Detecting Errors in Agent Behavior

Scott Wallace

University of Michigan

Soar Workshop – May 2002

The Problem of Correctness

- Agent’s must have correct, expert-level behavior
- Errors undermine project’s goals
- How can we ensure correctness?

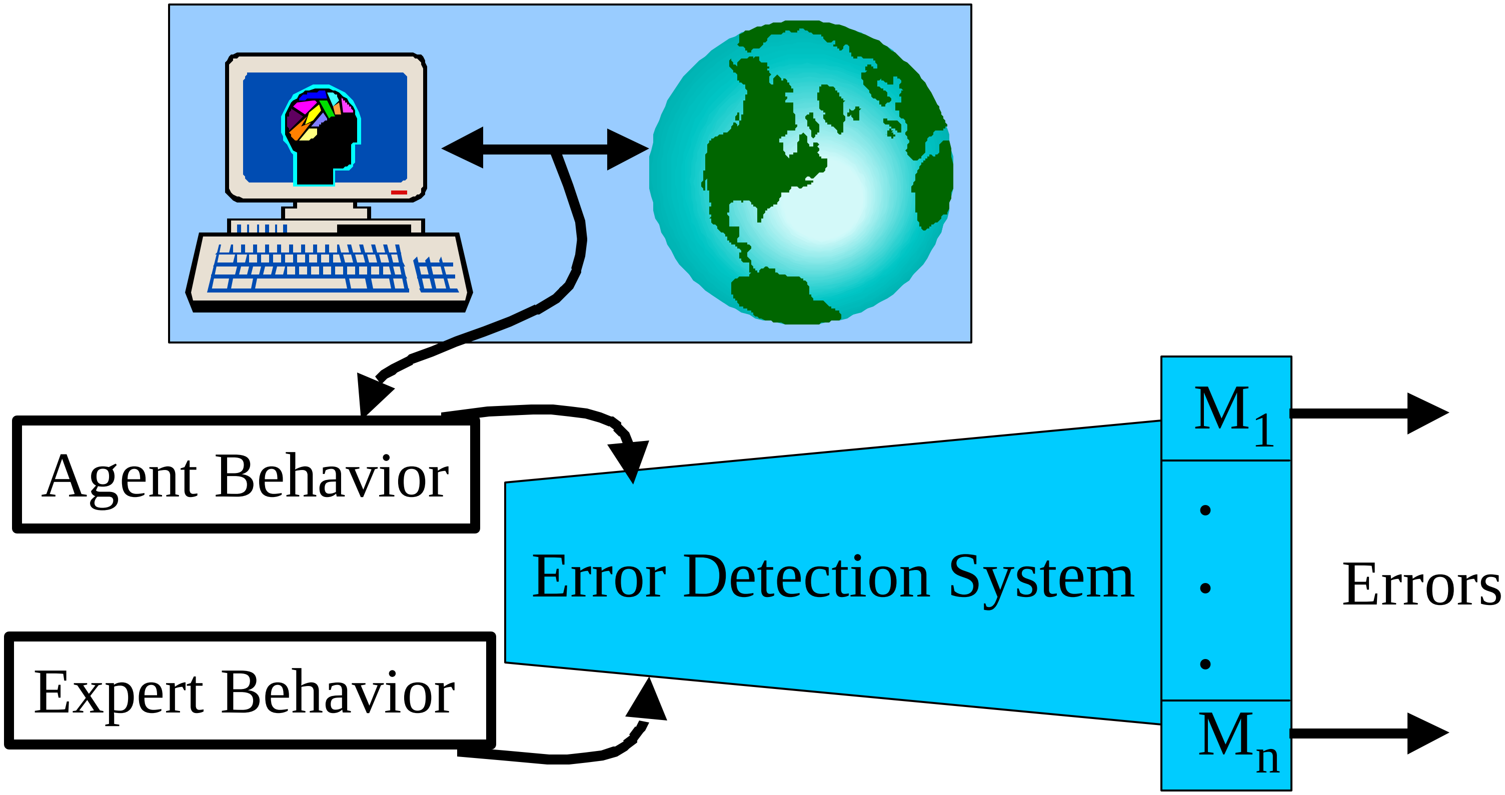
The Validation Bottleneck

- Our emphasis is on error detection

- Manual Validation: Expert critiques agent behavior
  - Requires significant human effort
  - Difficult to detect every error

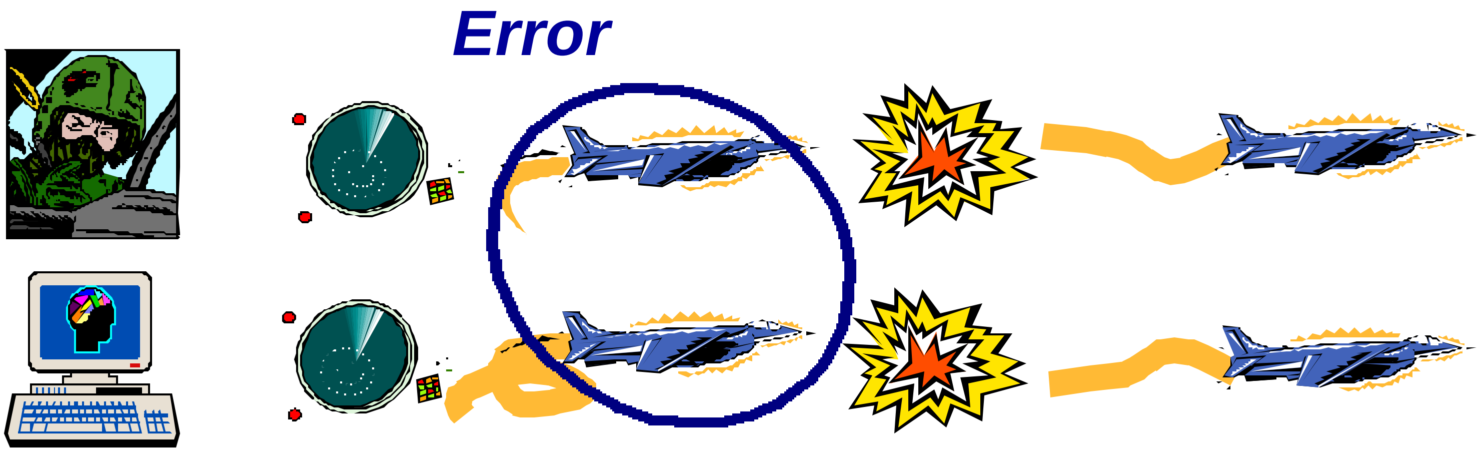
- Automated Validation
  - No precise definition of correct/incorrect behavior
  - "I can't tell you what's incorrect, but I know it when I see it."

System Overview



Detection can be performed online or offline

Initial Approach

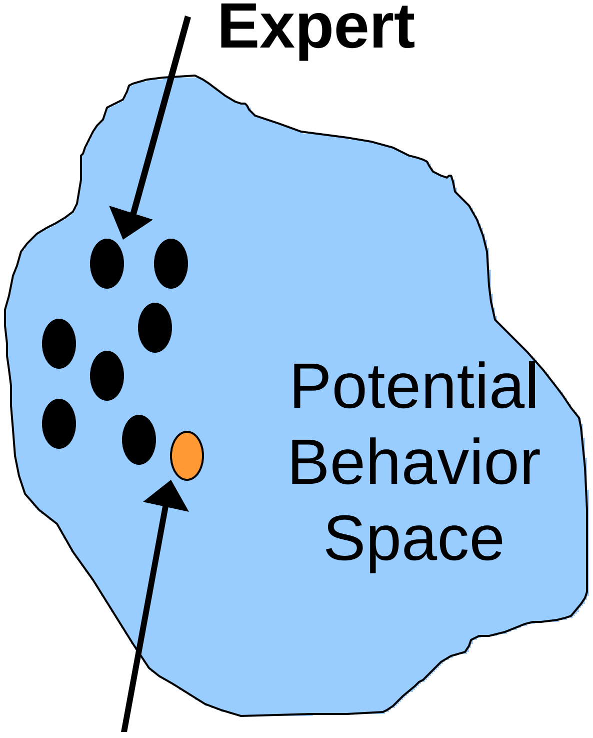


- Extract actions or goals from behavior traces
- Form a sequential representation
- Discrepancies between sequences indicate errors
- Works well for individual behavior traces

From Another Point of View

- Sequences represent instances of behavior
- Instances are points in the behavior space
- Want to represent aggregate behavior

Observations of Expert

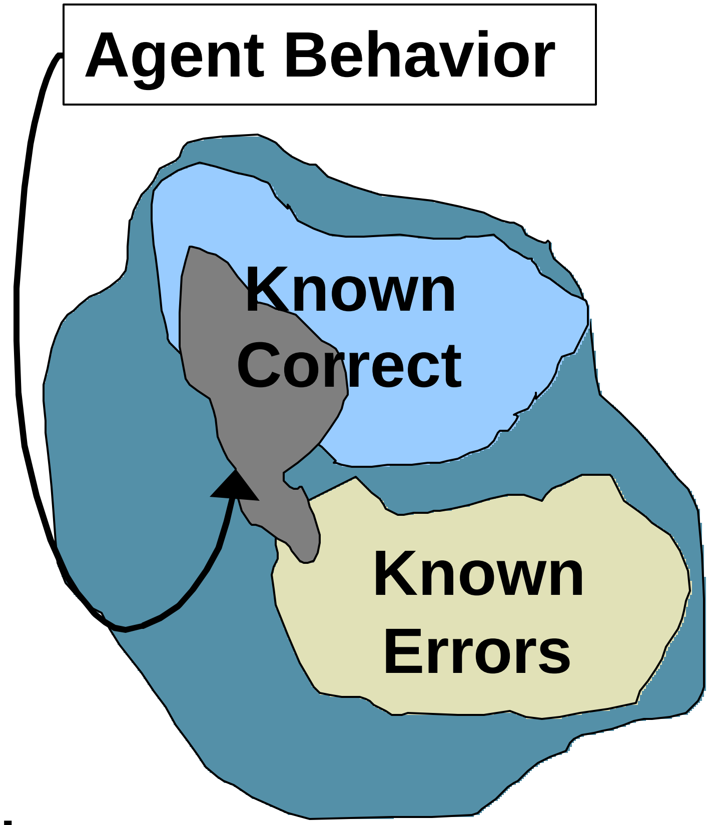


Potential Behavior Space

Observation of Agent

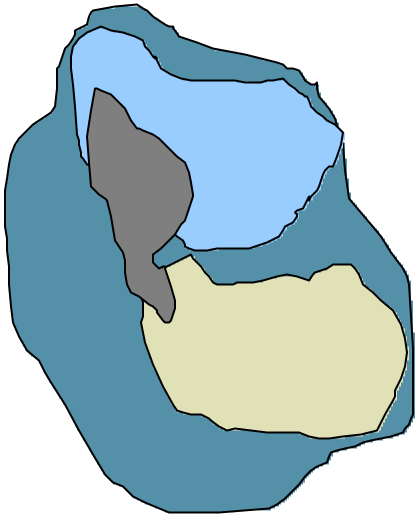
New Aggregation Approach

- Define boundaries in the space of potential behavior using:
  - – observations
  - – knowledge of task requirements
- Determine portion of agent behavior in each region

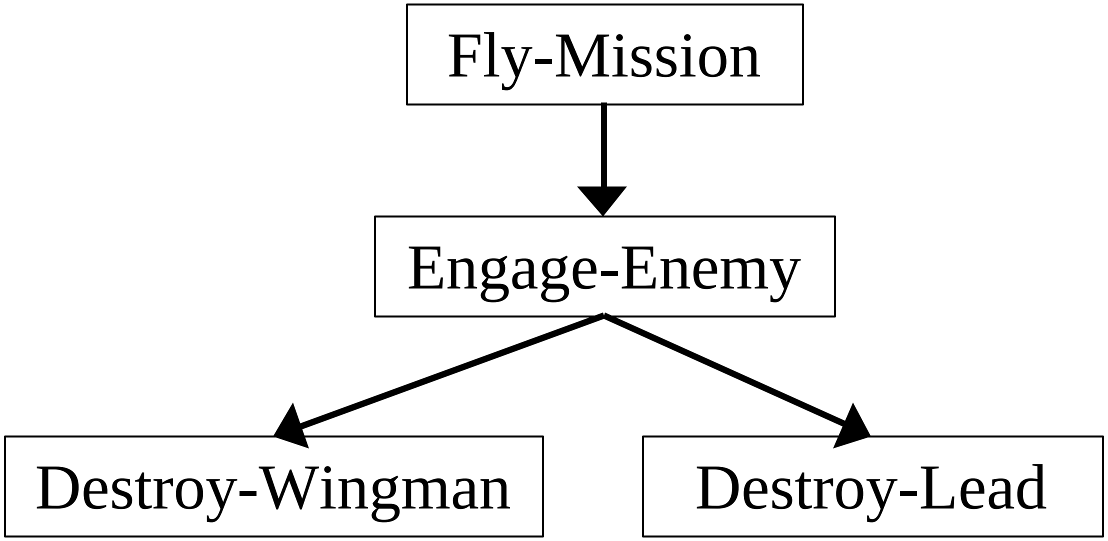


Defining Boundaries

- How can we construct a representation of an agent's aggregate behavior?
- How can we easily partition the behavior space?
- How can we identify how these partitions overlap?

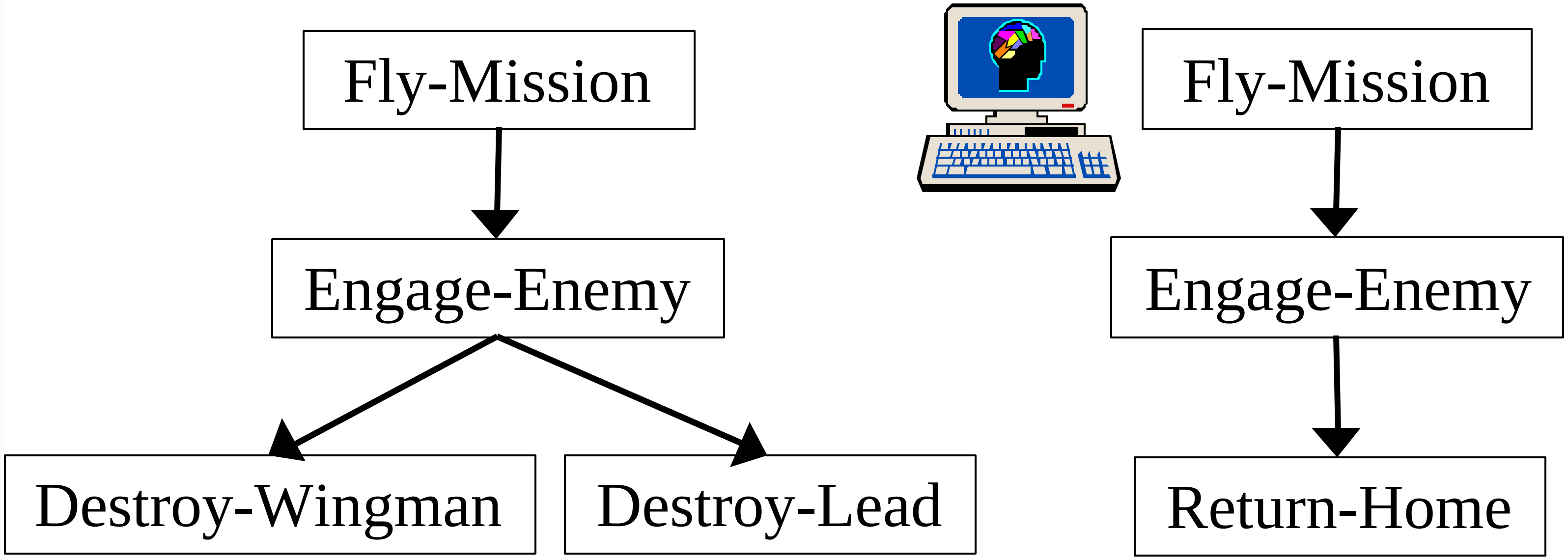


Enter the Goal Hierarchy



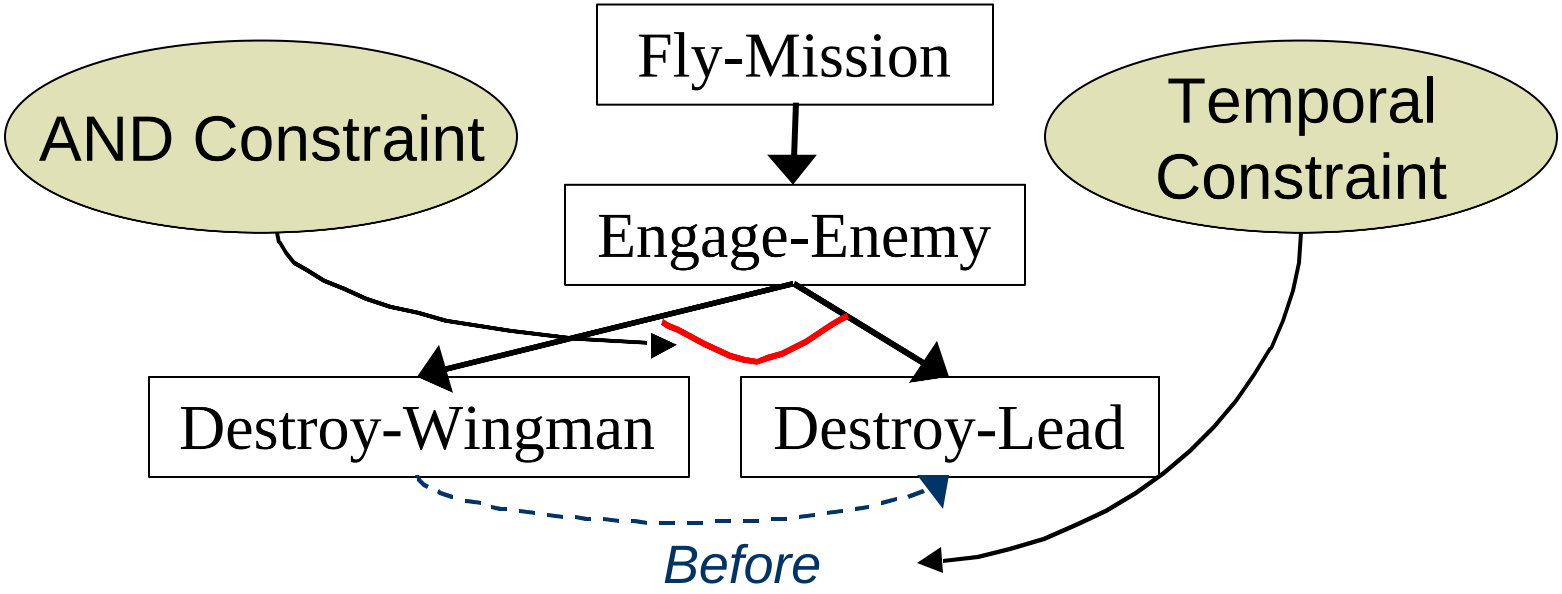
- Can be viewed as an outline of behavior
- Identifies relationship between goals, subgoals and actions
- Represents many potential behaviors

Goal Hierarchy As a Classifier



- Can be used to identify failure to meet minimal specifications

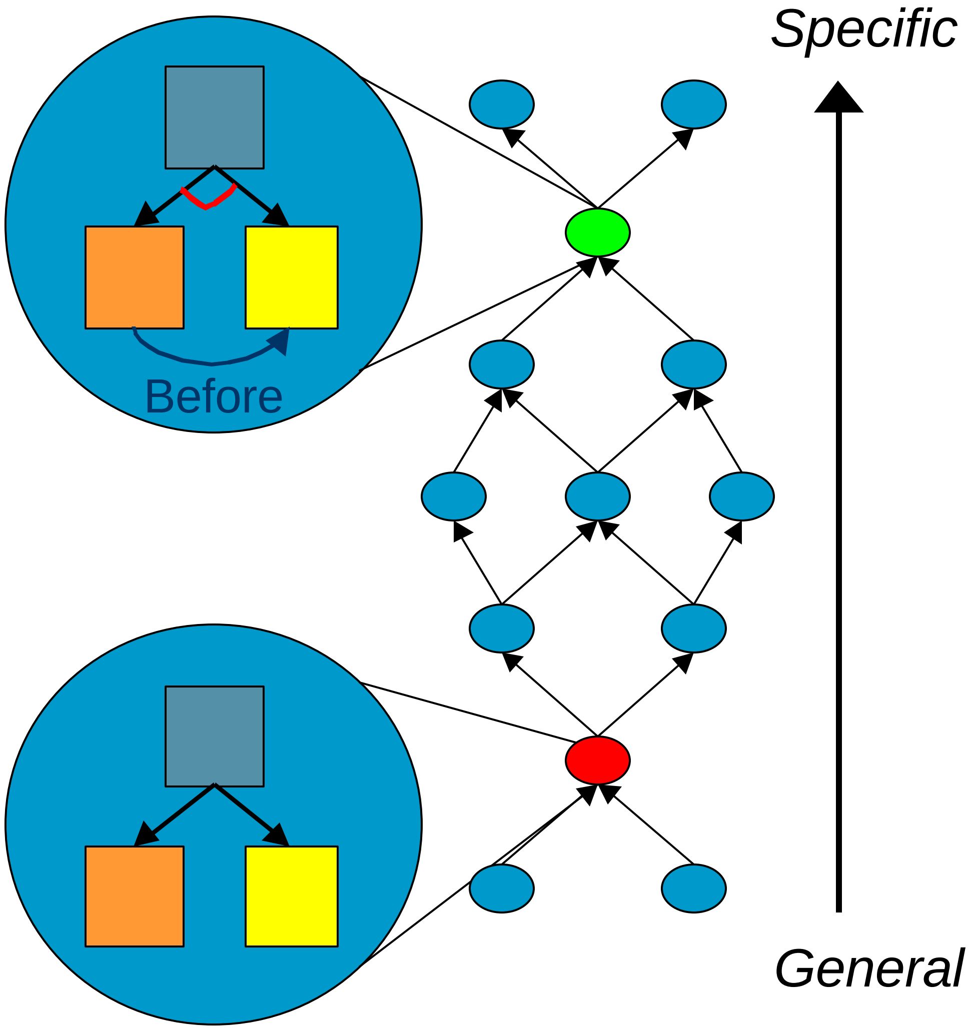
Constrained Goal Hierarchy



- Constraints reduce degrees of freedom
- Create specializations of original hierarchy
- Can also be used to classify behavior

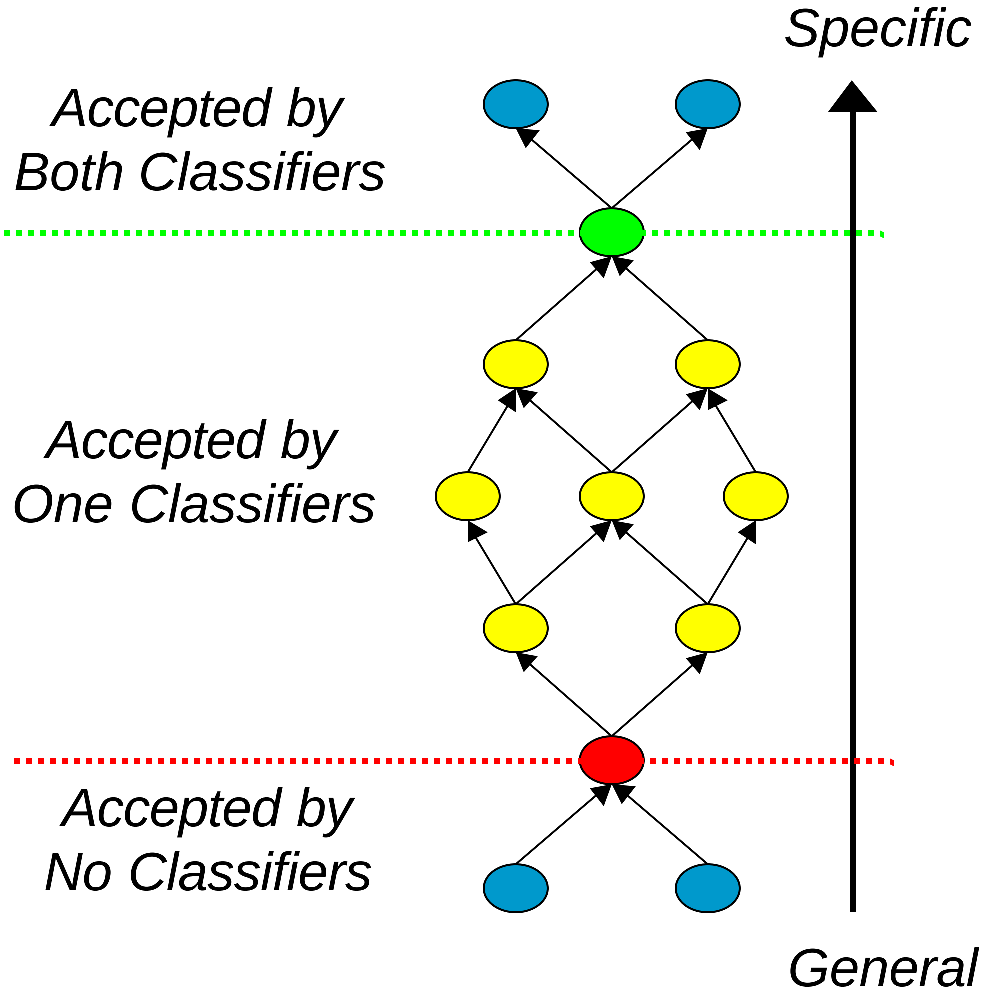
Hierarchies As Partitions

- Constraints impose an ordering on the behavior space



Hierarchies As Partitions

- Partitions space into three regions
- Paves way for a version-space approach to error detection



Specific

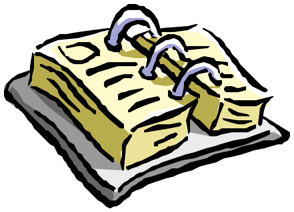
Accepted by Both Classifiers

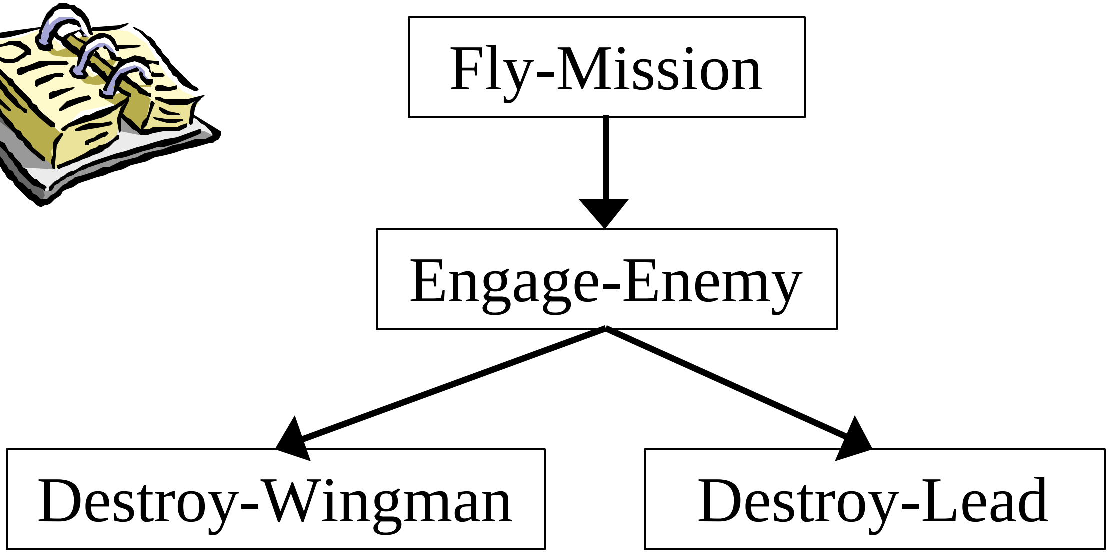
Accepted by One Classifiers

Accepted by No Classifiers

General

Putting Hierarchies to Work





Fly-Mission

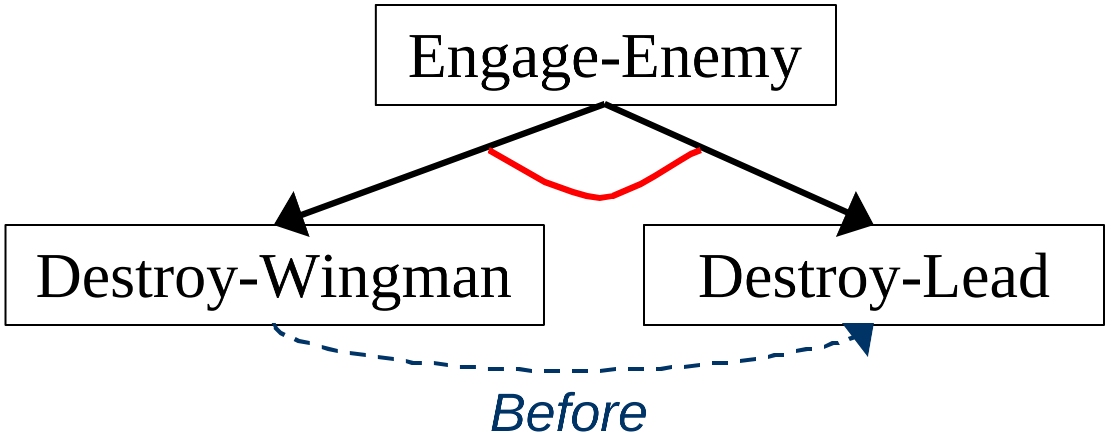
Engage-Enemy

Destroy-Wingman

Destroy-Lead

- Design begins with a specification
- Specification yields a basic goal hierarchy

Observe Expert Behavior



Engage-Enemy

Destroy-Wingman

Destroy-Lead

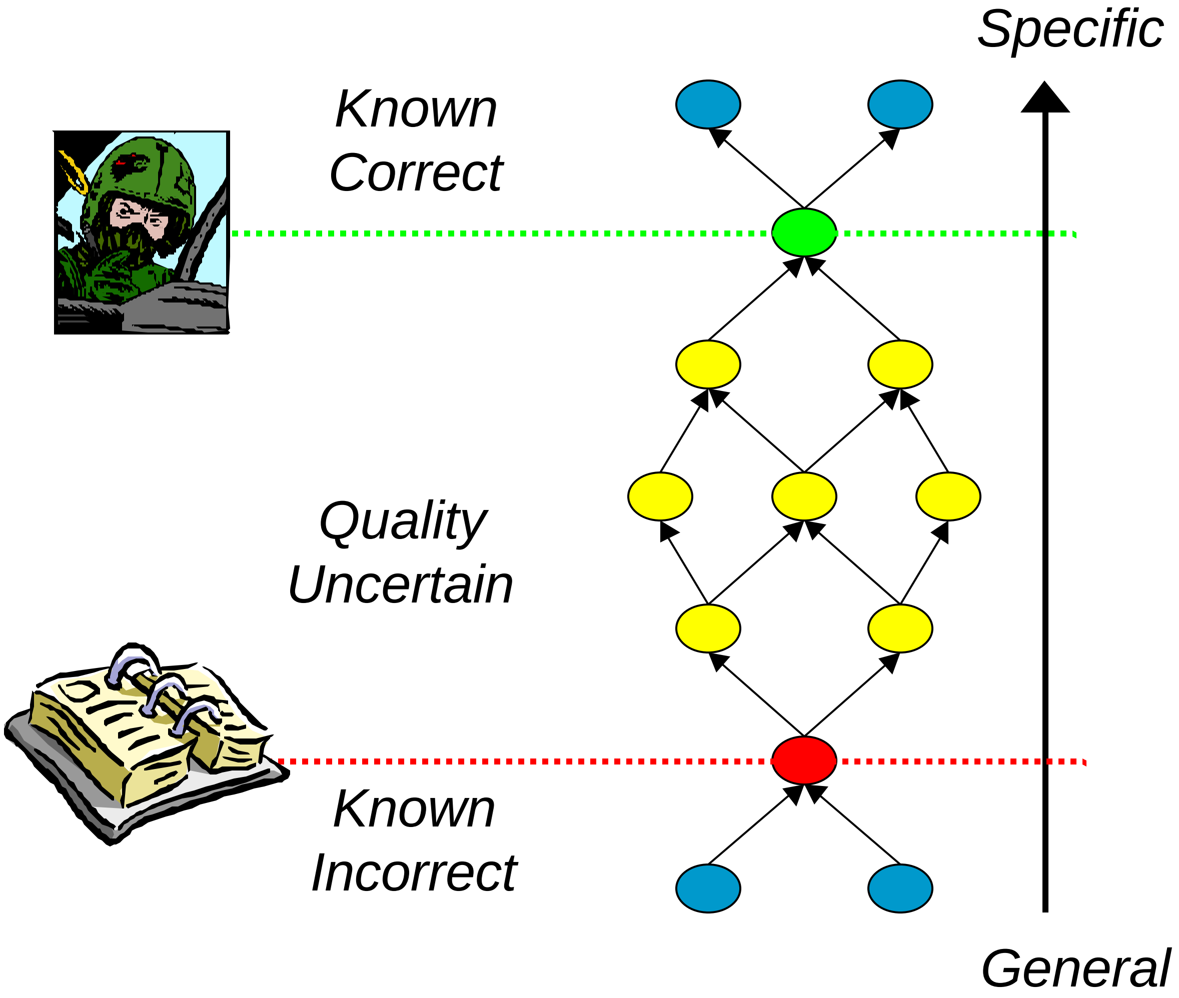
Before

Fly-Mission
Engage-Enemy
Destroy-Wingman

Fly-Mission
Engage-Enemy
Destroy-Lead

- Construct a maximally specific hierarchy covering the observations

Partition Behavior Space



Specific

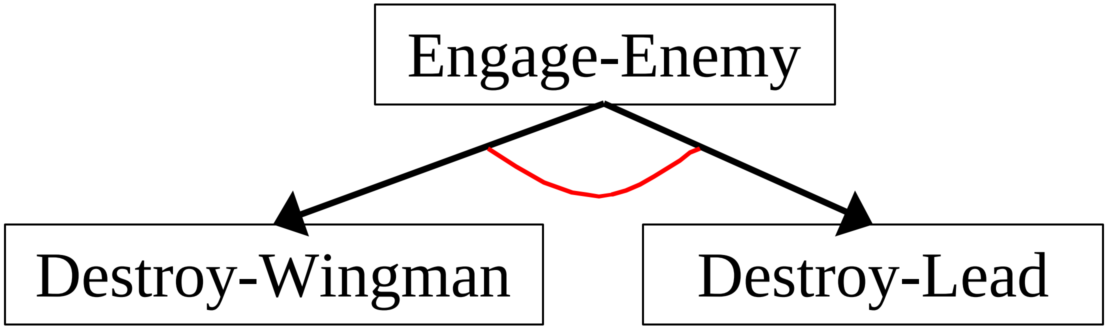
Known Correct

Quality Uncertain

Known Incorrect

General

Observe Agent Behavior

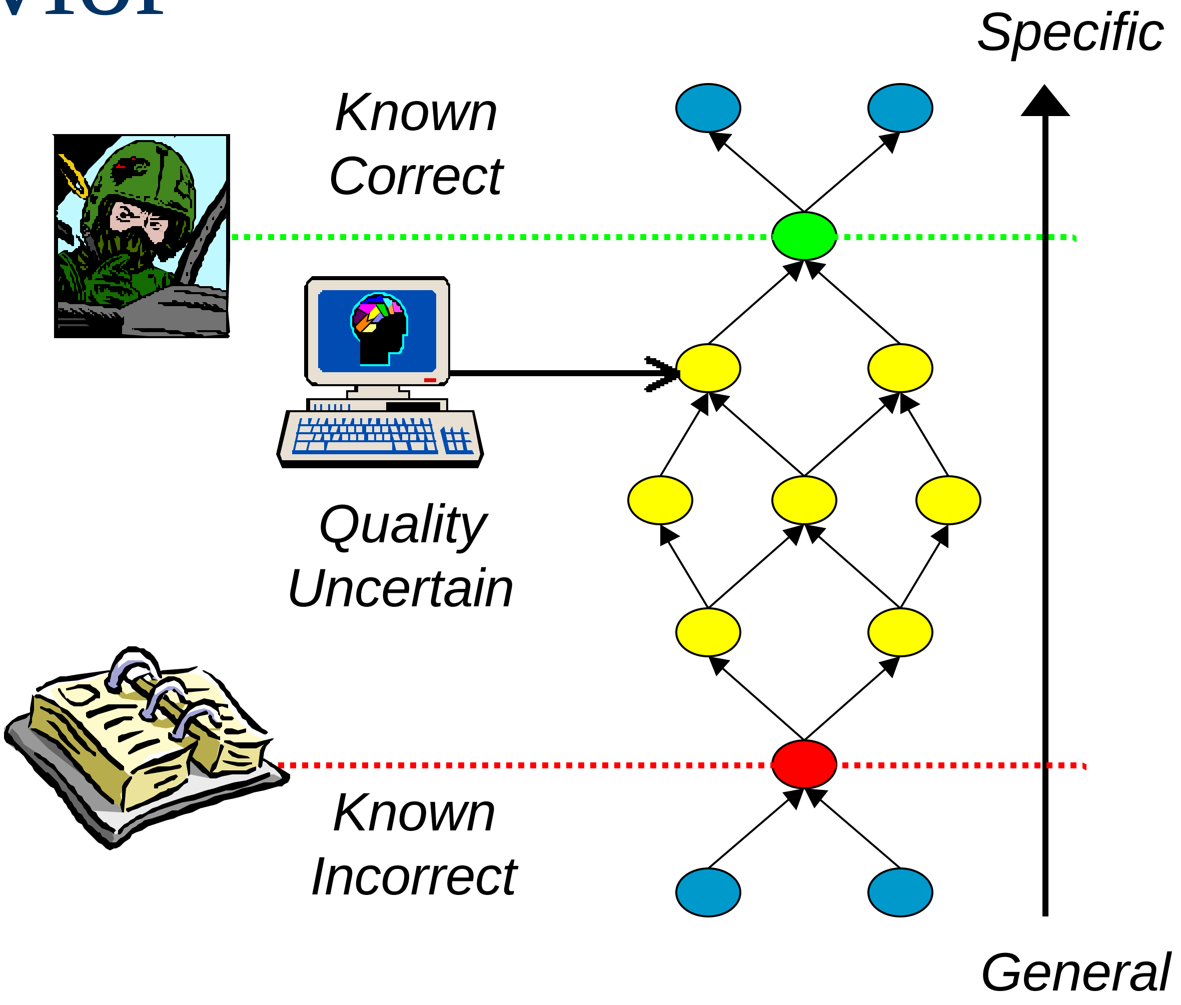


Engage-Enemy *Engage-Enemy*
Destroy-Wingman *Destroy-Lead*

Engage-Enemy *Engage-Enemy*
Destroy-Lead *Destroy-Wingman*

- Construct a maximally specific hierarchy covering the observations

Identify Quality of Agent Behavior



Specific

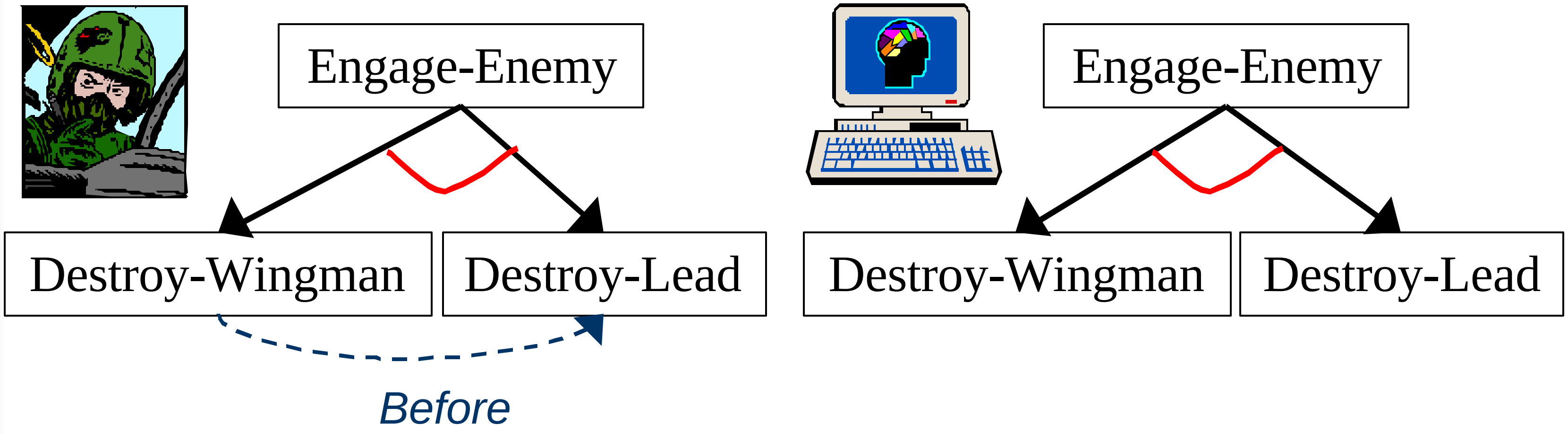
Known Correct

Quality Uncertain

Known Incorrect

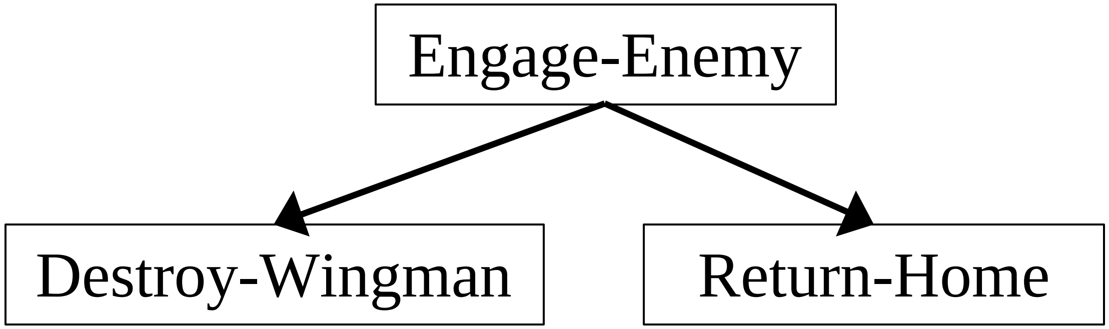
General

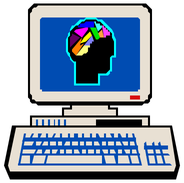
Identify Quality of Agent Behavior



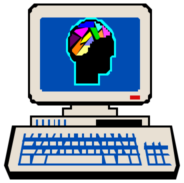
- Agent behavior is not a specialization of Expert behavior
- Looking at behaviors encapsulated by hierarchy gives details of differences
  - Only 50% of agent behavior is questionable

Identifying a Failure





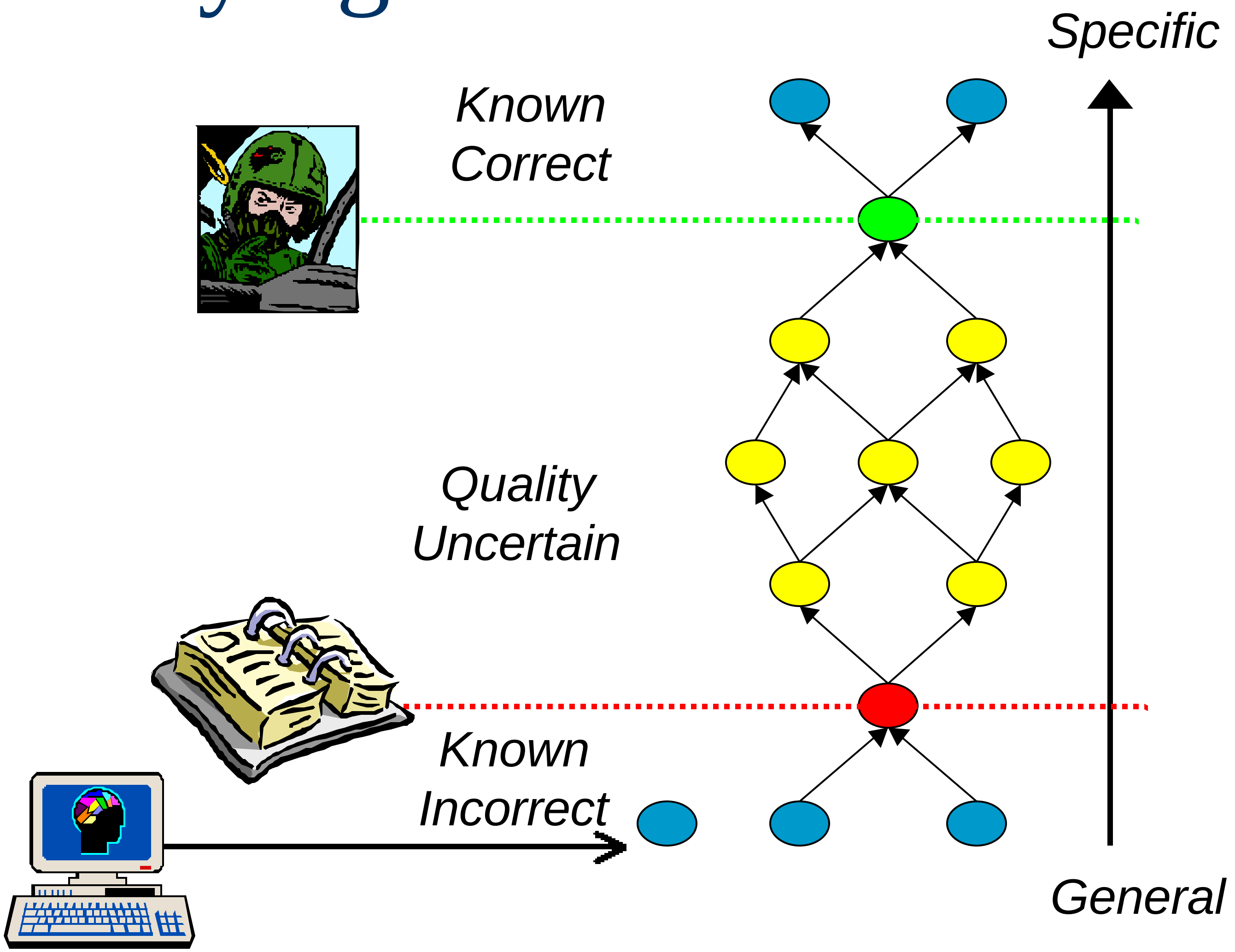
Engage-Enemy
    Destroy-Wingman



Engage-Enemy
    Return-Home

- Again, build the maximally specific hierarchy representing the observations

Identifying a Failure



Specific

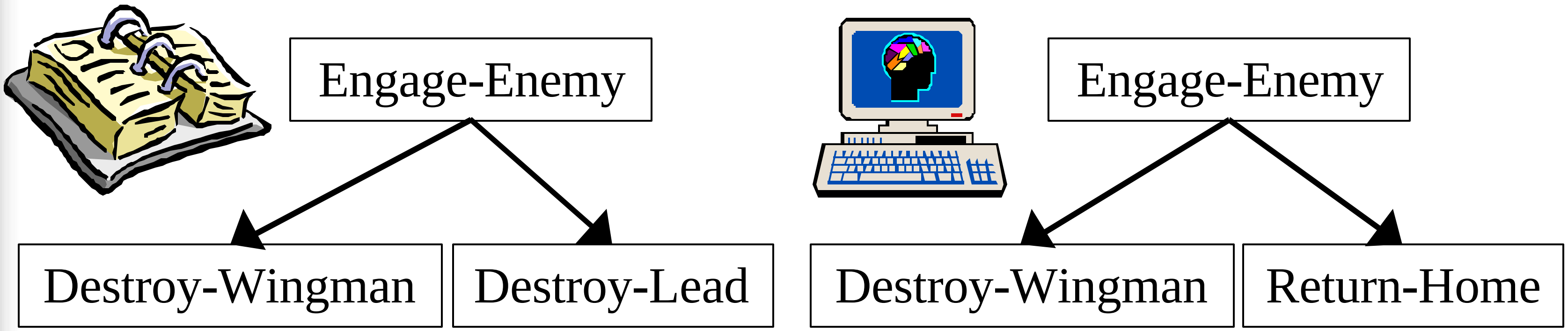
Known Correct

Quality Uncertain

Known Incorrect

General

Identifying a Failure



Engage-Enemy

Destroy-Wingman

Destroy-Lead

Engage-Enemy

Destroy-Wingman

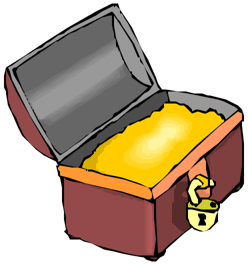
Return-Home

- Agent behavior *is not* a specialization of base goal hierarchy
- Looking at behaviors encapsulated by hierarchy gives details of differences
  - A portion of agent behavior *may be* correct

Current Status & Future Work

- Currently…
  - Implementation is 90%
  - Can build maximally specific hierarchies for a given set of observations
  - Testing will begin soon…
- Future…
  - Ability to use more knowledge to set boundary on *known errors*

Nuggets



- Can be viewed as a generalization of initial approach
- Generates new potential methods for detecting errors
- Can use method to validate project specification as well as agent behavior
- Provides a basis for:
  - efficiently dealing with aggregate behavior
  - determining when validation is complete
  - determining number of observations required for validation

Coal

- Requires induction
  - – May make invalid inductions under certain conditions
- Requires goal annotations from expert