Use of STPA for Analyzing Information Flows in Distributed Autonomous Systems (Can AIs Effectively Interpret Intent?)

Tom McDermott (Stevens Institute of Technology)

Dr. Dennis Folds (Lowell Enterprises)

AI4SE & SE4AI Research and Application Workshop

September 17, 2024

Concepts Explored

- Human-Machine Teaming
  - Intent, Rules, and **Transfers of Authority** (RECITAL)
- Visual Concepts
  - Progressive Disclosure
- Social Theories
  - Construal Level Theory (CLT)
- Human Systems Integration Methods
- STAMP/STPA Methods
- Modeling Human Machine Transfers

Vignette:

Steve is CEO of a growing services company that is learning to use data and artificial intelligence to **improve customer service**. Steve decides he **needs to hire an Executive** to manage corporate data collection and analytics processes to improve competitiveness. Steve **directs his Human Resources Director to find candidates** and hire this person within the next **2 months**. Steve asks the HR Director to assemble a search team and bring him the **top 3** candidates for his review and **selection within 30 days**.

purpose

action

order

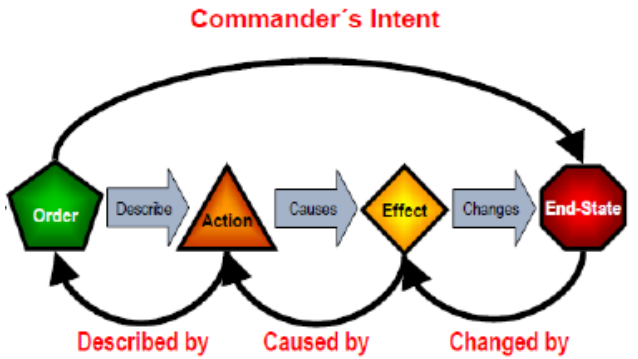
expressive: urgency

expressive: my decision

end state

The VP-Engineering and HR Director **proceed with the search process**. Based on the level of hire and the urgency they decide to use an executive search firm known to the HR Director for both its candidate networks and its speed. They provide the search firm a draft position description and a list of selection criteria they would like to emphasize.

effect



R Rules Of Engagement (RE)

E

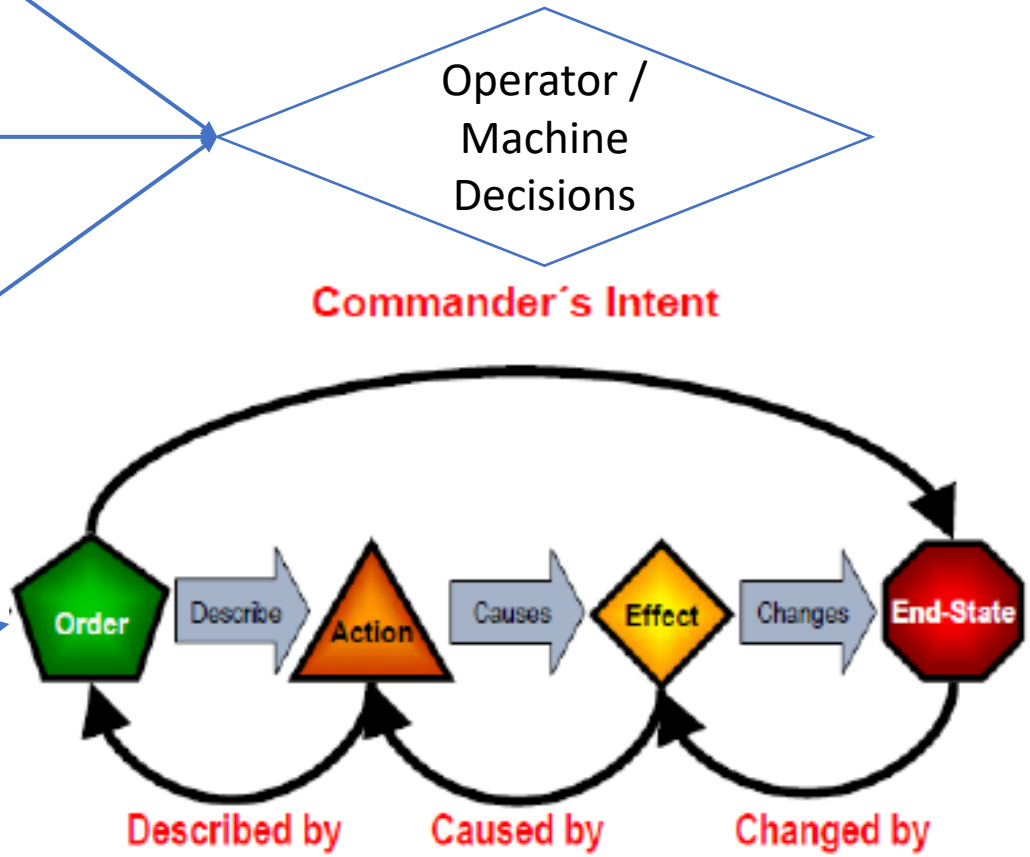
C Commander’s Intent (CI)

I

T Transfer of Authority (TA)

A

L Language (L)



We can define a semantic construct for CI where:

CI -> (Expanded Purpose) (Key Tasks) [End State] (Expressives)

Gustavsson, Per & Hieb, Michael & Moore, Philip & Eriksson, Patric & Niklasson, Lars. (2008). Machine Interpretable Representation of Commander's Intent. 13th International Command and Control Research and Technology Symposium (ICCRTS).

Transfer of Authority

- Transfer of authority to control may require agent (human/machine) to agent communication about:
  - Intent — Direct but high level and subjective
  - Rules — Usually located elsewhere and indirectly accessible
  - Operational plans — Need to adjust in near-real time
  - Capabilities and limitations — Inherent in the design/training and generally not accessible
- These communications may occur in an environment in which open-ended verbal communications are not afforded
- Transfer may occur between agents operating with differences in Rules or Rules interpretation and might involve assets with atypical capabilities and limitations
- We need is for standard UI components and underlying data structures that are machine readable and that humans can comprehend
- We need rigorous methods to assess safety and security of these transfers

WAZE Human-Machine Teaming

Rules

Expressives

Order

Other Contextual Information

Action

Transfer of Authorities

Effect

Interpretation of Information

Waze image: www.cnet.com/roadshow/news/google-assistant-waze-easier-reporting-less-distraction/

Information model based on Construal Level Theory

Construal Level Theory (CLT) -- a general cognitive theory that describes the extent to which a person's thinking about an object or event is abstract vs. concrete, as a function of psychological distance (Temporal and/or spatial distance, Relevance to tasks or other interests, and Probability of occurrence)

CLT	Construal	Measurement Model	User potential Information needs at this level	WAZE Example
CLT 1	The main Claim	This key outcome was/will be achieved as shown by these key indicators	provide further definition of entities in the plan; provide indicators of success/failure of the plan	I completed my trip by 3pm
CLT 2	The main reason	CLT1+because of these key causal effects	+ provide backstories for key entities, spatial and temporal aspects of the mission; drill down to additional details for critical selected mission aspects related to CI & RoE	Because I followed these routes and traffic behaved
CLT 3	The justification	CLT2+because in the full causal model these paths are of greatest importance (magnitude)	+ provide advisories and alerts related to changing context of mission as related to tasks at hand	Waze alerted a traffic incident, and I selected the alternate route
CLT 4	The basis of the justification	CLT3+because here is the full measurement model	+ provide links to alternative planning and tasking if determined by context	Other driver reports confirmed my decision
CLT 5	A full summary of the data	CLT4+because here is the time-step history of all the measurements	as supplied situationally to one of the levels above	The underlying Waze model
CLT 6	All the data	CLT5+and here are all the anomalies and alternatives considered	as supplied situationally to one of the levels above	All of the traffic, route, and user data

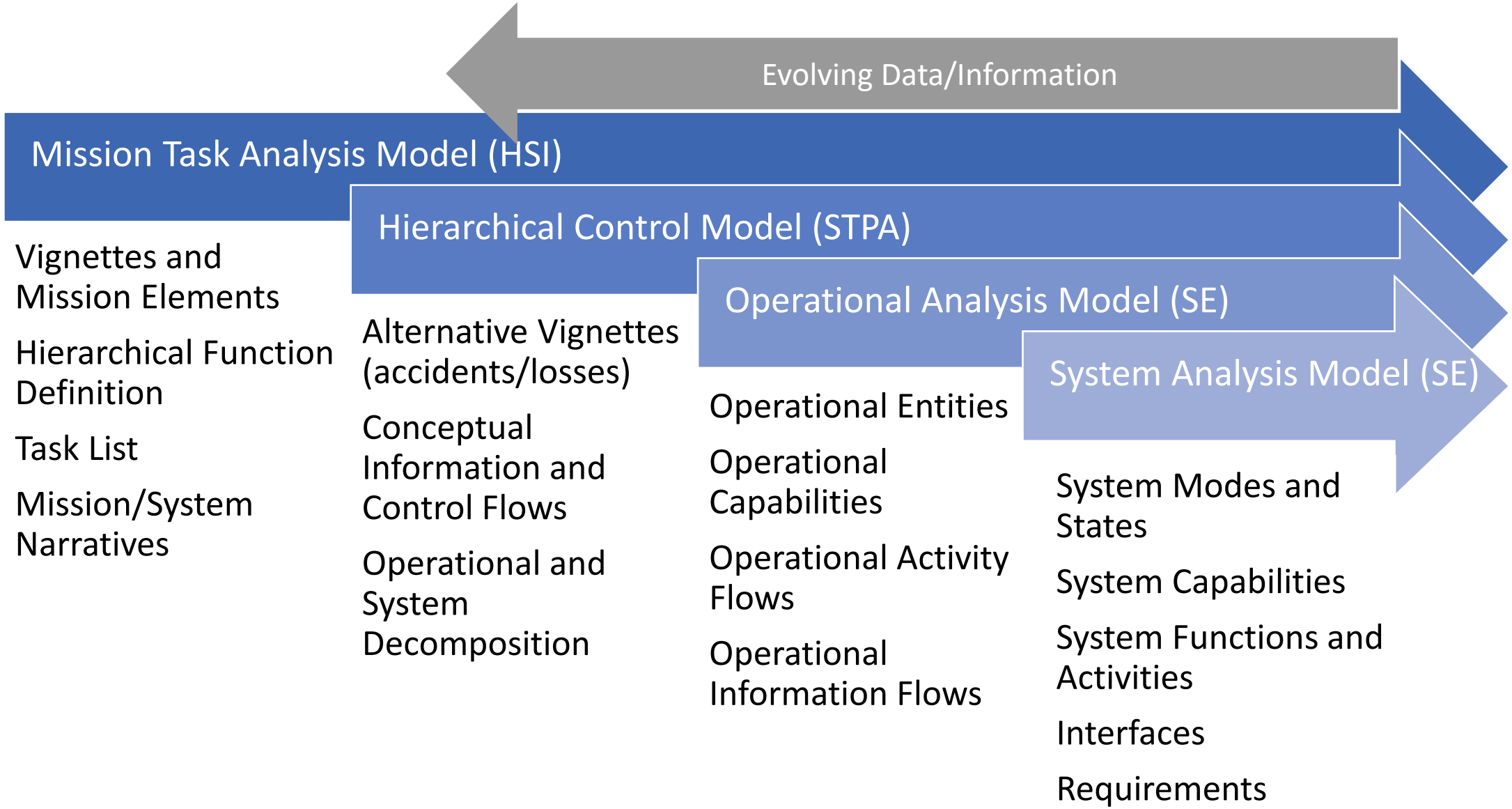
Why Model?

- Most accidents/mission failures will be caused by errors in interpretation of information by either the human or the machine
- Leading to errors transfer of control or authority made in the planning process
- Underlying concept of human informational transfer has subjectivity
  - **Intent**
  - **Rules**
  - **Authorities**
  - **Other Contextual Information**

- Consistently used in hierarchical control structures
- Lack of multi-disciplinary research

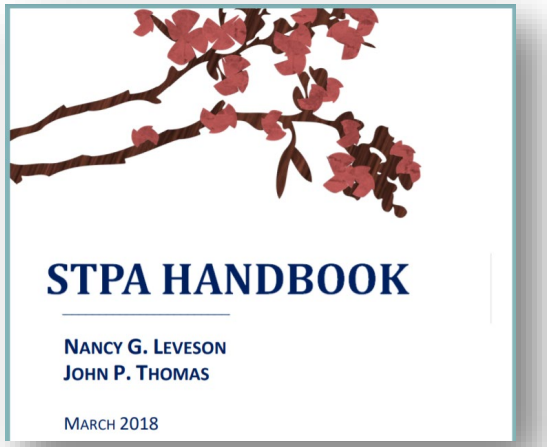
- Desire a Systems Engineering approach to address both information design and control mechanization across layers of hierarchy
- STAMP/STPA (System Theoretic Process Assessment) supports human reasoning about control flows and information flows in design

Example overall modeling flow

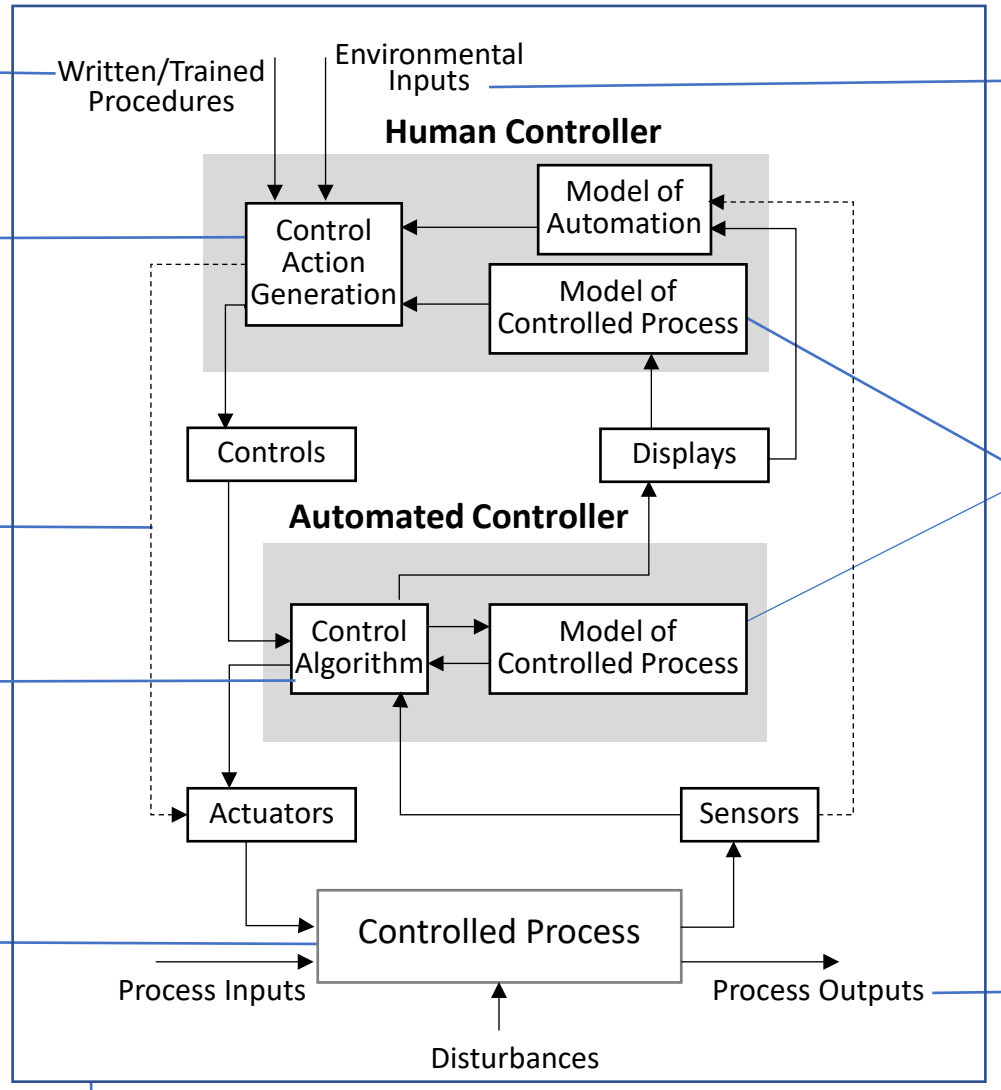


STPA Process Flow

1. **Defining the System**: Identifying components – both the human and machine agents in the system and their assets.
2. **Modeling the Control Structure**: Mapping how these entities interact and make decisions.
3. **Identifying Unsafe Control Actions (UCAs)**: Considering risks like missing information flow unauthorized transfer of authority.
4. **Deriving Safety/Security Requirements**: Implementing behaviors and controls; continuous monitoring of development activities.
5. **Continuous Monitoring**: Regular reviews and adapting strategies as the human-machine system progresses.



RECITAL Controller Model



Rules

Actions

Authority

Actions

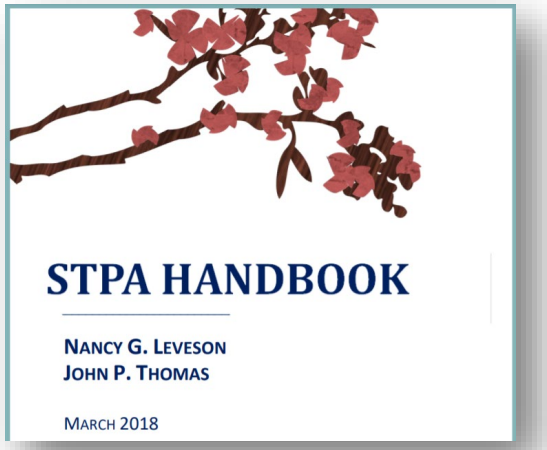
Next level Actions

INTENT: Order, Purpose, Expressives

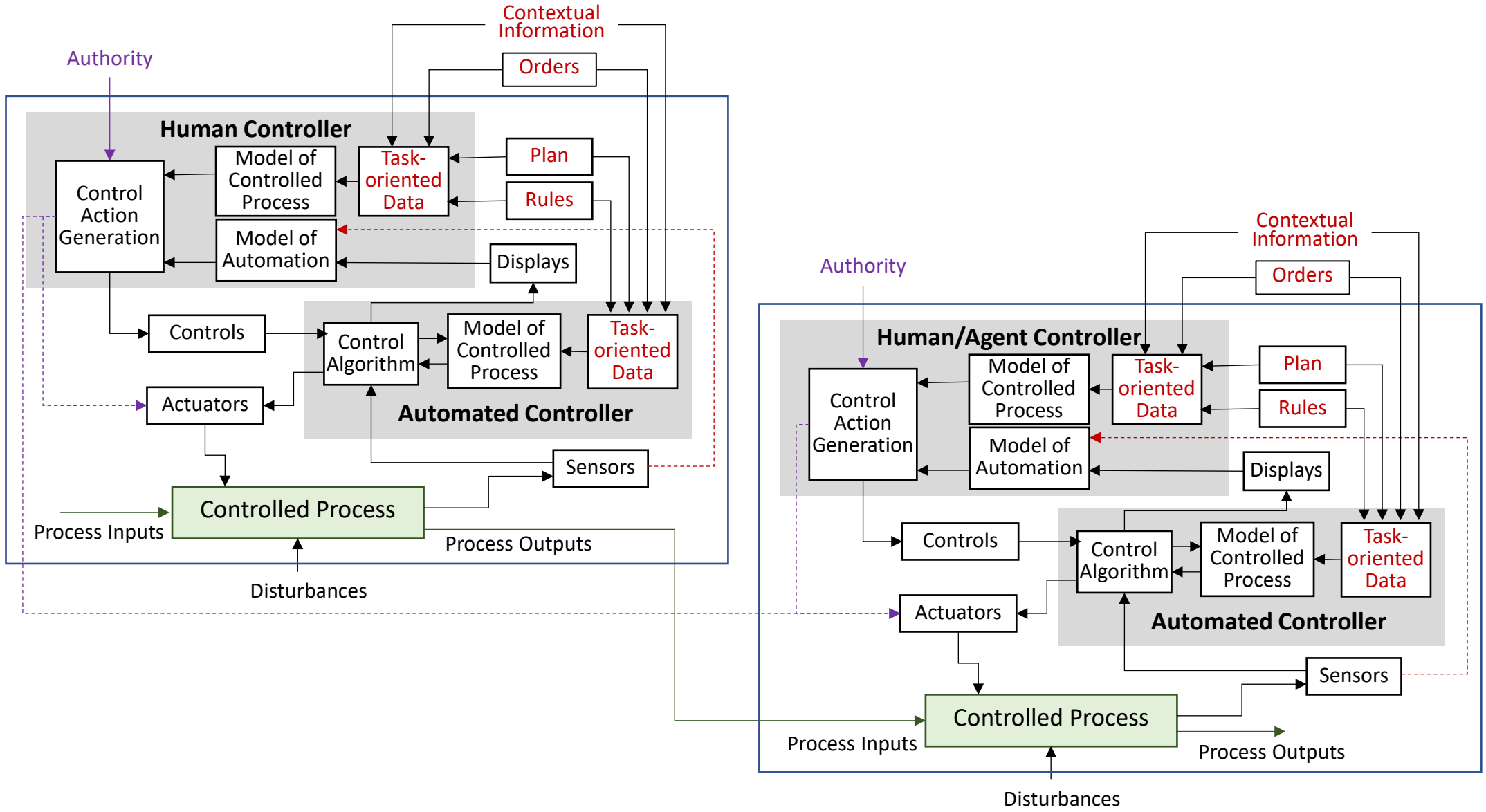
Interpretation of Information

Effect

Other Contextual Information



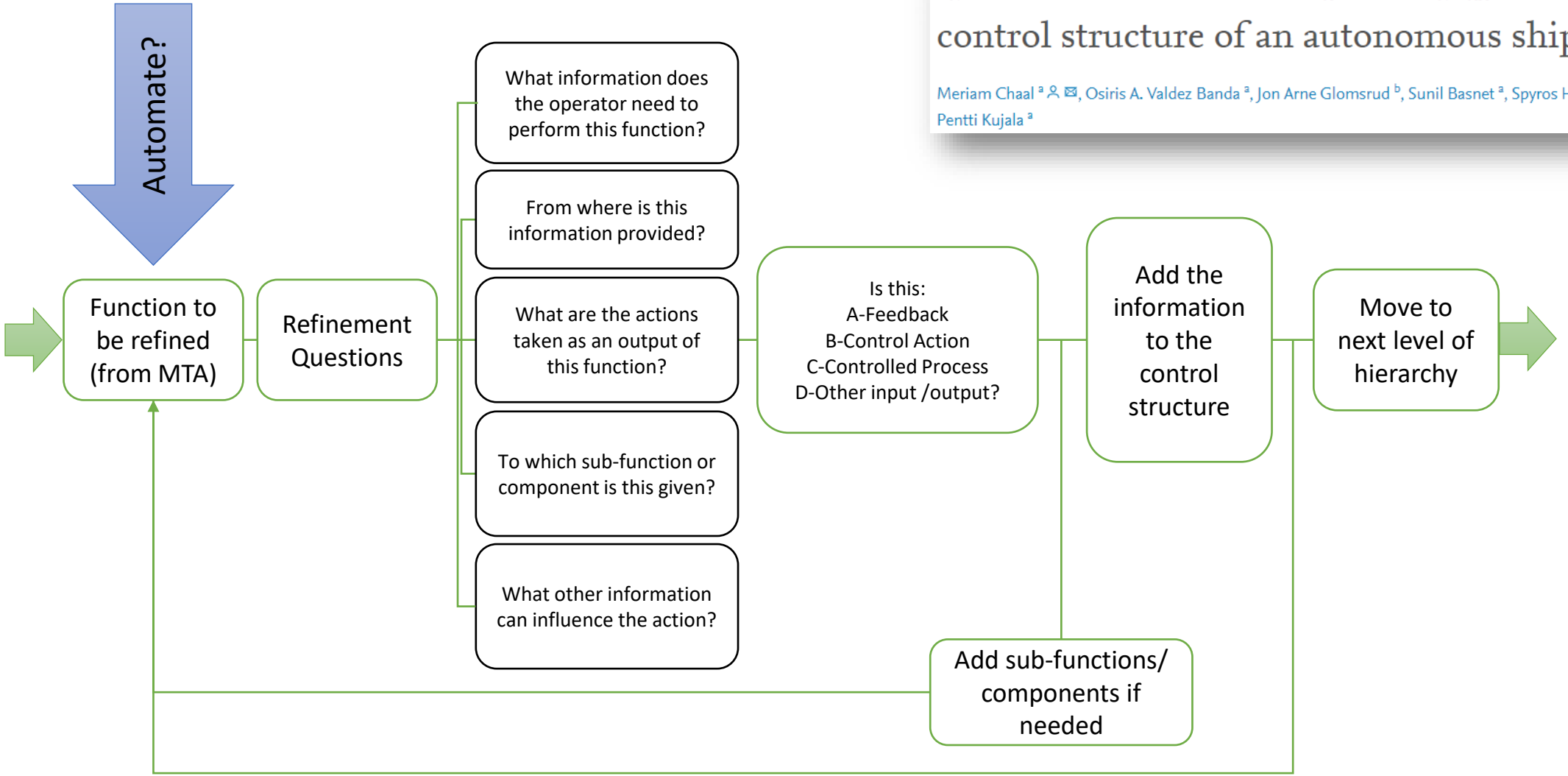
STPA-based Information Model



Decomposition Process

A framework to model the STPA hierarchical control structure of an autonomous ship

Meriam Chaal [a], Osiris A. Valdez Banda [a], Jon Arne Glomsrud [b], Sunil Basnet [a], Spyros Hirdaris [a], Pentti Kujala [a]



Example Functional Analysis for an automated driver

Function	What does the operator need to perform this function?	From where is this information provided?	What are the actions taken as an output to this function?	To which other function is it given?	What other information can influence the action?	*UCAs/Hazards*
8.X Navigate	Navigation data; location, speed, etc. Over the air data from Waze and other drivers	GPS; Vehicle navigation system	Navigate to next turn, or modify route	6. Driver Control/ Modify Vehicle Tactics	Driver initiated change of plans	• *Loss of or corruption of GPS data* • *Loss of vehicle* • *Driver injury*
	Vehicle navigation file w/ local area & possible routes	loaded navigation file	Update Navigation File	6. Driver Control/ Modify Vehicle Tactics	Driver initiated change of plans	(need defaults)
8.X Vehicle Selected Contextual Auto-routing	Contextual information that would instigate a change in navigation: weather, traffic changes, construction; as well as vehicle knowledge of its current statuses	Other drivers are monitoring observed changing conditions; Waze is dynamically adjusting potential routing choices (new scenarios)	The vehicle makes a decision to change its navigation based on selection of new routing alternatives as determined by the input data	8. Navigate; 6. Driver Control/ Modify Vehicle Tactics	Alternate instructions from the driver; Concern about an unauthorized source?	• *Above, plus* • *incorrect or missing contextual data* • *Intentionally misleading data feeds*

Questions?