Metareasoning for Comprehensive Troubleshooting

Center for Integrated Cognition

James Kirk

james.kirk@cic.iqmri.org

45th Soar Workshop

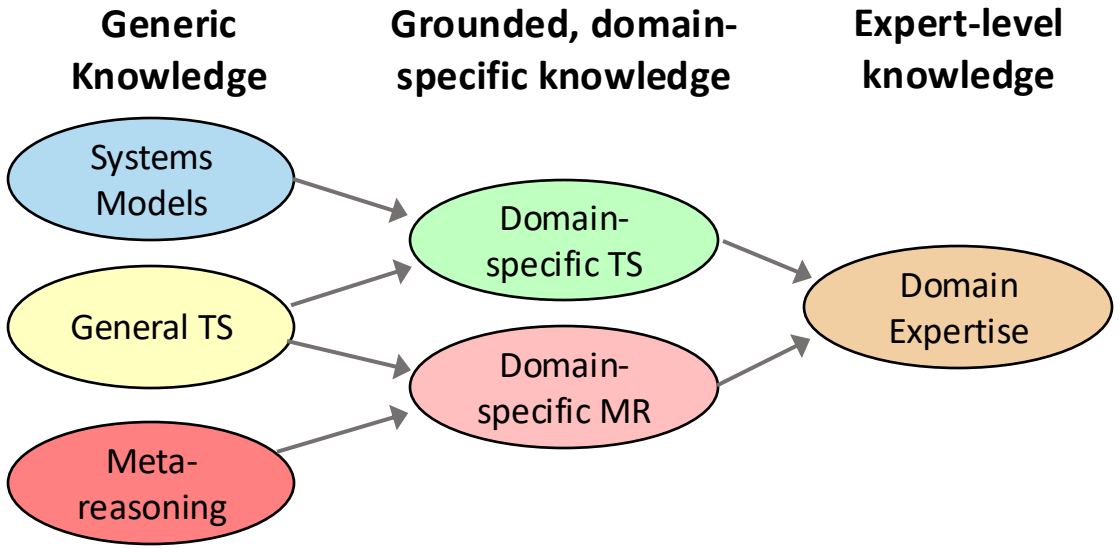
May 5, 2025

Comprehensive Troubleshooting (CTS)

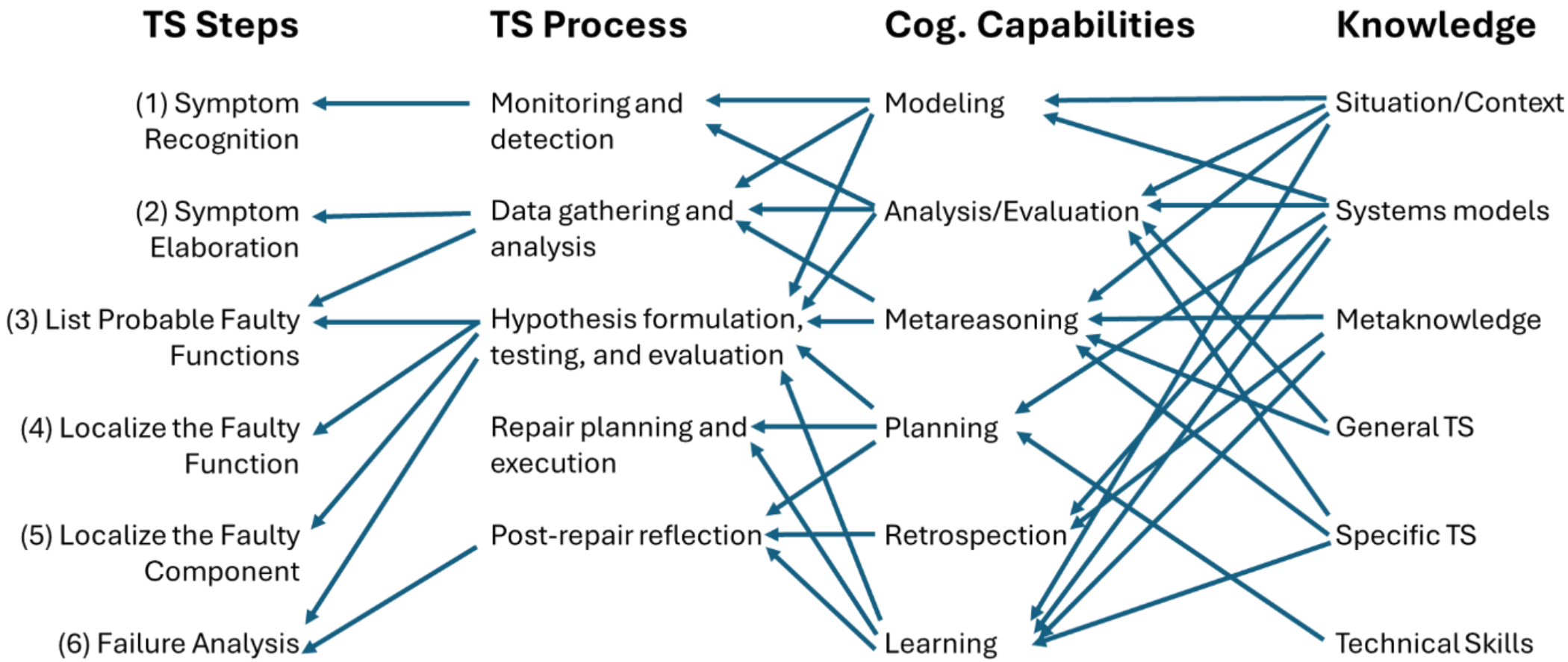
- CTS goes beyond routine troubleshooting, replicating expert human-level skills
  - In “real world” troubleshooting complex, novel problems and faults occur
  - Requires wide range of knowledge and cognitive skills, including metareasoning
- SOA limitations
  - Pre-planned responses (checklists, fault analysis trees) cannot address unpredicted complex issues
  - DL/ML strategies effective for limited TS (e.g., fault classification), but not for issues outside training data
  - Cognitive architectures support metareseaoning but limited in explorations to narrow domain/task specific scenarios
- Requirements for CTS
  - Capabilities/knowledge to identify, diagnose, and address novel, complex, unseen issues/faults
  - Need inexpensive/reliable methods for acquiring domain-specific TS knowledge
  - Need metareasoning and control knowledge to guide CTS

Metacognition for Self-reliability

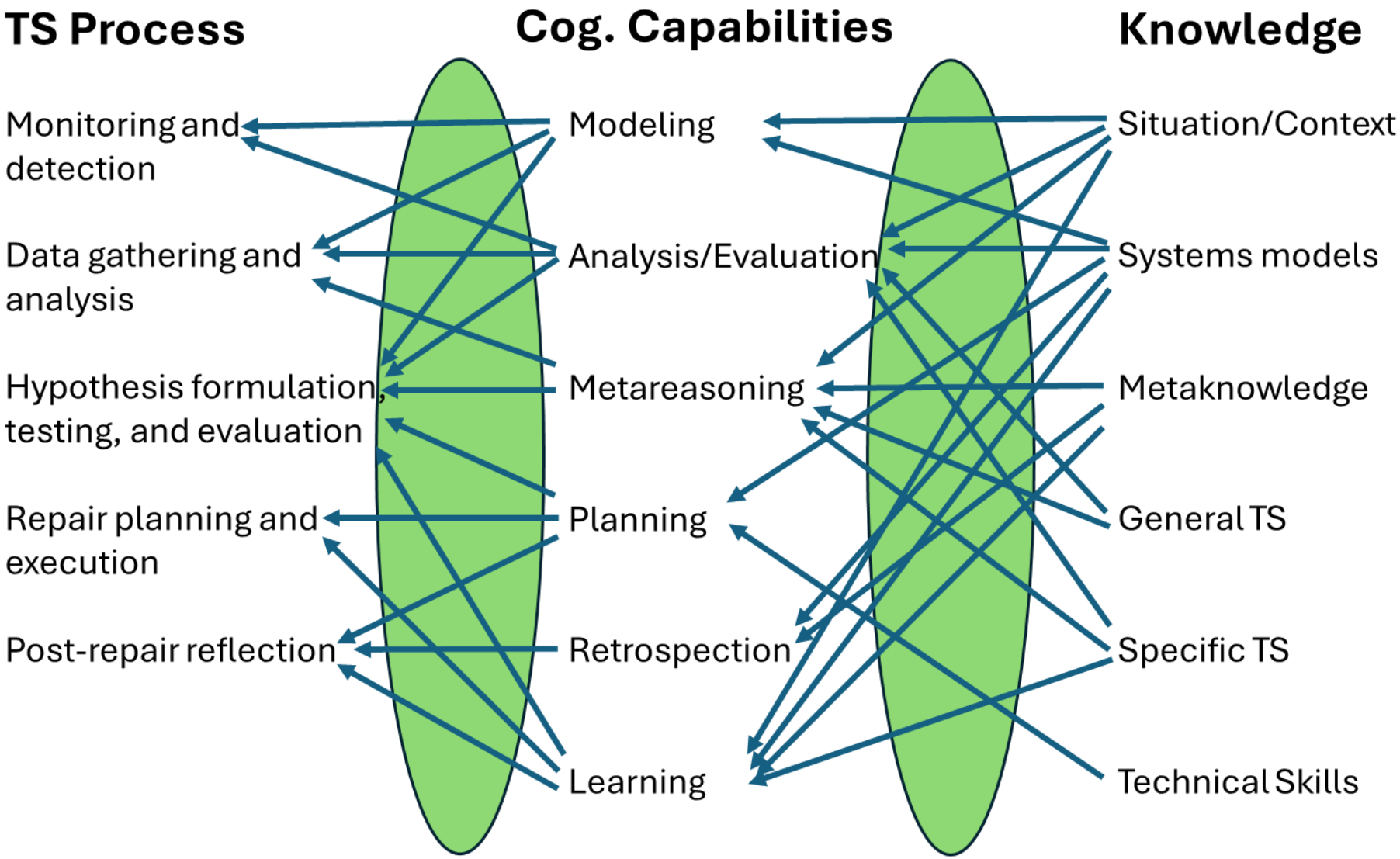
- Comprehensive TS is essential for reliability autonomy
  - Must identify and address unanticipated problems that inevitably occur
- Plan: Leverage LLMs and Cognitive Architectures
  - Build on cognitive capabilities to support TS steps
  - Develop metareasoning for control
  - Use LLMs to provide required knowledge
  - From generic system model, TS, and metareasoning knowledge: learn specialized domain-specific knowledge to achieve expert-level CTS



Knowledge and Processes for CTS



Metareasoning for CTS



Metareasoning Orchestration

Planned Approach

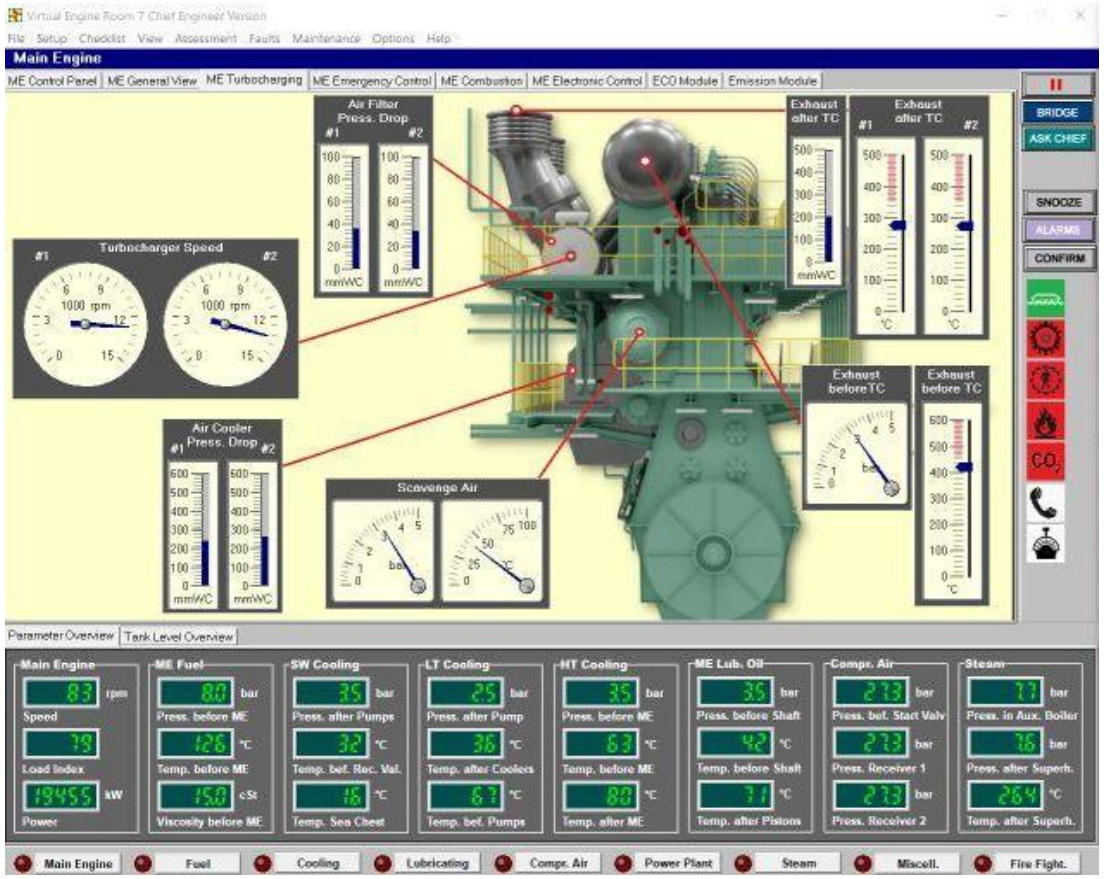
- Innovations
  - Leverage recent LLM advancement as source of domain-specific troubleshooting and metareasoning knowledge
  - Leverage our prior work on accessing and verifying task knowledge from LLMs for cognitive agents and current working on orchestration with LLMs
  - Build on existing CA capabilities, skills orchestrated via metareasoning and LLM
- Objectives
  - Acquire domain-specific troubleshooting knowledge for CTS
  - Acquire domain-specific metareasoning knowledge to orchestrate the usage of cognitive capabilities and access of knowledge during CTS
  - Learn expert-level CTS through the integration of knowledge and capabilities

Project Plan

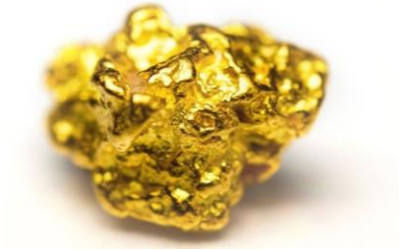
- **Year 1**
  - Develop TS processes in Soar
  - Explore simple CTS problems
  - Integrate with simulator(s)
- **Year 2**
  - Exploit LLMs for specializing metareasoning and TS knowledge
  - Leverage ongoing working using LangGraph for Metareasoning orchestration
  - Explore more complex CTS problems
- **Year 3**
  - Leverage Soar learning capabilities to learn long-term knowledge from specialized metareasoning and TS knowledge
  - Perform experiments on complex CTS problems

Experimentation Plan

- We will evaluate the agents CTS abilities in a simulation environment
- Initial investigations with simple simulation environment, possibly drawn from other work (Factorio?)
- Planning to use a ship-engine simulator, such as VIRTUAL ENGINE ROOM 7 (VER7) →
- Evaluation Metrics
  - **Solution performance:** percentage of problems it successfully diagnoses and solves
  - **Efficiency:** time, agent processing costs, and monetary cost of experiments (tokens)
  - **Transfer:** ability of the agent to use knowledge learned from past experiences to solve new troubleshooting problems more efficiently/directly



Nuggets & Coal



Nuggets

- Grant awarded!
- Synergy with other projects, continuing exploration of LLM usage

Coal

- Haven’t started yet, waiting to receive funding, hopefully in next couple months…

Questions?