Trusting Autonomous Machine Intelligence

Natalia Alexandrov

NASA Langley Research Center

AIAA SciTech Forum 360 Panel *Machine Intelligence and Autonomy Meet Aviation*, January 2021

n.alexandrov@nasa.gov

Fundamentals

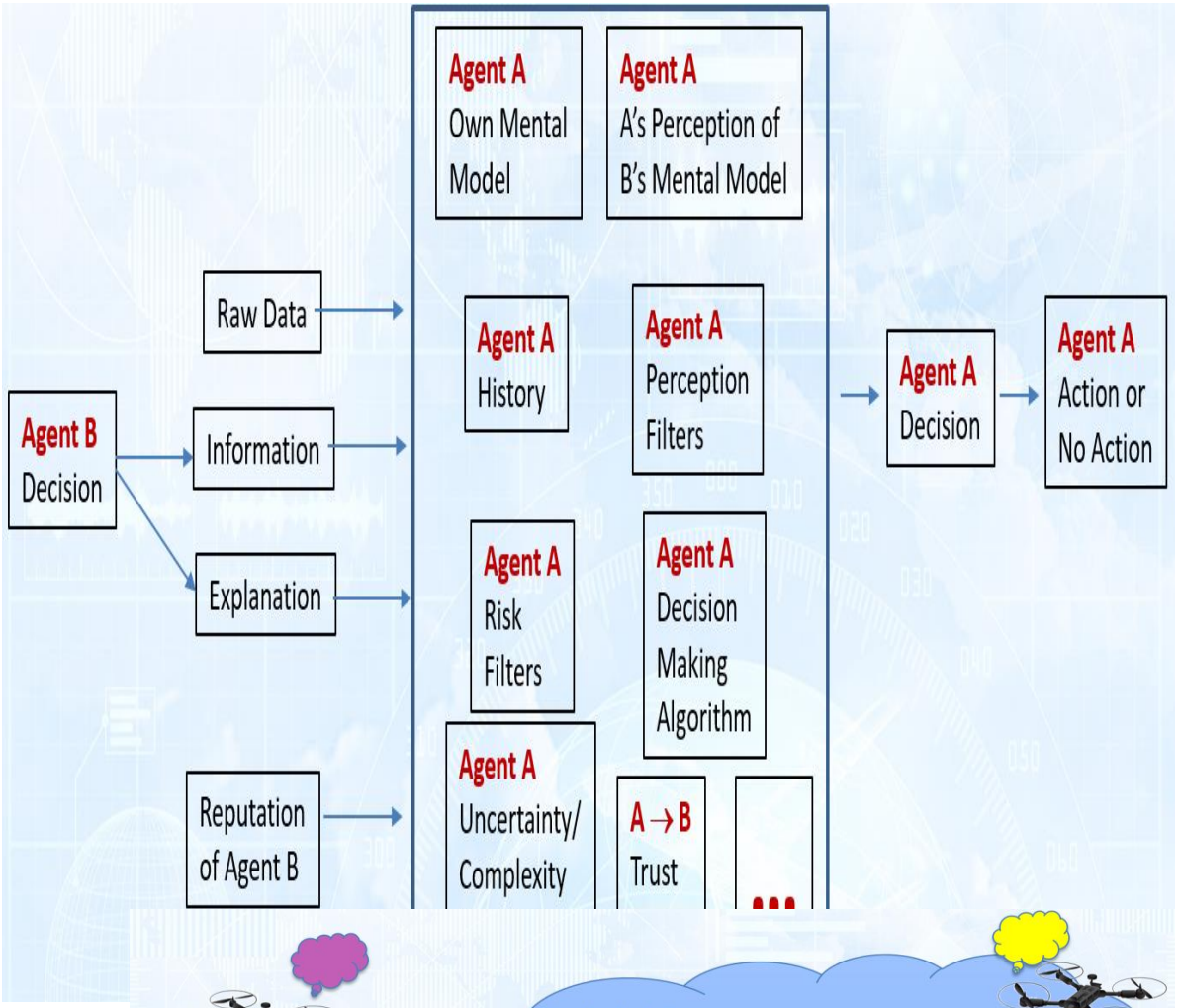
- Autonomy = decision-making and action without external control, regardless of the nature of the agent. Popular use: autonomy = machine autonomy.
- Authority for autonomy
  - Granted by a controlling agency (cooperative autonomy) or
  - Assumed (non-cooperative autonomy)
- Grounds for granting autonomy: Trust vs. Trustworthiness
  - Professor: "Who is willing to fly from Boston to LA on an airplane with no human pilot?"
  - Class: 
  - Professor: "What if the ticket cost $50?"
  - Class: 
- Trust is multi-objective and not strictly rational.
- Claim: ***Rational*** trust must be granted to ***trustworthy*** systems.

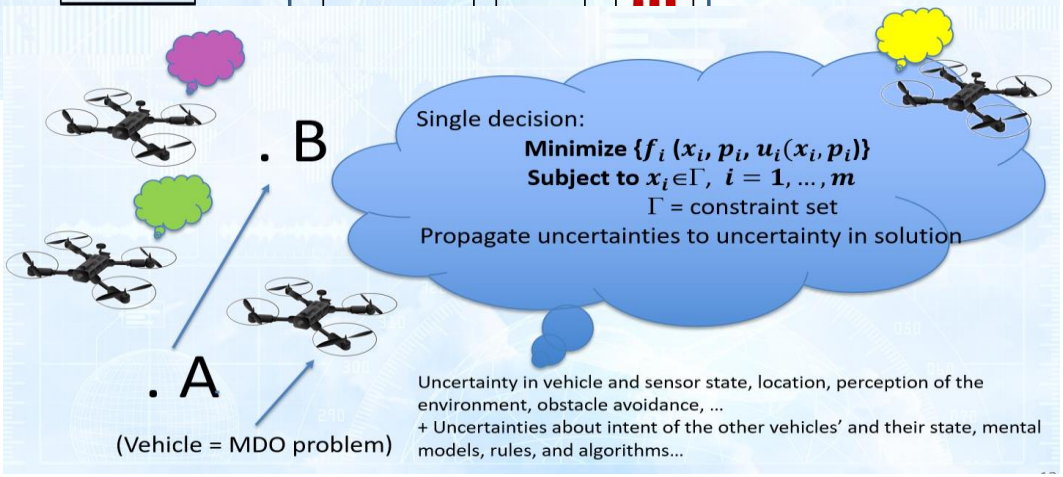
Trusting Full Machine Autonomy

- Control and Trust/Trustworthiness
  - $Control \sim \frac{1}{Trust} \sim \frac{1}{Trustworthiness}$
  - $Trustworthiness \sim \left(\frac{1}{Risk}, thresholds\right) \sim \left(\frac{1}{Uncertainty \times Impact}, thresholds\right)$
- Consequences:
  - Must have comprehensive risk/uncertainty estimation of decision-making, both for system design and real-time operations
  - Thresholds are subjective
- Functionally, a system is trustworthy if it has robustness, resilience, reliability, safety, efficiency, integrity, ability to coexist with non-cooperative agents …
  - The concepts are clear but a lot of work to map onto aviation domain
  - Context is everything

Technical Problems

- Big Problem 1: Implicit vs. Explicit
  - Human control involves many implicit and unknown mechanisms
  - Machine control requires explicit representation of mechanisms and uncertainties
- Big Problem 2: Heterogeneity
- Big Problem 3: Quality of Information for Decision-Making
  - Perception
- Big Problem 4: Complexity and Tractability
- Big Problem 5: V&V
  - Machine Learning – lack of correctness criteria and "safe envelopes"
  - Online Learning





Solution Directions

- Context: Physics-based; goal-based; not looking for AGI
- Math: Formalization of correctness and bounding for ML
- Control by complexity bounding: Relax as trustworthiness grows
- Formalization of decision making as optimization + comprehensive UQ
- New work on survivability in the air and on the ground
- Necessity: Which domains really need autonomy?

(Ad: NASA/ARMD/TACP/CAS: ATTRACTOR session, 20 January)