Verification for Trust

Human-centered and formal design artifacts

Ufuk Topcu

University of Texas

http://u-t-autonomous.info

Parts joint with Lu Feng, Laura Humphrey and Nils Jansen.

Outline

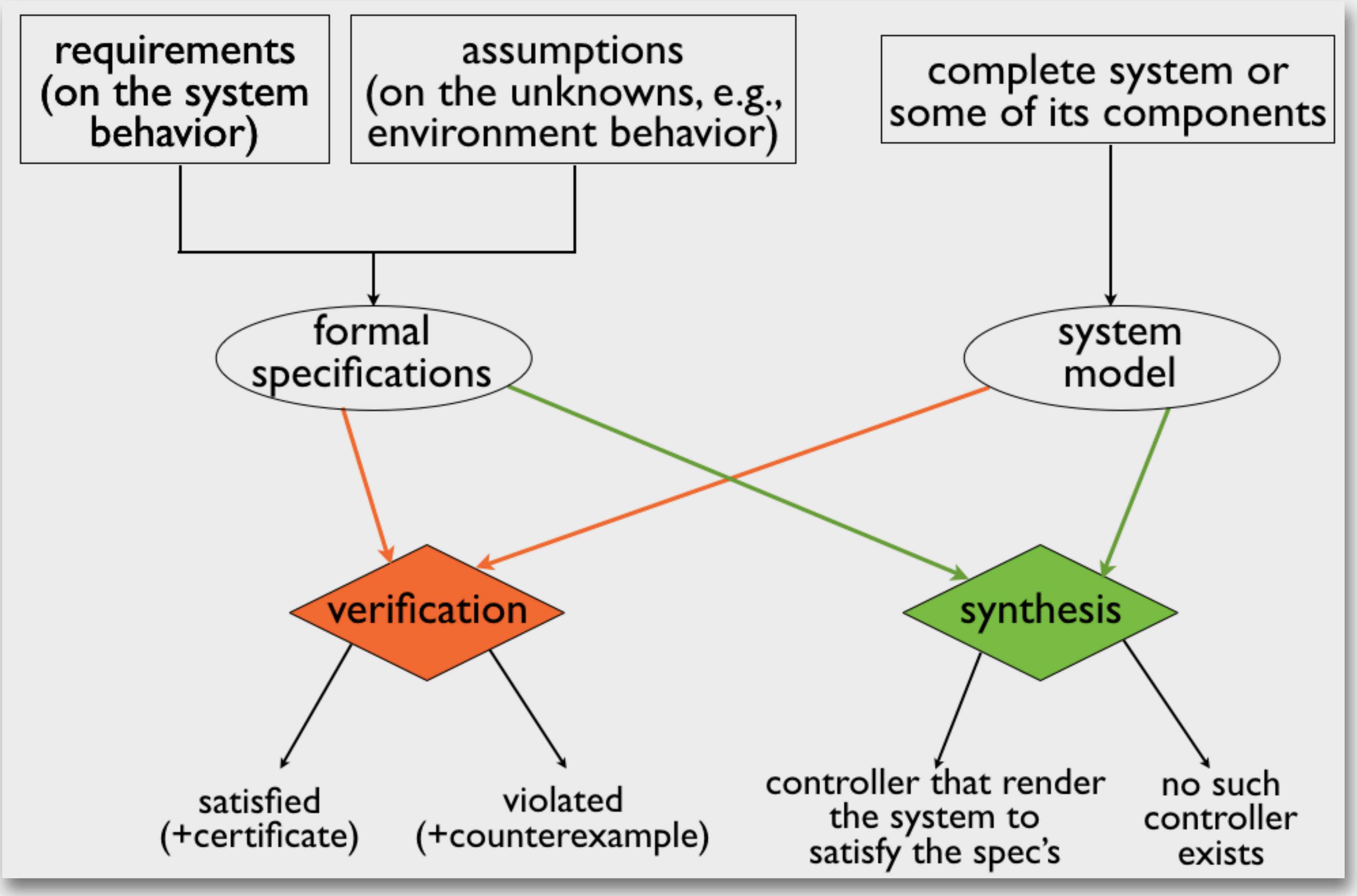
The organizers: In your talk, we would love to hear…

- your definition of trust,
- the role of trust in human-robot interaction/collaboration, and
- your view on "social trust in autonomous systems."

Thoughts on …

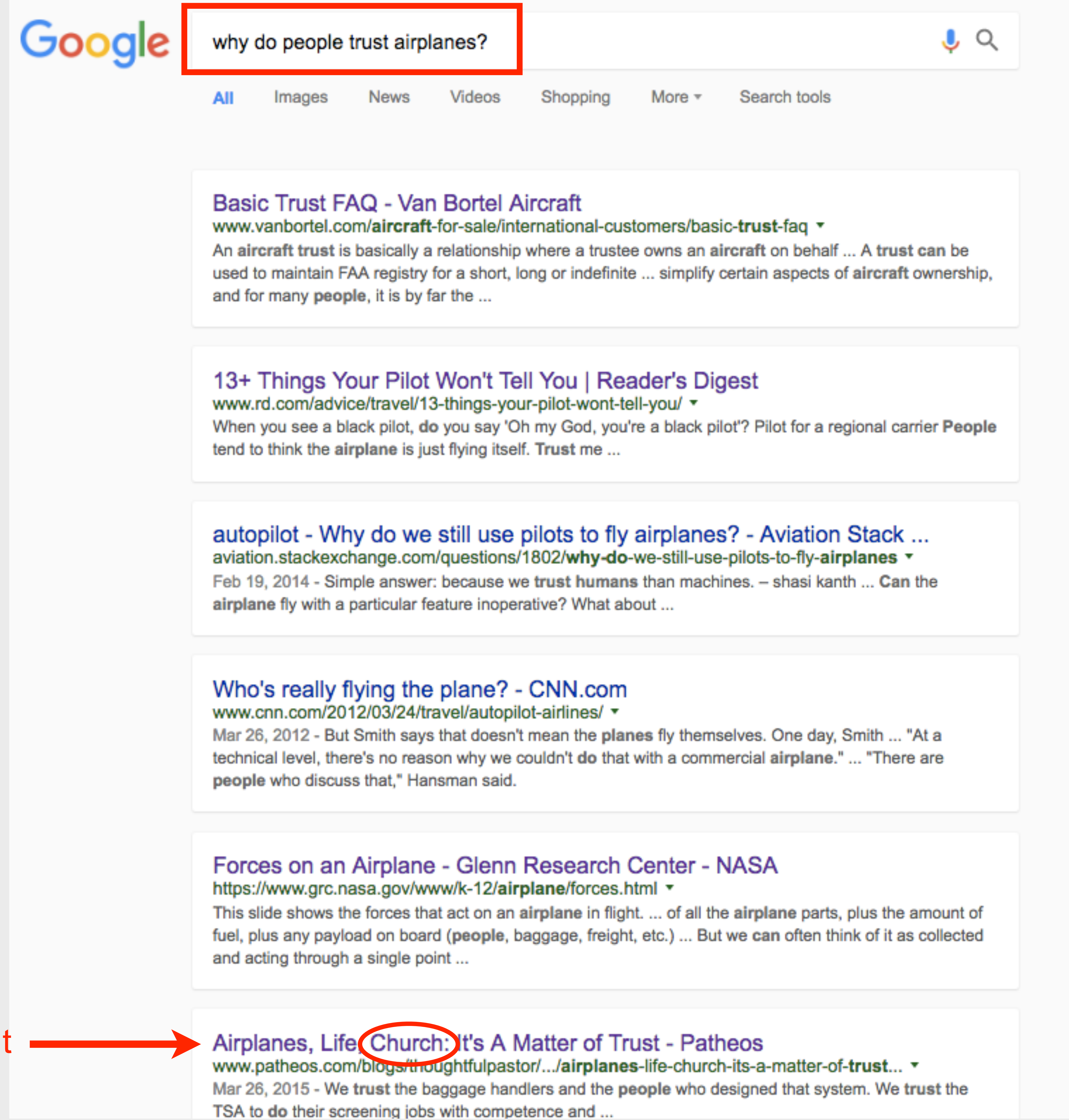
- How does trust factor in formal verification and synthesis?
- How do formal verification and synthesis help establish trust?

Formal verification and synthesis


requirements (on the system behavior)
assumptions (on the unknowns, e.g., environment behavior)
complete system or some of its components
formal specifications
system model
verification
synthesis
satisfied (+certificate)
violated (+counterexample)
controller that render the system to satisfy the spec's
no such controller exists

What will all this math have anything to do with social trust?

page 1 of Google search results



why do people trust airplanes?

Basic Trust FAQ - Van Bortel Aircraft
www.vanbortel.com/**aircraft**-for-sale/international-customers/basic-**trust**-faq ▾
An **aircraft trust** is basically a relationship where a trustee owns an **aircraft** on behalf ... A **trust can** be used to maintain FAA registry for a short, long or indefinite ... simplify certain aspects of **aircraft** ownership, and for many **people**, it is by far the ...

13+ Things Your Pilot Won't Tell You | Reader's Digest
www.rd.com/advice/travel/13-things-your-pilot-wont-tell-you/ ▾
When you see a black pilot, **do** you say 'Oh my God, you're a black pilot'? Pilot for a regional carrier **People** tend to think the **airplane** is just flying itself. **Trust** me ...

autopilot - Why do we still use pilots to fly airplanes? - Aviation Stack ...
aviation.stackexchange.com/questions/1802/**why-do**-we-still-use-pilots-to-fly-**airplanes** ▾
Feb 19, 2014 - Simple answer: because we **trust humans** than machines. – shasi kanth ... **Can** the **airplane** fly with a particular feature inoperative? What about ...

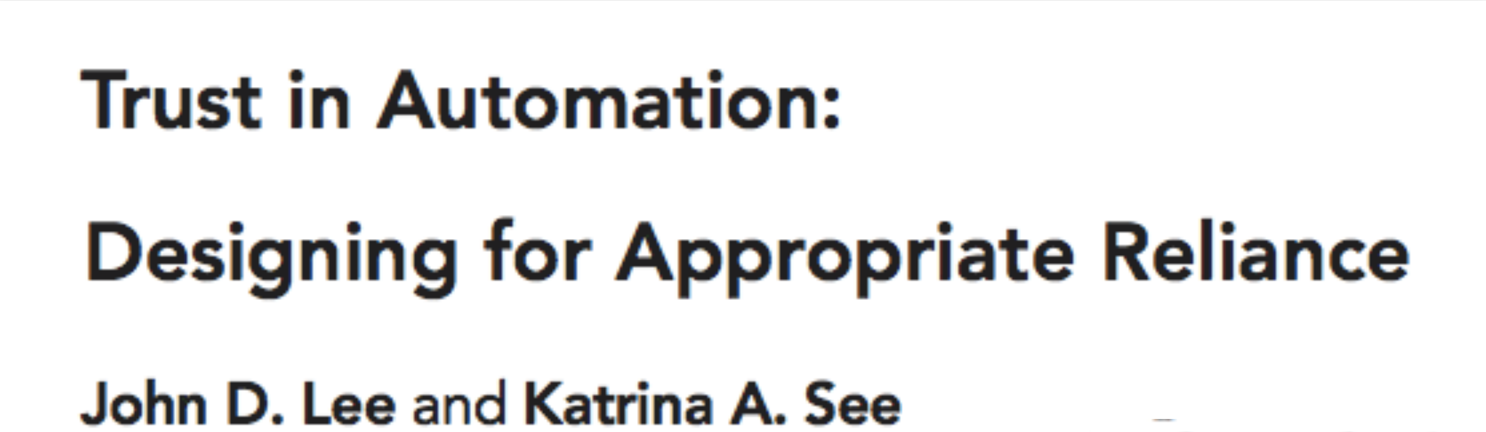
Who's really flying the plane? - CNN.com
www.cnn.com/2012/03/24/travel/autopilot-airlines/ ▾
Mar 26, 2012 - But Smith says that doesn't mean the **planes** fly themselves. One day, Smith ... "At a technical level, there's no reason why we couldn't **do** that with a commercial **airplane**." ... "There are **people** who discuss that," Hansman said.

Forces on an Airplane - Glenn Research Center - NASA
https://www.grc.nasa.gov/www/k-12/**airplane**/forces.html ▾
This slide shows the forces that act on an **airplane** in flight. ... of all the **airplane** parts, plus the amount of fuel, plus any payload on board (**people**, baggage, freight, etc.) ... But we **can** often think of it as collected and acting through a single point ...

6th result

Airplanes, Life, Church: It's A Matter of Trust - Patheos
www.patheos.com/blogs/thoughtfulpastor/.../**airplanes**-life-church-its-a-matter-of-**trust**... ▾
Mar 26, 2015 - We **trust** the baggage handlers and the **people** who designed that system. We **trust** the TSA to **do** their screening jobs with competence and ...

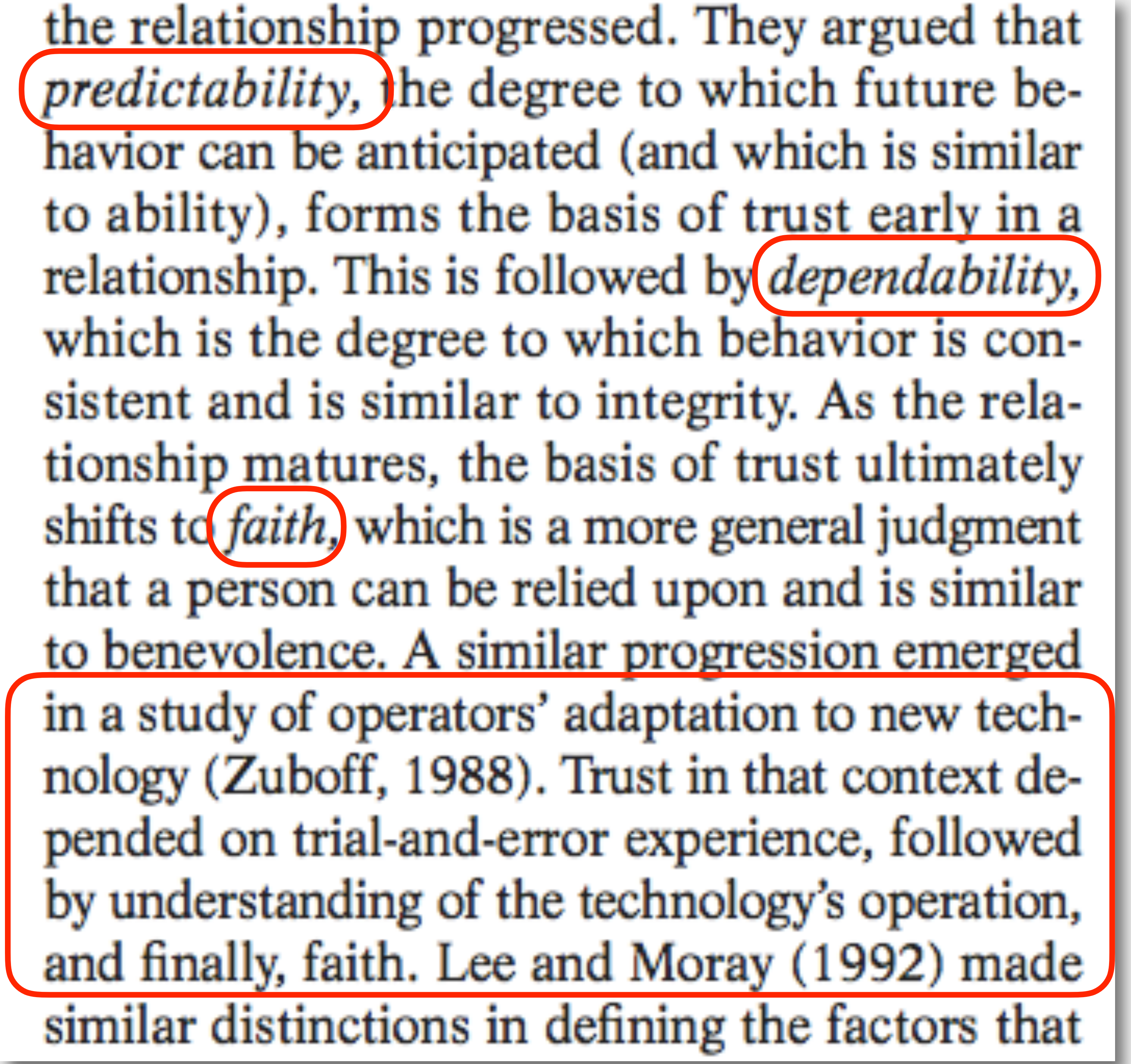
Ufuk Topcu



Trust in Automation:

Designing for Appropriate Reliance

John D. Lee and **Katrina A. See**



the relationship progressed. They argued that *predictability,* the degree to which future behavior can be anticipated (and which is similar to ability), forms the basis of trust early in a relationship. This is followed by *dependability,* which is the degree to which behavior is consistent and is similar to integrity. As the relationship matures, the basis of trust ultimately shifts to *faith,* which is a more general judgment that a person can be relied upon and is similar to benevolence. A similar progression emerged in a study of operators' adaptation to new technology (Zuboff, 1988). Trust in that context depended on trial-and-error experience, followed by understanding of the technology's operation, and finally, faith. Lee and Moray (1992) made similar distinctions in defining the factors that

UAV Mission Planning

A human operator remotely works with an unmanned air vehicle

- sensor tasks, e.g., steering the onboard sensor to capture imagery of targets
- high-level piloting commands, e.g., how many loiters to perform at each waypoint

Image source: AFRL

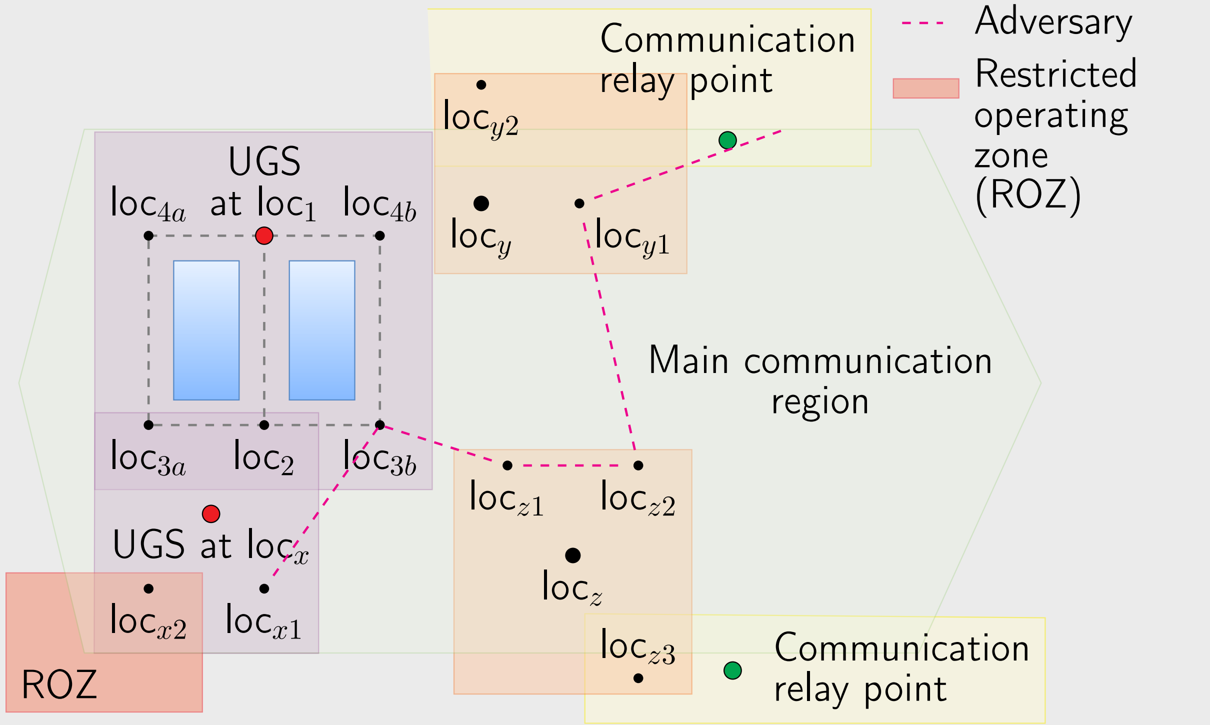
Autonomy in unmanned air vehicles

- low-level piloting (e.g., way-point navigation, loitering)
- automated mission planning

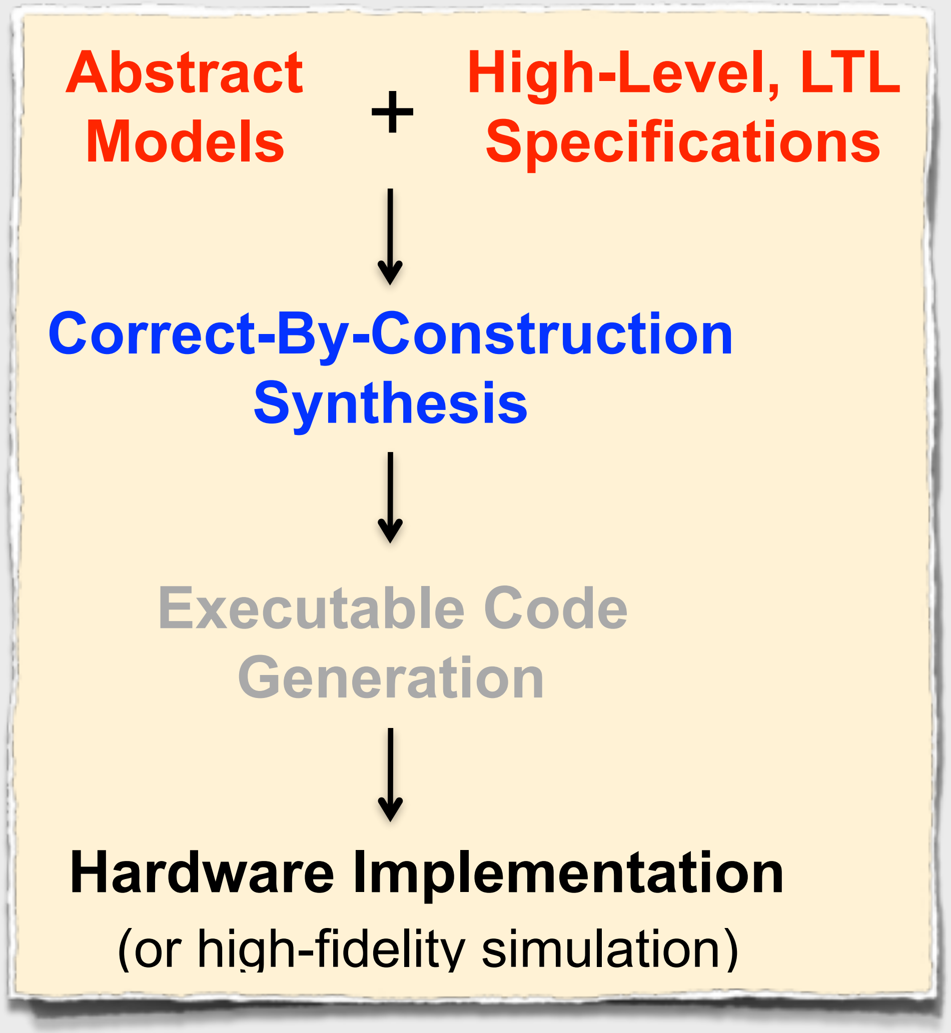
Mission specifications (to be) expressed in a formal language

- covering all the waypoints while avoiding restricted operating zones
- loitering over certain waypoint to capture sensor images of the target
- patrolling of certain road

Quantitative objectives: completion time, fuel usage, …



Specify + Synthesize + Implement



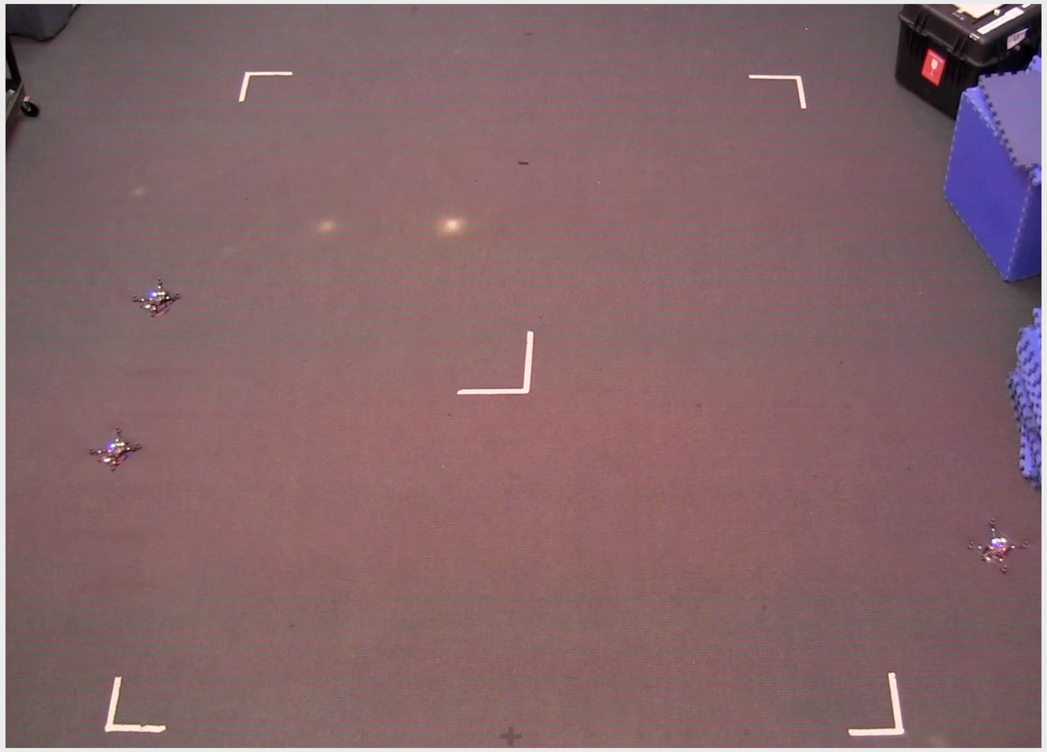
Abstract Models + High-Level, LTL Specifications

Correct-By-Construction Synthesis

Executable Code Generation

Hardware Implementation

(or high-fidelity simulation)

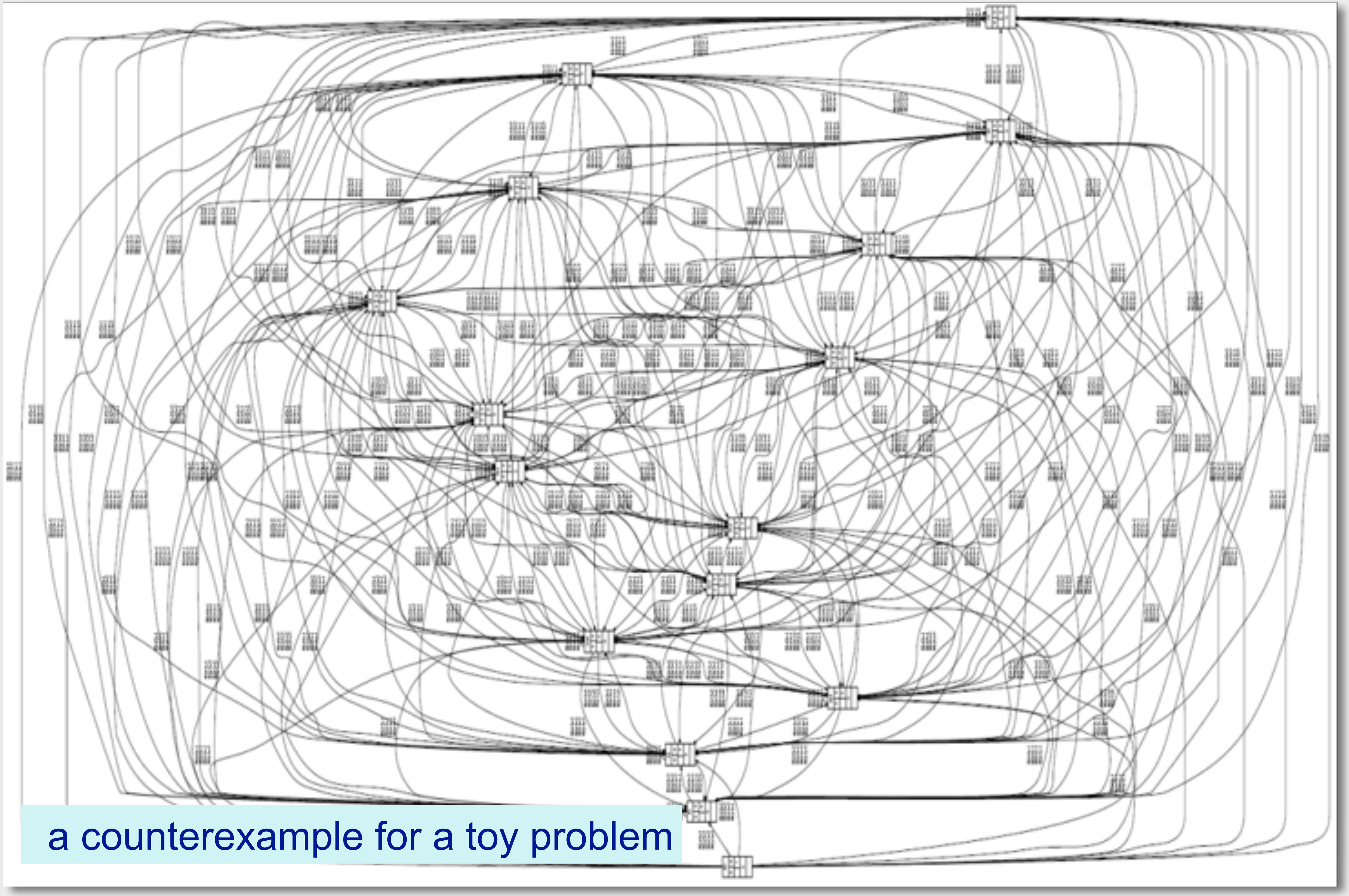


(with Kumar Lab at Penn)

on AMASE autonomy interface

What if things do not work out?

There usually is a reason. And, a **counterexample** that explains why.



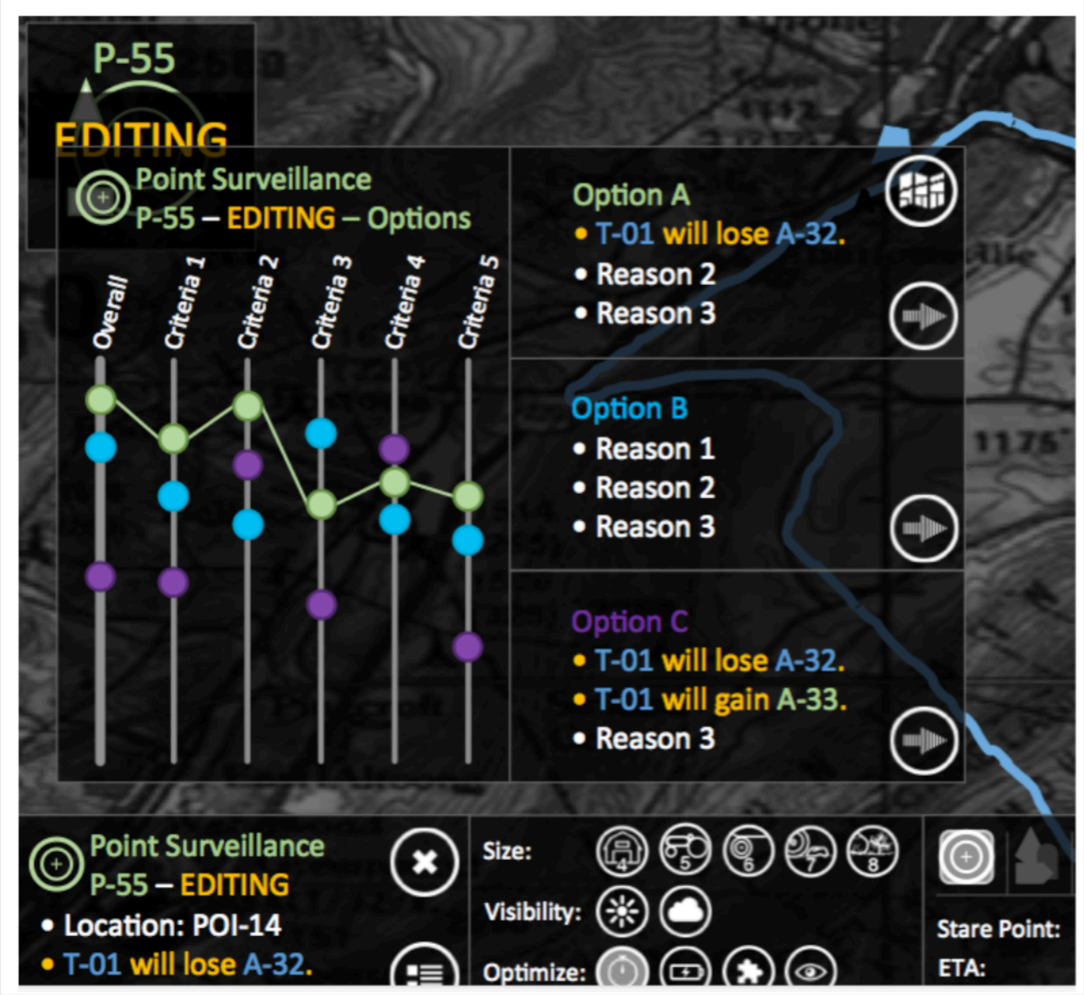
a counterexample for a toy problem

Structured Counterexamples

Compute counterexamples that can be understood by “humans”

- Uses the same alphabet and grammar with humans
- Respects the limitations (expressivity, bandwidth, etc.) of the interface

User interface based on “play calling”



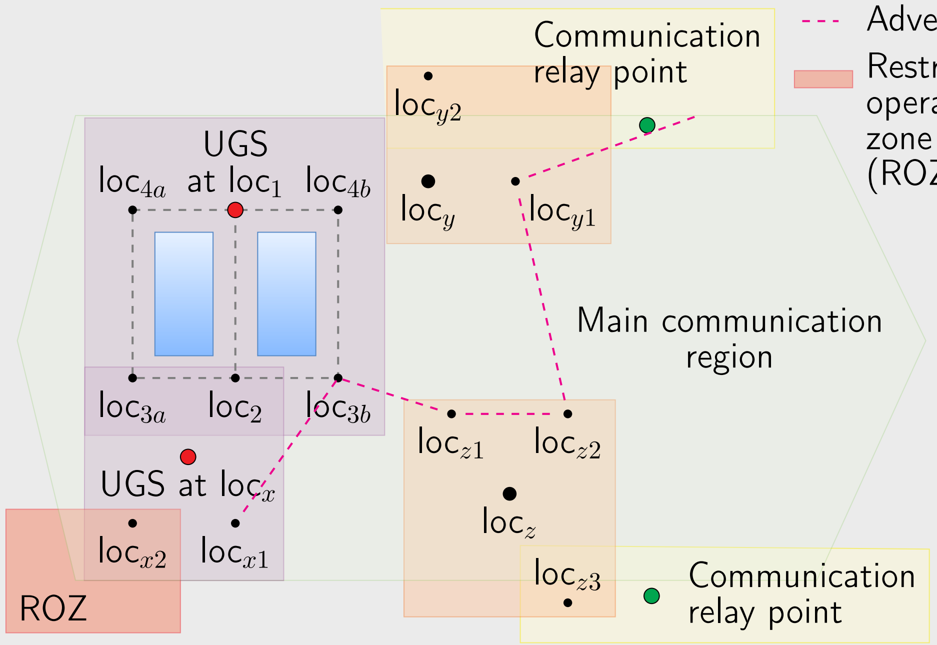
Point Surveillance
P-55 – EDITING – Options

Option A
- T-01 will lose A-32.
- Reason 2
- Reason 3

Option B
- Reason 1
- Reason 2
- Reason 3

Option C
- T-01 will lose A-32.
- T-01 will gain A-33.
- Reason 3

Point Surveillance
P-55 – EDITING
- Location: POI-14
- T-01 will lose A-32.



Example plays

- Random building patrol
- Detect target at loc_x
- Monitor loc_y or loc_z

Structures in Counterexamples in Terms of Plays

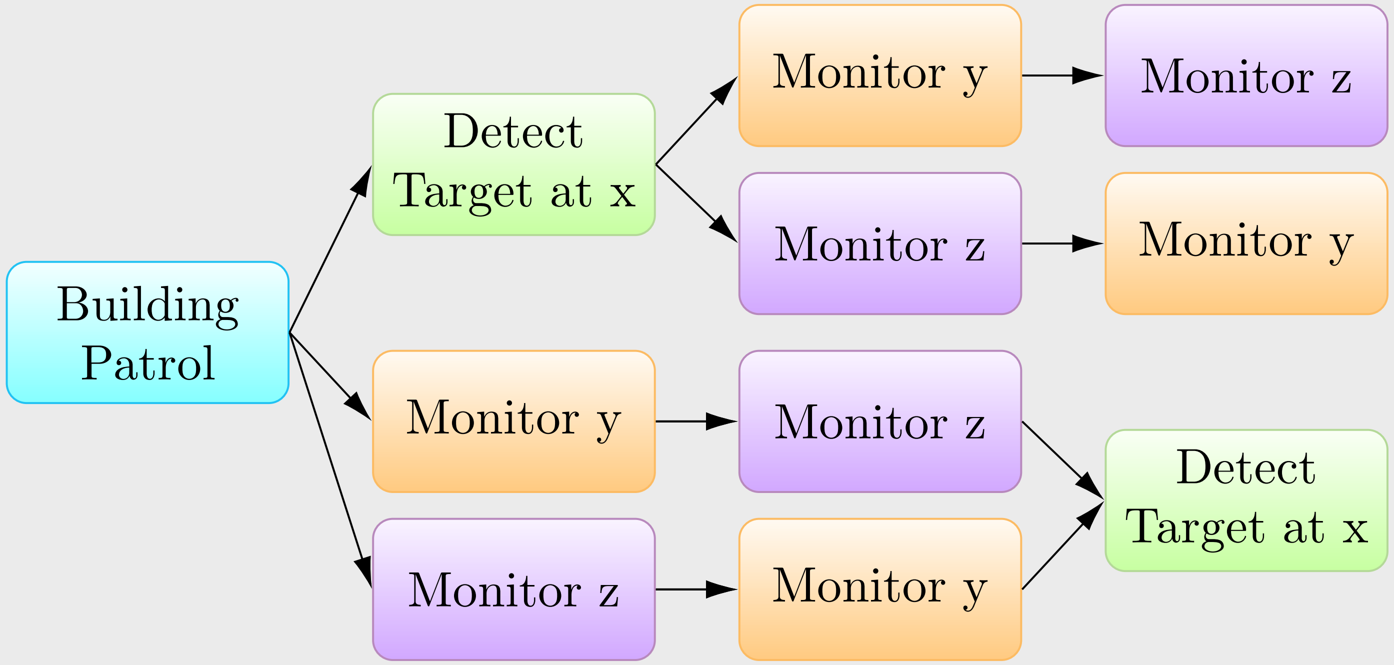
Process algebra statements to create missions out of plays:

“Building Patrol” · (“Detect Target at x” + (“Monitor y” || Monitor z”))

sequential

alternative

interleaving



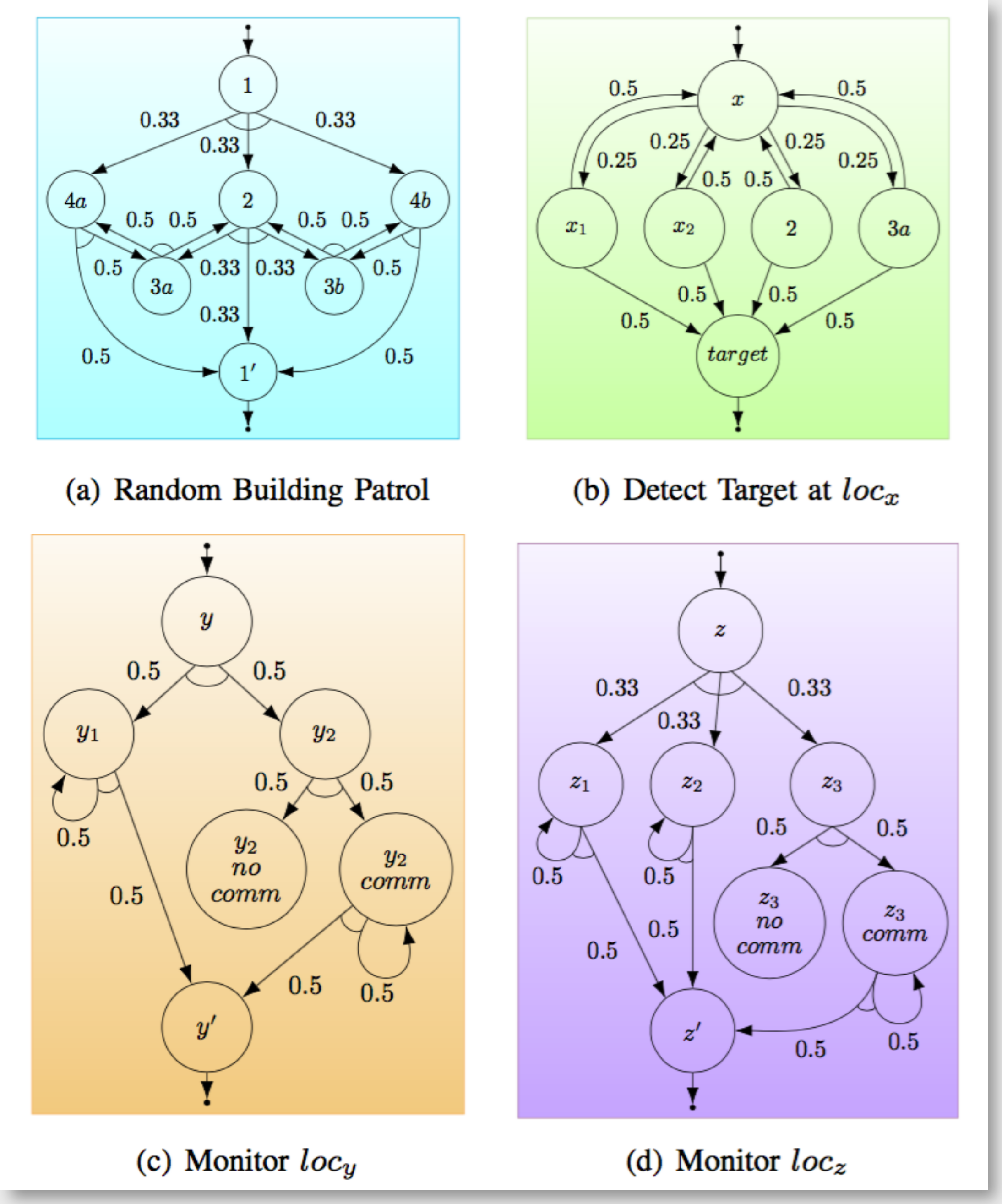
Structure:

- Minimal number of plays
- Temporal or logical relations at the play level

…

(An) Abstraction of Plays

- Model each play as a discrete-time Markov chain with special entrance and exit conditions, where probabilistic distributions are used to represent uncertainties in system behavior
- Compose plays into a Markov decision process (MDP), where the nondeterminism is introduced through the alternative and interleaving operators



(a) Random Building Patrol

(b) Detect Target at loc_x

(c) Monitor loc_y

(d) Monitor loc_z

- Find a subsystem of the MDP that violates the probabilistic specifications and involves a minimal number of plays

Counterexamples with minimal number of plays

as a mixed integer linear program

T: set of target states

X: set of exit states

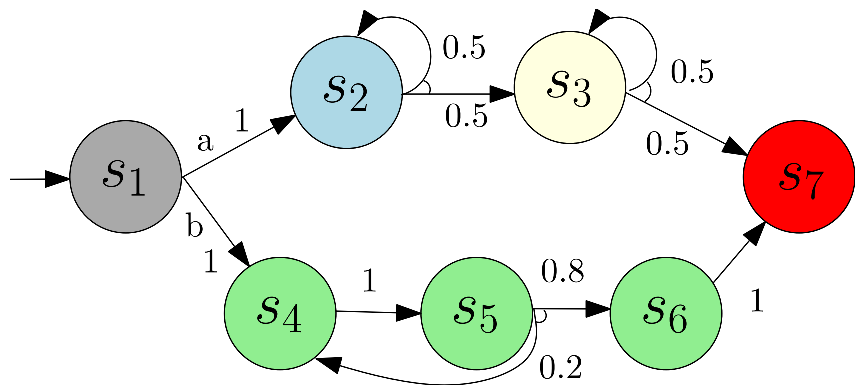
only one action is chosen at exit states

$$
\begin{aligned}
&\text{minimize} \sum_{1\le i\le n} \omega_i && \text{(1a)}\\
&\text{such that} && \\
&p_{\bar{s}} > \lambda && \text{(1b)}\\
&\forall s \in T,\ \text{for } s \in \Omega_i : \ p_s = \omega_i && \text{(1c)}\\
&\forall s \in S \setminus T,\ \text{for } s \in \Omega_i : \ p_s \le \omega_i && \text{(1d)}\\
&\forall s \in S \setminus (T \cup X) : \ p_s \le \sum_{s' \in \mathsf{succ}(s,\tau)} P(s,\tau,s') \cdot p_{s'} && \text{(1e)}\\
&\forall s \in X \setminus T, a \in \alpha : \ p_s \le (1-\theta_{s,a}) + \sum_{s' \in \mathsf{succ}(s,a)} P(s,a,s') \cdot p_{s'} && \text{(1f)}\\
&\forall s \in X,\ \text{for } s \in \Omega_i : \ \sum_{a \in \alpha} \theta_{s,a} = \omega_i && \text{(1g)}
\end{aligned}
$$

Binary variables indicate if a state partition is included in the counterexample

The probabilistic reachability property is violated

Intuition: encoding MDP transition probabilities



Property: $\mathcal{P}_{\le 0.1}(\mathsf{F}\ s_7)$

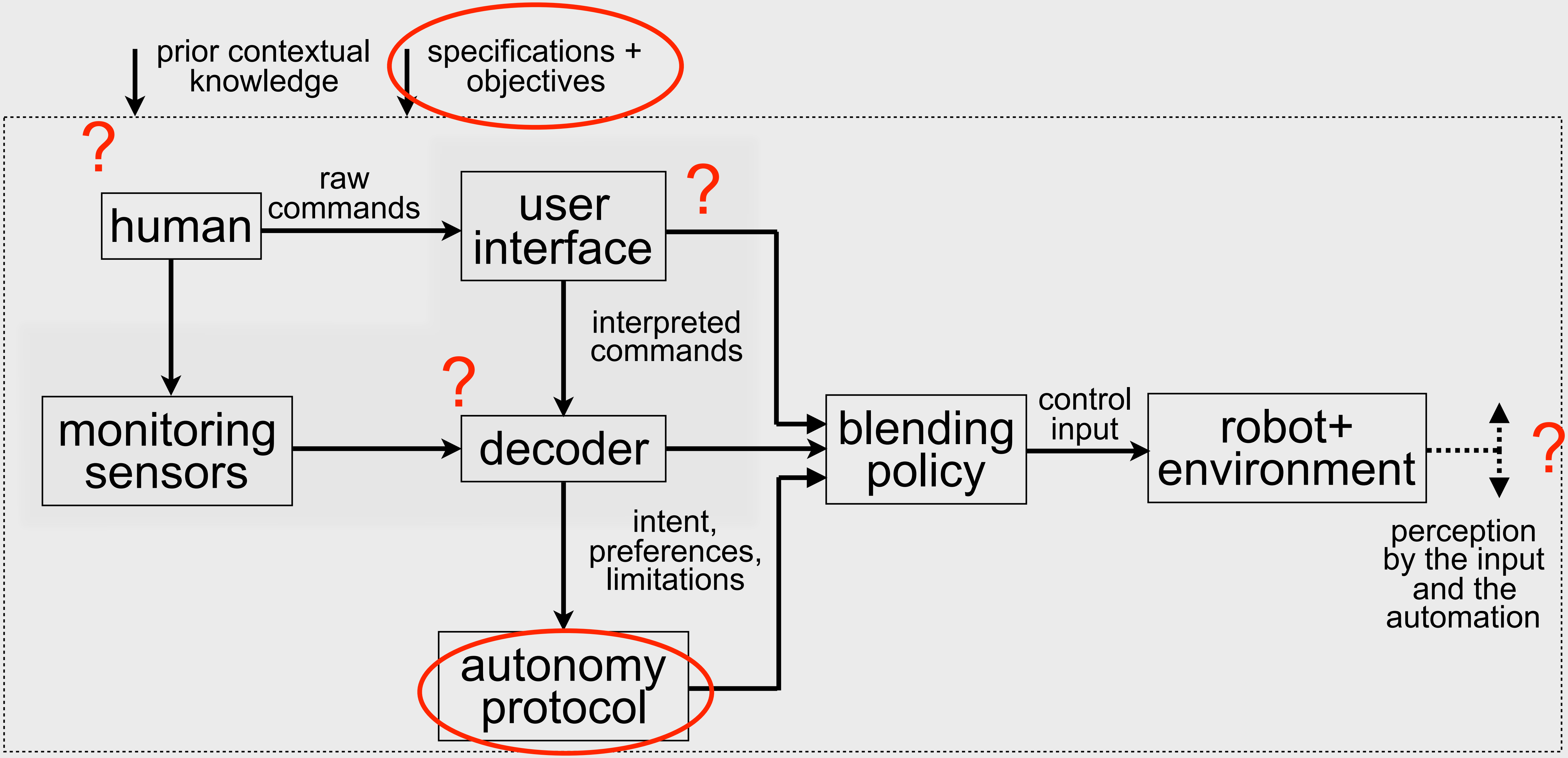
Structured counterexample:

$\text{grey}(s_1) \rightarrow \text{green}(s_4, s_5, s_6) \rightarrow \text{red}(s_7)$

Scales relatively well for the UAV example with more than one vehicle.

Humans and autonomous systems live together.

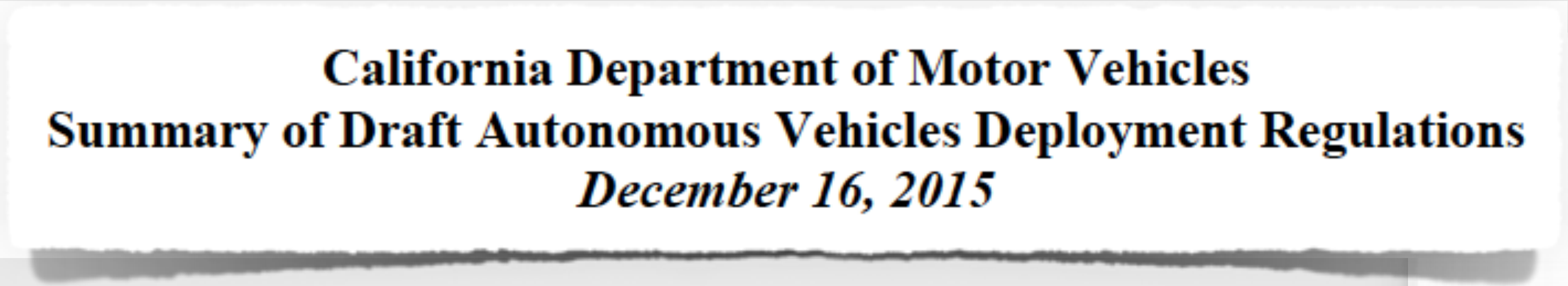
But, they don't in formal verification and synthesis yet.



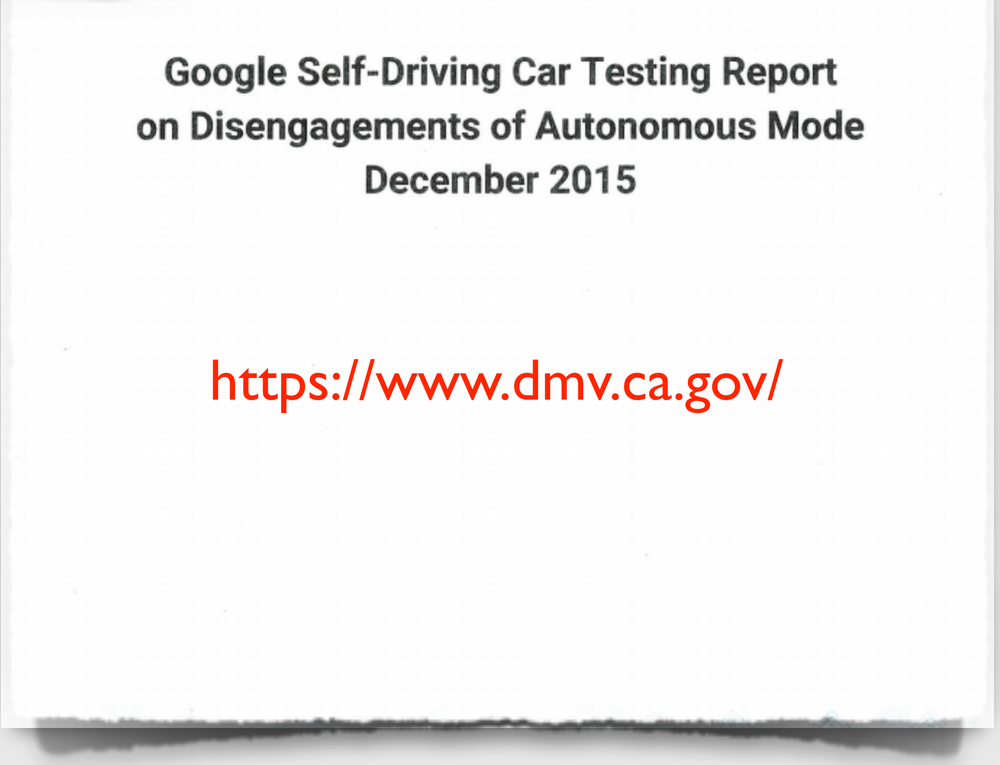
Will formal verification and synthesis ever have an impact on social trust?

I don’t know.

But, they **must (?)** have an impact on certification.



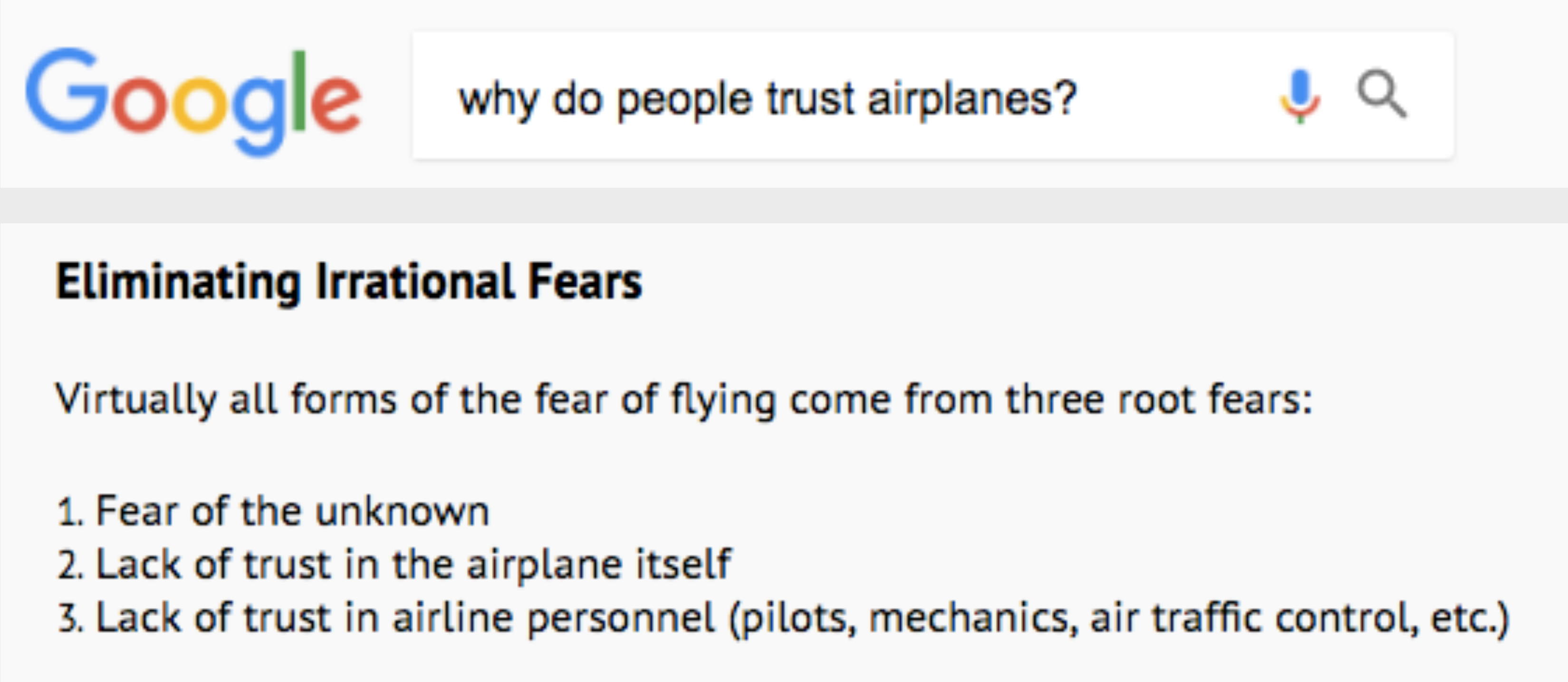
California Department of Motor Vehicles
Summary of Draft Autonomous Vehicles Deployment Regulations
December 16, 2015



Google Self-Driving Car Testing Report
on Disengagements of Autonomous Mode
December 2015

https://www.dmv.ca.gov/

And, certification is a precursor to trust.



why do people trust airplanes?

Eliminating Irrational Fears

Virtually all forms of the fear of flying come from three root fears:

1. Fear of the unknown
2. Lack of trust in the airplane itself
3. Lack of trust in airline personnel (pilots, mechanics, air traffic control, etc.)

In order to learn to trust, we need the autonomous systems on the street.

User study (based on surveys)?

Compare social trust in autonomy in Mountain View, CA and Austin, TX to other neighborhoods with similar demographics