Social Causality and Responsibility – Modeling and Evaluation

Wenji Mao & Jonathan Gratch
University of Southern California

institute for creative technologies

USC

The Blame Game

- Assigning blame is a central human past time
- Prerequisite for important social functions
  - Distributing rewards in collaborative settings
  - Learning from collaborative success and failure
  - Social planning: predicting and manipulating social agents
  - Explaining social motivations
- Prerequisite for generating and interpreting social dialogues
  - “It’s not my fault!” “Is too!”
- Core determinant of social emotions
  - Distinguishes Guilt from Anger

Many IVA Applications

- Informing models of emotions
  - Gratch, Mao & Marsella 2005
- Guiding mitigating dialogues
  - Martinovski et al. 2005
- Assessing performance in group training
  - Mao and Gratch, 2003

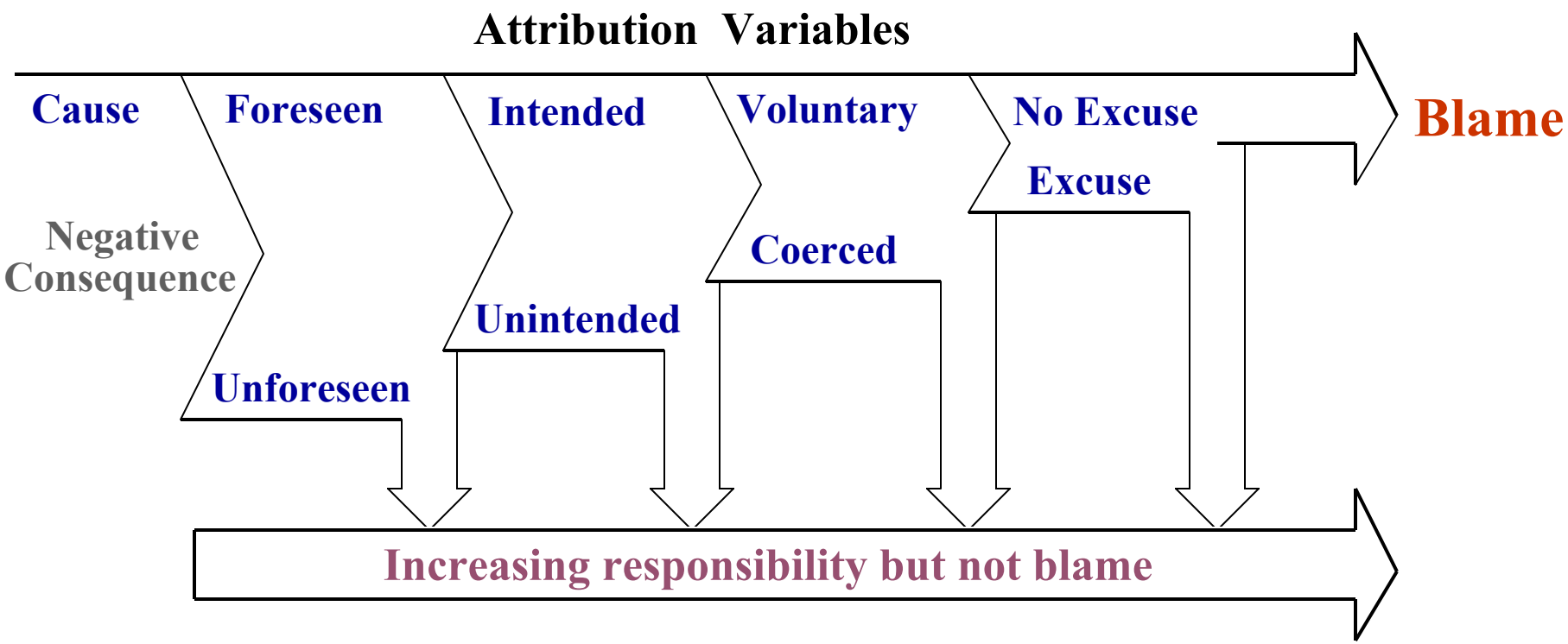
Outline

- Review theory of social causality
- Review computational model of social judgment
  - Presented at last IVA, AAMAS2004
- Evaluate ability to predict human judgments
  - Contrast with alternative models
- Summary and Conclusion

Social Attribution Theory

- Dominant psychological theory of how people form judgments of social cause
- Differs marked from physical causality
  - Incorporates epistemic variables (intent, coercion..)
- Appraisal-theoretic approach (for those that know emotion)
  - Emphasizes cognitive judgments
  - Emphasizes subjective perspective

Social Attribution Theory



Adapted from Shaver [1985]

Example

- An officer orders two marksmen to shoot a man. They strenuously object but the officer insists. The marksmen take careful aim, shoot the man and $\underbrace{\text{he dies}}_{\text{Negative Consequence}}$.

Example

- An officer orders two marksmen to shoot a man. They strenuously object but the officer insists. The marksmen take careful aim, shoot the man and he dies.

Direct Observation +
Knowledge of Action

Marksmen: Cause

Example

- An officer orders two marksmen to shoot a man. They strenuously object but the officer insists. The marksmen take careful aim, shoot the man and he dies.

Causal Inference

Marksmen: Cause → Intent

Example

- An officer orders two marksmen to shoot a man. They strenuously object but the officer insists. The marksmen take careful aim, shoot the man and he dies.

Dialogue Inference

Marksmen: Cause → Intent → Voluntary X

Example

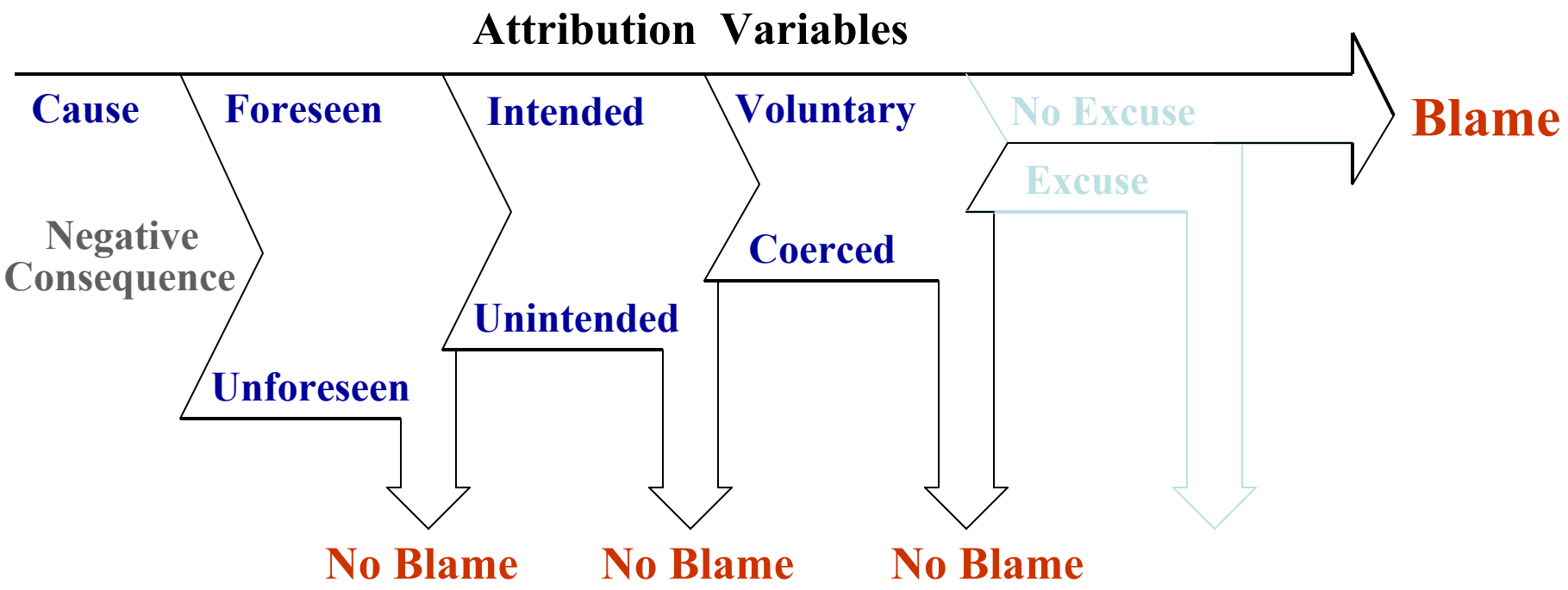
- An officer orders two marksmen to shoot a man. They strenuously object but the officer insists. The marksmen take careful aim, shoot the man and he dies.

Marksmen: Cause → Intent → Voluntary X

Officer: Cause → Intent → Voluntary? → Excuse?

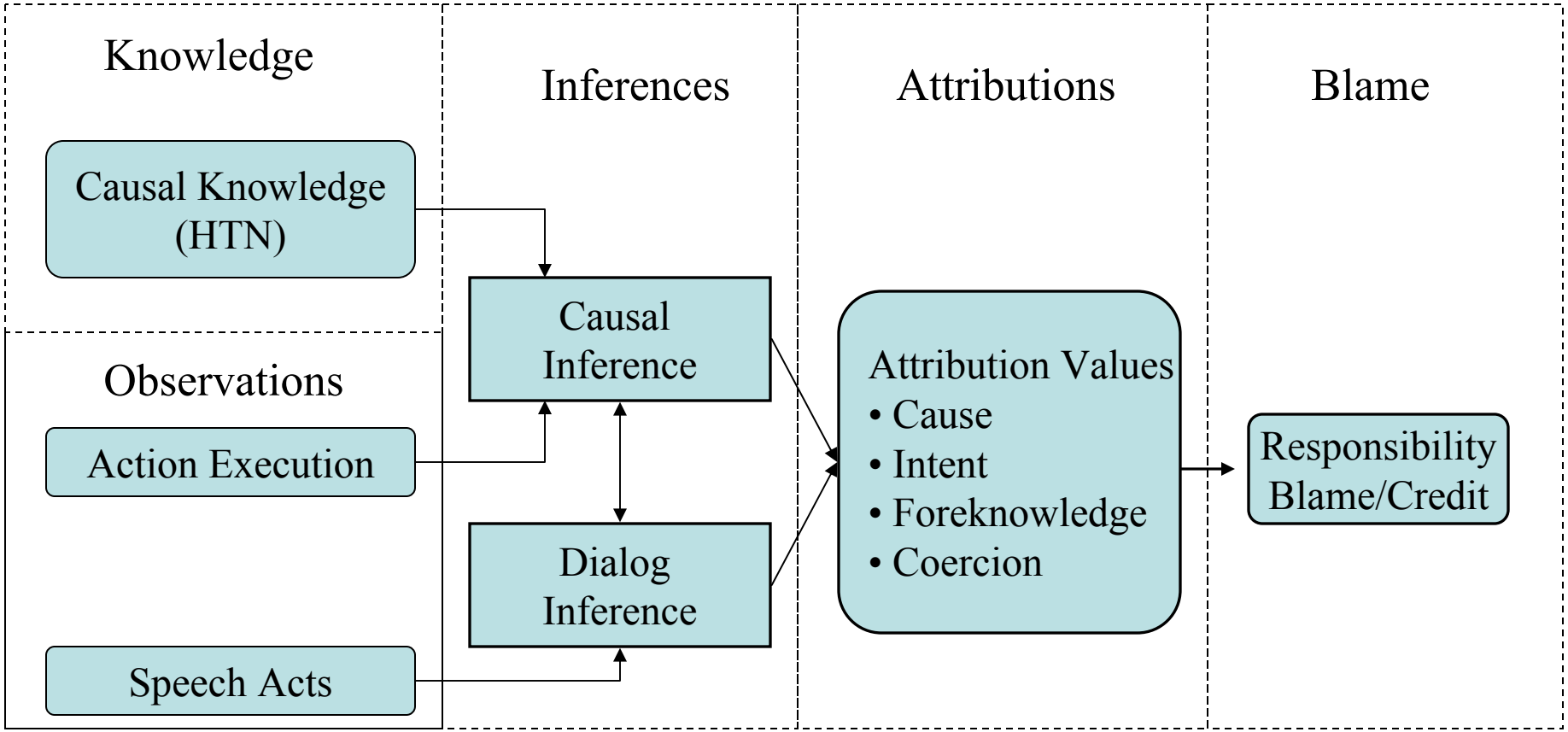
Blame depends on further judgments: was his decision was coerced, was the killing justified

Mao & Gratch 2003 (M&G)



Mao & Gratch 2003 (M&G)

Computational model of attribution theory



Causal & Dialogue Inference

- Foreknowledge

  e.g. if *e* is direct effect of intended act *A* → foreknowledge

- Intention

  Employing general Intention recognition algorithm

  [Mao & Gratch, 2004]

- Coercion

  Dialogue and Counterfactual inference

Example: Coercion

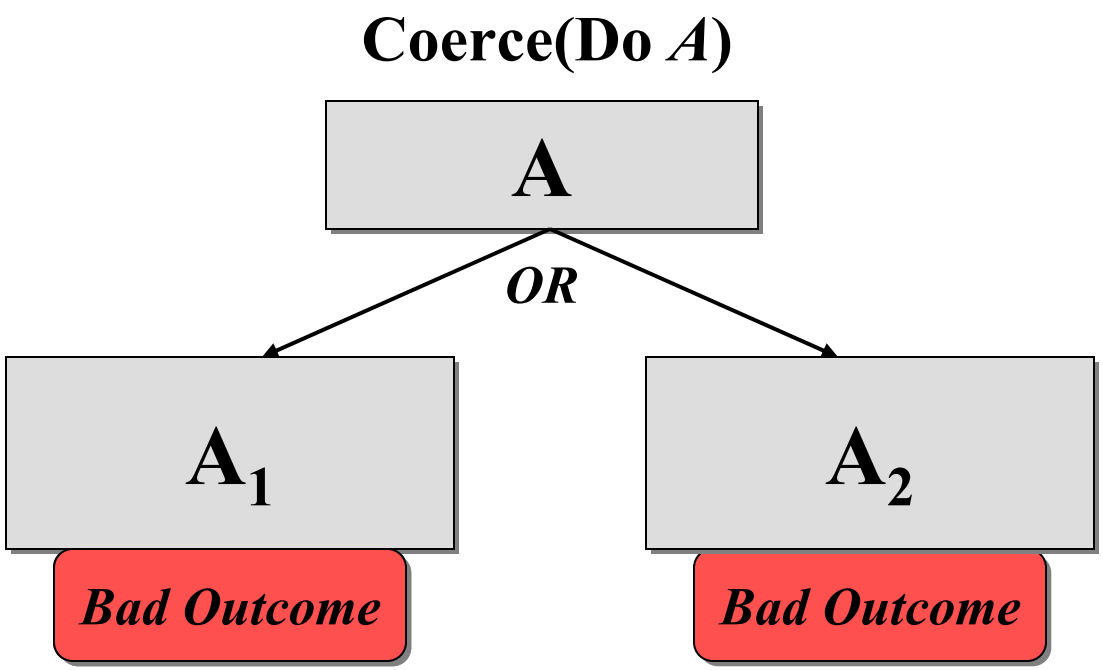
- Some inferences can be quite involved

Officer orders, marksmen reject, officer insists, marksmen accept

$$\textit{Rule}:\ \neg\mathbf{intend}(h, p, t1)) \wedge \mathbf{obligation}(h, p, t2) \wedge \mathbf{accept}(h, p, t3) \wedge t1<t3 \wedge t2<t3<t4 \Rightarrow \mathbf{coerce}(s, h, p, t4)$$

- But act coercion doesn't imply effect coercion
- Also need causal inference

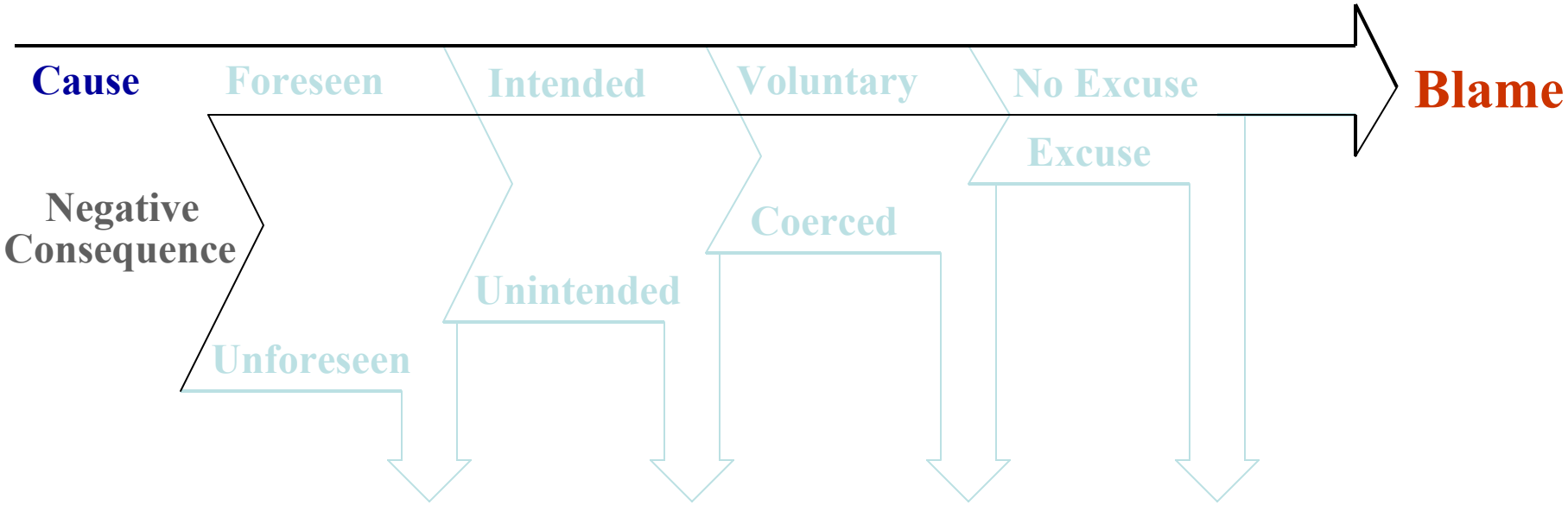
Coercion: counterfactual inference



Subordinate has meaningful choice

No Coercion

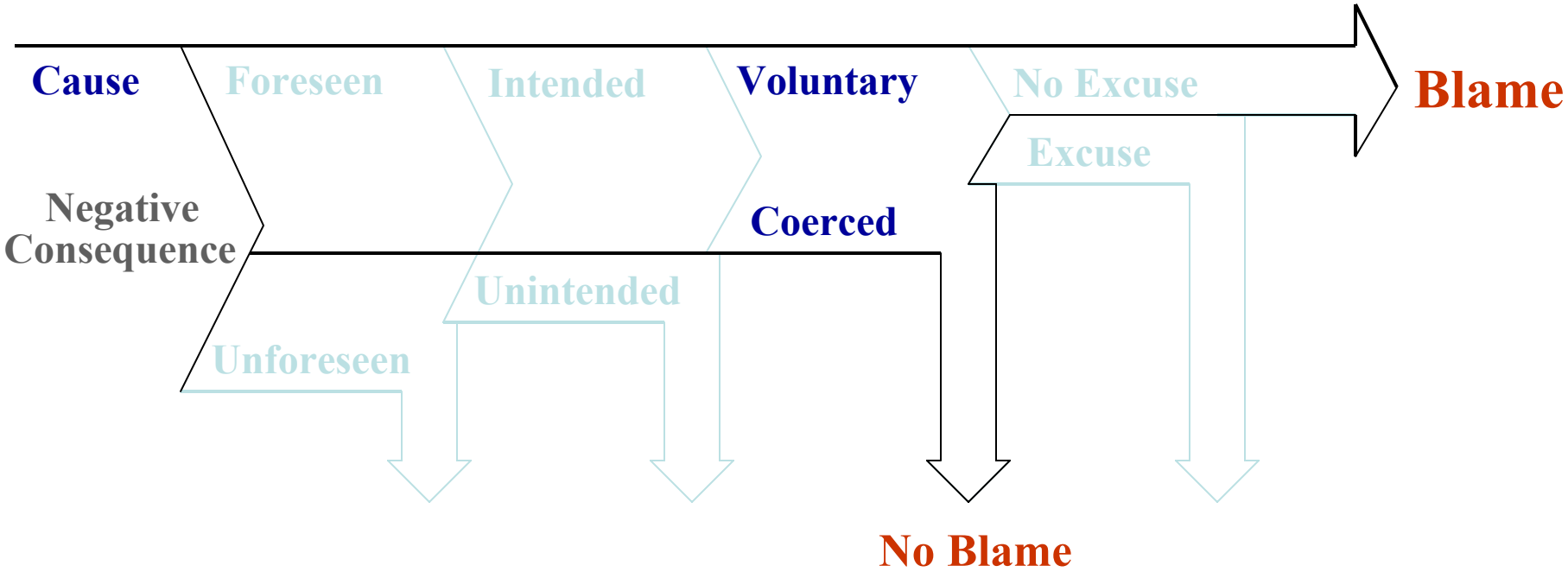
Alternative Models



Blame Physical Cause

- Blame agent that immediately caused consequence

Alternative Models



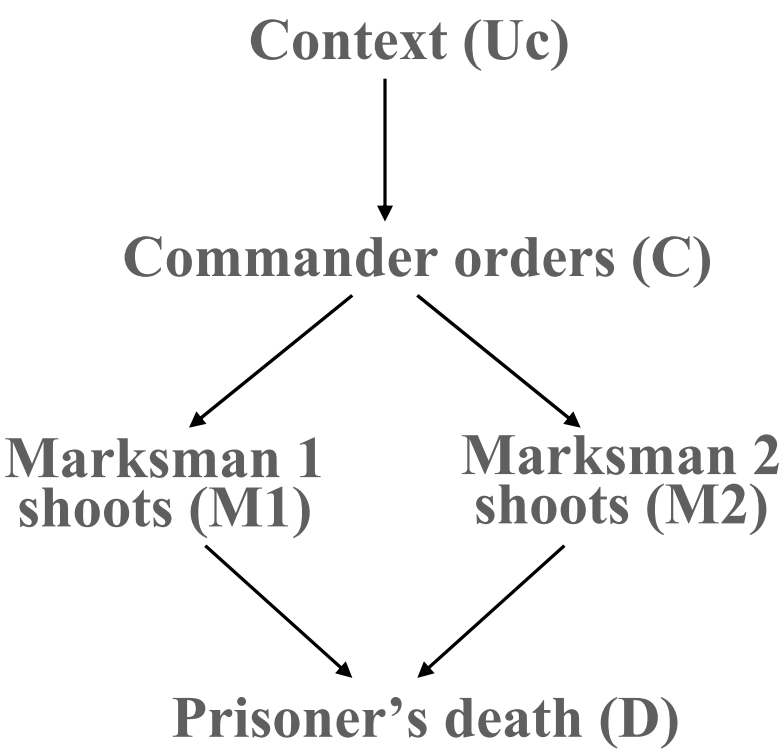
Blame Authority

- Blame highest authority associated with act

C&H (Chockler&Halpern04)

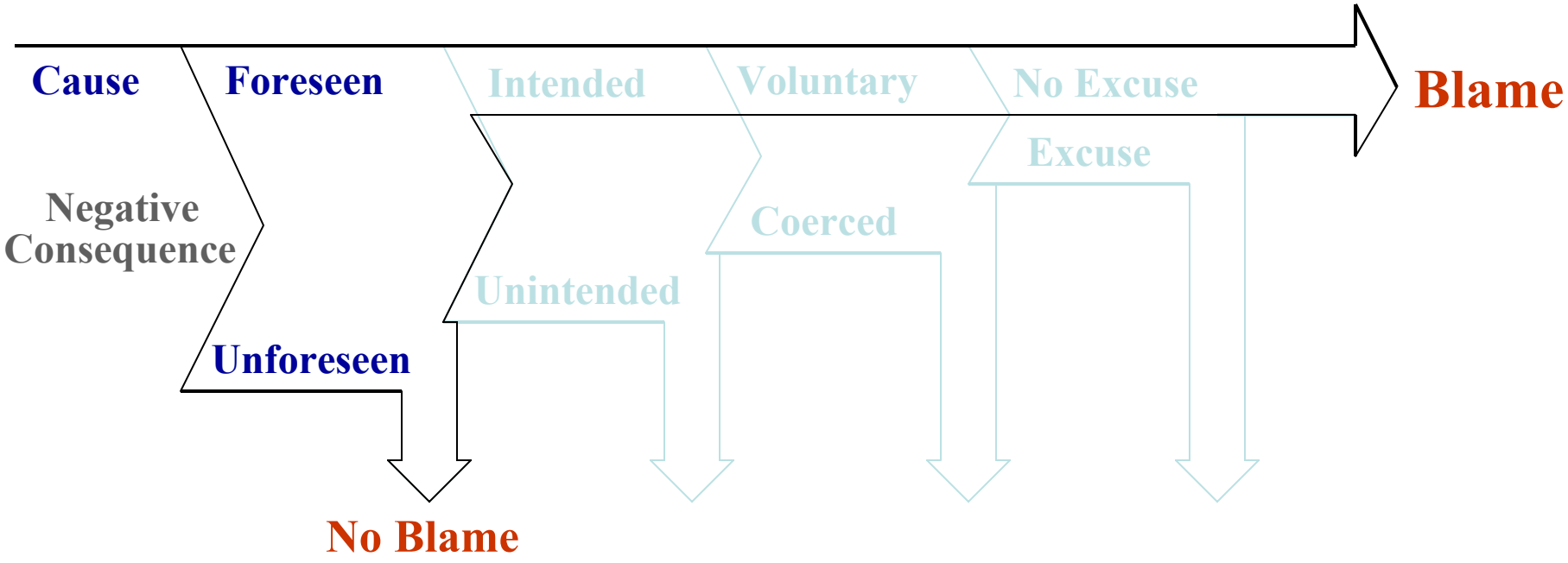
Extension of Pearl's structural-model approach

- Identifies indirect causal factors including social causes



- Causal equations
  $C = M1,\ C = M2$
- Speech treated as a physical action
- *M1=1*, *M2=1* and *C=1* are causes of the death
- Blame shared
  M1=1/2, M2=1/2, C=1

C&H (Chockler&Halpern04)



Claim

- M&G model will better predict human judgments than
  - **Simple causal model**

    Based on physical cause
  - **Simple authority model**

    Always choose highest authority
  - **C&H model**

    Structural model approach to shared responsibility and blame [Chockler & Halpern, 2004]

Method

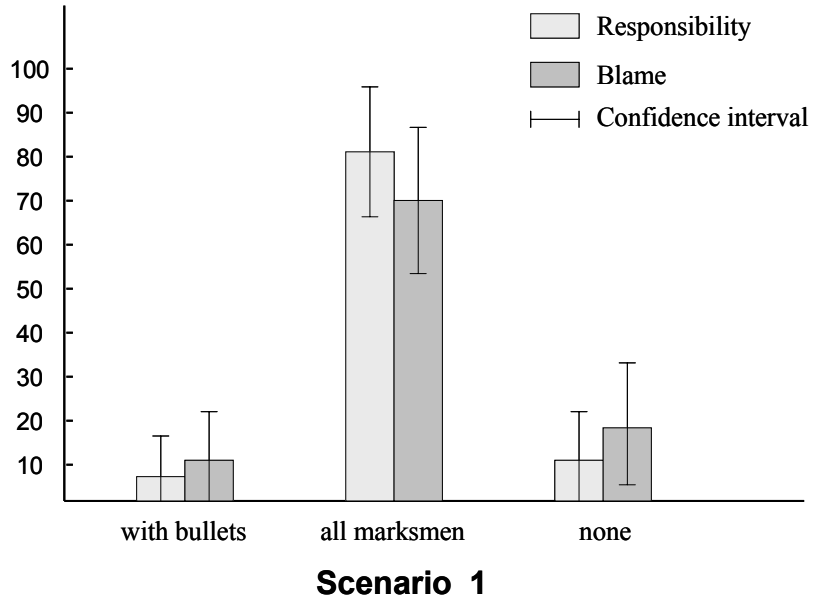
- Constructed 4 variations of C&H04 *firing squad*
  - Vary factors that should influence attributions
  - Emphasized factors influencing coercion
- Encoded variants in each model
  - Checked w/ C&H (personal communication)
- Generated predictions of blame
- Compare predictions with human data
  - Queried 27 Subjects on blame and epistemic variables
  - Judge model predictions against majority response

Scenario 1: No officer

- A firing-squad consists of ten excellent marksmen. Only one marksman has live bullets in his rifle; the rest have blanks. The marksmen do not know who has the live bullets. They shoot at the prisoner and he dies.
- Human majority agreement: *all marksmen*

Blame	Simple Cause Model	Simple Authority Model	C&H Model	Social Inference Model
Predict	marksman with bullets	N/A	all marksmen	all marksmen
Human				

Scenario 1: No officer



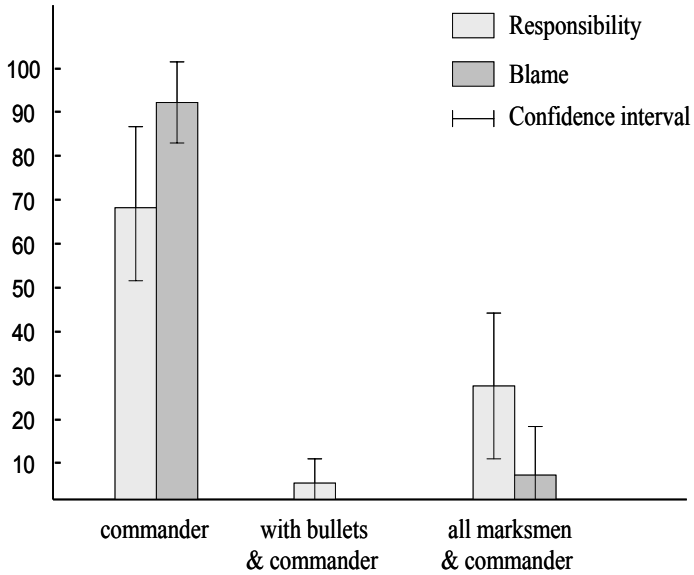
Blame	Simple Cause Model	Simple Authority Model	C&H Model	Social Inference Model
Predict	marksman with bullets	N/A	all marksmen	all marksmen
Human	No	no	yes ☺	yes ☺

Scenario 3: Insubordinate

- The same firing-squad. The commander *orders* the marksmen to shoot the prisoner. The marksmen *refuse* the order. The commander *insists* that the marksmen shoot. They shoot at the prisoner and he dies.
- Human majority agreement: *commander*

Blame	Simple Cause Model	Simple Authority Model	C&H Model	Social Inference Model
Predict	marksman with bullets	commander	all marksmen & commander	commander
Human				

Scenario 3: Insubordinate



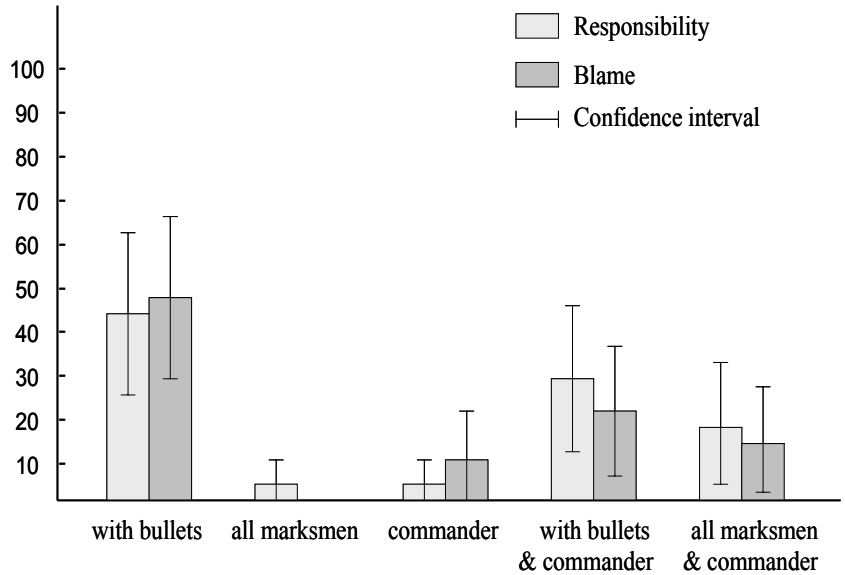
Blame	Simple Cause Model	Simple Authority Model	C&H Model	Social Inference Model
Predict	marksman with bullets	commander	all marksmen & commander	commander
Human	no	yes ☺	no	yes ☺

Scenario 4: marksmen choose

- The same firing-squad. The commander *orders* the marksmen to shoot the prisoner, and each marksman can *choose* to use either *blanks* or *live bullets*. The marksmen shoot at the prisoner and he dies.
- Human majority agreement: *marksman with bullets; commander & marksman with bullets*

Blame	Simple Cause Model	Simple Authority Model	C&H Model	Social Inference Model
Predict	marksman with bullets	commander	N/A	marksman with bullets
Human				

Scenario 4: marksmen choose



Blame	Simple Cause Model	Simple Authority Model	C&H Model	Social Inference Model
Predict	marksman with bullets	commander	N/A	marksman with bullets
Human	Yes* ☺	no	N/A	Yes* ☺

Results: Summary

BLAME	Simple Cause Model		Simple Authority Model		C&H Model		Social Inference Model		Human Majority Agreement
	Results	Match	Results	Match	Results	Match	Results	Match	
S 1	with bullets	no	N/A	N/A	all marksmen	yes	all marksmen	yes	**all marksmen**
S 2	with bullets	no	commander	yes	all marksmen& commander	no	commander	yes	**commander**
S 3	with bullets	no	commander	yes	all marksmen& commander	no	commander	yes	**commander**
S 4	with bullets	yes (partial)	commander	no	N/A	N/A	with bullets	yes (partial)	**with bullets/ w. bullets & commander**

Future Work

- Further refine the computational model
- Experiment on more social scenarios to validate the work
- Implement inference engine and other components of social inference module

Acknowledgements

- Thanks to Joseph Halpern and Andrew Gorden for helpful discussions

This work was sponsored by the U. S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.