IDA

# From Automation to Autonomous Systems: The Story of Human Trust

**Poornima Madhavan, Ph.D.**
**Institute for Defense Analyses**

IDA

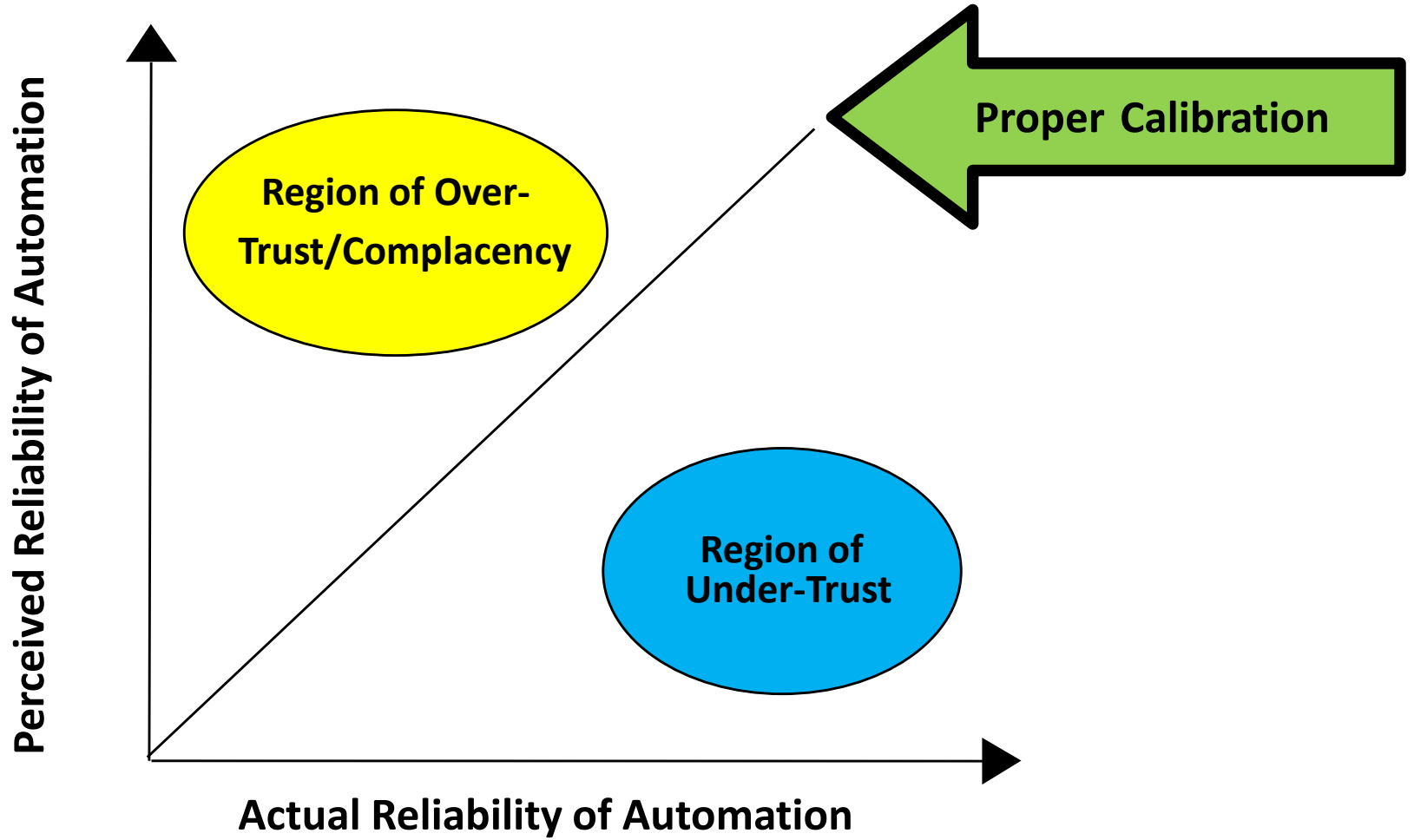
# Trust vs. Assurance

- **Trust** – the belief that an entity will do what it is supposed to do.
  - Based on beliefs or feelings about an entity -> subjective.
  - Can require extrapolation beyond the information that is immediately available to make useful broader judgments.
  - To determine trustworthiness, we focus on metrics that allow us to measure the degree of confidence that we can place in the entity under consideration.

- **Assurance** – the confidence that an entity meets its expectations and requirements based on specific evidence.
  - Built on concrete experiences; situational in nature -> subjective + objective
  - Examples of assurance techniques – formal methods for design, analysis, and testing.

# Model of Trust Calibration

(Wickens, Gempler, & Morphew, 2000)



IDA

# Factors Affecting Automation Trust

Error frequency: automation reliability or the probability of automation generating errors (e.g. Bliss, 1993; Metzger & Parasuraman, 2005).

Error type: false alarms create under-trust -> people ignore alarms or automation alerts; misses have different effects depending on context (Bliss, 1993; Parasuraman, et al., 1997).

Error salience: conspicuous or highly visible errors capture attention and trigger a decline in trust (e.g., Lee & Moray, 1992, 1994; Dzindolet et al., 2001; Wiegmann et al., 2001).

Error easiness: automation errors on tasks that can be easily performed by humans (without assistance from automation) can be especially damaging of trust (Madhavan, Wiegmann & Lacson, 2007).

IDA

# Salience and Easiness of Automation Errors

- Information that is **inconsistent with expectations** is likely to be well remembered -> unduly large role in information processing (Ashcraft, 1994; Ruble & Stangor, 1986; Smith & Graesser, 1981).
- **Perfect automation schema** -> expectation that automated systems will perform at nearly perfect rates (Dzindolet et al., 2001, 2002).
- **Salient** or **obvious** errors by an automated system capture attention -> trigger a rapid decline in trust and accompanying confidence (e.g., Lee & Moray, 1992, 1994; Dzindolet, et al., 2001; Wiegmann, Rich & Zhang, 2001).
- When human operators are **not allowed to view errors** made by an automated system, they might be more likely to trust the automation than when such errors are obvious to operators (Dzindolet et al., 2003).
- Other research has contradicted this; trust is higher when humans have **access to all "raw" information**, even if the information is negative (Madhavan & Phillips, 2008).

# Type of Automation Error: Misses vs. False Alarms

- When the base rate of a real world event is low (e.g., a weapon in a suitcase at an airport security checkpoint), the potential for false alarms by an automated aid is high, even for very sensitive systems.

- **False alarms** (i.e., repeatedly "cry wolf ") -> under-trust in automation -> operators not respond to alarms or other automation alerts (Parasuraman, Hancock, & Olofinboba, 1997).

- BUT, **misses** can be more damaging in the long run.

- **Example: Airport baggage screening**. False alarms (false positives) can lead to time wastage, inconveniences and missed/delayed flights. But misses (false negatives) can lead to a weapon getting aboard an aircraft leading to disastrous consequences.
  - Dilemma: In this context, misses are costlier and consequences are delayed; false alarms are less costly but consequences are immediate

IDA

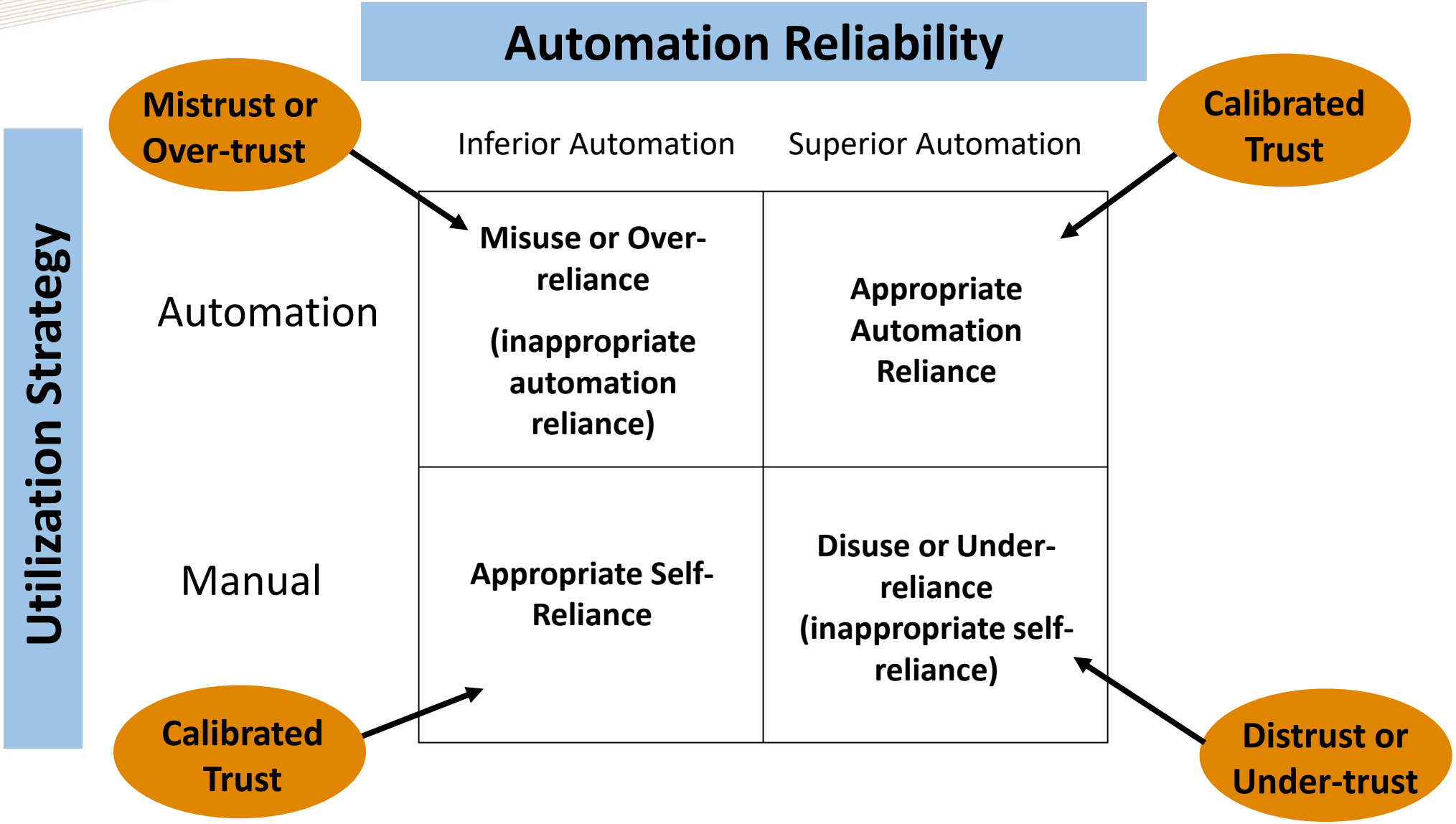
IDA

# A Closer Look at Trust: Humans vs. Automation (Trust – Self Confidence? Lee & Moray, 1992)

- Humans tend to rely on all available information while completing complex tasks.
- What happens when CONFLICTING information is presented by human and automated sources?
- **Example:** A Russian passenger jet and cargo plane in 2002 crashed in a mid-air collision. Automation told the two planes to change elevation in different directions, but so did ATC; these two messages directly conflicted in terms of the directions delivered to the pilots. One pilot listened to the automation and the other pilot listened to ATC. The planes collided.
- Human-automation has been studied as a special case of human-human trust.
- Trust in automation is determined (to an extent) by the degree of self-confidence in own ability to perform the task unassisted.

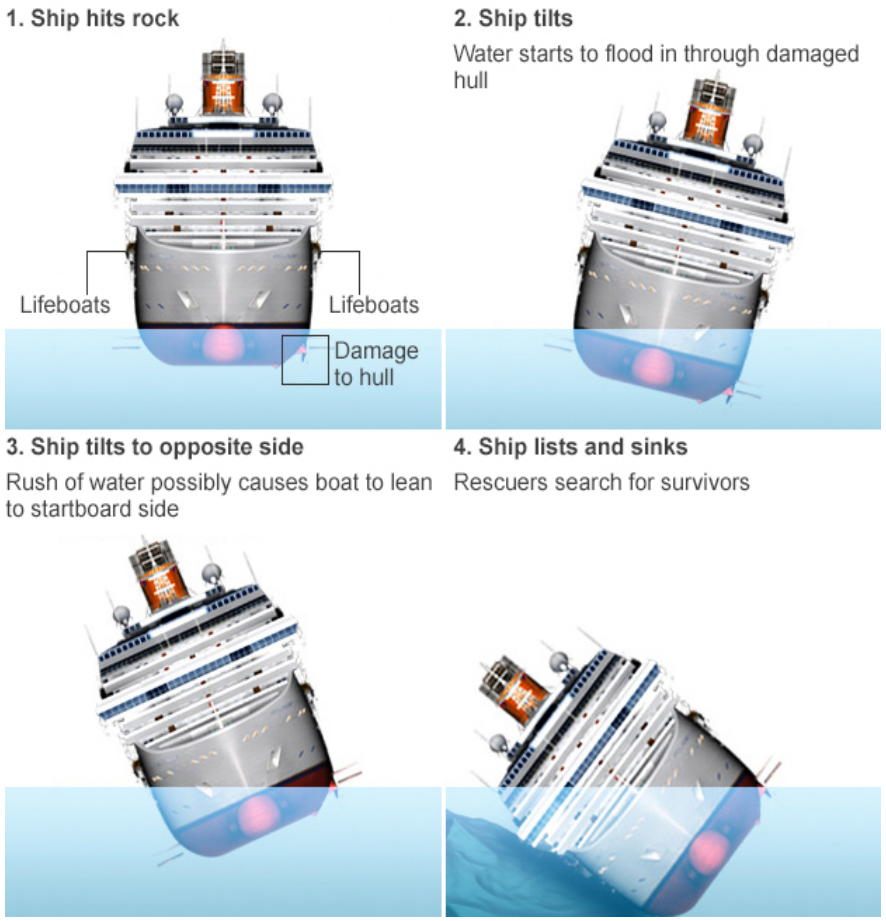
# Trust and Automation Utilization

Adapted from Parasuraman, Raja, and Victor Riley. "Humans and automation: Use, misuse, disuse, abuse." *Human factors* 39, no. 2 (1997).

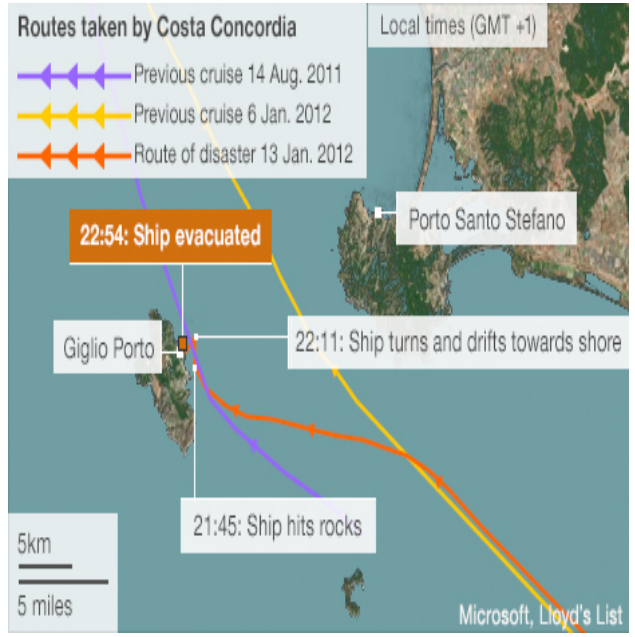


# Example of Under-Trust, Automation and Human Error, and Breakdown of Situation Awareness

- The Italian cruise ship ***Costa Concordia*** capsized and sank after striking an underwater rock off Isola del Giglio, Tuscany, on the night of 13 January 2012, resulting in 32 deaths.




- **21:30:** Captain issued override of ship's automated navigator – unapproved, unauthorized
- **21:45:** Concordia hit rocky outcrop; crew struggled to assess situation and failed to convey accurate information to authorities
- **21:52:** Chief engineer tried and failed to start emergency generator
- **22:33:** General emergency alarm activated – passengers told to await instructions
- **23:32:** Both captains abandoned ship leaving 300 passengers and crew on board




IDA

IDA

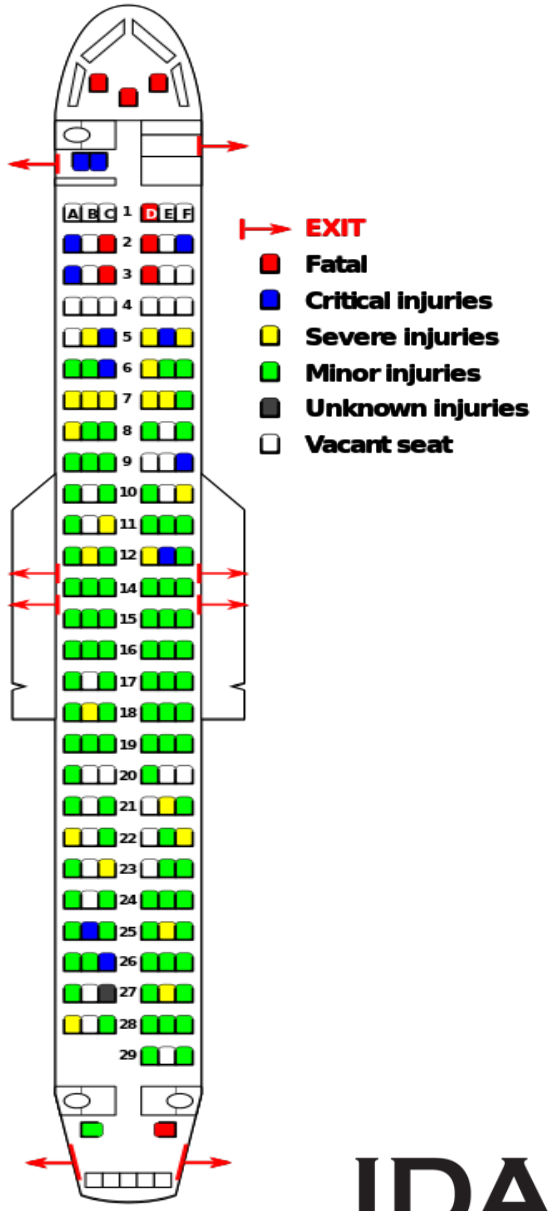
# Analysis of the ***Costa Concordia*** disaster

- **21:30:** Captain issued override of ship's automated navigator – unapproved, unauthorized → Under-trust of and under-reliance on automation
- **21:45:** Concordia hit rocky outcrop; crew struggled to assess situation and relayed incomplete information to authorities → Breakdown of situation awareness
- **21:52:** Chief engineer tried and failed to start emergency generator → Automation failure
- **22:33:** General emergency alarm activated on board – passengers told to await instructions → Faulty communication; Poor shared situation awareness
- **23:19:** First captain abandoned ship leaving second captain to coordinate emergency rescue operations → Human error; Poor shared situation awareness
- **23:32:** Second captain also abandoned ship leaving 300 passengers and crew on board → Human error
- **00:42:** Coastguard commanded captain(s) to go back on board; captain(s) did not comply → Human error; Poor shared situation awareness

# Example of Over-Trust, Automation and Human Error, and Breakdown of Situation Awareness

- Passenger aircraft Turkish Airlines Flight 1951 crashed during landing at the Amsterdam Schiphol Airport, Netherlands, on 25 February 2009, resulting in the death of 9 passengers and crew, including all 3 pilots.

- Final approach for landing – altitude ~2000 ft. above ground
- Left-hand (captain's) altimeter reading changed from 1950 ft. to -8 ft. altitude; right-hand (co-pilot's) altimeter functioned correctly
- Pilots received incorrect auditory warning – TOO LOW! GEAR! – indicating landing gear should be down
- Pilot pulled throttle back to slow aircraft, but autothrottle automatically reverted to "retard" mode
- Pilot failed to disengage autothrottle, which is the recommended course of action during altimeter failure




IDA

IDA

# Analysis of the ***Turkish Airlines*** disaster

- **Final approach for landing – altitude ~2000 ft. above ground**
- Left-hand (captain's) altimeter reading changed from 1950 ft. to -8 ft. altitude; right-hand (co-pilot's) altimeter functioned correctly → First automation failure; Poor shared situation awareness between pilot and co-pilot
- Pilots received incorrect auditory warning – TOO LOW! GEAR! – indicating landing gear should be down → Second automation failure; Over-trust
- Pilot pulled throttle back to slow aircraft, but autothrottle automatically reverted to "retard" mode; Pilots unaware of autothrottle action → Poor situation awareness; Over-trust
- Pilot failed to disengage autothrottle, which is the recommended course of action during altimeter failure → Poor situation awareness; Human error
- Aircraft landed short of runway, sliding through wet clay of a plowed field
- Pilot survived crash, but died due to inability of rescue operations to reach him quickly through locked cockpit door → Third automation failure

IDA

# Summary
# Trust, Assurance (or Confidence), Reliance

- Important to distinguish between trust and assurance (or confidence). Trust -> more holistic and dispositional judgment; confidence -> based on situational information and is pertinent to an entity's actual performance on a task.
- Although trust and confidence are distinctly different issues, the line between them is often blurred. Utilization decisions -> based on a combination of trust and confidence.
- Automation errors that typically influence trust (and confidence): frequency, type (miss vs. false alarm), salience and easiness.
- Types of trust: Distrust (or under-trust); Mistrust (or over-trust) leading to complacency; Calibrated trust (trust is proportionate to the actual reliability of automation).