

Of Tunnels and Wolves:
Over trust and Under trust in
automation" and its relation to
attentional
tunneling, alarms and alerts.
Chris Wickens

Professor Emeritus University of
Illinois, Adjunct Professor,
University of Colorado.

Automation Reliability. Unreliable (imperfect) automation may be:

- Truly faulty (software reliability)
- Asked to perform in unanticipated situations (stretching capabilities beyond the limit)
- Incorrectly setup (“dumb and dutiful”) KAL007
- Asked to make inferences with imperfect data (forecasting, prediction, medical diagnosis)..Therefore:
- Automation **WILL MAKE ERRORS.**

Attention-Centric

Switching Selective (SEEV) Resource allocation (divided)

In perception In performance

Automation-Centric

Imperfect Automation

Overtrust

Undertrust

Mental Models

Complexity

*

*

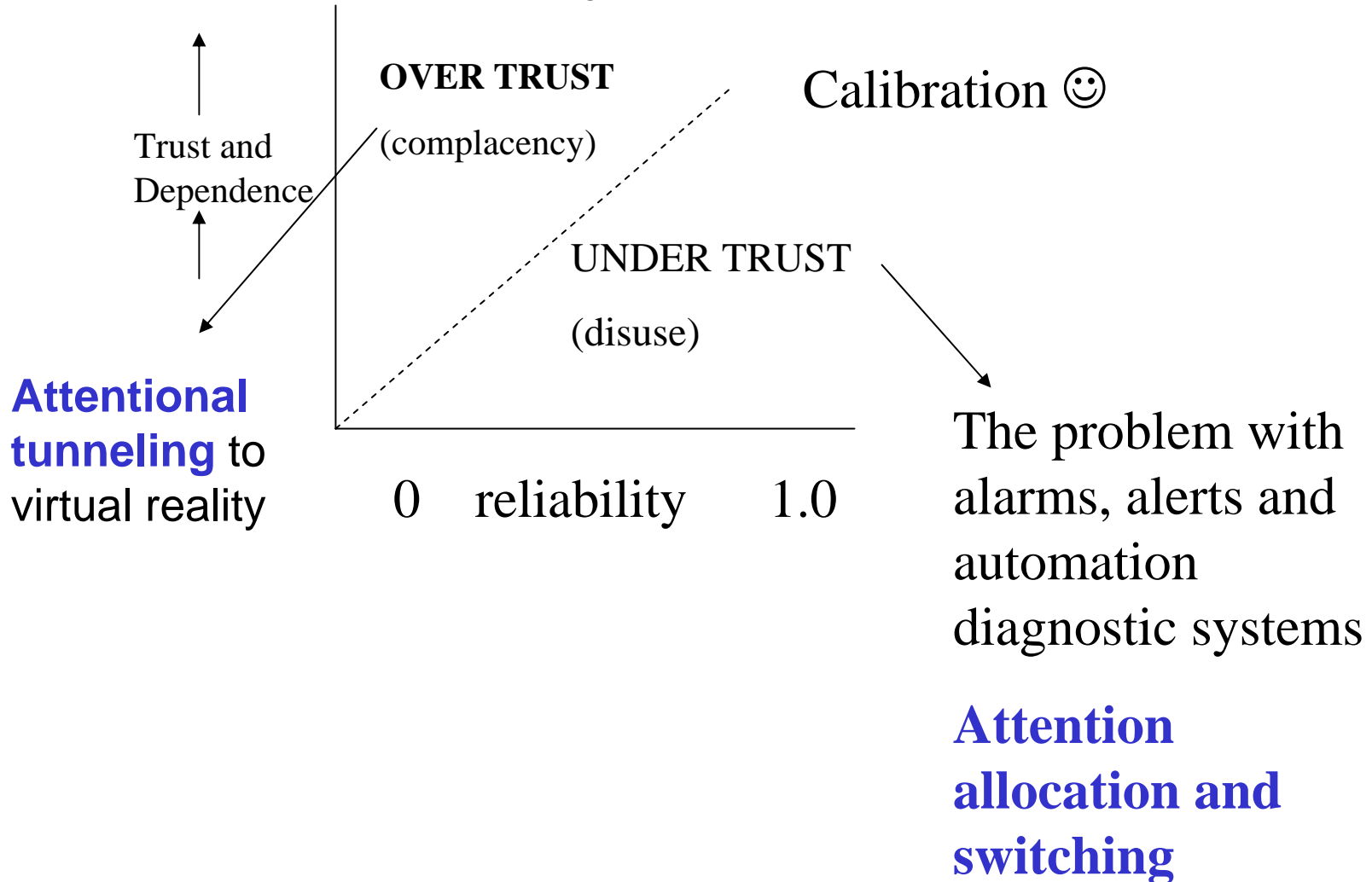
	Switching	Selective (SEEV)	Resource allocation (divided)	
			In perception	In performance
Imperfect Automation				
Overtrust		*	*	
Undertrust	*		*	*
Mental Models				
Complexity				
*				
*				

Reliability and Trust calibration

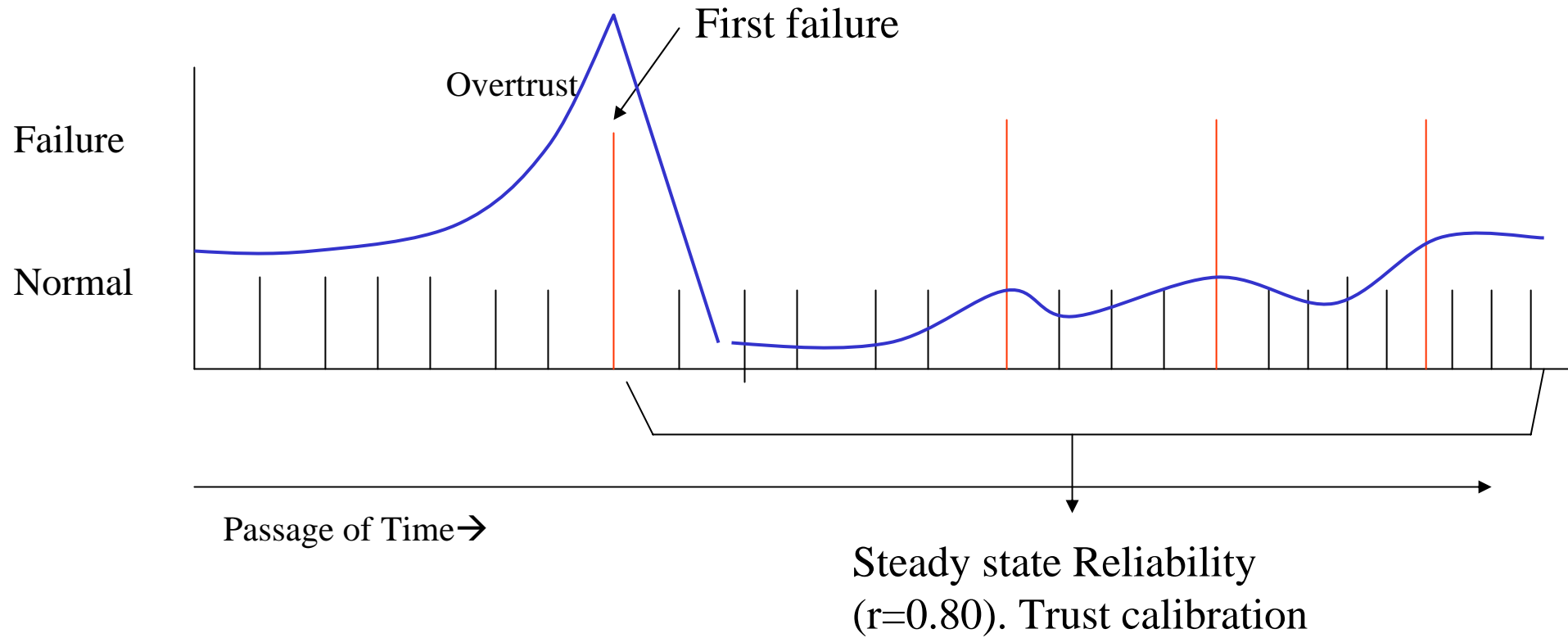
Automation reliability \rightarrow trust \rightarrow automation **dependence**

cognitive belief

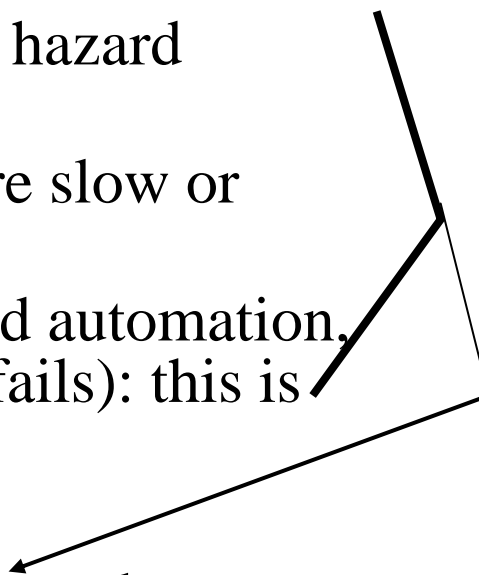
behavioral measure



The two Phases of Trust



The First failure: Overtrust and complacency: The China Airlines incident

- With highly reliable automation, on the very rare occasions when it **DOES** first fail, the human:
 - 1. Won't be **monitoring** the automation (or the hazard domain). (fail to notice)
 - 2. Won't be **aware** of the system state (therefore slow or incorrect in intervention)
 - 3. May be **less skilled** in “taking over” for failed automation, because “out of practice” (the exam calculator fails): this is “deskilling”
 - The “out of the loop unfamiliarity” (OOTLUF) syndrome.
 - Ironies of automation: (1) the more reliable is automation, the more severe is OOTLUF. (2) Automation is more likely to “fail” on the most complex problems, when the need for human situation awareness and manual skill is greatest.
- 

OVERTRUST & COMPLACENCY

Application to Automation-induced **attentional tunneling**
(Wickens, 2005)



Compelling

Important

AUTOMATION

“virtual reality” of the
3D SVS Display

OTHER SOURCES

Immersion

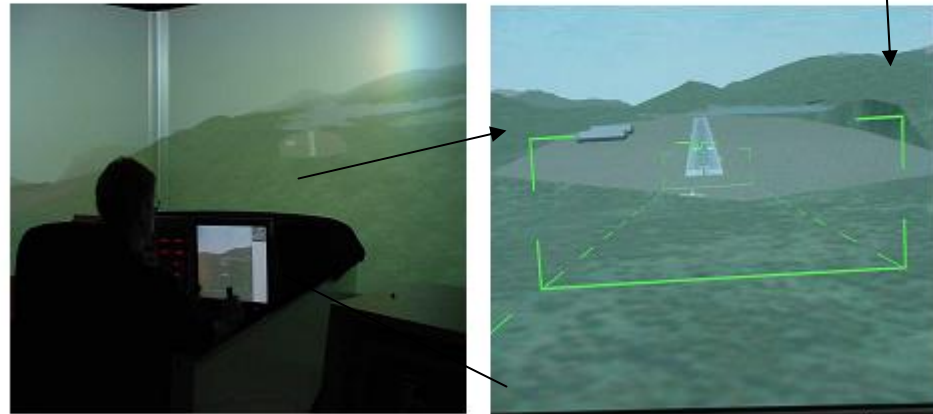
Fault management

ATTENTIONAL TUNNELING

The SVS display and the highway in the sky (HITS): (Stage 3 automation: It advises the pilot of where to fly)

Synthetic Vision Systems (SVS)

- Provide pilot with synthetic view of 3D terrain beyond the aircraft



2

A good human factors design: prevents **controlled-flight into terrain** (CFIT) accidents. HITS relieves the demands of information integration for trajectory choice **BUT.....**

Is it too **compelling** ? Does it lead the pilot “down the garden path” of complacency: What is on the display **is** reality. This is the problem of **attentional tunneling** and **change blindness**. We don't notice changes that are outside of the focus of attention.

Off-Normal Outside-World Traffic



Suppose automation **fails** to contain all pertinent information in its data base⁷. (The **virtual reality** of the display does not quite reflect the **true reality** of the outside world): An example of **imperfect automation**.

Pilots flying with the SVS HITS, don't notice the “rogue blimp” or tower, which is only visible in the outside world. This is automation-induced attentional tunneling to the compelling display. Overtrust in its information.

The **first failure effect** and **Attentional Tunneling**

Miss Rate: Off-Normal hazard Events outside; Flying with SVS

<u>STUDY</u>	<u>EVENT</u>	<u>HITS</u> (Tunnel)	<u>No HITS</u>	
<u>Attention Tunneling Evidence</u>				} Off- Normal is outside
SVS 3	Blimp	4/8	1/6	
SVS 3	Runway offset	5/12		
SVS 2	Missed approach blimp	14/17		
SVS 5	Tower	10/24		
SVS 6	Blimp	3/24		
SVS 8	Tower	6/24		

Average: 38% of events are missed !.

(17% missed when HITS is absent). **Attentional**

Tunneling is Real.

Linkage of attentional tunneling to:

- Change blindness
- SEEV Model of selective attention in visual scanning: The eyeball is driven by:
 - Saliency
 - Effort conservation
 - Expectancy (bandwidth of information source)
 - Value (of information).

Wake vortex modeling application

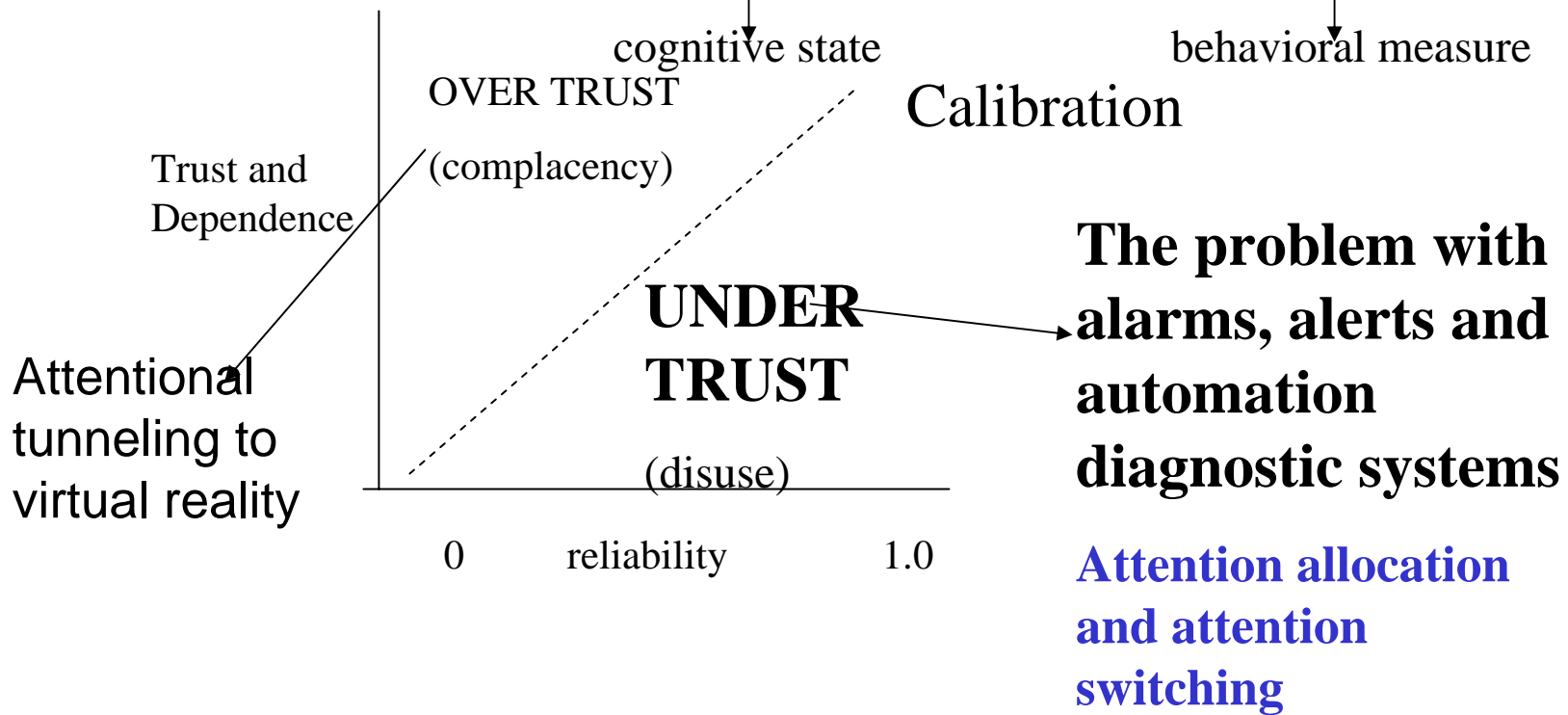
Solution to automation-induced attentional tunneling & complacency

- TRAINING in need to monitor the raw data more. (Spend more time looking out. Pilots with SVS only do this 6-10% of the time)
- DESIGN to present the raw data close to the automation guidance. (e.g., SVS and HITS display on a Head up Display,

3. Under trust

Reliability and Trust calibration

Automation reliability \rightarrow trust \rightarrow automation dependence

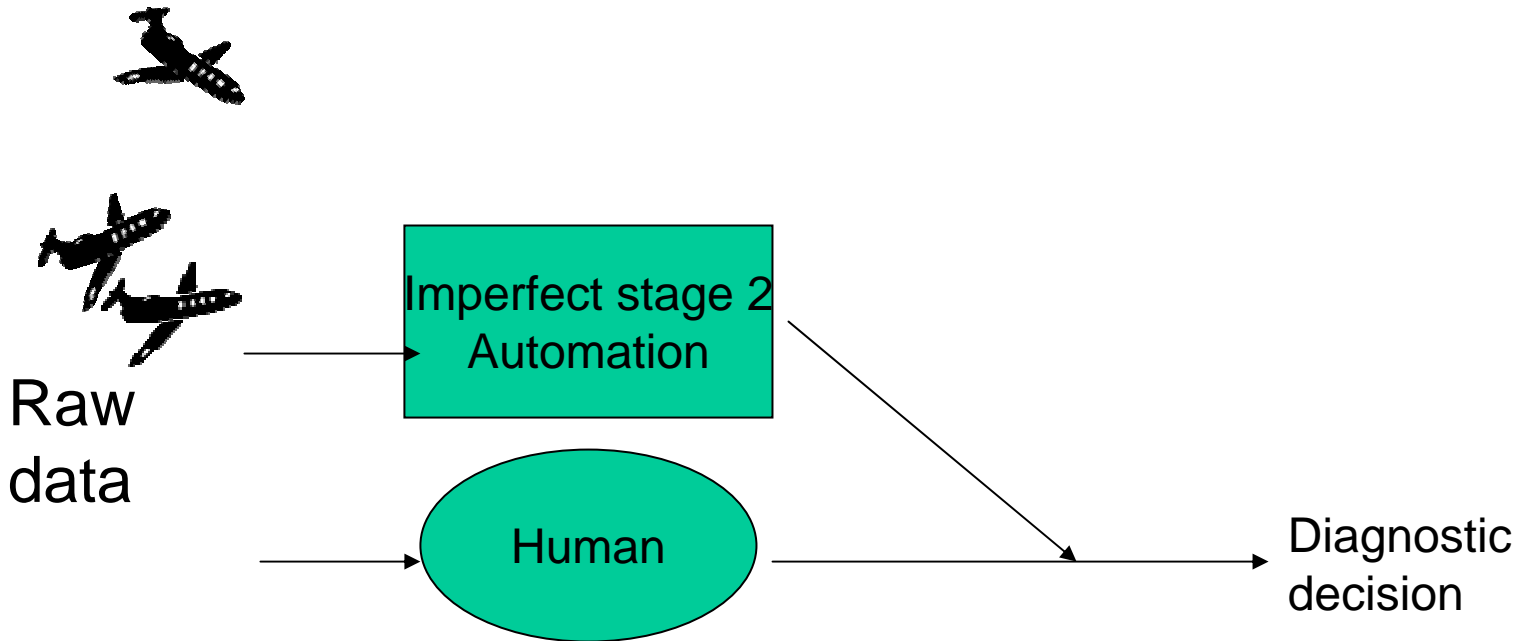


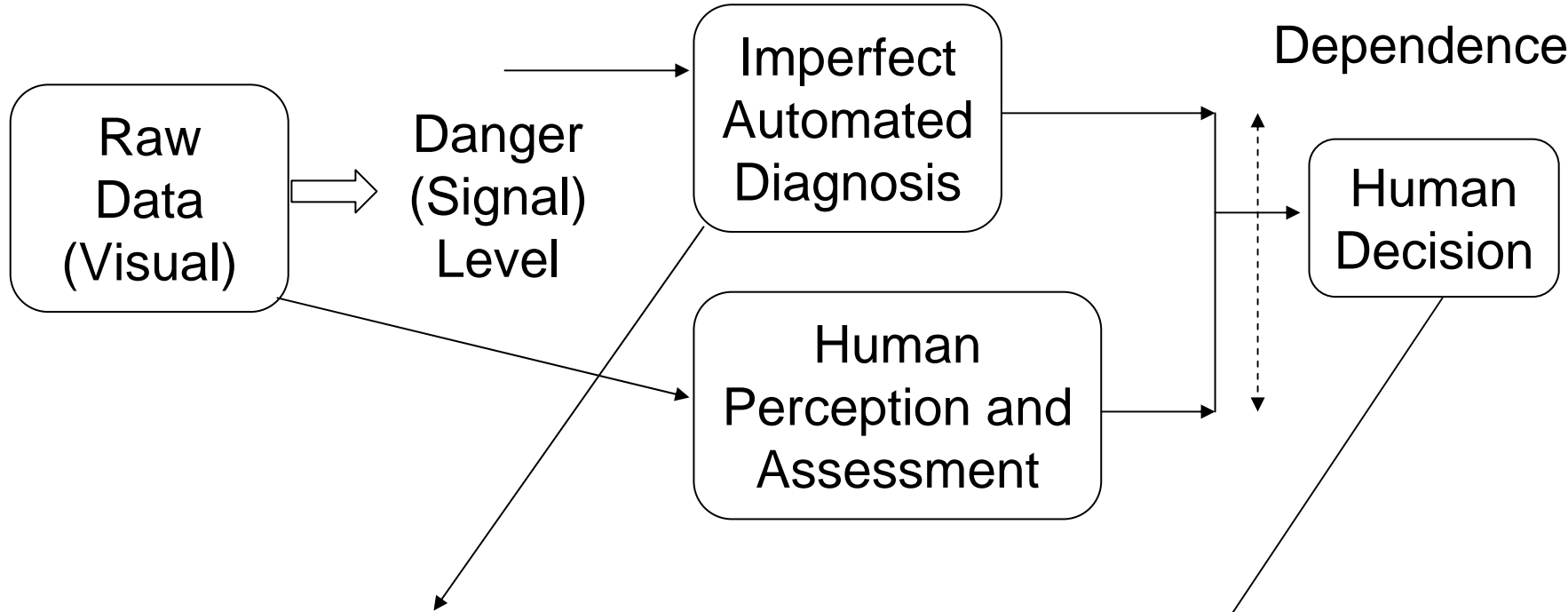
2003 First Energy Company: Utility operators “missed” a developing crisis because an alarm did not sound (it had previously been disabled because of **false alerts**): They were then “complacent”. The resulting crisis caused blackout in much of north-central North America.

2001: Guam air traffic controllers disabled minimum safe altitude warning system (MSAW) because it had issued too many **false alerts**. It “**cried wolf**” too often. A result: they missed detecting a low altitude descent. KAL flight: Controlled-flight-into-terrain. Crash → over 100 fatalities.

2006: NTSB Safety recommendation to FAA: Air traffic Controllers missed automated collision and terrain alerts in 11 aircraft accidents. **Cry Wolf?**

Parallel monitoring of automation and human diagnostic systems: imperfect diagnostic automation





State of the World

		Signal (Danger)	
		Y	No
Diagnosis	Y	"Hit"	False Alert ☹️
	N	☹️ Miss	Correct Rejection

Automation Errors (Unreliability)

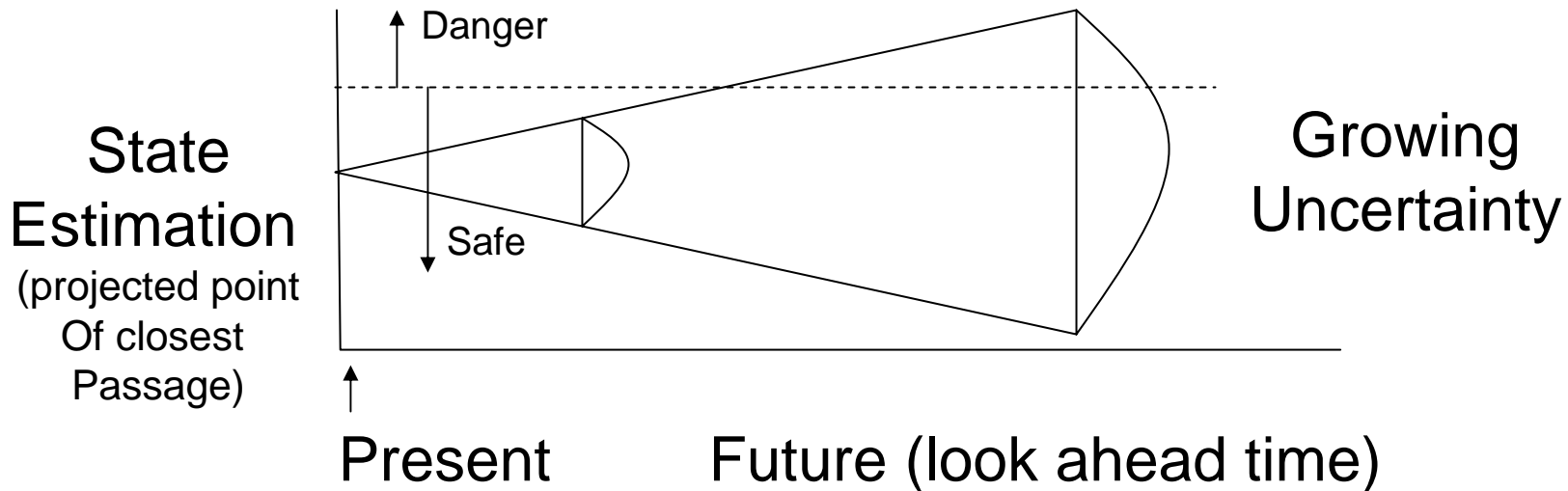
State of the World

		Signal (Danger)	
		Y	No
Diagnosis	Y		FA ☹️
	N	☹️ Miss	

Errors

Why is diagnostic automation unreliable (imperfect)?

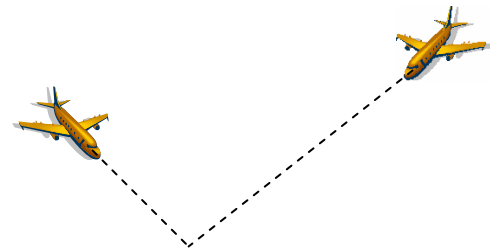
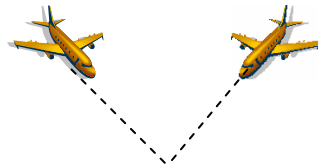
- Poor sensors
- Imperfect algorithms
- Probabilistic world, particularly with predictive diagnosis as in any **conflict detection system**. (TCAS (mid air collision), GPWS (ground collision))



Predict Collision With:

Certainty 😊

Uncertainty ☹️



Short

Medium

Long

—————> Look Ahead Time <—————

Rationale for longer look-ahead-time: Allow plenty of time for **Planning** & conflict avoidance maneuvering.

Growing Uncertainty with longer look-ahead time →
Diagnostic (alert) automation will have lower diagnostic
reliability and therefore:

EITHER more false alerts

OR More misses,

In conflict detection,

a “miss” is essentially a “late alert”).

	Signal	
	(Danger)	No
Y		FA ☹
N	☹ M	

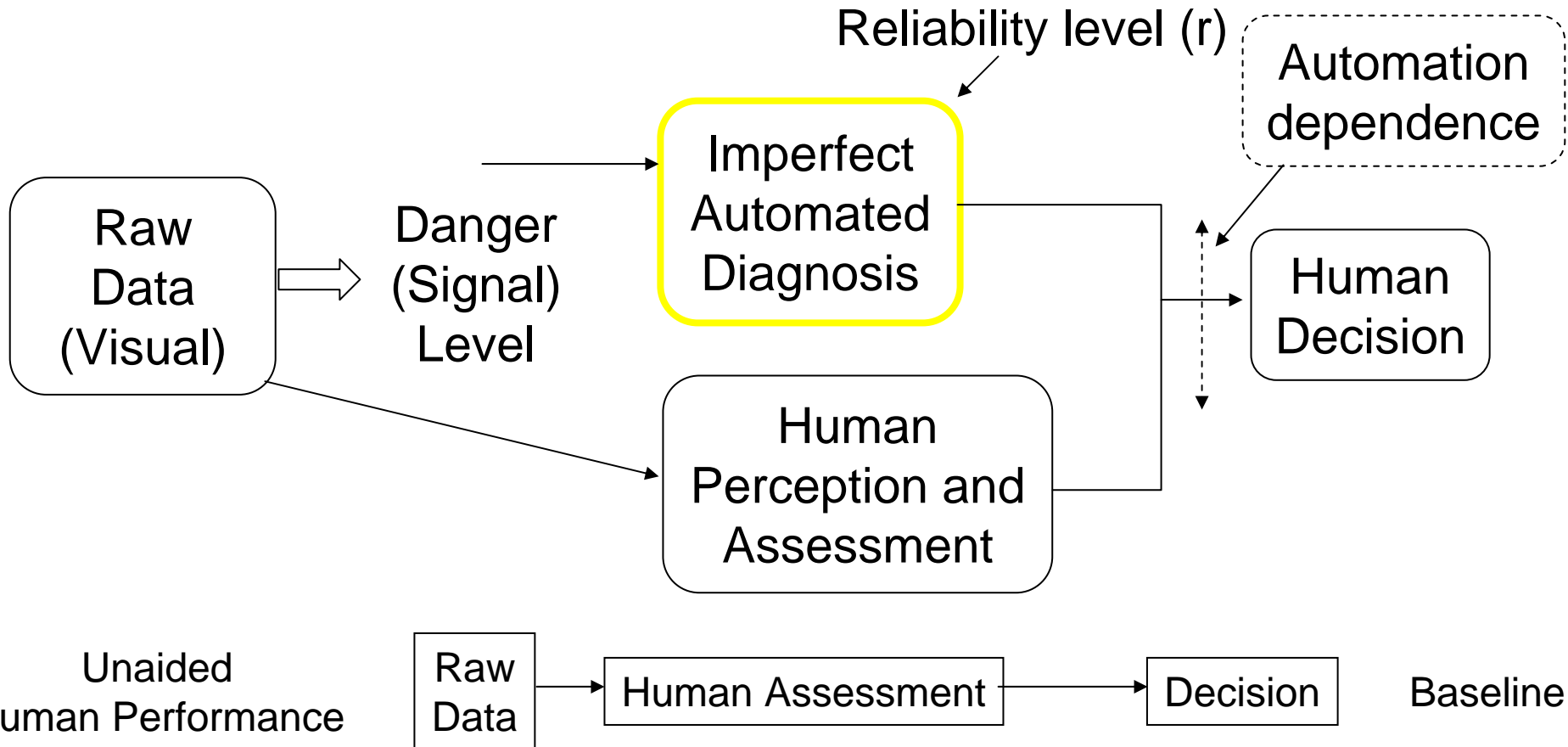
The Federal Aviation Administration has asked us to:

- Assess how low diagnostic (prediction) automation reliability can go (and still be useful). Certifying alerting systems with the cockpit traffic display. Longer LAT than with TCAS.
- Address the “false alarm problem” in air-traffic-control alerting systems

Two Research Issues

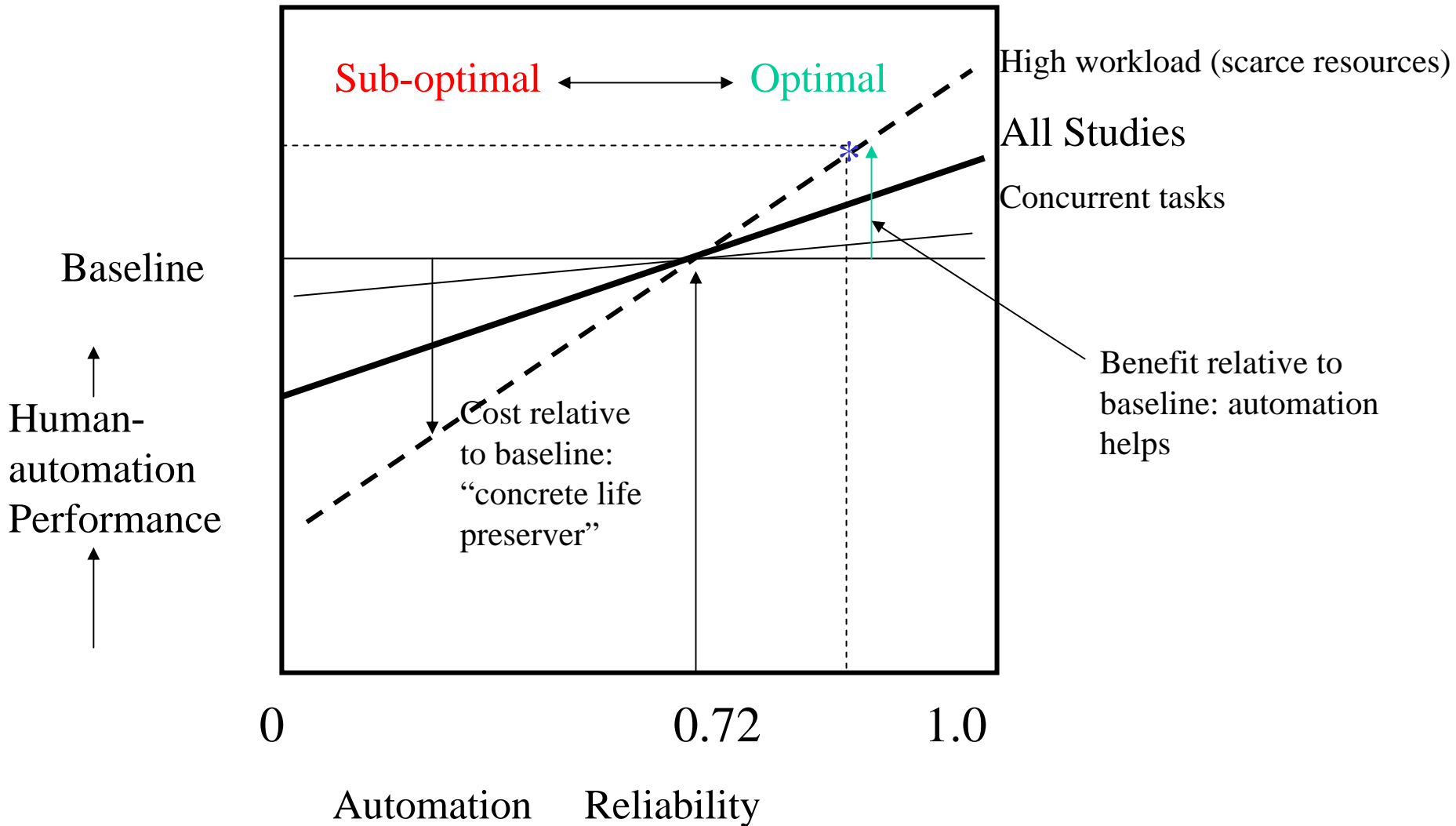
- 1. How low can reliability be, and still be useful to the human-automation system?
- 2. When reliability drops (because of longer look-ahead time), what should be the tradeoff between allowing automation misses and automation false alerts?
- How are these manifest in attention?

A. Given that automated diagnostic systems have necessary imperfections, how low can automation reliability be and still be better than unaided human performance?



The search for constants in human performance modeling.
(Like working memory limit: 7 ± 2)

Wickens & Dixon (2007): Regression meta-analysis of different imperfect diagnostic automation studies



Therefore

- Under high workload (difficult task or multi-task environment) diagnostic automation with $r > .080$ ($0.72 + .08$) will help relative to unaided performance.
- Below $r = 0.72 \rightarrow$ the “concrete life preserver”. Its better for the diagnostic task performance, to ignore automation entirely.
- But people don't: A non-optimal **attention allocation strategy**

Two Research Issues

- 1. How low can reliability be, and still be useful to the human-automation system?
- **2. When reliability drops (because of longer look-ahead time), what should be the tradeoff between allowing automation misses and automation false alerts?**
- **How are these manifest in attention?**

Different Systems and Research Paradigms.

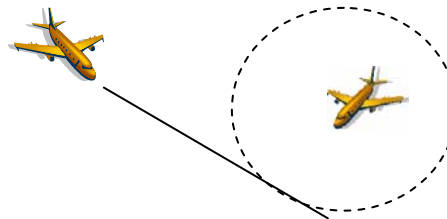
1. The cockpit display of traffic information (CDTI) alerting system

. TCAS: Short range prediction 30-40 sec

CDTI: Broader Traffic Picture. Longer range prediction (2-5 min)

How long can/should
look-ahead time be?

How should threshold
be set?



Oral Alert System

Research generalizes to
air traffic control conflict
alerts

2.Unmanned Air Vehicles (UAV). The Army's Hunter/Shadow



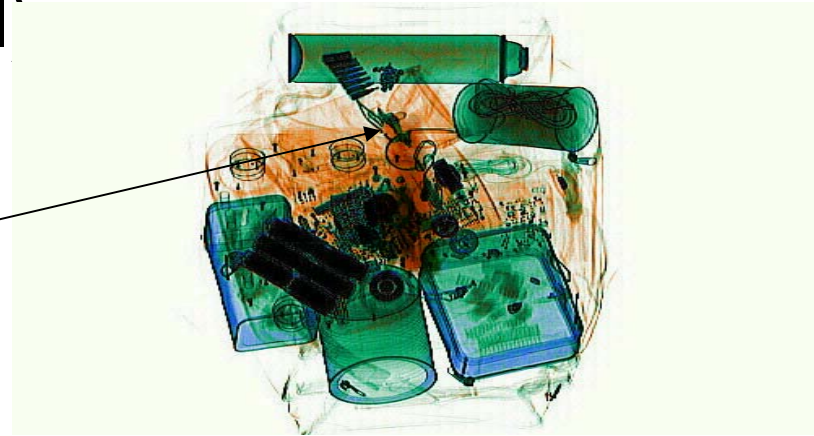
2. Unmanned Air Vehicle surveillance. Automatic ground target recognition



Automated Systems
monitor

3. Luggage Screening: an automated decision aid (Wiegmann et al`

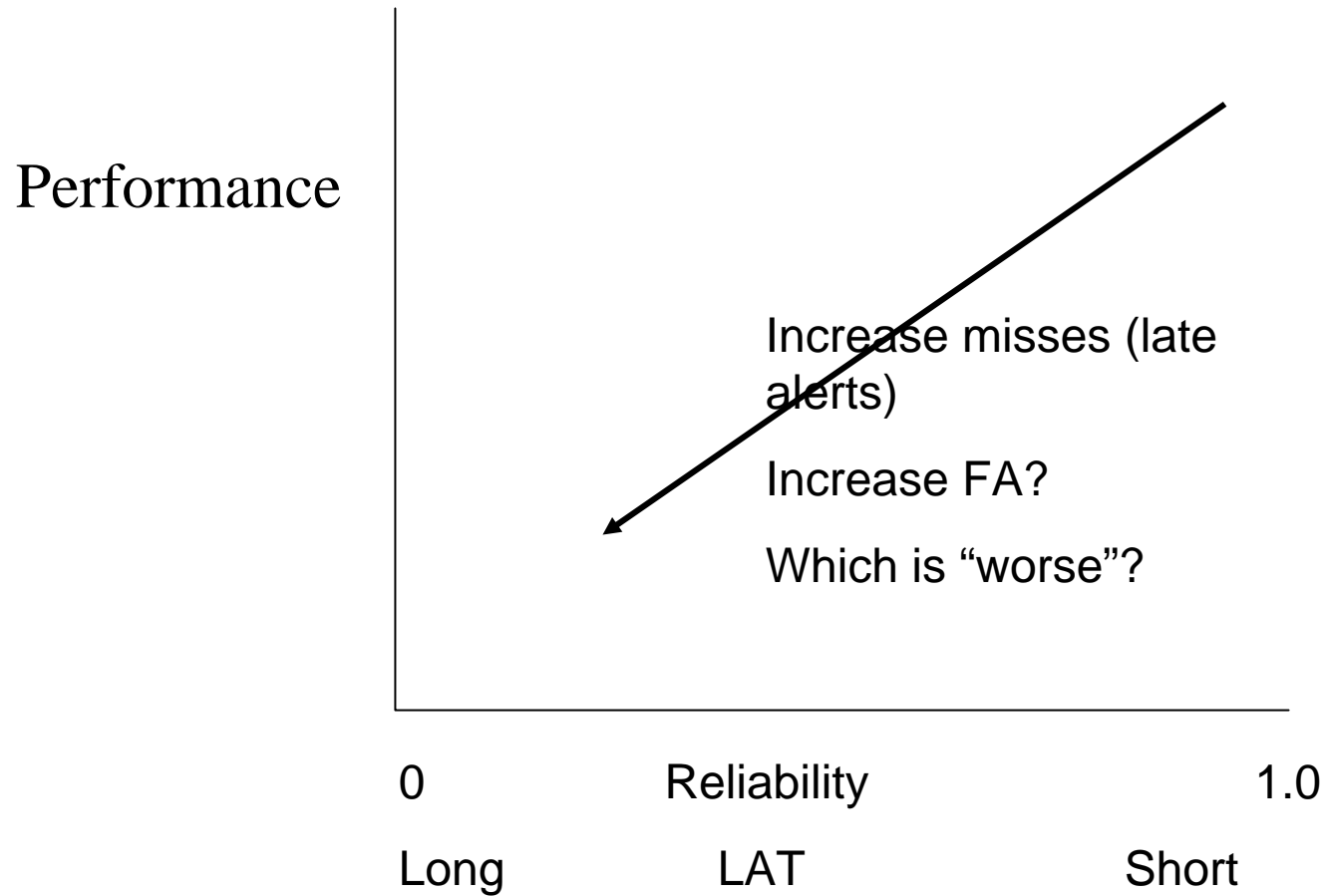
Knife?



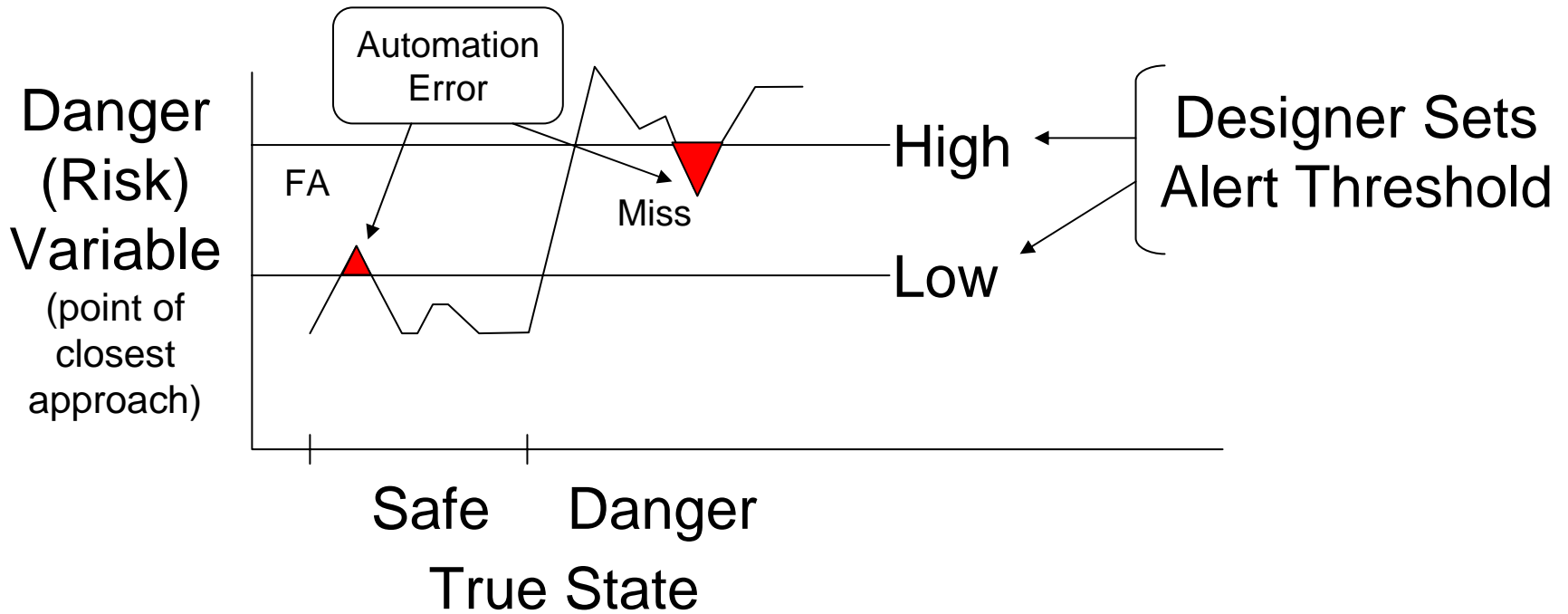
In all three systems, imperfect diagnostic automation reliability above 0.80 improves performance relative to non-automated baseline, PARTICULARLY as workload increases.

Threshold setting has different implications for reliance versus compliance.

The threshold setting: Misses versus False alerts



FA-Miss Automation Errors of Diagnostic Systems



Alert	Yes		FA ☹️	Low Threshold ↑ FA Rate
	No	☹️ Miss		High Threshold ↑ Miss Rate

Threshold is set low to avoid alert misses (late alerts)

To keep miss rate low, as **base rate** of dangerous events decreases (ie. Lower traffic density), FA-rate will increase greatly: i.e., $P(\text{conflict/alert}) \ll 0.50$.

BUT

- False alerts and alarms have a dangerous effect on human trust and dependence on automation:
- The “CRY WOLF” effect. True alerts may be ignored or disregarded.

Therefore:

- Is a false alert really that much better than a late alert, when the observer has perceptual access to the raw data?

Attention Issues:

Two aspects of Diagnostic

Automation Dependence

```
graph TD; A([Automation Dependence]) --> B[Reliance:]; A --> C[Compliance:];
```

Reliance:

What pilot does when alert is silent.

Rely on miss-free automation.

Plenty of spare attention to do other tasks.

Compliance:

What pilot does when alert sounds.

Complies with alert to take action.

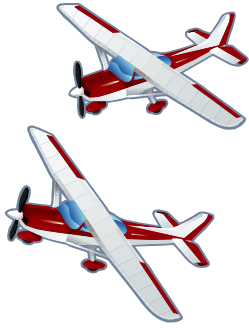
Rapid attention switch to alerted task.

The alert threshold.

Low vs. high.

High threshold → Misses: → Destroys Reliance

Low threshold → false alerts: → Destroys Compliance



Alert Domain
(Automated Task)

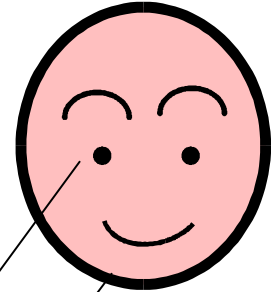
Raw Data

	FA
M	

Alert

Concurrent Tasks

Degraded by
False Alarm-Prone
Automation



Compliance

Switch
Attention

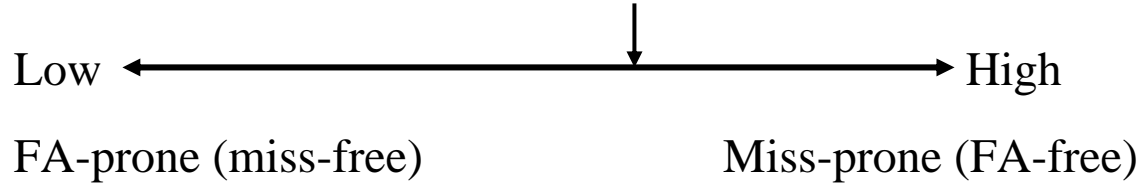
Reliance

Residual Attention

Degraded by
Miss-Prone
Automation

13 studies at University of Illinois. CDTI and UAV monitoring.

Threshold



Compliance	Low	High
Reliance	High	Low

What is the best tradeoff?

Compliance: Speed (or evidence) of switching attention to all alerts (whether true or false): Very fast with high compliance. Slow (or no) with low compliance (“Cry wolf”)

Reliance:

*1. Response to rare occasions when automation misses the event. Very slow with high reliance (“Complacency”). People don’t monitor the raw data when rely on automation

*2. Concurrent task performance. Very good. People re-allocation attention to concurrent tasks with high reliance.

False alarm disruptions?

THRESHOLD

FA-Prone

Miss-Prone

AUTOMATED TASK

Alerted Events

“CRY WOLF” ☹️
**Delayed (ignored)
response to all events
(true and false)**

Rare Automation Misses

“COMPLACENCY” ☹️
Detected poorly because
reliance on low-miss
automation is increased

NO COMPLACENCY 😊
Carefully monitor raw
data

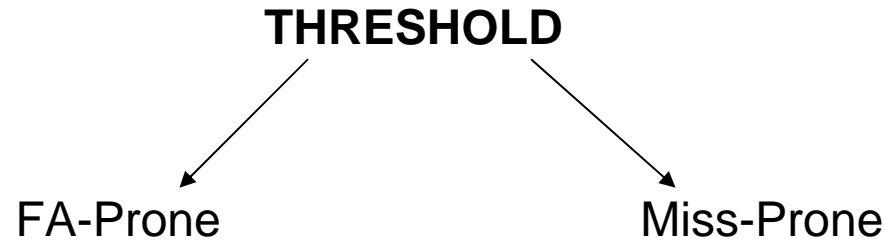
CONCURRENT TASK

INTERRUPTION ☹️
Switch attention to more
frequent alerts

GENERAL DISTRUST ☹️
False alerts are (more)
salient errors

LOW RELIANCE ☹️
**Need to monitor raw
data of automated task**

Strength of effect from
collective Illinois studies



AUTOMATED TASK

Alerted Events

Cry Wolf Cost
Strong Effect.
10/13 (77%) support
Only 1/13 (08%) refute.

Rare Automation Misses

Complacency Cost
Rare auto-miss is
undetected or delayed
Strong effect. 6/7 (86%)
No refute.

CONCURRENT TASK

Interruptions
Loss of trust
2/16 (12.5%)

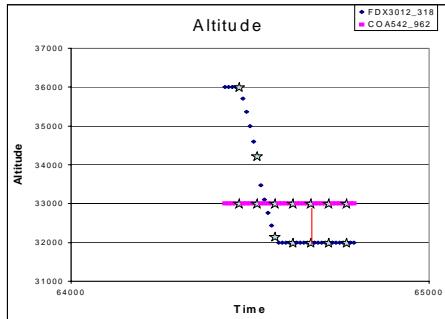
Low Reliance.
Diverted Resources to
raw data monitoring
(4/16 (25%))

No Difference: 7/16 (44%)

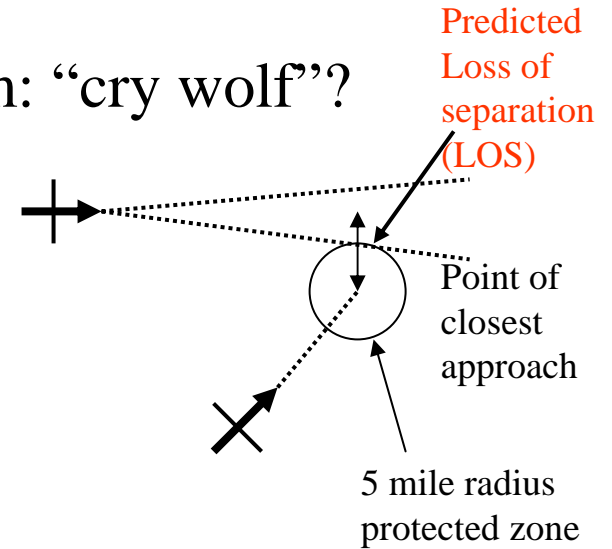
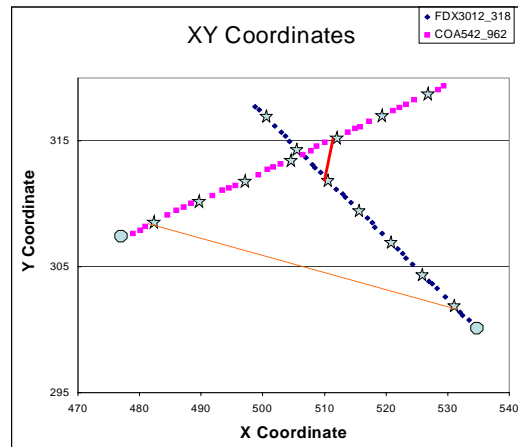
The FAA's Air Traffic Control FA-problem: "cry wolf"?

Conflict Alerts in the ATC Centers:

Real data with "live traffic" (Atlanta)



A false alarm CA; Zsl 09, C3



Actual min hor dist: 39.1

Actual min ver dist: 1600

Co-altitude reached and crossed well before min hor distance.

CA Rate: 0.5%

99.5% of encounters, the CA is silent

CA

Predicted LOS

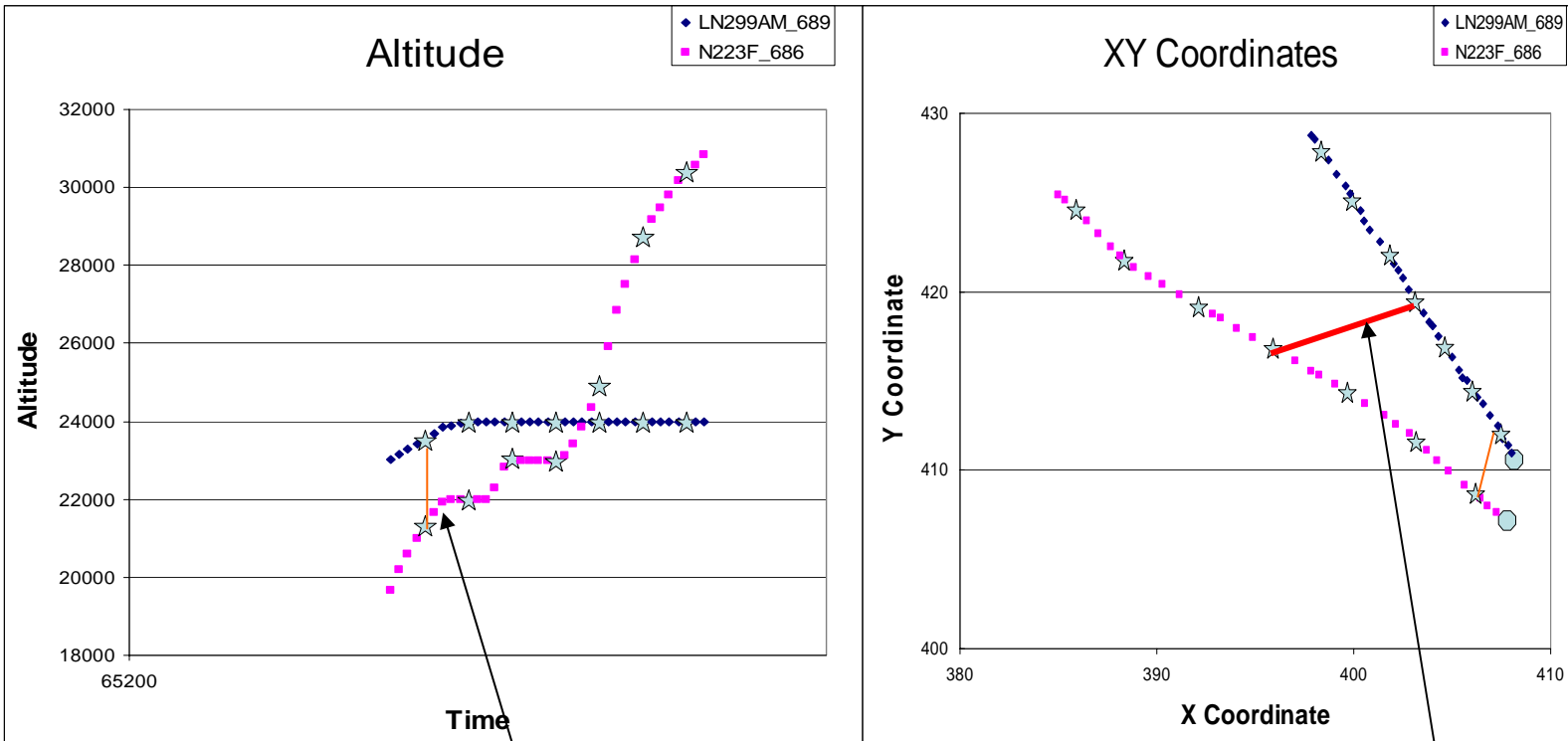
Safe trajectory

Silent

CA	0.80	0.20
Silent		

What drives the FA rate? How do controllers respond to a FA? Do they cry wolf when FA is more frequent?. The challenge of causal attribution with live data.

Definite controller response (level off pink)



Level off in response
to CA

Distance at co-altitude

SL 09 C9. Min max = 1.02

Acknowledgments

- Federal Aviation Agency
- NASA Ames Research Center
- Transportation Security Agency
- Army research Lab
- Jason McCarley, Stephen Dixon, Angela Colcombe, Xidong Xu, Ken Leiden.