

A scenic view of a rocky coastline. On the left, a steep, forested cliffside descends to a sandy beach. The water is a vibrant turquoise color, with white foam from waves crashing against the rocks. The sky is a clear, pale blue. The overall scene is bright and natural.

Computable Social Communication

David Pautler
Final Defense
February 8, 2007

Overview

- The Problem
- Our Objective
- Related Work
- Our Approach
- Empirical Evaluation
- Conclusions & Future Work

The Problem

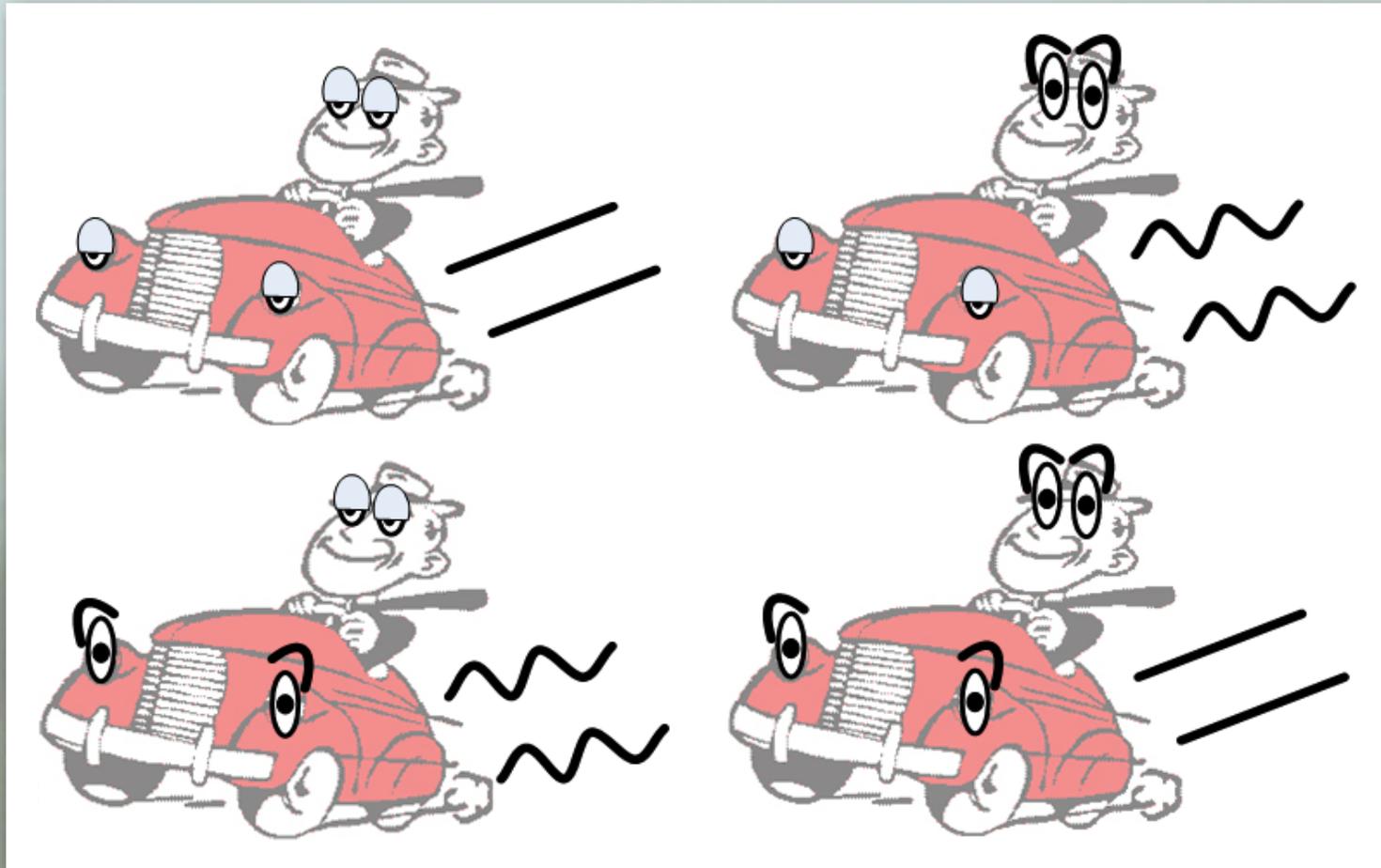
- Software that plays a social role is becoming embedded ever deeper in everyday life
- Example: Voice-driven “phone trees”
- But social interaction is too varied to be fully anticipated “at the factory”
- Socially-inept interfaces are frustrating, even dangerous

Ex: In-Car Driving Advisors do help...



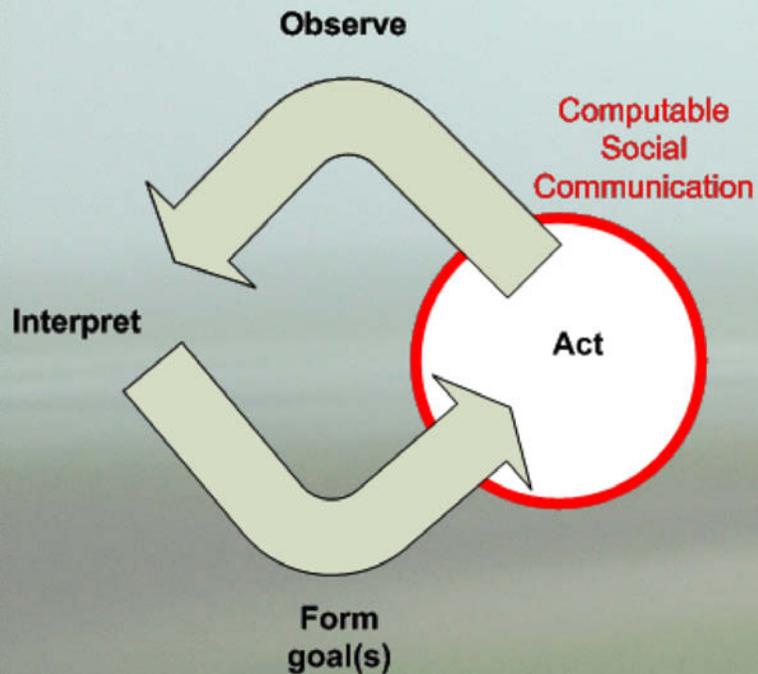
Source: Jonsson et al 2005

...But to do so they must match the mood of the driver



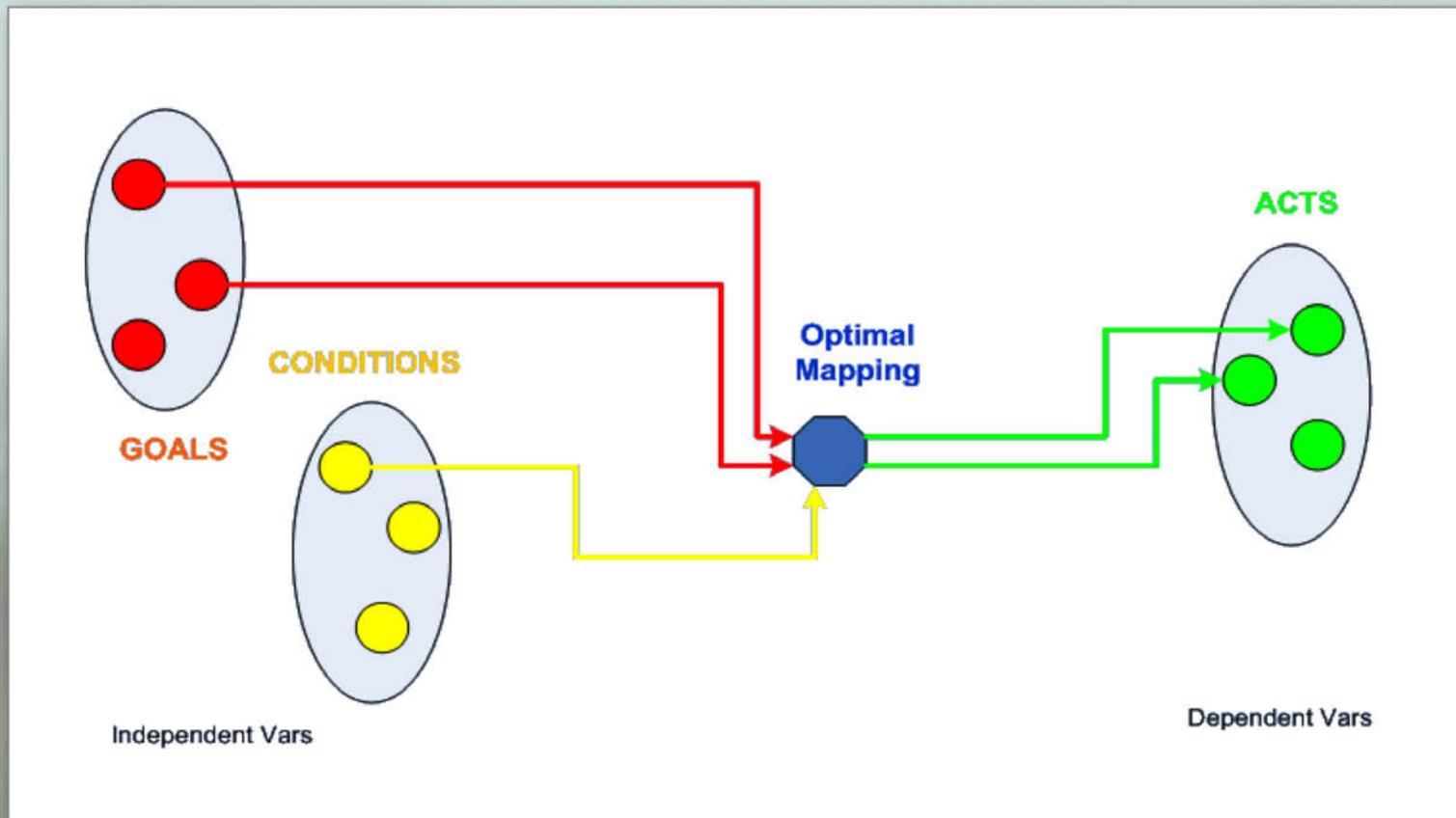
Source: Jonsson et al 2005

Choosing where to focus



- Overt, verbal communication
- Between two people
- Where one tries to persuade the other
- Allowing for a variety of social goals, conditions, and communicative tactics
- A contribution to theory, but with an eye for future application

Desired “shape” of our solution



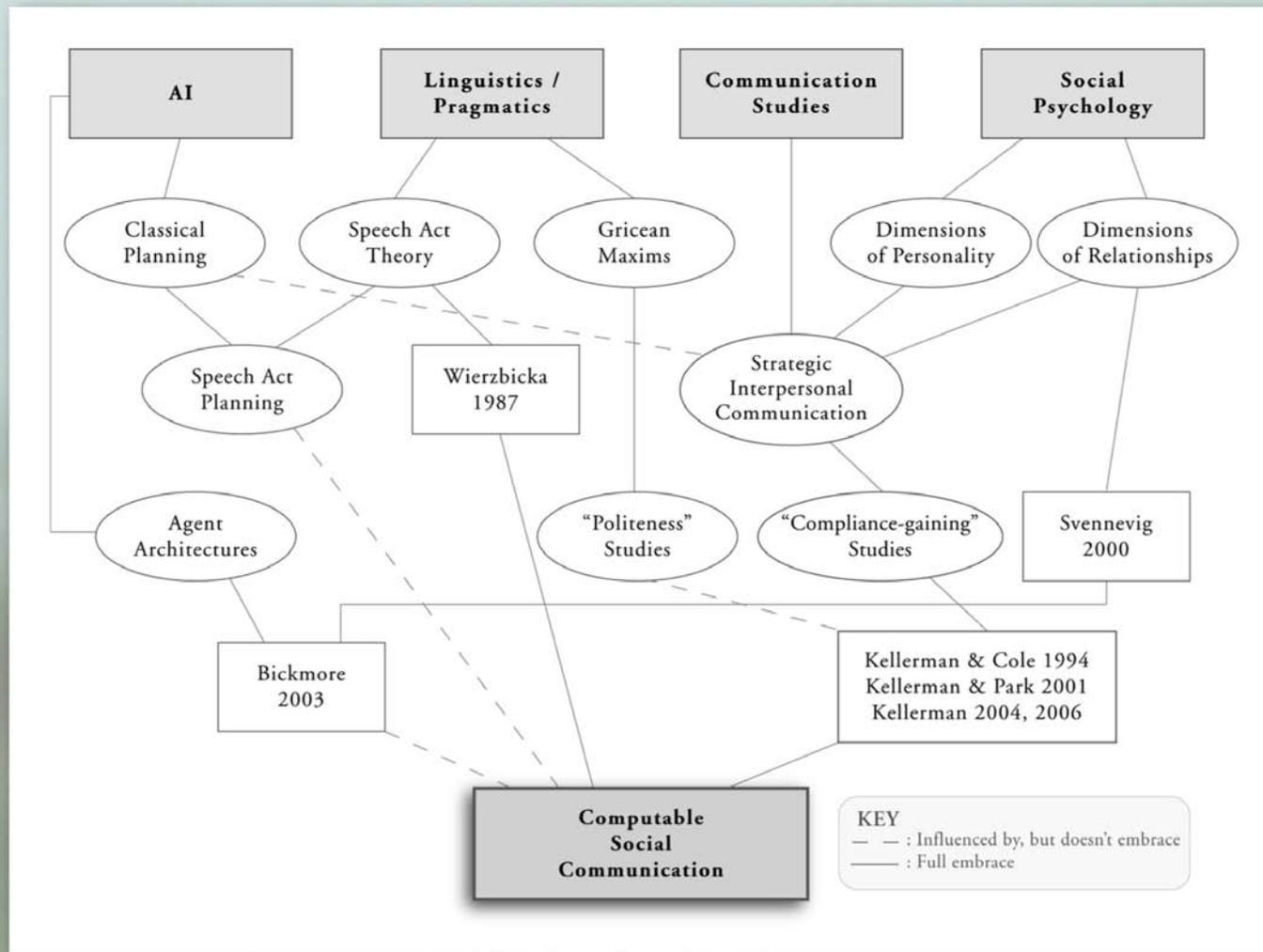
So the problem is

- What goals, conditions, and communicative tactics to simulate?
- How to find the optimal tactics for given goals and conditions?
- How to do it dynamically?



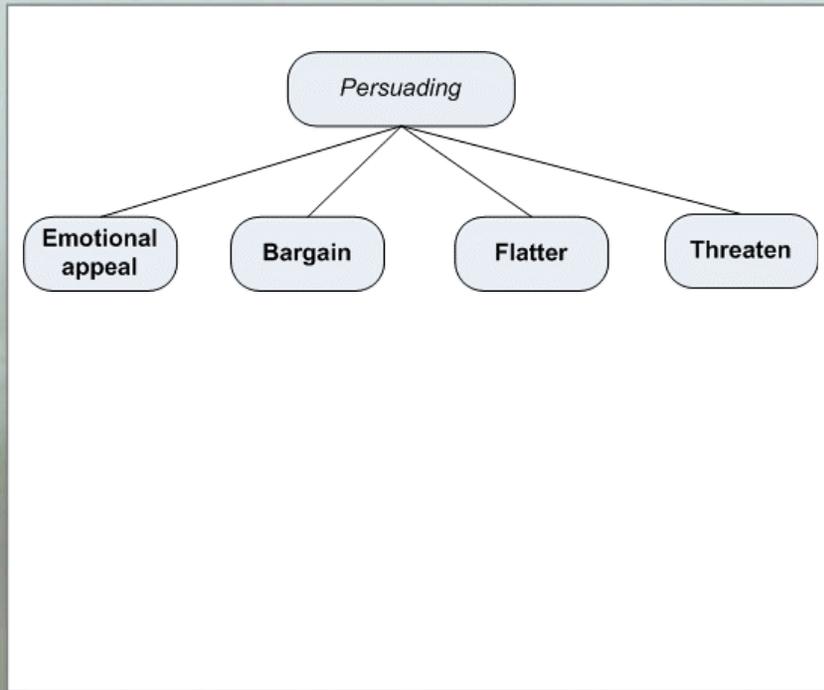
Related Work

Conceptual map of influences on our approach

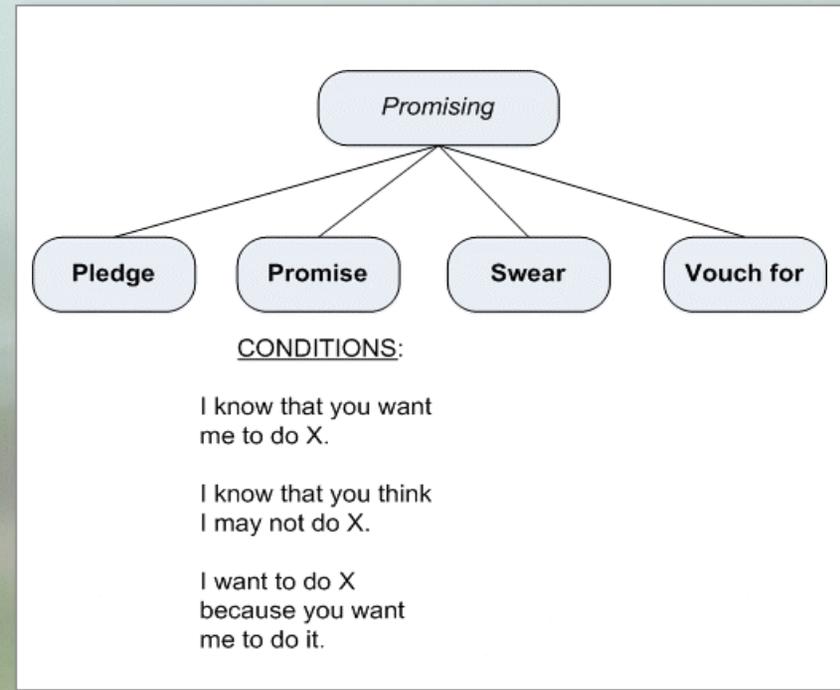


Social Sciences

Strategy Taxonomies

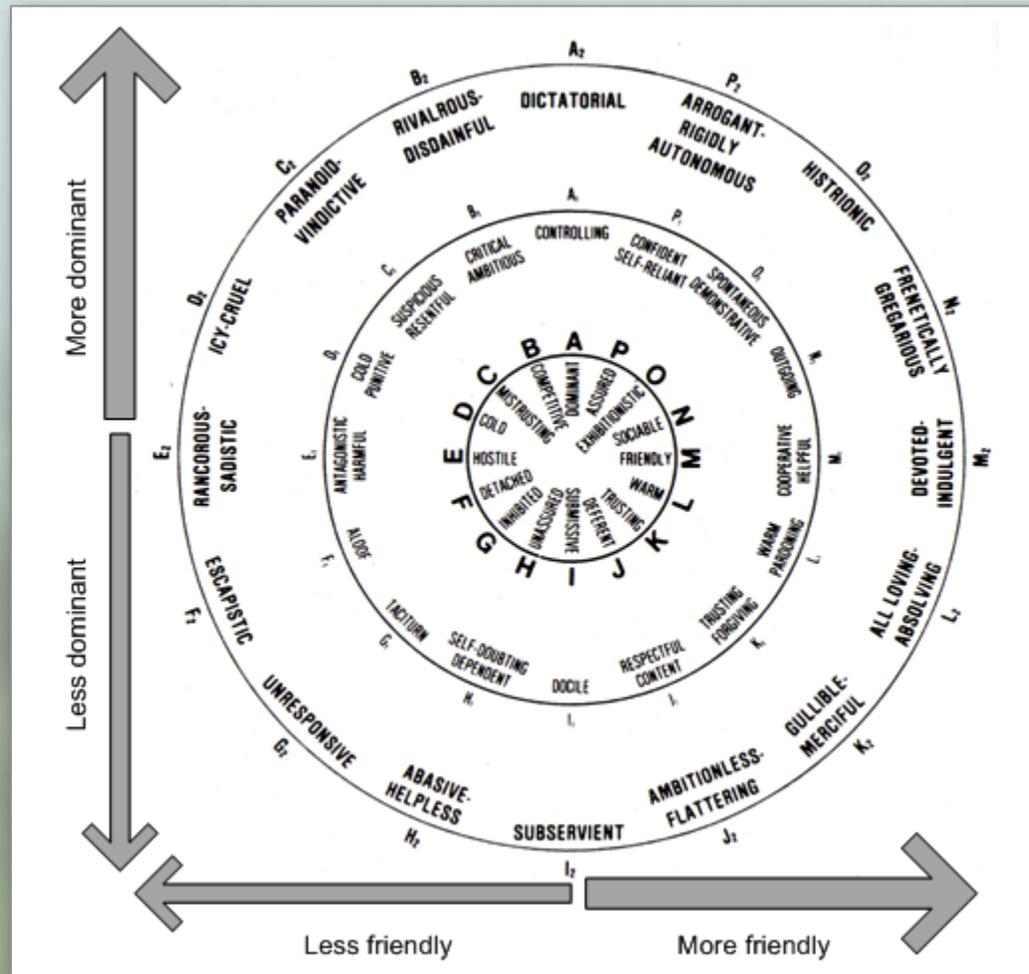


Act Taxonomies & Act Conditions

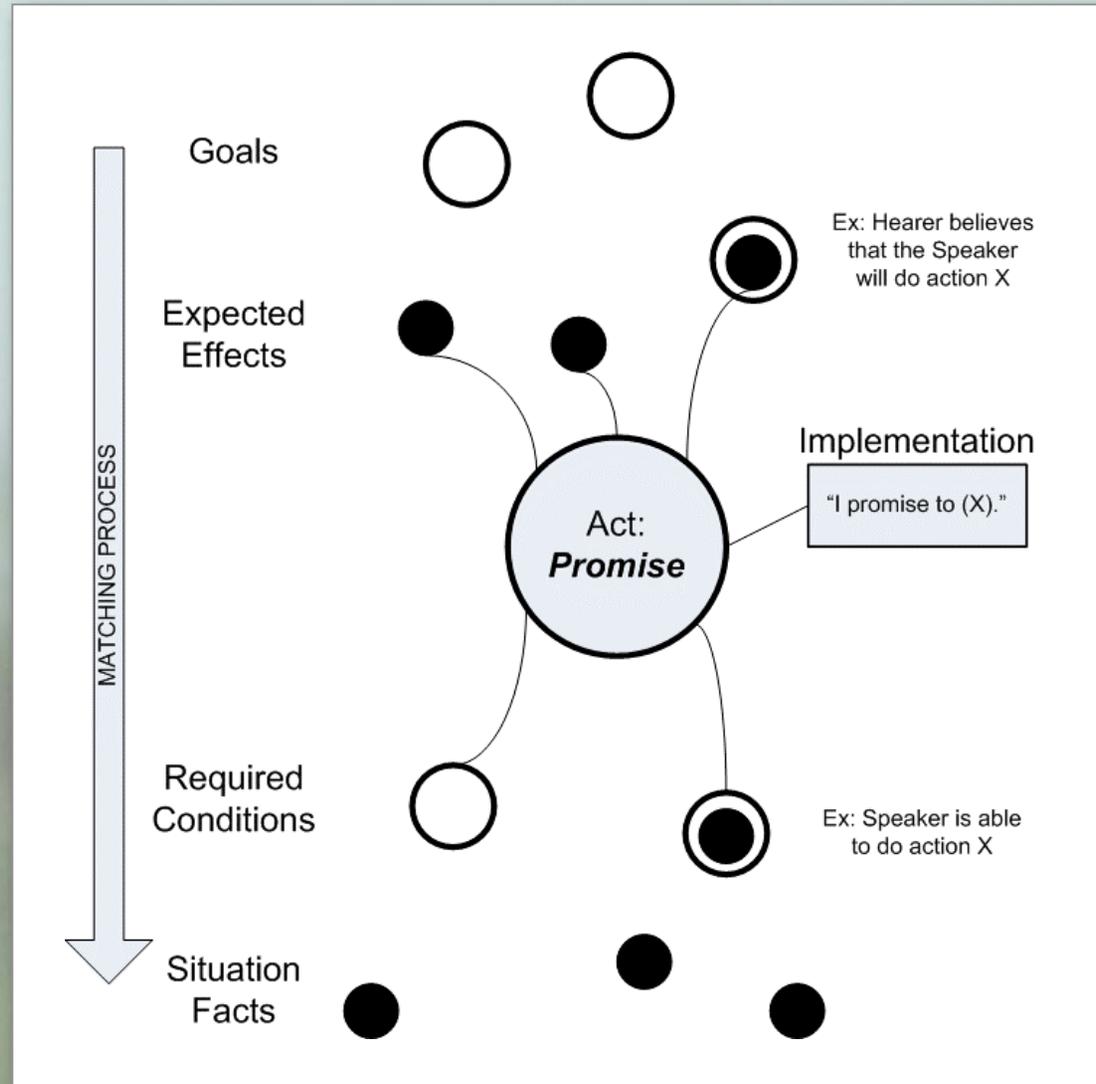


Social Sciences (continued)

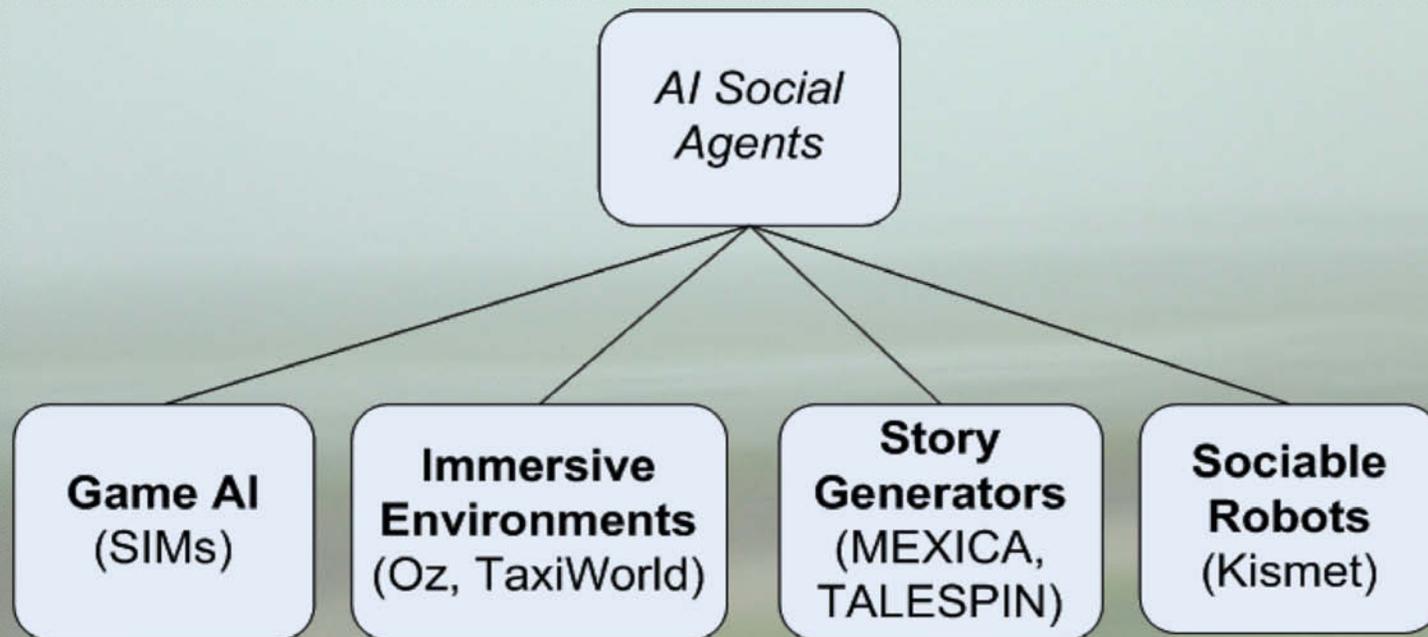
Dimensional Models of Reciprocal Behavior



AI: Speech Act Planning



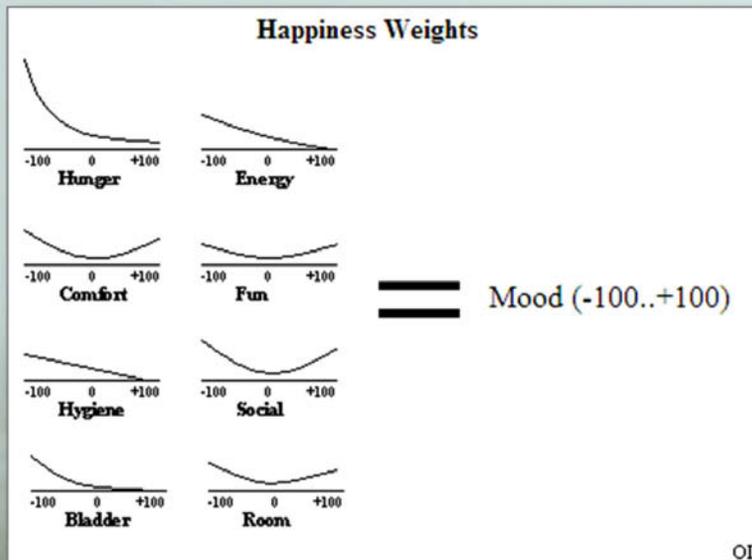
Types of AI Social Agents



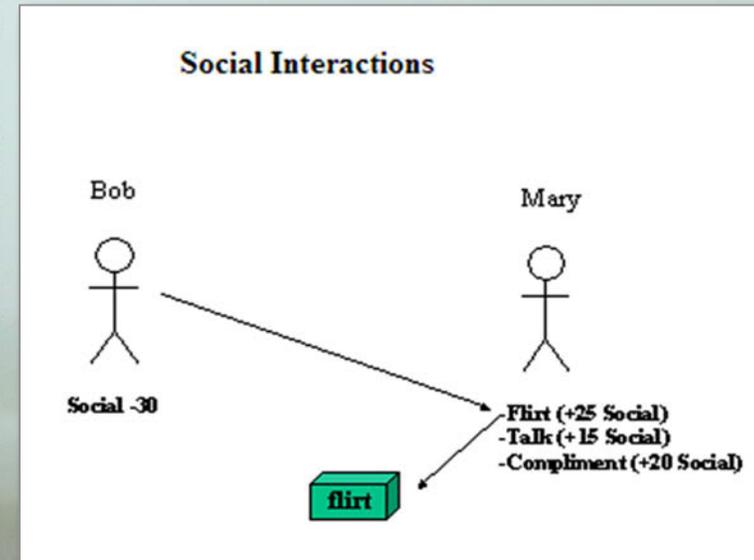
How do SIMs agents decide how to act?



Urges, Emotions, and Behaviors in SIMs NPCs



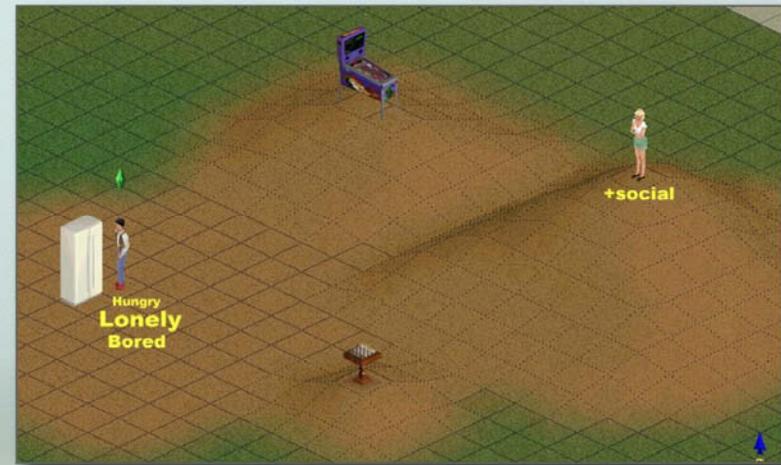
- Urges naturally increase (or decay)
- Urge levels determine emotions



- Urges and emotions drive behavior selection
- Behaviors affect the state of urges and emotions

Source: SIMs creator Will Wright, via Ken Forbus lecture notes

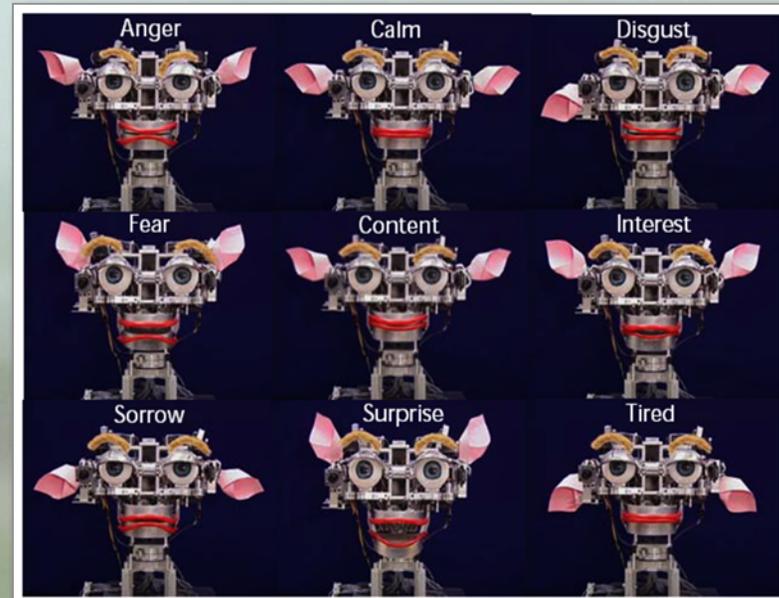
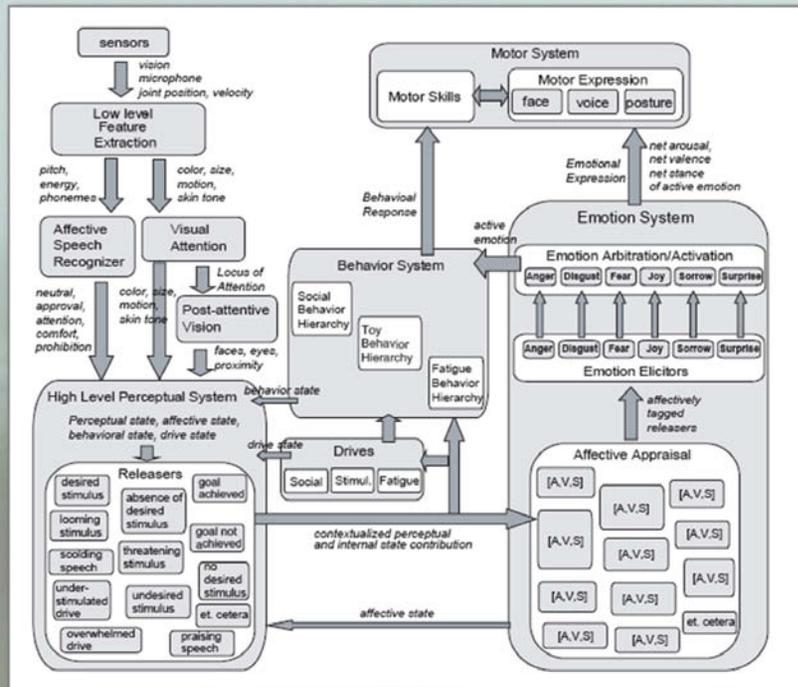
“SIMs life” emerges from the Urge-Action-Observation cycle



Sociable robot "Kismet" (MIT)

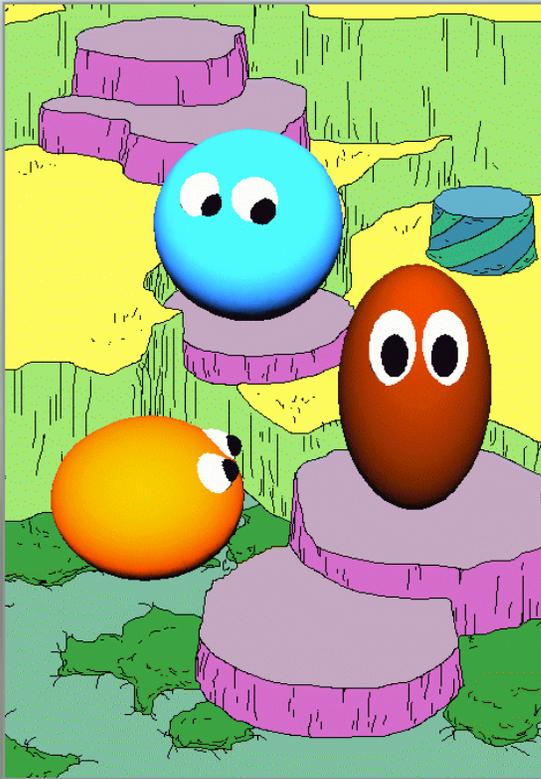
Similar Urge-Action-Observation architecture

Actions include facial expressions, head posture, and simple vocalizations



Interactive Drama:

Oz Project (CMU)



Story Generation:

MEXICA (U of Sussex)

Sample rule:

*“If a warrior is ill or wounded and a lady cures him, it is natural that he will develop very positive emotions towards the lady. In this way, when defining the action A CURED B, post-conditions might include that B develops an **Emotional Link** of type 1 intensity +3 towards A.”*

Reactions

- PRO's
 - Inter-linked representations of urges, emotions, actions, and observations allow for rich, believable social simulations.
- CON's
 - Believability depends on the numeric weights being “in tune”, but these are largely based on introspection (and are rarely published).
 - Evaluations apply to whole systems, which makes it difficult to appraise the social model itself.

Search for empirically-defined models

- Conversational Constraint Theory (CCT) is an act taxonomy with empirically-defined weights
- Models how one person persuades another to do something
- 13 persuasion goals, 56 speech acts, 10 rules for adjusting plans to the situation

CCT methodology

- For each of the 13 goals, 19-20 undergraduates rated each of 56 speech acts for their efficiency and appropriateness for that goal on a 1-7 scale
 - Appropriateness = proper, fitting, polite, and courteous
 - Efficiency = virtually immediate effect without much expenditure of time or effort
- Told to imagine they were engaged in a conversation, where the purpose of the conversation was to achieve the goal
- Told to focus solely on how appropriate (or efficient) the act is for the goal, regardless of:
 - Personal preference
 - Efficiency (or appropriateness)

Sample scores

Social Appropriateness scores

	Get Date	Stop Annoying Habit	Obtain Favor	End Relationship	Change Opinion	Share Time Together	Initiate Relationship	Get Advice	Move Relationship Forward	Obtain Information	Fulfill Obligation	Obtain Permission	Provide Guidance
insult	1.70	1.44	1.30	1.61	1.63	1.90	1.79	2.10	1.88	1.80	1.55	2.00	1.84
threaten	1.40	1.72	1.45	1.72	2.16	1.67	1.74	1.81	1.88	1.80	1.80	1.95	1.95
ridicule	1.55	1.68	1.35	2.00	1.95	1.76	1.95	2.05	1.76	2.10	1.75	2.15	1.95
attack	1.80	1.33	1.60	2.56	2.79	2.71	2.11	2.05	2.00	2.05	2.10	2.10	2.37
blame	1.65	1.61	2.15	2.39	2.68	2.14	2.11	2.38	2.35	2.15	2.00	1.95	2.10
boast	2.80	2.38	2.45	2.44	2.05	2.00	1.80	2.00	1.80	1.85	2.20	2.40	2.37

Efficiency scores

	Get Date	Stop Annoying Habit	End Relationship	Obtain Favor	Share Time Together	Move Relationship Forward	Get Advice	Change Opinion	Initiate Relationship	Obtain Information	Obtain Permission	Fulfill Obligation	Provide Guidance
insult	1.70	1.44	2.94	1.85	2.76	2.00	1.95	2.16	2.21	2.15	1.95	1.80	2.05
ridicule	1.75	1.67	2.83	1.85	2.76	2.24	1.81	2.26	2.00	2.20	2.20	1.75	2.53
blame	1.75	1.61	2.72	2.00	2.67	2.35	2.05	2.84	2.16	2.20	2.15	2.00	2.68
boast	2.80	2.50	1.78	2.15	2.48	1.76	2.57	3.21	2.32	2.35	2.45	1.85	2.79
threaten	1.80	2.33	2.56	2.25	3.05	2.53	1.76	2.68	2.16	2.55	2.65	2.20	2.79
attack	1.80	1.44	3.22	2.40	3.14	2.47	2.24	2.05	2.26	2.60	2.25	2.80	2.74

Ex: When Threatening someone in order to Stop an Annoying Habit, the expected efficiency is 2.33 but its appropriateness is only 1.72.

CCT Situation Factors

- **YOU:**

- Concerned about your bond with them?
- Is being proper and courteous generally a high priority?
- Is achievement and avoidance of waste generally a high priority?

- **YOUR NEEDS:**

- Urgent?
- Does it feel like special courtesy is called for?
- Does it feel it should be especially business-like?

- **THEM**

- Do they have higher social position?
- Is being proper and courteous generally a high priority?
- Is achievement and avoidance of waste generally a high priority?
- Do they expect special courtesy?
- Do they expect you to be especially business-like?

- **THE ENVIRONMENT**

- Formal?
- Private?

CCT decision process

- CCT is not a simulator
- CCT does not specify any algorithm
- But we can infer a decision process from the discussion of goals, act scores, situation factors, and thresholds

CCT decision process (continued)

Choose a goal. Ex: Stop Annoying Habit

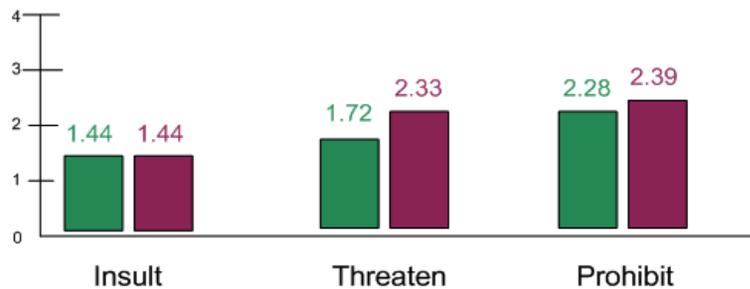
Note the scores that different speech acts have for this goal on both scales.

Appropriateness Scores

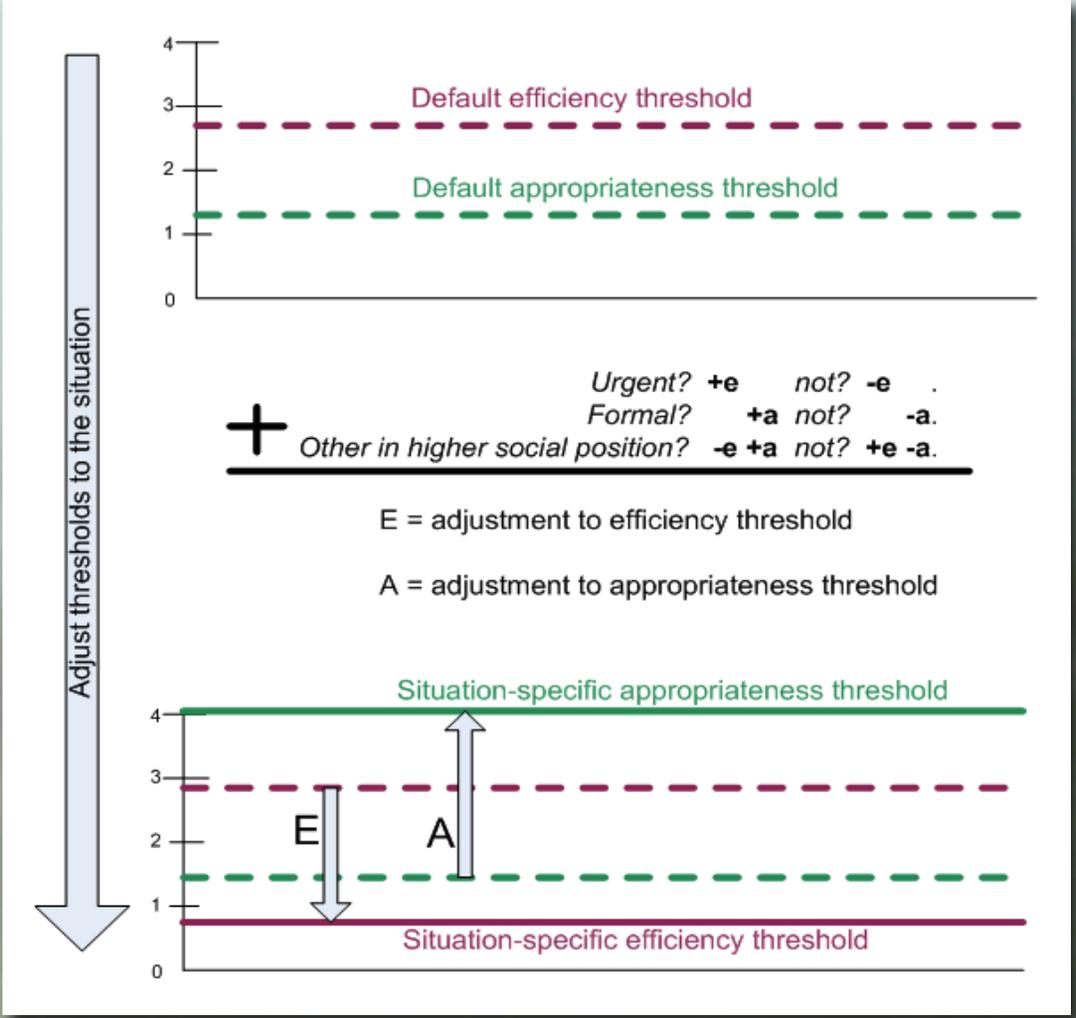
	Get Date	Stop Annoying Habit	Obtain Favor	Re
insult	1.70	1.44	1.30	
threaten	1.40	1.72	1.45	
ridicule	1.55	1.56	1.35	
attack	1.80	1.33	1.60	
blame	1.65	1.61	2.15	
boast	2.80	2.28	2.15	
accuse	1.70	1.83	2.20	
forbid	2.20	2.22	2.30	
prohibit	2.25	2.28	2.55	
order	2.10	1.88	2.35	

Efficiency Scores

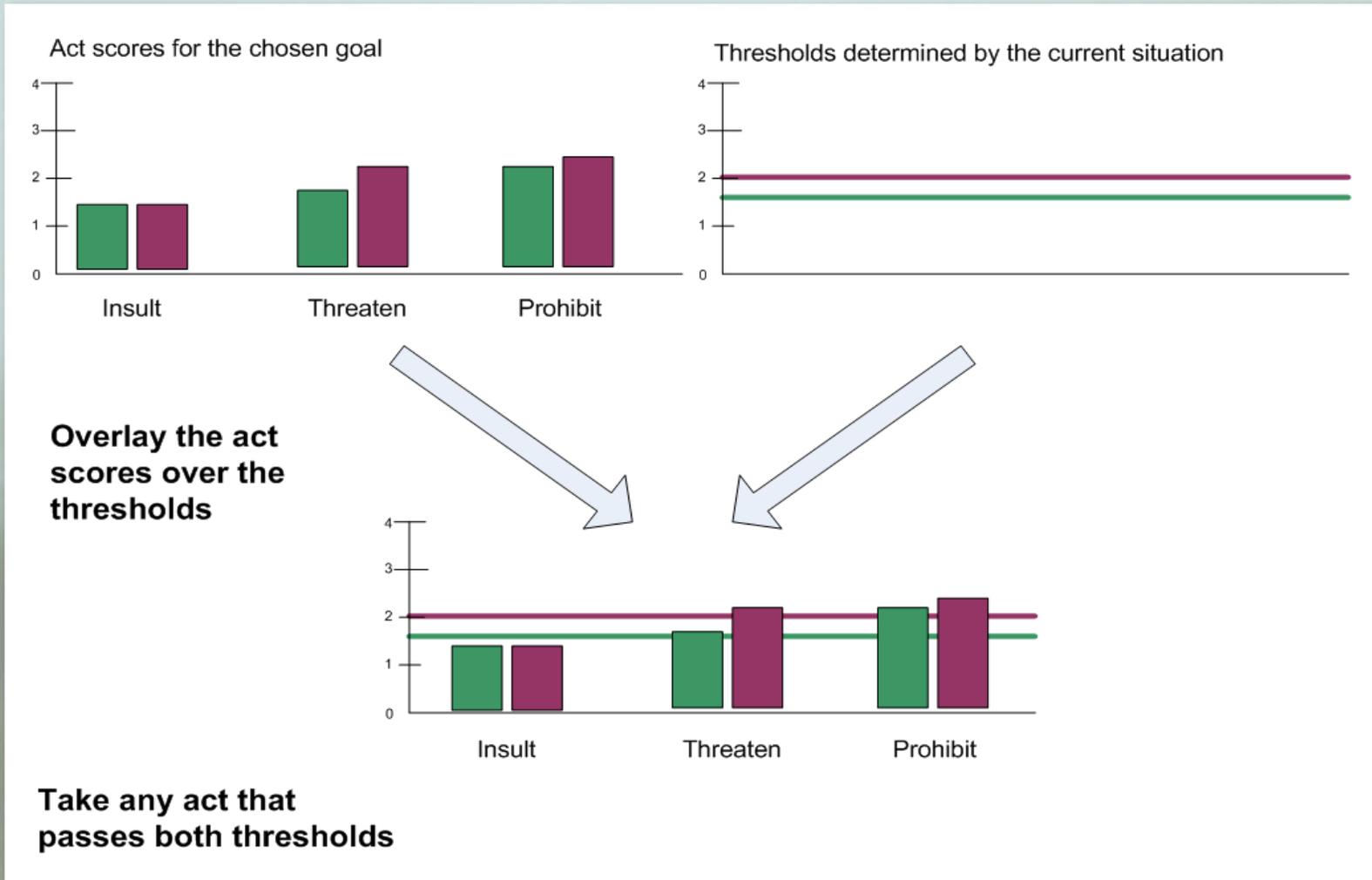
	Get Date	Stop Annoying Habit	End Relationship	Ot
insult	1.70	1.44	2.94	
ridicule	1.75	1.67	2.83	
blame	1.75	1.61	2.72	
boast	2.80	2.50	1.78	
threaten	1.80	2.33	2.56	
attack	1.80	1.44	3.22	
forbid	1.90	2.33	3.06	
accuse	1.55	2.44	3.28	
prohibit	2.20	2.39	3.33	
order	1.90	2.00	2.50	



CCT decision process (continued)

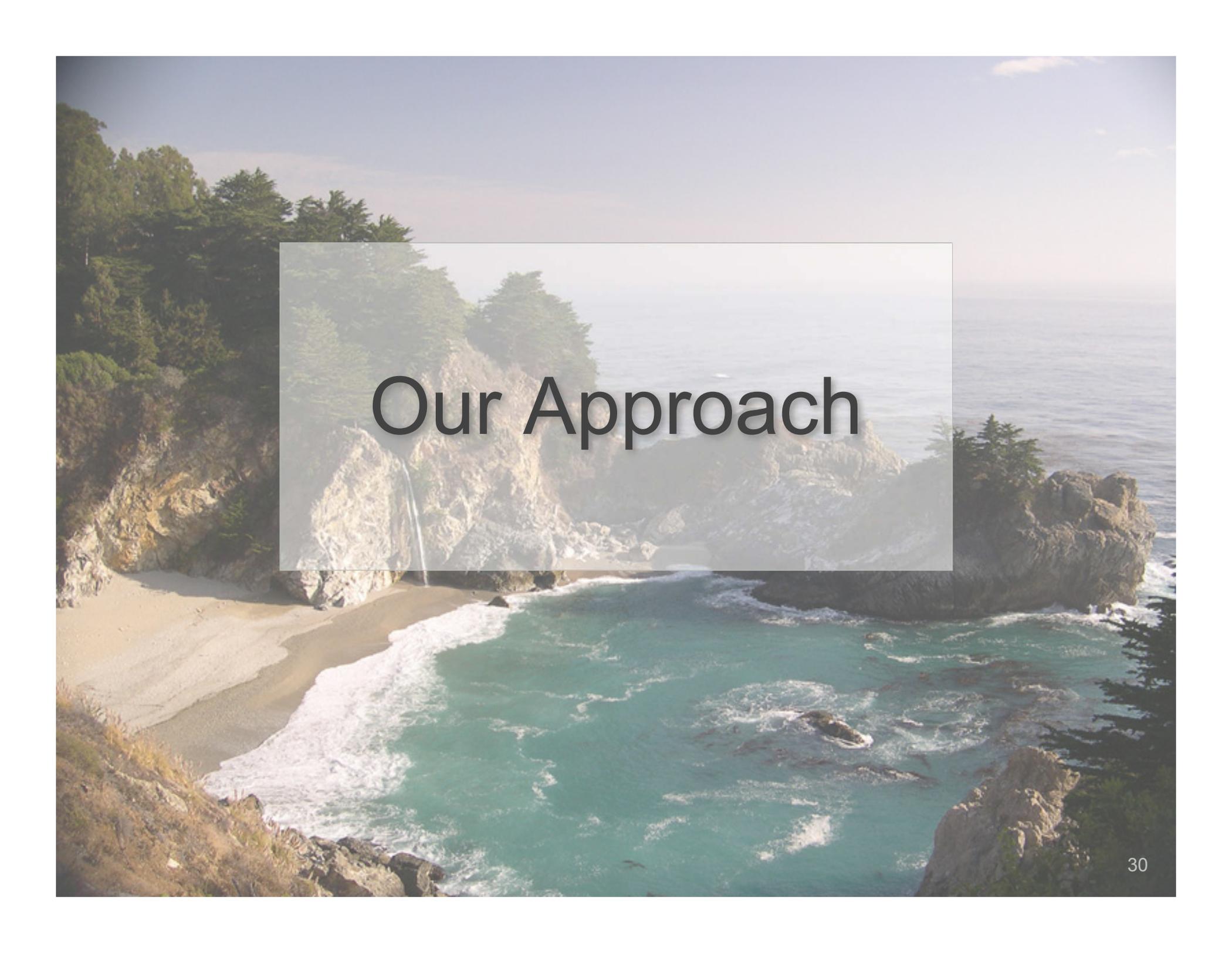


CCT decision process (continued)



Reactions

- PRO's
 - Similar architecture to agent work plus empirical basis
- CON's
 - No values for threshold defaults or deltas
 - What if no act can satisfy both thresholds on its own?
 - What if there are multiple goals?
 - Acts can be inappropriate due to unmet conditions
 - Ex: Promising to do X when one is incapable of performing X



Our Approach

Overview

- Create an algorithm from CCT's implicit decision process
- Extend the algorithm to handle multi-goal, multi-act plans
- Verify the algorithm through comparative tests with human participants

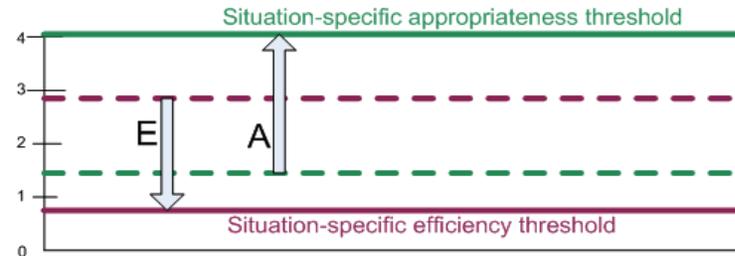
Designing an algorithm for CCT

- Setting thresholds
 - Depends on 10 situation factors, which can be asked about in any order
 - Each situation factor affects one or both thresholds
 - Need values for default thresholds and deltas
 - Since act scores range within $\{-3, 3\}$, assume a small default such as $+0.5$
 - Assume all factors use the same delta, regardless of whether the factor affects appropriateness or efficiency
 - Testing these assumptions would be a good area for future work
 - If these assumptions are wildly inaccurate, one would expect our empirical evaluation to show it

Designing an algorithm for CCT (cont.)

- Allowing for multiple goals and acts
 - Try an AI planner?
 - Probabilistic planners? (to allow for partial fulfillment of goals)
 - Turns out to be overkill
 - Planner only needed if plans expected to be tree-shaped
 - Integer Linear Programming is a better match
 - Solves a system of linear equations while optimizing a measure of your choice
 - CCT scores and thresholds define the equations; optimize by preferring plan with fewest acts

How it works



For every selected goal, the associated act scores can be used as equation coefficients

Once both thresholds are adjusted for the situation, the threshold values can be used as upper-bounds

Appropriateness Equation for goal 'Stop Annoying Habit':

$$1.44 * \text{Insult} + 1.72 * \text{Threaten} + 2.28 * \text{Prohibit} + \dots \geq 4.1$$

Efficiency Equation for goal 'Stop Annoying Habit':

$$1.44 * \text{Insult} + 2.33 * \text{Threaten} + 2.39 * \text{Prohibit} + \dots \geq 0.7$$

Insult, Threaten, etc are variables with value either 0 or 1. When the inequalities are solved, "1" means the act should be used and "0" means it shouldn't.

Sample session



- SCENARIO
 - Couple at fancy dinner
 - Husband is messy eater
- GOALS (of woman):
 - Stop Annoying Habit
 - Provide Guidance
- Recommends 3 specific acts
- Plausible when instantiated like this:
 - **Give:** Unfold your husband's napkin and hand it to him.
 - **Ask:** Ask your husband if he is aware that his trousers are being stained.
 - **Inform:** Tell your husband that he is slurping loudly and embarrassing you.

Empirical Evaluation

- Pilot
 - Web survey hosted by U of Zurich Psychology Dept
 - 54 participants over 2 months
- Survey
 - Paper questionnaire distributed at Leeward
 - 23 participants in 3 morning classes, same day

Pilot questionnaire

- Your goals:
 - Obtain a favor – You are going on a trip and need to find someone to feed your pet
 - Get advice – Your flight is during rush hour, and you'd like to know a quick way to the airport
- The situation:
 - You are moderately goal-oriented (5 on a 1-7 scale)
 - You are moderately attentive to social cues in general (4 on a 1-7 scale)
 - The situation is fairly formal (6 on a 1-7 scale)
 - The situation you are talking about is fairly private (6 on a 1-7 scale)
 - Your goals are fairly urgent (6 on a 1-7 scale)

Thank - I'm so grateful.

Warn - I would avoid that if I were you.

Apologize - I'm sorry to impose on you.

Forbid - Don't mess with that.

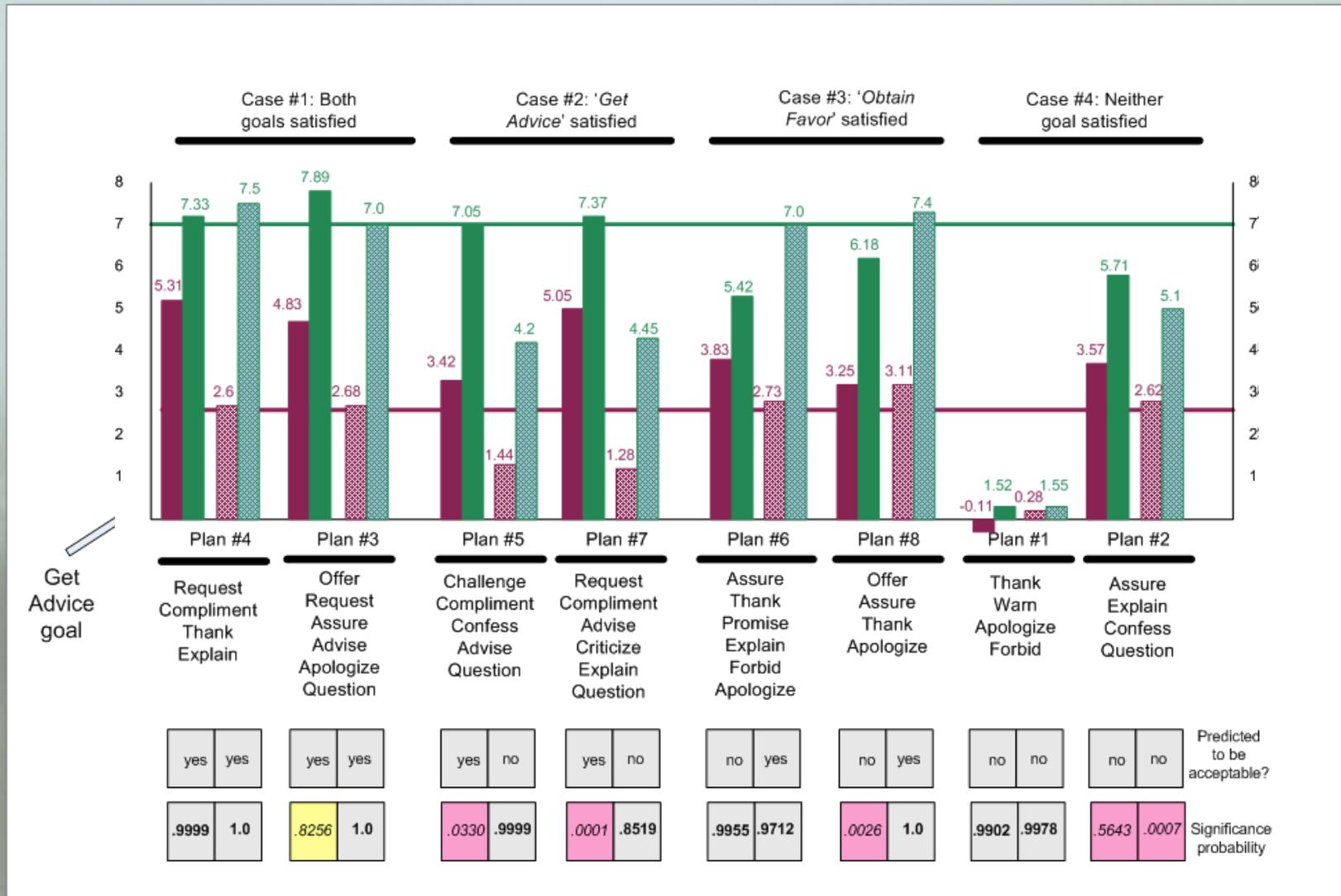
1. Is this set an acceptable way to Get Advice?

- 1 (Not at all)
- 2
- 3
- 4
- 5
- 6
- 7 (Absolutely)
- Not sure

2. Is this set an acceptable way to Obtain a Favor?

- 1 (Not at all)
- 2
- 3
- 4
- 5
- 6
- 7 (Absolutely)
- Not sure

Pilot design and results



Could the results have been better?

- Did “redundant” hypotheses have similar support?
- Some act scores might not be reliable
- Wording of sample sentences might be biased
- Assumed thresholds might need adjustment

Faulty scores?

- We are using 1456 score values (13x56x2) from a 2004 study
- A 1994 study has 34 scores for the same goal/act/effect combinations
- Only 18 of the 34 are within 1.0 of the corresponding 2004 value
 - Answers range in {1..7}
- 4 scores used to create the pilot are cast into doubt by this
- But only 2 of those are associated with failed hypotheses
- => Possible source of error, but too little evidence

Biased wording?

Explain - ...and that's why I need your help?

Offer – Let me know if I can return the favor.

Forbid - Don't mess with that. $\square ? \Rightarrow$ Don't do it that way.

Looked at:

- Acts that appear only in unsupported “yes” hypo’s
- Acts that appear only in unsupported “no” hypo’s

=> No simple pattern emerged

Adjust thresholds?



5.1

...helps plan 2



6.18

...helps plan 8

7.05



OR

3.42



...helps plan 5
...but hurts others

7.37



OR

5.05



...helps plan 7
...but hurts others

=> Choose either of the top two

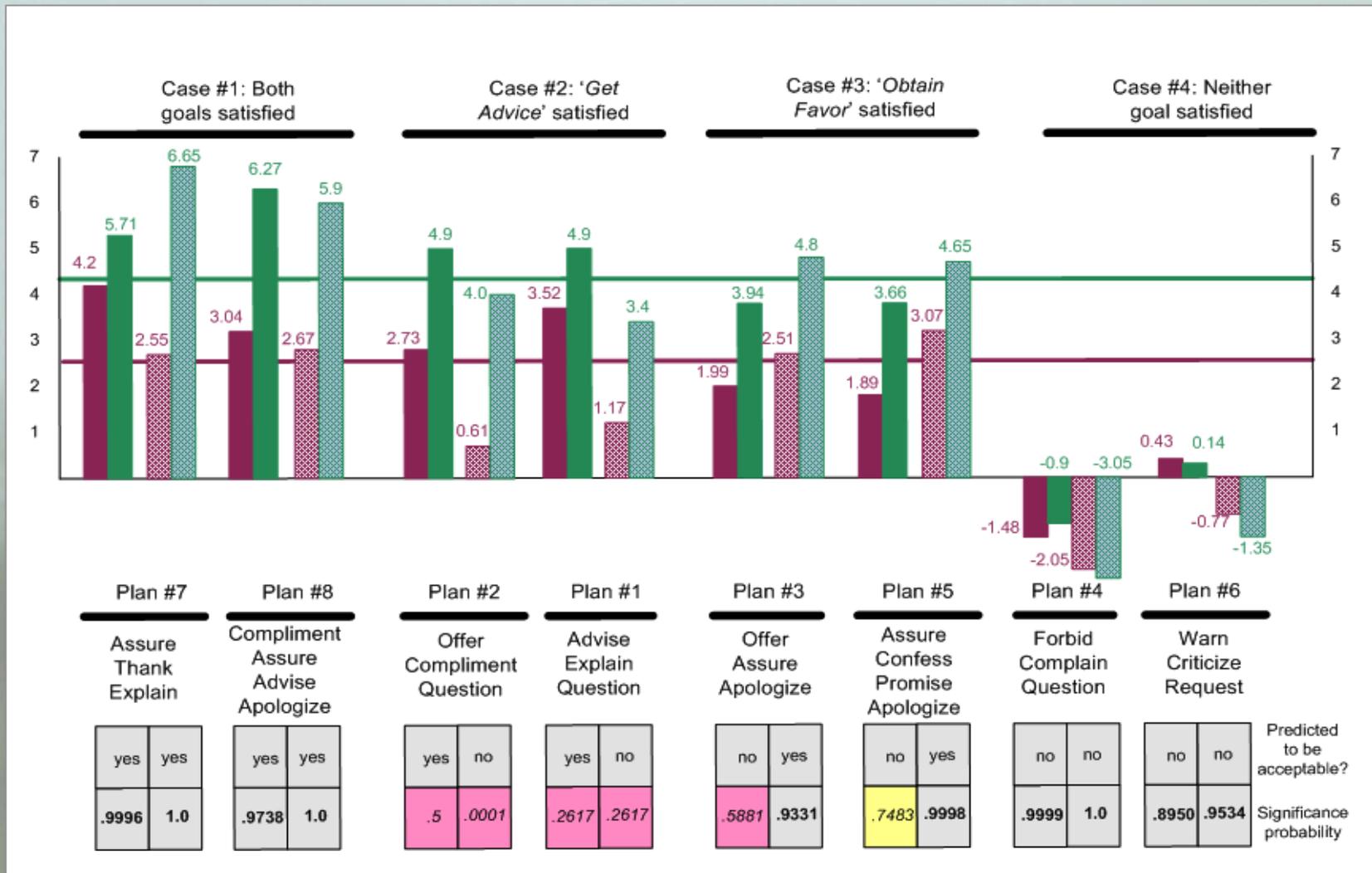
Lessons learned from pilot

- Likert scale required norming responses to yes/no
 - Took the max and min for all questions for one user and halfway between those was the cut-off for yes versus no for that one user.
 - Assumption #1: With 12 questions that were selected to present both obviously unacceptable and obviously acceptable plans, every participant would have at least one question that he/she would label with his/her min value and another question that would be labeled with his/her max value.
 - Assumption #2: Participants have a linear scale and the acceptable threshold is halfway between min and max.
 - => Switched to yes/no options in the study

Lessons learned from the pilot (cont.)

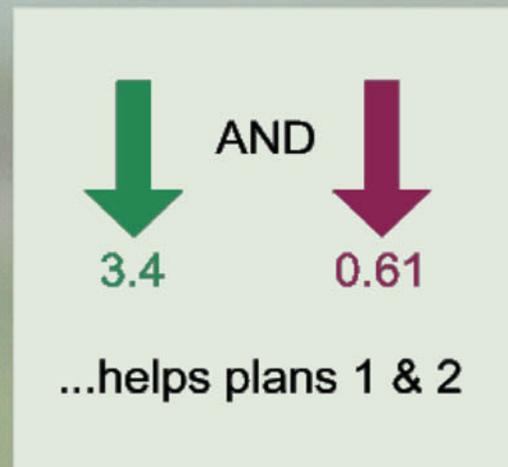
- Appropriateness threshold should be adjusted
 - Move appropriateness threshold down from 7.0 to 5.1 or 6.18
 - (Accidentally used 4.45 in survey, but didn't cause a problem there.)

Survey design and results



Could results be better?

- Do “redundant” hypotheses get similar support?
 - Yes
- Is there obvious bias toward some speech acts?
 - No
- Should thresholds be adjusted?
 - Yes



...and shows that using 4.45 instead of 5.1 made no difference



Conclusions

Conclusions

- CSC supports more goals and acts than any other speech act planner
- CSC allows for partial fulfillment of goals
 - Has empirical support for its representation via CCT
- CSC optimizes multi-goal plans by minimizing # acts
- CSC shows good support in empirical comparisons with human judgments of plan quality

Future work

- Verify assumptions
 - Threshold defaults and deltas
 - Use of addition to combine scores vs. other n-ary functions
- Make the specifics of the situation part of the plan
 - Add conditions to each act
 - Tie the conditions to the expected effects
 - => Would help prune plans and flesh out text realization
- Integrate with social interpretation research



Q & A

David Pautler
Final Defense
February 8, 2007