WASHINGTON STATE UNIVERSITY
VANCOUVER

World Class. Face to Face.

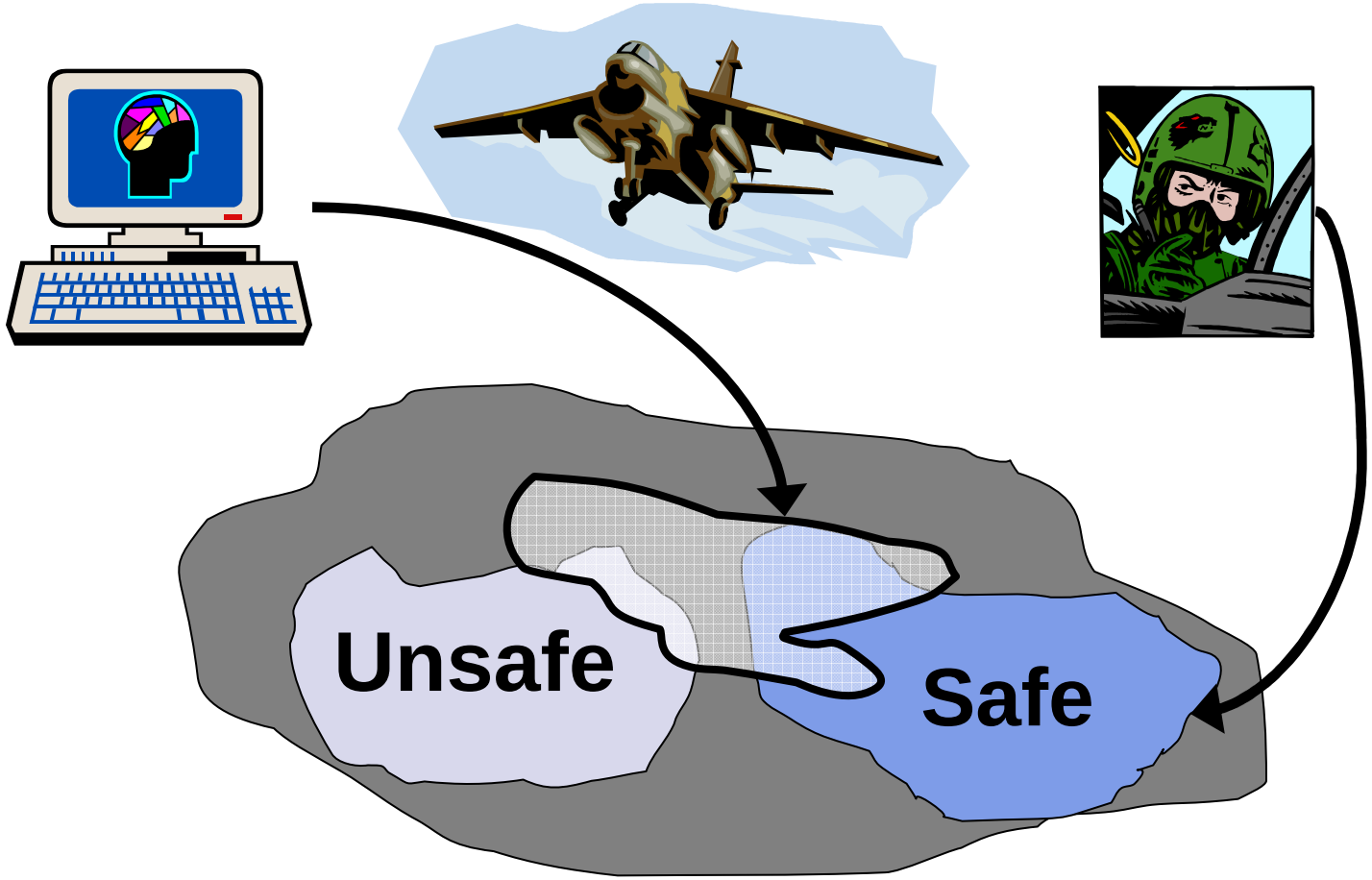
S-Assess:

Self-Assessment with Soar

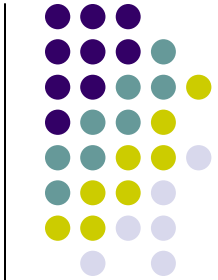
Scott Wallace

WSU Vancouver

Building Human-Level Agents



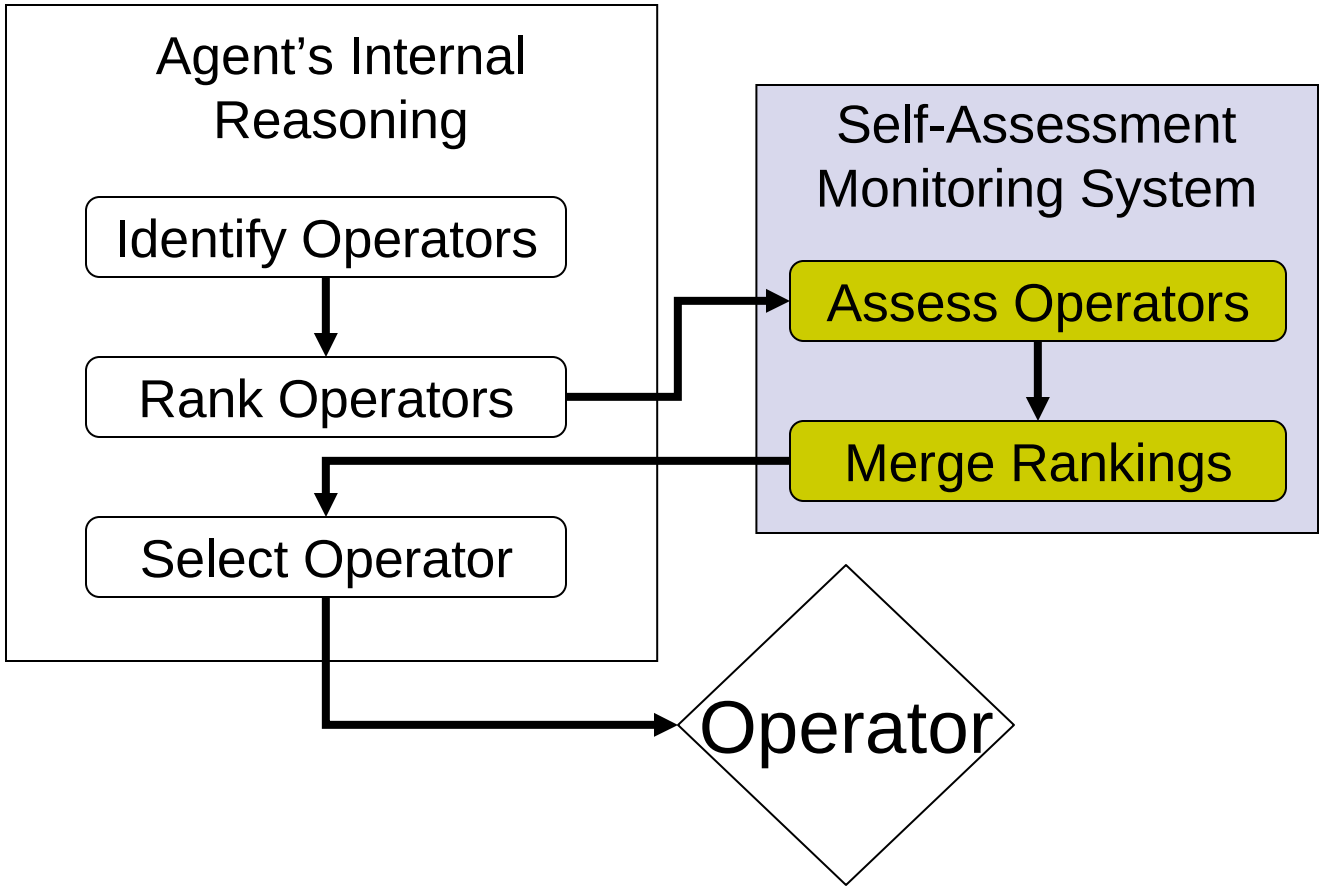
Detecting Errors in the Lab



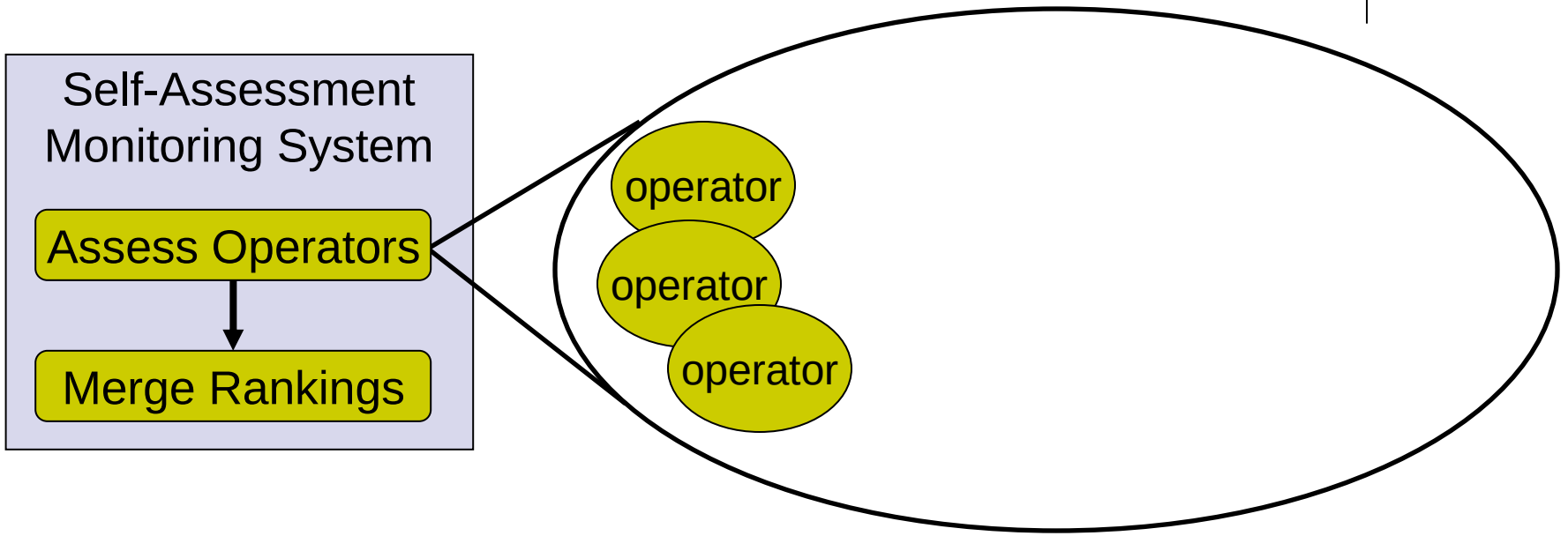
- Abstract Behavior Representations
  - Concisely represent patterns in behavior
  - Provide a basis for automated comparison
  - Allow aggregate behavior to be examined efficiently

- Yet, in lab testing is *incomplete by nature*
  - Creates trust problems
  - Need to bring validation into the field

Self Assessment Framework

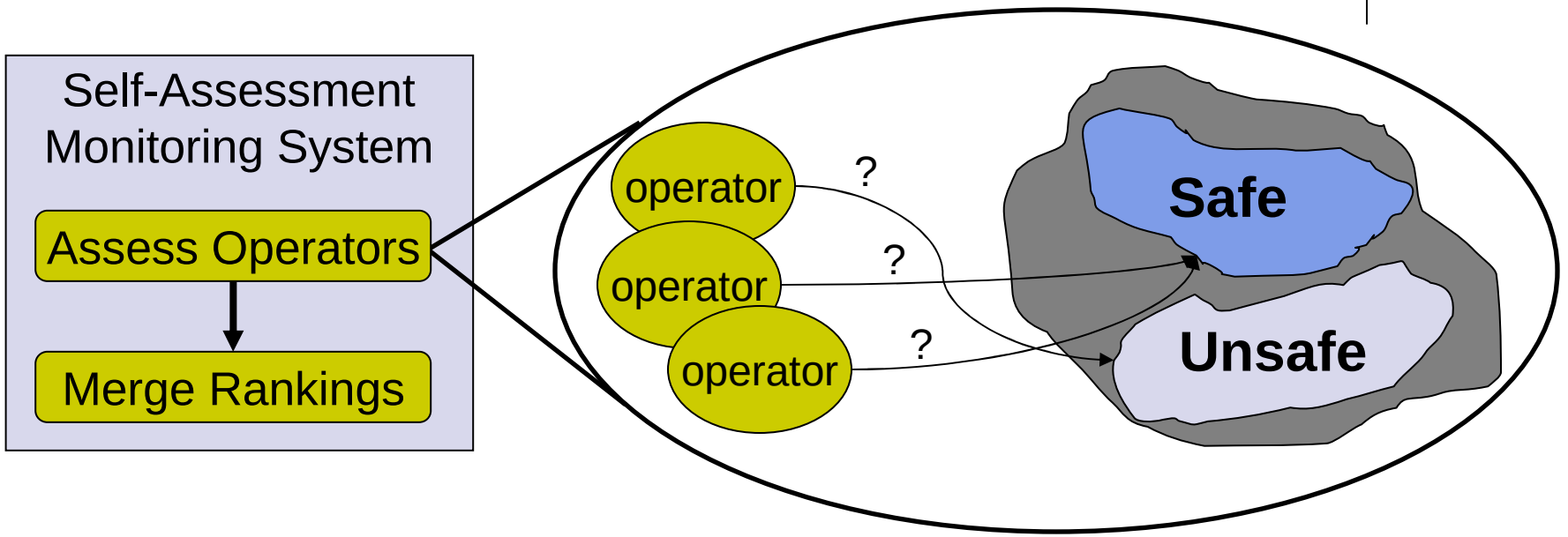


Assessment



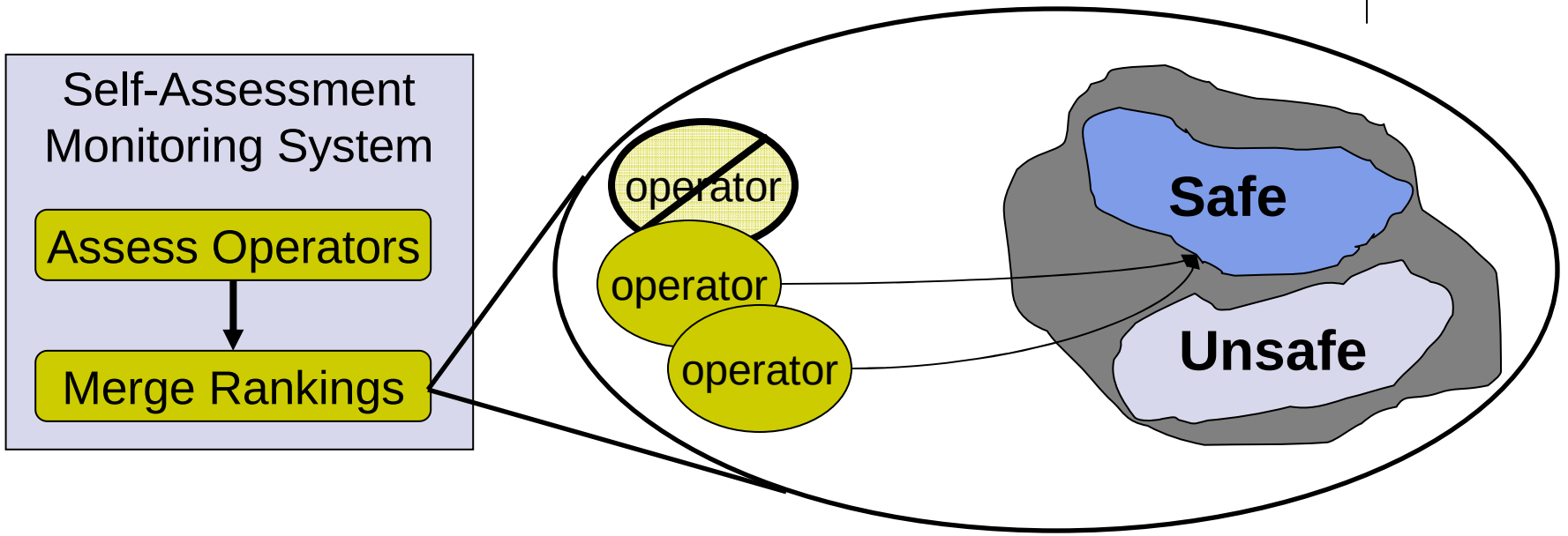
- The Agent is allowed to suggest a set of operators and their relative rankings

Assessment



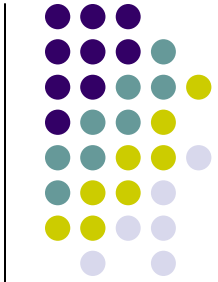
- Framework evaluates potential operators by comparing against domain constraints

Assessment



- Unsafe operators are rejected
- Partial ordering remains otherwise intact

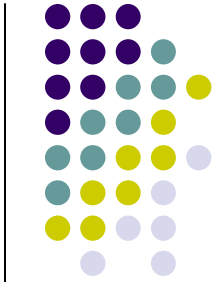
S-Assess via Knowledge



- Domain independent
- Set of 62 Soar rules
- Minor adjustments required to existing agents
  - Preferences encoded as operator augmentations

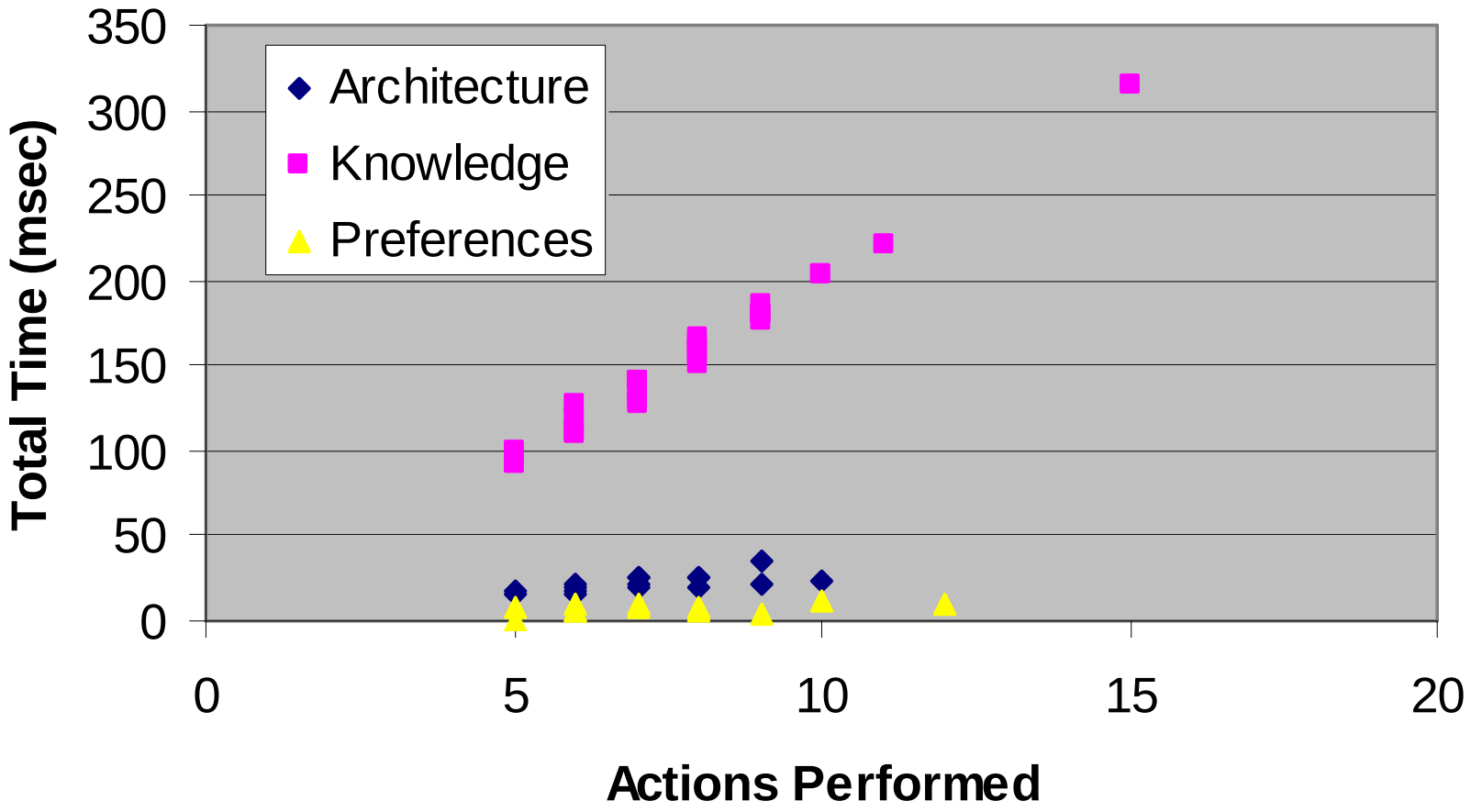
- Impasses trigger assessment operation
  - Assessment creates new symbolic preferences

S-Assess via Architecture

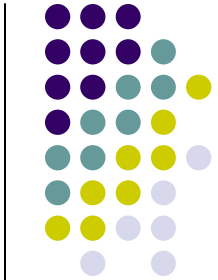


- Domain independent
- Architectural Modification
- Intercepts Preference Calculations
  - For each preference, S-Assess returns either:
    - ALLOW (preference consistent with rule book)
    - DENY (preference inconsistent with rule book)
  - Based on assessment result Soar's own preference calculations may be short circuited

S-Assess Overhead



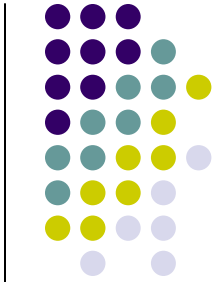
Beyond Safety



- S-Assess acts as a high-level control

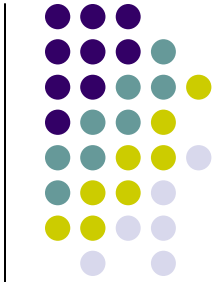
- Potential uses for such a control:
  - Safety/Correctness (specified by designers)
  - Environmental Policy (specified by environment)
  - Social Policy (specified by other agents)
  - As an exception mechanism
    - Adjustable Autonomy

S-Assess and Soar-RL



- S-Assess
  - Targets offline learning
  - Provides a mechanism for “universal preference computers”
- Soar-RL
  - Targets online learning
  - Provides a clean integration of reinforcement learning with Soar’s knowledge representation

Nuggets & Coal



- Nuggets:
  - An interesting counterpart to Soar RL
  - Ready for real experimentation
    - Soar + Bayes Nets?
- Coal:
  - Some efficiency is sacrificed