How Different Sources of Knowledge Shape Normative Decision-Making

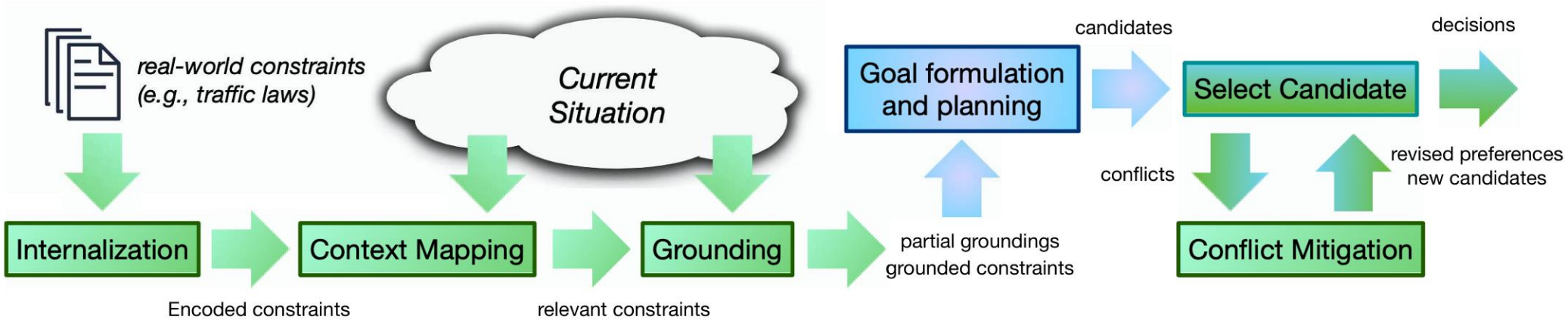
Center for Integrated Cognition

Steven Jones, Robert Wray, John Laird

45th Soar Workshop

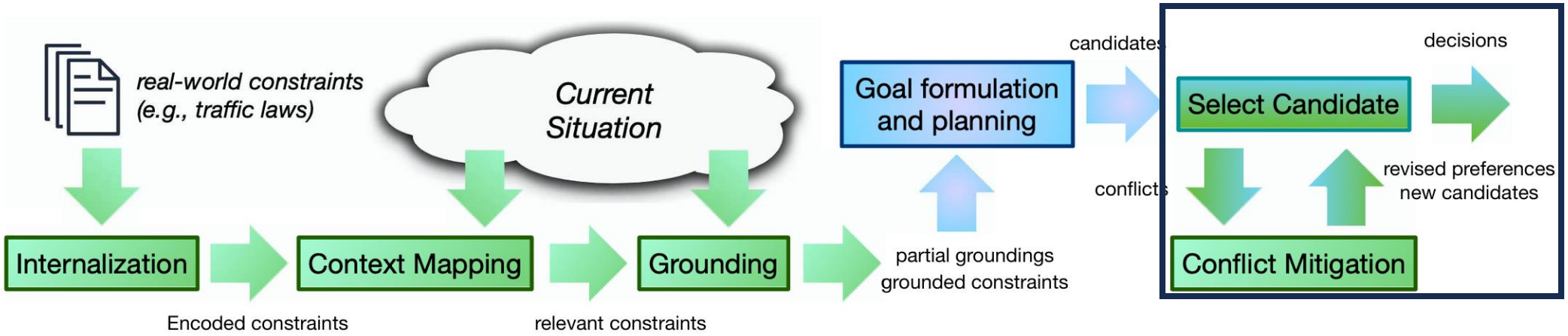
5 May 2025

Constraint Compliance



Problem: “Conflict Mitigation”

- Cannot foresee every combination of constraint instantiations and train everything before.



- May have incomplete constraint knowledge as well.
- Constraints in the world are of different “types”

Examples of humans handling constraints

- Traffic laws
- Choosing among TVs
- Manners
- Following an example

Soar Approach: Bring all available knowledge to bear

Agent should use all available:

- Rules
- Costs
- Examples
- …?

Strawman approach:

- Use *only* rules
- Use *only* costs
- Use *only* virtuous exemplars
- …

Example: Sailor Overboard

- Suppose losing a sailor costs x.
- What about when you can see them?

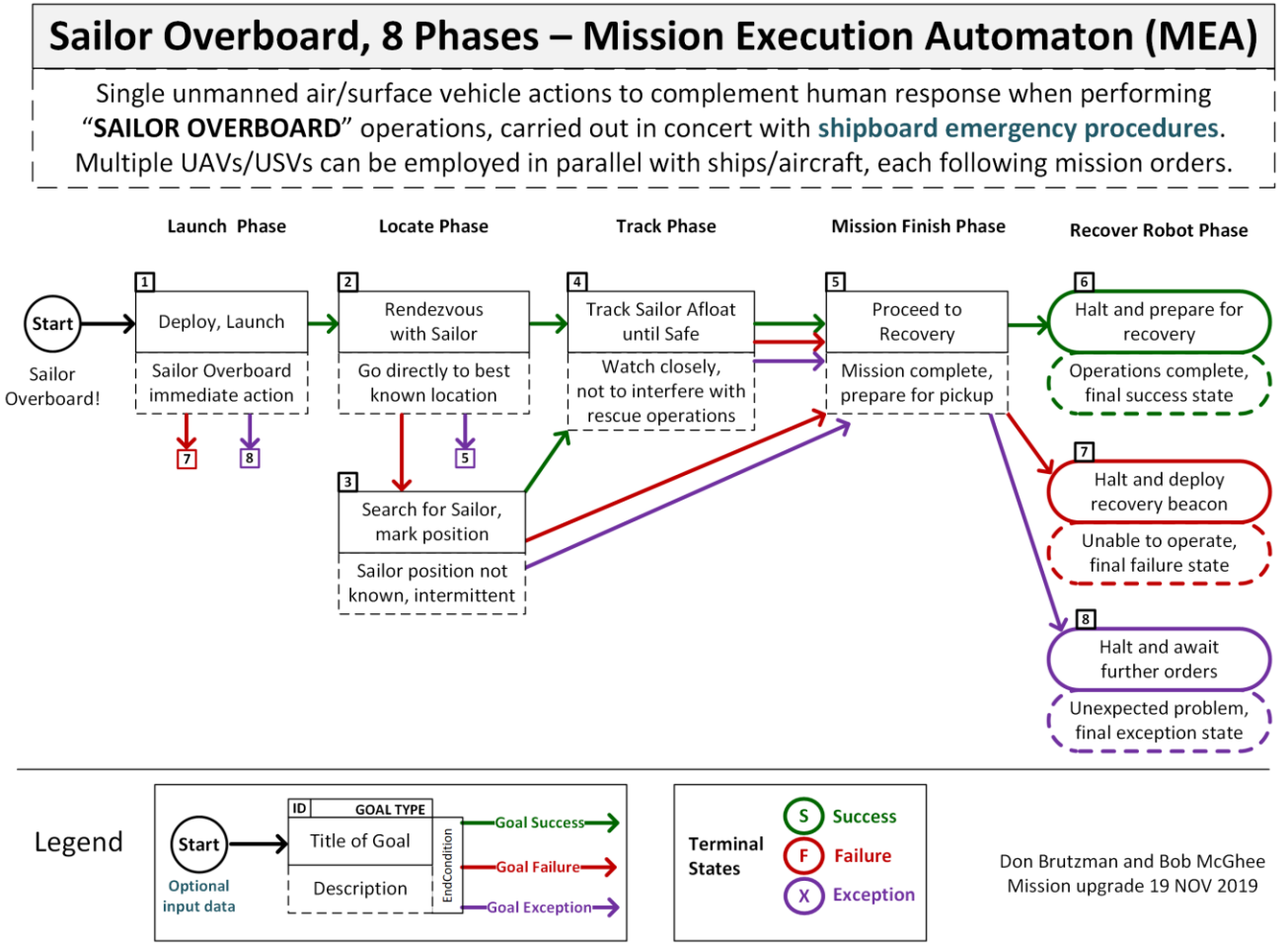
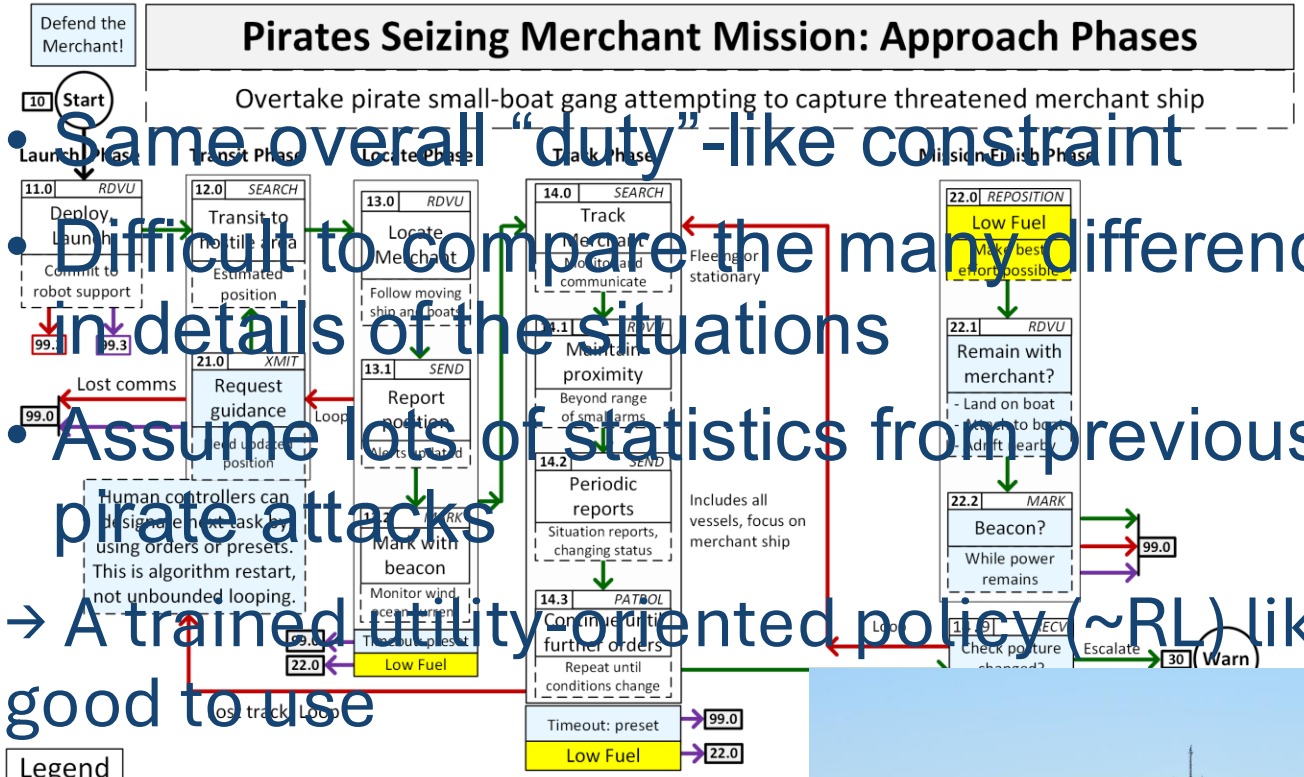


Image from NPS

Example: Multiple Pirate Attacks

P(interdict) = f(distance)

Utility(interdict) = g(crewsize)



- Same overall "duty"-like constraint
- Difficult to compare the many differences in details of the situations
- Assume lots of statistics from previous pirate attacks

→ A trained utility-oriented policy (~RL) likely good to use

Example: Multiple Pirate Attacks, but you saw a water cannon demo

Requirements for any approach to deal with this

- integration of a variety of constraint knowledge sources and types, such as examples, rules, costs, abstract virtues
  - Need an expressive knowledge representation that can match to the diversity of constraint knowledge present in the world.
- real-time constraint knowledge incompleteness and/or conflict detection.
- real-time mitigation of the above.

Important Aspects of our Approach in Soar

- Conflicts and knowledge incompleteness are explicit
- Metareasoning guides decision-making in those cases
  - You can tell why an agent picked what it picked
- Soar supports this as part of its real-time reasoning

Nuggets

- Plays into the strengths of Soar: About bringing all available knowledge to bear on a problem
- Doesn’t depend on taking a moral stance to champion one ethical frame at design-time
- More legible reasoning behind constraint-related decisions

Coal

- Inability to simplify problem around a prescribed form of constraint knowledge