Calibrating Trust with Assured Self-Confident Autonomy

Nisar Ahmed

Assistant Professor

Cooperative Human-Robot Intelligence (COHRINT) Laboratory

Department of Aerospace Engineering Sciences

University of Colorado at Boulder

RSS Workshop on Social Trust in Autonomous Robots

Ann Arbor, MI

June 19, 2016

An Autonomous Robot Appears Before You, and Says…

"Come with me if you want to live…"

What do you do?

Would you **trust** this robot?

What shapes and influences your trust?

Wait – what's this whole "trust" thing?

Trust = one autonomous agent's willingness to depend on another

Many mixed contextual influences, including:

- background disposition and beliefs
- look and feel of interaction
- expected vs. actual behavior/capabilities and knowledge (for a robot from future)

…

→ perceived ability to do what we want and as advertised

What if a robot told you how much it trusted itself to get something done,
i.e. its **"self-confidence"**?
(would only an "expert" appreciate this?)

Which of these inspires trust?

vs.

Overview

Definitions and User Trust Modeling

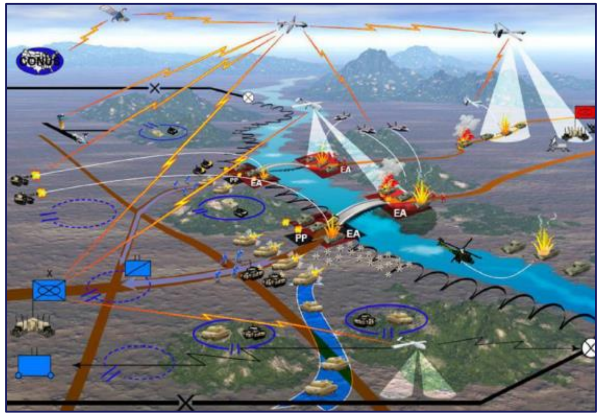
– Assurances and actions between user-autonomy

“Self-confidence”

– A possible assurance for calibrating trust

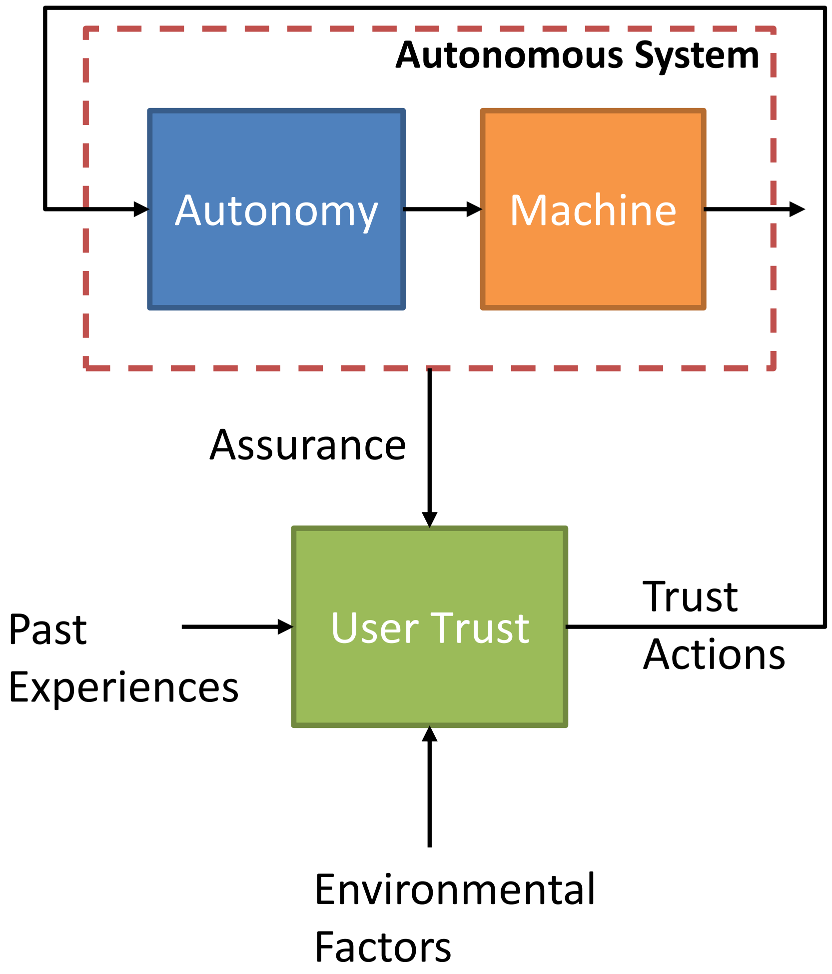
“Trustors/Trustees” for autonomous UXV systems in aerospace

– many conceptual overlaps with social, medical, industrial, etc. robotics



Terminology

- **Trust**: User's willingness to depend on the autonomous system
  - Depends on situation and initial disposition/beliefs
  - Affects the user's actions
- **Assurance**: Autonomous system's ability to affect the user's trust
  - Positive or negative
  - Can affect trust in several key categories
  - Ultimately affects user's actions



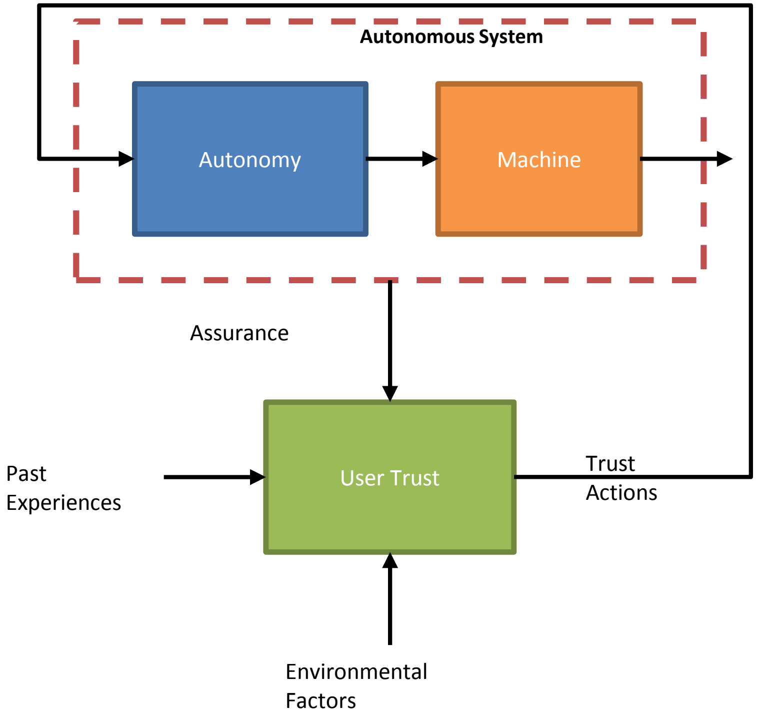
Barriers to Everyday Acceptance of Sophisticated Autonomy

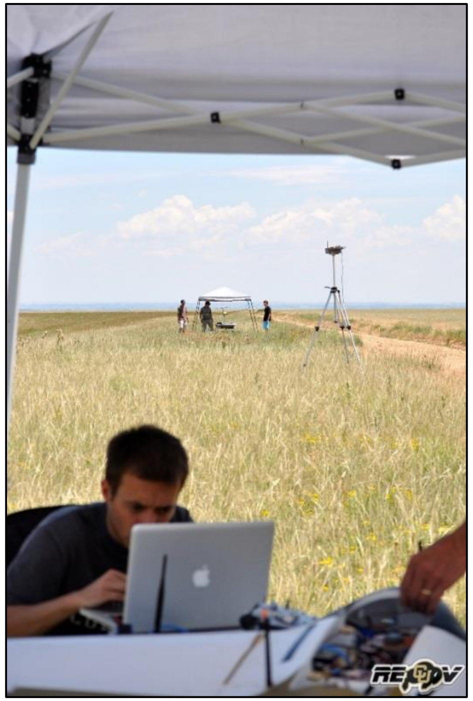
Some barriers are driven by notion that autonomy must be **type certified**

- Non-Determinism
- Perception / Big Data
- Security / Networked
- Human-UAS Interface
- Modeling and Simulation
- Verification and Validation

Yet, people are **licensed** with relatively sparse assessment

- This is possible because of "trust" of other pilots/users, regulators, societies, etc.



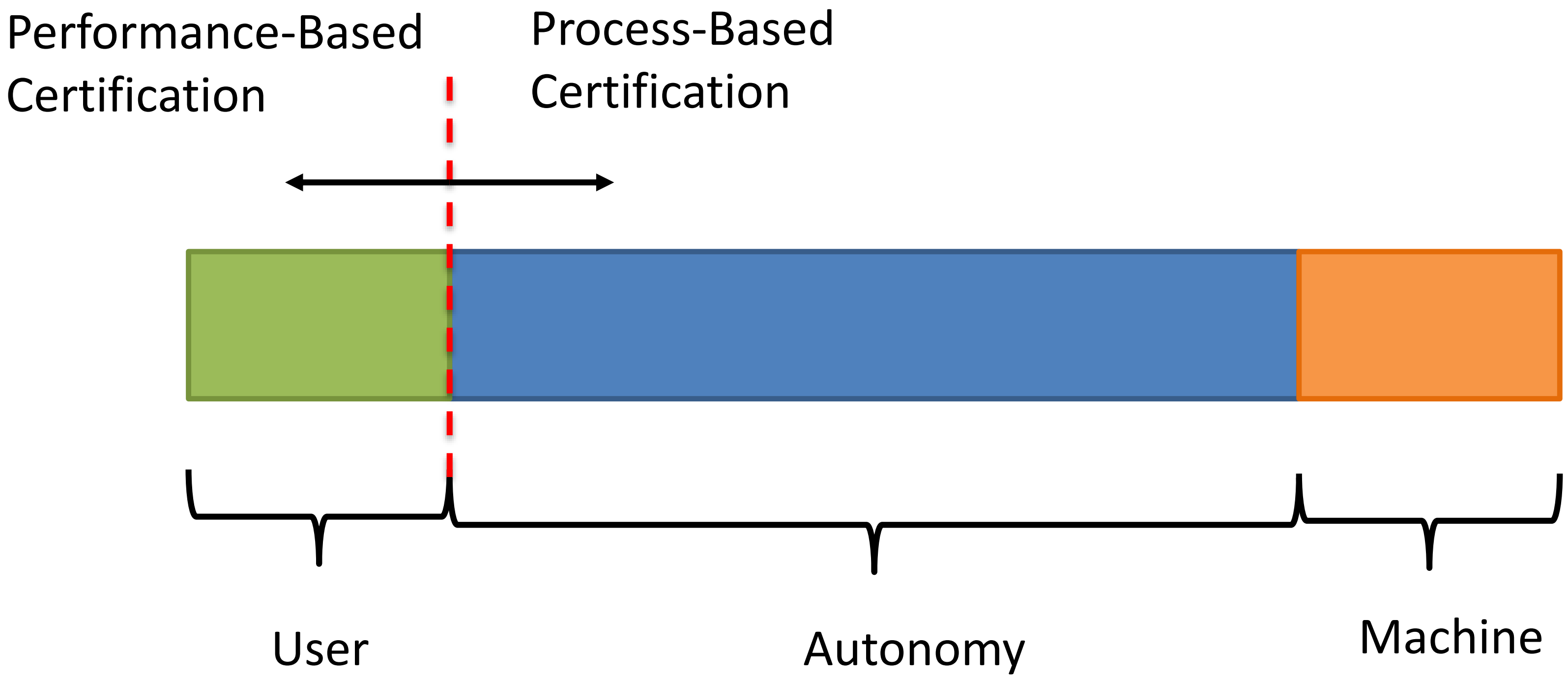


Private Pilot / General Aviation Aircraft

Example: **user** = pilot, **autonomy** = autopilot, **machine** = aircraft

- Autonomy would be type certified
- Regulated by FARs (federal aviation regulations)

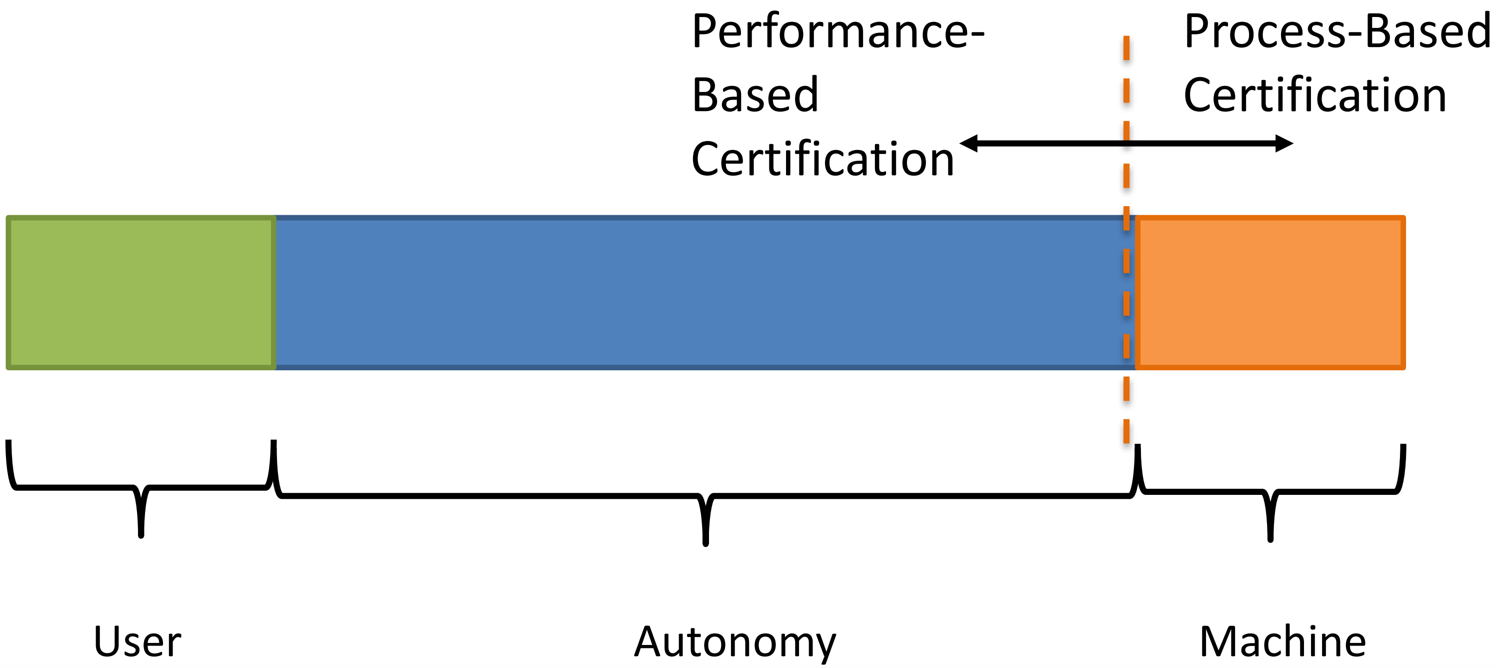
  -Pilot is licensed, autopilot is type certified, aircraft is type certified



Mapping Commercial Farm

Ex: **user** = farmer, **autonomy** = R/C operator, **machine** = aircraft

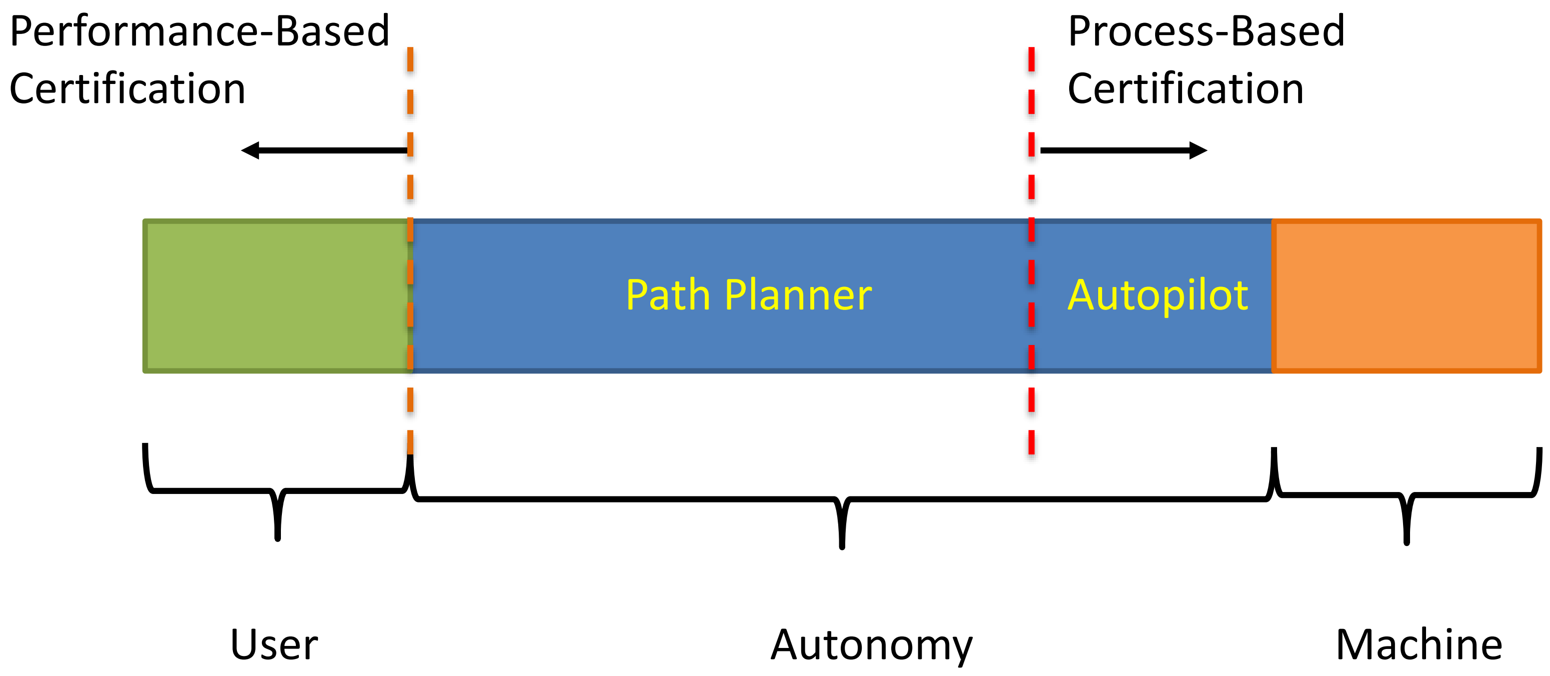
- Autonomy would be licensed
  -i.e. Driver's License, Pilot's License, etc.
- Enabled by Sec 333 / COA (waiver to some FARs)
  -Farmer is unlicensed, R/C operator is licensed, aircraft is type certified



Pipeline Inspection

Ex: **user** = dispatcher, **autonomy** = path planner + autopilot , **machine** = aircraft

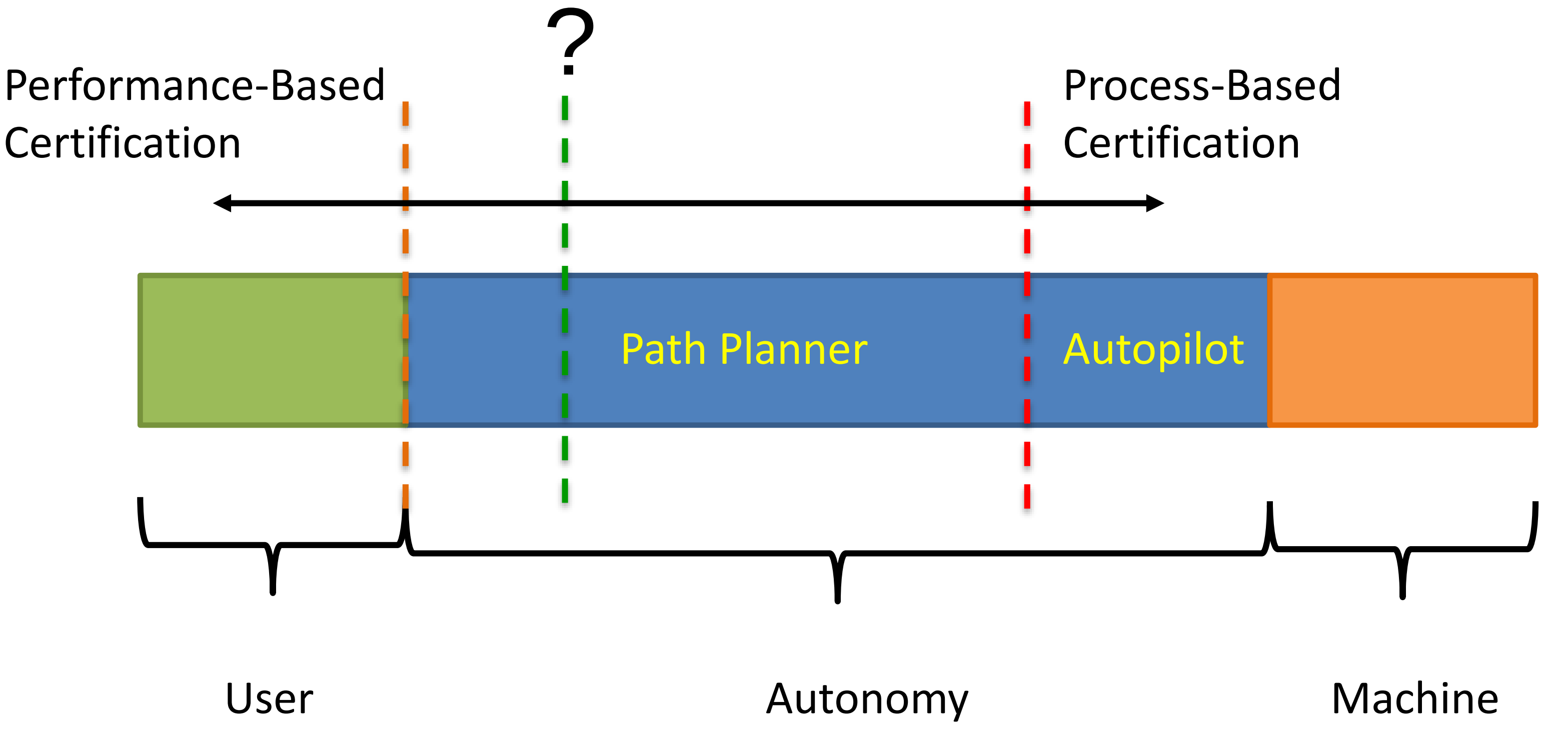
- Assured autonomy requires acceptable human, hardware, software interfaces
- Execution and Algorithmic features would be certified
- Interpreting/Intention Alignment features would be licensed



Pipeline Inspection

Ex: **user** = dispatcher, **autonomy** = path planner + autopilot , **machine** = aircraft

- Assured autonomy requires acceptable human, hardware, software interfaces
- Execution and Algorithmic features would be certified
- Interpreting/Intention Alignment features would be licensed



What "Exactly" is Trust?

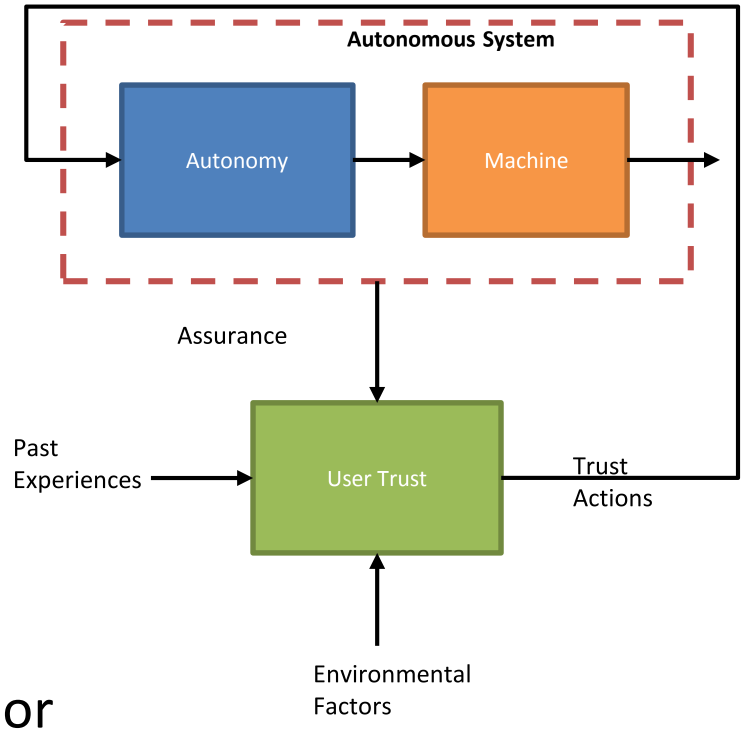
- Trust is a squishy concept
  - terminology has multiple meanings
  - know it when you see it, but hard to define it
- How do you define in such as way that we can develop design principles for autonomous systems?

- An engineering approach [Lillard, Frew, Argrow, Lawrence, and Ahmed, submitted to IROS 2016]
  - perform literature review
  - apply related work to unmanned aerial systems (UAS)
  - develop a model with clearly defined terms

Literature Review Summary

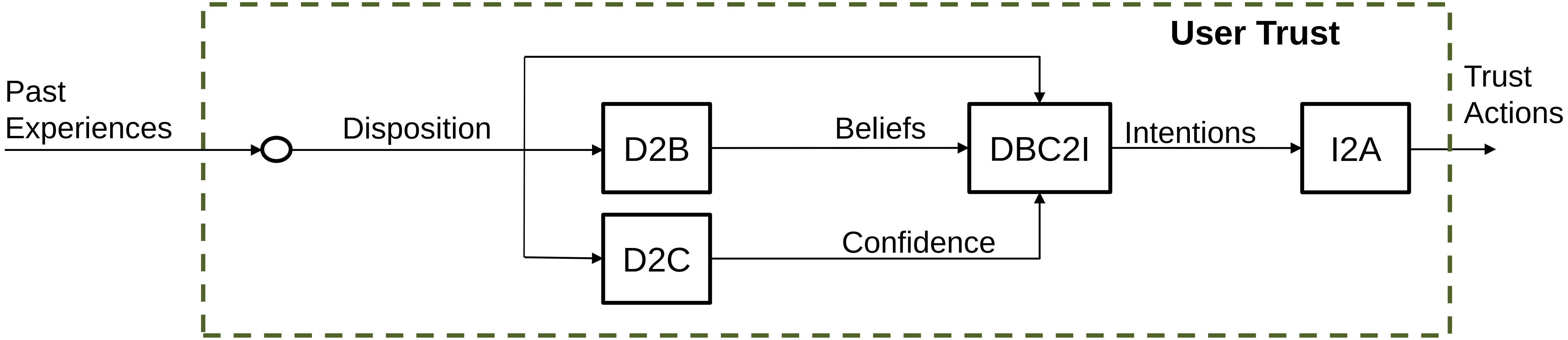
- Psychology, sociology, business [McKnight and Chervany, 2001], engineering [Lee and See, 2004]

- Literature states that trust is…
  – dynamic
  – multivariate
  – responds to feedback mechanisms
  – should be calibrated, ideally by design



- Our work adds:
  – **Assurances** also multivariate (from different parts of autonomy)
  – Slow outer loop dynamics often ignored
  – Explicit desire to calibrate trust, esp. for non-deterministic behavior
    - Synthesis, not just analysis
  – Multiple overlapped types of trust
    - Trust for customer, trust for operator/pilot, trust for certifier, trust for community/society,…

User Trust Model

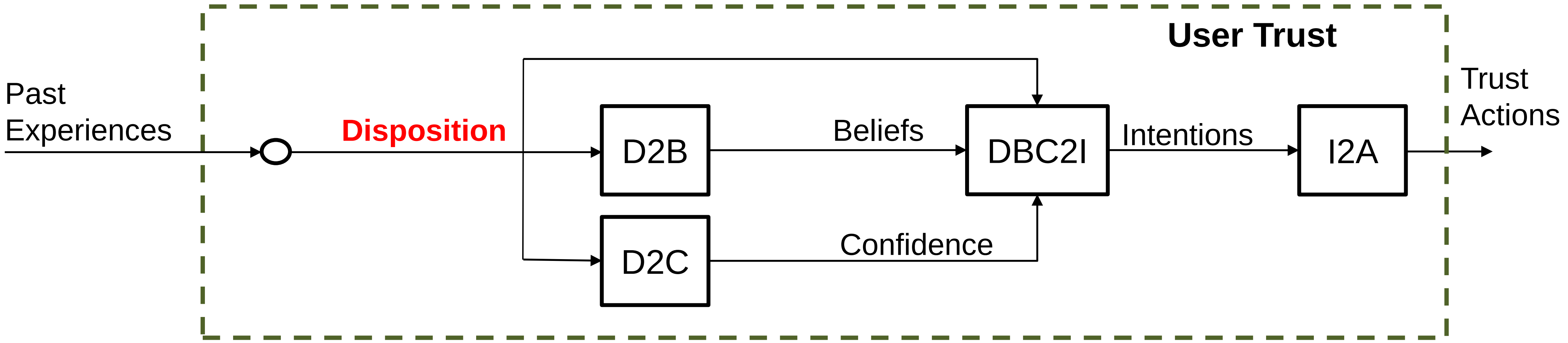


McKnight and Chervany, "What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology.," International Journal of Electronic Commerce, vol. 6, no. 2, pp. 35–59, 2001

Example: Driving Instructor's Trust in a Student Driver

- User = instructor
- Autonomy = student driver
- Machine = Automobile

Trust - User



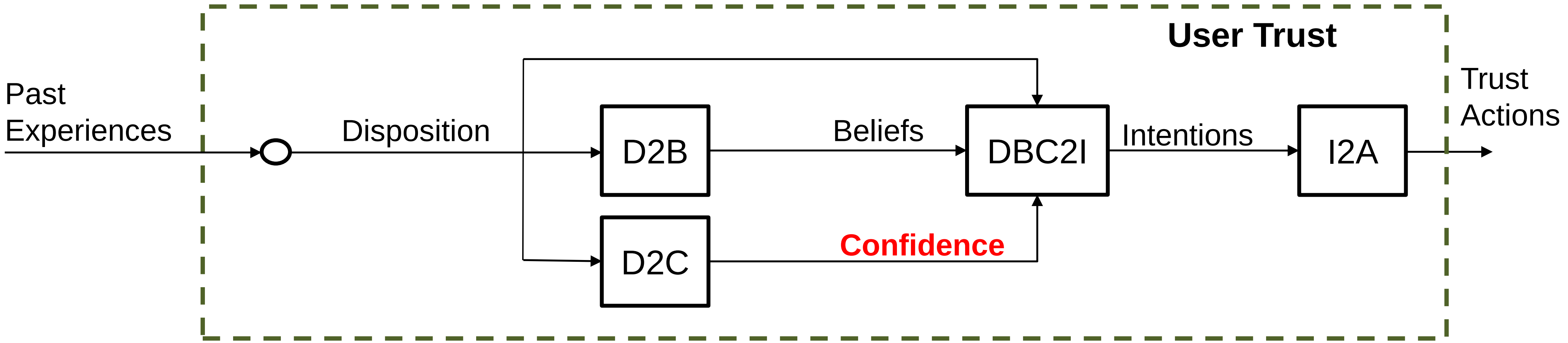
- **Faith in Autonomy** – Disposition to generally assume the autonomy is good-willed
- **Trusting Stance** – Disposition that trusting the autonomy will lead to better outcomes

Background attitudes. Slow to form/change

Example: The instructor's disposition that…

- People generally wish to drive safely to avoid possible injury to themselves and others.
- If students are never trusted, cannot fulfill requirements of being an instructor.

Trust - User



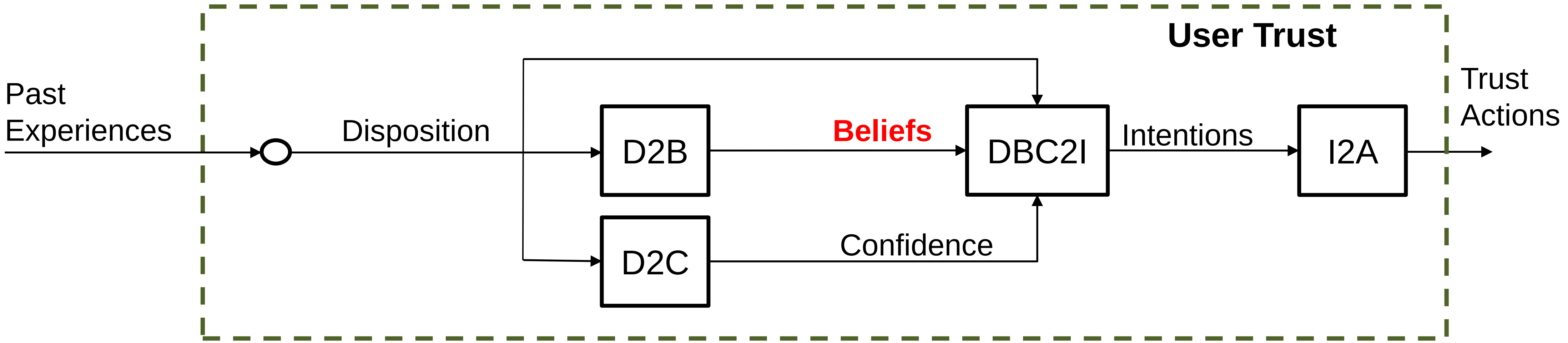
- **Structural Assurance** – confidence that the system that supports the autonomy will promote expected outcomes
- **Normality** – confidence that operating conditions will be normal

Relates to support systems (as opposed to specific autonomous system)

Example: The instructor's confidence that…

- The written exam is a good test of knowledge. Traffic lights will function properly.
- The weather appears to be normal and the car is operating correctly.

Trust - User



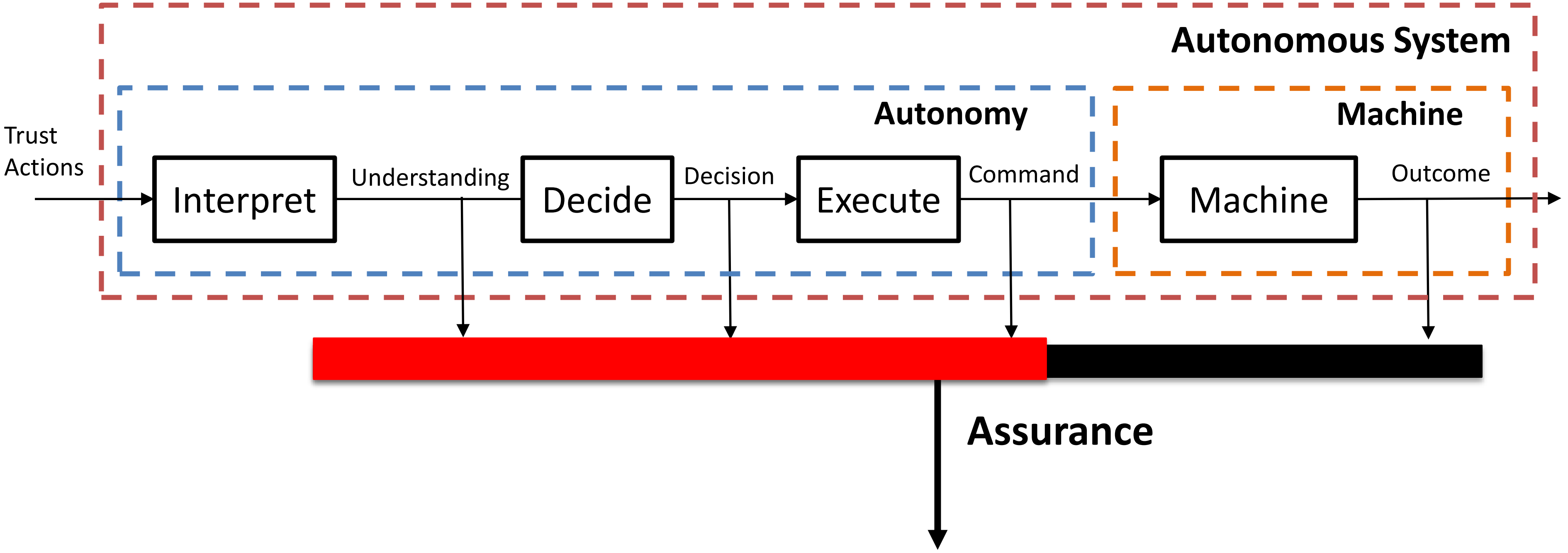
- **Benevolence** – Belief that the autonomous system will behave in your best interest
- **Integrity** – Belief that the autonomous system will tell the truth and keep agreements
- **Competence** – Belief that the autonomous system will be able to achieve what is required
- **Predictability** – Belief that the behavior of the autonomous system can be forecasted

Relates to specific autonomous system

Example: The instructor's belief that…

- The student driver will not intentionally crash the car.
- The student driver has not lied about previous experience and practice.
- The student driver has the motor skills to operate the vehicle.
- The student driver will make similar/repeatable decisions in typical driving scenarios.

Autonomous System Model



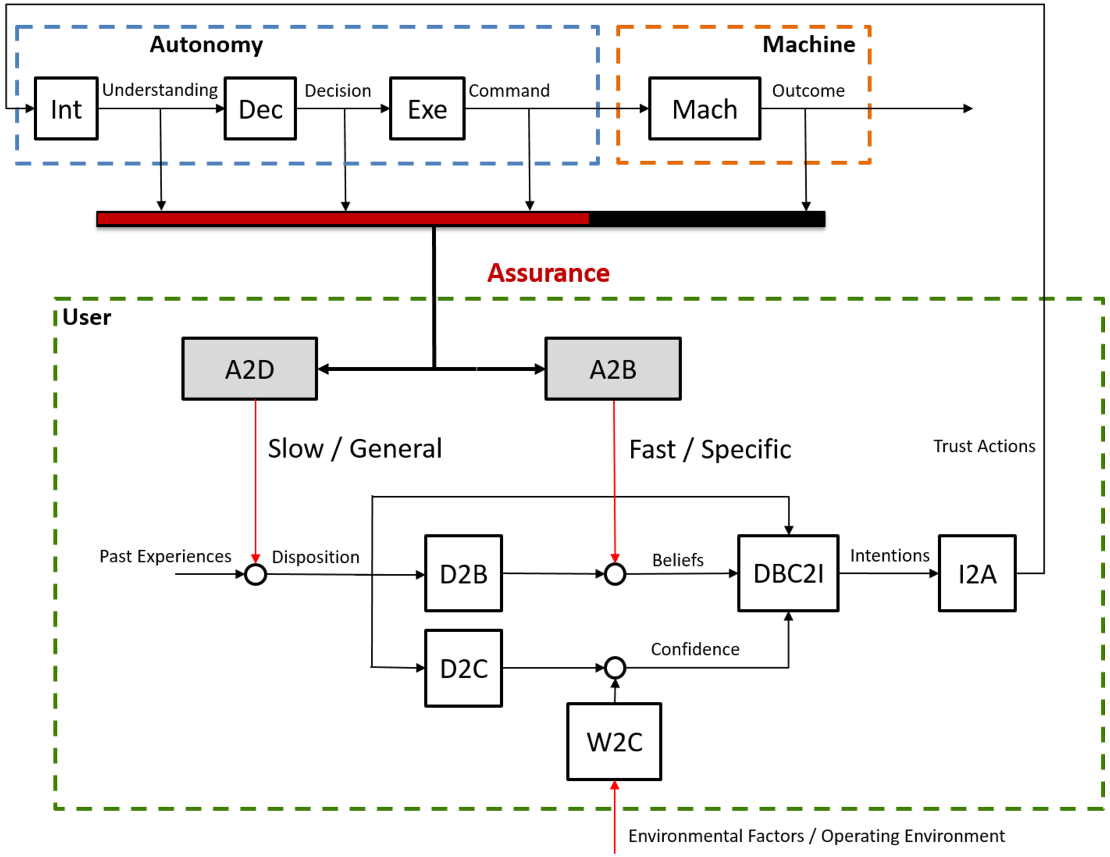
Multiple types of assurances

- Sheridan's "Levels of Autonomy" = one version of "Understanding – Decision – Command" assurances [Parasuraman, et al. 2000]
- Focus on **assurances that come from the autonomy**

Implications

- Assumes that: increased trust → appropriate use → best performance
- Trust only important if it leads to **Trust Actions**
  - thoughts are not enough!
  - → need interfaces that allow for trust actions
- **Assurances are what can be designed**

- Feedback loop
  - assurances from Autonomy to Human
  - actions from Human to Autonomy
- Dependent of specific application
- Multiple instances of trust for the same system
  - User, structure, bystander, etc.



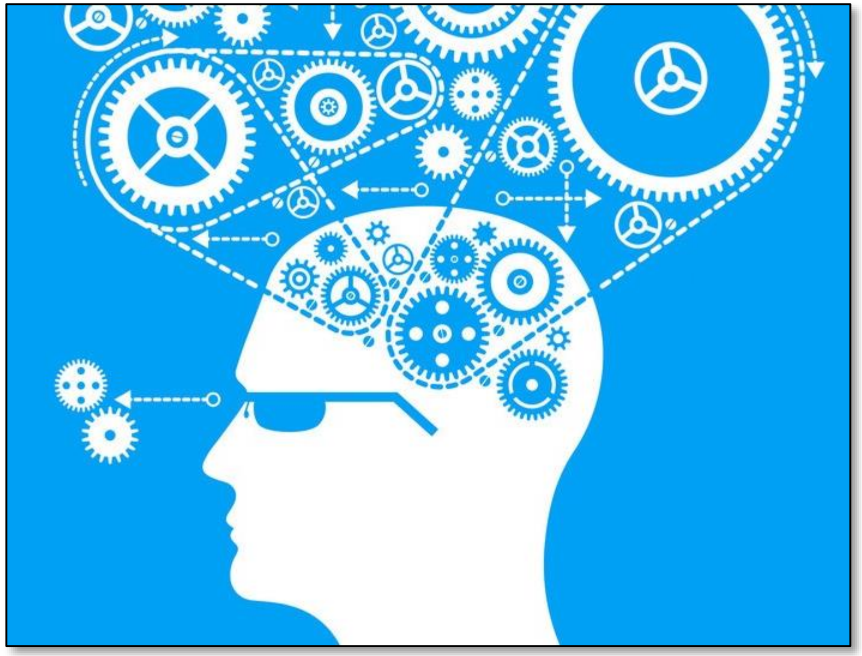
Hypothesis

Trust can be **calibrated** by more **insightful assurances about autonomy's internal processes**

– **What** it's doing at various levels to interpret human actions, decide on a course of action, implement strategy/tactics/operations/functions

– **Why** it's doing it or why it's doing it this way or how clear is it on the situation/objectives/risks

– **How** will it be able to conduct itself to achieve the objective, how experienced/competent/robust is it in this situation



Caveats with "Just" Calibrating Trust

(Amy Pritchett, GA Tech)

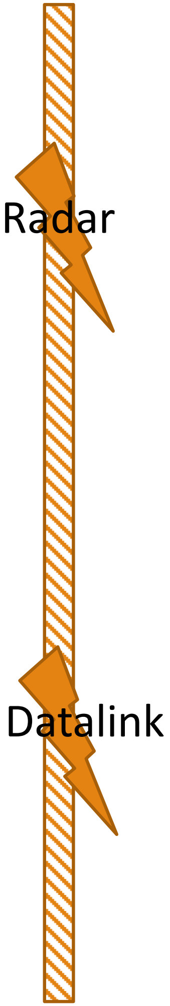
- If autonomy is supposedly correct 80% of the predicted use cases, does that mean users should rely on it 80% of the time?
- How does user know whether an immediate case is part of the 80% or 20%?
  - Shouldn't 100% of cases be seen to see whether the autonomy is correct in *this* case?
- Does user always use same criteria to judge if autonomy is correct?
- Feasible for user to make judgment *in situ*?
- Even if user *believes* autonomy to be correct, do other factors impact whether the human *relies* on it?
  - Ease of doing task oneself, and confidence in ones performance
  - Ease of intervening in automated execution once allowed
  - Responsibility for outcome

Caveats with "Just" Calibrating Trust

- Difference between *belief* and *reliance*
  - *Reliance* involves cost-benefit analysis
- Difference between *aggregate performance* and *confidence in immediate situation*.
- Bases in human-human trust (belief):
  - Without frequent interaction: *faith*
    - Influenced by credentials, recommendations
  - With frequent interaction: perceived *dependability* and *predictability*
    - Can be shaped by experience – requires understanding the machine and seeing consistent behavior…
- If a user is responsible for the outcome, (e.g. airline pilot), then her/his job is to "trust, but verify" – can they verify from where they are sitting?

The "Turing Test for Aviation" (Pritchett)

What would one aviation agent expect from another?



Ability to do a task

Ability to report when it can't do a task

Ability to flex the task structure to achieve desired ends

Ability to adapt its goals to the situation

Ability to communicate and coordinate in way that makes sense to other agent

Ability to ignore other agent when necessary

Ability to recognize and use interdependencies in inter-agent activities

Ability to operate at many levels of abstraction simultaneously

Autonomous Agent "Self-Confidence"

Explore machine **"self-confidence"** (**self-trust**) as a possible assurance [Sweet, Ahmed, Kuter, and Miller, InfoTech 2015; Hutchins, Cummings, Draper and Hughes, HFES 2015]

- Self-confidence = *perceived* ability to execute assigned tasks (within defined scope of autonomy), despite uncertainties in…
  - knowledge of world
  - own/self state
  - reasoning process and execution abilities
- How to quantify/qualify?
  - Task competency: boundaries of execution/reasoning?
  - Info adequacy: data sources/models good enough?
  - Communication: how to relate to users?
  - Implementation: how to actually encode in machines?

Come with me if you want to live

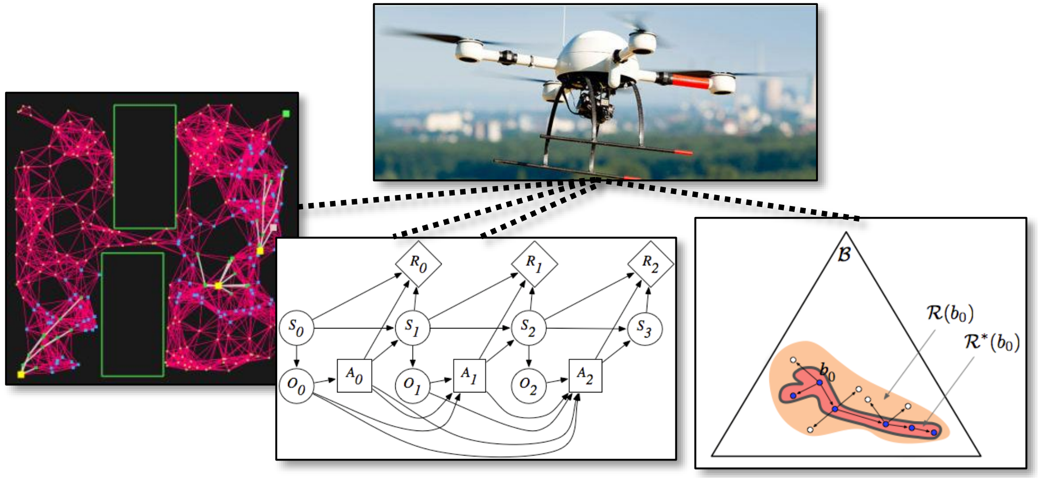
Your chances of living are about 73% if you come with me

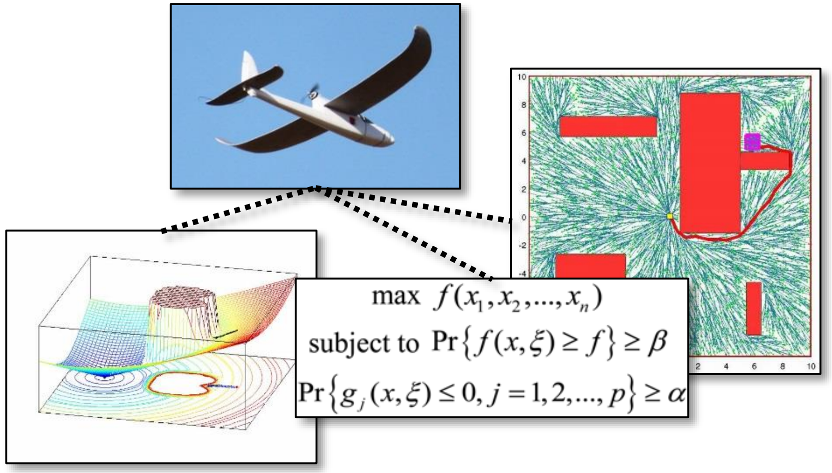
7 out of the last 10 people who came with me lived

Trust me…

Self-Confidence: Motivation for Definition

- Probabilistic reasoning ubiquitous in autonomy
  - Stochastic models can capture/handle many kinds of uncertainties and nondeterminism





  - …but models are still only approximations of reality
  - …and more approximations needed to solve planning/perception problems
  - What insights to give to users (who are not engineers, statisticians, etc.) on soundness of underlying models, data and decisions?
  - What to do if reasoning/execution competency boundaries reached? How to stay within/away from these? How to know if reached?

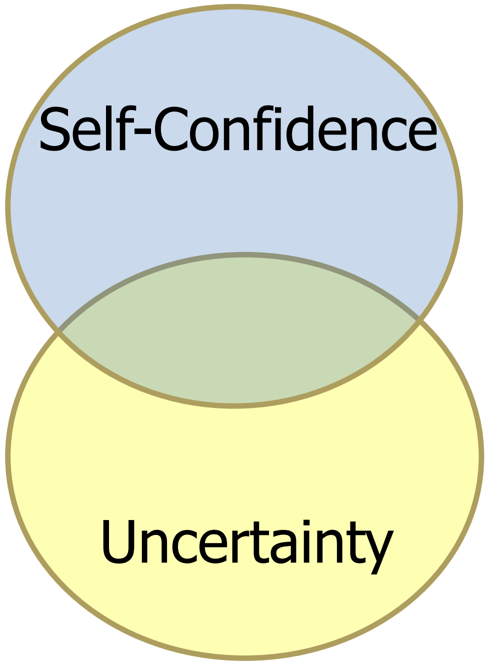
Self-Confidence: Implications of Definition

- Machine self-confidence related to, but not same as, uncertainty used for task reasoning
  - One can be certain that probability of success for goal/context pair is 0.02– & **thus be confident of failure**
  - One can be very uncertain about the contexts one encounters during plan execution, but confident of robustness with regards to them – & **thus be highly confident in face of "known unknowns"**
- Self-confidence assessment could be "skewed" if understanding of uncertainty is wrong

e.g. if no/too many "unknown unknowns" assumed

→ over/under confident

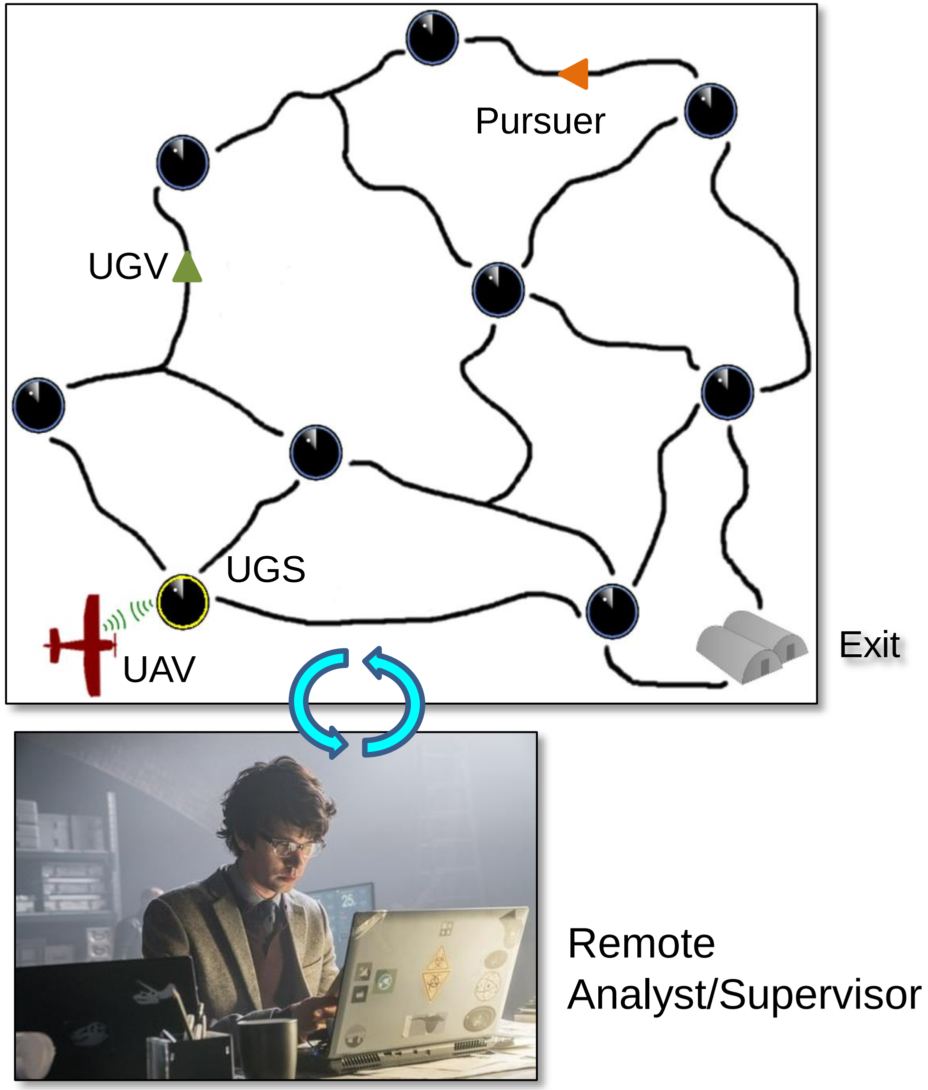
- Nevertheless, uncertainty will (and should) tend to undermine self-confidence



A Concrete Application Scenario: Autonomous Pursuit-Evasion on a Road Network

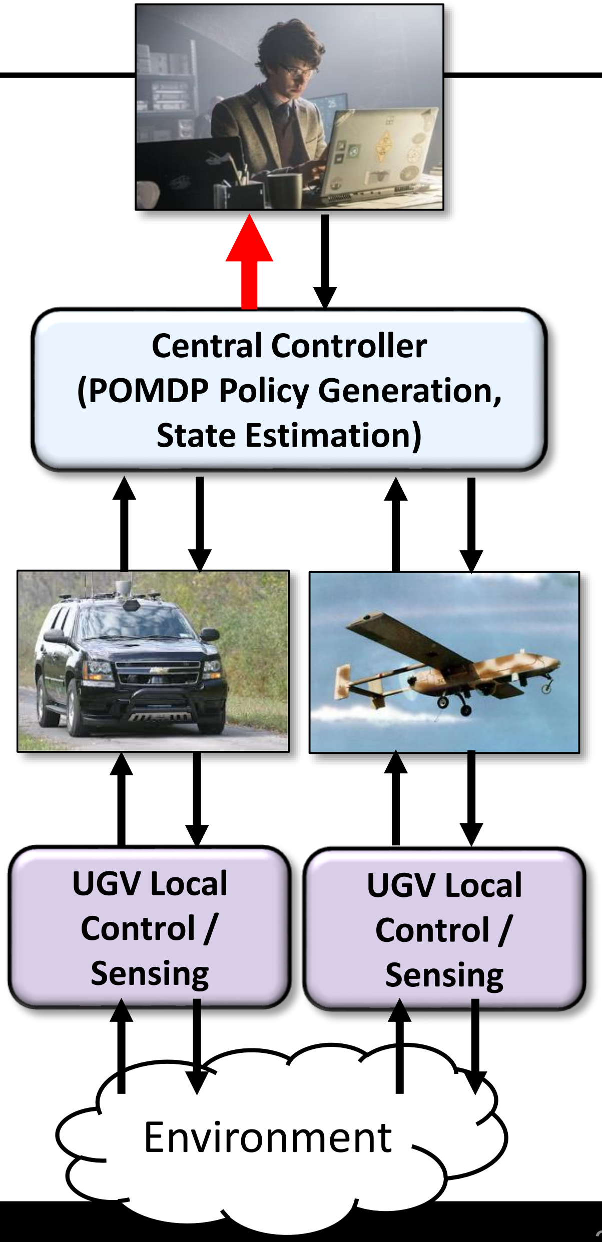
Goal: Autonomous UGV to exit w/o capture

- **Chaser's** *location and behavior uncertain*
- **Autonomous UAV** gathers info
  - interrogating **unattended ground sensors (UGS)**, taking pictures
  - UGV processes noisy data
  - short-range comm links
- **Central controller** updates movement/sensing policies for UGV and UAV
  - network represented as finite grid along roads
- **Remote human analyst** recommends high-level stances to steer planning
  - **Operating stance**: "get out fast", "wait and see", "stay close to safest route," etc.
  - **Enemy's stance**: "trying to ambush", "searching", "trying to corner", etc.

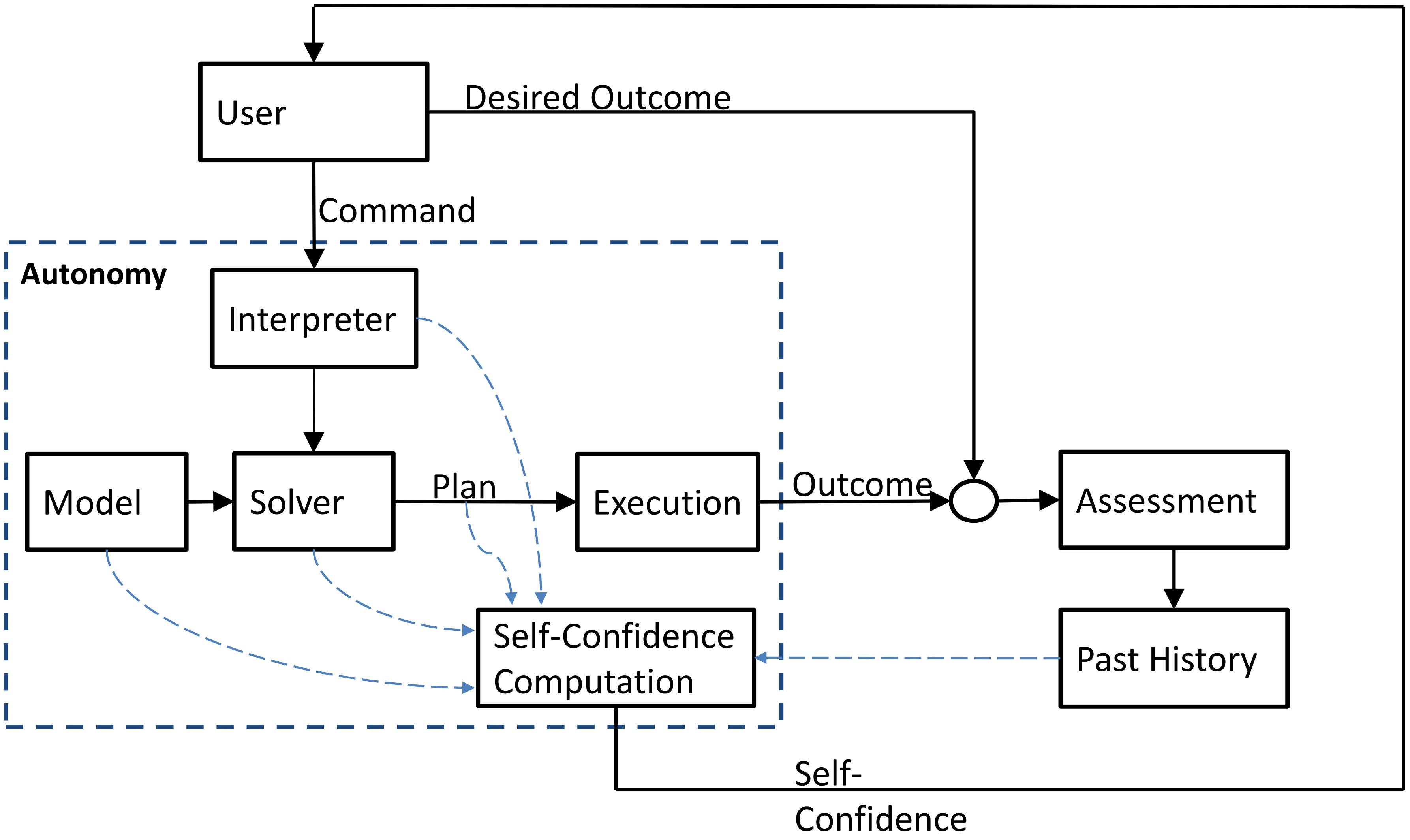


Scenario Features

- Highly complex planning space
  - **suboptimal planning heuristics** practically necessary
  - **no guarantees** that autonomy always "get it right"
  - **human helps narrow down** uncertain decision/info space
  - incentives/rewards can be shaped by human input
- Multiple layers of uncertainty
  - **noisy data** from UGS/camera
  - **models**: true chaser dynamics/behavior unknown
  - **"fog of war"**: limited situational awareness, variety of "actual" operating conditions to piece together
- Strategic high-level interaction with human analyst
  - low-level UAV/UGV autonomy not assessed
  - **self-confidence of central planner as assurance**: feedback to provide better insight into and usage/guidance of autonomous planning



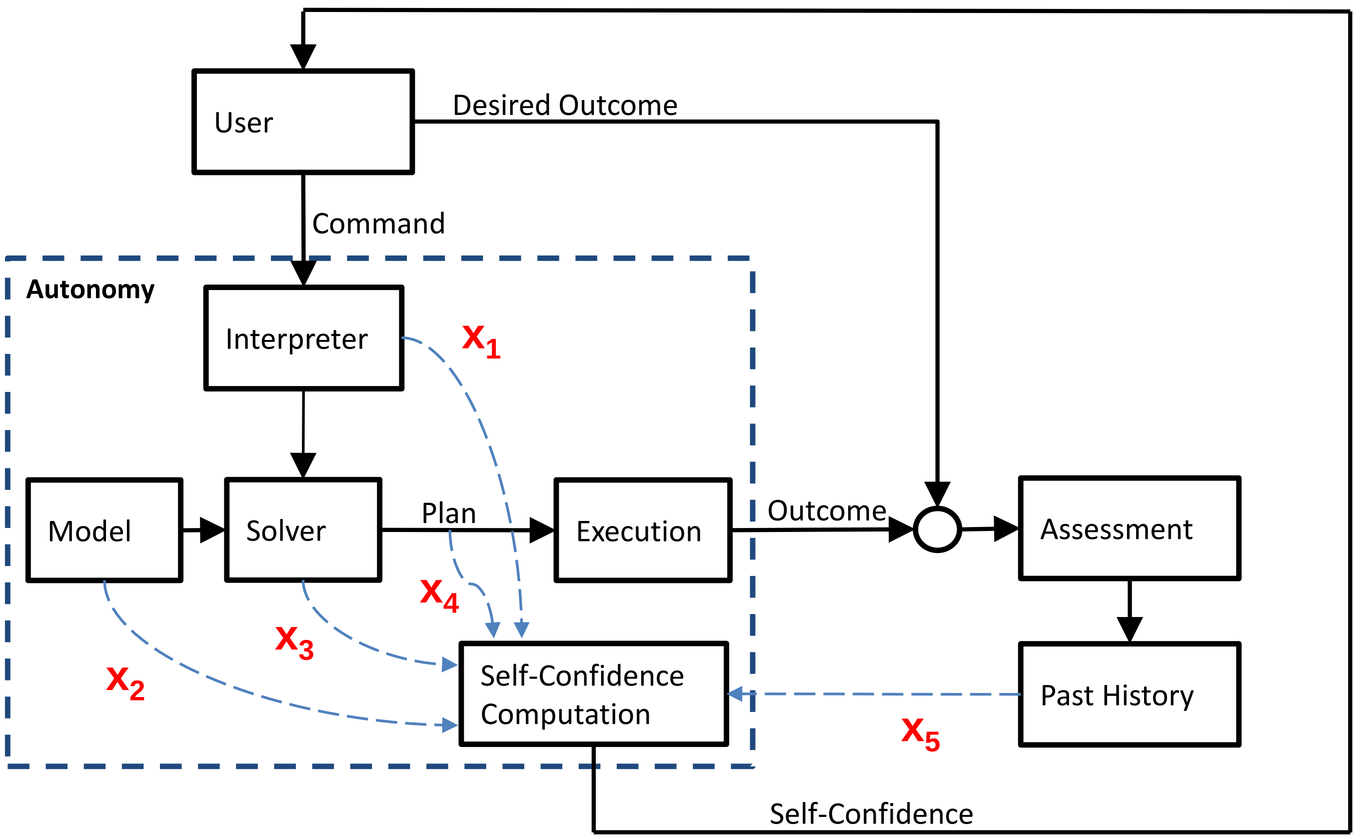
One Possible Diagram for Computing Self-Confidence



Self-Confidence Factors

$$X_{sc} = [x_1, x_2, x_3, x_4, x_5, \ldots]$$

- x_1 Command Interpretation
  - are autonomy and user 'on the same page'?
- x_2 Model Validity
  - how well does autonomy's model reflect reality?
- x_3 Solver Quality
  - how well can the solver use the model to generate policies/plans?
- x_4 Outcome Assessment
  - how 'good' is expected distribution of results?
- x_5 Past Performance
  - how well has the autonomy done in similar circumstances?



Some Other Recent Related Algorithmic Work

- Qualitative visualization of plan/sensor-related uncertainties

[Hutchins, Cummings, Draper and Hughes, HFES 2015]

- Counter-planning for self-assessment and plan repair

[Kuter and Miller, AAAI FS 2015]

- Statistical model residuals for object recognition in robotic sorting

[Kaipa, Kankanhalli-Nagendra, and Gupta, AAAI FS 2015]

- Surprise index for assessing Bayesian network models

[Zagorecki, Kosniewski, and Drudzdel, AAAI FS 2015]

For more:

see AAAI 2015 Fall Symposium on Self-Confidence in Autonomous Systems:

scas2015.recuv.org

(proceedings available at https://www.aaai.org/Press/Reports/Symposia/Fall/fall-reports.php)

Conclusions

Trust = willingness of one autonomous agent to depend on another

- something to be calibrated to appropriate level

User trust model: engineering perspective and attempt to develop autonomous system design principles from UXV perspective

- assurances and trust actions: key "input/output signals" of trust
- online/offline flavors: live interaction, certification, etc.

Machine self-confidence: possible assurance for calibrating trust

- introspective self-reporting of "competency boundaries"
- wide open -- a possible basis for licensing sophisticated "everyday" autonomy?

Credit Where Credit is Due

Many thanks to the following collaborators for their valuable insights and contributions


Amy Pritchett
(GA Tech)


Chris Miller
(SIFT)


Missy Cummings
(Duke)


Ugur Kuter
(SIFT)


Mark Draper
(AFRL)


Eric Frew


Dale Lawrence


Brian Argrow


Matthew Aitken


Austin Lillard


Nicholas Sweet

Thanks!

A National Science Foundation Industry/University Cooperative Research Center (IUCRC)