FRAZER-NASH CONSULTANCY
A KBR COMPANY

Trustworthy AI: An operator's view…

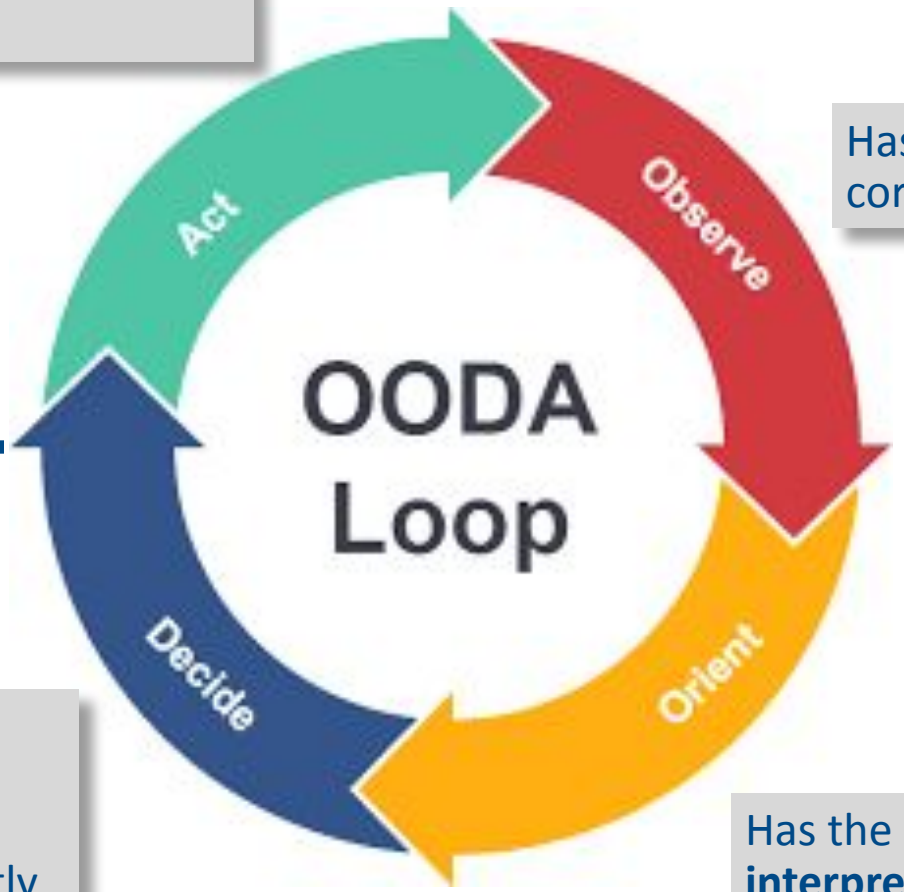
- 'Black Box' Decision Making – the challenge of assurance
- Calibrated trust – what happens when it goes wrong?
- Trustworthy AI – an enabler

SYSTEMS · ENGINEERING · TECHNOLOGY

Ben Keith
Ships Business Manager

Decision Making in a System of Systems

Based on a calibrated and appropriately supervised system, is the **Action** correct?



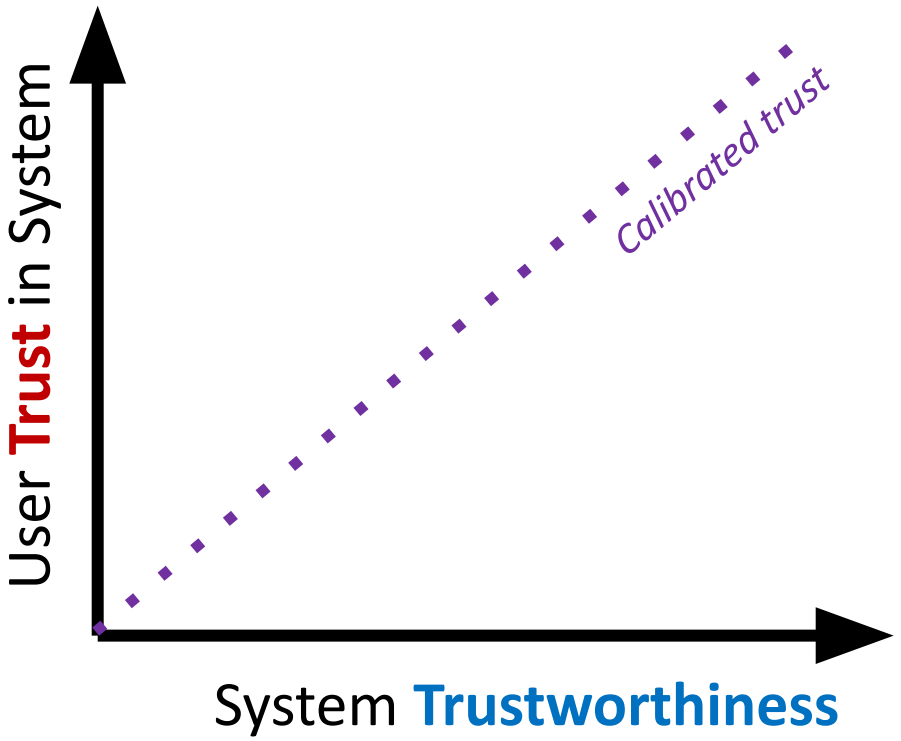
Has the system **observed** the correct things?

Decision based on *trust* in Observe and Orient phases

Level of trust has to be *calibrated* correctly – and *supervised* accordingly

Has the system made the correct **interpretation** of the situation based on what it has observed?

Trust and Trustworthiness



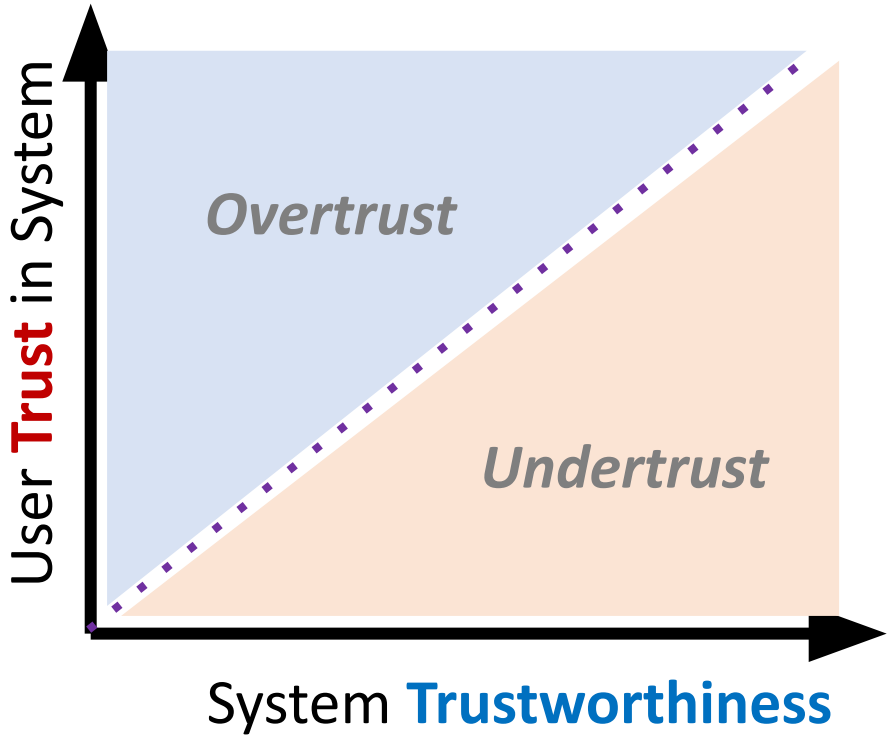
Trust = response of a user in a situation of uncertainty or vulnerability. ***Subjective***

Trustworthiness = measure of trust qualities in a system (autonomous or AI). ***Objective***

User **Trust** must be commensurate with the **Trustworthiness** of the system (well calibrated)

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

Trust and Trustworthiness



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

- Over-reliance on AI/automation
- Taking inappropriate or misguided action

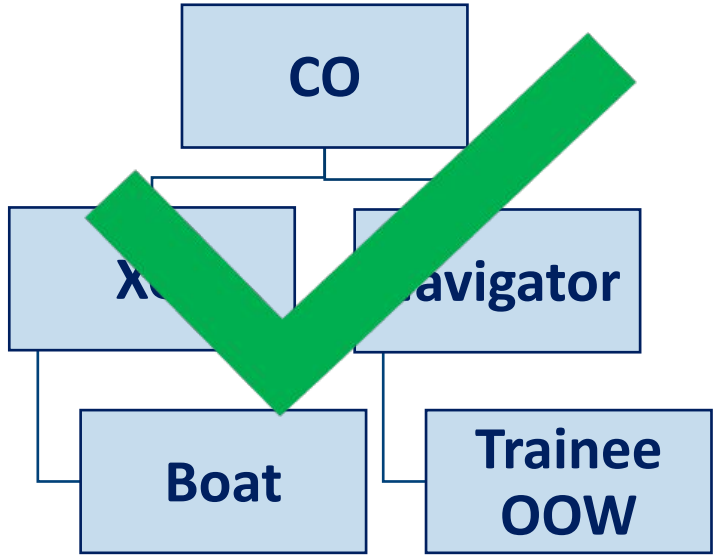
Undertrust. System performs better than supervisor allows for:

- Supervisor ‘knows better’
- Taking alternative, contrary or abortive action

Human example – HMS DIAMOND, Akrotiri Bay 2018

Trust/Trustworthiness

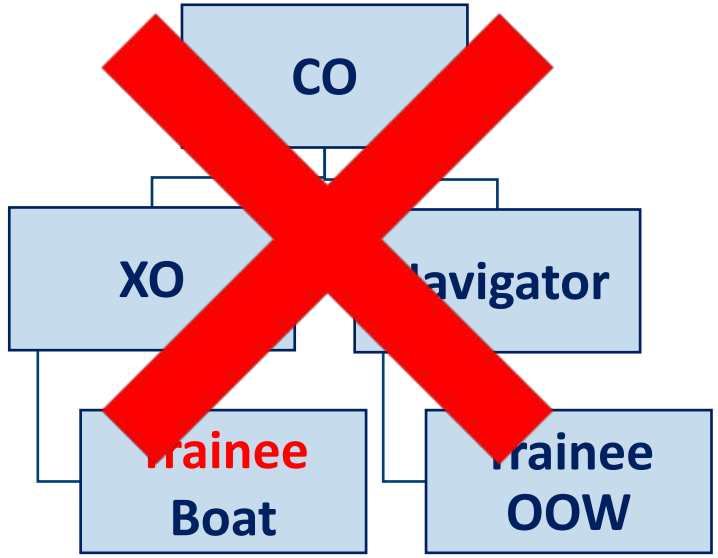
System of Systems



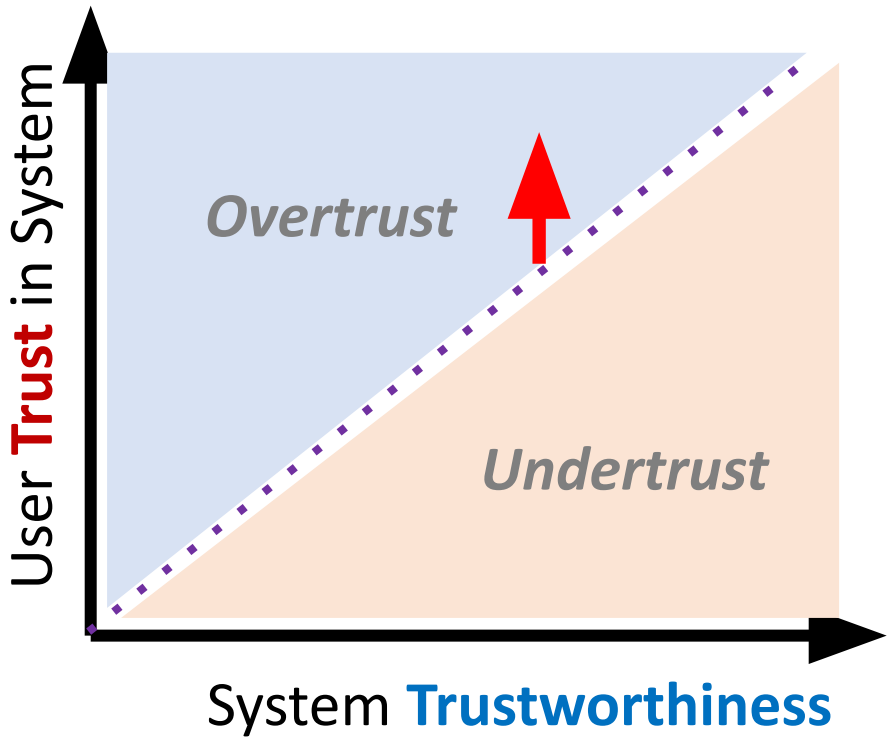
Human example – HMS DIAMOND, Akrotiri Bay 2018

Trust/Trustworthiness

System of Systems

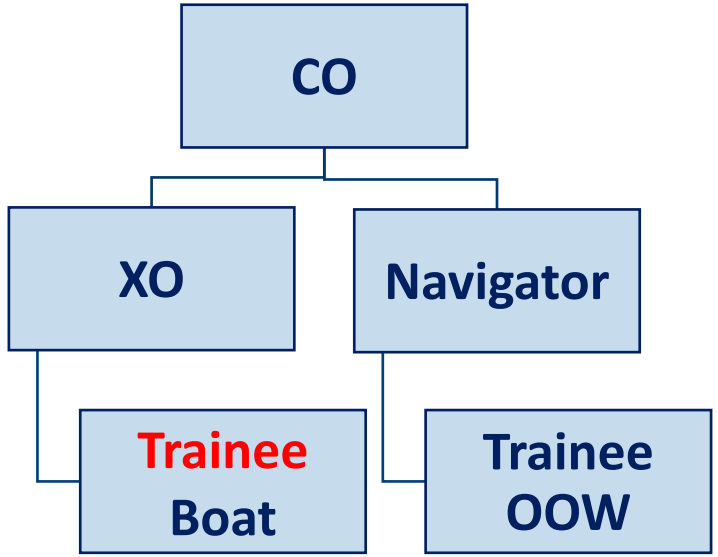


Trust and Trustworthiness

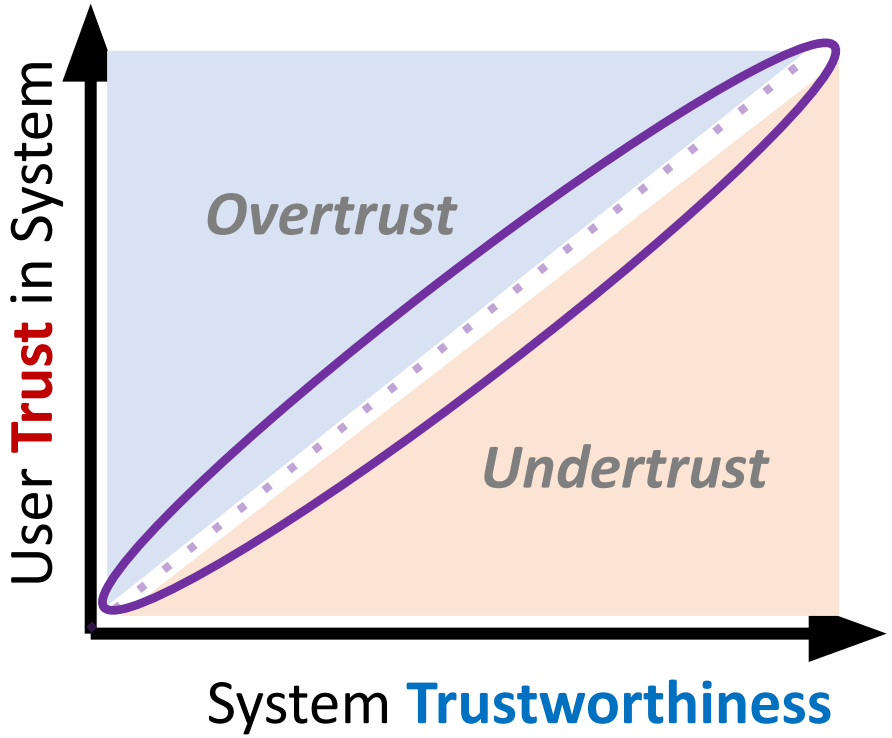


Moved from calibrated to miss-calibrated trust = ***overtrust***

Human System: *Calibrated trust* mix of intuition, experience, qualification, external validation

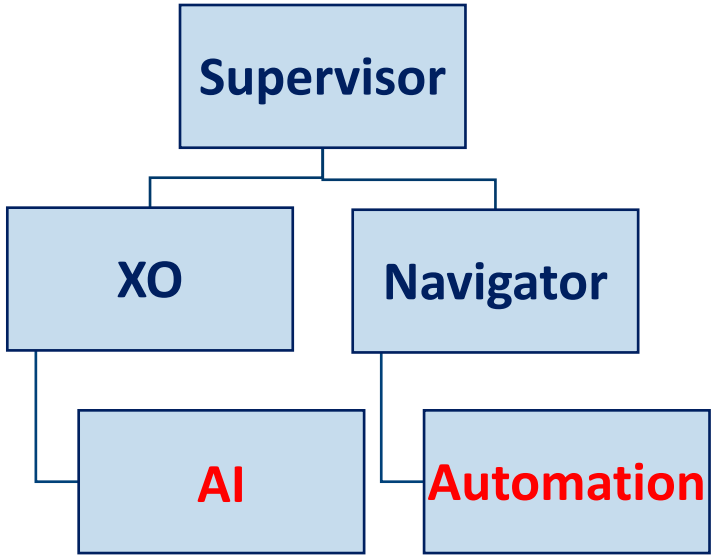


Trust and Trustworthiness

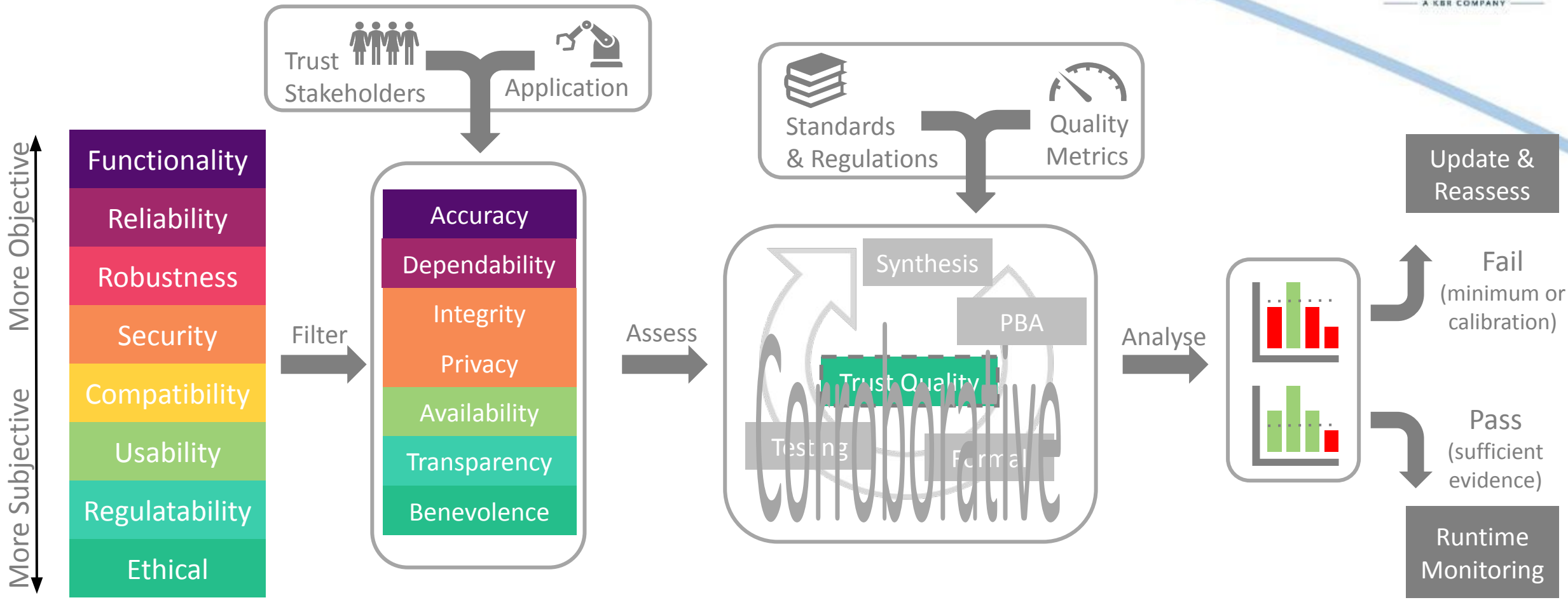


AI Applicable lessons:

- Increase automation in the system – *calibrated trust* vital
- **But** harder to define the line = area
- Need to measure *System Trustworthiness*



Assuring AI / Autonomy: Frazer-Nash Research and Development



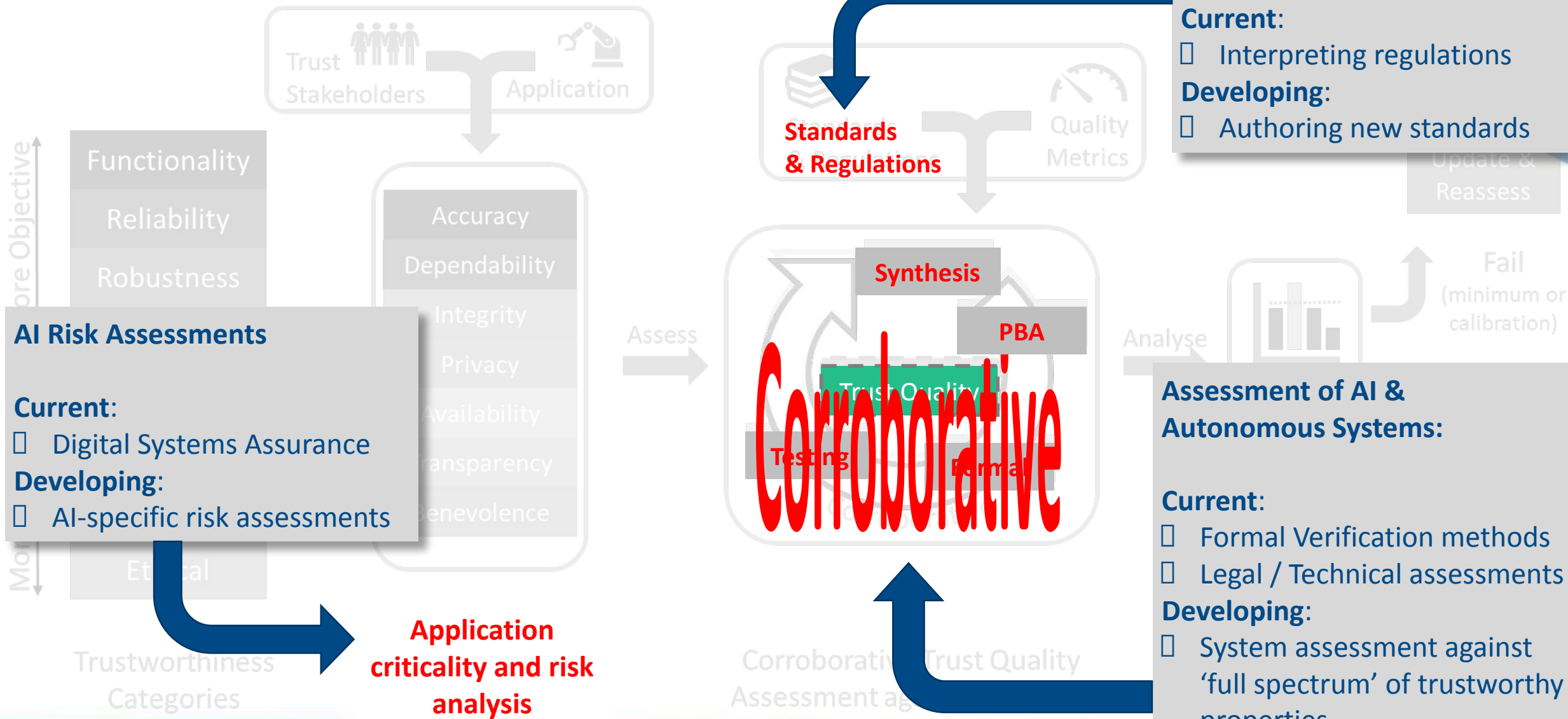
Trustworthiness Categories

Application criticality and risk analysis

Corroborative Trust Quality Assessment against Standard

Quality Assessment & User Calibration Analysis

Trustworthiness as an 'AI Enabler'



AI Risk Assessments

Current:
- Digital Systems Assurance

Developing:
- AI-specific risk assessments

Standards and Regulations

Current:
- Interpreting regulations

Developing:
- Authoring new standards

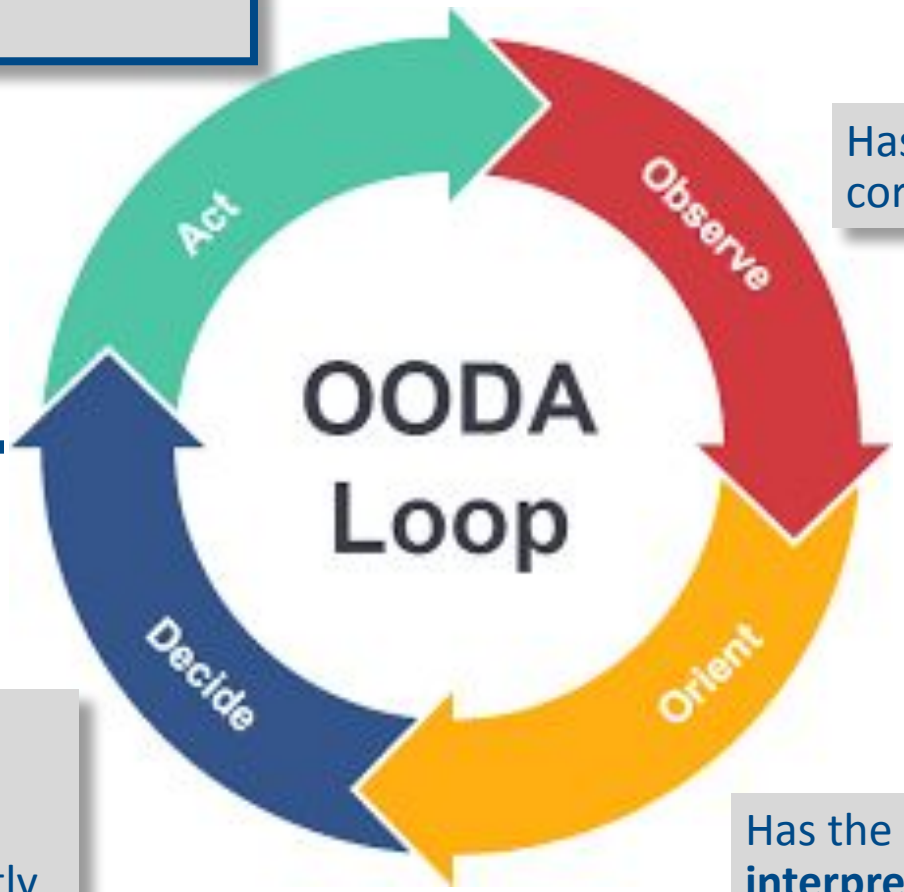
Assessment of AI & Autonomous Systems:

Current:
- Formal Verification methods
- Legal / Technical assessments

Developing:
- System assessment against 'full spectrum' of trustworthy properties

Decision Making in a System of Systems

Based on a calibrated and appropriately supervised system, is the **Action** correct?

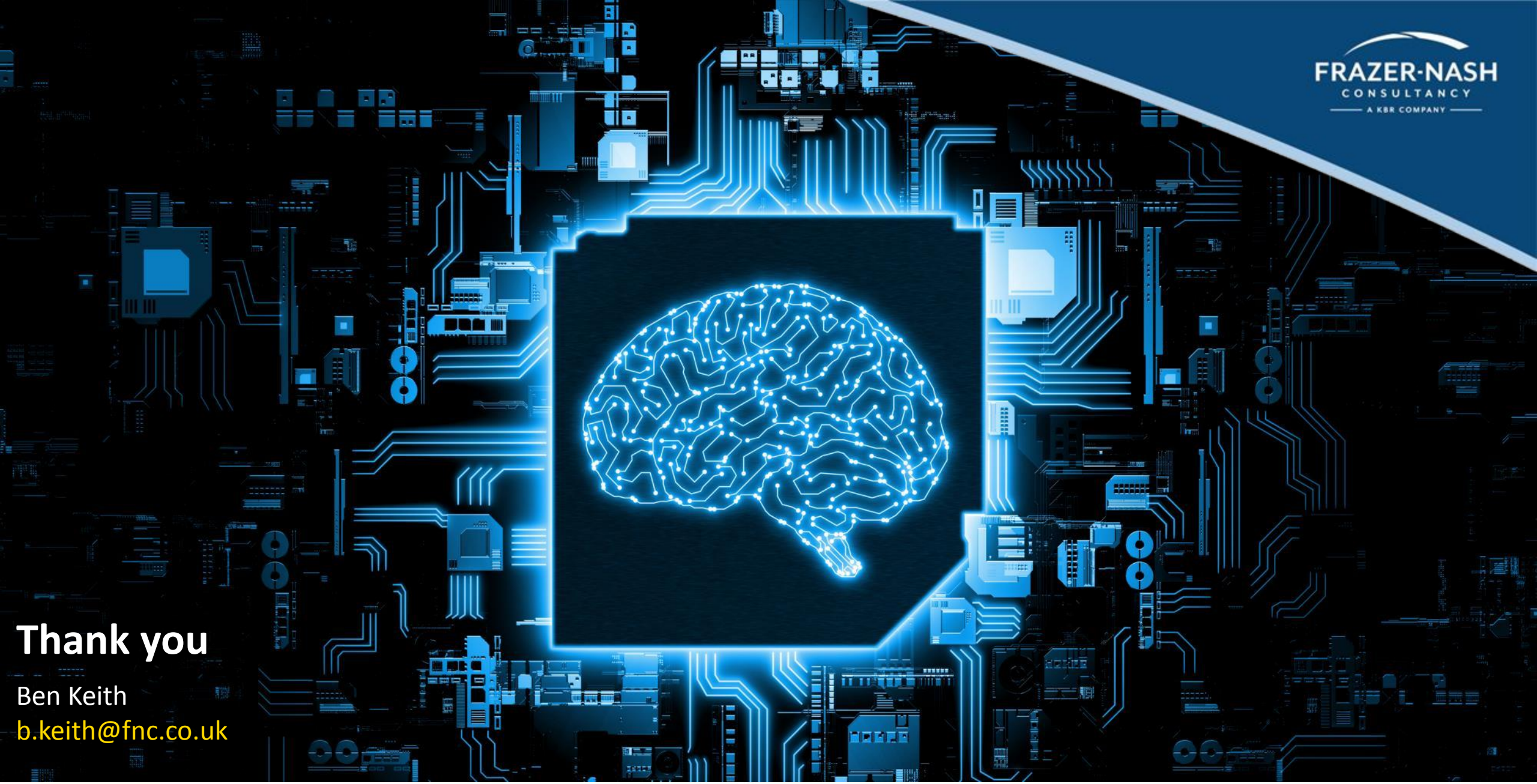


Has the system **observed** the correct things?

Decision based on *trust* in Observe and Orient phases

Level of trust has to be *calibrated* correctly – and *supervised* accordingly

Has the system made the correct **interpretation** of the situation based on what it has observed?



Thank you

Ben Keith
b.keith@fnc.co.uk

FRAZER-NASH CONSULTANCY
A KBR COMPANY

SYSTEMS · ENGINEERING · TECHNOLOGY