An Optimization Perspective on Trustworthiness and Trust in Autonomous Systems

Natalia Alexandrov

NASA Langley Research Center

9 October 2019

n.alexandrov@nasa.gov

Outline

- Overall idea:
  - Describe a few hard and timely problems in the area of trustworthiness/trust of autonomous systems that would benefit from mathematical attention
  - Point to specific needs in modeling and optimization
  - Encourage subscription to NASA NSPIRES

- Problem setting for trustworthiness/trust
- “Body” problem (survivability)
- “Mind” problems (ATTRACTOR)

Definition of “Autonomous System”

- **Extensive literature and heated arguments on definitions and the meaning of autonomy**

- **A definition (mine):**
  - **Autonomous System ≡ Cyber(-physical-human) system with a capacity for independent decision making and a degree of authority to act on decisions in a specified environment**

- **E.g., my vacuum cleaner is an autonomous system**

Problem Setting in Aero/Airspace



Environment:

Safety-critical

Time-critical

Certification-dependent

3D is harder than 2D!



Density + heterogeneity + autonomy + non-cooperative agents ⇒
System complexity increases / uncertainty grows / safety decreases ⇒
Control must transition to human ∪ machine, with increasing machine authority
Machine authority is a major source of uncertainty

Overarching Trustworthiness Questions

- In future airspace, how do we ensure a high degree of safety for humans in dense, autonomous, non-cooperative traffic when the risk of collisions grows?

- When can we (justifiably) trust an autonomous system in safety-critical and time-critical environments, to inform certification and public trust?

Outline

- Overall idea:
  - Describe a few hard and timely problems in the area of trustworthiness of autonomous systems that would benefit from mathematical attention
  - Point to specific needs in modeling and optimization
  - Encourage subscription to NASA NSPIRES

- Problem setting for trustworthiness/trust
- "Body" problem (survivability)
- "Mind" problems (ATTRACTOR)

Effects of Density on Safety



- Probability of collisions as a function of density
  - Not enough data to formulate a historical model
  - Gas particle models give an estimate (e.g., Alexander 1970)
- Consider randomly distributed aircraft

  r = a given interaction distance

  S = distance traveled by an aircraft

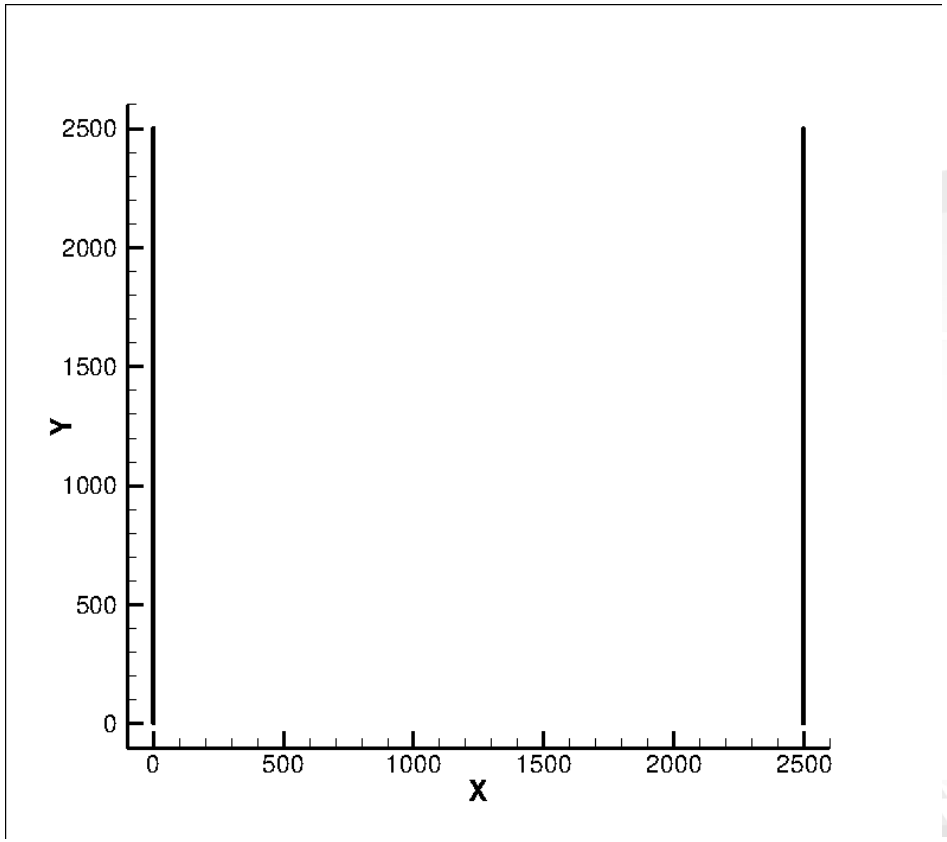
  N = average number of aircraft/unit land area

  L = number of altitude layers; aircraft equally divided among layers

  $P = e^{-\frac{2rSN}{L}}$ = the probability that an aircraft moving horizontally a distance S will not come within distance r of another aircraft in the same layer
- Example: for r = 1 mi, L = 10, velocity = 150 mi/h, and 3000 mi^2
  - **N= 0.01 aircraft per mi^2**: hours for 2-body interaction and 1000 hours for 3-body interactions
  - **N= 0.1 aircraft per mi^2**: evasive maneuvers required every 20 minutes
  - **N= 1.0 aircraft per mi^2**: nearly continuous maneuvering
- Supported by frequency of near misses in dense airspace near airports
- Airspace structure is an option, but not in door-to-door on-demand mobility (ODM)

The Biggest Problem: Heterogeneity

- With shared rules, great densities can be accommodated (joint work with NRL)



- With heterogeneity, e.g.:
  - Interference with CA firefighting efforts
  - “Drone Close Calls, Sightings by Airliners up Fivefold” (3/28/2016, ATCA News)
  - “Miracle on the Hudson”: 1.5 sec between “bird!” and “collision!”



Related Areas

Hazard Environment	Discipline
Normal operations • Mechanical failure • Human error • Atmospheric conditions	Design for system safety
Hostile, intentional • Air defense • Combat	Combat survivability (comprehensive reference: Ball, 2003)
Hostile, unintentional • Severe weather • Midair collisions, including bird strikes • Crashes	System safety + survivability

Civil domain: avoid collisions via highly structured traffic control + bird strike tolerances ; vehicle design separate from airspace load

Military domain: “tougher” aircraft, greater maneuverability; survivability complex , inferred

Consider Survivability in Military Aviation

$$P_S = 1 - P_{NS}$$

- P_S - probability of survival
- P_{NS} - probability of non-survival ("killability") = f(susceptibility, vulnerability)

$$P_S = 1 - P_{NS} = 1 - P_H P_{NS|H}$$

- The probabilities are decomposed further, based on specific scenarios and events.
- The quantification of probabilities in the military domain is done via construction of scenarios, event tree diagrams, simulations, and tests. Improving P_S requires reductions in susceptibility and vulnerability.

Susceptibility and Optimization

- Susceptibility = f(threat, the aircraft, and the scenario)
- Threat
  - Encountering a non-cooperative UAS in civilian airspace would likely be a random event (density, non-cooperativity, equipment faults)
- Aircraft
  - Awareness, countermeasures and performance
  - No obvious opportunity for design optimization to contribute to awareness (a function of intelligent systems, both onboard and remote) or countermeasures.
  - Optimization can play a large role in modeling and design for performance, given a set of requirements imposed by the likely composition of the airspace; e.g., the density and diversity of vehicles, combined with the range and accuracy of detection may lead to new requirements for aircraft agility.
  - Supported by appropriate sensing and automation, small improvements in agility may suffice to weaken collisions with critical sections of the aircraft.
  - Flocks of drones would present a more difficult problem, raising the question of adding countermeasures and structural features.
  - For GA vehicles, the windshield is the most susceptible part of the aircraft. A clear opportunity for optimization is to contribute through design of materials and cockpit re-design.

Vulnerability and Optimization

- In the military domain: Vulnerability Program (e.g., Ball 2003)
  - Three tasks (practiced in other safety-critical areas):
  1. Critical component analysis: identification of critical components and kill modes
     - Relies on data analysis of loss statistics
  2. Vulnerability assessment: quantification of the individual component and entire aircraft vulnerability; start early in design cycle and update continually
  3. Design for low vulnerability, preferably early in design; location and size of components, material, construction, redundancies (e.g., engines), etc.
- SOA: discrete technologies in accordance with vulnerability reduction principles; evaluation of resulting vulnerability; or hardening after design
- Need to convert vulnerability reduction principles to computable design optimization problem

Vulnerability Reduction

- Based on general principles:
  - Critical component redundancy with separation. For example, multiple engines, control surfaces. NASA's SCEPTOR concept with its multiple engines is an example of this principle.
  - Critical component location and separation to prevent cascading damage.
  - Passive damage suppression via fault tolerant design and such technologies as high-strength and self-healing materials.
  - Active damage suppression via technologies that incorporate sensors, e.g., fire detection and suppression.
  - Critical component shielding by other components or additional structure. Model development is needed here to avoid parasitic shielding.
  - Critical component elimination. Replacing pilots with automation comes under this category.
- Overall approach used in civil aviation to reduce vulnerability with respect to, e.g., mechanical failure, but no formal design optimization problem, to my knowledge

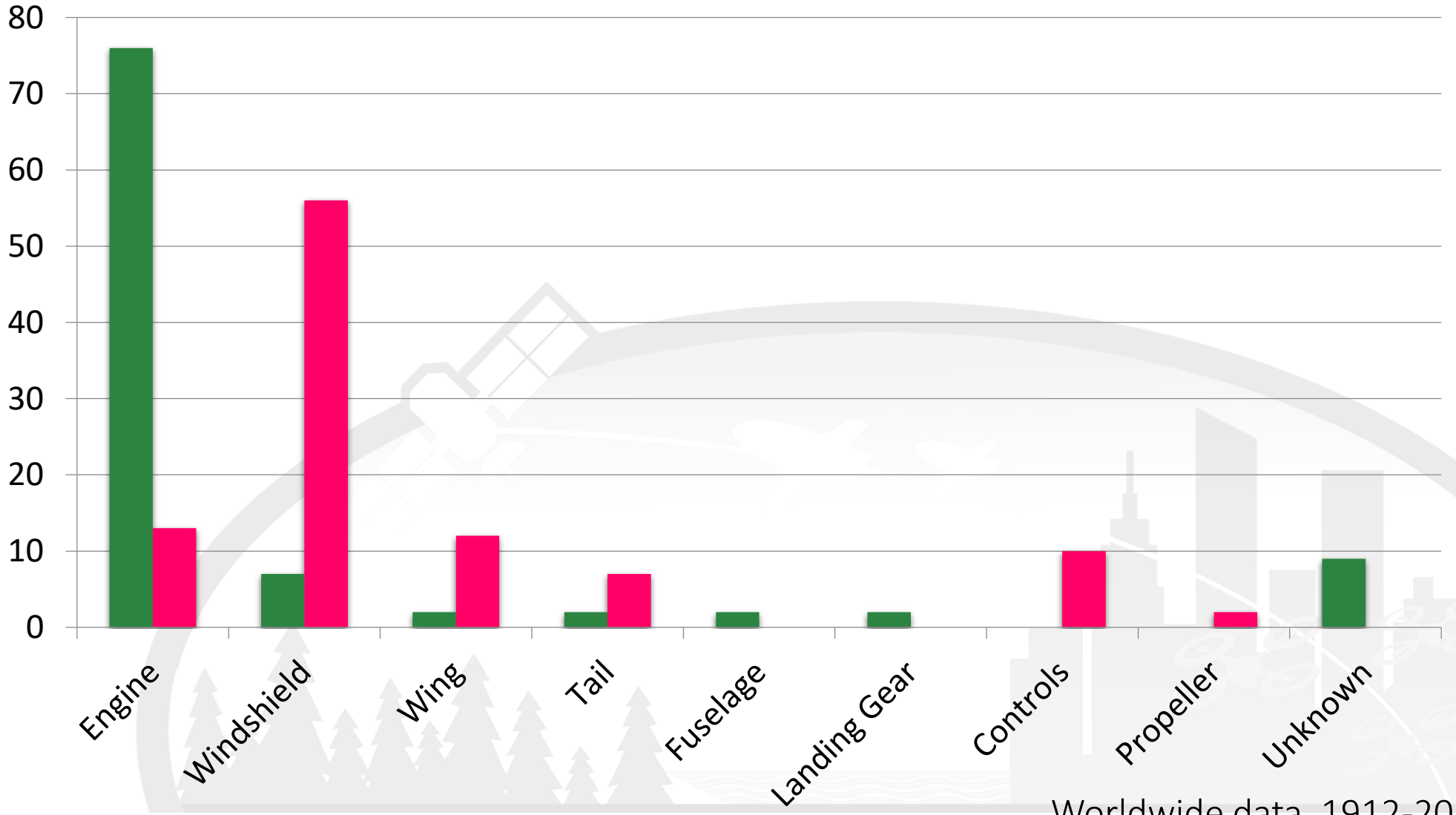
Optimization Opportunities

- ***Vulnerability as Objective/Constraint***
  - Vulnerability value is an outcome of applying heuristic principles during design and an *a posteriori* evaluation; followed by corrections, which might require re-design. A purely probabilistic, *post factum* evaluation of vulnerability cannot be easily used as an objective function or a constraint in a design optimization formulation.
- ***Uncertainty Quantification***
  - Given the need for a higher degree of survival assurance of each civilian aircraft carrying passengers, a great deal of accuracy in uncertainty quantification of impact for possible collision sources in civilian domain is needed.
- ***Dimensionality***
  - Given thousands of components on an aircraft, the approach to identifying critical components, damage modes and potential mitigation via discrete enumeration is likely to give rise to problems of very large problem dimension. Appropriate problem decomposition and aggregation is a fruitful area of investigation for design optimization.

On Collision Damage in the Civil Domain

- A recent activity, with a lot of unknowns
- Some efforts, e.g., Virginia Tech CRASH lab FEM modeling of collisions (Bayandor et al., ongoing; Radi 2013)
- ID the areas of aircraft that require first priority
- Use wildlife collisions as a model for frequency of incidence with an object
- Rough estimates for where to focus, to minimize damage
- Cannot completely extrapolate from birds to UAV

Location of Bird Strikes (%)

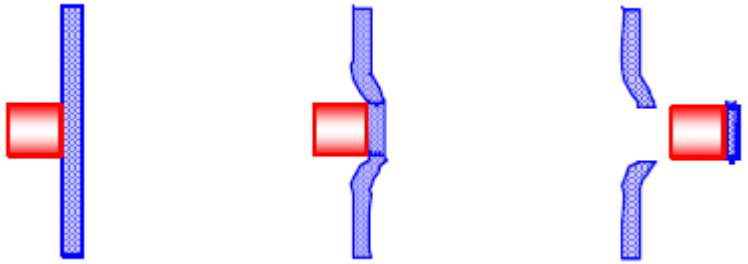


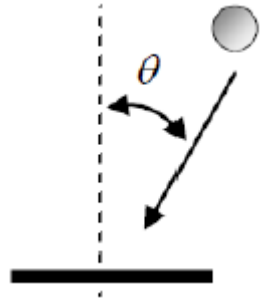
Notes:

- In the military domain, the order of importance is different
- For GA, windshield collisions result in most fatalities

Worldwide data, 1912-2011, Thorpe, 2012

Impact Damage





- Studies, e.g.,
  - Bird impacts, based on historical data (e.g., Thorpe 2012)
  - Studies of UAV impact damage, using the FAA penetration equation $V_{50} = \sqrt{\frac{2LCSt^2}{mcos^2\theta}}$ = the ballistic limit = velocity required to make a hole in a sheet of metal. Here m is the mass of the projectile, θ is the angle of the impact, $C_S$is a material property constant, L is the perimeter of the presented area of the projectile, t is the thickness of the metal sheet (e.g., Radi 2013)
  - Many assumptions; large uncertainties, but overall trends relevant

Impact Damage, Cont.

- Commercial aircraft
  - Engine ingestion – the likeliest event, with probable loss of engine; unlikely to be catastrophic (N.B. However, consider a flock of UAV)
  - Penetration of Al airframe predicted for velocities 200 mi/h and above, regardless of UAV size
  - Penetration predicted for fixed-wing UAV at very low velocities
  - Windshield results highly uncertain and depend on material properties and angle of the windscreen
- GA aircraft
  - Penetration likely at cruise speeds, regardless of UAV size; and likely for large UAV during approach

An Approach to Problem Formulation

- SOA: Rule-based point designs with inferred vulnerability; no direct or inverse design formulation
- Problem: Growing potential for collisions, with extreme epistemic uncertainties
- Calls for robust design optimization
- Approach now in development: application of info-gap theory (e.g., Ben-Haim 2001) to design for survivability
  - Info-gap is related to Wald's maximin decision theory
  - Does not require definite *a priori* uncertainty quantification
  - Problem is expressed as nested uncertainty analyses, e.g.,

$$U(\alpha, \hat{u}) = \{u(t): |u(t) - \hat{u}(t)| \leq \alpha\varphi(t)\}, \alpha \geq 0,$$

  - where $U(\alpha, \hat{u})$ is the set of functions $u(t)$ that deviate for the nominal function $|\hat{u}(t)|$by no more than $\alpha\varphi(t)$ for some known envelope function $\varphi(t)$ and α is the uncertainty parameter

Cont.

- Info-gap: when information is scarce or the system undergoes changes, so that the past may not be a good predictor of the future
- For a vector of design variables x, robustness is expressed as

$$\hat{\alpha}(x) = \max\{\alpha\text{:minimal requirements are always satisfied}\}$$

- Robustness = resistance to uncertainty; optimization not with respect to performance; satisficing at critical survival conditions
- If success is measured by a merit function M and m_c is a critical merit value, then robustness is

$$\hat{\alpha}(x, m_c) = \max\{\alpha: (\min_{u \in U(\alpha, \hat{u})} M(x, u)) \geq m_c\}$$

- For multidisciplinary (multiobjective) performance, robustness is

$$\hat{\alpha}(x, m_c) = \max\{\alpha: (\min_{u \in U(\alpha, \hat{u})} M_i(x, u)) \geq m_{c,i}, i = 1, \dots, N\}$$

- where m_c is a vector of critical merits

Example of Info-gap application

- Dealing with model uncertainty
- Consider application to a small subproblem
- Using the FAA penetration equation $V_{50} = \sqrt{\frac{2LCSt^2}{mcos^2\theta}}$ to compute robust skin/armor thickness

Cont.

- Choose or certify the thickness t, so that $x = \frac{1}{V_{50}}$ is acceptably small (or V_{50} is acceptably large), subject to uncertain mass m, i.e.

$$x \leq x_c \text{ for some } x_c > 0$$

- The nominal mass $\hat{m}$ is known (doesn't have to be; may be a distribution)
- The actual mass m is unknown
- The mass uncertainty is represented by an interval info-gap model

$$U(\alpha, \hat{m}) = \{m : |m - \hat{m}| \leq \alpha\}, \alpha > 0$$

- Unbounded family of nested sets, where α is an unknown uncertainty parameter
- Robustness $\hat{\alpha}$ = the greatest value of α for which the performance requirement $x \leq x_c$ holds

Example, cont.

- To compute robustness:
- Evaluate maximum x, up to uncertainty α

$$\max_{m \in U(\alpha, \hat{m})} x = \sqrt{\frac{(\hat{m} + \alpha)cos^2\theta}{2LS_c t^2}}$$

- Equate maximum to critical value x_c and solve for α

$$\hat{\alpha}(t, x_c) = \begin{cases} \frac{2x_c^2 LS_c t^2}{cos^2\theta} - \hat{m}, \text{ if expression} \geq 0 \\ 0 \; otherwise \end{cases}$$

- Require that $\hat{\alpha}(t, x_c) \geq \hat{\alpha}_d$, some minimum design value, i.e.,

$$\frac{2x_c^2 LS_c t^2}{cos^2\theta} - \hat{m} \geq \hat{\alpha}_d$$

Example, cont.

- And t must satisfy

$$(*) \qquad t \geq \sqrt{\frac{(\hat{\alpha}_d + \hat{m})cos^2\theta}{2x_c^2 LS_c}}$$

- Interpretation: acceptable V_{50} is guaranteed if the value of the uncertainty parameter α is smaller than the value of the robustness threshold $\hat{\alpha}_d$, i.e., if t satisfies (*), then

$$\left|\frac{1}{V_{50}}\right| \leq |x_c| \forall m \in U(\alpha, \hat{m})$$

- The choice of threshold robustness $\hat{\alpha}_d$
  - Experience
  - Dimensional: in the example – same units; $\hat{\alpha}/\hat{m} \ll 1 \Rightarrow$design vulnerable to small mass variations; $\hat{\alpha}/\hat{m} \geq 1 \Rightarrow$insensitive to large variations in mass.
  - Based on acceptable levels of risk (evaluation of consequences; qualitative)
  - Other methods

Survivability on the Ground

- Increased UAS traffic + projected appearance of air taxis and private air vehicles ⇒increase risk to people on the ground, animals, and property in densely populated areas
- Ensure lift, thrust, control, and structural integrity, for higher collision probabilities
- If damaged, Safe2Ditch (Glaab et al.)
- Design for landing harmlessly is a promising area for design optimization techniques in the case of aircraft carrying humans, because such techniques may involve a re-design of the aircraft.
  - E.g., design ballistic recovery systems, such as Cirrus Airframe Parachute System (CAPS)
  - Larger vehicles may be amenable to parachuting if separable modules are equipped with ballistic recovery systems.
- Cost is a problem. Optimization could contribute by incorporating comprehensive cost modeling in the analysis of new configurations resistant to crashing and amenable to harmless landing.

Survivability Summary

- Survivability is affected by contributions from all design disciplines. Connection between the traditional vehicle design variables and the measures of survivability is implicit and inferred.
- Typical operational scenario
  - Sensory systems of the aircraft continually monitor the external environment and internal state of the aircraft
  - Inform the intelligent autonomous perception and decision making system
  - Issue commands to the aircraft, to avoid a collision, weaken it, or recover sufficiently to deliver the passengers to the ground safely
  - The traditional disciplines are responsible for the actions dictated by the intelligent system to be physically implementable. In order to achieve this, survivability must be directly connected, through variables, objectives and constraints, with all relevant disciplines.
- Including survivability into design optimization on par with traditional computational disciplines presents a difficult challenge: need developments in computational modeling, manufacturing, materials, and design optimization problem synthesis, robust design, among other disciplines.
- If not done, could remove the business case

Outline

- Overall idea:
  - Describe a few hard and timely problems in the area of trustworthiness of autonomous systems that would benefit from mathematical attention
  - Point to specific needs in modeling and optimization
  - Encourage subscription to NASA NSPIRES

- Problem setting for trustworthiness/trust
- "Body" problem (survivability)
- "Mind" problems (ATTRACTOR)

ATTRACTOR Team

Autonomy Teaming & TRAjectories for Complex Trusted Operational Reliability

LaRC

Danette Allen, co-PI
Edward Adcock
Bryan Barrows
Jeremy Castagno[1]
Meghan Chandarana[1,6]
John Cooper
Chester Dolph
Brian Duvall[2]
James Ecker
James Farrington
Alwyn Goodloe
Lynn Hadley
Zachary Johns[2]
Benjamin Kelley
Lisa Le Vie
Kyle McQuarry
Andrew Miloslavsky
Paul Miner
Daniel Moerder
James Neilan
James Plant
Paul Rothhaar
Ivan Perez[2]
Javier Puig-Navarro[3]
Loc Tran
Matthew Vaughan
Ralph Williams
UAS Operations Office

Jon How[4]
Nick Roy[4]
Sebastian Scherer[6]
Matthias Scheutz[5]

ARC

Adrian Agogino
Kamalika Das
Dimitra Giannakopoulou
David Iverson
Ritchie Lee
Nikunj Oza
Donald Soloway
Shawn Wolfe

AFRC

Manuel Castro
Jason Holland
Otto Schnarr

1. Pathways students
2. National Institute of Aerospace
3. University of Illinois Urbana Champaign
4. Massachusetts Institute of Technology (MIT)
5. Tufts University
6. Carnegie Mellon University (CMU)

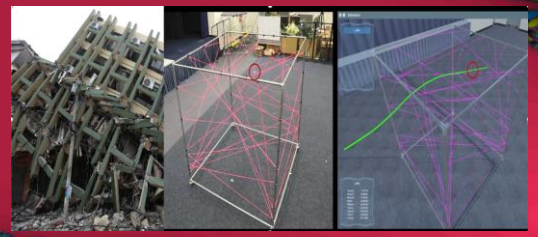
The Problem

Build a basis for certification of autonomous systems via establishing metrics for trustworthiness and trust in multi-agent team interactions, using AI explainability and persistent modeling and simulation, in the context of mission planning and execution, with analyzable trajectories.

Components of Complex Multi-agent Systems

Measurable **trustworthiness** of fundamental functional components

Context: Trajectories



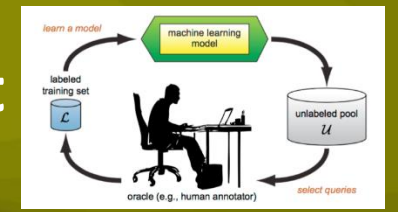
Heterogeneous multi-agent (including human-machine) teaming is essential



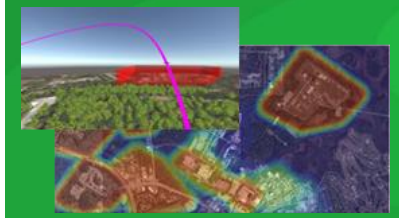
Teaming

Efficient two-way learning/ training loop.

Learning & Justifiable Trust

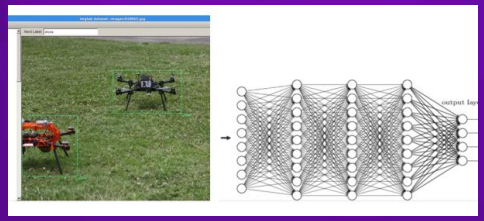


Good Outcomes: Basis for Trust

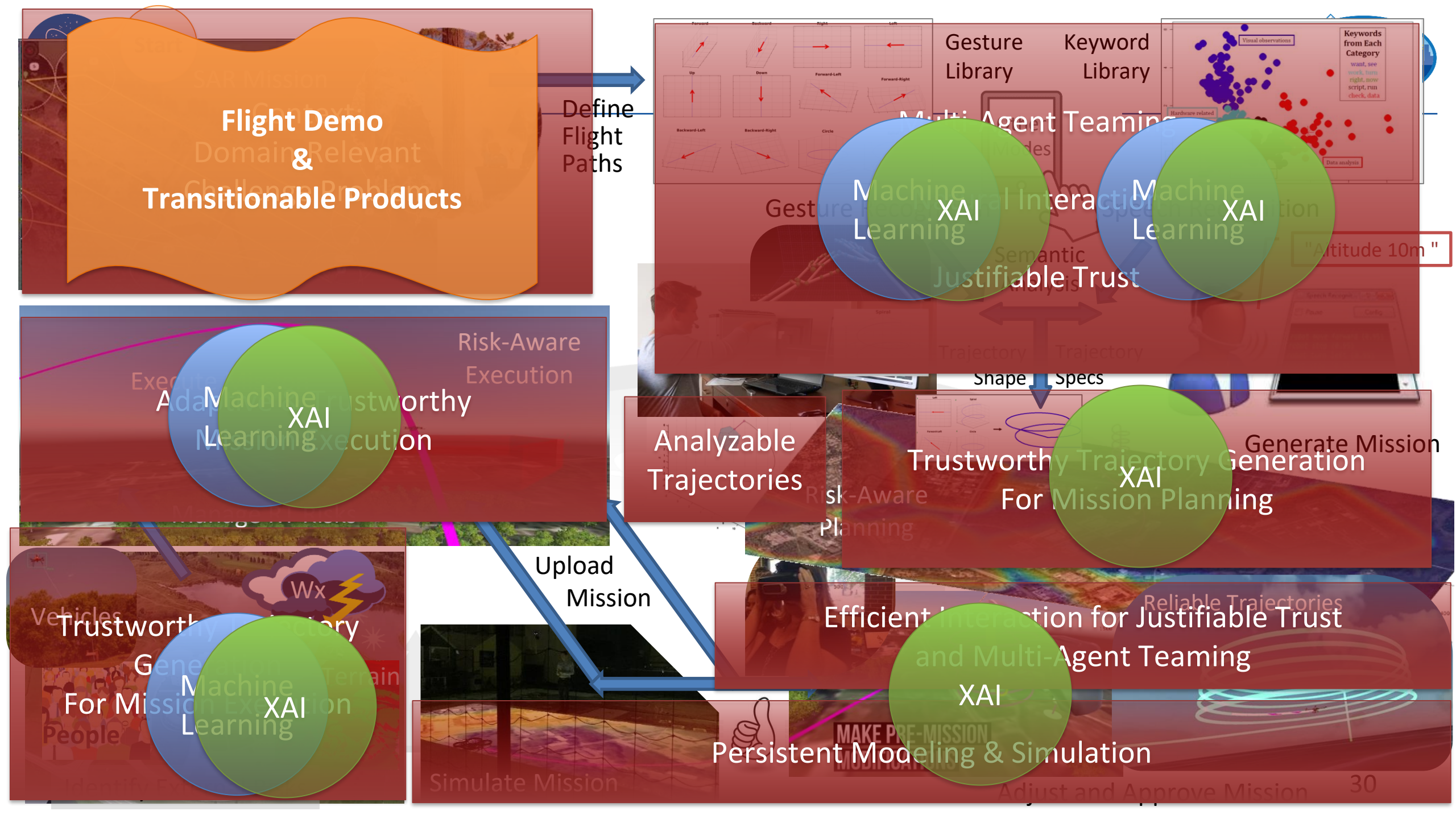


Sound multiobjective decision-making under uncertainty

Explainability

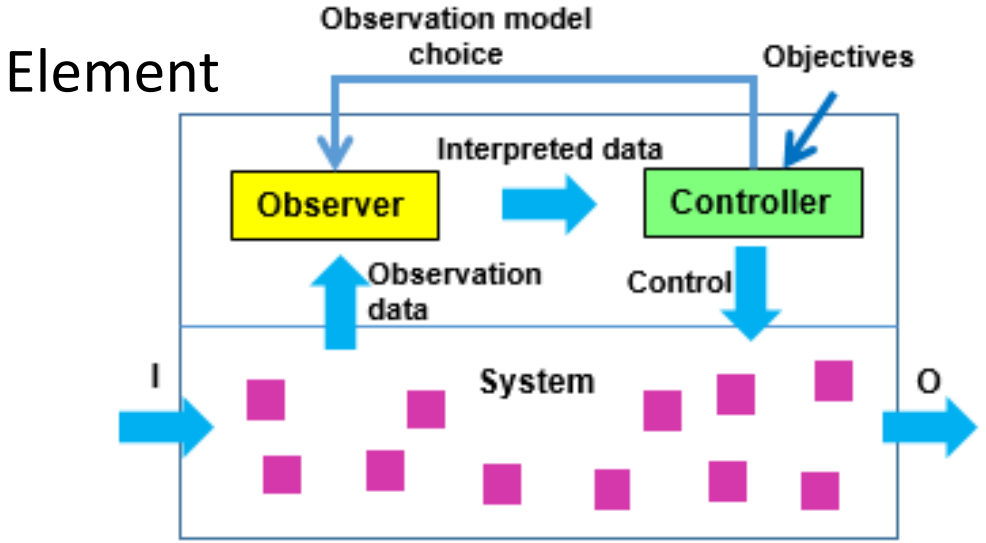


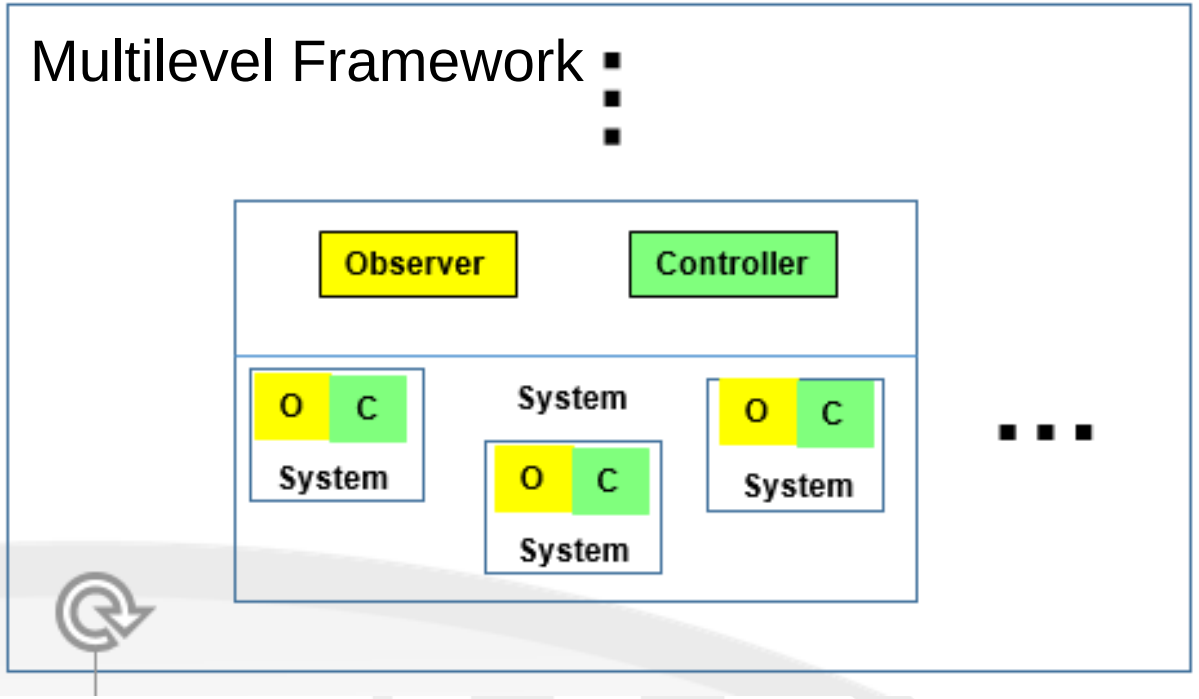
These are two UAVs. Because… They are quadrotors. They have four propellers. They have four legs.


Flight Demo
&
Transitionable Products
Define
Flight
Paths
Gesture
Library
Keyword
Library
Keywords
from Each
Category
Multi-Agent Teaming
Machine
Learning
XAI
Justifiable Trust
"Altitude 10m "
Risk-Aware
Execution
Trustworthy
Execution
Analyzable
Trajectories
Shape
Specs
Generate Mission
Trustworthy Trajectory Generation
For Mission Planning
Upload
Mission
Wx
Trustworthy
Generation
For Mission
Efficient Interaction for Justifiable Trust
and Multi-Agent Teaming
Reliable Trajectories
Persistent Modeling & Simulation
Simulate Mission
Adjust and Approve Mission
30

Overall System Scheme:

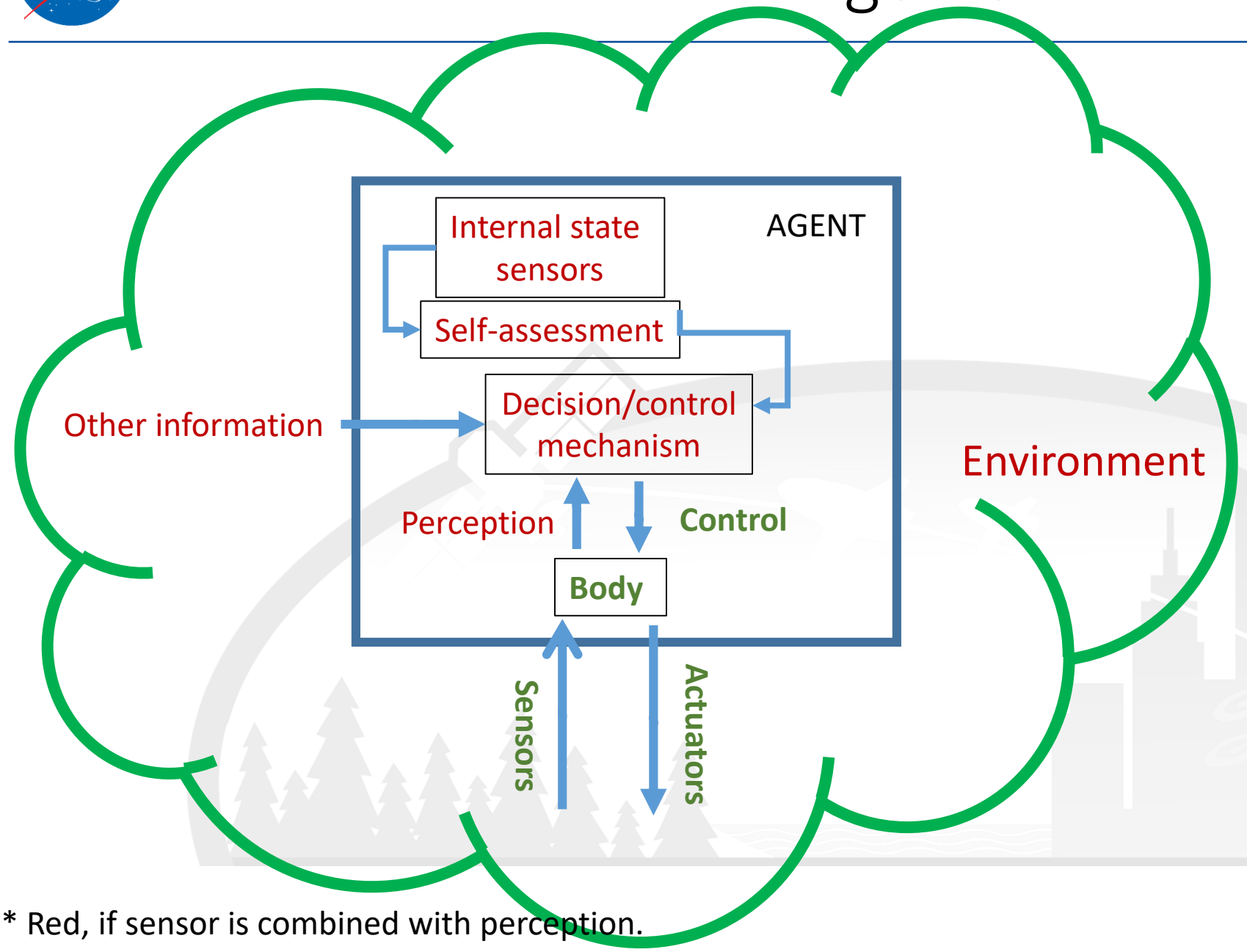
Generic Multi-level Multi-agent Observer-Controller Framework (e.g., Tomford et al.)





- Observer
  - Monitors, perceives and analyses observation data; “plays forward” to predict potential states; conveys information to Controller
- Controller
  - Detects divergence from “known knowns”; updates states based on system performance; updates knowledge base; explores new actions; updates objectives

Agents



- trust $\wedge$ trustworthiness $\sim$ {reliability, efficiency, robustness, resilience, predictability of behavior, explicability, timely prediction of phase transitions, survivability} $\sim$ 1/Risk
- Conjecture:
  - **Green:** Amenable to traditional safety-based design
  - Red: New risks; new approaches needed or bound complexity and achieve trust and trustworthiness

* Red, if sensor is combined with perception.

Background

- **Overarching feasibility question: How/when can we trust an autonomous multi-agent cyber-physical-human system?**
- Must define the problem and terms (over 20 definitions of trust alone) to enable active design for trust and trustworthiness
- In technology, trust has usually been associated with reliability; separated with the advent of software-controlled complex systems
- *Trustworthiness* – attribute of the system, but a function of trust as well
- *Trust* – attribute of an agent; readiness to rely on another agent for decision and/or action
  - H-H and H-M: subjective, sometimes not rational
  - M-M: similar mechanisms but hope to rely on objective criteria
  - H-M and M-M: rely on imposed thresholds

Trust Components

- Trust = subjective probability on the part of agent A that agent B will give direction or perform an action that will result in a positive outcome (or will not result in a negative outcome) for A
- Trust = subjective expectation by A of behavior by B, based on the history of their interactions
- Trust implies dependence, reliability (trustworthiness), confidence, subjectivity, control (e.g., Kofta 2007), risk, expected benefit to the agent that exhibits trust
- In distributed systems trust is related to reputation (e.g., online systems)
- Cognitive aspects of trust (e.g., Castefranchi and Falcone 2000) are beliefs in competence, intent, persistence, dependence, realization

Trust, Confidence, and Control Related

Consider agents A and B

- Confidence $CE_{A \rightarrow B} = f_A (TWB_B)$
- Trust $T_{A \rightarrow B} = g_A (TW_B)$
- Control $CL_{A \rightarrow B} = CE_{A \rightarrow B} - T_{A \rightarrow B}$
- Where

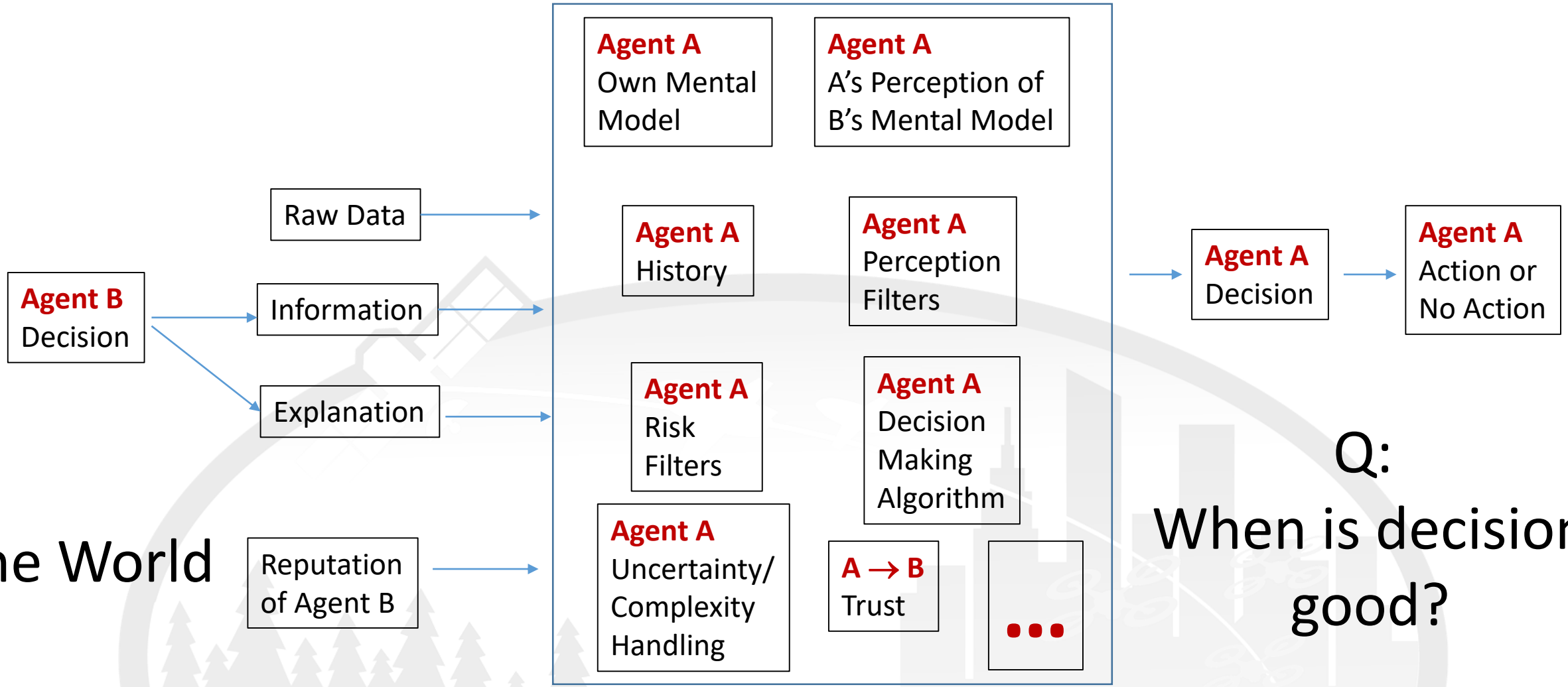
  f_A and g_A are agent A's subjective estimates
  TWB is trustworthy behavior
  TW is trustworthiness

- Trustworthiness and control are at odds – A may not exercise control if B is trustworthy
- The nature of confidence depends on the environment and goals of the agent
- One representation:
  - All system components form a directed network
  - Nodes are agents
  - Edges contain evidence on which trust/confidence are based
  - Confidence of an agent in an event = sum of confidences into relevant nodes

Evidence

- General:
  - Good decision outcomes for a long period of time
  - Explicability during training and forensics
    - Requires a shared mental model
  - Adaptability or, at least, graceful degradation in the face of unanticipated conditions
  - Recognition and warning of “no solution”
  - Prediction of phase transition from controllability to non-controllability
  - Risk minimization
- Can be accumulated via
  - Forensics, analysis, explanation
  - V&V
  - Statistically in practice (e.g., the current air traffic system)
  - Games
  - Simulations

Problem Setting



The World

Q:
When is decision good?

A Sample of Working Hypotheses

- Explicability or interpretability during training, forensics and operations—except in time-critical operations—increases justified trust and trustworthiness
  - (Working on explaining any computational decision making, with special emphasis on explaining machine learning used in the visual component of ATTRACTOR)
- Shared mental models support explicability and interpretability
- Shared mental models support sound decision making

ATTRACTOR Example: Decision Conflict



A Possible Conversation

- M: I must change a planned portion of trajectory
- H: Why?
- M: I detect children in the area. Risk rises from X to Y.
- H: Are you sure?
- M: Yes, here is the image of children.
- H: What is your new trajectory?
- M: Here is the image and associated risk.
- H: Are there alternative trajectories?
- M: Yes, but their associated risks are higher and the associated rewards are small.
- …
  - Explanations implied that the goals and risk assessment are shared
  - Q: Who has the final decision authority?
  - N.B. Representation of risk and uncertainty information to the human is a big problem (e.g., Monty Hall problem)

Shared Mental Models

Collaboration with Tufts University, Matthias Scheutz

- Mental model: a mechanism for *describing, explaining, and predicting* the behavior of the *system*
- Shared mental model: Knowledge structures shared by members of the team to enable coordination for a task and adaptation of the task (*describe, explain and predict the behavior of the team*)
- Shared models are necessary for explanation
- Tufts team:
  - Formalization of mental models (UML representation and logical notation)
  - Formalization of model similarity and compatibility
  - Extension of cognitive concepts of human teams to include software agents
  - Consider combining physical and mental models to form an “extended mind” (Rouse and Morris 1986)

Shared Mental Models

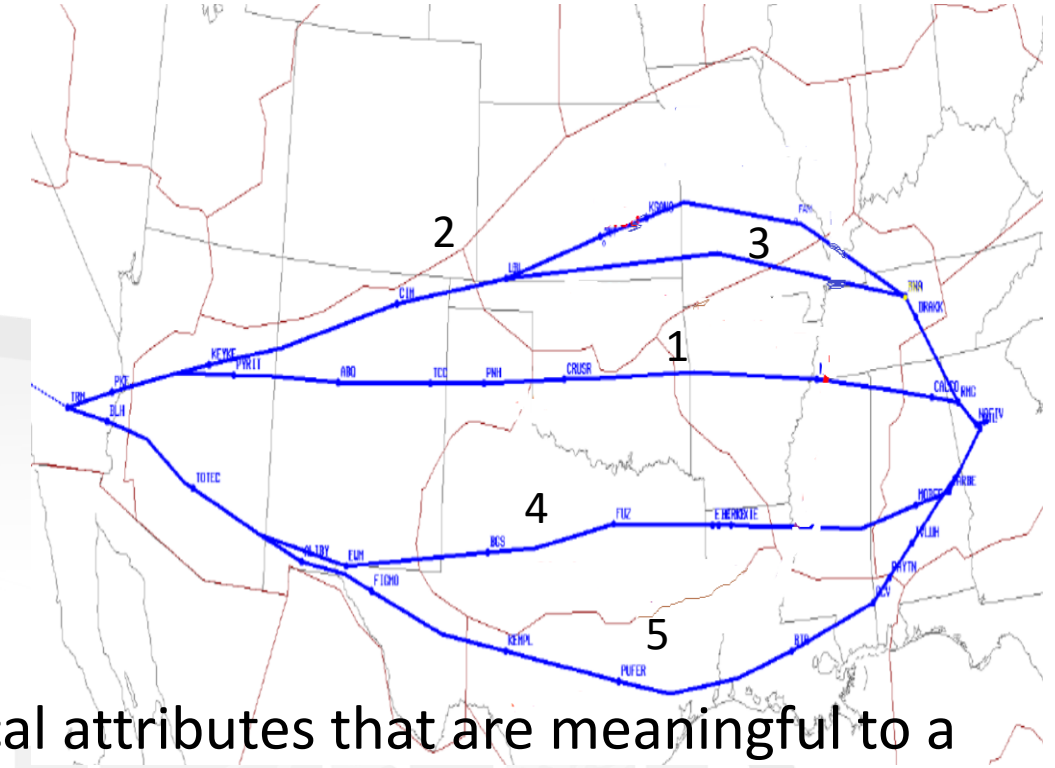
Complementary Approach in ATTRACTOR

- Claim: All decisions at all scales are solutions of optimization problems, subject to constraints, some unstated (e.g., incomplete formulation, algorithmic limitations)
- All difficulties in “Concrete Problems in AI Safety” (Amodei et al. 2017) can be traced to reward (objective) function definition
- N.B. “On the Folly of Rewarding A, While Hoping for B” S. Kerr, 1975
- View mental model as a formal optimization problem formulation
- Model similarity = shared variables, objectives, constraints, and multiobjective formalization

Proposed Approach to Explicability of Decision Making

- Consider explanations/justification/interpretation of a set of proposed trajectories:
  - Planned trajectory 1 is infeasible because violent weather is in the path of the old trajectory
  - Trajectories 2-5 are flyable and conflict-free
  - Trajectory 3 saves more fuel but flies through denser traffic that trajectory 2
  - …
- Common features of all acceptable explanations:
  - The old trajectory violated constraints
  - New trajectories are flyable and conflict-free (satisfy constraints); objectives have better values

Image courtesy FAA. Used for notional depiction of trajectories.



- Regardless of the algorithm, the trajectory has physical attributes that are meaningful to a human decision maker in terms of value functions associated with moving from point A to point B.
  - E.g., minimizing fuel expenditure, minimizing delays in reaching destination, maintaining safe distance from other aircraft, objects, and weather, etc.

Context and Goal Driven Explanation

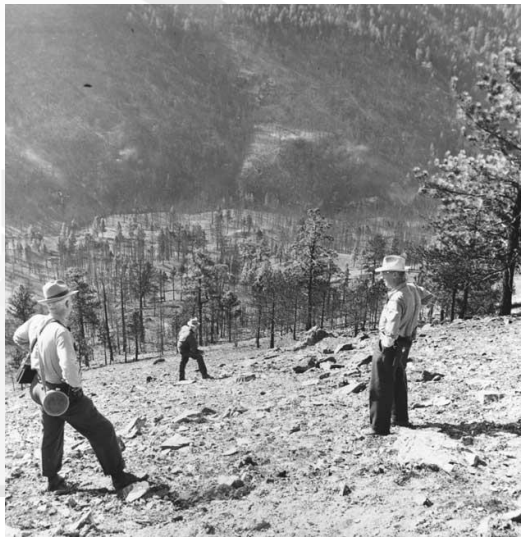
- Convincing explanation must contain information about
  - Constraint violations and comparative values of the objectives and constraints between the current and proposed decision, such as a trajectory
  - Sensitivity (robustness information)
  - Overall risk information (uncertainty horizons, consequences implied or explicit)
  - Information about alternative decisions via what-if scenarios if sims are available
- Note: In M-M interactions, resolving information constraints also requires physical “explanation”, appropriately represented.

Correct / Complete Objectives Support Trustworthiness

- Good adaptation
  - Robustness to distributional shift (good behavior in environments that differ from training ones)
- Bad Adaptation
  - E.g., reward hacking (if, during cleaning, you get a reward for not seeing a mess, don't look)
- Peppered moth
  - Slow
  - Not assured
  - Salient point: potential variable existed, in principle
- Decision making: Mann Gulch Fire
  - Want this
  - Fast
  - Not assured
  - Seemingly out of nowhere
  - Salient point: potential variables/functions existed

Fermi Lab





Nat. Interagency Fire Center

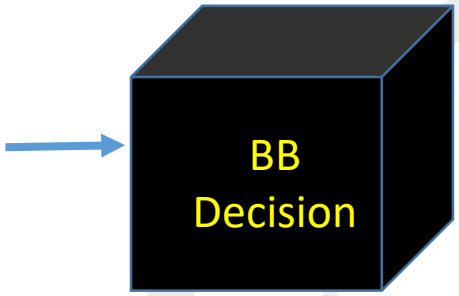
Finding Objectives

- Conjectures
  - Objectives will not arise unless they have been defined *a priori*
  - Objectives can only be “activated” or “de-activated” via adaptive adjustment of preferences during execution
  - Preference adaptation (adjustment) can be learned, given “complete” objective formulation

- In progress:
  - Using preference adjustment to establish common “cognitive” model between human and machine decision makers for execution and explanation
  - Multiobjective preference adaptation for moving from point A to point B via interactive learning

Common Cognitive Model for Explanation

- Working Hypothesis: Recipients of explanations would find others' explanations convincing if they fit within their own explicability framework (although the framework may require expansion via learning)
- Local (reward) and global (utility) objectives must be interpretable to H in M-H interactions
- Initial approach to forming a common model, for both objectives



x* = argmin F(x), where F(x) = $\sum w_i f_i(x)$ is a decision objective, local or global, $f_i(x)$ are known; e.g., {distance, fuel, delays,…}

E.g., Use interactive multiobjective optimization and inverse reinforcement learning to ID common values of weights

Mixed Reality in Flight



SAR under the Canopy (SARUC)

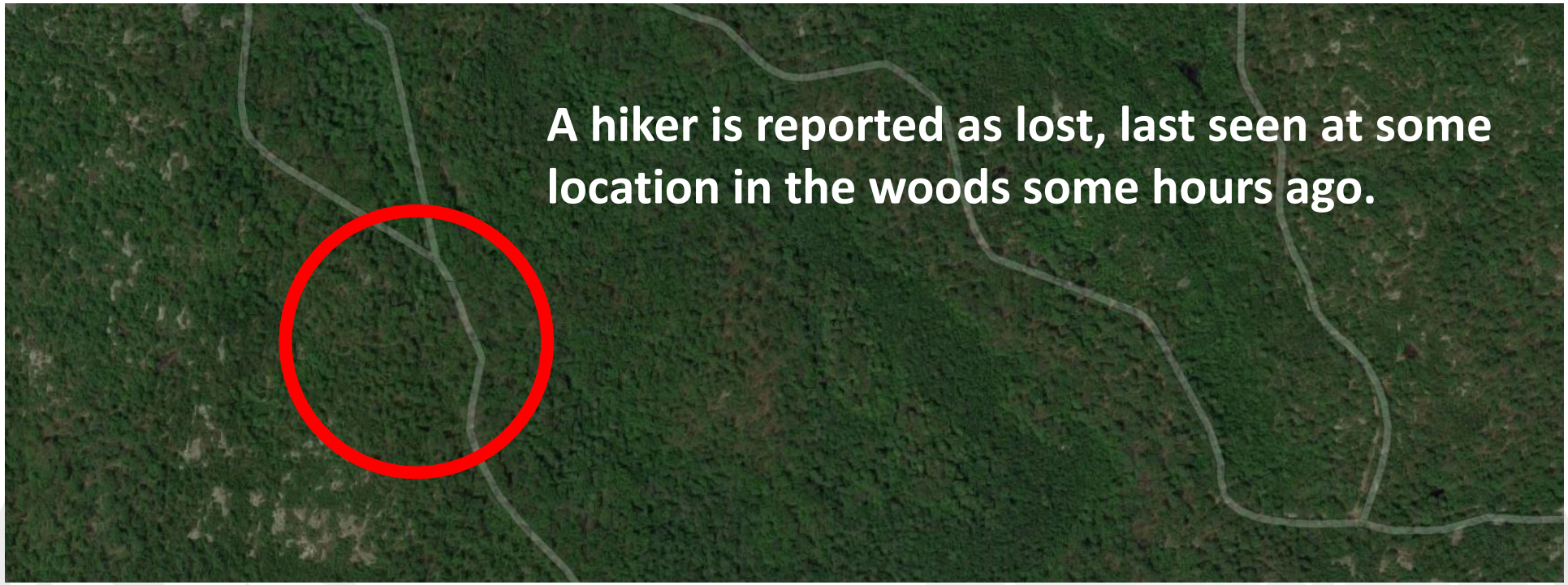


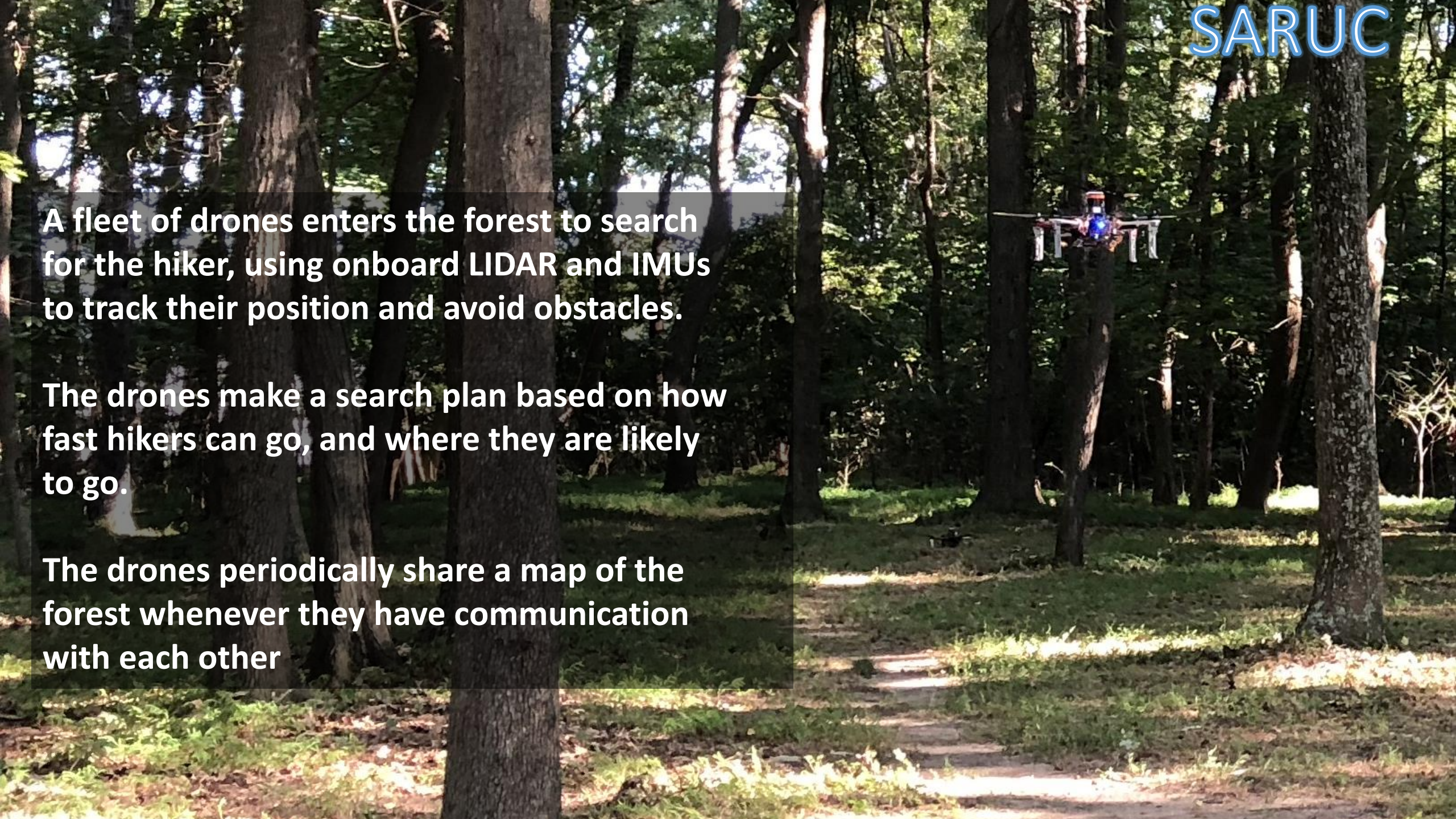
Exploration and Detection
March 2018

Collaboration with
Massachusetts Institute
of Technology (MIT)

Exploration and Map Closure
October 2017

ATTRACTOR DRM





A fleet of drones enters the forest to search for the hiker, using onboard LIDAR and IMUs to track their position and avoid obstacles.

The drones make a search plan based on how fast hikers can go, and where they are likely to go.

The drones periodically share a map of the forest whenever they have communication with each other

Required Capabilities (UML 4-6)

SARUC



Each vehicle must be able to:

- Estimate its position, attitude and velocity robustly from onboard sensors instead of GPS, since the tree canopy will interfere with GPS-based estimation;
- Locate the obstacles around it, such as trees and the forest floor;
- Decide where to search next and plan collision-free trajectories;
- Keep track of the vehicle energy state and maintain efficient egress paths;
- Share information and vehicle status with other vehicles, and subdivide the search and rescue tasks amongst the team appropriately.

LaRC CERTAIN Range




SARUC
Time
ROS Time: 1538662099.07
ROS Elapsed: 3763.58
Wall Time: 1538662099.13
Wall Elapsed: 3763.58
Reset
Left-Click: Rotate. Middle-Click: Move X/Y. Right-Click/Mouse Wheel: Zoom. Shift: More options.
nasa1.rviz* - RViz
Interact
Move Camera
Select
Focus Camera
Measure
2D Pose Estimate
2D Nav Goal
Publish Point
Time
ROS Time: 1538662099.13
ROS Elapsed: 3760.28
Wall Time: 1538662099.15
Wall Elapsed: 3760.21
Experimental
29 fps
CAUTION DO NOT ENTER

Touchpoint Demo 10.4.2018



SARUC Demo Technologies




Autonomous
Multi-Modal
Multi-Agent
Search & Rescue

ATTRACTOR Concluding Remarks

- Overviewed two examples from a range of work toward justified trust
- Ongoing work:
  - Comprehensive identification of uncertainties, including sensors, perception, and computational tractability in a multi-agent system (lacking data, lacking safe envelope notion for ML)
  - Roll up of uncertainties and risk to quantify expectation of good outcomes in the system
  - Capture differences between H and M decision making under uncertainty
  - Rigorously partition the domain of method WRT to their amenability to traditional V&V For heterogeneous autonomous multi-agent systems, especially in a dense environment, justified trust is more complicated; a system question
  - Anomaly detection
  - Trust and trustworthiness through bounding complexity
  - Stress testing algorithms
- At the end of ATTRACTOR we are anticipating a general scheme for system design for trustworthy autonomous systems

<u>Acknowledgment:</u>

Thanks to NASA's Convergent Aeronautics Solutions (CAS) Project of the Transformative Aeronautic Concepts Program (TACP)

Backup materials

Statistical Estimates from the Automobile Domain

- How many miles of driving would it take to demonstrate autonomous vehicle reliability? (Kaira & Paddock)

Miles/Years to be driven…	1.09 fatalities per 100 million miles	77 reported injuries per 100 million miles	190 reported crashes per 100 million miles
Without failure to demonstrate with 95% confidence that the failure rate is at most…	275 million miles (12.5 years)	3.9 million miles (2 months)	1.6 million miles (1 month)
To demonstrate with 95% confidence their failure rate to within 20% of the true rate of…	8.8 billion miles (400 years)	125 million miles (5.7 years)	51 million miles (2.3 years)
To demonstrate with 95% confidence that their failure rate is 20% better than the human driver failure rate of…	11 billion miles (500 years)	161 million miles (7.3 years)	66 million miles (3 years)