
Pawsitive Persuasion
Using Cat Comics to Claw Your Organization Closer to Responsible AI Adoption
Kiowa Scott-Hurley | Slide Enthusiast
Dr. Chris Hines | Cat Comic Connoisseur
eResearch Australasia 2024 | Oct 31st

TABLE OF CONTENTS

ICYMI

What's GenAI again?

RAI THIS, RAI THAT

Let's dig in a bit here

KTHXBYE

Summary, questions, etcetera


1
ICYMI
What's GenAI again?

JUST IN CASE...

Ye Olde ML

- Learn and extract insights from data.
- Can be used to:
  - Classify data.
  - Make predictions.
  - Make recommendations.
  - Perform detection: images, anomalies.
  - Perform pattern recognition.
  - Discover labels for data.

Generative AI

- Learn about a dataset and create new data from what is learned.
- Has been applied across data types:
  - Text
  - Image
  - Video
  - Audio
  - Multimodal

WHY RAI ALL OF THE SUDDEN?

Accessibility

Even Snapchat has an LLM chatbot now!

Expertise

Despite broad access, user knowledge is still limited…

Explainability

If you thought the users were confused by the outputs, you should chat to the devs!

Human? AI?

It's getting harder to know who's who… perhaps I am a bot.

Potential Disaster

GenAI has increased risk of misuse - generative malware, deepfakes, misinformation…

WHY RAI ALL OF THE SUDDEN?

Privacy

Turns out everyone having access to these tools is predicated on big corporations and cloud hosting - who knew!

More Data Problems

Copyright? Bias? IP? Anyone???

T&E

The ways to effectively test and evaluate these models are still being established - a test might show a model with low bias, but unexpected user interactions may prove otherwise.


11
RAI THIS, RAI THAT


WHAT IS
RESPONSIBLE AI?

Government

Department of Industry, Science and Resources: Australia's AI Ethics Principles, National Framework for the assurance of AI

Human, societal and environmental wellbeing: AI systems should benefit individuals, society and the environment.

Human-centred values: AI systems should respect human rights, diversity, and the autonomy of individuals.

Fairness: AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

Privacy protection and security: AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

Reliability and safety: AI systems should reliably operate in accordance with their intended purpose.

Government

Transparency and explainability: There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.

Contestability: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.

Accountability: People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

INDUSTRY

Microsoft Responsible AI Standard

- Fairness and inclusiveness
- Reliability and safety
- Transparency
- Privacy and security
- Accountability

Google - Responsible AI practices

- Fairness
- Interpretability
- Privacy
- Safety and Security
- General practices
  - Use a human centered design approach
  - Identify multiple metrics to assess training and monitoring
  - When possible, directly examine your raw data
  - Understand the limitations of your data and model
  - Test, Test, Test
  - Continue to monitor and update the system after deployment

CSIRO

Their documentation was too comprehensive (read: scary) to include in a list: https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/

Process Patterns

We identify process-oriented patterns (i.e. best practices) that can be incorporated into development processes, so the developers could consider to apply them during the development lifecycle. Fig.4 describes the software development lifecycle and the potential ethical risks and breaches corresponding to each stage, while Fig.5 presents the summarized patterns for different stages.

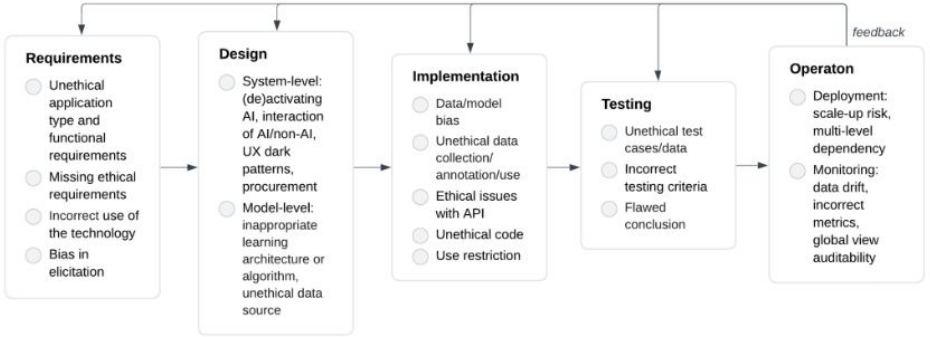


Fig.4. Development process lifecycle and potential ethical risk.

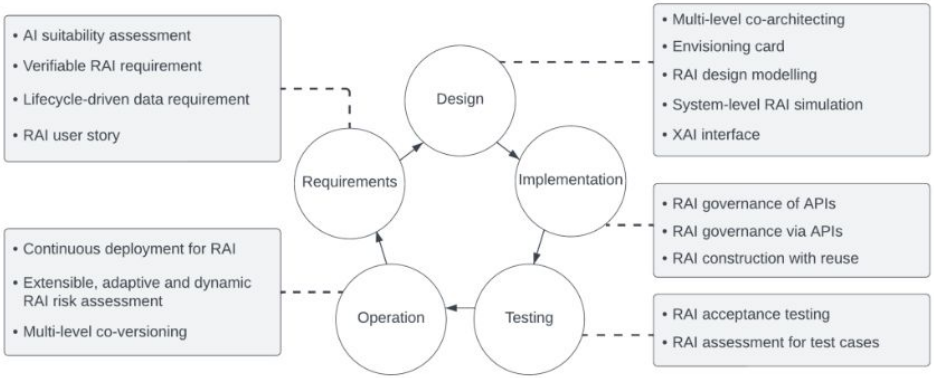


Fig.5. Process patterns for responsible AI system development.

NIST - AI Risk Management Framework

Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

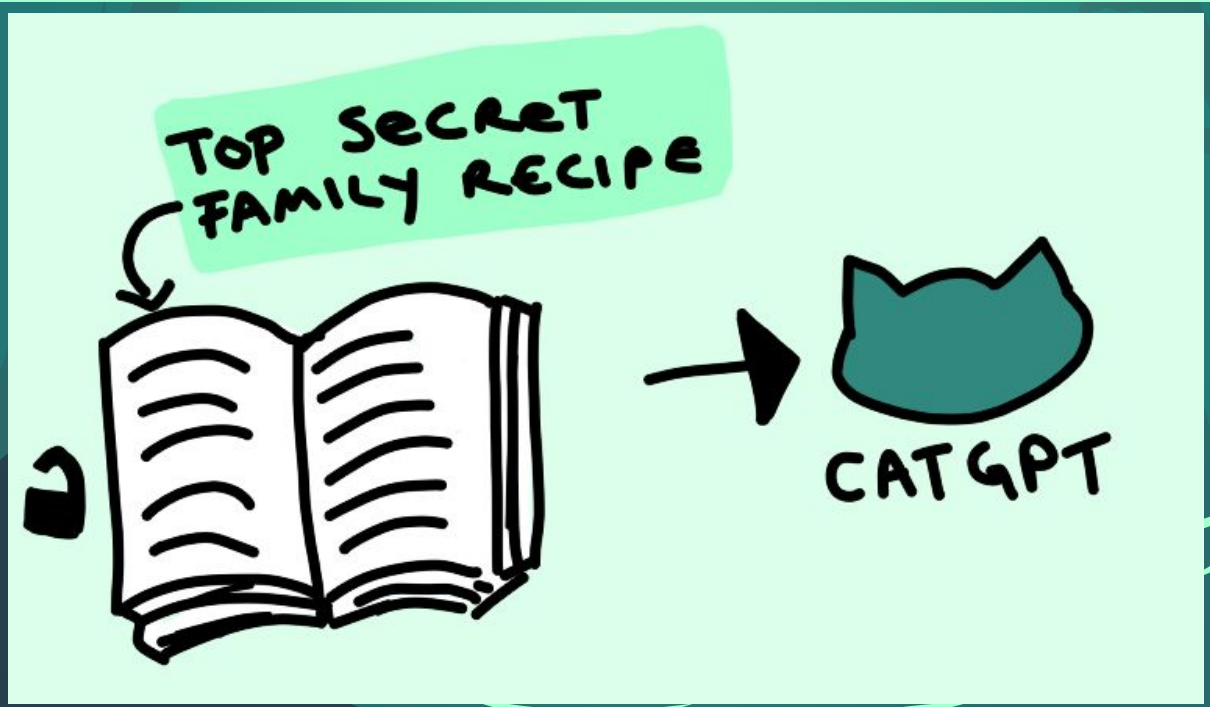
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Safe

"AI systems should “not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered” (Source: ISO/IEC TS 5723:2022)"

Secure and Resilient

Secure: Meeting standards of availability, confidentiality and integrity.



Secure and Resilient

Resilient: Under adverse conditions, the models fails "gracefully".

Transparent

Transparency can answer the question of “**what happened**” in the system.

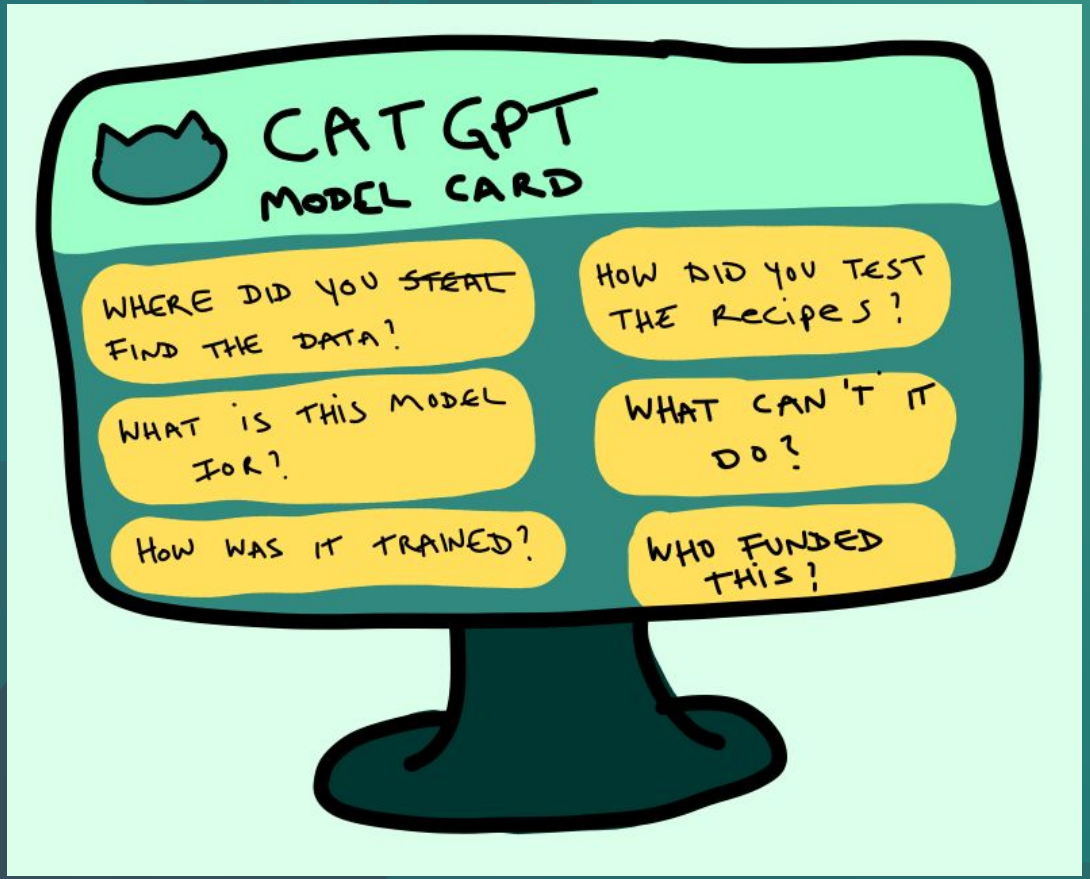
Explainable

Explainability can answer the question of “**how**” a decision was made in the system.

Interpretable

Interpretability can answer the question of “**why**” a decision was made by the system and **its meaning or context to the user"**

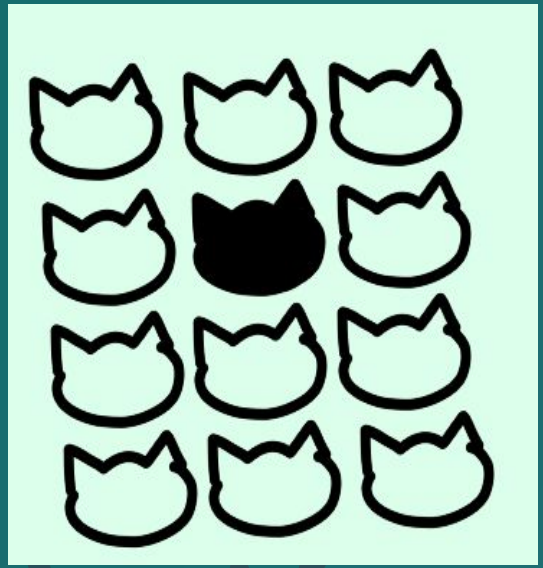
TRANSPARENT



Explainable & Interpretable

Privacy-Enhanced & Fair

"**Privacy** refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity."

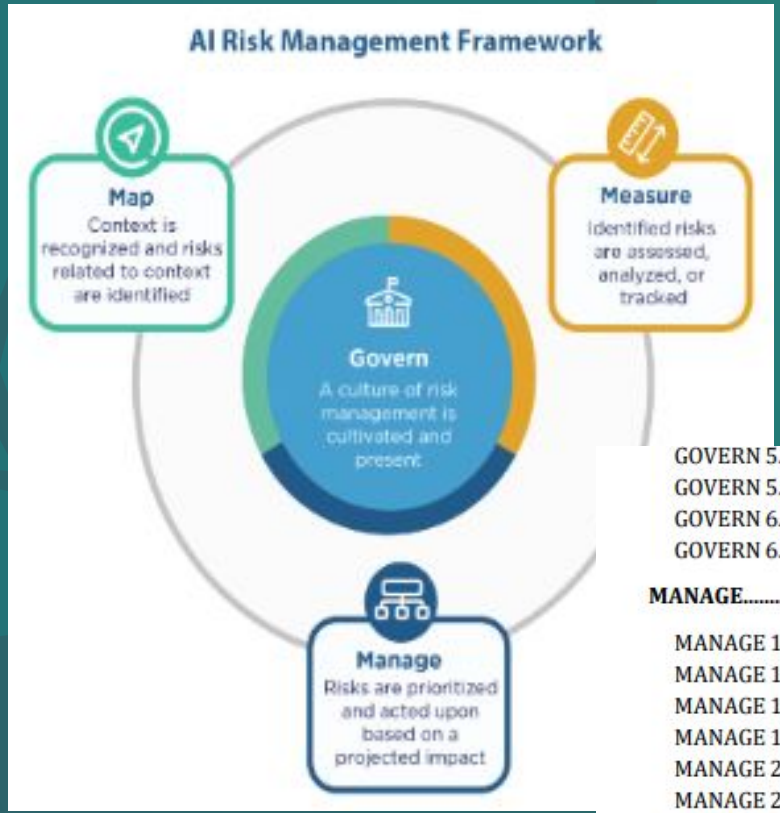


"**Fairness** in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination"

Valid and Reliable

"Validation is the “confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled” (Source: ISO9000:2015)"

"Reliability is defined in the same standard as the “ability of an item to perform as required, without failure, for a given time interval, under given conditions” (Source: ISO/IEC TS" 5723:2022)"




III
KTHXBYE

CHALLENGES: FRAMEWORKS

WHICH ONE?

It's challenging to actually compare frameworks - they aren't exactly fun to read either!

DETAIL

Frameworks vary wildly on the detail level they offer - principles or playbooks?

MISSING PIECES

Frameworks are not comprehensive (i.e. missing things like the Privacy Act and other regulations).

EXPECTATIONS

How do I *actually* get a diverse team? What is "expected" behaviour for GenAI?

OTHER CHALLENGES

PACE

- Do the frameworks keep up with the technology?
- Do I want to prototype new things or meet RAI guidelines?

RESEARCH

- Resourcing
- Funding

KEY TAKEAWAYS

IT'S HARD

This stuff is way harder than it looks!

GET STARTED

Do what you can and document the rest.

WHY?

Remember what you're trying to achieve!

EXPERTS

Drag in experts from other areas to help when you can.

SELECTION

Just pick one - they all overlap anyway.

Thanks!

...Questions?