Incorporating Abstract Behavioral Constraints in the Performance of Agent Tasks

Bob Wray, John Laird

Jun 2021

CIC @ IQMRI

Motivation: Long-lived, Adaptive Autonomous Systems

Increasingly intelligent automation ("autonomy") is becoming part of our lived experience….

- Autonomous systems pose many difficult requirements:
  - Do the task: Responsive, robust task execution
  - Economical feasibility: Easy to develop and to extend
  - Work with users: Easy to customize and to interact with
  - Adaptive to new tasks, task contexts, and user requirements
  - Be Safe: "Follow the rules", be predictable/understandable, make ethical decisions

Requirements (generally) must be met across all the dimensions

Image: Tesla

Image: US Navy

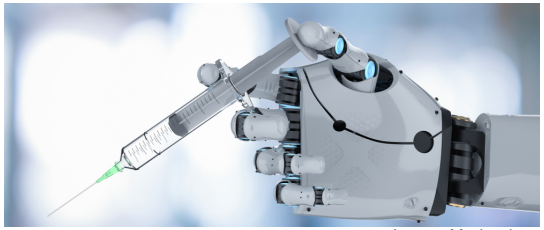


Image: Medgadget

CIC @ IQMRI

Approaches vs. Requirements (High-Level Analysis)

	Efficient Task Execution	Cheap to Develop	Easy to Interact	Adaptive to New Tasks	Predictable and Safe
Knowledge Engineering	✓	⊖	✓	⊖	±
Policy Learning	✓	±	±	±	⊖
Interactive Task Learning	±	±	✓	✓	⊖

No current technology/methodology satisfies all these requirements

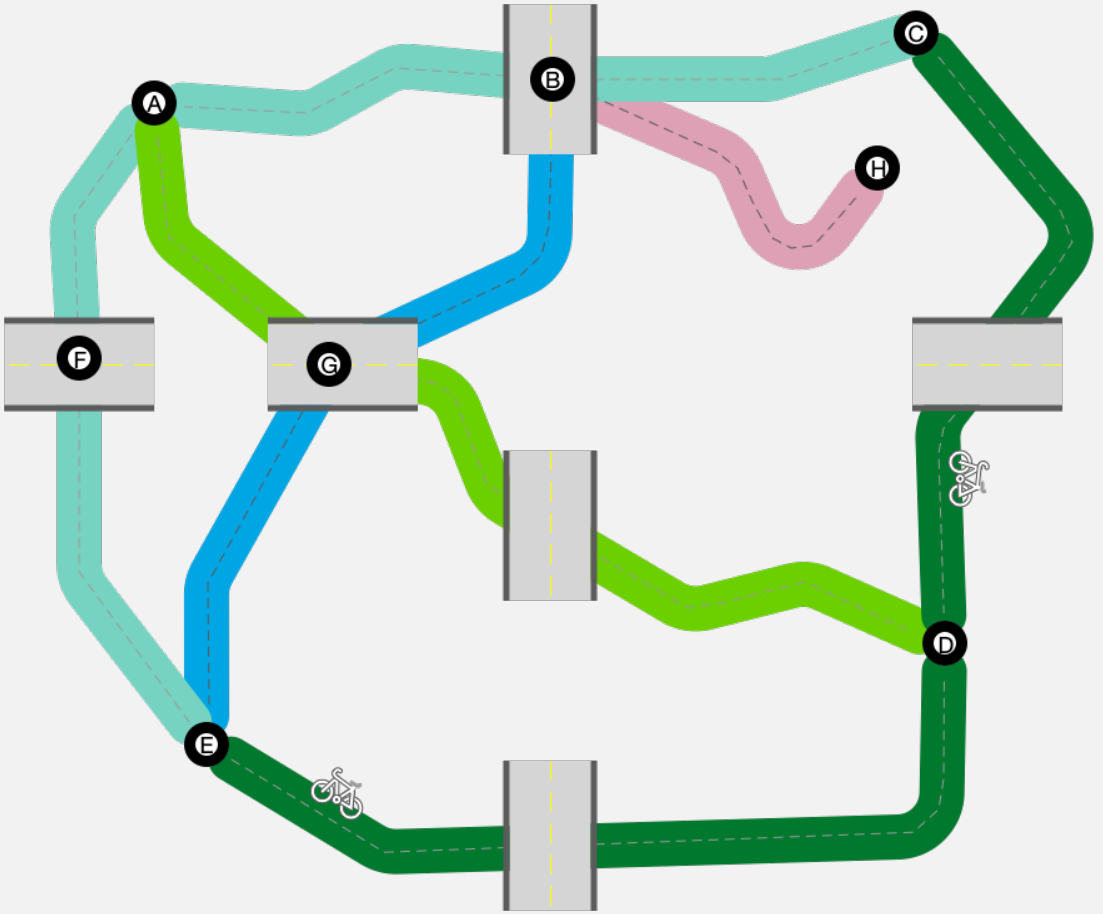
Long-term goal: Develop methods that allow agents to incorporate new/different behavior constraints while also continuing to perform tasks efficiently and robustly (within the constraints)

Work-to-date: Problem definition to better understand requirements

Illustrative Example

Robot system that uses a network of multi-use bike paths for rapid, goal-directed movement (e.g., deliveries)

What would be involved in the design of a general system for the problem (e.g., deployable to anyplace there is a network of bicycle paths)?



Abstract Behavioral Constraints

Many sources and kinds of constraints on task behavior

Operational conditions: The task-performance context influences how the task should be executed

- Examples:
  - Rainy vs. icy vs. sunny weather
  - Braking distance (wear on brake pads)
- Operational conditions change (not inherent)
  - Changes can be rapid or slow

Norms and obligations: Informal, implicit prescriptions; "the way it's done around here"

- Examples:
  - Cruising speeds around other path users
  - (Calling out) "On your left"
- Norms express typical/frequent behavior patterns
- Norms often imply social expectations (one "ought" to follow norms…)

Rules and laws: Formally defined prescriptions on behavior.

- Examples:
  - (US) Pass on the left
  - (US) Stop before entering a roadway
- There can be many, many laws and rules governing some behavior
  - Ex: Military doctrine and the laws of war

Safety and Ethics: Bounds on allowed behavior that attempt to minimize injury or cost to self and others

- Example: "Safe" distance when passing
- Some safety concerns look like ethical decisions
  - Crash oneself or strike an unavoidable obstacle?

Abstract Behavioral Constraints

- Many sources and kinds of constraints on task behavior
  - Operational constraints
  - Rules and laws
  - Norms and obligations
  - Safety and Ethics
  - …

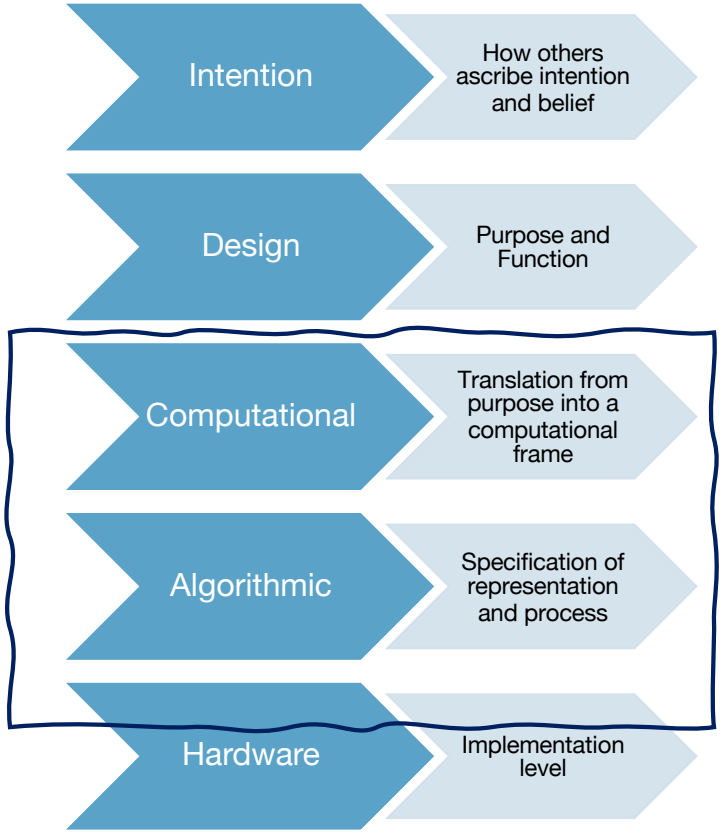
- Ensuring behavior consistent with all constraints can be challenging
  - Many constraints
  - Hard constraints vs. soft constraints
  - Constraints interact
    - Worst case: combinatorically many interactions
  - Incompleteness
    - Contingencies
  - Inconsistent constraints

CIC @ IQMRI

Are there general solutions for handling systems of constraints?

- Goal: Research what kinds of knowledge and capability are necessary for an agent to incorporate and conform to abstract behavioral constraints in its behavior
  - How have others dealt with these constraints?
  - Operationalization problem
  - Requirements for a solution

- Caveat: There is increasing critique of AI about what bounds and assumptions are/are not included in AI systems
- One framework: Artificial Morality (Misselhorn, 2019)
- Our aim is to explore solutions at the computational levels and below
  - Search for solutions that realize intention and design in the implementation as close to the design intent as feasible

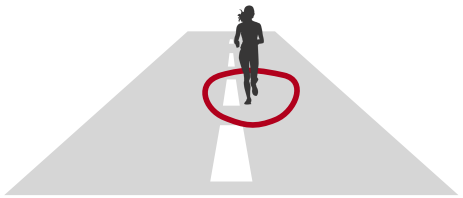


CIC @ IQMRI

Implementation of Behavior Constraints in Prior Agent Approach

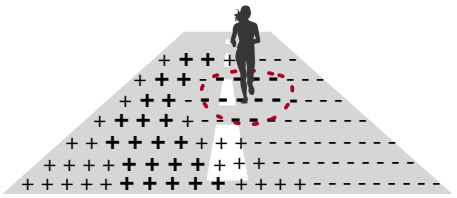
Explicit Procedural Encoding:

- Constraints are directly encoded in the task knowledge itself
- Standard, common approach in many agent systems (BDI architectures, Soar 8)
- Advantage: Efficiency/speed
- Limitations:
  - Functional fixity/non-adaptive
  - Incomplete specification (or very costly)
  - Limited “scrutability”



Implicit Procedural Encoding:

- Constraints are indirectly encoded within task performance policies and utilities
- Standard approach in RL agents
- Advantages: Efficient and adaptive
- Limitations:
  - Incomplete constraints (and less understanding of incompleteness)
  - Unpredictable behavior (unexplored spaces)
  - Largely inscrutable



- **Declarative Encoding**:
- Encoding of constraints in declarative form
- Logic-based approaches (e.g., Arkin’s Ethical Governor)
- Advantages:
  - Scrutability
  - Adaptability
- Limitations:
  - Recurring expense of interpretation
  - Unpredictability

Declarative Encoding

1. Pass on the left of the entity.
2. Leave a safe distance between yourself and the entity as you pass.
3. After passing, return to the right side.

CIC @ IQMRI

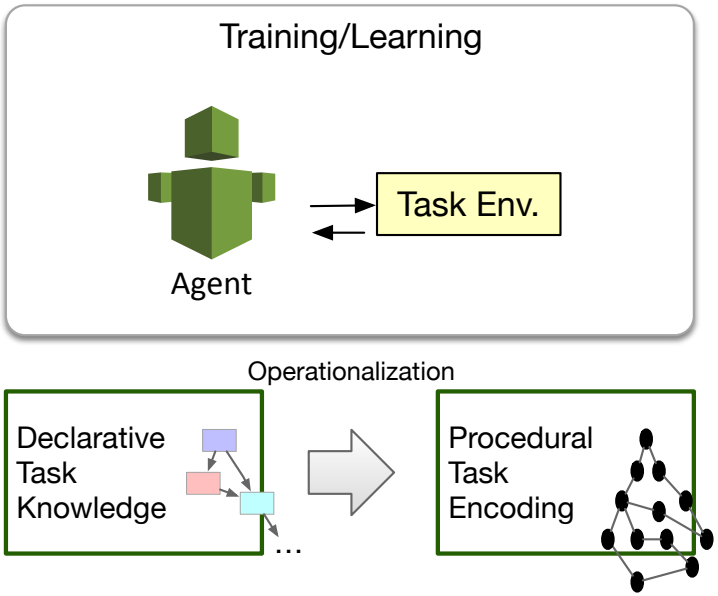
The Operationalization Problem

Why doesn't chunking suffice?

Compilation of declarative knowledge via EBL/chunking during a "training period" would speed up declarative approach via proceduralization of the constraints

Limitations

- Incompleteness: Training will not cover all cases/all possibilities
- Consistent Generalization: Is generalization reliable over the constraints?
- Changing constraints: What does the agent do when constraints change?
  - Changing operational characteristics
  - New rules/laws
  - …



Insights from Interactive Task Learning

- Agents benefit from a “dual” representation of a task
  - Procedural: Efficient, compiled
  - Declarative: (Self)inspectable, enables (straightforward) understanding of “how I do the task”

- Anticipation via mental simulation
  - The agent can evaluate what is likely to happen before taking a step
  - Identify potential gaps in knowledge

- Apprenticeship learning model
  - “Fail soft” learning/training
  - Use the expertise of others to “fill in missing gaps” before having to perform the task autonomously

Priority Requirements for a Solution

- Dual representation of tasks
  - Agent supports both declarative and procedural representations of task
  - Agent must maintain “parallel and consistent” representations

  → Self-evaluation of representations and active steps to anticipate/resolve conflicts
- Retrospective assessment
  - Agent must evaluate actual performance against expected performance
  - Identify potential misalignments
  - “Confidence” in a particular task context
- Metacognitive reasoning
  - Anticipate future states and act to avoid problematic ones
  - Balance processing needs (task performance, retrospective assessment, anticipation)

Nuggets & Coal

Nuggets

- Abstract behavior constraints have significant impact on what specific agent behavior is appropriate (and when)
- Architectural perspective pushes toward a more systematic and general approach
- Formulation of a unified approach to constraints is (seemingly) novel and will have large payoff if successful

Coal

- Big problem. Just taking the first steps/getting familiar with the terrain.
- Complexity of evaluation (task performance, changing constraints, multi-domain?)

For more details: Wray, Robert E., and John E. Laird. “Incorporating Abstract Behavioral Constraints in the Performance of Agent Tasks.” In *Proceedings of the International Conference on Artificial Intelligence*. Las Vegas, NV: Springer, 2021.