Presentation to the Human-AI Teaming
Flex-Tech International Industrial Spring School
Biarritz, France — May 2024

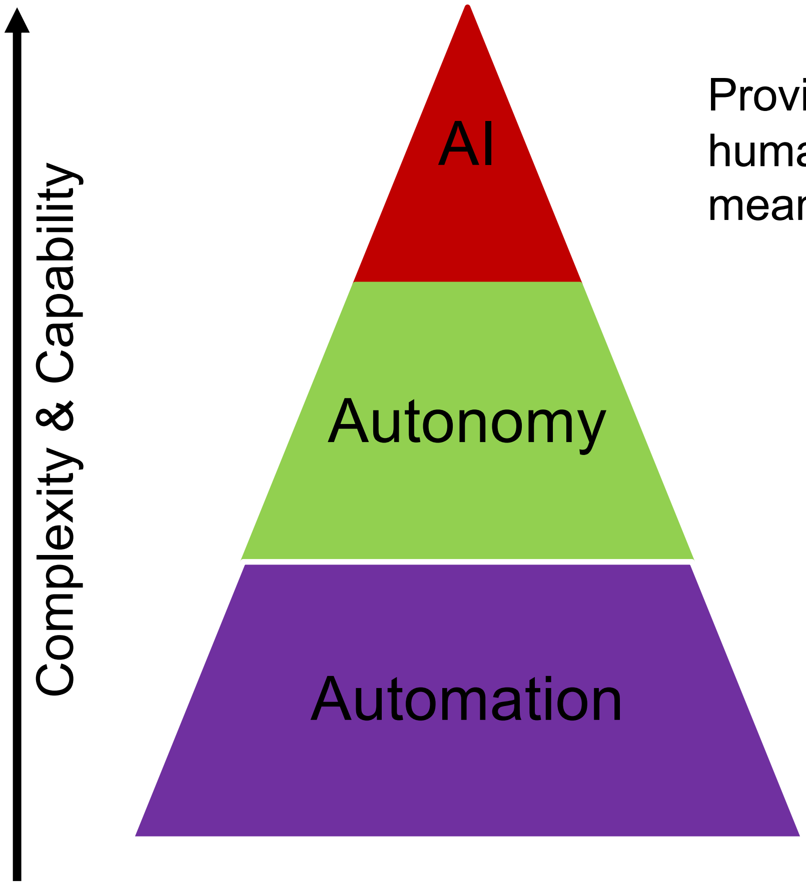
Supporting Human-AI Teaming:

Transparency, Explainability, and Situation Awareness

Mica R. Endsley, PhD
SA Technologies

Some Definitions

SATechnologies



Provides intellectual processes similar to those of humans, including the “ability to reason, discover meaning, generalize, or learn from past experience”

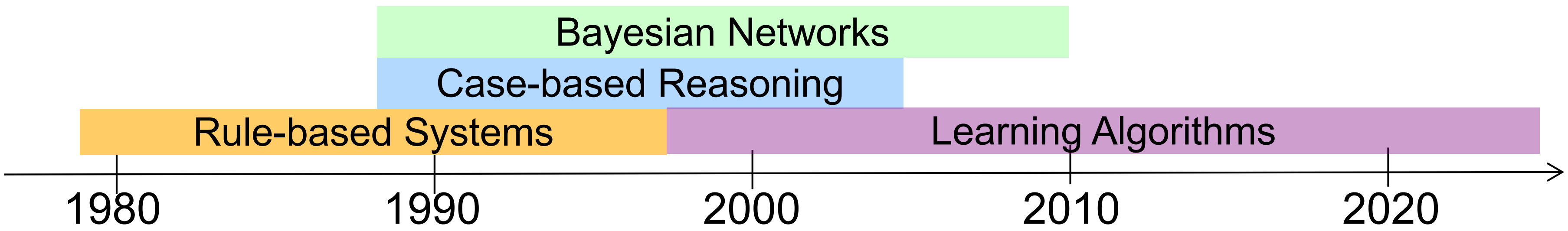
Intelligence-based capabilities that can respond to situations that were not explicitly programmed or were not anticipated in the design

Technology that performs tasks independently, without continuous input from an operator

AI describes a form of highly capable automation directed at highly perceptual and cognitive tasks.

The Rise of Artificial Intelligence

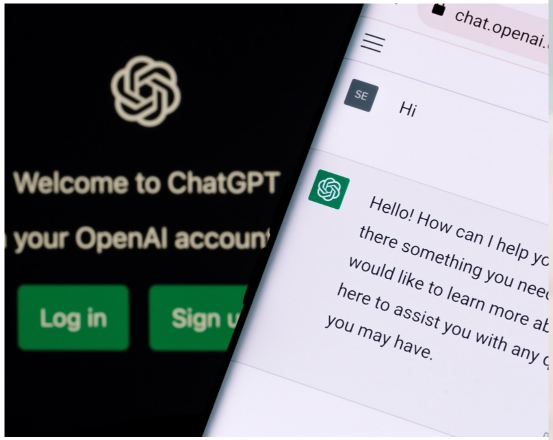
- Dartmouth Summer Research Project on AI – 1956
  - John McCarthy: Artificial Intelligence is "the science and engineering of making intelligent machines"
- Increase in computational power
- Rise of big data
- Deep learning

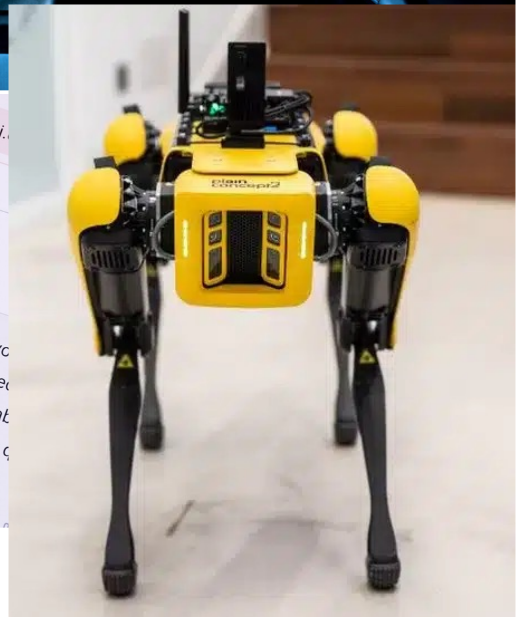


An intelligent system is defined as one that recognize situations, adapts to changes and generates solutions to even novel problems, and can act to optimize performance

AI Applications

- E-Commerce
- Information Systems
- Driving
- Healthcare
- Robotics
- Finance
- Aviation
- Military
- Policing & Security
- Manufacturing





Challenges with AI

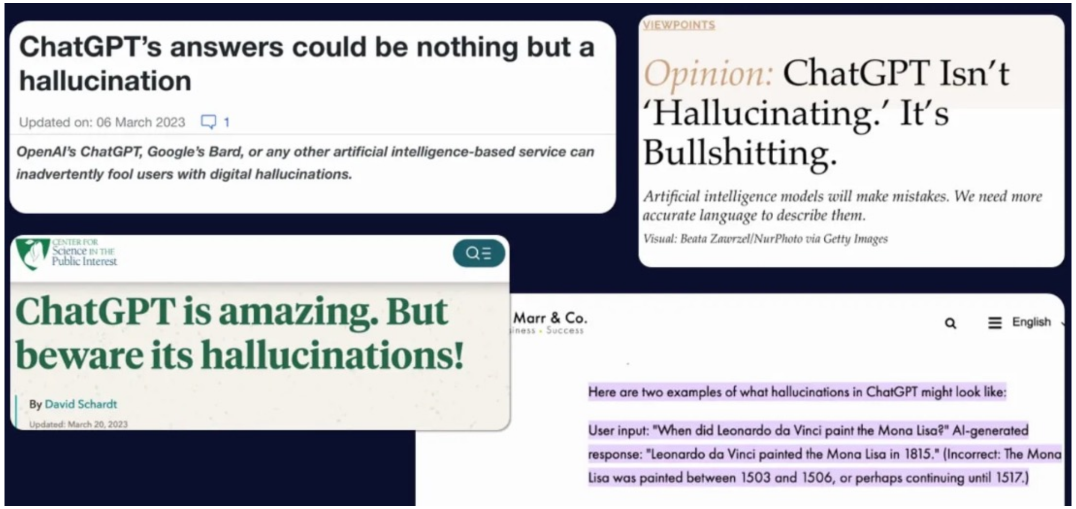
- Expectation that AI will improve quality and efficiency of operations
  - Automatically
  - As input to current decision making and processes
- Danger
  - AI is just an advanced form of automation
    - No understanding
    - No common sense or reasoning
    - No knowledge of its own limits
  - Long history of automation helping with routine tasks, but also increasing the likelihood of catastrophic failures

AI May Not be Accurate

SA Technologies

- Deep Fakes
  - Images
  - Video
  - Speech
  - Text

- Hallucinations



ChatGPT's answers could be nothing but a hallucination

Updated on: 06 March 2023 1

OpenAI's ChatGPT, Google's Bard, or any other artificial intelligence-based service can inadvertently fool users with digital hallucinations.

VIEWPOINTS

Opinion: ChatGPT Isn't 'Hallucinating.' It's Bullshitting.

Artificial intelligence models will make mistakes. We need more accurate language to describe them.

Visual: Beata Zawrzel/NurPhoto via Getty Images

Center for Science in the Public Interest

ChatGPT is amazing. But beware its hallucinations!

By David Schardt

Updated: March 20, 2023

Marr & Co.

English

Here are two examples of what hallucinations in ChatGPT might look like:

User input: "When did Leonardo da Vinci paint the Mona Lisa?" AI-generated response: "Leonardo da Vinci painted the Mona Lisa in 1815." (Incorrect: The Mona Lisa was painted between 1503 and 1506, or perhaps continuing until 1517.)

AI Has Limited Reliability & Robustness

- **AI and system autonomy will be unable to handle many unforeseen (unlearned) situations for the near future**
  - **Perceptual limitations**
    - Continue to struggle with reliable and accurate object recognition in “noisy” environments
  - **Brittleness**
    - Only capable in situations that are covered by its training
    - Learning “lag”
  - **Hidden biases**
    - Hidden biases from using a limited set of training data, or from biases within that data itself.
  - **No model of causation**
    - AI cannot use reason to understand cause and effect, it cannot predict future events, simulate the effects of potential actions, reflect on past actions, or learn when to generalize to new situations. (Pearl & Mackenzie, 2018)

Will you be ready for the unexpected?

SA is Critical to Autonomy Oversight & Interaction

SA Technologies

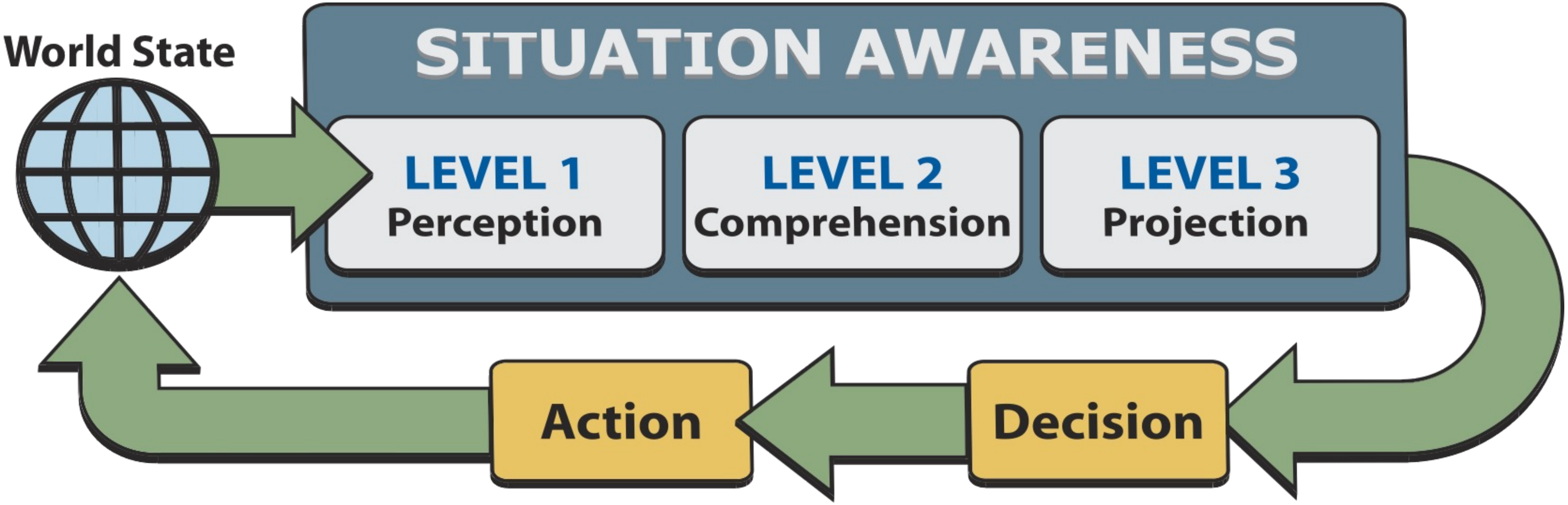
- Understanding of its status
- How well is it functioning
- When interventions are needed and what kind
- How the system's status effects operator tasks and vice-versa
- Is it meeting my goals?

Automation State

Task State

System Environment

What is Situation Awareness?



Situation Awareness is the *Perception* of elements in the environment within a volume of time and space, the *Comprehension* of their meaning, and the *Projection* of their status in the near future.*

*Endsley, 1988

Effects of Automation and AI on Human Performance are Well Known

- **Automation Confusion**
  - ***What is it doing? Why? What next?***
  - **Poor mental models of AI**
  - **Poor understanding and projection**

- **Automation confusion is most likely to occur when:**
  - **The automation acts on its own without immediately preceding directions from the operator,**
  - **The operator has gaps in knowledge of how the automation will work in different situations,**
  - **Weak feedback is provided on the activities of the automation and its future activities relative to the state of the world**

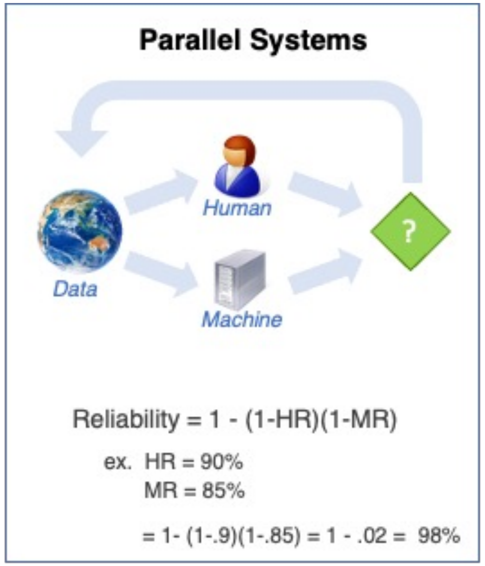
Out-of-the Loop Loss of SA

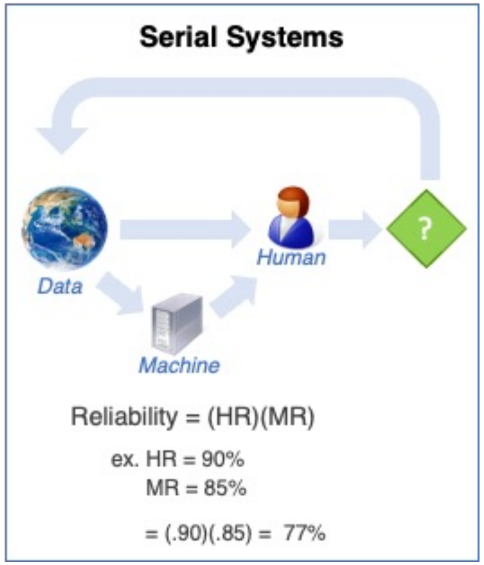
SATechnologies

- Low SA on how the automation is performing
  - Slow to detect problems with system or automation
  - Slow to regain understanding of what it is doing and taking over manually
- Loss of Situation Awareness
  - Vigilance , Monitoring and Trust
  - Changes in information feedback
    - Intentional
    - Unintentional
  - Level of Engagement
    - Active vs. Passive processing

Decision Biasing

- **Even when the system just makes recommendations, it affects performance**
  - **If system is correct → human performance better**
  - **If system is incorrect → human performance is worse**
- **People are not independent cross-checkers of AI recommendations**
  - **They include system inputs into their decision process**





Overall human-system performance is degraded

Irony of Automation

- People are increasingly unable to perform when they need to take over for automation (Bainbridge, 1983)
  - Increases in Cognitive Workload
    - More complex system
  - Reduction of Manual Skills
  - Less Understanding of What is Happening

- Increase in Catastrophic Failures
  - Lumberjack effect (Wickens)

B737-Max8

- **Lion Air 610 (October 2018)**
  - Crashed 13 minutes after take-off from Jakarta
  - 189 fatalities
- **Ethiopian Airlines 302 (March 2019)**
  - Crashed 6 minutes after take-off from Nairobi
  - 157 fatalities

- **Maneuvering Characteristics Augmentation System (MCAS)**
  - Larger engines installed
  - MCAS provided pitch stability
  - Insufficient reliability – produced erratic, uncontrollable actions
    - **No redundancy - based on only 1 of 2 AOA sensors**
    - **Repeated, inaccurate trim corrections made it difficult for pilots to correct and overcome**

Estimated $19B Cost to Boeing and Air Carriers

Boeing 737-Max8: Automation Implementation Gone Wrong

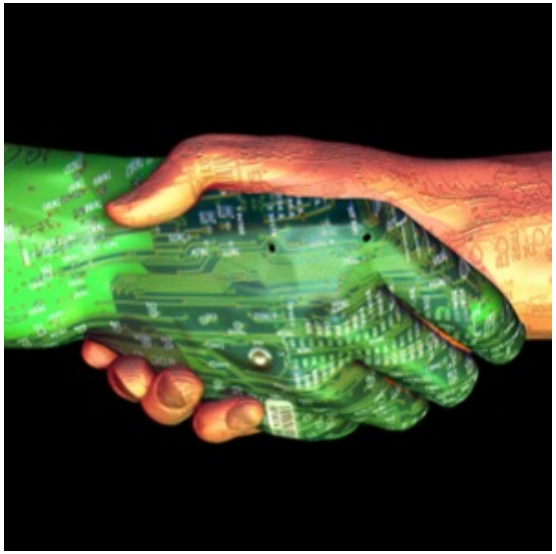
SATechnologies

- MCAS automation not in manuals and not trained
  - Assumption that it will always work perfectly
  - Created significant automation confusion
- High pilot workload
  - Manually controlling aircraft and trying to troubleshoot problem
- No indication of MCAS actions on displays
  - Silent actor leaves pilots confused as to what system is doing and why
  - No AOA sensor display (sold only as upgrade)
- Confusing array of alarms unrelated to AOA sensors or MCAS.
  - Assumption that pilots will respond immediately (3 seconds) with correct work around procedure (Stab Trim Cut-off)
  - Altitude disagree and indicated airspeed disagree did not point to real problem and sent them on the wrong path
  - Responses to alarms and alerts are affected by many factors including the salience of the alert for gaining attention, form of presentation, agreement/ disagreement with other indicators, and prior experience with the alert
    - NASA Study found that the probability of responding correctly for non-trained aircraft emergencies was only 7%, as compared to highly trained "text-book" emergencies at 86%
- Loss of SA
  - Ethiopian Airline crew performed Stab Trim Cut-off procedure but still could not control aircraft manually – auto-throttle kept airspeed high

Need Effective Oversight of AI and Autonomous Systems

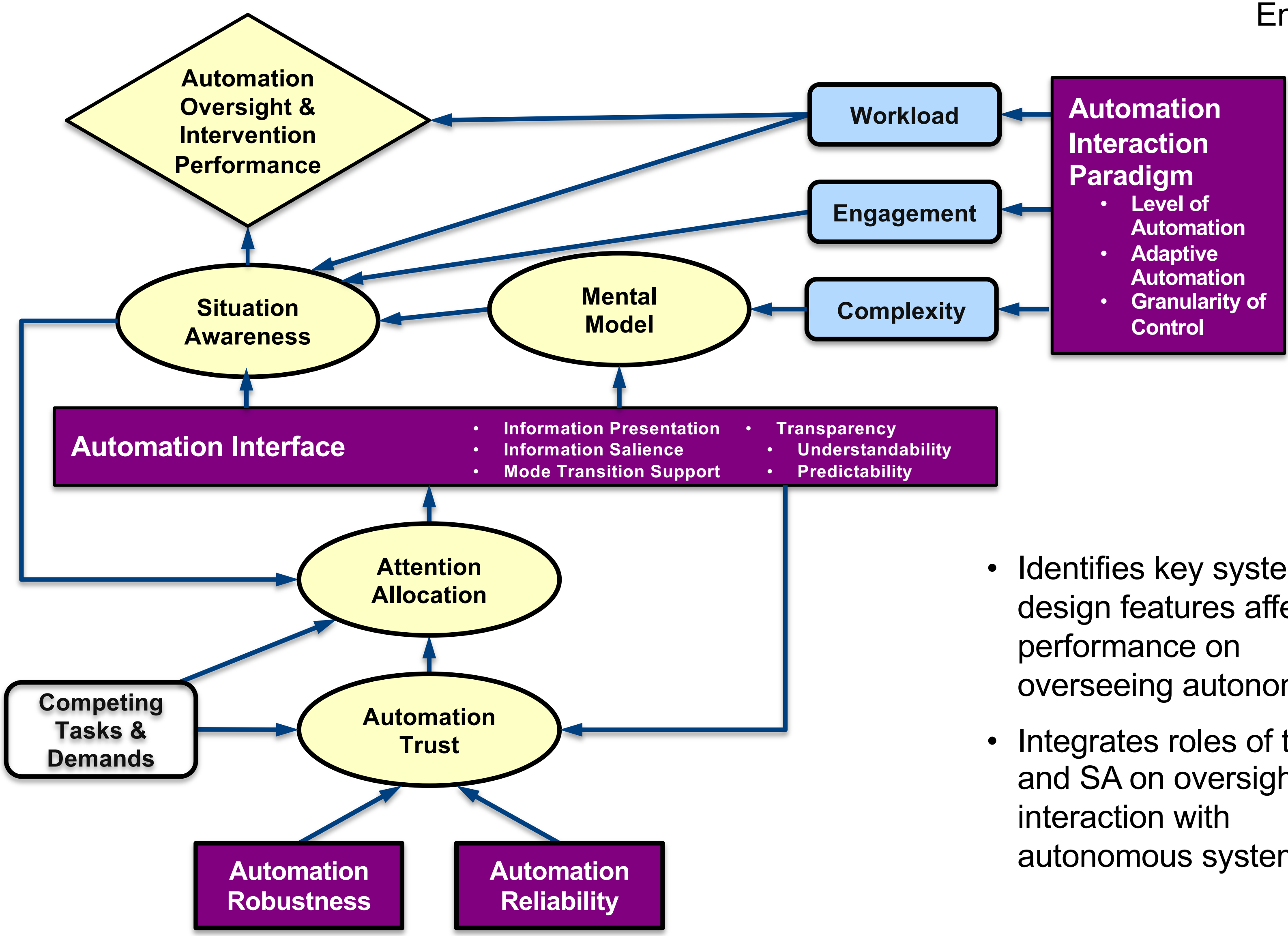
- AI and system autonomy will not be able to handle many unforeseen (unlearned) situations for the near future

- Synergistic human & AI team is critical to success
  - Overseeing what system is doing
  - Intervening when needed
  - Coordination and collaboration on functions



Human Autonomous System Oversight (HASO) Model

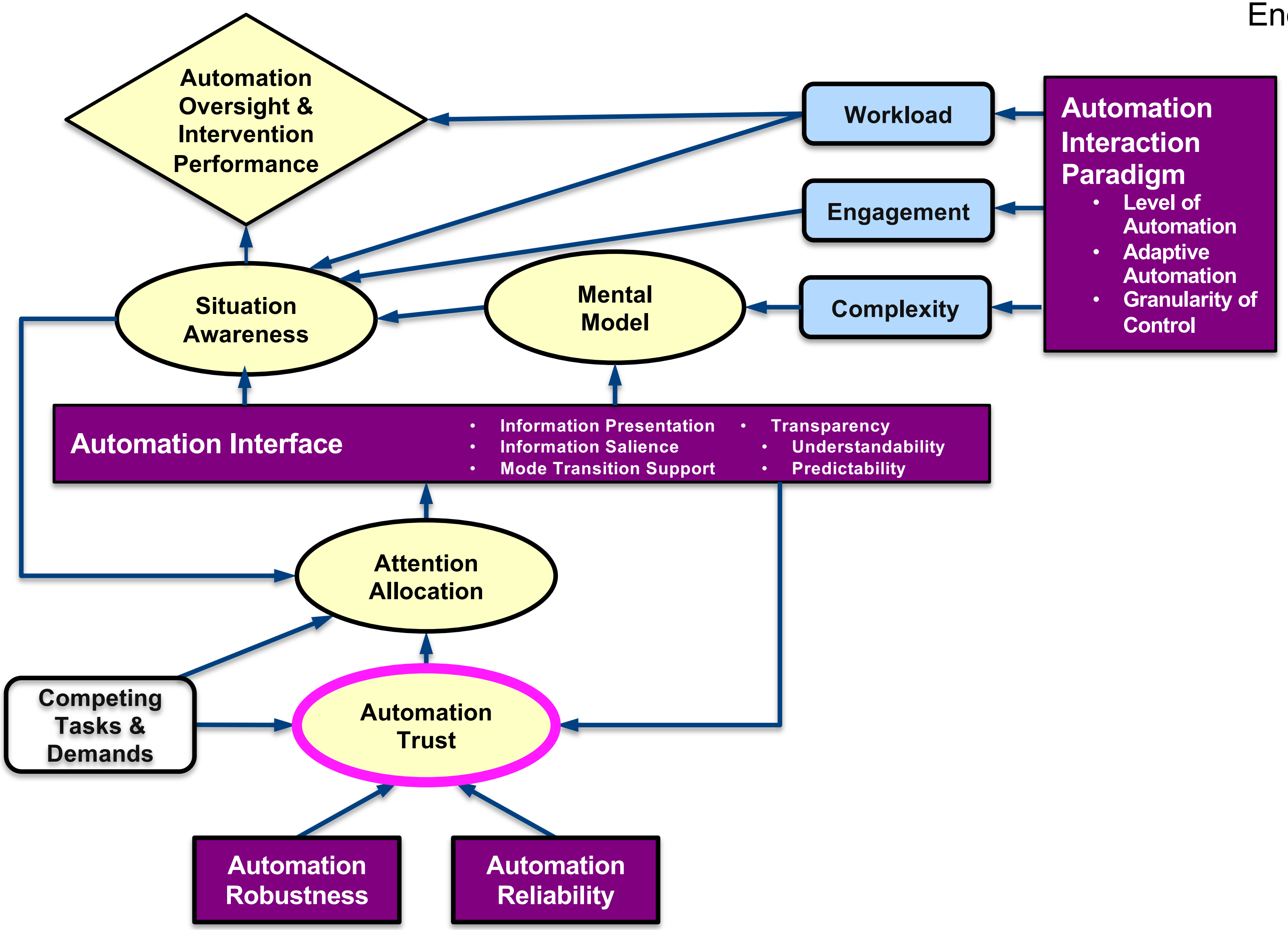
Endsley, 2017



- Identifies key system design features affecting performance on overseeing autonomy
- Integrates roles of trust and SA on oversight & interaction with autonomous systems

Human Autonomous System Oversight (HASO) Model

Endsley, 2017



Informed Trust Requires SA

SA Technologies

- Oversight
- Intervention
- Coordination

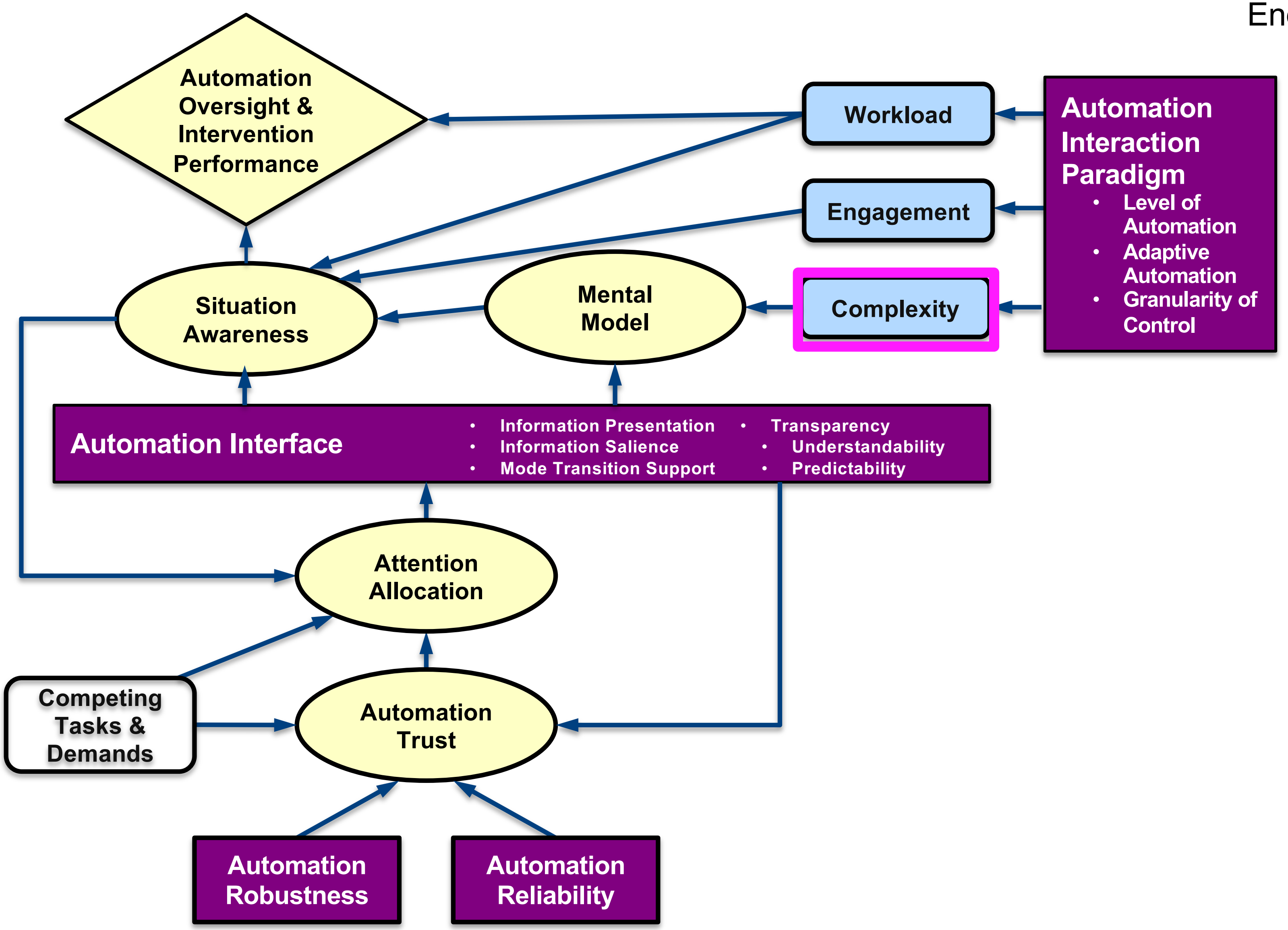
How much confidence do I have in the system?

- Generically
- Situationally
  - Is it working?
  - Is it getting good data?
  - Is it within its programmed envelope?
  - Will its actions meet my intended goals?

Calibrated Trust is Dynamic and Situational

Human Autonomous System Oversight (HASO) Model

Endsley, 2017

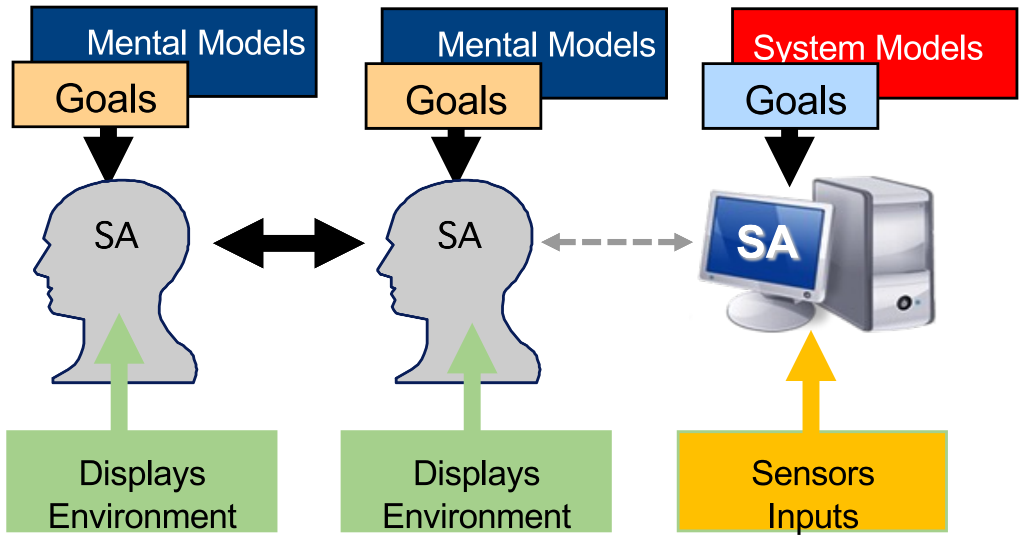


The Complexity Problem

SA Technologies

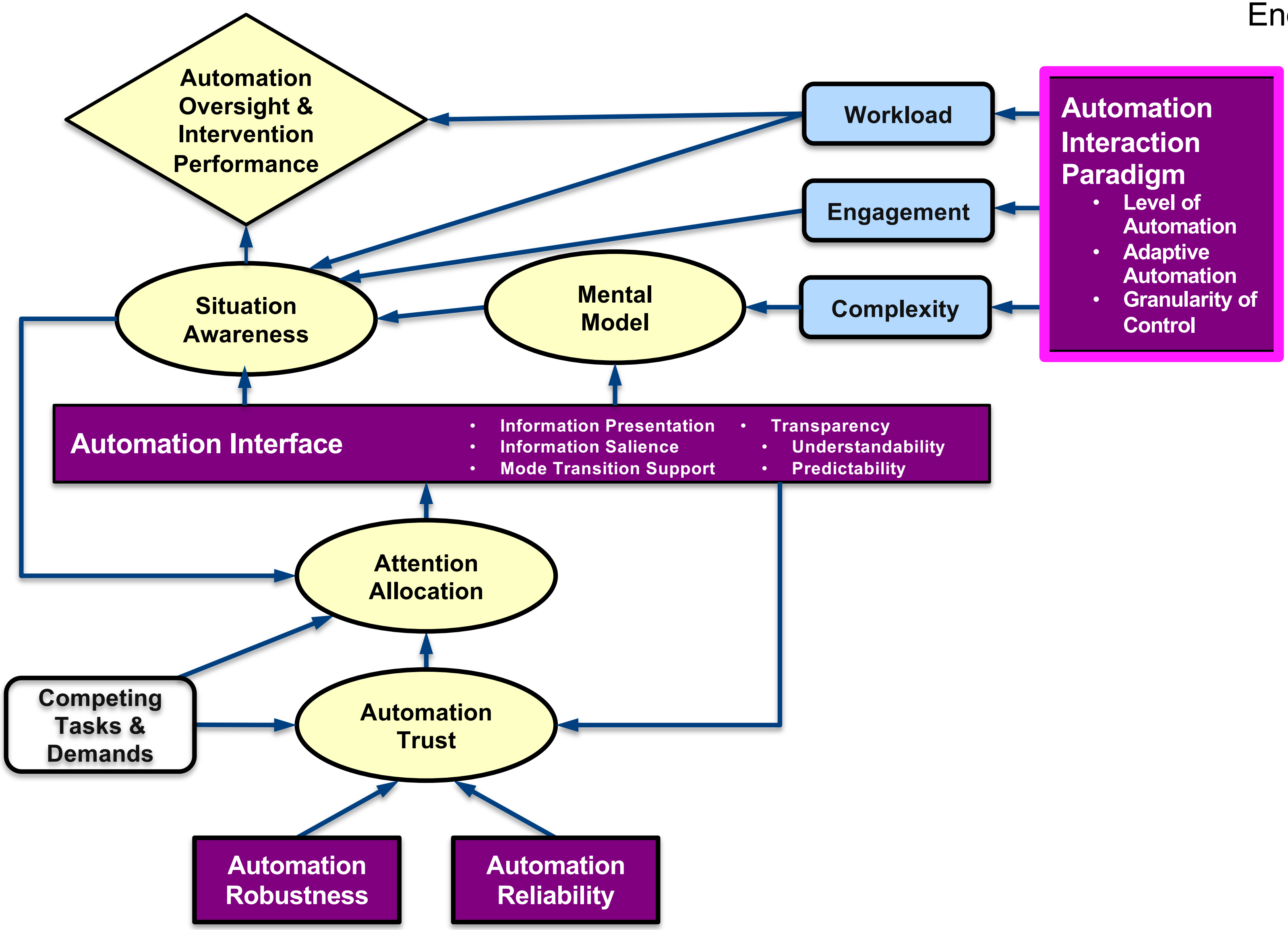
- **Mental models critical for understanding and comprehension**
  - **As system complexity increases it is more difficult to develop a good mental**
- **AI system models will be very different than human mental models**
  - **SA of AI system likely to be different that persons**
- **People are very poor at developing good mental models of how AI works**
  - **Opaque systems**
  - **With AI, system models can change frequently & mental models will be out of date**

Mental models of AI Likely to be poor Contributing to SA problems



Human Autonomous System Oversight (HASO) Model

Endsley, 2017



The Automation Conundrum

SATechnologies

Endsley, 2017

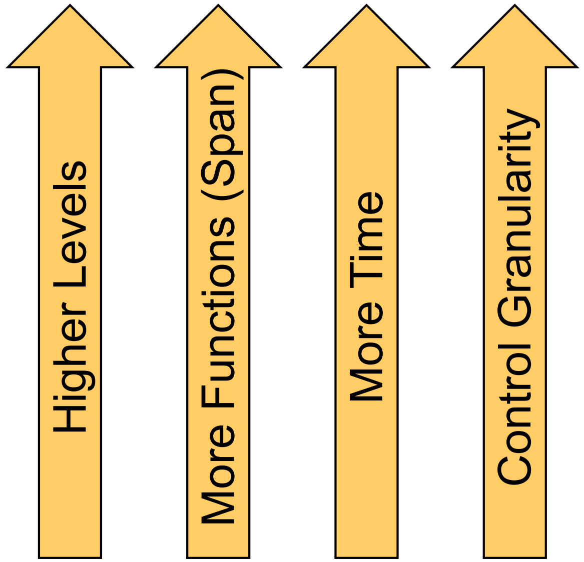
The more automation is added to a system,

and the more reliable and robust that automation,

the less likely that human operators overseeing the automation will be aware of critical information

and able to take over manual control when needed.

More Automation



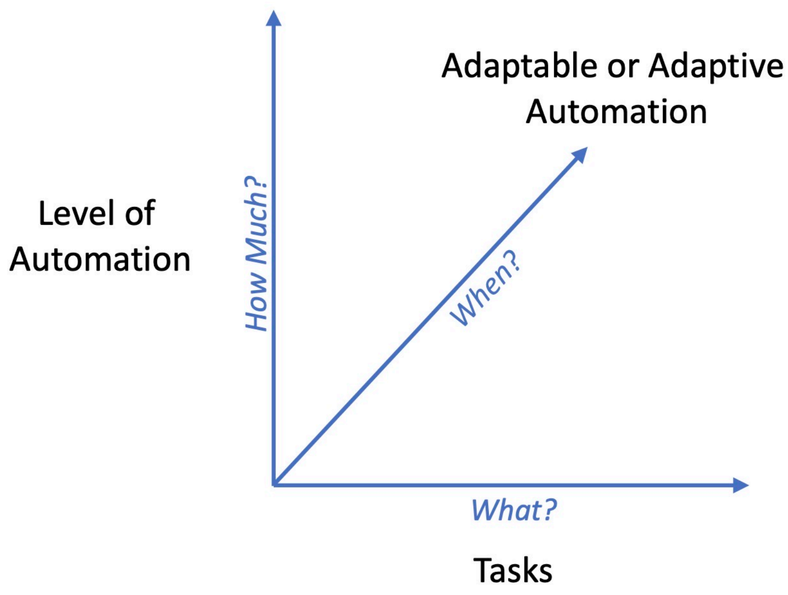
Attention Allocation

Engagement

Level of Automation Taxonomies

Effect of Automation on Human Performance Varies Based on What Aspect of the Task is Being Automated

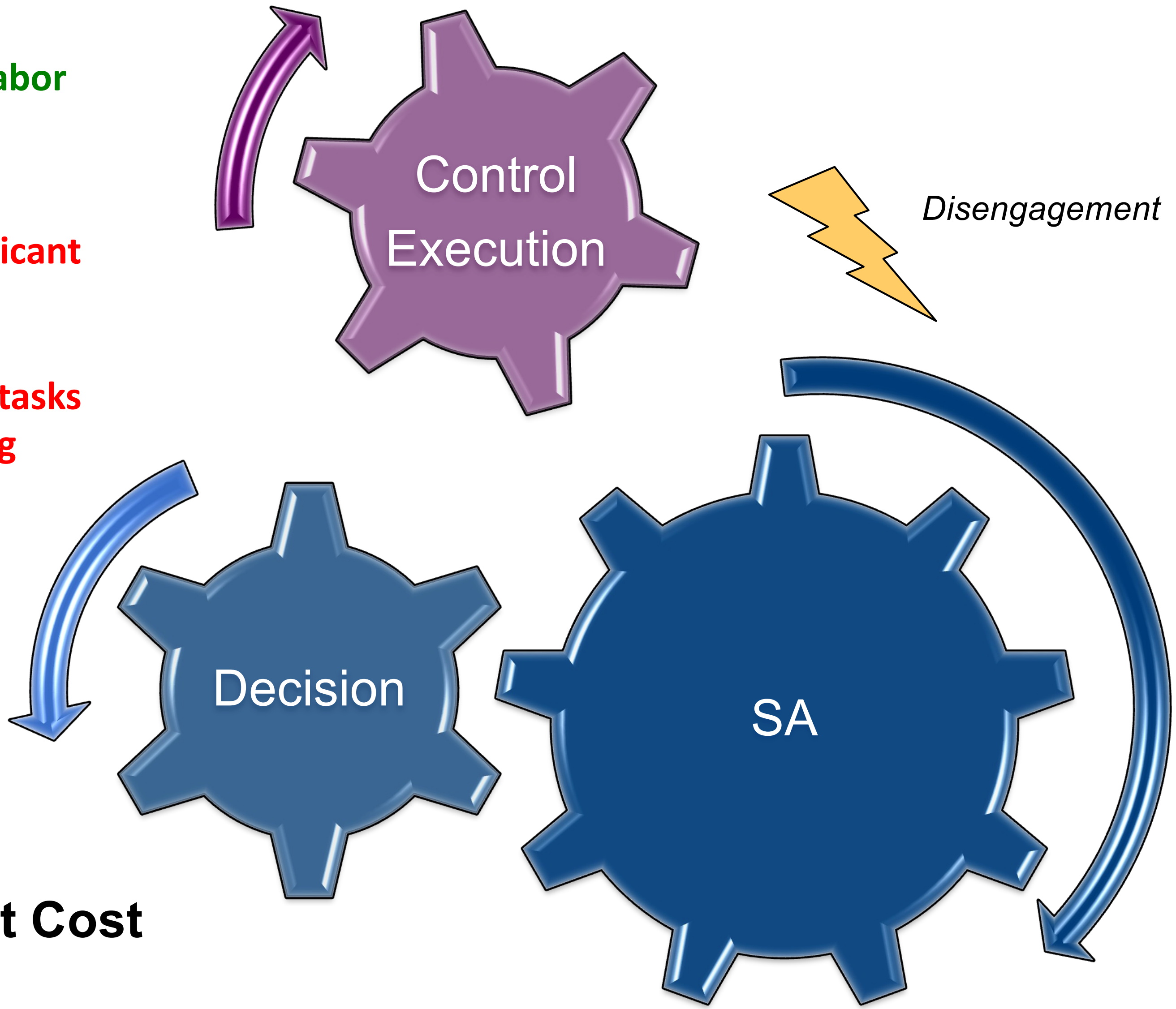
Onnasch, et al., 2014
Endsley, 2017



Processing Stage	Situation Awareness		Decision Making		Implementation
Kaber and Endsley (1997)	Monitoring & Information Presentation		Option Generation	Action Selection	Implementation
Parasuraman, et al. (2000)	Information Filtering	Information Integration	Action Selection		Action Implementation

Automation of Task Execution

- Ok for routine, repetitive manual labor (no intervention needed)
- Lower SA and significant OOTL problems for automation of continuous control tasks & advanced queuing
- Increases in cognitive workload when interventions needed



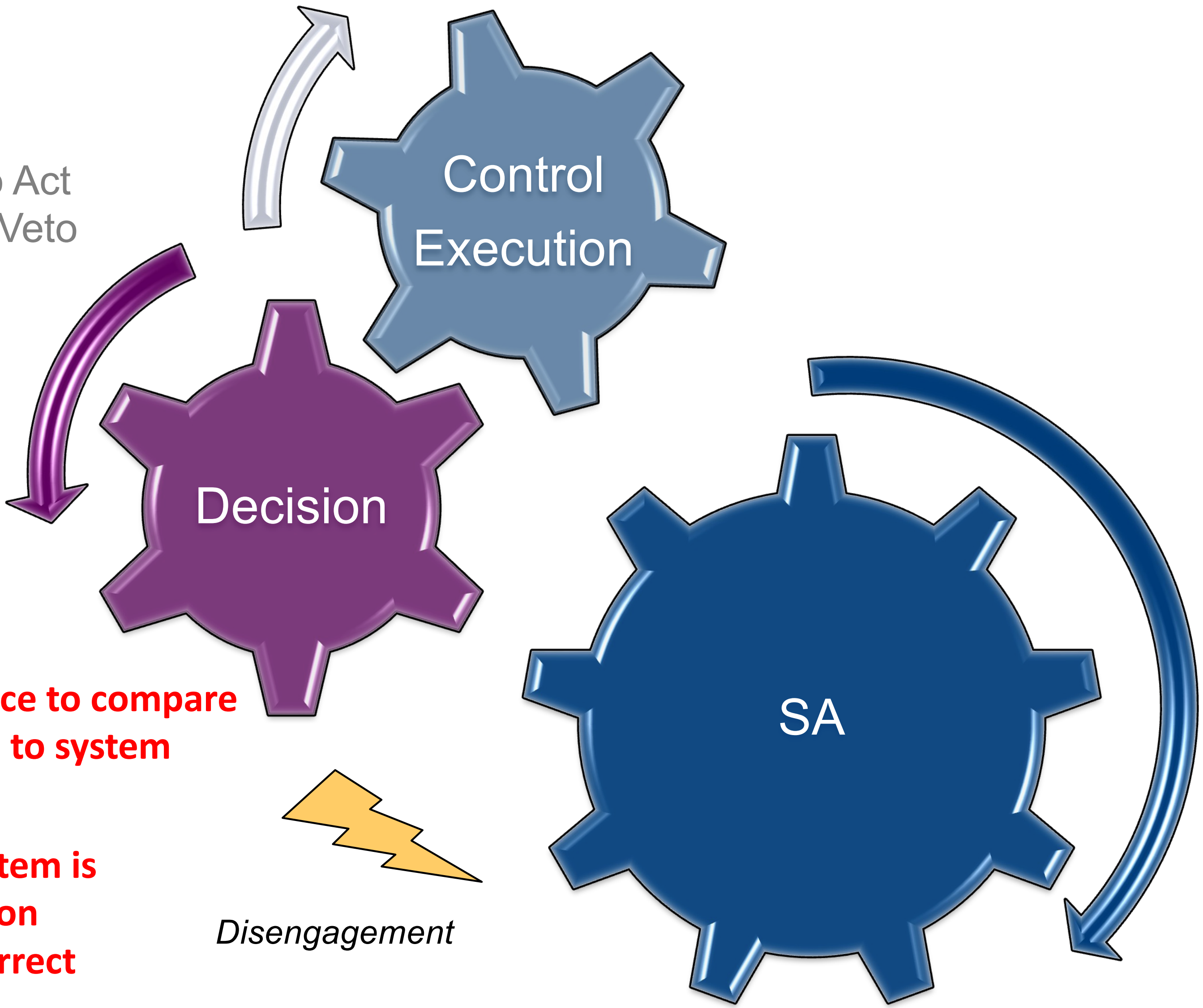
Re-engagement Cost

Automation of Decision Making

Option Generation
Option Selection
- Approval to Act
- Act unless Veto

Re-engagement Cost

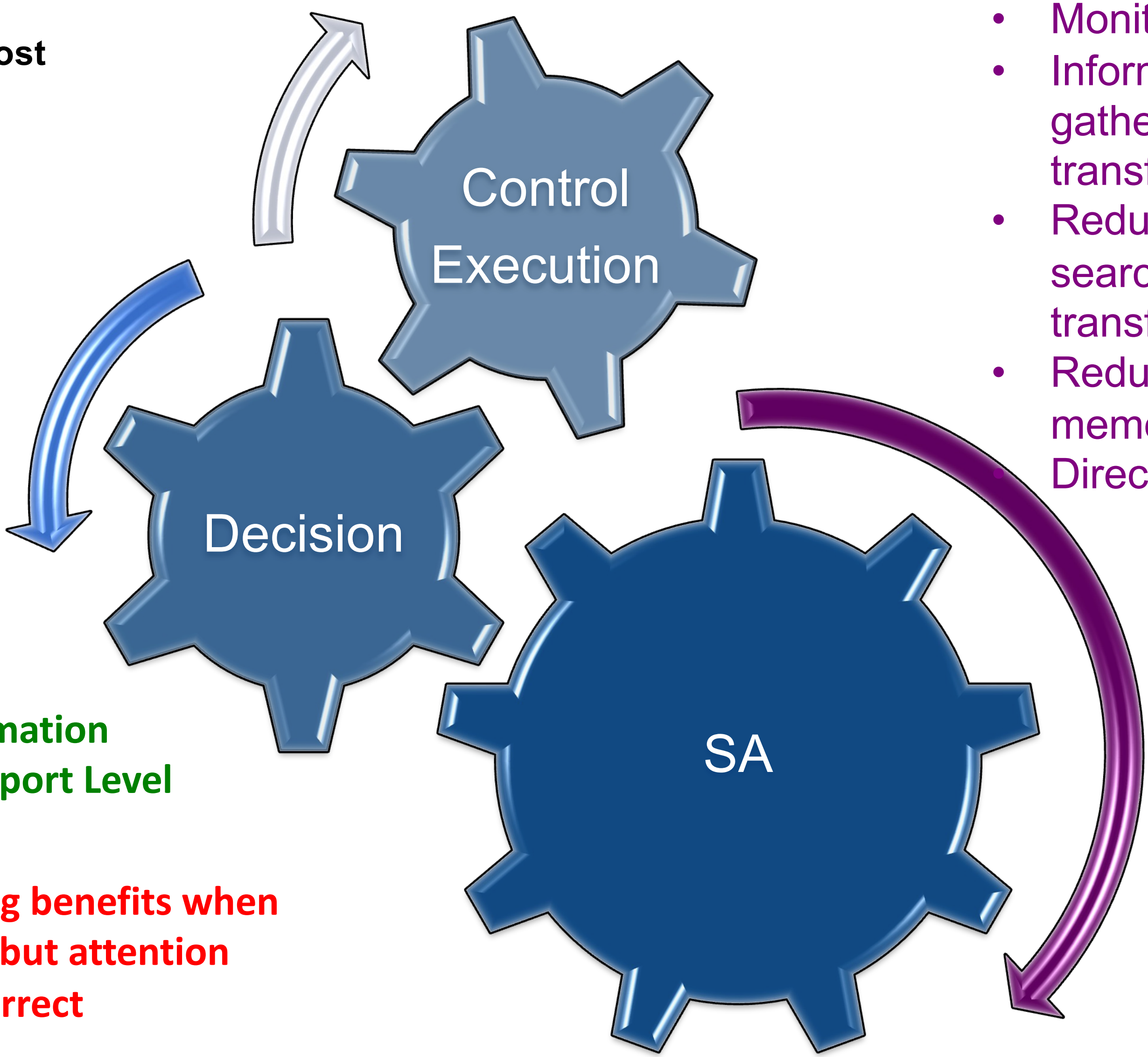
- **Slower performance to compare recommendations to system information**
- **Benefits when system is correct, but decision biasing when incorrect**



Automation to Support SA

SA Technologies

No Disengagement
No Re-engagement Cost



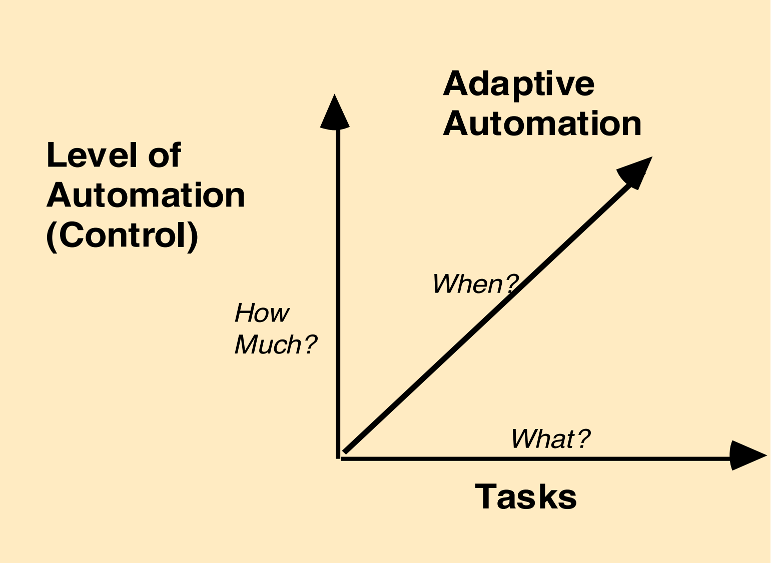
- Monitoring support
- Information gathering and transformation
- Reduce unnecessary searching, sorting, transformations
- Reduce working memory demands
- Direct attention

- **Benefits for Information Integration to support Level 2/3 SA**
- **Information cueing benefits when system is correct, but attention biasing when incorrect**

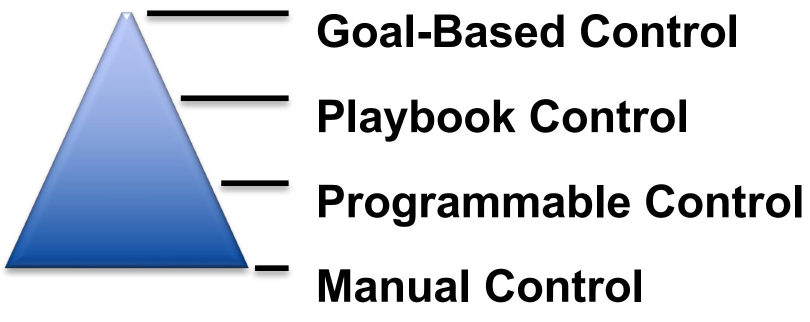
Human – AI Interaction Affects Engagement, Workload and SA

SATechnologies

- **Level of Automation**
  - Worse for Automation of :
    - Action – carrying out tasks, continuous control, advanced queueing
    - Decision Making – creates bias towards system recommendations
  - Not a problem for automation that is focused on improving SA



- **Adaptive & Adaptable Automation**
  - Primarily affects workload
  - SA decreases with more time in auto
- **Granularity of Control**
  - Reduces workload & SA?



Timely Intervention

SA Technologies

- **Automation Inertia**
  - **Tendency to stay in automated mode, not recognizing that over-ride is called for**
  - **From 1 decision step to 2**
- **Manual Performance –> Event Response**

$$RT = t_{detect} + t_{decide} + t_{execute}$$

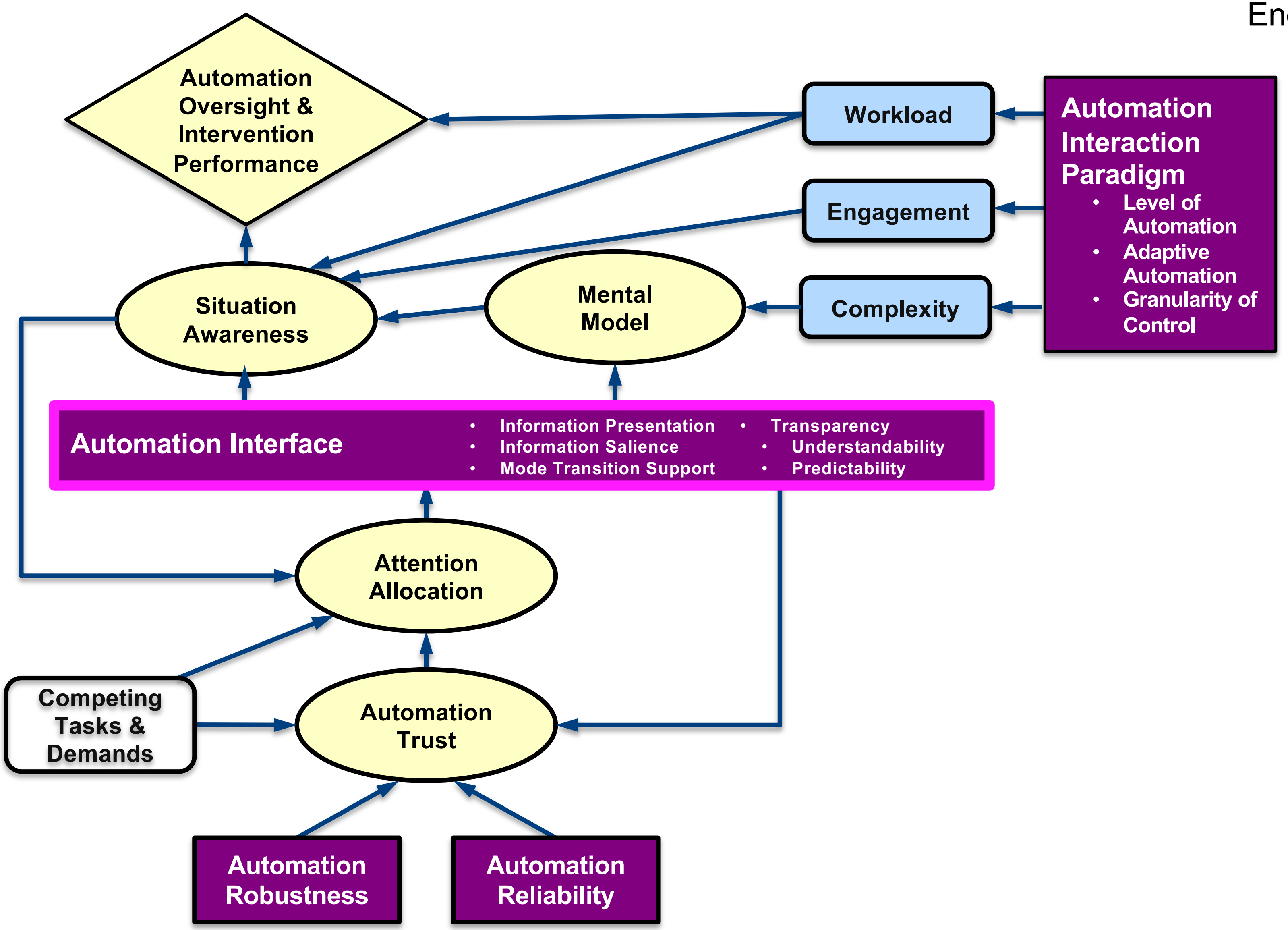
- **Automated Performance -> Event Response**

$$RT = t_{detect} + t_{decide} + t_{over\text{-}ride} + t_{execute}$$

Is the system handling it?
Do I need to do something?

Human Autonomous System Oversight (HASO) Model

Endsley, 2017



Explainability vs Transparency

- **With AI Systems Mental models may be poor**
    - Learning approaches opaque
    - Changes over time
- **Approaches**
    - **Training to improve mental models**
        - Frequent training on how system works, capabilities, changes
    - **Explainable AI**
        - Often backwards looking
        - Focused on why (mental model)
        - May be done in low workload periods, pre-mission, post-mission
        - Hard to do in time demanding scenarios
    - **Real-time display transparency**
        - Real-time support integrated with operator displays
        - What it is doing and will do (SA)
        - Make obvious so don't need to rely on mental models

Transparent AI & Autonomous System Interfaces

SA Technologies

- **Transparency Benefits**
  - Effective in reducing negative effects of OOTL
  - Improved performance, SA & trust
- **Transparency Goals**
  - Understandability
  - Predictability
  - System Reliability
    - How well it is functioning
      - Sensors/Data, Algorithms
    - Level of confidence in fused data
    - Level of confidence in system assessments
  - System Robustness
    - Ability to handle current and upcoming situations

What it is doing?

What will it do?

How well is it working?

What does it know about the situation?

Do I need to intervene?

Transparency is a key mechanism for supporting SA & Team SA in Human AI Teams

Value of Automation Transparency

- **Significantly reduces out of the loop performance problems**
  - Meta-analysis of 15 studies (Wickens, Helton, Hollands, Banbury, 2022)
- **Improves oversight of automation and performance**
  - 10 studies
- **Improves SA and performance**
  - Meta-analysis of 17 studies
  - (Van de Merwe, Mallam and Nazir, 2022)
- **Improves calibration of trust**
  - 10 studies

AI Transparency

Understandability & Predictability of the System

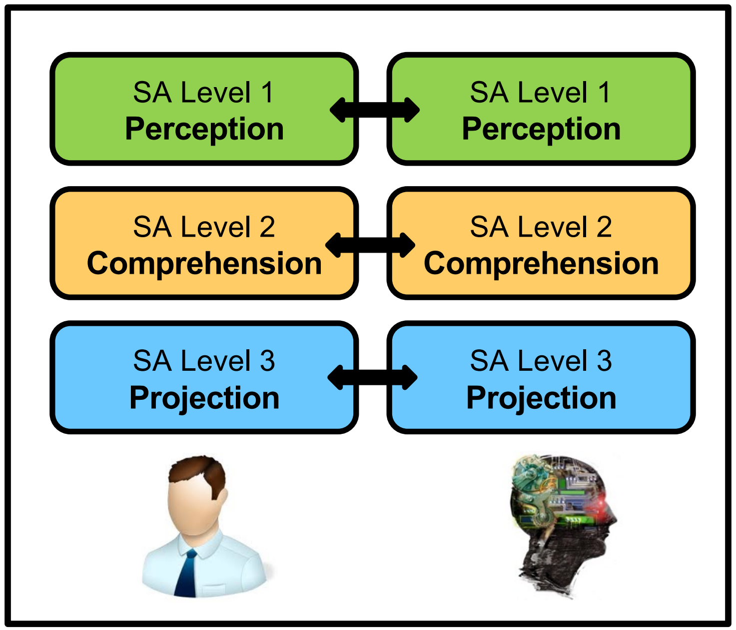
Taskwork SA	SA of the World and the System	Level 1 SA – Basic Data Level 2 SA - Understanding Level 3 SA - Projection	**Task Performance**
Agent SA	SA of the AI	State/Mode SA Sufficiency Ability to Perform Tasks Task Status & Performance Projected Actions Impact of Actions on Others	**Oversight, Intervention & Backup**
Teamwork SA	SA of the Team	Goals Functions Decisions, Strategies, Plans	**Team Coordination & Cooperation**

Transparency of AI Situation Model

AI representation of the state of the world

- It's interpretations
- It's projected actions

Shared Goals



- ***Does the AI detect the same information that I do?***
- ***How does it interpret the information it has?***
- ***What projections is it making?***
- ***How confident is it?***

Transparency of State of Automation (AI)

- State/Mode
- SA Sufficiency
- Task Status & Performance
- Ability to Perform Tasks
- Projected Actions
- Impact of Actions

Explainable AI

SATechnologies

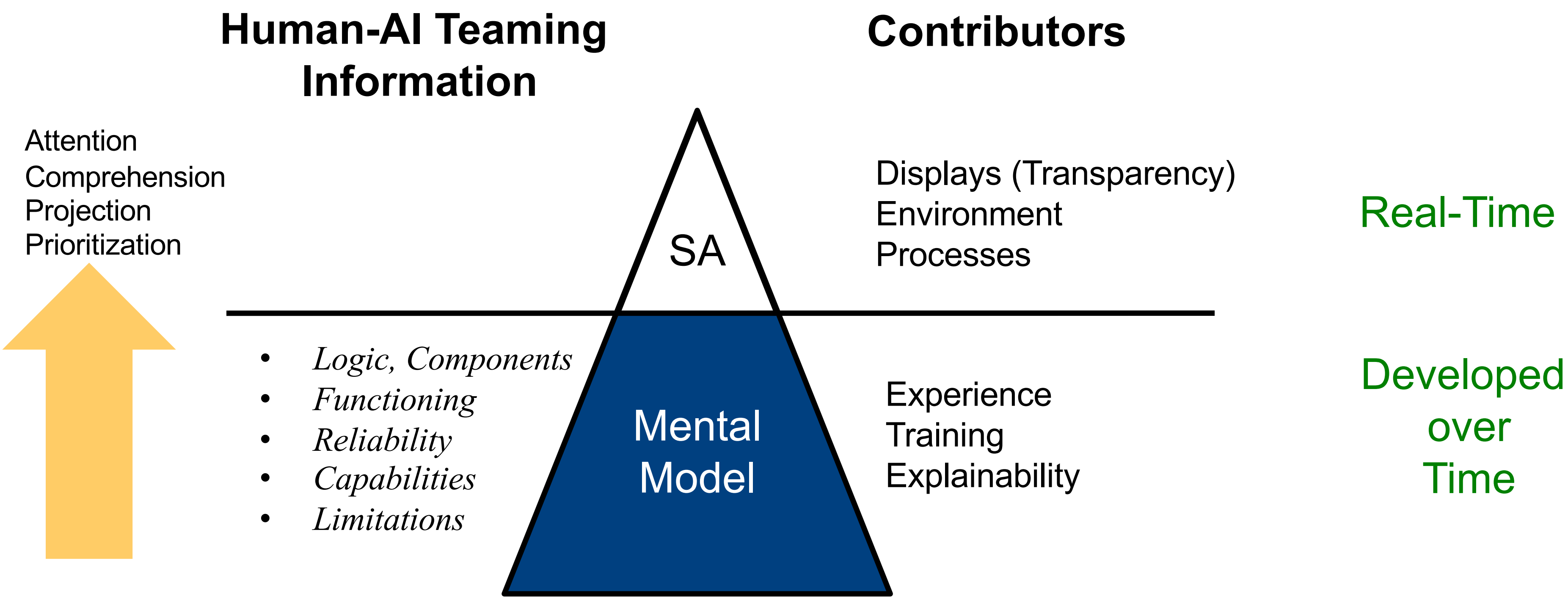
- Rule-based systems
  - Why feature: List of rules executed
- AI Systems
  - Opaque
  - No Rules
  - Explainability Approaches
    - Generate a rule set from the neural nets to convey logic (Huang & Endsley, 1997)
    - Aspects of images being used by AI to make a classification (Goebel et al, 2018)
      - People still need to create a mental model of what it is doing over time
- What people need from Explanations (Miller, 2019)
  - Answers for why it did something
  - Causative and Contextual
  - Contrast – Why A and not B?
  - Operationally relevant descriptions of how it will perform in different circumstances
    - Need both cues used (L1 SA) as well as explanation

Can slow decision making

Transparency even more important with AI-based systems

SA Technologies

SA is fed by both real time information and by mental models of system

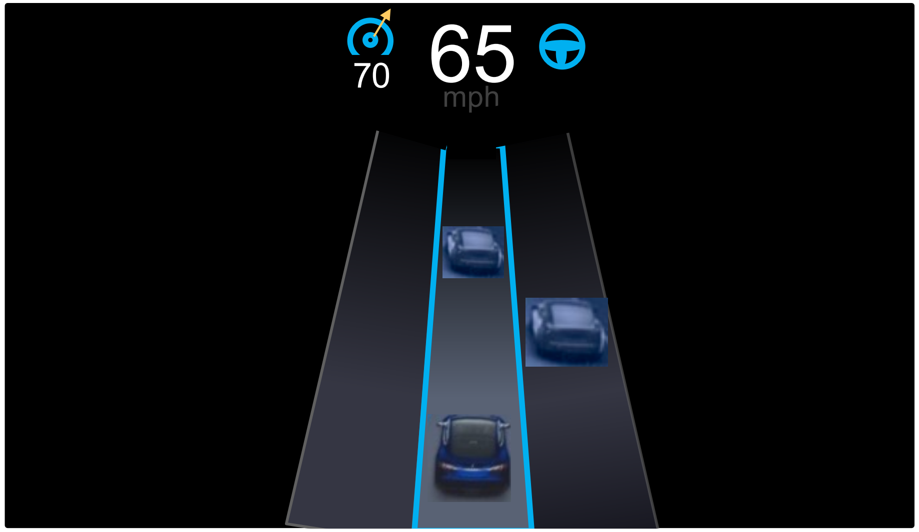


AI makes it more difficult to develop and maintain an accurate mental model

Example of Transparency
737-Max8

AOA Disagree

Example: Tesla Autopilot



Mental model of AI
Limited – Monthly uploads of S/W
No training or instructions

Current Challenges for SA

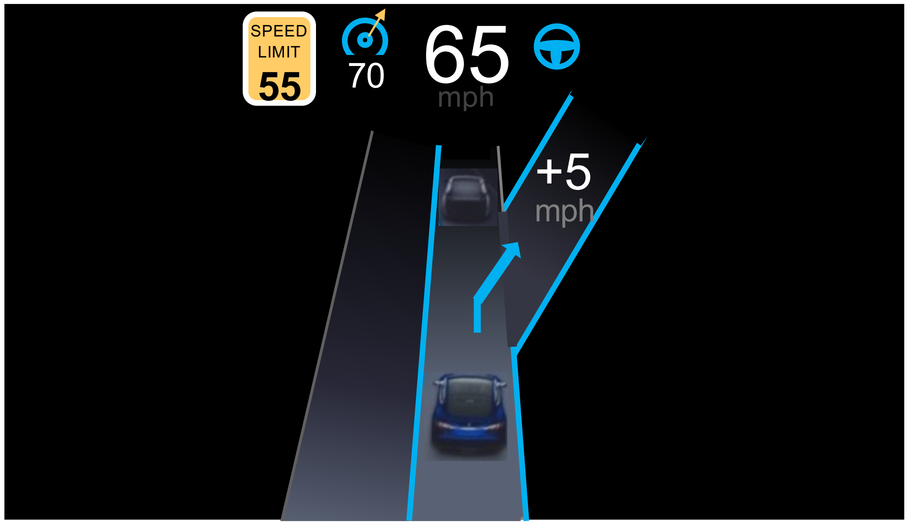
- Missing some key Level 1 information
  - Blind spot
  - Vehicle jitter
- Level 2 and 3 Info must be mentally derived
  - Distance to other vehicles, lane deviations, projected deviations in speed or lane keeping
- Mode changes not salient
- Emergent behaviors
  - Unexpected speed surges
- Future behaviors & capabilities unknown
  - Ability to perform in upcoming conditions
  - Future trajectories
- No information is provided on system confidence levels

Example of Automation Transparency: Tesla Autopilot

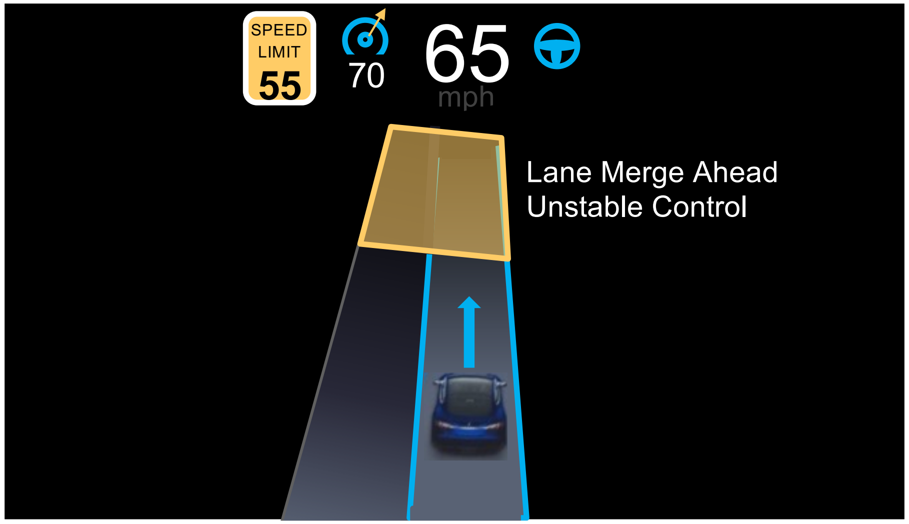
Add System Knowledge & Actions

Salient mode changes

Add Projection of Actions



Add Capabilities in Context



Achieving AI Transparency
- Design the System to Support SA

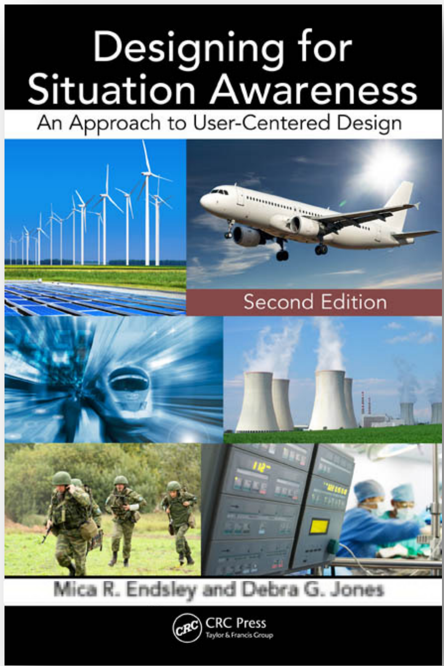
SA Requirements Analysis → SA-Oriented Design → SA Measurement

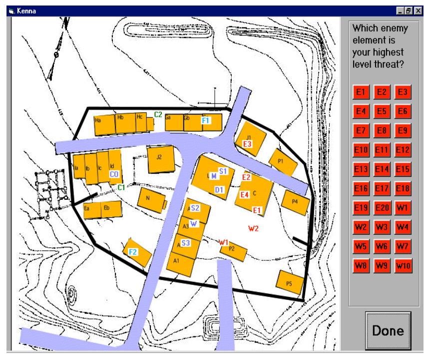
Goals

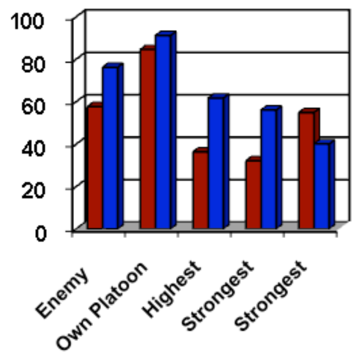
Decisions

- Projection Requirements
  - Comprehension Requirements
    - Data Requirements

50 Design Principles







Need A Clear Roadmap Of What Information People Really Need

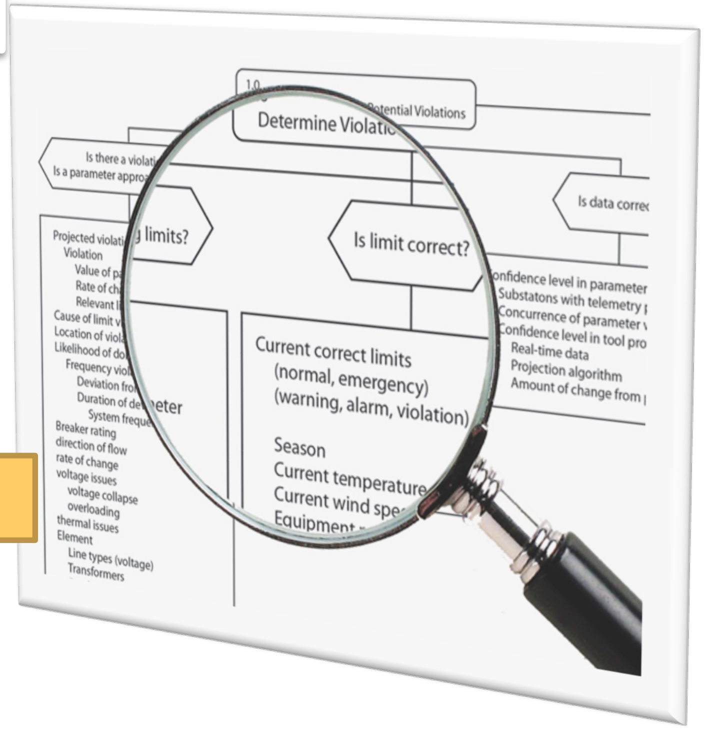
SA Technologies

Goal Directed Task Analysis

- **Goals**
  - **Subgoals**
    - **Decisions**
      - **Projection Requirements**
      - **Comprehension Requirements**
      - **Perception Requirements**

What is Meaningful?

Provides Detailed Analysis of What People Really Need to Know for Decision Making



Many SA Design Principles Directly Address Transparency

SA Technologies

General Principles	Complexity Principles	Automation Principles
Organize information around goals	Minimize logic branches	Use automation for assistance in carrying out routine actions rather than higher level cognitive tasks
Present Level 2 information directly—support comprehension	Map system functions to the goals and mental models of users	Provide SA support rather than decisions
Provide assistance for Level 3 SA projections	Provide system transparency and observability	Keep the operator in control and in the loop
Make critical cues for schema activation salient	Group information based on Level 2/3 SA requirements and goals	Avoid the proliferation of automation modes
Support transmission of different comprehensions and projections across teams	Reduce display density, but don't sacrifice coherence	Make modes and system states salient
		Enforce automation consistency
Uncertainty Principles	**Alarm Management Principles**	Avoid advanced queuing of tasks
Support sensor reliability assessment	Don't make people rely on alarms—provide projection support	Avoid the use of information cueing
Explicitly identify missing information	Support alarm confirmation activities	Use methods of decision support that create human/system symbiosis
Support assessment of confidence in composite data	Support the rapid development of global SA of systems in an alarm state	Provide automation transparency

Human-AI Teaming with High Levels of Collaboration

SATechnologies

Dynamic Function Allocation

- **Goal Alignment**
  - Desired goal state actions need to support
  - Requires active goal switching based on prioritization
- **Function Allocation/Re-allocation**
  - Assignment of functions and tasks across team
  - Dynamic reassignment based on capabilities, status
- **Decision Communication**
  - Selection of strategies, plans and actions needed to bring world into alignment with goals
- **Task Alignment**
  - Coordination of inter-related tasks for effective overall operations

Shared Situation Awareness

Increased emphasis on the importance of creating effective team SA and shared SA within the human-AI team

- **Support collaboration (including anticipation and back-up)**
- **Ensure goal alignment**
- **Share status on functional assignments and task progress**

Teamwork skills

Conclusions

SA Technologies

- **AI is being developed for a wide variety of applications**
- **High levels of SA of the state of the system, environment and automation will be needed**
- **Need to develop robust, reliable and transparent autonomy**
  - **Requires careful consideration of information that needs to be made transparent**
  - **Design of displays to support transparency needs**
  - **Evaluation of effectiveness of displays to provide needed SA**
- **Effective integration of human-AI team will be critical to successful implementation**
  - **Shared SA to provide effective Human-AI Teaming**

Endsley, M. R. (2023). Supporting human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*, 107574.