SYSTEMS ENGINEERING RESEARCH CENTER

AIRC | ACQUISITION INNOVATION RESEARCH CENTER

SERC PERSPECTIVE: ARCHIMEDES WORKSHOP ON TRUSTWORTHY AI

Tom McDermott, SERC CTO, Stevens Institute of Technology
Zoe Szajnfarber, SERC Chief Scientist, the George Washington University

SERC and AIRC Collaborating Universities

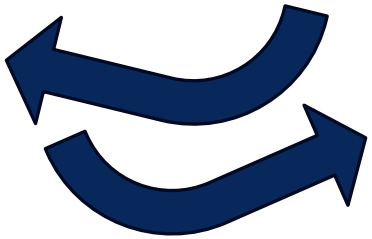
AFIT Air Force Institute of Technology
Carnegie Mellon University
NPS Praestantia Per Scientiam
The George Washington University Washington, DC
Georgetown University
Stevens Institute of Technology
PennState
The University of Arizona
University of Massachusetts Amherst
Massachusetts Institute of Technology
Old Dominion University
University of Virginia
USC
Virginia Tech
Purdue University
Auburn University
Georgia Tech
George Mason University
North Carolina Agricultural and Technical State University
University of Maryland
The Ohio State University
Texas A&M University
The University of Alabama in Huntsville
NC State University
UCF
University of South Florida

Role for Systems Engineers in AI space

AI4SE and **SE4AI**

Focuses on **application of AI in support of systems engineering processes**, enabling enhanced decision-making, optimization, and efficient effort allocation.

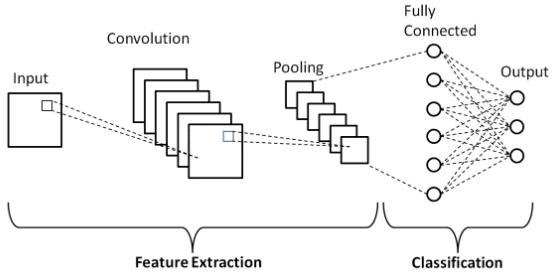
Focuses on **leveraging systems engineering principles to develop AIES that are safe,** robust, and efficient AI systems, while extending them in response to the nature of AI enabled systems.



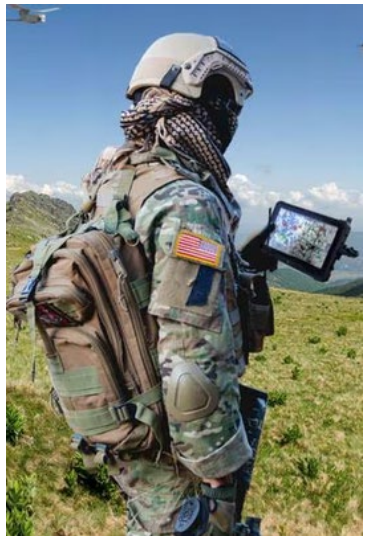
SE4AI applies to AI4SE too, but types of AI tools tend to be different

… and AI4SE might change what SEs do too.

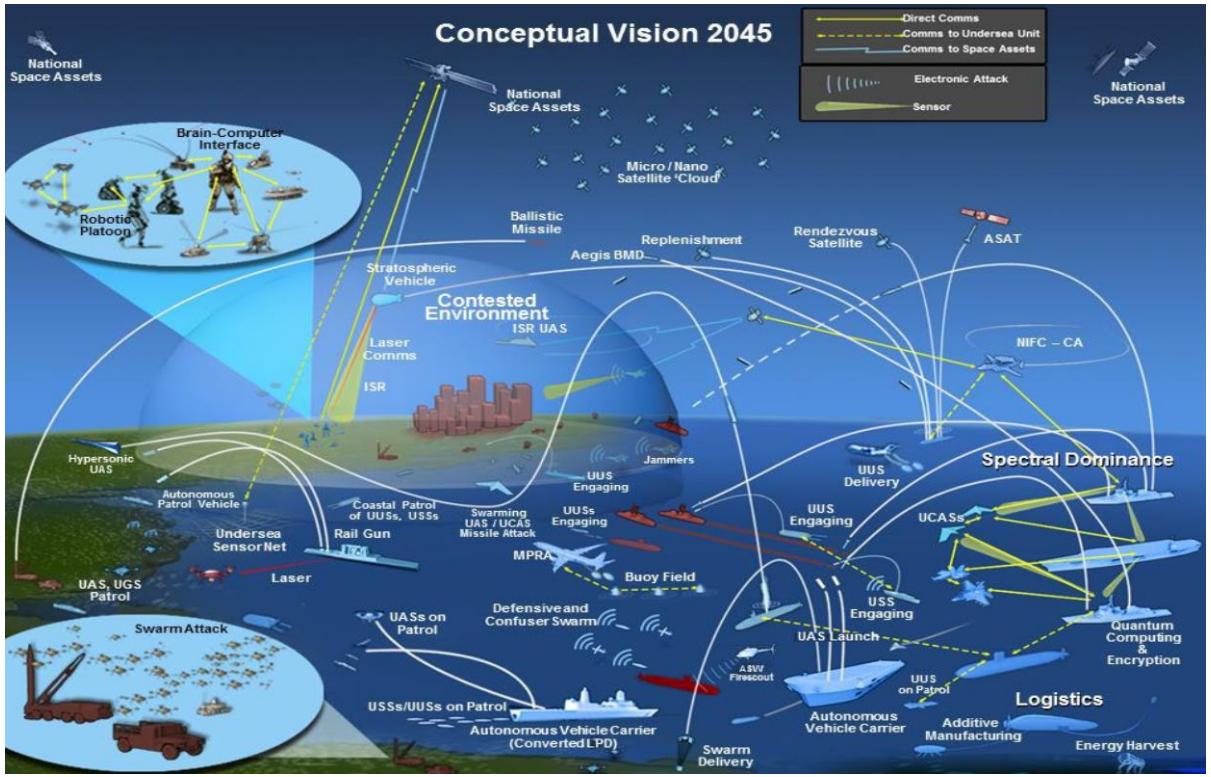
Why do we need SE to ensure safe/robust/trustworthy AIES?



System behavior is not **only** about the algorithm…



… or even one human working with one cognitive assistant…



… it's an embedded socio-technical systems problem, which is an extension of SE practices

ARCHIMEDES WORKSHOP 2024

The Archimedes Initiative was co-founded in 2022 to encourage cross domain SE research

- DLR: mobility systems
- TECOSA: telecom & edge computing systems
- TNO-ESI: hi-tech equipment & manufacturing
- SERC: defense systems

Download the report =>

PART 1: PARTNER PERSPECTIVES ON TRUST+AI

DEFINING TRUST(WORTHY) AI

Trust is by the user and is a property of the relationship.

"attitude that an agent (automation or another person) will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."[1]

Trustworthiness is a property of the artifact.

"ability to meet stakeholders' expectations in a verifiable way; an attribute that can be applied to services, products, technology, data and information as well as to organizations."[2]

Trustworthy AI combines both concepts

emphasizing properties that generate "AI that can *[should?]* be trusted by humans"[3] Those properties typically include valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.[4]

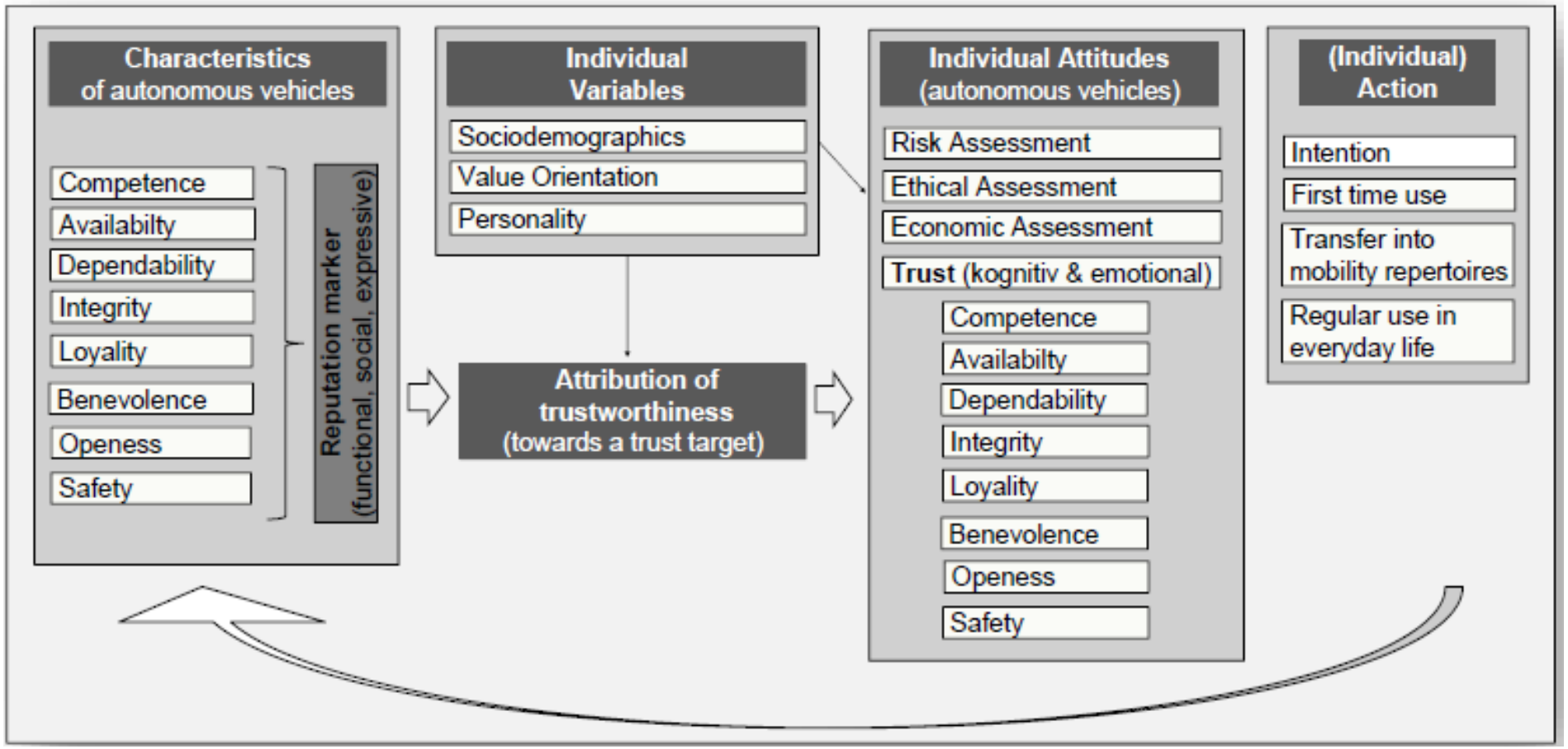
[1]Cited in NIST RMF Glossary: John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **46**(1):50–80, 2004
[2]Cited in NIST RMF Glossary: ISO/IEC_TS_5723:2022(en)
[3]Cited in NIST RMF Glossary: Mark Coeckelberg (2020) " AI Ethics" MIT Press; [4]NIST RMF

Summary

- The role of trust has to be considered increasingly important as “human factor” in systems engineering.
- Key ingredients of trust: abilities, benevolence, integrity and explainability
- Technical and non-technical understanding necessary for implementation
  - Non-technical understanding for defining social and ethical norms
  - Interdisciplinary research to identify corresponding indicators
  - Definition of metrics and sensing mechanisms needed
- More autonomy of systems needs more interdisciplinary research

TRUST IN AUTONOMOUS VEHICLES

"Autonomous Vehicles represent a completely new class of systems."
Axel Hahn, DLR

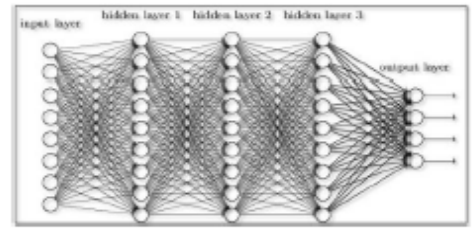


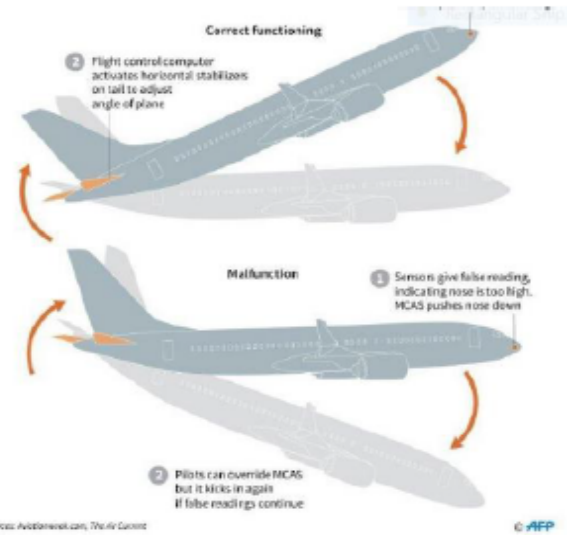
Technology
+
Humanities
+
Social Sciences

What could go wrong in an AI-based CPS?

- **Complexity**
  - Billions of transistors, LOC's and 100's of billions of (DL) parameters, and … thousands of engineers across multiple supply chains and organizations!
- **The world of software and bugs**
  - Industry average code ~ 15– 50 errors /KLOC
  - Safety critical systems ~ 0.1 error/KLOC at very high cost
  - Single event upsets (transient HW errors, bit-flips)
- **Deep learning: breakthroughs but brittleness & explainability**
  - Limited contextualization beyond training data
  - An engineering discipline yet to emerge (M. Jordan, UC Berkeley)
- **Cyber-security threats and attacks**
  - Dynamic threat landscape
- **The billions miles - environment & interaction complexity**
- **Automation surprises and pitfalls**
  - Humans in- and on- the loop

    Lisanne Bainbridge, 1983: Ironies of automation

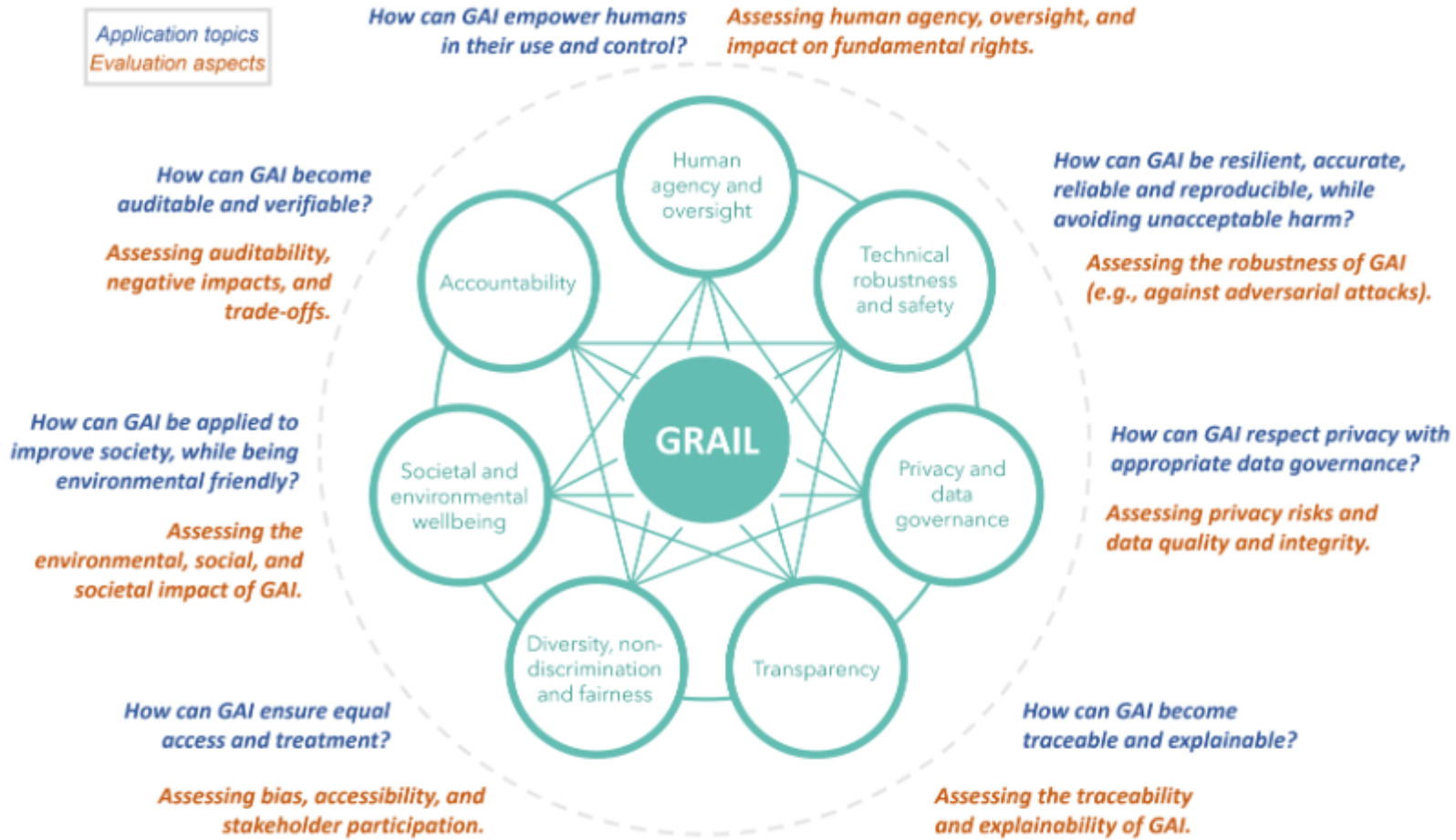




Generative Responsible AI League

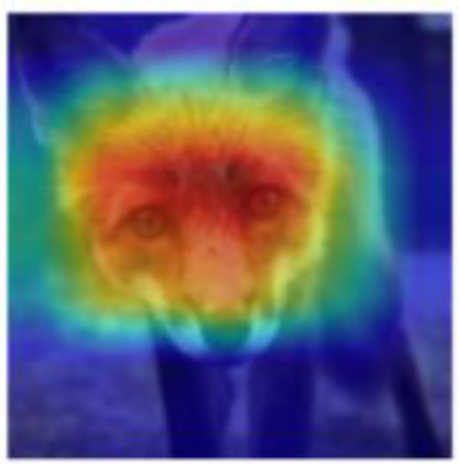
Explicitly addresses the 7 principles

- Application: risk scanning & maximize positive impact
- Evaluation: metrics & assessment tool
- Governance: blueprint based on best practices



WHAT MAKES YOU TRUST (OR NOT TRUST) "THE AI"?

Developer



[1]

Domain Expert

[2]

End User

U.S. Department of Transportation

[3]

Accuracy:

If you're a computer scientist you hate this phrasing, and want to see the math of this specific algorithm or at least a visualization of the prediction.

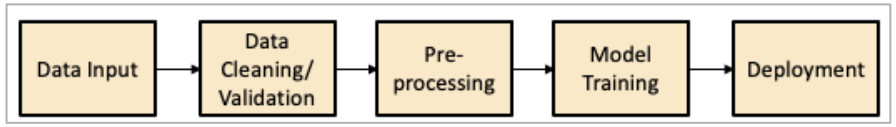
Agrees with me:

If you're a radiologist diagnosing pathology on an image, you might want to see the tool agree with you often enough.

Trusted 3rd Party:

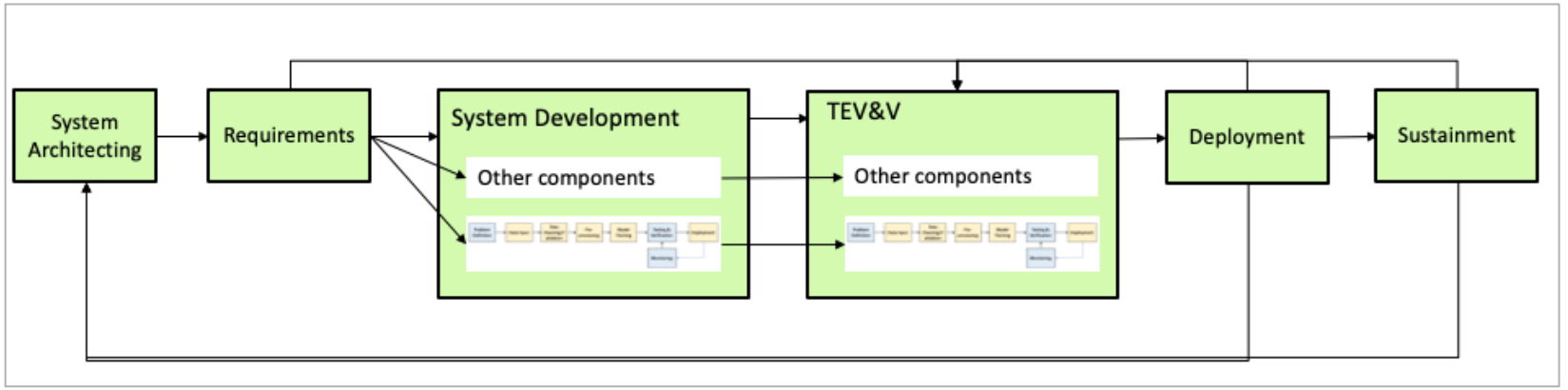
If you're an AV passenger, you might want to be told that someone reputable certified it's safety… and not have heard of any fiery crashes lately!

RETURNING TO TRUST

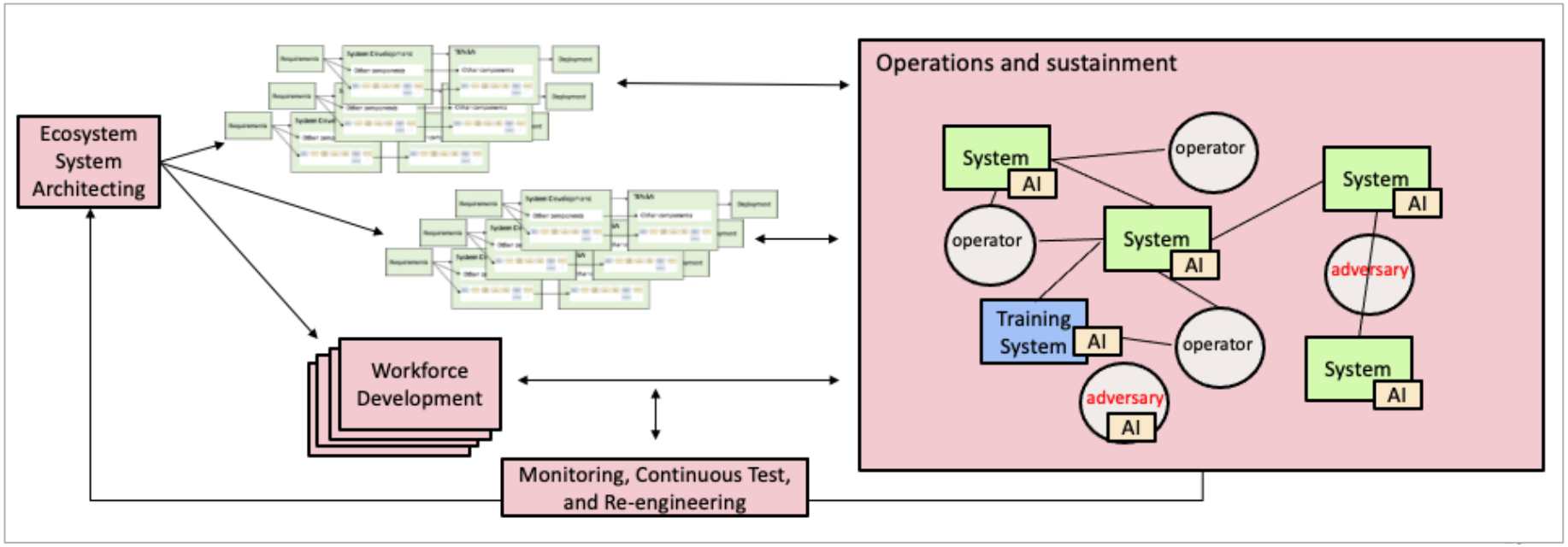


Correct and without bias

Emphasizes tradeoffs in performance and risk; unplanned behavior

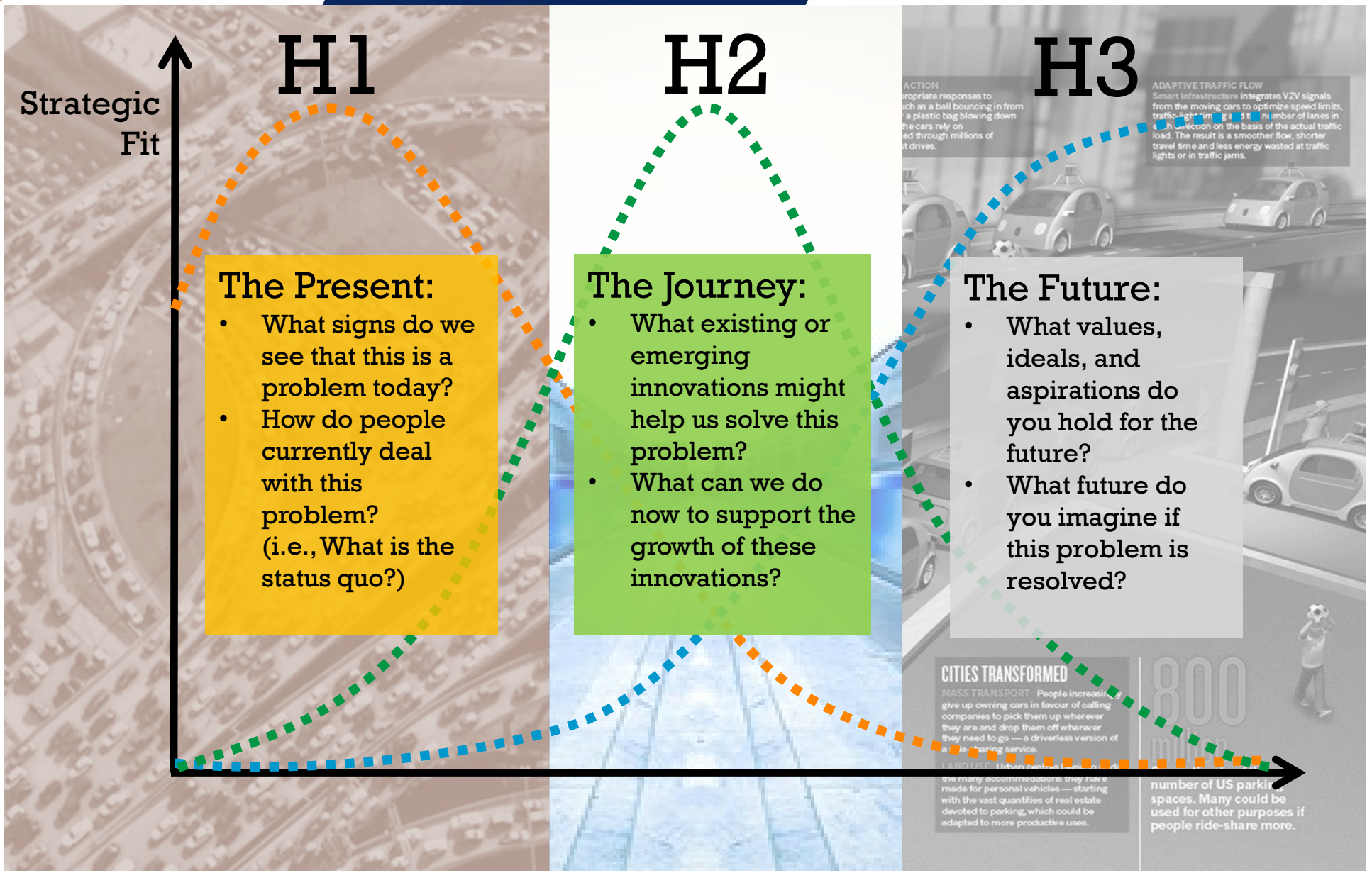


Complex interactions among humans and systems that were not always intended to work together in a constantly changing environment.



PART 2: THREE HORIZONS WORKSHOP

THE THREE HORIZONS



HORIZON 3: ASPIRATIONS

Development Decentralized

- Fewer dominant players
- Beyond wealthy nations
- Decentralized

Produces good outcomes

- Healthier lives
- More interesting work
- Better justice & equality

AI and automation disappear in background

Future perceptions

- Push back on AGI/eminent takeover narrative
- Perceived like water – always there
- Cannot automate human care

AI as human collaborator

- Acts as a valued partner to humans
- Complements human expertise
- Makes humans more productive
- Creates more mid-skill labor
- Super-human capabilities

New governance systems

- Protected labor rights
- More democratic governance
- Balancing regulation & innovation

A revolution in how we learn

- Unlimited access
- Individualized

A revolution in how we work

- Accessible expertise
- More inclusive
- Enables meaningful discourse
- Use of AI for good

Long term Outcomes:

- Healthier Society
- More Productive Society
- More profits to employees
- Help marginalized/at risk groups

HORIZON 2: RESEARCH

Focusing on **trust** forces us to think about the people who interact

- Trust is by the user and a property of the relationship
- Trustworthiness is a property of the artifact

Major trends

- Computing power – chip manufacturing
- New policy/rulemaking
- CoPilots for work tasks
- Healthcare Apps
- Car automation
- AI-enabled "new classes of systems"

Research on AI Trust:

- T&E, testbeds & test ranges
- Measurement of trust
- Hybrid AI/ML tech
- Explainability approaches
- Bias mitigation
- AI oversight of AI
- Digital ID, privacy
- Realtime correctness checks
- Resilience
- Assurance, Assured
- Security not vulnerability
- Risk assessment
- Trust as a Human Factor
- Benevolence
- Moral models for AI safety

Building Trust:

- Awareness of beneficial use and risks
- Report mistakes/remedies
- Accessible opt-outs
- Community stakeholders in the development
- "no-AI" bootcamps
- Economics incentivize responsible AI
- Transparency & reporting of training data
- Energy consumption labels
- Start with analysis, then assisted, then augmented, then autonomous
- A "dial" that supports trust
- "Public models" create common understanding

The Ecosystem of Trust

- Achieves everyday use
- Human- centric
- Explainable
- Regulated vs. unregulated
- Deterministic
- Accountability
- Organizational trust
- Social standards for "integrity"
- Hold developers accountable

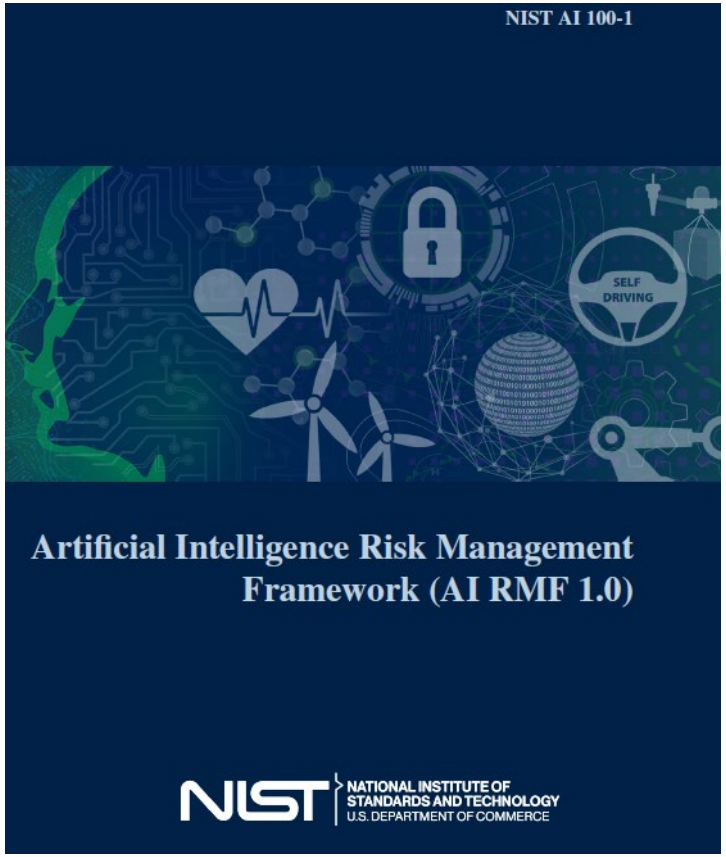
Recurring themes:

- Large number of frameworks and terms, not always applied consistently, especially working across US and Europe.
- The interactions of AI and humans will create new classes of systems that require a more interdisciplinary (socio-technical) perspective to design and manage them.
- Important role for Systems Engineers in this discussion, particularly in terms of metrics, methods and testbeds

Next steps: collaborate on SE research roadmap for TAI…

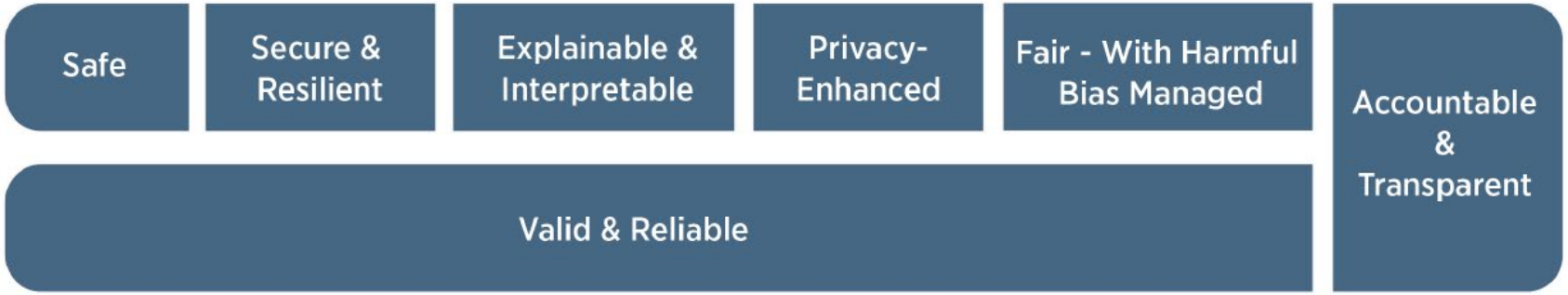
Back-up

NIST AI RMF 1.0



"Responsible AI is meant to result in technology that is also equitable and accountable."

Characteristics of trustworthy AI systems.



THE AI ACT (EU)

The AI Act: The Main Operational Elements High-Risk AI systems

New Legislative Framework (NLF) Product Safety Legislation +

↓

Sets

Mandatory Requirements for high-risk AI system before they can be used

↓

To address AI specific risks triggered by AI characteristics, such as, Complexity, Opacity, Unpredictability, Autonomy and Data

risks to health, safety and fundamental rights

1. **risk management system** for AI systems *[Art. 9 AI Act]*
2. **governance and quality of datasets** used to build AI systems *[Art. 10 Data and data governance]*
3. **record keeping** - built-in logging capabilities in AI systems *[Art. 11 Technical documentation and Art. 12 record-keeping]*
4. **transparency and information** to the users of AI systems *[Art. 13 Transparency and provisions of information to users]*
5. **human oversight** of AI systems *[Art. 14 Human oversight]*
6. **accuracy** specifications for AI systems *[Art. 15 Accuracy, robustness and cybersecurity]*
7. **robustness** specifications for AI systems *[Art. 15 Accuracy, robustness and cybersecurity]*
8. **cybersecurity** specifications for AI systems *[Art. 15 Accuracy, robustness and cybersecurity]*
9. **quality management system** for providers of AI system *[Art. 17]*
10. **conformity assessment** for AI systems *[Art. 19 + Art. 43 Conformity Assessment]*

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Courtesy of Tatjana Evas, European Commission

TRUSTWORTHY AI: GENERATIVE RESPONSIBLE AI LEAGUE (TNO)

Application topics
Evaluation aspects

How can GAI empower humans in their use and control?

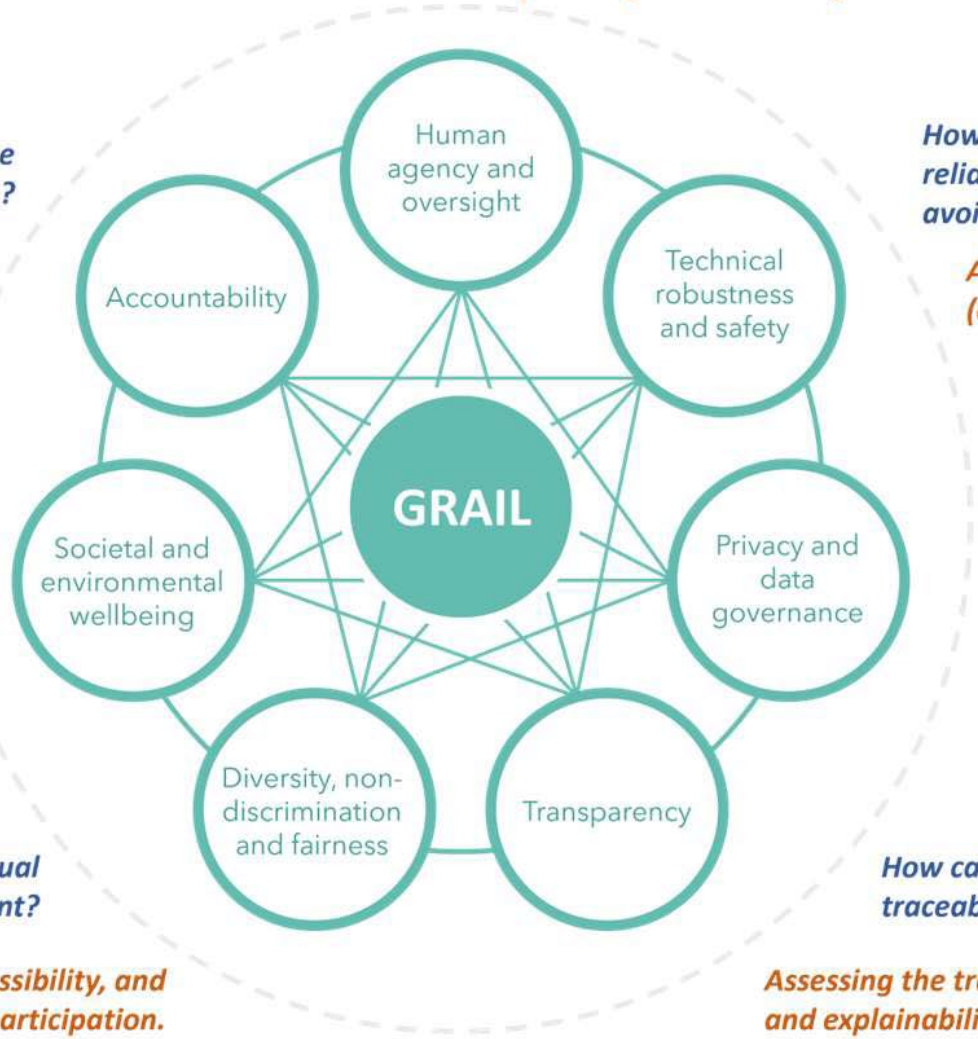
Assessing human agency, oversight, and impact on fundamental rights.

How can GAI become auditable and verifiable?

Assessing auditability, negative impacts, and trade-offs.

How can GAI be resilient, accurate, reliable and reproducible, while avoiding unacceptable harm?

Assessing the robustness of GAI (e.g., against adversarial attacks).



How can GAI be applied to improve society, while being environmental friendly?

Assessing the environmental, social, and societal impact of GAI.

How can GAI respect privacy with appropriate data governance?

Assessing privacy risks and data quality and integrity.

How can GAI ensure equal access and treatment?

Assessing bias, accessibility, and stakeholder participation.

How can GAI become traceable and explainable?

Assessing the traceability and explainability of GAI.

TAILOR
The TAILOR Handbook of Trustworthy AI