Explaining Automated Policies for Sequential Decision Making

February 16, 2010
University of Kentucky, Lexington

University of Waterloo

Presented by Pascal Poupart
University of Waterloo, Canada

Joint work with **Omar Zia Khan** and Jay Black

Sequential decision making

- Fault diagnosis, inventory management (OR)
- Medical diagnosis (health informatics)
- Course selection advising (recommender systems)
- Robotic control
- Web optimization

- Difficult to optimize policy
  - Uncertain action effects
  - Multiple/complex objectives
  - Repeated/sequential decision points

Automated Policy Generation

- Solution:
  - Harness the power of machines
  - Automated policy optimization
- Problem:
  - How can we ensure user trust?
  - How can we verify the correctness of the model and resulting policy?
- Contribution: policy explanation
  - Generic approach to explain the choice of action

Outline

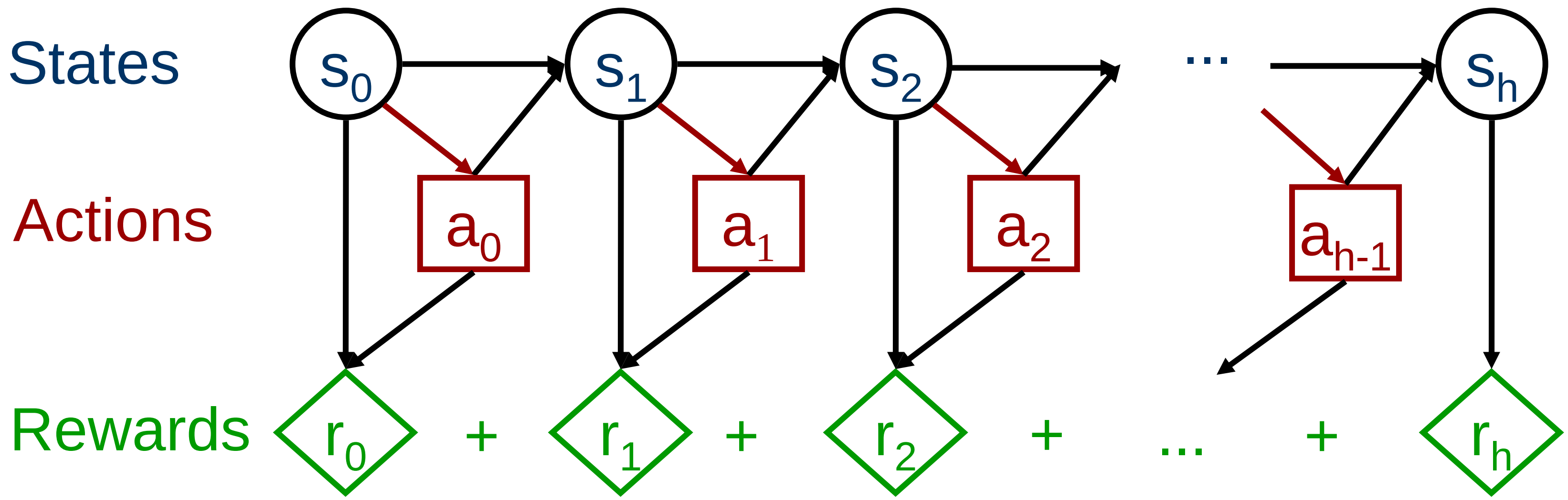
- Background
- Automated policy explanation
- Experiments and Sample Explanations
- User Study
- Conclusion and Future Work

Markov Decision Processes

- General framework for sequential decision making
- Formalized in Operations Research in the 1950s
- Automated policy optimization
- Today: one of the most popular approaches
- But, no generic technique to explain resulting policies

- This talk: automated explanation of MDP policies
  - Generic, problem independent technique
  - Minimal and sufficient explanations

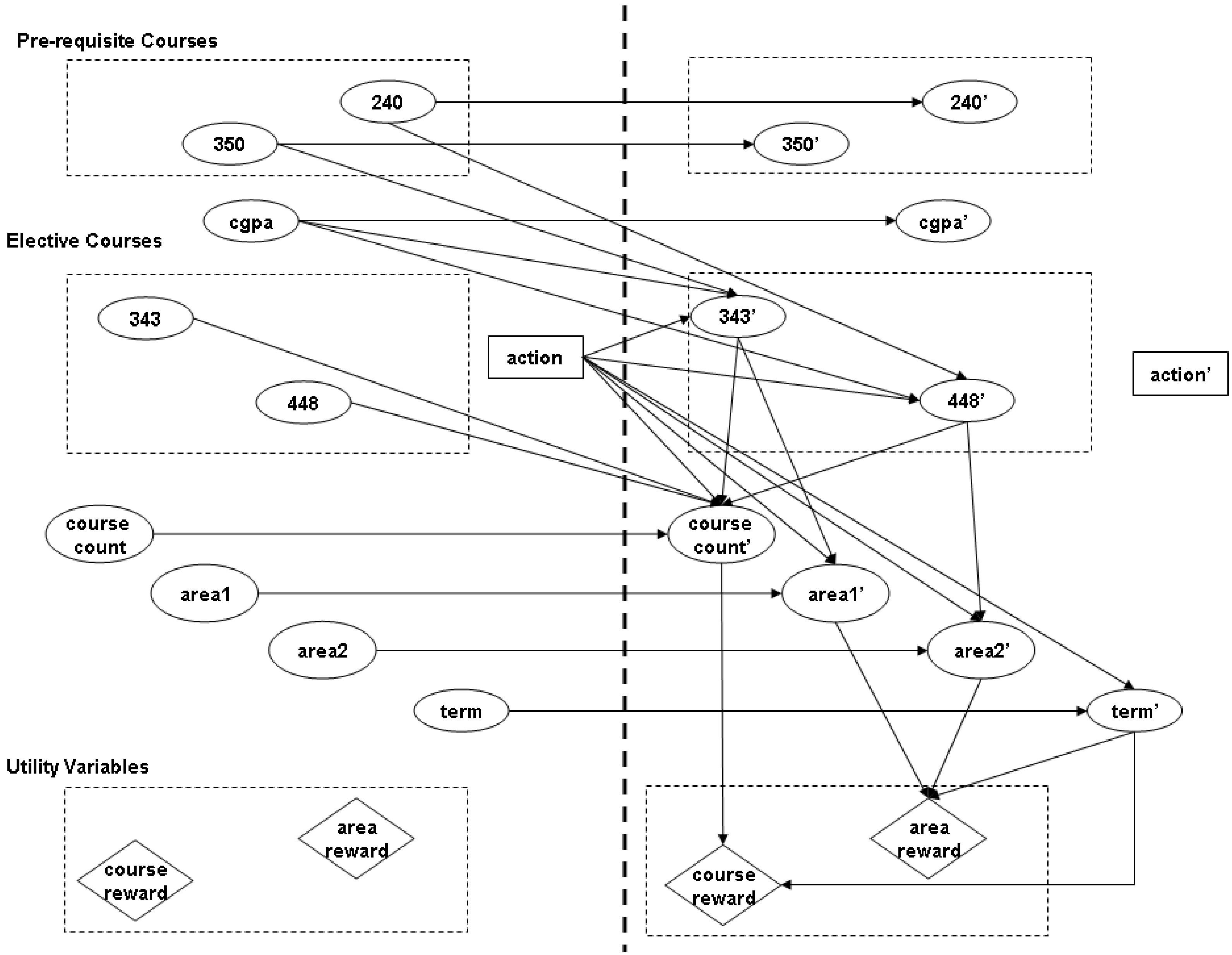
MDP Graphical Representation



Transition function: $Pr(s_t|s_{t-1},a_{t-1})$

Reward function: $R(s_t,a_t)$

Solution: policy π maximizes expected total rewards


Pre-requisite Courses
240
350
240'
350'
cgpa
cgpa'
Elective Courses
343
448
action
343'
448'
action'
course count
course count'
area1
area1'
area2
area2'
term
term'
Utility Variables
course reward
area reward
course reward
area reward

Policy Evaluation

- Optimal policy maximizes expected rewards

$$V^*(s) = \max_a \left[R(s,a) + \gamma \sum_{s' \in S} \Pr(s' \mid s,a) V^*(s') \right]$$

- Occupancy frequencies
  - Expected number of times a state is visited by executing a policy given a starting state

$$\lambda_{s_0}^{\pi}(s') = \delta(s', s_0) + \gamma \sum_{s \in S} \Pr(s' \mid s, \pi(s)) \lambda_{s_0}^{\pi}(s)$$

- Value of policy can be computed using terms that are products of occupancy frequencies and rewards

$$V_{s_0}^{\pi}(s) = \sum_{s \in S} \lambda_{s_0}^{\pi}(s) R(s, \pi(s))$$

Difficulty in Explanation

- Policy computed using complex numerical techniques
- Most primitive explanation
  - Action maximizes expected utility

- Issues
  - Many factors contribute to utility
  - Numerical value of utility only reflects preference (unless it represents something tangible like money or time)
  - Computation of expected utility is complex

Overview of Approach

- Use pre-defined templates populated at run-time
  - Not concerned with natural language generation
  - Number of templates identified such that explanations are sufficient yet minimal
- Report occupancy frequency of certain states
  - Focus on states with high/low reward
  - But do not report numerical utilities (harder to grasp)

Templates

- T1: Action ***actionName*** is the only action that is likely to take you to ***var1=val1, var2=val2, var3=val3*** about ***x*** times which is higher (or lower) than any other action

- T2: Action ***actionName*** is likely to take you to ***var1=val1, var2=val2, var3=val3*** about ***x*** times which is as high (or low) as any other action

- T3: Action ***actionName*** is likely to take you to ***var1=val1, var2=val2, var3=val3*** about ***x*** times

Minimal Sufficient Explanations

- Multiple templates possible for non-optimal actions
  - Non-optimal action may have highest frequency of reaching a state/scenario
  - May not guarantee highest expected utility

- Explanation with single template may be insufficient
- Explanation with all templates may be overwhelming
- Identify optimal number of templates to create a "*Minimal Sufficient Explanation*"

Minimal Sufficient Explanations

- Utility of explanation

$$V_{Explanation} = \underbrace{\sum_i r(s_i)\lambda_{s_0}^{\pi^*}(s_i)}_{\text{templates}} + \underbrace{\sum_j r_{\min}\lambda_{s_0}^{\pi^*}(s_j)}_{\text{no template}}$$

- Minimal sufficient explanation
  - Fewest templates with utility greater than any other action choice

$$V^{\pi^*} \geq V_{MSE} > V^{\pi'}$$

MSEs for Factored MDPs

- State space defined by a set of variables
  - Scenarios defined as set of states resulting from assigning values to a subset of variables
- Reward function can also be decomposed
- Value of policy can be computed using scenarios

$$V^{\pi}(s) = \sum_{k} \sum_{r \in dom(R_k)} r \lambda_{s_0}^{\pi}\left(sc_{R_k = r}\right)$$

- Value of explanation can also be computed using scenarios

$$V_{Explanation} = \sum_{i} r_i \lambda_{s_0}^{\pi^*}\left(sc_i\right) + \sum_{j} r_{\min} \lambda_{s_0}^{\pi^*}\left(sc_j\right)$$

Numerical Example

- Rewards
  - $R(Courses=6) = 100$, $R(Courses \neq 6) = 0$
  - $R(Areas=3) = 100$, $R(Areas \neq 3) = 0$
- Optimal action
  - $\lambda(Courses=6) = 0.67$, $\lambda(Courses \neq 6) = 0.33$
  - $\lambda(Areas=3) = 0.95$, $\lambda(Areas \neq 3) = 0.05$
  - $V^* = 100*0.67+0*0.33+100*0.95+0*0.05 = 162$
- 2nd best action
  - $\lambda(Courses=6) = 0.25$, $\lambda(Courses \neq 6) = 0.75$
  - $\lambda(Areas=3) = 0.68$, $\lambda(Areas \neq 3) = 0.32$
  - $V^{2nd} = 100*0.25+0*0.75+100*0.68+0*0.32 = 93$
- Minimal sufficient explanation
  - $V_{MSE} = (100*0.95) + (0*0.67 + 0*0.33 + 0*0.05) = 95$

Algorithm

1. For each R do
   a. For each r $\in$ dom(R) do
      i. Compute occupancy frequency: $\lambda(sc_{R=r})$
      ii. Template value: r $\lambda(sc_{R=r})$
2. Order templates in decreasing value
3. Show minimal # of templates to ensure sufficient explanation

Invariance of MSEs

Proposition: MSEs remain invariant under affine transformations of reward function

Proof:

- Occupancy frequencies add up to horizon h
- Substitute *r* with *r+c*

$$\widehat{V}^{\pi}(s) = \sum_{\mathrm{k}} \sum_{r \in dom(R_k)} (r + c) \lambda_{s_0}^{\pi}\left(sc_{R_k = r}\right)$$

$$= V^{\pi}(s) + c \sum_{\mathrm{k}} \sum_{r \in dom(R_k)} \lambda_{s_0}^{\pi}\left(sc_{R_k = r}\right)$$

$$= V^{\pi}(s) + cKh$$

Experimental Setup

- Course Advising MDP
  - Choose best combination of courses
  - 117.4 million states with 21 actions
  - Transition model generated from historical data
  - Reward for different requirements of degree
  - Horizon is 3, no discounting

- Handwashing MDP (adapted from Hoey et al 2007)
  - Assist people with dementia in handwashing
  - 207,360 states, 25 actions
  - Horizon is 100, and discount factor is 0.95

Sample Explanations

- Action ***TakeCS343&CS448*** is the best because:
  - It is likely to take you to ***CoursesCompleted=6, TermNumber=Final,*** about ***0.86*** times which is as high as any other action

- Action ***DoNothingNow*** is the best because:
  - It is likely to take you to ***handswashed=yes, planstep=Clean&Dry,*** about ***0.71*** times which is higher than any other action
  - It is likely to take you to ***prompt=NoPrompt*** about ***12.71*** times which is as high as any other action

Experimental Results

Course Advising Domain (Max Terms =4, Experiments=182)

Terms in MSE	1	2	3–4
Frequency	134	48	0
Mean (STD) $\frac{V^{\pi'}}{V^{\pi^*}}$	0.46 (0.41)	0.81 (0.24)	-

Handwashing Domain (Max Terms =19, Experiments=382)

Terms in MSE	1	2	3	4	5	6	7–19
Frequency	0	142	94	119	2	25	0
Mean (STD) $\frac{V^{\pi'}}{V^{\pi^*}}$	-	0.51 (0.22)	0.62 (0.10)	0.68 (0.04)	0.61 (0.15)	0.69 (0.05)	-

User Study

- Recruited volunteers to evaluate automatically generated explanations for course advising MDP
- Objective
  - Evaluate our explanations
  - Compare with advisor explanations
- Demographics
  - 37 undergrad and grad students participated from CS
  - 5 explanations shown to each student
    - 3 generated using our technique
    - 2 similar to those offered by human advisors

User Study Setup – Existing State

Book-keeping Information			Core			Electives		
Term Number	=	4A	cs246	=	Good	cs343	=	Good
CGPA	=	Good	cs251	=	Good	cs445	=	Not Taken
Systems/SE Area Covered	=	Yes	cs341	=	Average	cs446	=	Not Taken
Applications Area Covered	=	No	cs350	=	Good	cs348	=	Not Taken
Math Area Covered	=	No				cs448	=	Not Taken
Electives Completed	=	2				cs486	=	Not Taken
						cs360	=	Not Taken
						cs370	=	Not Taken
						cs372	=	Not Taken
						cs450	=	Average

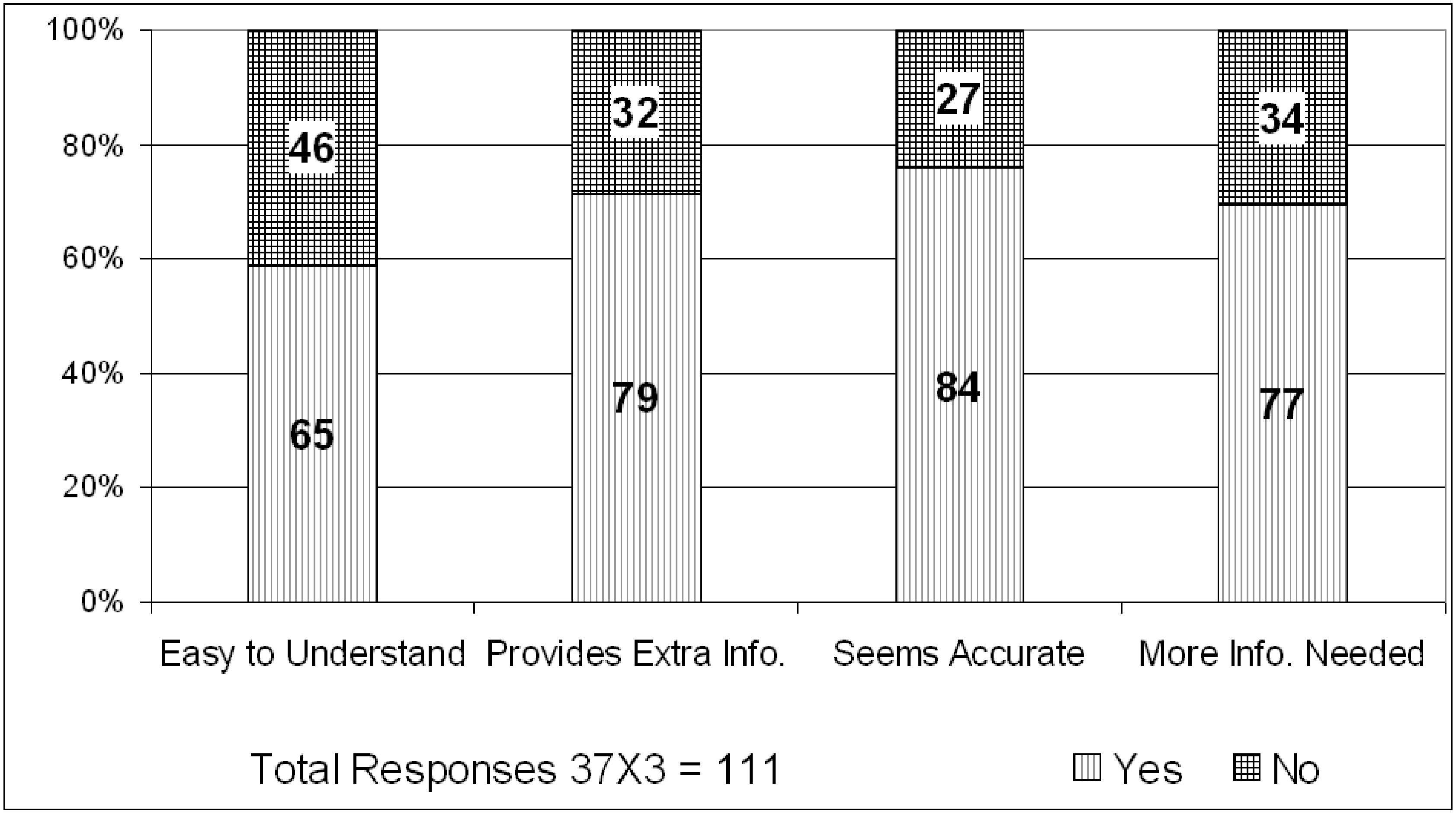
User Study Setup – Sample Explanation

Taking ***cs348 & cs370*** is the best action because:

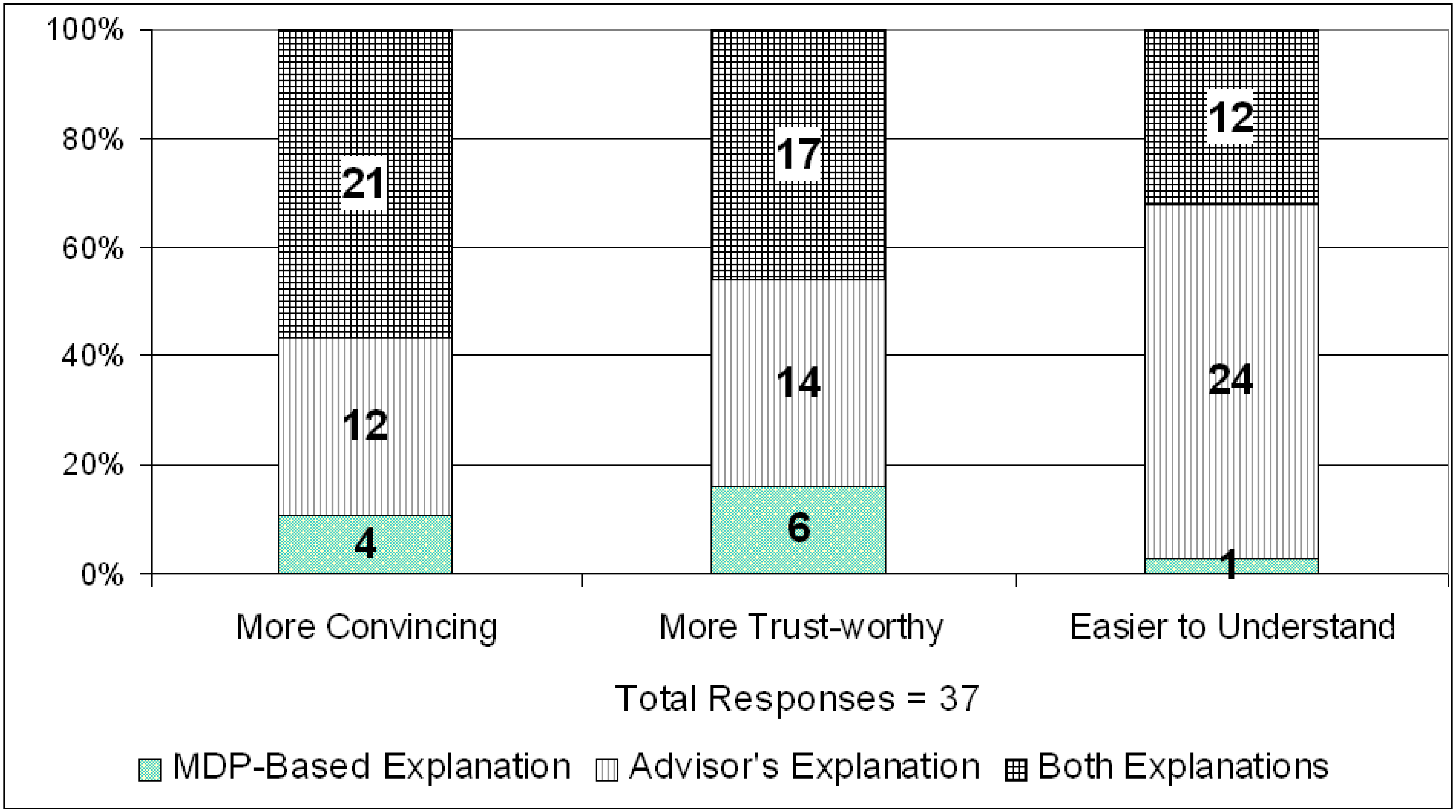
- You would be in a state with ***Electives Completed=6*** by end of ***Term Number=4B*** with about ***79%*** chance which is as high as any other combination of courses.

- You would be in a state with ***Systems/SE Area Covered=Yes*, *Mathematics Area Covered=Yes*, *Applications Area Covered=Yes***, by the end of ***Term Number=4B*** with about ***74%*** chance.

Effectiveness of MSEs



Comparison with Human Advisor Explanations



User Study – Results

- MSEs provide extra information and are trustworthy
- Human advisor explanations easier to understand
- Combination of MSE with human advisor is most preferred option
- Useful as a planning tool for students

Conclusion

- Domain-independent explanations for recommendations from MDP policies
  - Generated by populating pre-defined templates
  - No reference to numerical value of utility
  - Computed minimal set of explanations that completely justify the recommendation

- No additional effort needed from MDP designer
- User study indicates benefits of explanations

- O. Z. Khan, P. Poupart, J. Black, **Minimal Sufficient Explanations for Factored Markov Decision Processes**, *ICAPS*, Thessaloniki, Greece, 2009.

Future Work

- Inject domain-specific information in explanations
  - Represent domain-specific information in a domain-independent manner

- Explain effect of discount factor in explanations

- Extend explanations to POMDPs
  - Cater for observation function and distribution over initial state instead of single starting state

My Research Interests

- Areas
  - Reasoning under uncertainty
    - Sequential decision making (MDPs, POMDPs)
  - Machine learning, vision, NLP

- Application domains
  - Health informatics
    - Smart walker project
    - Symptom monitoring for Alzheimer's disease
  - Document clustering
    - Unsupervised cluster labelling

Graduate Studies at U of Waterloo

- CS endowment of $25 million
  - Donor: David Cheriton (Waterloo PhD, Stanford prof.)
  - **$1 million/yr for research & graduate studies**
- CS is in the **Faculty of Math**
  - In AI, statistics and optimization are key
  - Easy interaction with dept. of Statistics and Combinatorics & Optimization.
- Start your own company
  - **IP belongs to the creators (not the University)**
  - Spinoffs: RIM, Maple, Open Text, etc.
  - Technopark on campus