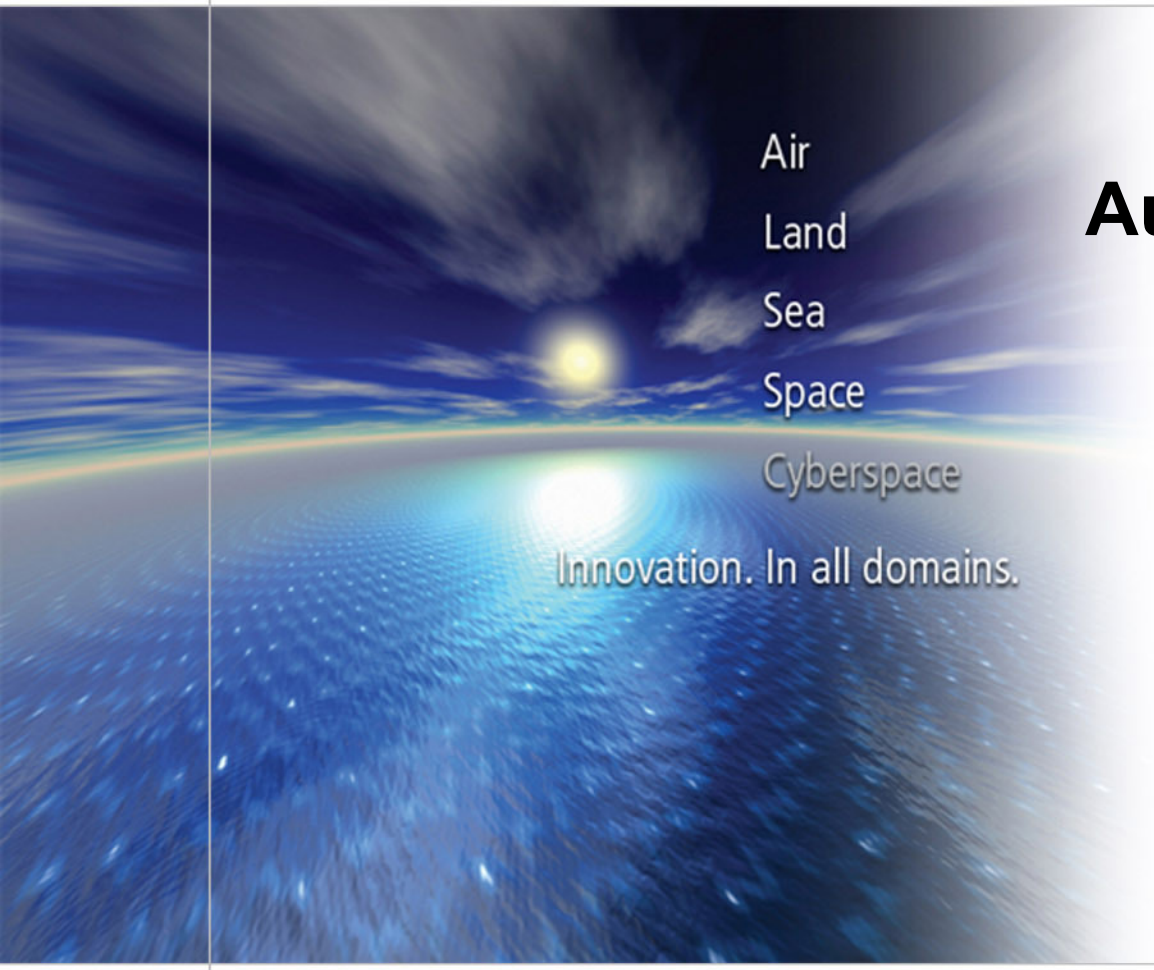
Raytheon

Customer Success Is Our Mission

"Trust in Autonomy is to DoD what safety is to FAA"



Creating Trust in Autonomous Systems – the Trust V

Dr. Dan Zwillinger
Gari Palmer
Anne Selwyn

Safe & Secure Systems & Software Symposium

12 June 2014

Trust In Autonomous Systems – Overview

What it is

- Develop and prototype methods for creating both Systems Trust and Operational Trust via the "Trust V"; enhancing system acquisition and use.
- Develop and prototype methods for trust measurement and assessment

Why it's important

- As Raytheon's systems move from automated to autonomous, successful deployment and adoption will require both Systems Trust and Operational Trust.

Key Technologies

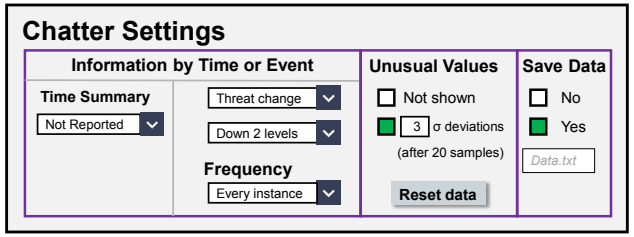
- Raytheon developed Trust techniques
  - "Chatter" and "What Agent"
  - "Future Risk Prediction"
- Trust Calibration

Key 2014 Deliverables

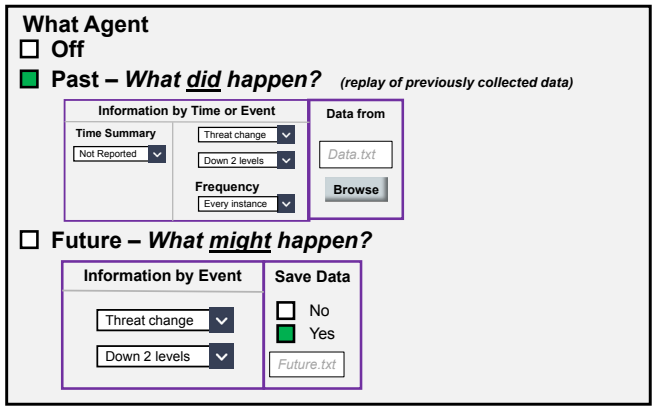
- Design, develop and add trust to 2+ Raytheon programs
- MOEs for "relative" trust calculation

Notational User Interfaces

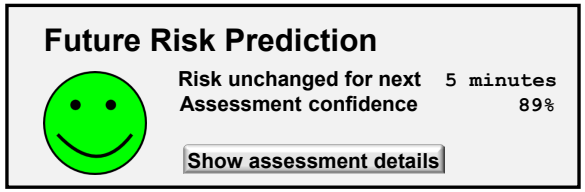
Chatter



What Agent



Future Risk Prediction



Need an affordable way to create Trust for autonomous systems

Raytheon

Trust – What / Why

What is Trust?

"The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."

Why do we care?

"For commanders and operators in particular, these challenges can collectively be characterized as a *lack of trust* that the autonomous functions of a given system will operate as intended in all situations."

Defense Science Board – The Role of Autonomy in DoD Systems – emphasis in the original

Trust goals

- *must* deliver value added capabilities
- *must* address the needs for different autonomy types
  - … does system have a learning capability?
  - … does system work in unstructured environments?
  - …
- *should* work within existing development process
- *should* support creation of modular components

While "Trust" adds belief in a result, it does not change the result

Systems Engineering V and "Trust V"

Autonomous Characteristics
- Mission
- Internal / System
- External / Environment

Identified trust attributes & needs

Trust techniques

T&E/V&V techniques

Requirements driven system

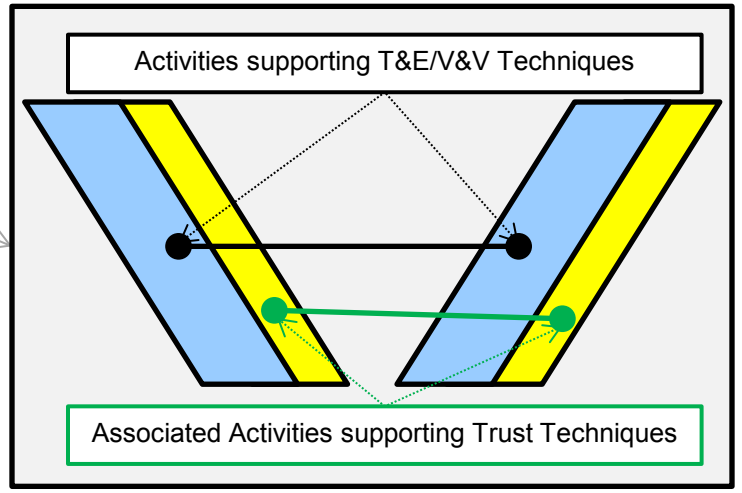
Operations

Trust
- If Trust V completed, can create operational trust
- When systems Engineering V completed, have system trust

Systems Engineering V

Trust V

Development



Key insight: treat "Trustability" as a non-functional requirement

18 Dimensions of Trust

1. Adaptive / Learning
2. Adversarial
3. Dynamic
4. Human proxy
5. Self-directed
6. Self-governed
7. Uncertain
8. Unstructured

ARPI "autonomous"

(Autonomy Research Pilot Initiative)

9. SoS (systems of systems) integration safe

Raytheon introduced

10. Perceived competence
11. Benevolence
12. Understandability
13. Directability
14. Reliability
15. Validity
16. Utility
17. Robustness
18. False-alarm rate

"automated"
(foundation for autonomous)

Need to create Trust in each dimension

Autonomy Trust V techniques

Raytheon

		Analysis preceeding Requirements	2 Requirements Analysis Process	3 Architectural Design Process	4 Implementation Process	6 Verification Process	8 Validation Process	9 Operation Process
	broadly applicable methods	For most of these, can use: (*) Simulation;	For most of these, can use: (*) Dynamic Fault Trees For most of these must have (*) Incorporation of MOEs			For most of these, can use: (*) Semantic Q&A Capability; (*) Future scenario prediction	For most of these, can use: (*) Semantic Q&A Capability; (*) Future scenario prediction	For most of these, can use: (*) Semantic Q&A Capability; (*) Future scenario prediction (*) Automated capture, analysis and feedback loop for Measures of trust
Autonomy characteristics	Adaptive / Learning							
	Adversarial	Information Assurance; Cyber Protection / Security; IFF Capability	Information Assurance; Cyber Protection / Security; IFF Capability	Cyber Protection / Security; IFF Capability	Information Assurance; Anti-tamper	Negative testing		
	Dynamic		Boundary definitions; Temporal considerations; Goal driven based on Beliefs, Desires, and Intention		Graceful degradation; Boundary testing	Testing to failure		
	Human interaction		Flexible autonomy; Human profiling; Require human adjustable interaction		Unified mission-based language for human / machine interaction			Performance confirmation
	Self-directed	Create positive statements of preventation of negative	Build in Directability; Include positive statements;	build in Directability	Run time assurance	Run time assurance (purpose); Formal methods based on Beliefs,	Test against CONOPS	
	Self-governed							
	Uncertain	Prediction of Mission Threats	Measure of uncertainty & confidence; Reasoning of Situational Awareness for dynamic threat prediction	Sensor reliabiity analysis			"Turing" Test Future Risk Prediction method	
	Unstructured		Environmental abstraction and characterization of mission threats; Measure of environment modeling		Inference engine to categorize Mission Threats; Boundary testing; Safety mechanisms	Extreme testing to failure; Hierarchical tested across abstraction levels; State transitions		
	SoS integration safe	TLYF (test like you fly); ConOps Analysis				TLYF	SoS certification; Thread based system testing	
Automation - trust attributes	Perceived competence		Executable Models; Calibrated trust		*Model CONOPS -- needed for "Semantic Q&A Capability"; *Identified Human trust needs -- needed for "ATCI"			ATCI (Calibrated Trust)
	Benevolence	Hazard analysis	SMRI (SW Mishap Risk Index); Adaptable safeguards	Domestically manufactured components	Adaptable safeguards	Negative testing		
	Understandability		Restrictive language (e.g., Behavior Trees); Model Based Engineering; DoDAF artifacts	DoDAF artifacts	Traceability; *Model DoDAF artifacts -- needed for "Semantic Q&A Capability"; Machine transparenc; "dead code" removal; traceability		Semantic Q&A	
	Directability		Requirements specification			Demonstration of override		Demonstration of override
	Reliability	MTBF analysis; MTTF analysis; FMECA	Independent concurrent SW/HW efforts; use of "taint checking" use of "trademarking"	Redundancy; Fault tolerant processing	Run time assurance; Reuse; Mean time to Recovery; Graceful Degradation; Model Based Engineering; FMECA; autonomous footprint as small as possible; Automated Unit Test; Continuous Integration; Automatic code generation	Hybrid systems verification; Accelerated aging; Test Optimization; Run time assurance; DO178C		
	Validity	External SME Validation	Formal Methods Acceptance; Model Based Engineering			Formal Methods Acceptance (purpose)	External SME Validation;	
	Utility		Statement of performance		Run time assurance	Proof of performance; Run time assurance	Test against CONOPS	
	Robustness	Monte Carlo simulations; Fault Tree Analysis		Redundancy; Fault tolerant processing	Reuse; DFMA; Graceful Degradation	Longevity Testing; Fault injection; Black Box Testing;		

Many table entries – Trust Dimension versus Development Step

Raytheon

Autonomy Trust V techniques – *top left corner*

	Analysis preceeding Requirements	2 Requirements Analysis Process	3 Architectural Design Process	4 Implementation Process
broadly applicable methods	**For most of these, can use:** **(*) Simulation;**	**For most of these, can use:** **(*) Dynamic Fault Trees** **For most of these must have** **(*) Incorporation of MOEs**		
Adaptive / Learning				
Adversarial	Information Assurance; Cyber Protection / Security; IFF Capability	Information Assurance; Cyber Protection / Security; IFF Capability	Cyber Protection / Security; IFF Capability	Information Assurance; Anti-tamper
Dynamic		Boundary definitions; Temporal considerations; Goal driven based on Beliefs, Desires, and Intention		Graceful degradation; Boundary testing
Human interaction		Flexible autonomy; Human profiling		Unified mission-based language for human / machine interaction
Self-directed **Self-governed**	Create positive statements of preventation of negative	Build in Directability; Include positive statements;	build in Directability	Run time assurance
Uncertain	Prediction of Mission Threats	Measure of uncertainty & confidence; Reasoning of Situational Awareness for dynamic threat prediction	Sensor reliabiity analysis	
Unstructured		Environmental abstraction and characterization of mission threats; Measure of environment modeling		Inference engine to categorize Mission Threats; Boundary testing; Safety mechanisms
SoS integration safe	TLYF (test like you fly); ConOps Analysis			
Perceived competence		Executable Models; Calibrated trust		Model CONOPS -- needed for "Semantic Q&A Capability"; Identified Human trust needs -- needed for
B		Adaptable safeguards	components	

Many table entries – Trust Dimension versus Development Step

Raytheon

Autonomy Trust V MOEs

	type	Operational Trust Metrics		
		passive	**active**	**historical**
Autonomy characteristics	**Adaptive / Learning**			Determine if system response to similar inputs has changed over time
	Adversarial		Evaluate results of system operation applied to operator/SME supplied what-if examples	Number of adversaries identified, and percentage of those addressed
	Dynamic	If operator must accept tainted results before they can be used then number of tainted results accepted by operator		*(if has occurred)* Assessment of quality of graceful degradation
	Human interaction		SUS score	
	Self-directed			Percentage of times that system "took appropriate action" without operator intervention given an
	Self-governed			
	Uncertain			Operator/SME rating of historical examples of system response to actually observed
	Unstructured			
	SoS integration safe			
Attributes	**Perceived competence**	Percentage of time operator chooses to not override system (decrease);	Evaluate results of system operation applied to operator/SME supplied what-if examples	Number of "functions" executed per unit time by computer (vs humans) (desire increased); Ratio for completion of "Mission-Critical Objectives" vs. "Secondary Objectives" (increase); Time to respond to critical events (decrease); Time to respond to non-critical events (decrease).
	Benevolence		"Lead it into temptation and see if it delivers evil" ==> Create "tempting" scenarios and assess if malevolence results	Percentage of time that system operates outside of "safeguarded" regions
	Understandability	Number operator queries per unit time; Percentage of queries which are followed immediately by another query	Evaluate results of query usefulness by operator/SME; SUS score	Time required for operator training (desire decrease)
		Response time when operator overrides system; back to system		

Many table entries– Trust Dimension versus MOE type

"Chatter" and "Why/What Agent"

1. Operator trust in the system is enhanced when

- The **user understands** why the system performs as it does
- Communications are in the **language of the user**
- The user is pleasantly **surprised** by meaningful yet unexpected information
  - Think "automated performance monitoring"

2. The speed of trust is enhanced when

- The operator controls the **frequency of communication**
- The operator controls the **information content level**
  - Layered control in many systems
  - HW: assemblies to sub-assemblies to components
  - SW: libraries to modules to functions
- The operator controls the **information details**

3. Communications between system and user

Attribute Tool	Chatter	What Agent	Why Agent
Query structure	pre-defined	scripted	free-form
Content level	operator selects	script driven	operator driven
Information details	operator selects	script driven	operator driven
Frequency of communications	operator selects	operator driven	operator driven
Driven by	system	operator	operator

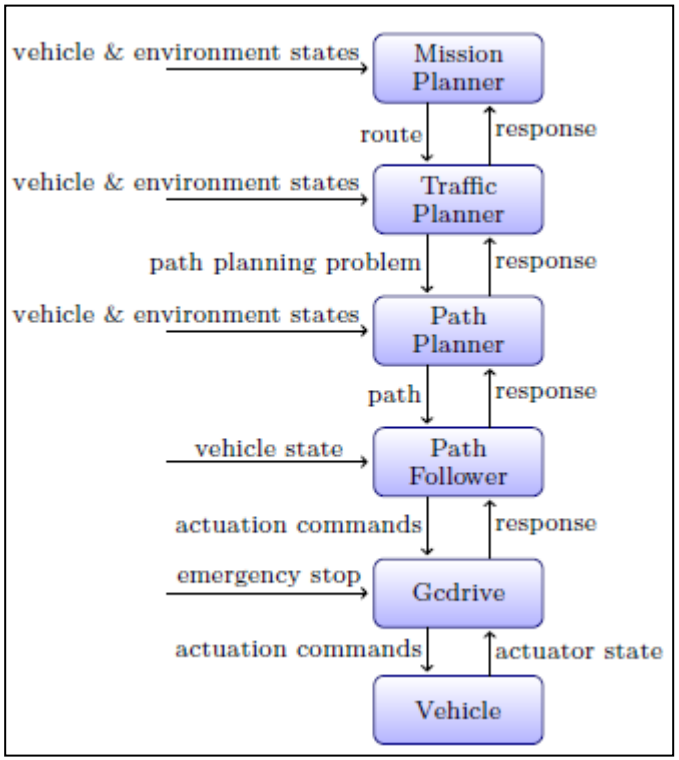
Many systems use a layered control model

BASIS - Fast Attack Craft / Fast Inshore Attack Craft (FAC/FIAC)

1. Threat Change
2. Behavioral Change
3. Last Trigger Change
4. Input Change

Alice – Caltech's entry in DARPA'a 2007 Urban Challenge – "navigation protocol stack" is

1. Mission Planner
2. Traffic Planner
3. Path Planner
4. Path Follower
5. Gcdrive
6. Vehicle



Want to know – at level (1-4) or (1-6)

- What changed?
- Why did it change? (recursively)

"Synthesis of Control Protocols for Autonomous Systems", Wongpiromsarn, Topcu, & Murray, December 2012, Figure 1,
https://www.google.com/url?q=http://www.cds.caltech.edu/~murray/preprints/wtm12-us_s.pdf&sa=U&ei=k9ZGU66qPlamsASt54CoBw&ved=0CBsQFjAA&usg=AFQjCNEb4KRpsDhTx9cqrthDzCZ7_zJnzQ

Chatter naturally supports hierarchical control systems

"Why/What Agent" and "Chatter" graphical representation

Why/What Agent ➔ Operator initiated queries

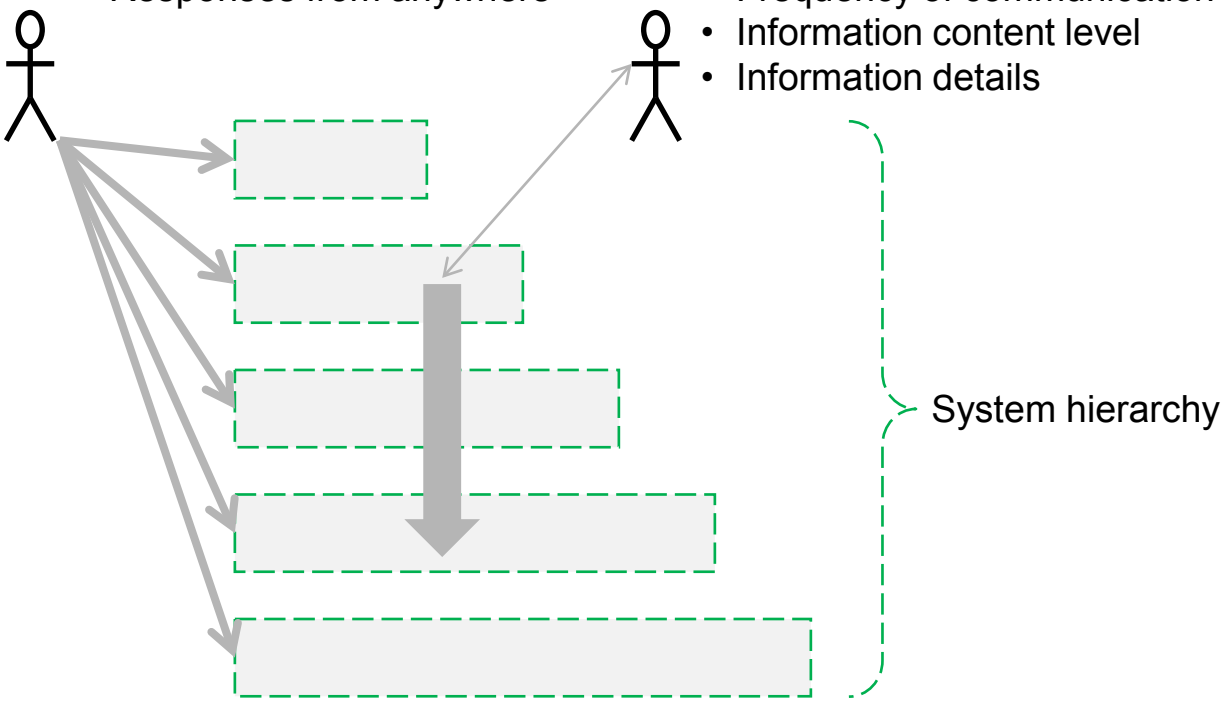
- Either
  - What ➔ Template driven ➔ "What is the reason for _____?" *easier*
  - Why ➔ Free form input ➔ "Why did the system do _____?" *harder*
- Operator can drill down to desired hierarchy level as needed
- Operator controls "level" of response

Why/What

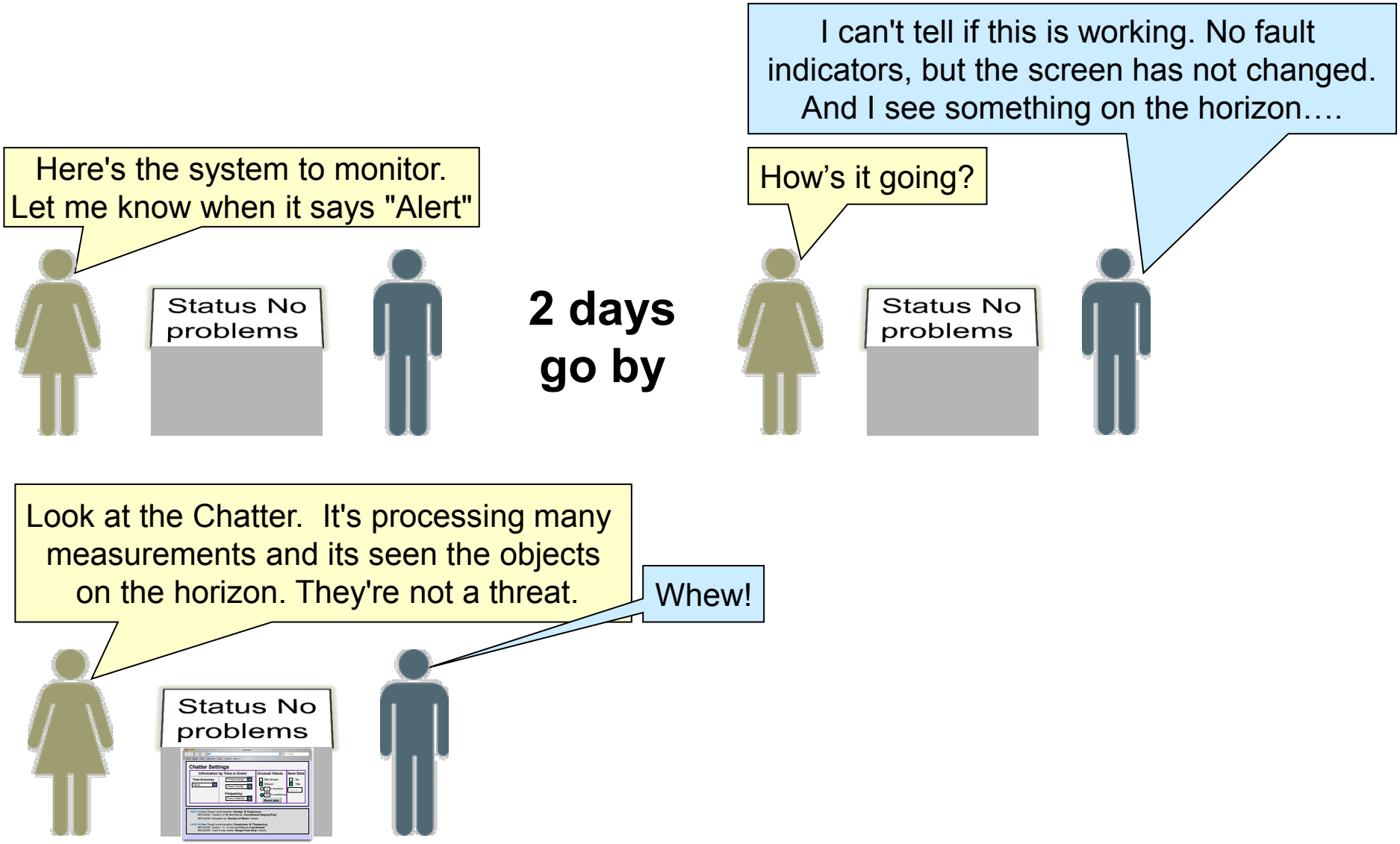
- Responses from anywhere

Chatter ➔ System initiated communications

- Frequency of communication
- Information content level
- Information details



Example of Chatter use



“Chatter” – notional interface

Untitled

Apple .Mac Amazon eBay Yahoo! News

Chatter Settings

Information by Time or Event

Time Summary

Not Reported

Threat change

Down 2 levels

Frequency

Every instance

Unusual Values

☐ Not shown

☑ 3 σ deviations

(after 20 samples)

Reset data

Save Data

☐ No

☑ Yes

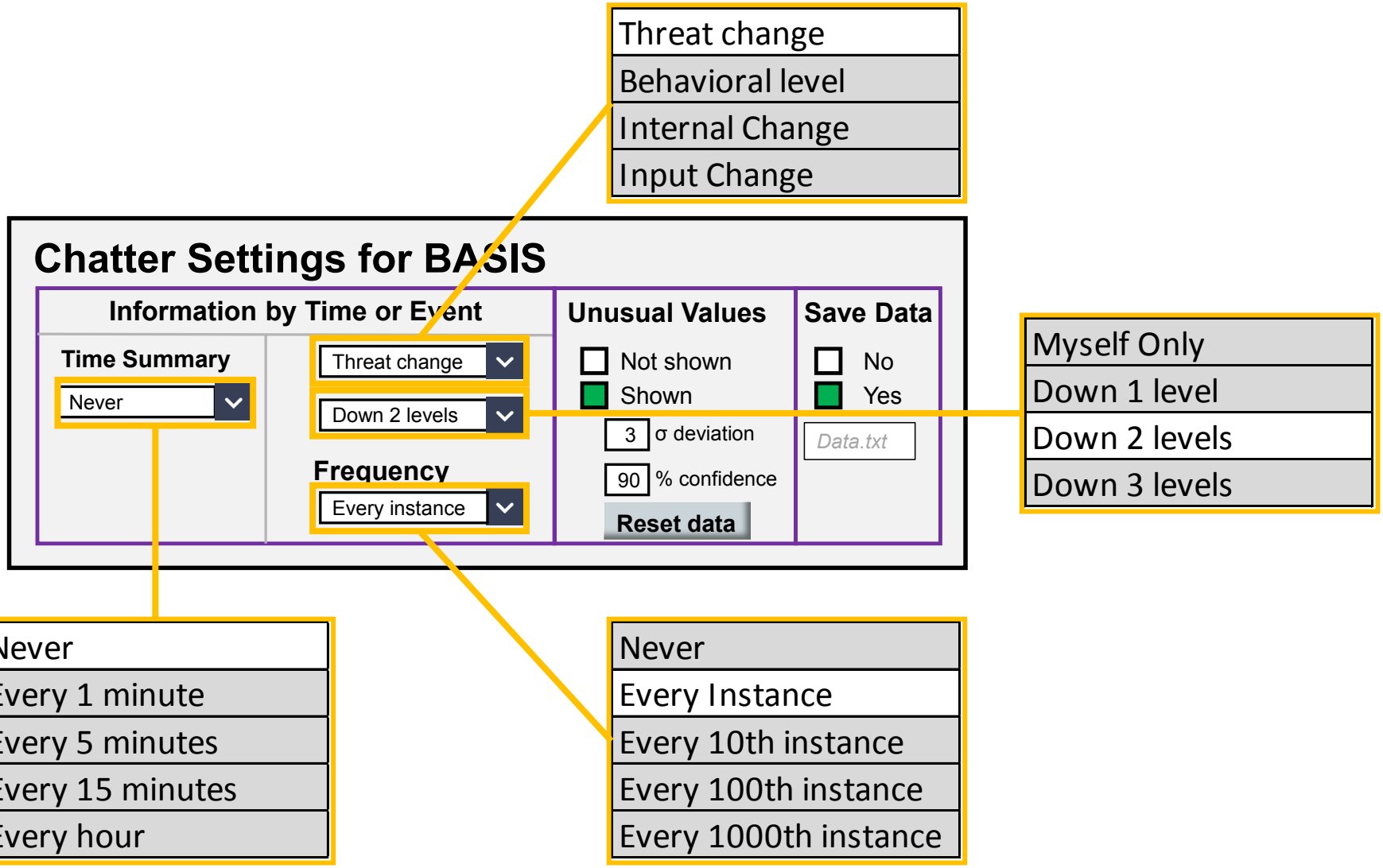
Data.txt

18:07:18 New Threat Level transition: **Benign ➔ Suspicious**
BECAUSE: Tracks 4, 8 ➔ identified as “**Criteria ABC**”
BECAUSE: Activated via “**Boat Criteria 3**” criteria

18:05:16 New Threat Level transition: **Suspicious ➔ Threatening**
BECAUSE: Tracks 1, 5, 12 now identified as **Coordinated**
BECAUSE: Track 5 now meets “**Criteria DEF**” criteria

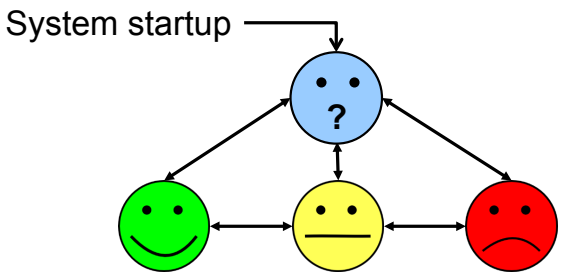
Applied to Fast Attack Craft / Fast Inshore Attack Craft (FAC/FIAC)

"Chatter" – Operator controls



Operator controls the level/details/frequency of communications

Future Risk Prediction – notional interface



Future Risk Prediction

Risk unchanged for next unknown
Assessment confidence unknown

Show assessment details

Future Risk Prediction

Risk unchanged for next 5 minutes
Assessment confidence 89%

Show assessment details

Future Risk Prediction

Risk unchanged for next 4 minutes
Assessment confidence 80%

Show assessment details

Future Risk Prediction

Risk unchanged for next 2 minutes
Assessment confidence 92%
Risk driven by *Diverse possibilities*

Show assessment details

Future Risk Prediction

Risk unchanged for next 5 minutes
Assessment confidence 89%

Hide assessment details

Number of similar states 23
1 had negative results 4%

Current Pursuit: Identifying target
Number of states required 15
State match threshold 0.1
Time increment (min) 1

Explanation of high risk assessment

Output shows independent risk assessment of system capability

BACKUP

Definitions from ARPI *(Autonomy Research Pilot Initiative)*

Automation

The system functions with no/little human operator involvement. However, the system performance is limited to the specific actions it has been designed to do. Typically these are well-defined tasks that have predetermined responses, i.e. rule-based responses in reasonably well-known and structured environments.

Autonomy

Systems which have a set of intelligence-based capabilities that allow it to respond within a bounded domain to situations that were not pre-programmed or anticipated in the design (i.e., decision-based responses) for operations in unstructured, dynamic, uncertain, and adversarial environments. Autonomous systems have a degree of self-governance and self-directed behavior and must be adaptive to and/or learn from an ever-changing environment (with the human's proxy for decisions).

Definitions of *Autonomous* and *Automation*

Dimensions Describing the Basis of Trust

TABLE 1: Summary of the Dimensions That Describe the Basis of Trust

Study	Basis of Trust	Summary Dimension
Barber (1983)	Competence	Performance
	Persistence	Process
	Fiduciary responsibility	Purpose
Butler & Cantrell (1984)	Competence	Performance
	Integrity	Process
	Consistency	Process
	Loyalty	Purpose
	Openness	Process
Cook & Wall (1980)	Ability	Performance
	Intentions	Purpose
Deutsch (1960)	Ability	Performance
	Intentions	Purpose
Gabarro (1978)	Integrity	Process
	Motives	Purpose
	Openness	Process
	Discreetness	Process
	Functional/specific competence	Performance
	Interpersonal competence	Performance
	Business sense	Performance
	Judgment	Performance
Hovland, Janis, & Kelly (1953)	Expertise	Performance
	Motivation to lie	Purpose
Jennings (1967)	Loyalty	Purpose
	Predictability	Process
	Accessibility	Process
	Availability	Process
Kee & Knox (1970)	Competence	Performance
	Motives	Purpose
Mayer et al. (1995)	Ability	Performance
	Integrity	Process
	Benevolence	Purpose
Mishra (1996)	Competency	Performance
	Reliability	Performance
	Openness	Process
	Concern	Purpose
Moorman et al. (1993)	Integrity	Process
	Willingness to reduce uncertainty	Process
	Confidentiality	Process
	Expertise	Performance
	Tactfulness	Process
	Sincerity	Process
	Congeniality	Performance
	Timeliness	Performance
Rempel et al. (1985)	Reliability	Performance
	Dependability	Process
	Faith	Purpose
Sitkin & Roth (1993)	Context-specific reliability	Performance
	Generalized value congruence	Purpose
Zuboff (1988)	Trial and error experience	Performance
	Understanding	Process
	Leap of faith	Purpose

Review of trust dimensions

- *Trust in Automation: Designing for Appropriate Reliance* by Lee & See
- `http://www.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf`

List of trust dimensions

- *benevolence*
- *directability*
- *false-alarm rate*
- *perceived competence*
- *reliability*
- *robustness*
- *understandability*
- *utility*
- *validity*

- *Trust in Automation*, Hoffman, Johnson, Bradshaw, Underbrink
- `http://www.jeffreymbradshaw.net/publications/50.%20Trust%20in%20Automation.pdf`