The Nature, Challenges and Diversity of AI Assurance

Dr Greg Chance

Consultant, Digital Systems Assurance, Frazer-Nash Consultancy

Honorary Research Fellow, Trustworthy Systems Lab, University of Bristol

Document ref: 143746V

26-27th March 2024

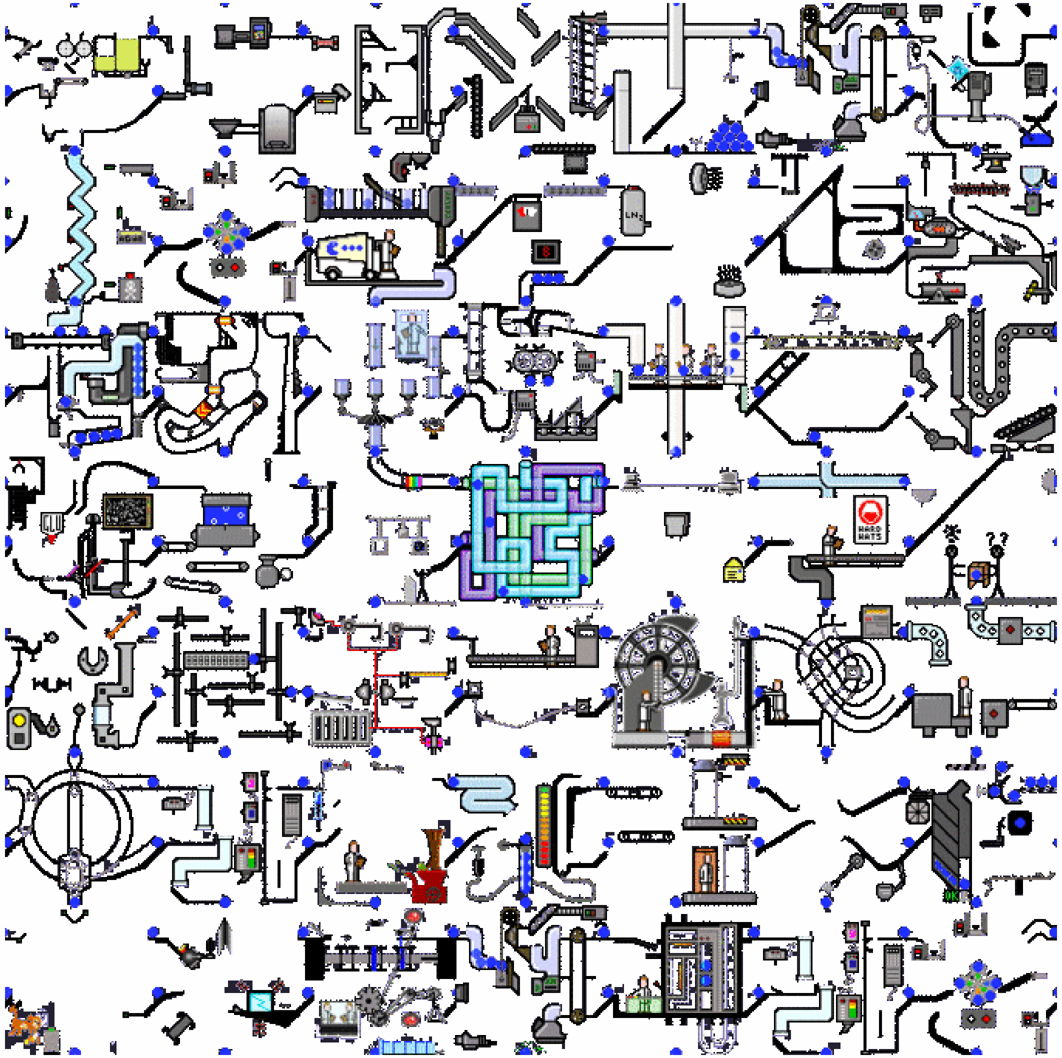
Security Classification

Agenda

- What is AI Assurance?
- Assurance Techniques
- Trust and Trustworthiness
- Final Thoughts

What is AI Assurance

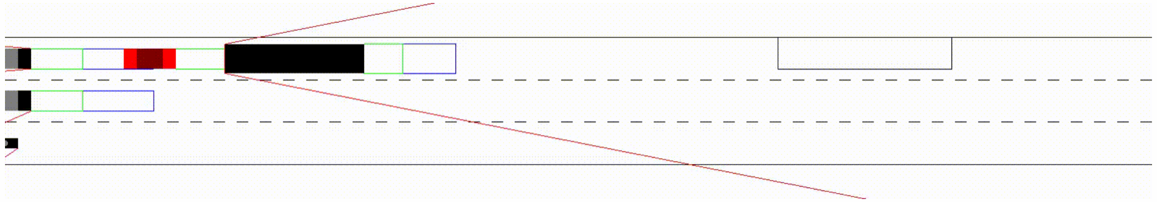
- What is AI Assurance?
  - A **proof** of a system design or property – a positive **declaration** of **certainty**
  - *Confidence in the correctness of a system*
  - Systems we can **Trust**
- How can we gain confidence in the system?
  - **Design** simple systems are understandable,
    - Design not retrofit
  - **Transparency** gives insight into decisions and behaviours



What is AI Assurance

- What is AI Assurance?
  - A **proof** of a system design or property – a positive **declaration** of **certainty**
  - *Confidence in the correctness of a system*
  - Systems we can **Trust**
- How can we gain confidence in the system?
  - **Design** simple systems are understandable,
    - Design not retrofit
  - **Transparency** gives insight into decisions and behaviours
  - **Verification and Validation** rigorous sub-system proof, simulation-based testing and advanced test generation methods, high level of automation

Simulation: Carla 3D Environment – Junction Safety Test



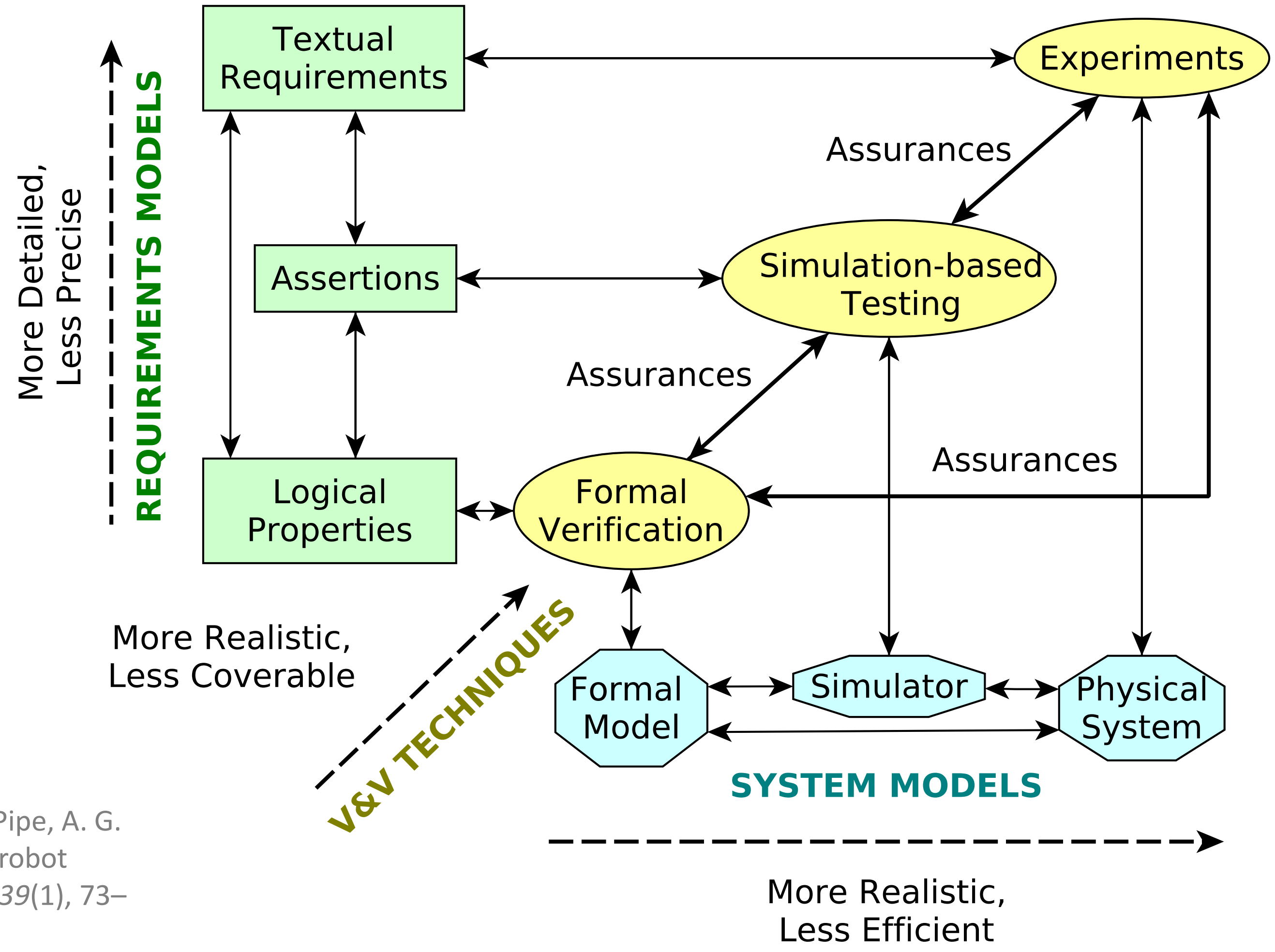
Test Generation: Gym Environment – Bus Stop Test

Chance, Greg, et al. "An agency-directed approach to test generation for simulation-based autonomous vehicle verification." *2020 IEEE International Conference On Artificial Intelligence Testing (AITest)*. IEEE, 2020.

Assurance Techniques

How can we gain confidence in the system?

In practice we need to gather **mutually consistent** evidence using a **variety** of verification **techniques** because there is no single approach to verify an entire design



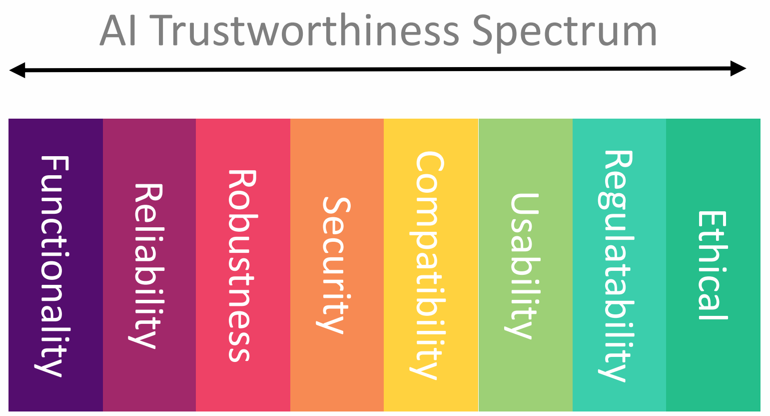
Webster, M., Western, D., Araiza-Illan, D., Dixon, C., Eder, K., Fisher, M., & Pipe, A. G. (2020). A corroborative approach to verification and validation of human–robot teams. [arXiv:1608.07403](https://arxiv.org/abs/1608.07403) *The International Journal of Robotics Research*, *39*(1), 73–99. https://doi.org/10.1177/0278364919883338

Assurance Techniques

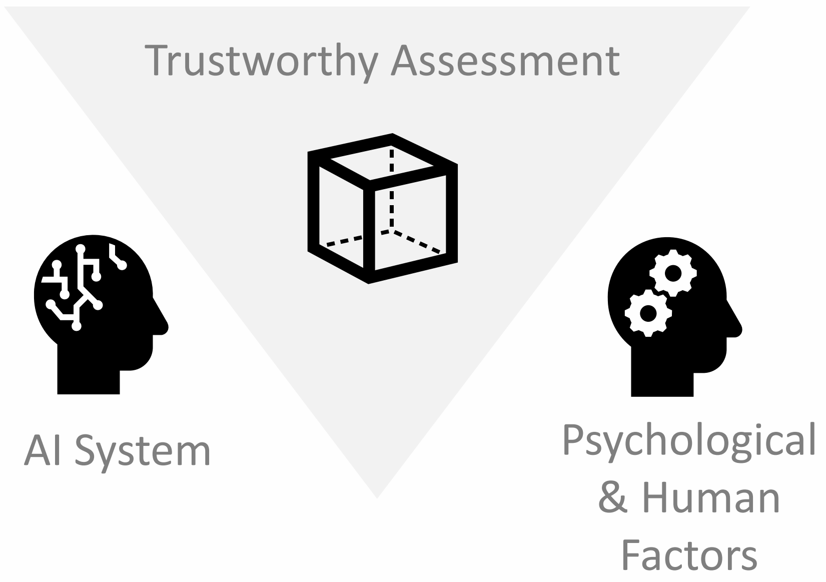
- What if the design is too complex, tools are inappropriate, or the environment too varied?
- Assuring Autonomous Vehicles is a good example
  - AI control system is highly complex
  - Tools inappropriate, e.g. unseen data issue
  - Environments are varied, high dimensional
    - Roads in central London
- Better to constrain environment, scope etc.
  - Parking shuttle Heathrow Car Park
  - Much more trustworthy!

Trust & Trustworthiness

- **Trust** in AI system required for adoption
- Trust is a **diverse spectrum** of qualities
- Part of assessment must account for **the user**



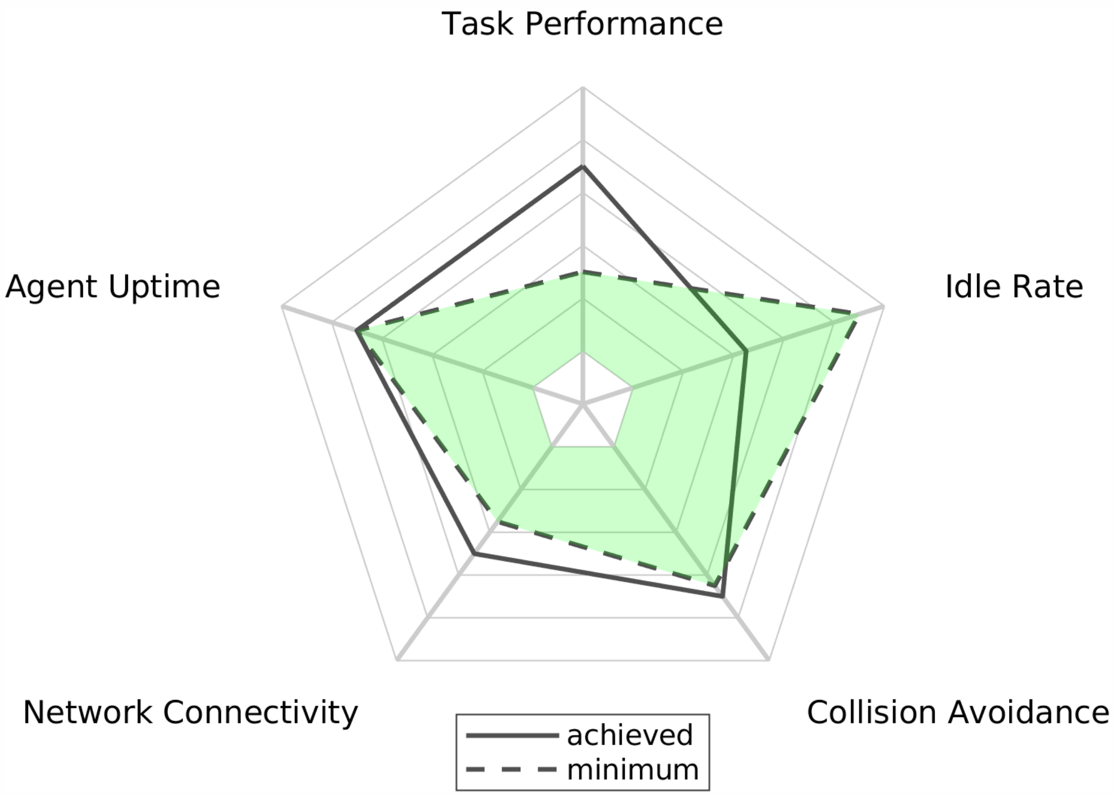
- **Functionality**: To prevent system failure or faults and maintain liveness.
- **Reliability**: To perform specified functions in a consistent manner.
- **Robustness**: To overcome adverse conditions and be maintained or modified.
- **Security**: Protection from subversion, forced failure or malicious use; and maintaining confidentiality, availability, accountability, authenticity and integrity.
- **Compatibility**: To exchange information, be able to transfer to other shared environments and to share the environment with other autonomous agents.
- **Usability**: To be available and responsive to achieve specified goals in a specified context with effectiveness and satisfaction.
- **Regulatability**: To be verifiable, readable, explainable, transparent, understandable and to support ease of verification and regulation.
- **Ethical**: To demonstrate fair and reasonable behaviour, beneficence, non-maleficence, preserve human autonomy and be easily understood.



Chance, G., Abeywickrama, D. B., LeClair, B., Kerr, O., & Eder, K. (2023). Assessing Trustworthiness of Autonomous Systems. arXiv preprint arXiv:2305.03411.

Trust & Trustworthiness

- **Criticality**
  - Harm from failure (physical, psychological etc.)
  - Vulnerable to violating trust
- **Automation Scope**
  - Ambition of the AI
  - Autonomous Vacuum or AV?
- **Authority Level & Decision Making**
  - Correct authority
  - Decision making level correct
- **Stakeholder risk**
  - Risk appetite
  - Failure mode
- **Metrics**
  - Monitor trust



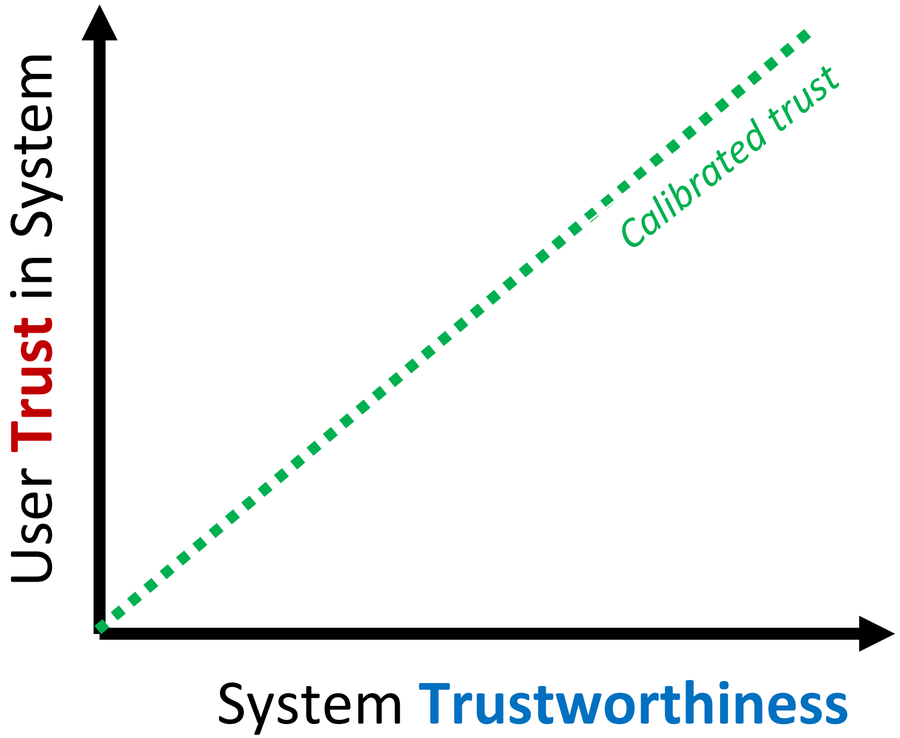
Metrics for an automated swarm robot agent

Chance, G., Abeywickrama, D. B., LeClair, B., Kerr, O., & Eder, K. (2023). Assessing Trustworthiness of Autonomous Systems. arXiv preprint arXiv:2305.03411.

Trust and Trustworthiness

FRAZER-NASH CONSULTANCY
A KBR COMPANY

Trust = response of a user in a situation of uncertainty or vulnerability



Calibrated Trust = User **Trust** is commensurate with the **Trustworthiness** of the system which leads to:

- higher adoption rate
- appropriate use
- utilising the capability

Trustworthiness = measure of trust qualities in the AI system

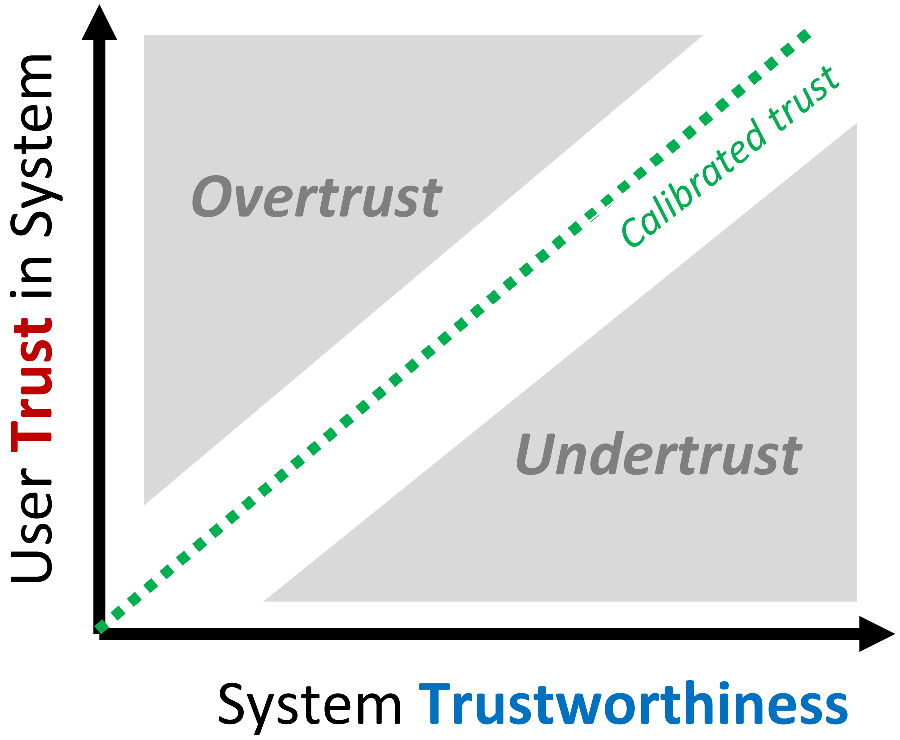
Sullins, J. P. (2020). Trust in robots. *The Routledge Handbook of Trust and Philosophy*, 313–225.

Trust and Trustworthiness

FRAZER-NASH CONSULTANCY
A KBR COMPANY

Overtrust Trust in the system is greater than the system can deliver:

- Violation in functionality, safety or critical assumptions
- Inappropriate reliance on AI
- Taking inappropriate or misguided action



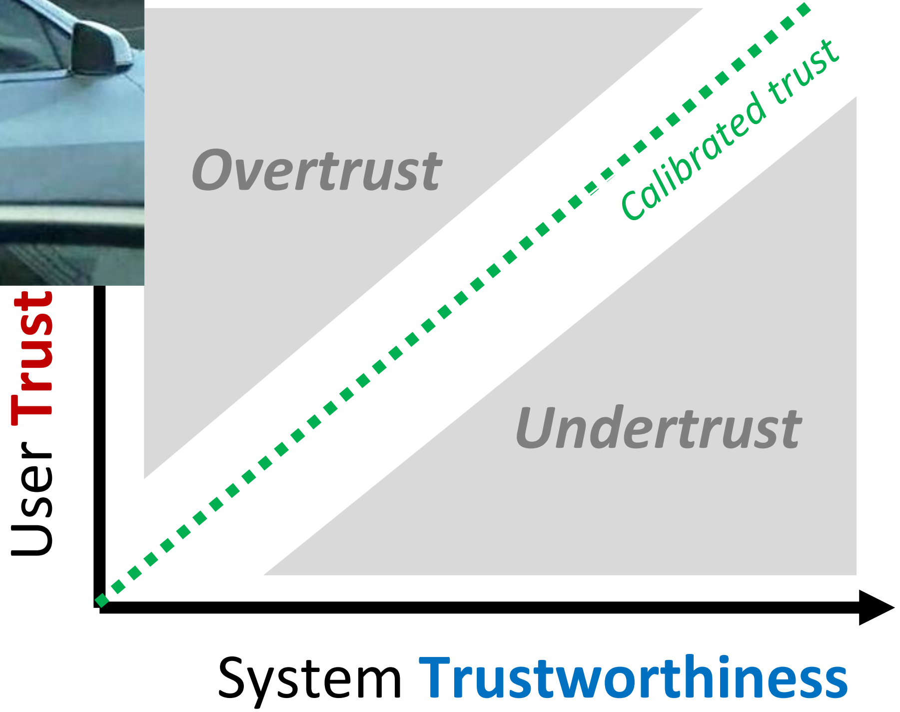
Undertrust System performs better than supervisor allows for:

- User defers to preexisting beliefs
- Taking alternative, contrary or abortive action
- Reject capability

Trust and Trustworthiness

Overtrust Tesla Model S60 driver in passenger seat on M1 using autopilot

Overtrust Elain Herzberg killed by self-drive Uber with safety driver

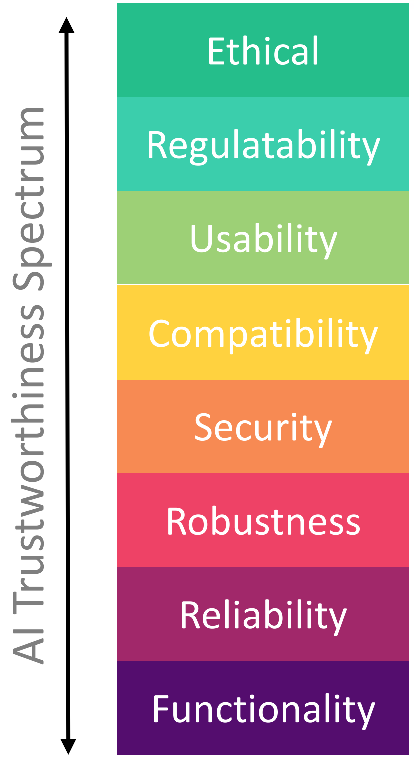


Undertrust in windshear alert system, 37 fatalities (USAir Flight 1016, 1994).

Final Thoughts

I hope you have learnt how we can build and **assure AI** systems and tools!

- Simple designs, using automated V&V helps build **confidence in correctness** of AI systems
- **Assurance techniques**, formal, simulation and physical
  - Scope definition & constraint
- **Calibrate** user Trust with system Trustworthiness
- Trustworthiness is a **spectrum of properties**
  - Design for **robustness**, build for **usability**
  - Best practice for **security** and **cybersecurity**
  - Understand the **standards and regulations** for AI systems in this sector
  - **Demonstrate to** regulators with **accessible evidence** and **explainable logic**
  - Understanding **ethical issues** and demonstrating acceptable behaviour



Thank you

Dr Greg Chance
g.chance@fnc.co.uk

Digital Systems Assurance
Frazer-Nash Consultancy