specialists, “App” developers, users

Orgs

Start-ups, Foundation labs, Enterprise companies

Potential formal mechanisms

Training, credentialing, licensing, registers, incident reporting

Inspections, certification, maintenance, testing, audits (shared infrastructure for above)

Potential informal or indirect mechanisms

Norms, culture around use, generational wisdom

Voluntary standards, Best practices about roll-outs

Conclusion and next steps

- This work already shows how many more risk-mitigation mechanisms exist than are currently being leveraged. Need for careful mapping of types of mechanisms to types of risks, by system class.
- Analogy is challenging, but this work provides a first step:
 - Recognize the different types of “AI” (not just high vs. low risk, but different modes of interacting with end-users)
 - Unpack space of stakeholders and organizations along the value chain
 - Clarifies what levers apply when
- Continuing to synthesize generalized governance ecosystem to more directly inform other novel technologies.

Thank you!

zszajnfa@gwu.edu

