Agent Response to Situations Outside of Design Scope (OODS)

Center for Integrated Cognition

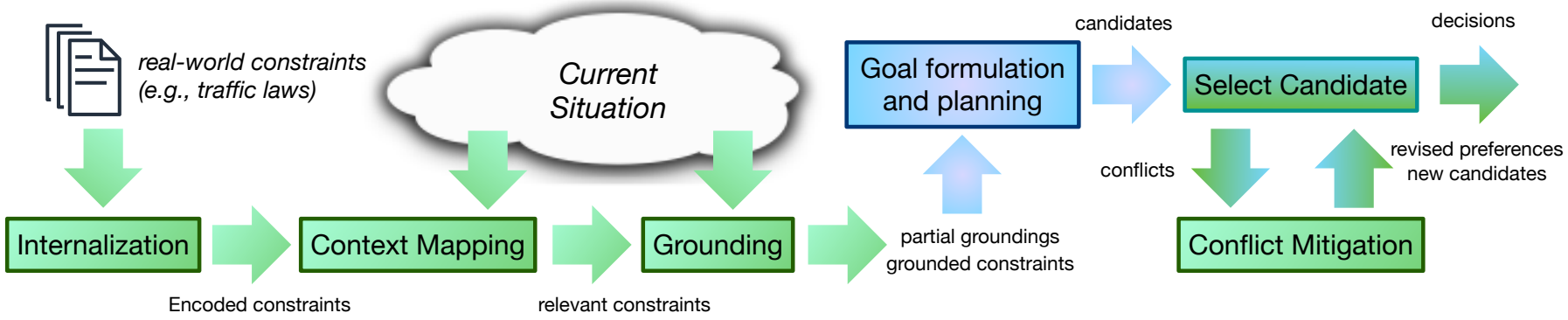
Robert Wray, Steven Jones, John Laird

45th Soar Workshop

5 May 2025

Preprint: Wray, R. E., Jones, S. J., & Laird, J. E. (2025). *Heuristic Recognition and Rapid Response to Unfamiliar Events Outside of Agent Design Scope* (arXiv:2504.12497). arXiv. https://doi.org/10.48550/arXiv.2504.12497

Point of Departure: Operationalizing Constraints



Prior presentations: Constraint Compliance

- Abstract specifications
- Grounding in specific situations
- Balancing conflicts between conformance and performance

Key Observation/Lesson: Constraint Compliance challenges **greatly** lessened via familiarization

- Training (as in RL)
- Instruction (as in ITL)
- Experimentation/exploration

Decision Making in Unfamiliar Situations

New Problem Focus: Dynamic Decisions in Unfamiliar Situations

An agent's prior experiences have not (fully) prepared it for the unfamiliar situation it encounters…

- Is it in any danger?
- Can it continue its mission/goals?
- What should it do in a specific unfamiliar situation?
- What can it do (what relevant affordances does it have?)
- How much time does it have to make a decision?

Image: ChatGPT

Out of Design Scope (OODS) Situations

Isn't an agent encountering an unanticipated situation a failure of design?

→ Reality: Environments are unbounded

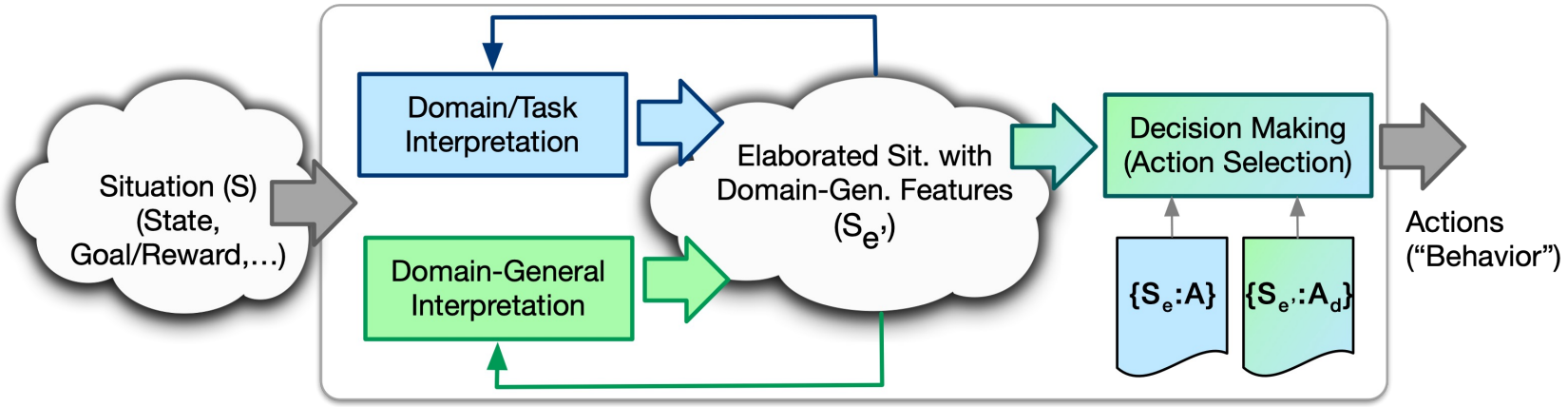
- Designers cannot anticipate (and prepare the agent for) everything
  - Economics generally dictates omitting preparation for very rare events
  - Designers often assume some boundedness (limited operational settings)
- Environment will change during deployment
  - New objects and properties, new constraints (change in laws), new contexts

General autonomous agents will encounter situations outside of their design scope (*OODS situations*)

Types of OODS Situations

Type	Description	Example
Out of Distribution (OOD)	Familiar but policies fail to cover specific combination of features and values	Autonomous driver trained on North American roads asked to drive at 120 mph.
Out of designed-feature scope	Observations contain data salient to effective response, but recognized/known features are inadequate (or inapt).	Camera-based, lane-following system that was not trained on or exposed to snow being used on snowy roads.
Out of observation	Relevant situation features not directly perceptible given agent's embodiment. Agent ability to survive and adapt is limited.	Agent deployed on roads near active volcanoes but lacks perceptual capability to distinguish lava from rocks.

What might a solution look like?



Requirements

- Survive
- Continue mission (zero-shot adaptation)
- Accommodate irreversibility
- Scale to complex tasks
- Adapt to changing environment
- Improve with experience

Other Approaches to the Problem

Lots of related work on “open worlds” decisions… Solution categories:

- Seek more knowledge
  - Deliberation (e.g., plan from first principles)
  - Ask for guidance
  - Explore/Experiment/Imagine (including modeling)
- Run away!
  - Default behaviors (return to base, return to safe/familiar state)
- Make Assumptions
  - Apply existing models
  - Ignore (maybe it will go away)
- Ensembles/combinations of these approaches

What Do Humans Do in Unfamiliar Situations?

Various theories… Appraisal theory (Scherer et. al.)

- People continually assess the situations they are in
- Assessments (“appraisals”) add context about the current situation
  - Novel, unpleasant, etc.
  - Set of useful, general meta-properties that characterize situations (tuned by evolution)
- Appraisal processes (generally) are:
  - Rapid
  - Hierarchical (some appraisals depend on earlier ones)
  - Knowledge-lean (modulated by, but not dependent on contextual understanding/knowledge)
- Appraisal processes:
  - Produce action tendencies (“fight or flight”, “closed” vs. “open” posture)
  - Inform production of emotion

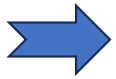
Human Appraisal

- Recognize this situation?
- Pleasurable or painful? Feel in danger?
- Does the situation help or hurt my overall goal(s)?
- Sense time pressure to act?
- What levels of power/control might I have?
- How does the situation impact others?

Image: ChatGPT

Functional Roles of Appraisal Processing

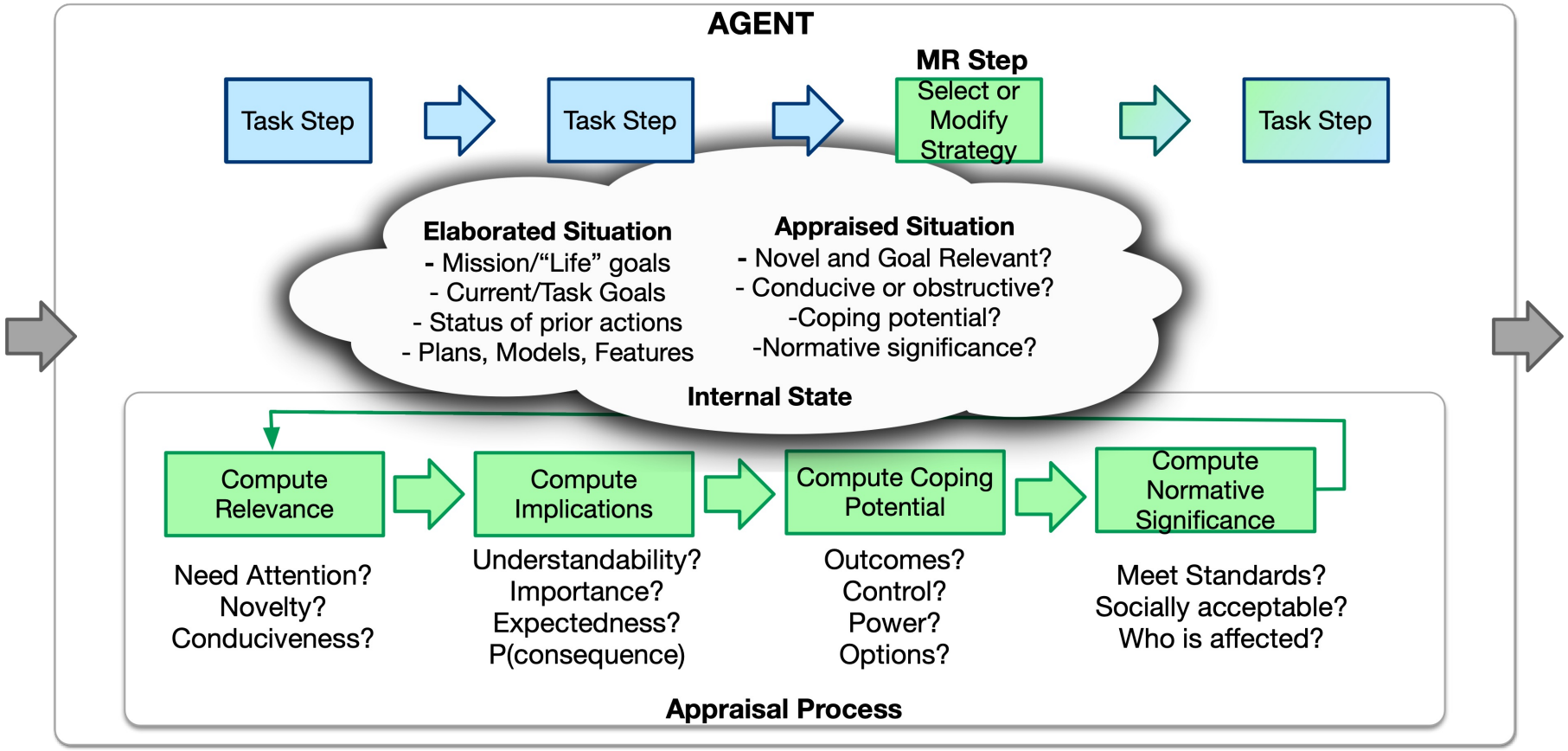
- Appraisal processes in psych are generally seen as precursors to emotion
- For autonomous agents, we are largely interested in the **functional roles** of appraisal processing:
  - Appraisal processing produces meta-knowledge that characterizes key properties (appraisals) of the agent's situation
  - Specific combinations of appraisals inform:
    - Choice of specific actions (attend, flee, etc.)
    - Choice of decision-making strategy



Input for Metareasoning

- Search for more knowledge
- Make some assumptions
- Run away!

Current Vision



Conclusions

Nuggets

- Human “technology” again looks very relevant to open problem in general intelligence
- Research problems at the edge of content/architecture interface
  - Parallelizing task reasoning with appraisal processing (chunking?)
  - “Knowledge-Modulated” architectural processes?

Coal

- Large scope; breadth or depth?
- Large body of related work might make it hard to differentiate appraisal story
- Also need methodology advances
  - Create situations “outside of design scope” that are rigorous, defensible

Acknowledgements

This work was supported by the Office of Naval Research, contract N00014-22-1-2358. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.