Extending the CMC with Metacognition

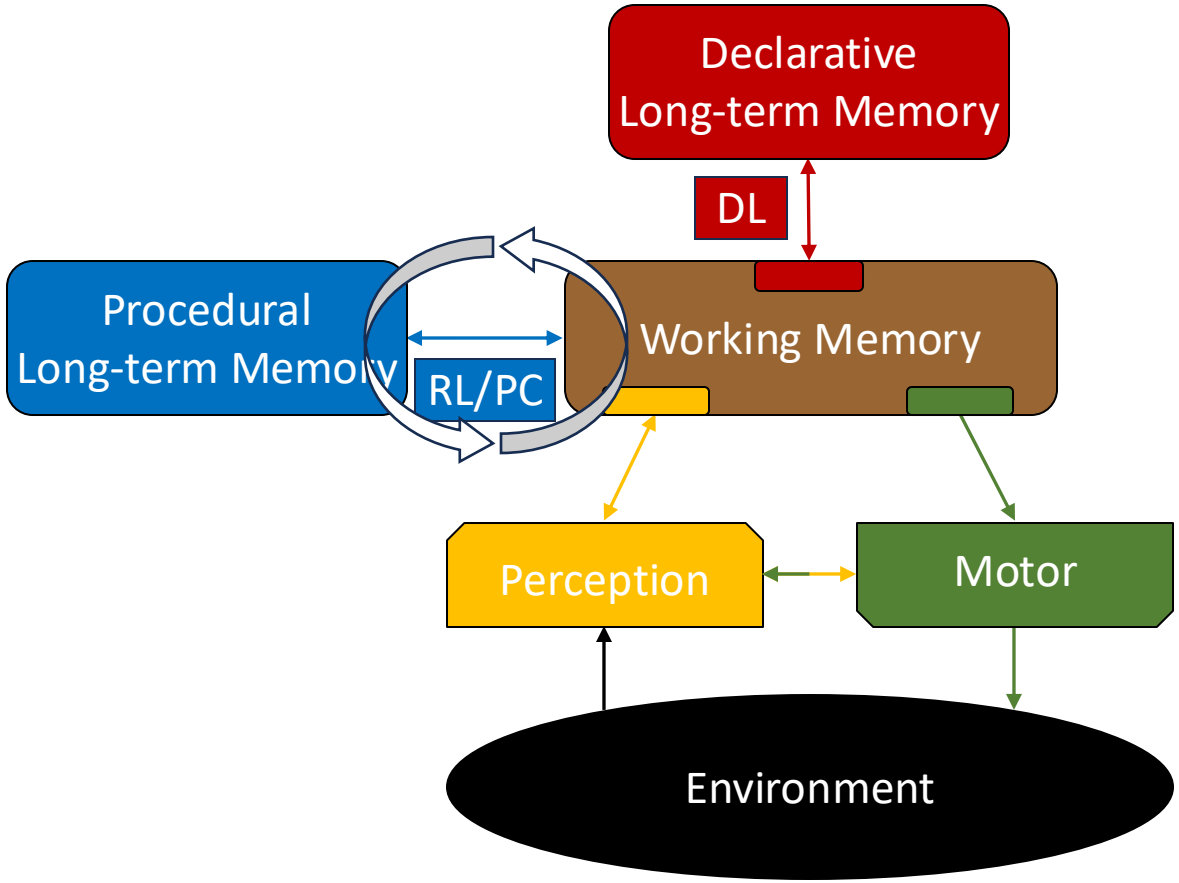
Center for Integrated Cognition

John Laird,
CMC Mafia, CIC, …

May 5, 2005

Soar Workshop

Common Model of Cognition



- Consensus abstract model of cognition architectures for human-like behavior
- Reasoning cycle driven by procedural memory interacting with working memory
- Focus on routine performance & learning
  - No metacognition

Metacognition and Metareasoning

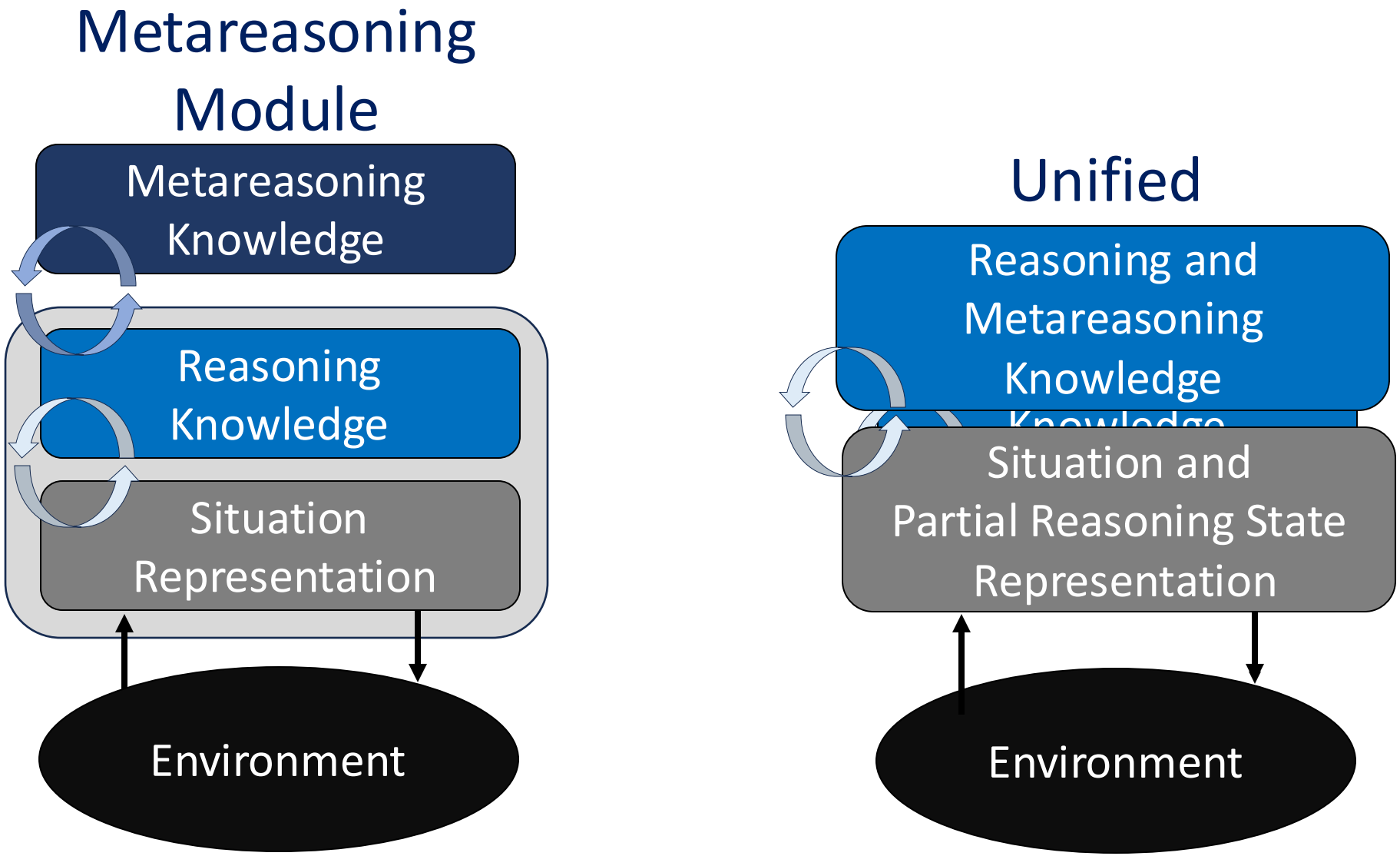
- Metacognition:
  - *Reasoning about all aspects of cognition*
  - CMC: Reasoning about reasoning, learning & memory, perception, motor
- Metareasoning:
  - *Reasoning about reasoning*
  - CMC: Reasoning about what happens in the reasoning cycle

- How can CMC be extended to include meta-reasoning?
  - Take inspiration from CMC architectures: Soar, Sigma, ACT-R

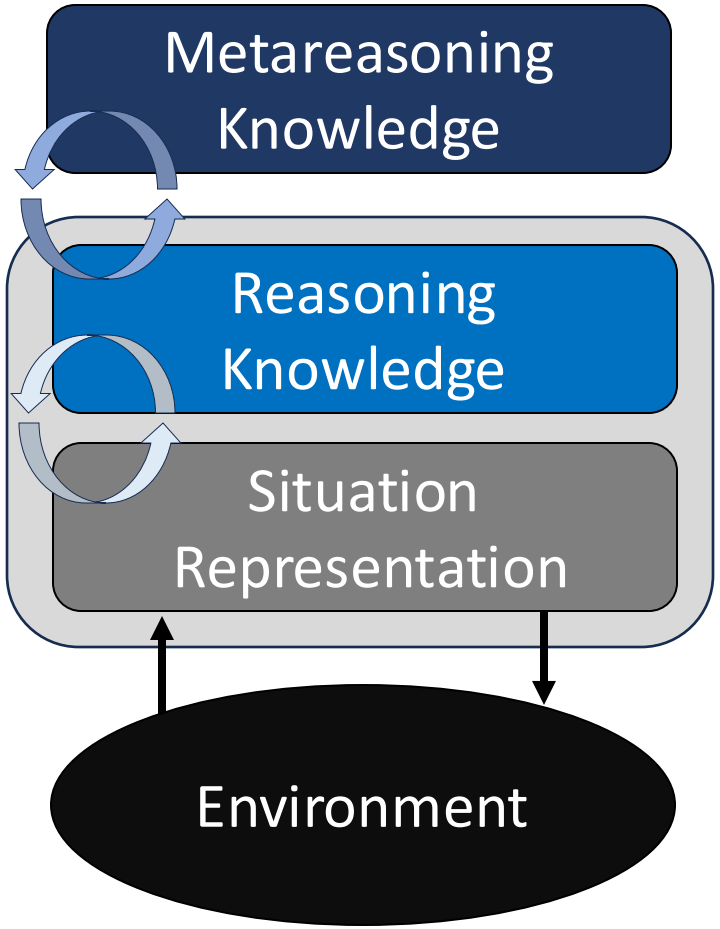
Examples of Metareasoning

- Introspective Monitoring
- Reasoning Failure Recovery
- Deliberate Decision Making
- Predictive and Hypothetical Reasoning
- Retrospective Reasoning
- Strategy Selection
- Self Explanation

Alternative Models of Metacognition



Metareasoning Module



Merits

- Clear separation of reasoning and metareasoning
  - Allows continual, parallel metareasoning
  - Avoid interference
- Focus on domain-independent metareasoning

Challenges

- Added complexity
- Limited reuse of cognition capabilities
- Assumes access to modules' internals
  - Incompatible with neural models of memory

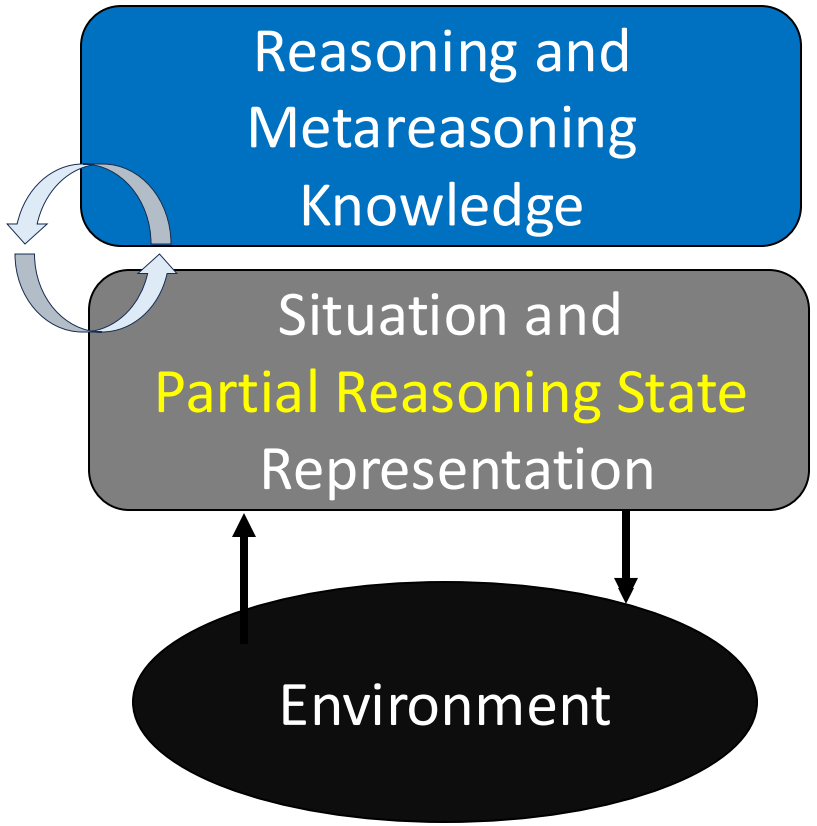
Merits

- Simpler architecture
- Reuse of cognitive capabilities
- Intermixes reasoning and metareasoning
- Compatible (?) with neural models of memory

Challenges

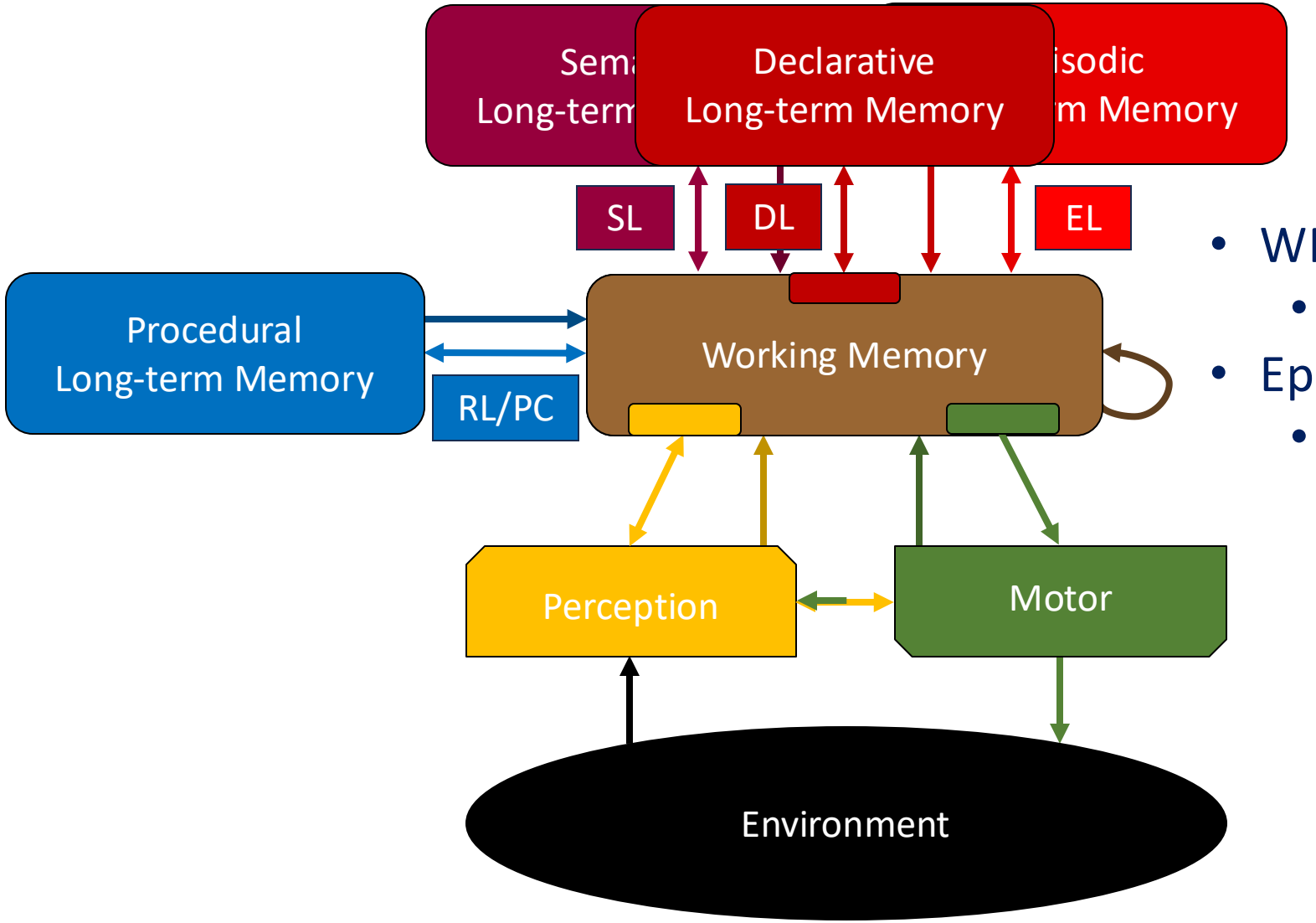
- No parallel metareasoning and task reasoning
- No pre-existing self-model to access
- Must learn incrementally from direct and indirect sources
- What is source of the partial reasoning state?

Unified



CMC Architecture Extensions: Direct Sources

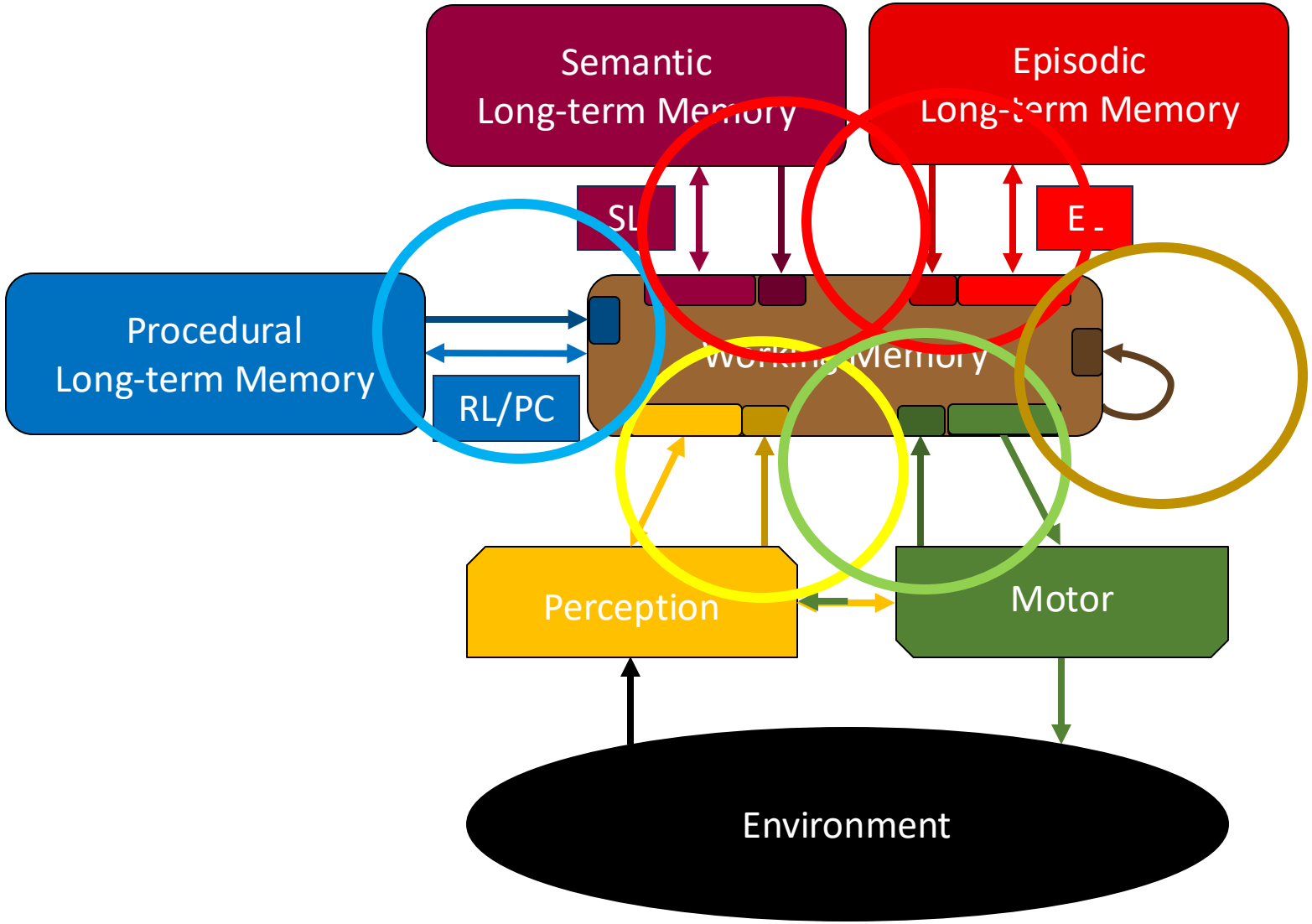
- Add data to WM that it can reason over



- WM Module Process State buffers
  - Success/failure/certainty/…
- Episodic Memory
  - History of reasoning

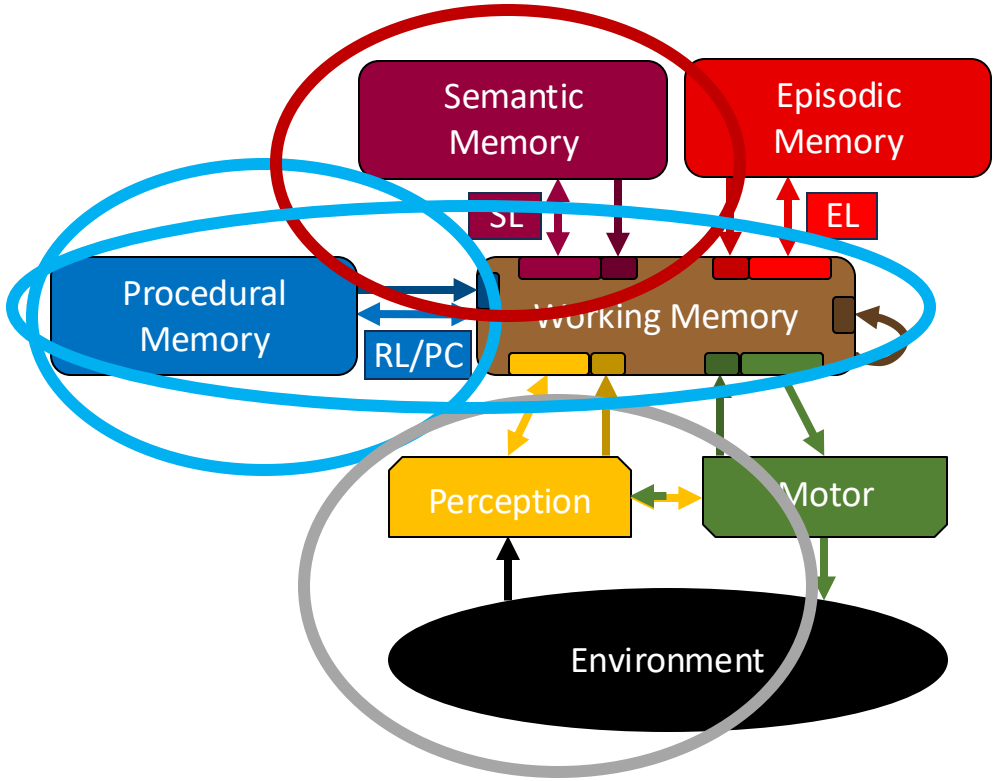
CMC Architecture Extensions: Direct Sources

- Add data to WM that it can reason over



Indirect Sources of Reasoning Information

- Procedural Memory
  - Acquired from reasoning over other sources
  - Compiles metareasoning to reasoning!
- Semantic Memory
  - Saves WM info from other sources for future
- Environment
  - Self Observation
  - Other Agents
  - Externally recorded Information
- Metareasoning
  - Composition of other sources



Example Types of Metareasoning

- Introspective Monitoring
  - Procedural memory, working memory metacognitive appraisals
- Reasoning Failure Recovery
  - Impasses/failures from module retrievals
- Deliberate Decision Making
- Predictive and Hypothetical Reasoning
  - Impasse in decision making
  - Internal simulation of alternative choices
- Retrospective Reasoning
  - Impasses/failures from module retrievals
  - Episodic memory to reconstruct reasoning trace
- Strategy Selection
- Self Explanation

What's Missing? (But in Soar)

- Explicit representation of operators in working memory
  - And acceptable preferences
  - Supports micro-metareasoning for decision making
  - Is this necessary to support prospective and hypothetical reasoning?
- Substates
  - Provides representation independent of the current task state
  - Provides indirect access to task state – can (meta)reason without modifying it.
- Should these be part of a CMC proposal?

Do LLMs have Metacognition / Metareasoning?

- They have only *indirect* sources of knowledge about reasoning.
- They can “reason” about what they’ve been trained on about reasoning, but they cannot reason directly about their own reasoning.