Differential Assessment of Black-Box AI Systems

Rashmeet Kaur Nayyar*, Pulkit Verma*, Siddharth Srivastava, Arizona State University

How can we efficiently assess an AI that has changed, without expensive reassessment from scratch?

Introduction

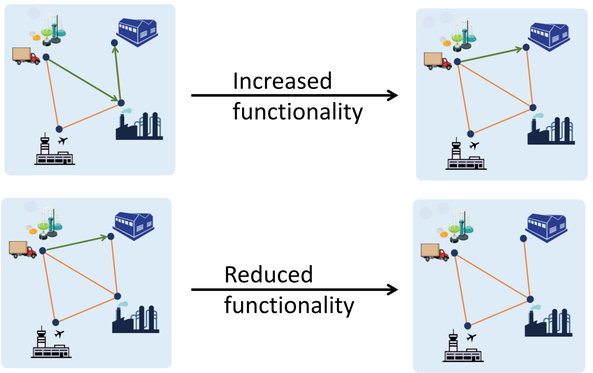
Objective: Assess and learn model of true functionality of an adaptive black-box AI agent that has drifted from its previously known functionality.

Key technical challenges:

- How to identify what has changed?
- How to identify how has it changed?

What Can Change?

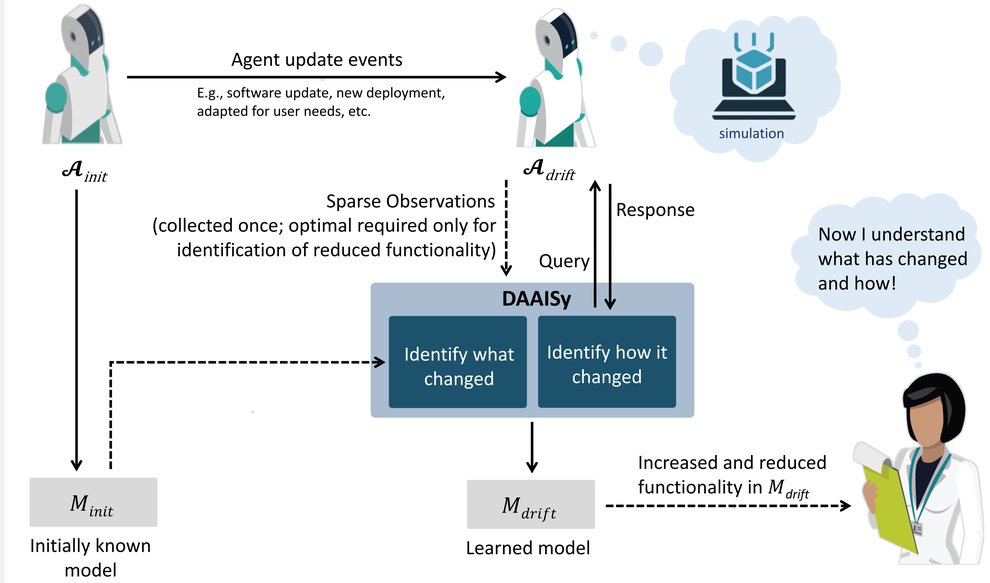
- A predicate in an action's precondition or effect (also called a PAL-tuple) is either present in the positive (+)/negative (-) form or is absent (ϕ).
- Possible mode changes: + → -/ϕ, - → +/ϕ, ϕ → +/-



Identify Increased Functionality

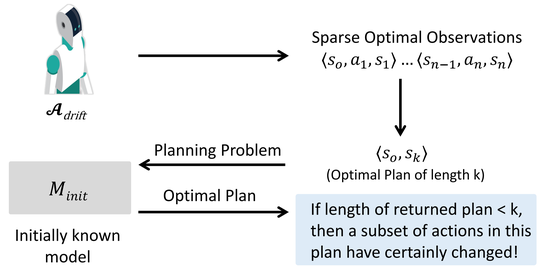
- Infer using available observations and the previously known model.
- If all possible modes for a PAL-tuple that are consistent with the observations are inconsistent with the previously known model, it has changed!

Differential Assessment Of AI Systems (DAAISy)



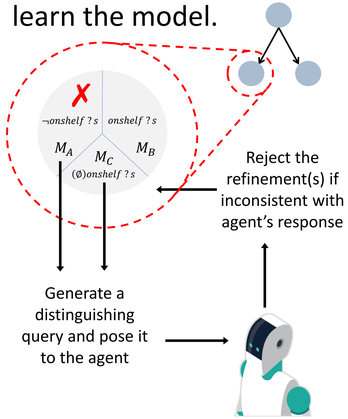
Identify Reduced Functionality

- Cannot infer directly as observations of reduced applicability are almost never available!
- Placing the agent in an optimal mode resolves this challenge.

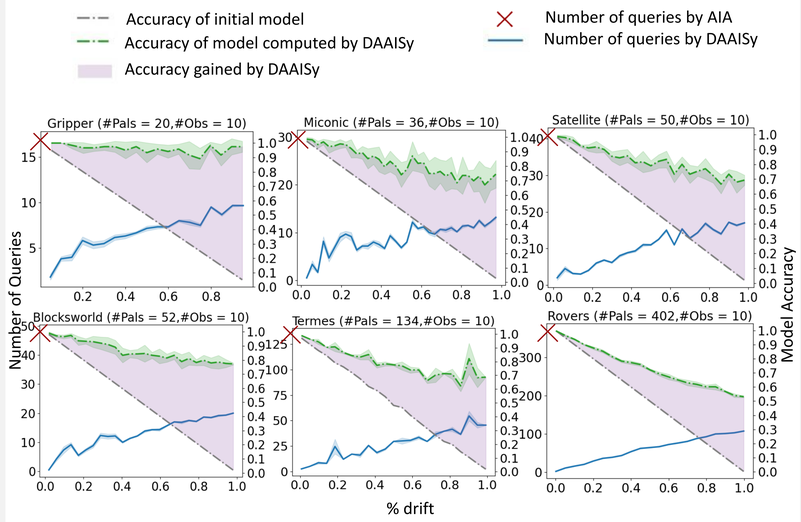


Identify The Change

Selectively query the agent and use the responses to learn the model.



Results



Domain	#Max PAL-tuples	Relative #Queries*
Gripper	20	0.43
Miconic	36	0.24
Satellite	50	0.26
Blocksworld	52	0.29
Termes	134	0.23
Rovers	402	0.19

*#Queries by DAAISy/#Queries by reassessing from scratch (to achieve the same level of accuracy for 50% drifted models)

Theorem†: Models learned by DAAISy are guaranteed to be consistent with the observations of the agent after the updated events and the responses of the queries posed to it.

†Theorem 1 in the paper.

Conclusion

DAAISy efficiently learns highly accurate models of an agent's functionality, using a significantly lower number of queries as opposed to relearning from scratch.

bit.ly/3so0nrx