Helping Humans and Agents Avoid Undesirable Consequences with Models of Intervention

Sachini Weerawardhana

Dissertation Defense
Advisor: Prof. Darrell Whitley

October 4, 2021

Agenda

- **Introduction**
- Motivational study from cyber-security
- Intervention models
  - Intervention by recognizing actions that enable multiple undesirable consequences
  - Intervention as planning
  - Human-aware Intervention
- Intervention recovery model
  - The Interactive Human-aware Intervention

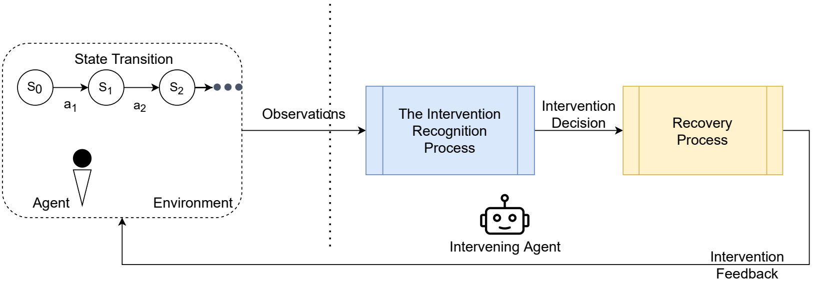
The Intervention Problem

- A human user (or an agent) is doing something online that may have an undesirable outcome that it can not recognize
- Two sub-problems:
  - **Intervention Recognition**: Identify what the user is doing is "bad"
  - **Intervention Recovery**: Help the user decide what to do next
- **Use automated planning as a framework to model and understand**:
  - Human user behavior in cyber-security
  - Problem solving in the Rush Hour puzzle

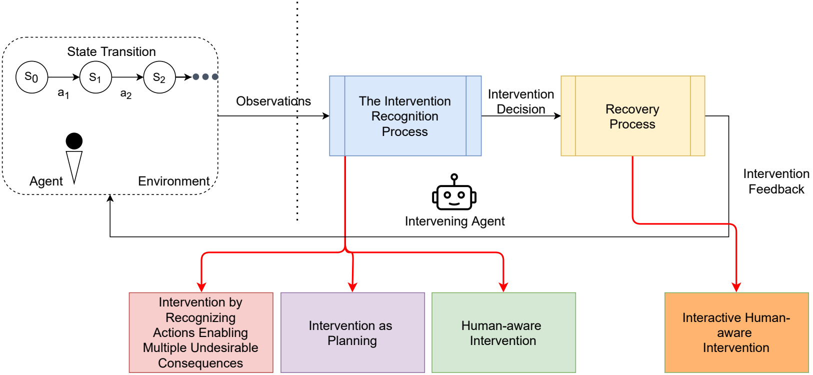
Why is Intervention Important?

- The actor is working in an unfamiliar environment
- Examples:
  - Learning to use a new software application
  - Use a computer having hidden security vulnerabilities
- Intervention is a utility for online assistive agents

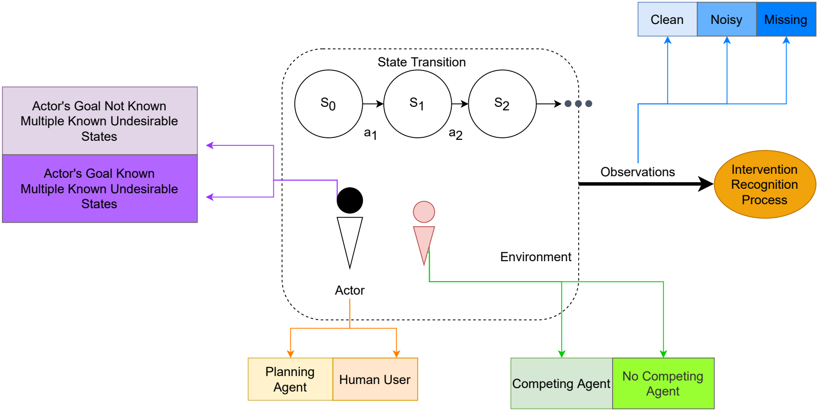
The Intervention Framework



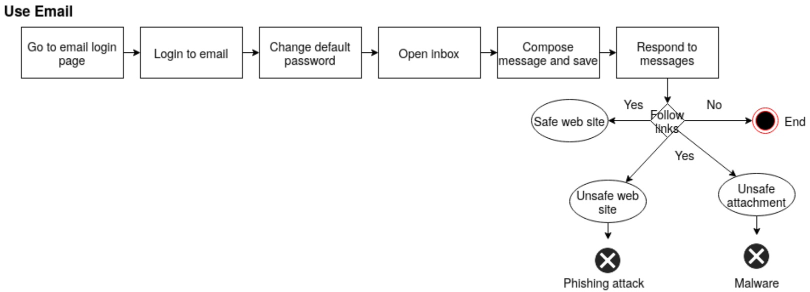
Thesis Outline



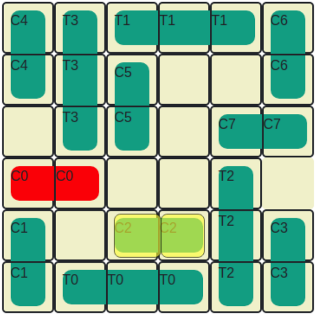
The Intervention Problem Dimensions



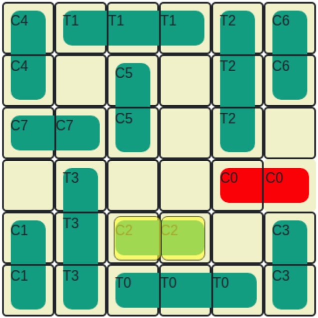
Undesirable States: Cyber-security Domain



Undesirable States: The Rush Hour Domain



(a) Initial game state



(b) End game state

The Gap in Existing Work

- The goal/plan recognition problem
  - The intervening agent can infer the goal/plan of the actor using the observations as evidence
  - May not work if the actor's likely goals are too similar
- We use machine learning to learn the differences between safe and unsafe plan suffixes.
- The Intervention Problem:
  - Online
  - Actors may have different views of the domain
  - Intervene on time but allow the actor to pursue his own goal
  - Intervention recovery

Research Questions

- **1: What are the salient characteristics for deciding when to intervene**?
  - the actor's goals are known (Rush Hour)
  - the actor's goals are not known (Cyber-security)
- **2: How to help task continuation following intervention**?
  - Probe the search space and inform the actor about the probes
- **3: How to design tools to study intervention with human user participation**?
  - Planning Domain Definition Language (PDDL) models for the two domains
  - Software tools to study human users in-situ.

Agenda

- Introduction
- **Motivational study from cyber-security**
- Intervention models
  - Intervention by recognizing actions that enable multiple undesirable consequences
  - Intervention as planning
  - Human-aware Intervention
- Intervention recovery model
  - The Interactive Human-aware Intervention

Home Computer User Behavior in Questionable Security Situations - A Motivational Study

- Objectives:
  - To capture actions taken by users when asked to perform tasks that provided “opportunities” to trigger security vulnerabilities
  - To assess consistency in users answers and actions.
- Contributions:
  - Cyber-security planning domain model for home computer security vulnerabilities
  - Software framework to support home computer user security/privacy studies

Key Findings

- Usage patterns
  - 9+ years of experience (66%)
  - Owned multiple devices (82%)
  - 67% found it challenging to identify harmful actions and take safety steps
- Using software
  - Users incorrectly assess self-efficacy for antivirus software installation
  - Users correctly assess self-efficacy for choosing legitimate software
- Twitter/Email safety
  - Users incorrectly assess self-efficacy for recognizing phishing attempts on Twitter/email
  - Users incorrectly assess self-efficacy for identifying malicious attachments

Key Findings

- Recognizing system cues for safety
  - Web page content helps users select safe software to download
  - HTTPS/padlock icon helps users recognize safe web pages to download software
  - HTTPS/padlock icon does not help users recognize phishing links on Twitter
  - HTTPS/padlock icon does not help users recognize phishing links while using email

Designing Intervention for Cyber-security Domain

- Home computer users might not think about the post-conditions of actions while performing common tasks
- Home computer users might not perceive preconditions that exist in the state (e.g., padlock icons/HTTPS)
- Intervention solution needs to address:
  - Users do not complete tasks methodically (e.g., repeated actions, skips) ⇒ `noise in observations`
  - Intervention must be decided as and when observations become available ⇒ `online operation`
  - Need to evaluate proximity to an exploit regardless of the user's goal ⇒ `user's desirable goal is unknown`

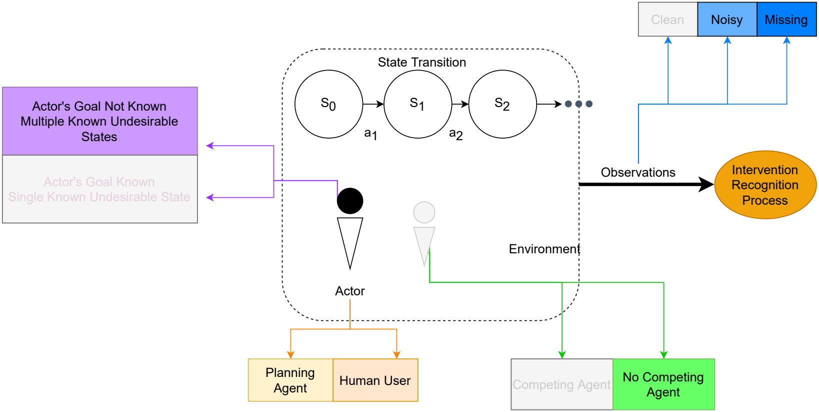
Agenda

- Introduction
- Motivational study from cyber-security
- **Intervention models**
  - **Intervention by recognizing actions that enable multiple undesirable consequences**
  - Intervention as planning
  - Human-aware Intervention
- Intervention recovery model
  - The Interactive Human-aware Intervention

Intervention by Recognizing Actions That Enable Multiple Undesirable Consequences

- Need to identify an action that causes the most damage and least interferes with the user's needs
- Contribution:
  - Undesirable Consequences Recognition Function
  - Domain-independent metrics to measure the importance of an observed action towards contributing to multiple undesirable states

Intervention Problem Dimensions



Salient Characteristics for Deciding to Intervene

- Certainty (C)
  - How many plans contained the action over the number of sampled plans
  - Highlight frequently occurring actions in plans as important
- Timeliness (T)
  - Maximum normalized steps remaining in the sampled plans
  - Quantifies how soon the undesirable state may occur
- Desirability (D)
  - Number of times the action appears in the sampled plans over the sum of actions in the sampled plans
  - Separate common harmless actions that further the user's actual goal from harmful actions to be avoided
  - Negative metric

Undesirable Consequences Recognition Function

- Critical Trigger Action is an observed action that maximizes $V(a)$:

$$V(a) = \alpha_1 * Certainty(a|\Pi_U) + \ \alpha_2 * Timeliness(a|\Pi_U) \\ - \ \alpha_3 * Desirability(a|\Pi_U)$$

- a candidate action from the sampled undesirable plans plans
- Π_U sampled undesirable plans
- $(\alpha_1, \alpha_2, \alpha_3)$ metric weight assignments

Experiments

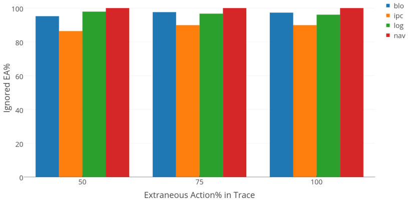
- Planning domain from the home computer cyber-security study
- Four benchmark planning domains (Blocks words, navigator, ipc-grid+, logistics)
- Four undesirable states for each domain
- Observation traces of actions
  - Activity logs (n=61) captured during the human subject study
  - Synthetic traces generated with controlled levels of noise and missing actions for benchmark domains
- Metric weights
  - 7 classes of discrete weight assignments for the three metrics
  - (1,0,0), (0,1,0), (0,0,1), (.33,.33,.33), (.5,.5,0), (.5,0,.5), (0,.5,.5)

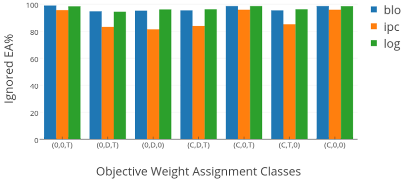
Key Findings - Cyber-security Domain

- For each decision cycle, the selected critical trigger action is correct if it is in a ground truth undesirable plan
  - Mean accuracy = 59.53% (SD=30.79) across the 7 metric weight assignment classes
- Effect of metric weights on accuracy is significant ($F = 40866, p << 0$)
- Highest accuracy (mean=95.59%,SD=2.13) for two classes
  - Equal weights for C, T, D
  - Equal weights for C, T ignoring D
- Dominant metrics: **Certainty** and **Timeliness**

Key Findings - Benchmark Synthetic Domains

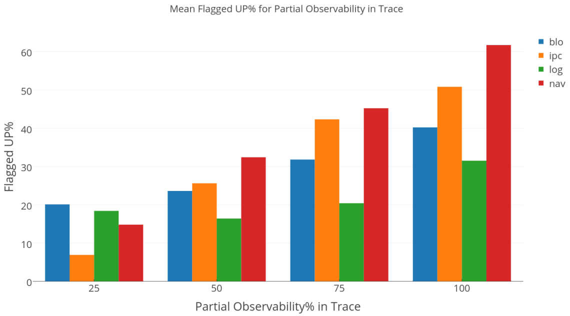
- Percentage of extraneous actions not flagged as critical





- Metric weights significantly influence ignoring extraneous actions
  - Dominant metrics: **Certainty**, **Desirability**

- Percentage of ground truth undesirable actions flagged as critical
- Low rates indicate that other factors in addition to the three metrics may influence flagging undesirable actions
- Metric weights significantly influence flagging undesirable actions
  - Dominant metrics: **Timeliness**



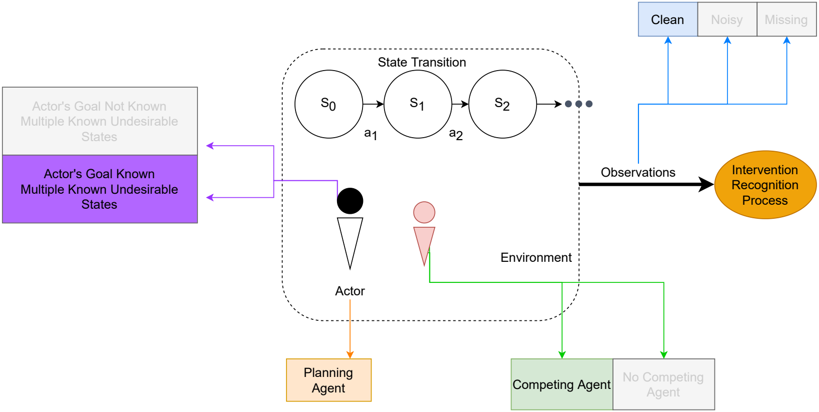
Agenda

- Introduction
- Motivational study from cyber-security
- **Intervention models**
  - Intervention by recognizing actions that enable multiple undesirable consequences
  - **Intervention as planning**
  - Human-aware Intervention
- Intervention recovery model
  - The Interactive Human-aware Intervention

Intervention as Planning

- Combines machine learning and automated planning
- Solution is based on plan suffix analysis
- Identify different characteristics between solutions obtained from an automated planner that contain undesirable actions and solutions that do not
- Hidden effects in the environment
  - a competitor using a hidden object to subvert the actor's goal
  - a pothole the actor can not see

Intervention as Planning: Problem Dimensions

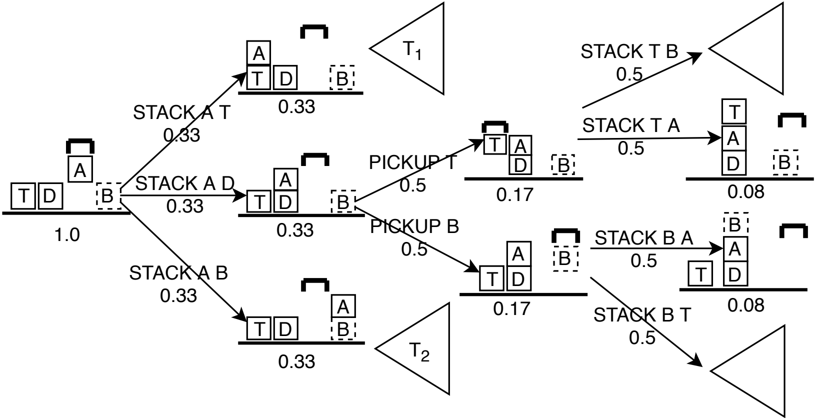


Learning to Intervene

- Two feature sets:
  - Metrics from **The Intervention Graph**
  - Plan distance measures from the sampled plans
- Use the feature sets to train classifiers to recognize actions that should be flagged for intervention
  - Naive Bayes, K-nearest neighbor, Decision tree, Logistic regression

The Intervention Graph

- Produce the intervention graph from current state (root) to G_1 (BAD) and G_0 (TAD) (leaves)



Intervention Graph Features

- **Risk** - Posterior probability of reaching G_1, when the user is trying to reach G_0
- **Desirability** - Posterior probability of reaching G_0, without passing G_1
- **Distance to G_0** - Mean number of edges between the root of the tree and G_0, which doesn't pass through G_1
- **Distance to G_1** - Mean number of edges between the root of the tree and G_1
- **Active attack landmarks%** - From the total number of predicates that must be true in any valid solution to the planning problem $\langle M, G_1 \rangle$, how many are true in the current state?

Plan Distance Measures From Sampled Plans

- Instead of computing exact distances and probabilities **compute an estimated proximity** to G_0 and G_1
- Use an automated planner to find two sets of solutions for $\langle M, G_0 \rangle$ and $\langle M, G_1 \rangle$
- Compute a **reference plan** $= \{observations + \pi_*\}$
- Compute the plan distances between the **reference plan** and the plans in $\langle M, G_0 \rangle$ and $\langle M, G_1 \rangle$

Classifier Performance

Reporting $F - score = \frac{TP}{TP+1/2(FP+FN)}$
Matthews Correlation Coefficient (MCC)

Domain	Logistic Regression						K-Nearest					
	Test Set 1		Test Set 2		Test Set 3		Test Set 1		Test Set 2		Test Set 3	
	F-score	MCC	F-score	MCC	F-score	MCC	F-score	MCC	F-score	MCC	F-score	MCC
Intervention Graph Method												
Blocks-1	1	1	1	1	1	1	1	1	1	1	1	1
Blocks-2	1	1	1	1	1	1	1	1	1	1	1	1
EasyIPC	.88	.87	.88	.87	.86	.86	1	1	1	1	1	1
Ferry	1	1	1	1	1	1	1	1	1	1	1	1
Navigator	1	1	1	1	.99	.99	1	1	.96	.96	.99	.99
Plan Space Sampling Method												
Blocks-1	.25	.33	.25	.33	.25	.33	1	1	1	1	1	1
Blocks-2	1	1	1	1	1	1	1	1	1	1	1	1
EasyIPC	.64	.63	.46	.44	.67	.66	.05	-.04	.04	-.03	.05	-.02
Ferry	.31	.32	.23	.22	1	1	.33	.40	.13	.15	.81	.82
Navigator	.60	.59	.98	.94	.97	.97	.61	.65	1	1	1	1

Intervention Using Existing Goal Recognition Algorithms

RG (LAMA) - Probabilistic goal recognition using a satisificing planner

Domain	Test Set 1		Test Set 2		Test Set 3	
	F-score	MCC	F-score	MCC	F-score	Mcc
Blocks-1	.38	.45	.43	.49	.40	.47
Blocks-2	1	1	.9	.9	1	1
EasyIPC	.13	.05	.21	.17	.23	.19
Ferry	.17	.18	.22	.23	.15	.17

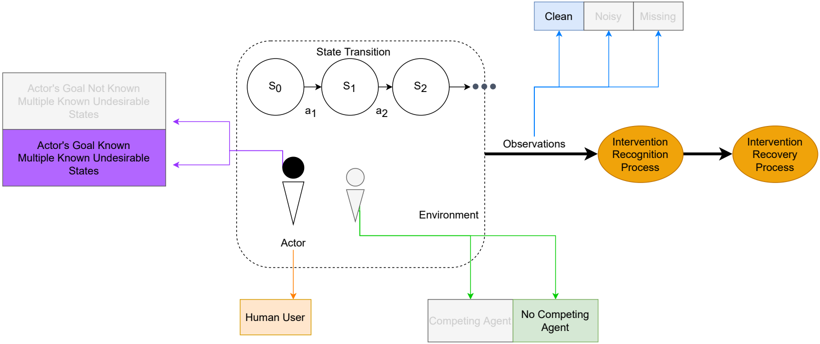
Agenda

- Introduction
- Motivational study from cyber-security
- **Intervention models**
  - Intervention by recognizing actions that enable multiple undesirable consequences
  - Intervention as planning
  - **Human-aware Intervention**
- Intervention recovery model
  - The Interactive Human-aware Intervention

Human-aware Intervention

- The actor is a human user. We cannot approximate plan suffixes using automated planners
- Observe how human users solve Rush Hour puzzles
- What did the users who did not move the forbidden vehicle do differently than those who moved the forbidden vehicle?

Human-aware Intervention - Problem Dimensions



Human-aware Intervention - Behavior Study

- In Web-based puzzle simulator app, subjects solve one randomly assigned Rush Hour puzzle.
- Subjects are told the puzzle has one forbidden vehicle that need not be moved
- No alerts if the forbidden vehicle is moved
- Post-study survey of demographics and puzzle solving habits
- 136 university students from different departments

Key Findings

- Huge enthusiasm for puzzle solving tasks (78%)
- 49% moved the forbidden vehicle
- Behavior patterns in unsafe solutions
  - Moving the same car back and forth in succession (do/undo)
  - Making moves that clears space around the forbidden vehicle
  - Lengthy solution : **statistically significant positive correlation** between the solution length and the number of times the forbidden vehicle was moved

Learning Human-aware Intervention

- Game state based features
  - number of times a move increased the number of cars blocking the goal car's path
  - number of times a move freed up empty spaces around the forbidden vehicle
  - number of times the number of empty spaces around the forbidden vehicle blockers increased
  - mean number of empty spaces around the goal and forbidden car blockers
- User action based features
  - number of moves in the user's solution
  - difference of number of moves to the cost optimal solution
  - number of vehicles moved
  - number of times a move was immediately undone

Classifier Performance

- Intervention accuracy while offering three levels of freedom $k = \{1, 2, 3\}$
- 70-30 split for training and test sets
- Classifiers trained with 10-fold cross validation

Classifier	$k = 1$			$k = 2$			$k = 3$		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Decision Tree	0.70	0.90	0.89	0.80	0.95	0.87	0.89	0.81	0.85
KNN	0.89	0.76	0.82	0.86	0.86	0.86	**0.95**	**0.90**	**0.93**
Logistic Regression	**0.91**	**0.95**	**0.93**	**0.87**	**1**	**0.93**	**0.91**	**0.95**	**0.93**
Naive Bayes	0.73	0.90	0.81	0.74	0.86	0.83	0.68	0.90	0.78

- Predict intervention using the Probabilistic Plan Recognition Algorithm (Ramirez and Geffener, 2010)

Goal Priors	$k = 1$			$k = 2$			$k = 3$		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Uniform	0.67	0.56	0.61	0.67	0.56	0.61	0.56	0.67	0.61
$P(\mathrm{u}) = 2\times P(\mathrm{d})$	0.69	0.61	0.65	0.69	0.61	0.65	0.69	0.61	0.65
$P(\mathrm{d}) = 2\times P(\mathrm{u})$	0.67	0.56	0.61	0.67	0.56	0.61	0.67	0.56	0.67

Summary

- Introduced a family of Intervention Problems
  - Intervention for a single actor (planning agent)
  - Intervention for an actor in the presence of a competitor (planning agents)
  - Human-Aware Intervention for a single actor (human user)
- Solutions
  - If the actor is a planning agent - Plan Suffix Analysis
  - If the actor is a human user - Observed History Analysis
- Proposed learning based intervention outperforms existing plan recognition algorithms

Agenda

- Introduction
- Motivational study from cyber-security
- Intervention models
  - Intervention by recognizing actions that enable multiple undesirable consequences
  - Intervention as planning
  - Human-aware Intervention
- **Intervention recovery model**
  - **The Interactive Human-aware Intervention**

Interactive Human-aware Intervention

- Study intervention recovery in a Rush Hour planning task
- The intervening agent helps the user modify the current trajectory of the plan by providing the hints about the search space of the planning problem.
- Hints:
  - The minimum remaining number of moves
  - The next best move
  - The vehicles that must be moved
  - Restart puzzle

Gap in Existing Work

- Improving human-agent collaborations with explanations
- When the intervening agent (has knowledge advantage) does something the human user does not expect, Explainable AI has been used to enable transparency.
- Explain the surprise using different modalities
  - plan visualization techniques to help human users understand the solution produced by an automated planner
  - question answer dialog ("why did you do A?", "why not B?")
- Hints are designed to help the user uncover information about the Rush Hour planning problem

Interactive Human-aware Intervention - Behavior Study

- In Web-based puzzle simulator app, human subjects solve one randomly assigned Rush Hour puzzle assisted by the Human-aware Intervention agent
- Participants were randomly assigned to also watch a help video to learn how to avoid the forbidden vehicle
- Subjects are told the puzzle has one forbidden vehicle and the puzzle can be solved without it
- When forbidden vehicle is moved subjects see an alert message
- Post-study survey of demographics and hint helpfulness rating
- 135 university students from different departments

Key Findings

- Statistically significant positive correlation between the solution length and the number of times the forbidden vehicle was moved
- Most requested hint : **Show the Next Best Move**
  - Slow/medium/fast solvers
  - Three puzzle classes C, E, M

Hint Request Distribution

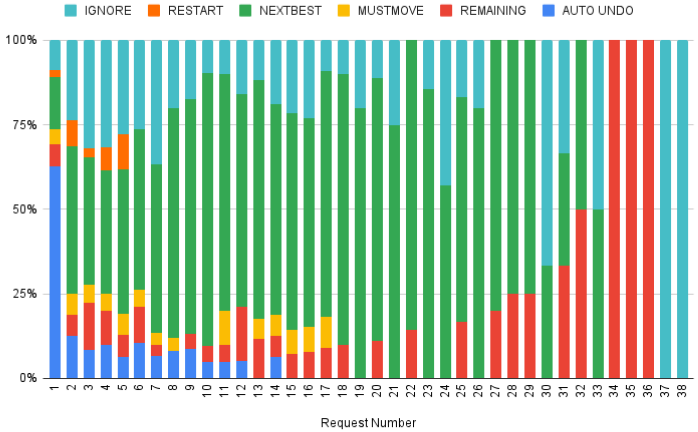


Figure: Percentage split for hints for each request number

Qualitative Evaluation

Table: **H1**: Show the minimum remaining number of moves, **H2**: Show the next best move, **H3**: Show the vehicles that must be moved, **H4**: restart puzzle

			Mean (std. dev.) helpfulness rating			
Type	**Puzzle ID**	**Count**	H1	H2	H3	H4
C	P2	6	1.5 (2.5)	**4.3** (2.6)	3.8 (2.3)	2.0 (2.3)
	P4	11	1.3 (2.1)	**2.3** (2.1)	1.5 (2.2)	0.8 (1.7)
	P6	3	2.0 (1.2)	**3.3** (2.6)	1.7 (2.9)	0
	P8	1	0	**3.0**	**3.0**	1.0
E	P1	4	1.5 (1.3)	**4.3** (1.0)	3.8 (1.9)	2.0 (2.2)
	P3	13	1.3 (1.6)	**2.3** (2.2)	1.5 (1.7)	0.8 (0.9)
	P5	13	1.4 (1.3)	**2.2** (1.9)	**2.2** (2.0)	1.8 (1.9)
	P9	7	0.9 (1.2)	**1.7** (2.2)	1.1 (1.7)	1.3 (2.0)
	P10	8	0.9 (1.0)	**2.8** (2.3)	1.6 (1.5)	**2.8** (2.1)
M	P7	6	2.0 (2.1)	**3.5**(2.1)	2.5 (1.9)	1.8 (1.3)
	P11	3	2.0 (2.6)	**3.3** (2.9)	2.0 (2.6)	**3.3** (2.9)
	P12	1	1.0	**5.0**	**5.0**	1.0
	P13	11	1.2 (1.8)	**1.7** (1.8)	0.9 (1.4)	0.9 (1.6)

Quantitative Evaluation

Evaluation Metrics

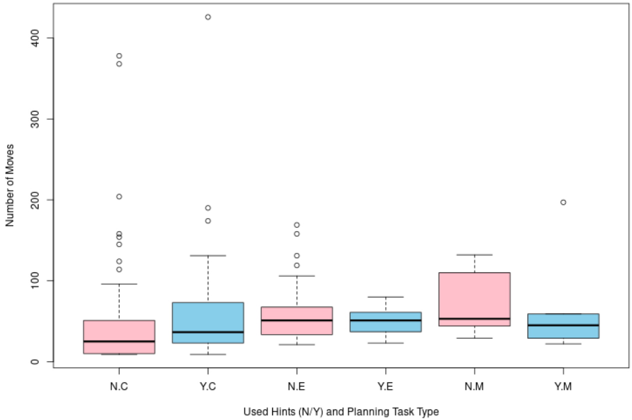
1. The number of moves in the human user's solution
2. The difference from the cost optimal solution
3. The latest time a fact landmark is eventually achieved (landmark achievement)
4. The number of times a fact landmark is lost and regained (landmark regain)

Evaluation Questions

1. Does the Interactive Human-aware Intervention have an effect on the solution length? (Metric 1)
2. Does the Interactive Human-aware Intervention help move the user closer to the optimal solution? (Metric 2, 3, 4)
3. Does seeing a help video affect the solution length? (Metric 1)

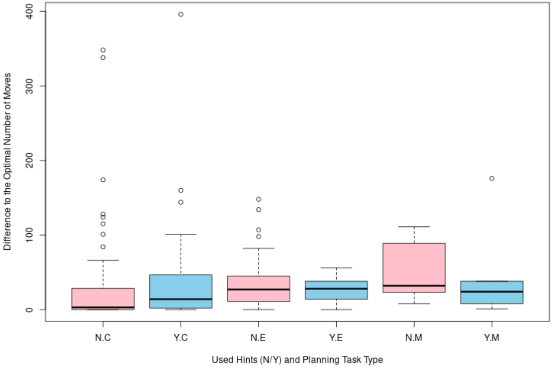
Effect on Solution Length - Metric 1

- Solution length difference between the condition (Y) and the control (N) groups **is not statistically significant**



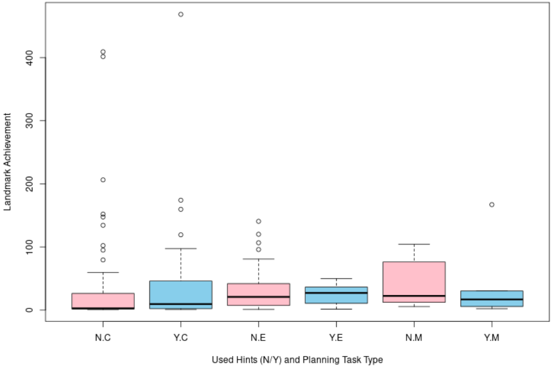
Moving the User's Solution Closer to the Optimal Solution - Metric 2

- Solution length difference to the cost optimal solution between the condition (Y) and the control (N) groups **is not statistically significant**



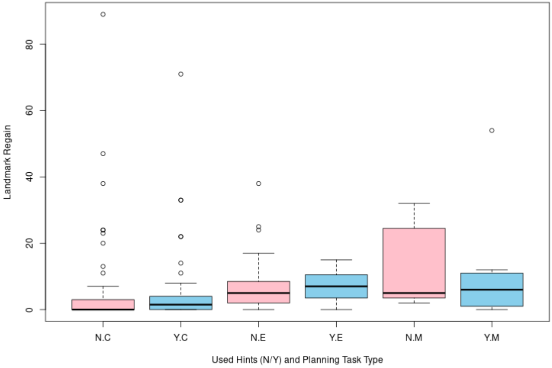
Moving the User's Solution Closer to the Optimal Solution - Metric 3

- The latest times until the landmarks are achieved between the condition (Y) and the control (N) groups **is not statistically significant**



Moving the User's Solution Closer to the Optimal Solution - Metric 4

- The number of times landmarks are lost and regained between the condition (Y) and the control (N) groups **is not statistically significant**



Effect of Using Different Types of Help on Solution Length

- Four help categories
  - participant watched the video and used the Interactive Human-aware Intervention (YY)
  - watched the help video but did not use the Interactive Human-aware Intervention (YN)
  - participant did not watch the help video but used the Interactive Human-aware Intervention (NY)
  - participant used neither type of help (NN)
- Mean number of moves between different help types **is statistically significant**

Planning Task Type	Number of Moves median (mean)			
	YY	YN	NY	NN
C	78 (112)	15 (27)	54 (70)	26 (25)
E	57 (62)	33 (35)	50 (56)	34 (38)
M	41 (55)	29 (32)	32 (36)	26 (28)

Summary

- Qualitatively human users prefer the "**Show the next best move**" hint
- Quantitatively the use of Human-aware Intervention did not statistically significantly change the solution length, the difference to the optimal, landmark achievement time and landmark regain
- Use of different help types significantly affect the solution length
- Need to strike a balance between revealing too much information (i.e., complete solution) and too little information (i.e., next best move) about the planning problem.

Concluding Remarks

- **Contributions:**
  - Intervention is viewed as two sub processes
    - Intervention Recognition (proposed 3 solutions)
    - Intervention Recovery (proposed 1 solution)
  - Solutions combine automated planning and machine learning
  - Evaluated on synthetic domains and realistic data from human subject studies
- **Future Work:**
  - Explore different models of the actors' environment
  - Domain abstraction techniques to support intervention recovery
  - Ensuring longevity of interactive intervention models