面向智能体的认知与社会模拟

毛文吉
中国科学院自动化研究所

认知与社会计算交叉研究

人与社会相关理论的计算模型

人与社会相关理论的计算嵌入

理论模型驱动

结合认知/社会因素的数据处理

人与社会相关数据的分析处理

数据驱动为主

理解

计算

认知与社会计算交叉研究

- **数据驱动的认知与社会计算交叉研究：**
  - 揭示认知和社会现象
  - 从数据中发现相关规律性
  - 提升对认知和社会相关规律的认识

结合认知/社会因素的数据处理

人与社会相关数据的分析处理

数据驱动为主

认知与社会计算交叉研究

人与社会相关理论的计算模型

人与社会相关理论的计算嵌入

理论模型驱动

- **理论模型驱动的认知与社会计算交叉研究：**
  - 人与社会相关理论在计算系统中的嵌入和使用，增强计算系统的认知和社会能力
  - 基于人与社会相关理论的普适性计算模型建立最为困难

关于人与社会的理论

- 理论/模型解释人类行为的某些方面是如何组织的，提供概念描述及其相互关联
- 来自不同学科领域，一般性理论 (General theory) 为主

- Social cognition
- Psychology
- Sociology
- Linguistics
- Decision science
- Anthropology
- Communication

… …

- Motivation theory
- Decision theory
- Emotion theory
- Pragmatics
- Attribution theory
- Organization theory
- Personality theory
- Social network theory
- Social interaction theory

… …

关于人与社会的理论

- 最为重要的理论来自**认知和社会心理学领域**，关于人类对社会行为和心理（情绪）的评估模型
- 以**描述性模型**（Descriptive model）为主
  - **Descriptive model**（描述性模型）
    Describe what people actually do in their everyday lives
  - **Normative/Prescriptive model**（规范性模型）
    Prescribe what people should do, i.e. the ideal criterion they ought to follow
- 支持**符号主义**智能观

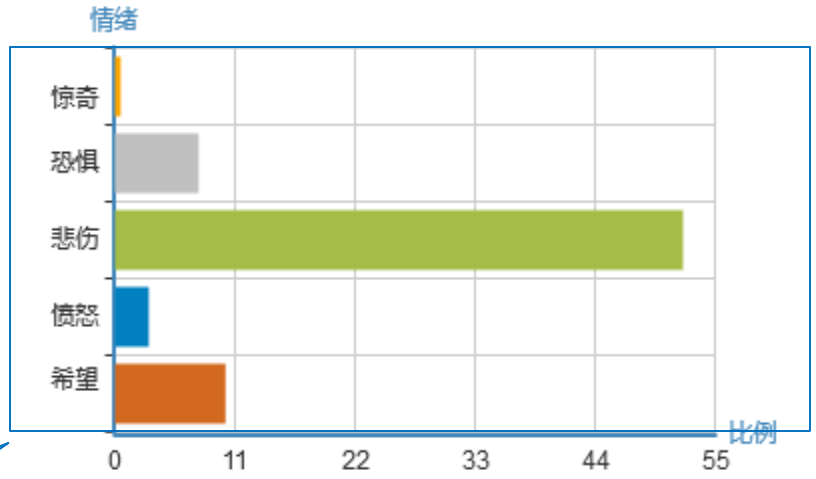
计算嵌入：社会群体情绪归因

事件详细

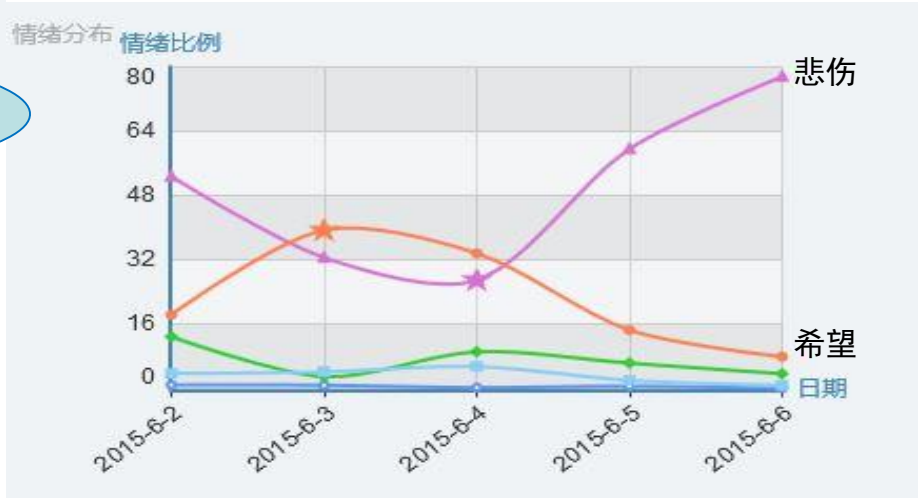
长江游轮倾覆

讨论人数：556008

事件简介：6月1日22时左右，一辆东方之星号游轮在长江水域湖北境内突遇龙卷风发生倾覆。出事船舶载客454人，船长反映船舶在航行途中突遇龙卷风瞬间翻沉。因现场有大风暴雨，搜救困难。



社会情绪



情绪：希望？　恐惧？　悲伤　　愤怒

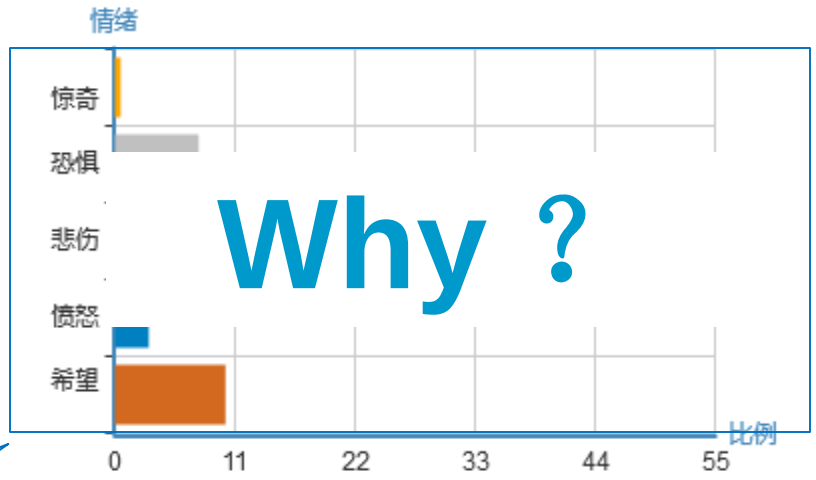
计算嵌入：社会群体情绪归因

事件详细

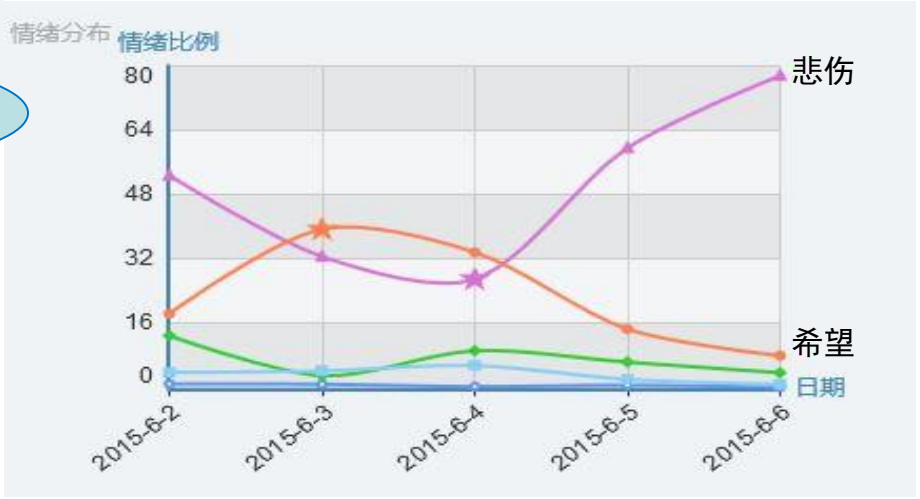
长江游轮倾覆

讨论人数：556008

事件简介：6月1日22时左右，一辆东方之星号游轮在长江水域湖北境内突遇龙卷风发生倾覆。出事船舶载客454人，船长反映船舶在航行途中突遇龙卷风瞬间翻沉。因现场有大风暴雨，搜救困难。



归因分析



情绪：希望？　恐惧？　悲伤　　愤怒

计算嵌入：社会群体情绪归因

- 与传统情感分析及其原因抽取的区别

	传统情感分析及其原因抽取	社会情绪归因
角度	作者自身	他人/受众
个体/群体	个体	群体
任务类型	抽取	推断
输出	情感正负极性	不同情绪类型
因果解释	原因子句	情绪维度

客观报道 vs. 主观情感表达

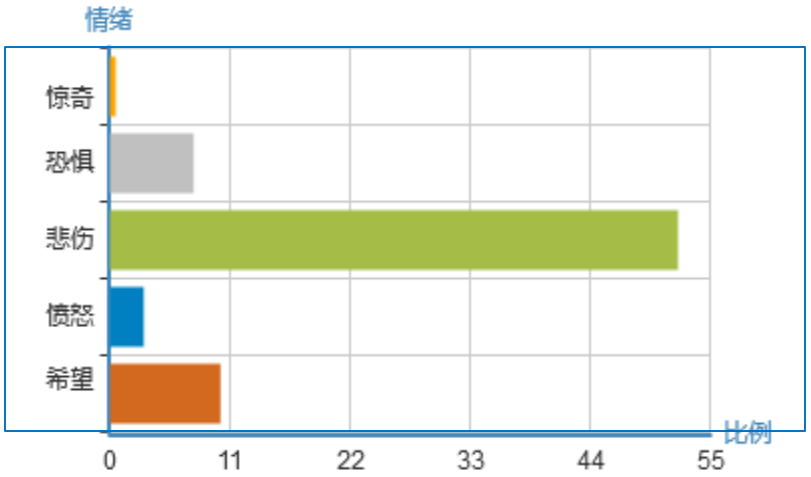
计算嵌入：社会群体情绪归因

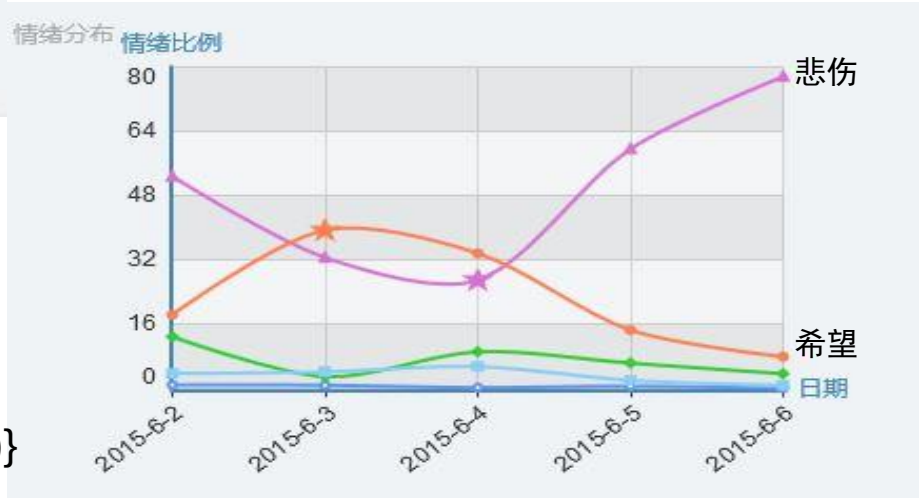
事件详细

长江游轮倾覆

讨论人数：556008

事件简介：6月1日22时左右，一辆东方之星号游轮在长江水域湖北境内突遇龙卷风发生倾覆，出事船舶载客454人，船长反映船舶在航行途中突遇龙卷风瞬间翻沉，因现场有大风暴雨，搜救困难。





归因：已发生 → 不合意 → 应受责备

情绪：希望？ 恐惧？ 悲伤 愤怒

IF {情绪维度，取值} = {(合意性, ***D/U***); (可能性, ***U***)}

THEN 情绪类型 = “**希望/恐惧**”

基于理论的计算模型建立

理论模型

只有抽象的概念及其关系描述

计算建模

计算模型

自动化

计算模型

外部事件

个体信念

心智状态

Mapping

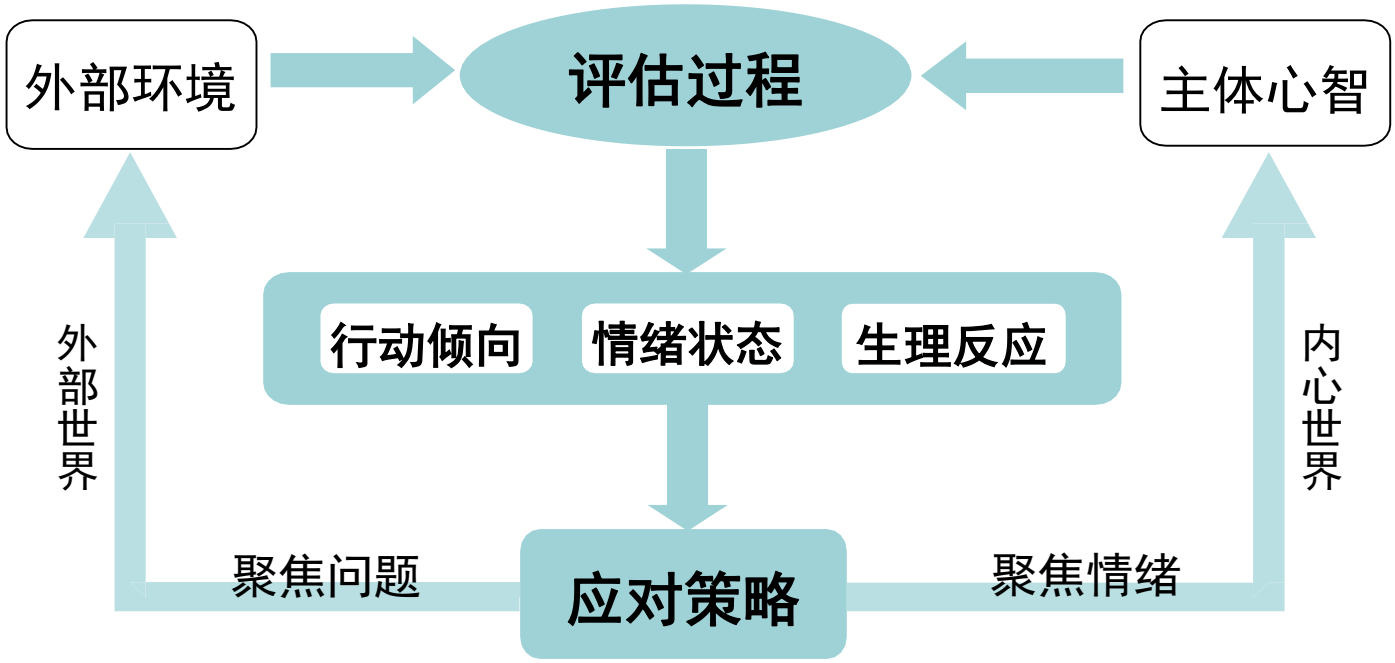
理论模型的变量取值

建立计算模型的困难性

- **计算模型的建立涉及多方面研究挑战：**
  - 理论中的描述性概念（变量）的形式化表示
  - 建立从智能系统获取的输入到变量的自动推演
    （Ideally）尽可能利用智能系统中已有的知识和表示
  - 领域适用的**普适性计算模型**的建立
    采用领域无关的表示和自动推演机制
  - 对所建立的计算模型进行**人的实验验证**
- 目前真正意义上基于描述性理论模型、适用于不同领域且经过人的实验验证的**计算模型尚非常少**

情绪计算模型

- **EMA计算模型**（Gratch & Marsella）



S. Marsella and J. Gratch. EMA: A Process Model of Appraisal Dynamics. *Journal of Cognitive Systems Research*, 10(1):70-90, 2009.

Mission Rehearsal Exercise (MRE)

- 大型虚拟演练环境（aka **Serious Games**）
  - 基于VR/AR战场仿真的作战能力
  - 高风险/高压下的指挥决策才能
  - 冲突解决、战术语言、谈判能力等

Carmen's Bright IDEAS (*e*-drama)

- 帮助儿科肿瘤患者的母亲学习应对和解决问题技能
- 通过交互式场景逐层展开故事情节：
  - 根据母亲作出的选择展开不同的情节
  - 虚拟教师引导母亲学习解决问题技能

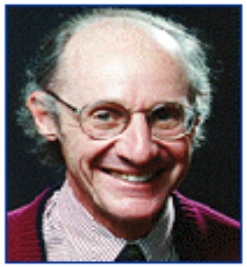
提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

Weiner's Theory of Achievement Motivation



B. Weiner

- An appraisal (评估) theory for a particular human motivation domain
  - Achievement-related setting (1986, 1995, 2001, 2006)
- A lifespan approach to the ***Cognition-Emotion- Behavior*** linkage
  - From thinking to feeling to acting
  - Both intrapersonal (内心的) and interpersonal (人际间的) theory
- Empirically ***validated*** theory and capable of many issues
  - Large empirical efforts; meta-analytic approach (2004)
  - Achievement evaluation, responsibility judgment, excuse giving

B. Weiner. *An Attributional Theory of Motivation and Emotion*. Springer-Verlag, 1986.
Social Motivation, Justice and the Moral Emotions: An Attributional Approach. Lawrence Erlbaum, 2006.

Human Motivation

- No single agreed-on definition on motivation
- Analysis of motivation involves the explanation of why people and animals initiate, choose, or persist in specific actions in specific circumstances
- The spring of all motivational behavior is to increase pleasure and reduce pain (i.e. hedonism)

Intrapersonal Motivation

- Beliefs about causal properties (i.e. dimensions) determine motivational consequences
- **Three causal dimensions:**
  - ***Locus*** of causality: internal/external
  - ***Stability***: duration of a cause
  - ***Controllability***: volitional alteration

	Internal	External
Stable	Aptitude	Task Difficulty
Unstable	Short-term Effort	Luck

	Stable	Unstable
Controllable	Laziness	Short-term Effort
Uncontrollable	Aptitude	Temporary Illness

From Thinking to Feeling

- Stability to ***expectancy*** of success/failure
- Internal cause

  Success → ***pride*** and increments in ***self-esteem***

  Failure → decrements in ***self-esteem*** 自尊
- Internal cause for negative outcome

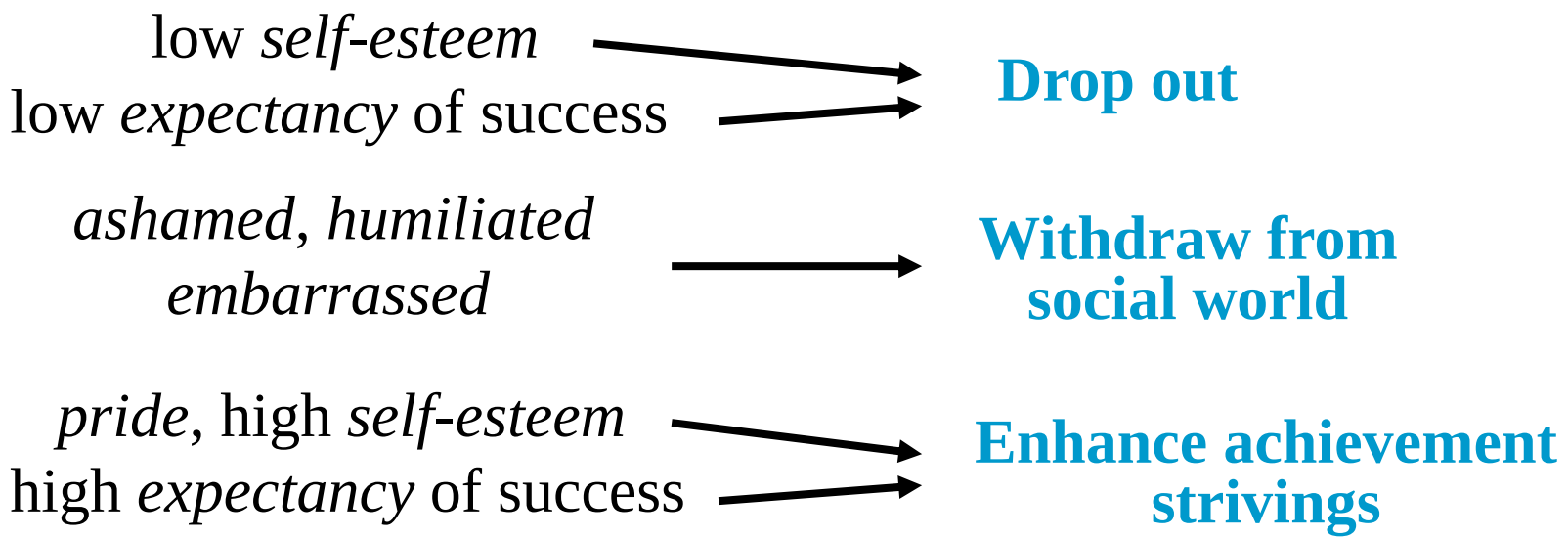
  Controllable → ***guilt***

  Uncontrollable → ***shame***

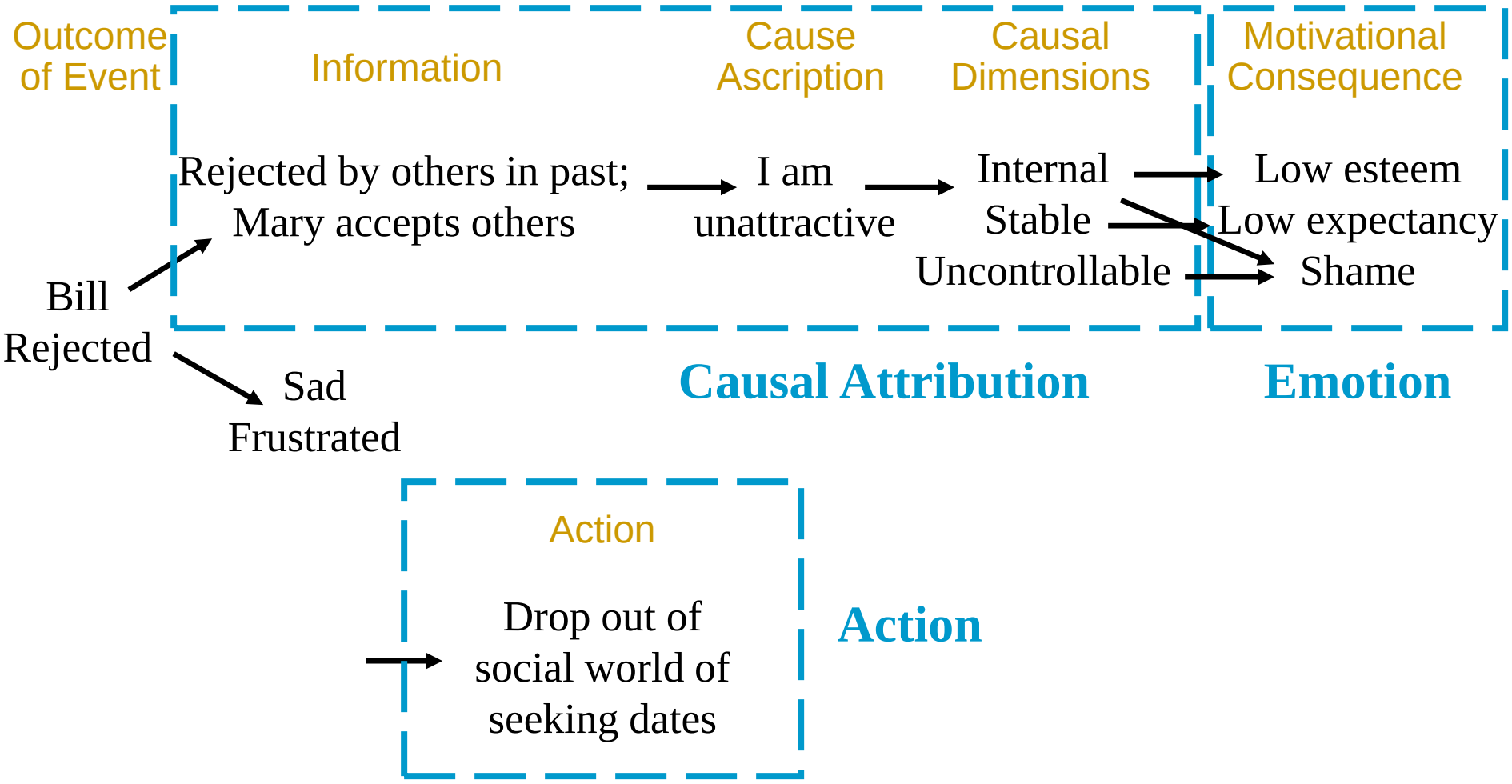
  (Other-affects will be examined later.)

From Feeling to Acting

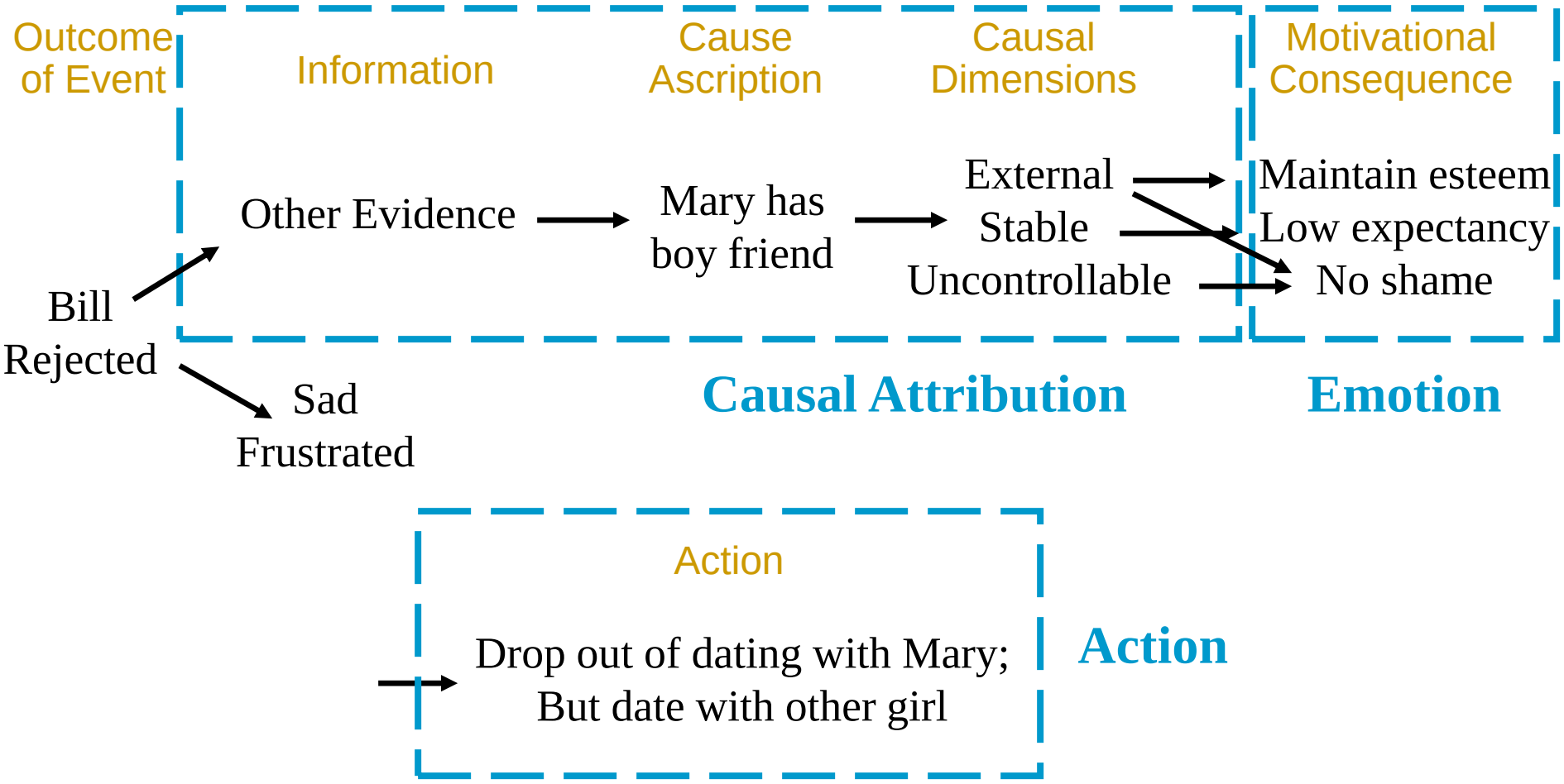
- *Expectancy* of goal attainment, together with *emotions* determine subsequent behavior, e.g.



Dating Example



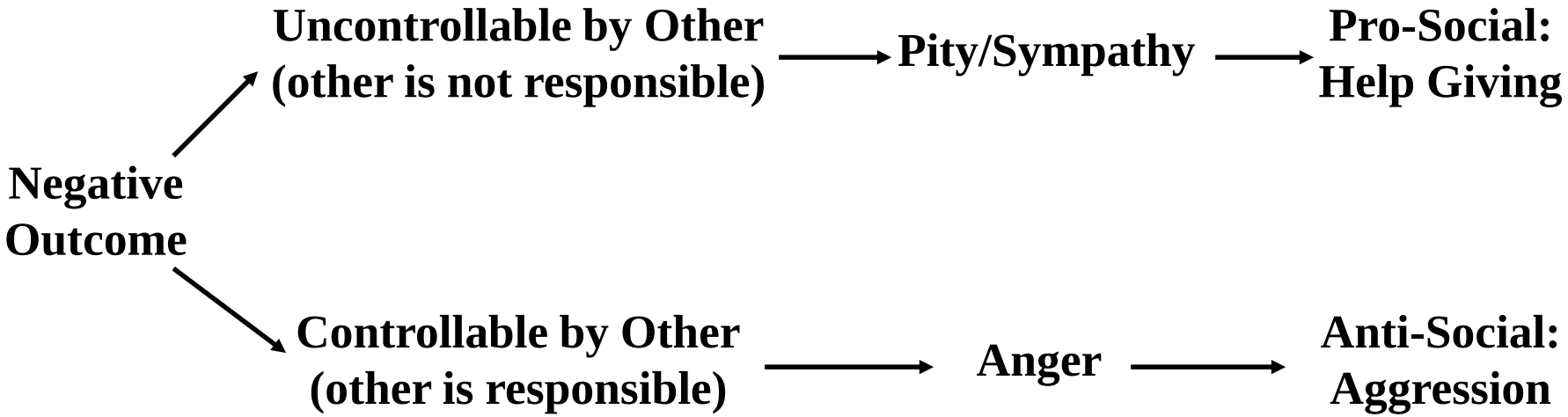
Dating Example



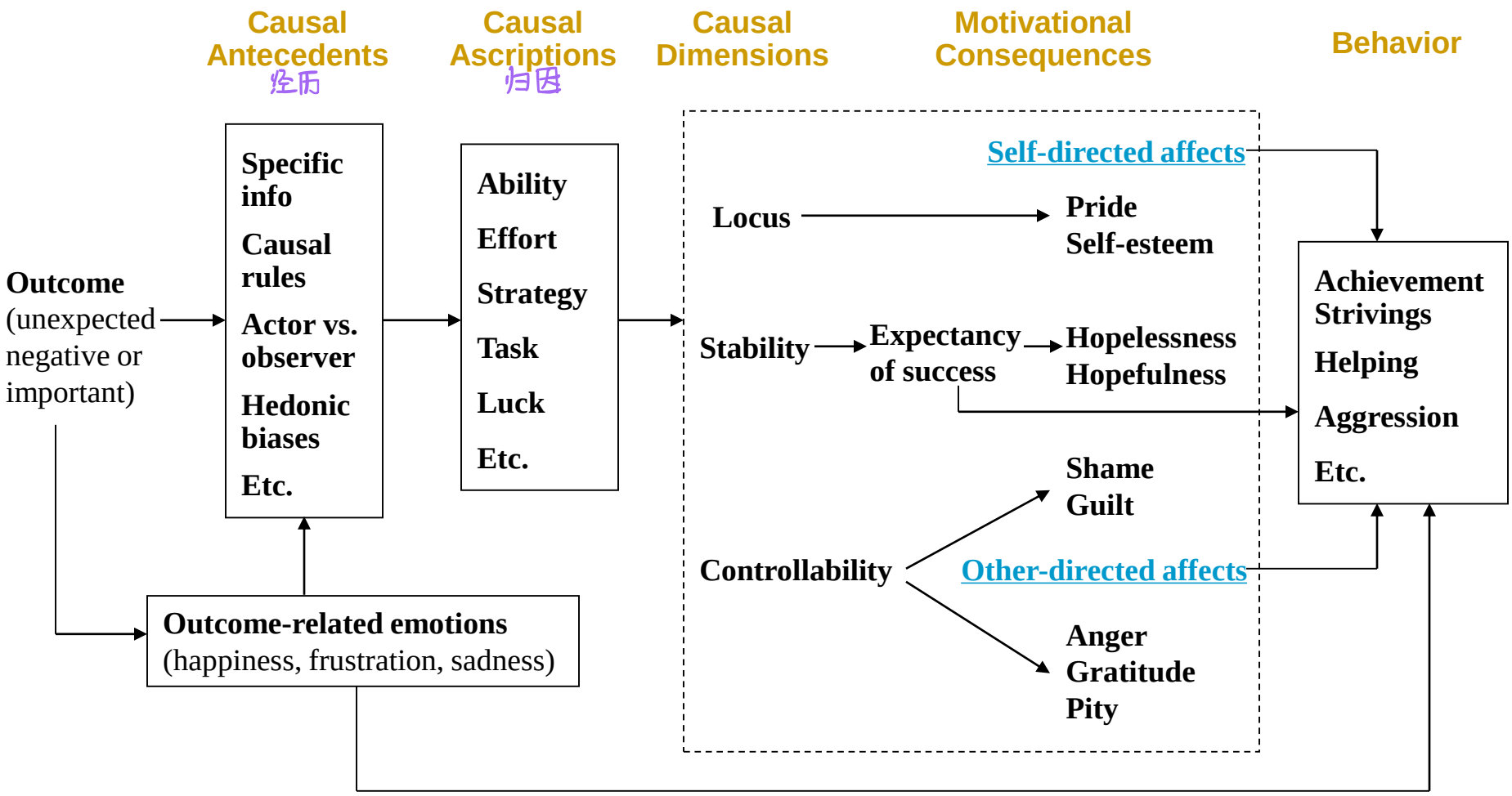
Interpersonal Motivation

- Causal search by an involved ***other***, not actor
- Same causal dimensions for attribution
- Same thinking-feeling-acting sequence
- Eliciting ***social*** motivations and ***social*** actions
- Perception of ***responsibility*** mediates between outcome and reaction to individual

Thinking-Feeling-Acting Linkage



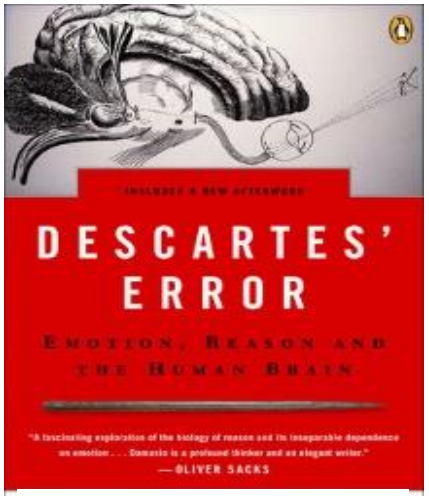
Integrating Theoretical Frameworks



提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

Descartes' Error (1995)



Antonio Damasio

Human organisms are endowed from the very beginning with a **spirited passion for making choices**, which the social mind can use to build *rational behavior*

- Emotions are not a luxury, they are **essential** to rational thinking and to normal social behavior
- Emotions as the *source* of a person's true being

Emotion Theories in Cognitive Psychology

- **Bottom up theories** 自底向上
  - Emotion influences cognition 情感影响认知

    Influence of happy/sad music (Clore)

- **Top down theories**
  - Cognition influences emotion

    Appraisal Theory (Arnold, Lazarus, Frijda, Scherer):

    Emotion arises from an ***evolving subjective interpretation*** of person's relation to their environment and informs behavior

进化的 主观 解释

Cognitive Appraisal Theory

- **Influential and well-established theory**

  Arnold; Frijda; Lazarus; Ortony, Clore & Collins; Scherer; Smith; etc

- **Emphasizes tight coupling between**
  - Emotion
  - Cognition
  - Motivation

Appraisal Theory Perspective

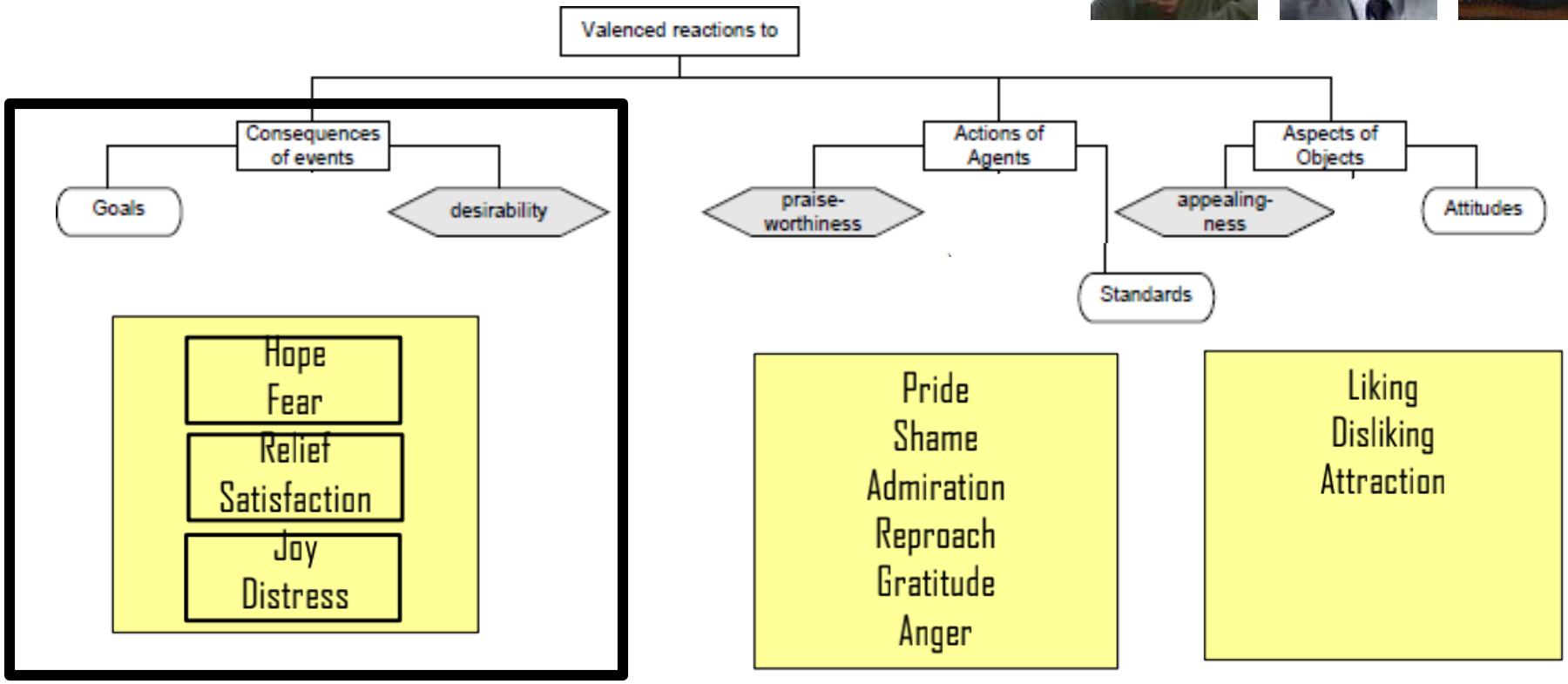
- Emotion is "**goal relevant**"
- Emotion arises from how events impact goals
- Emotion prepares body and mind to address goal threats or opportunities

- Emotion is said to be "**endogenous**" 内生的
  - Meaning that it is goal/task relevant

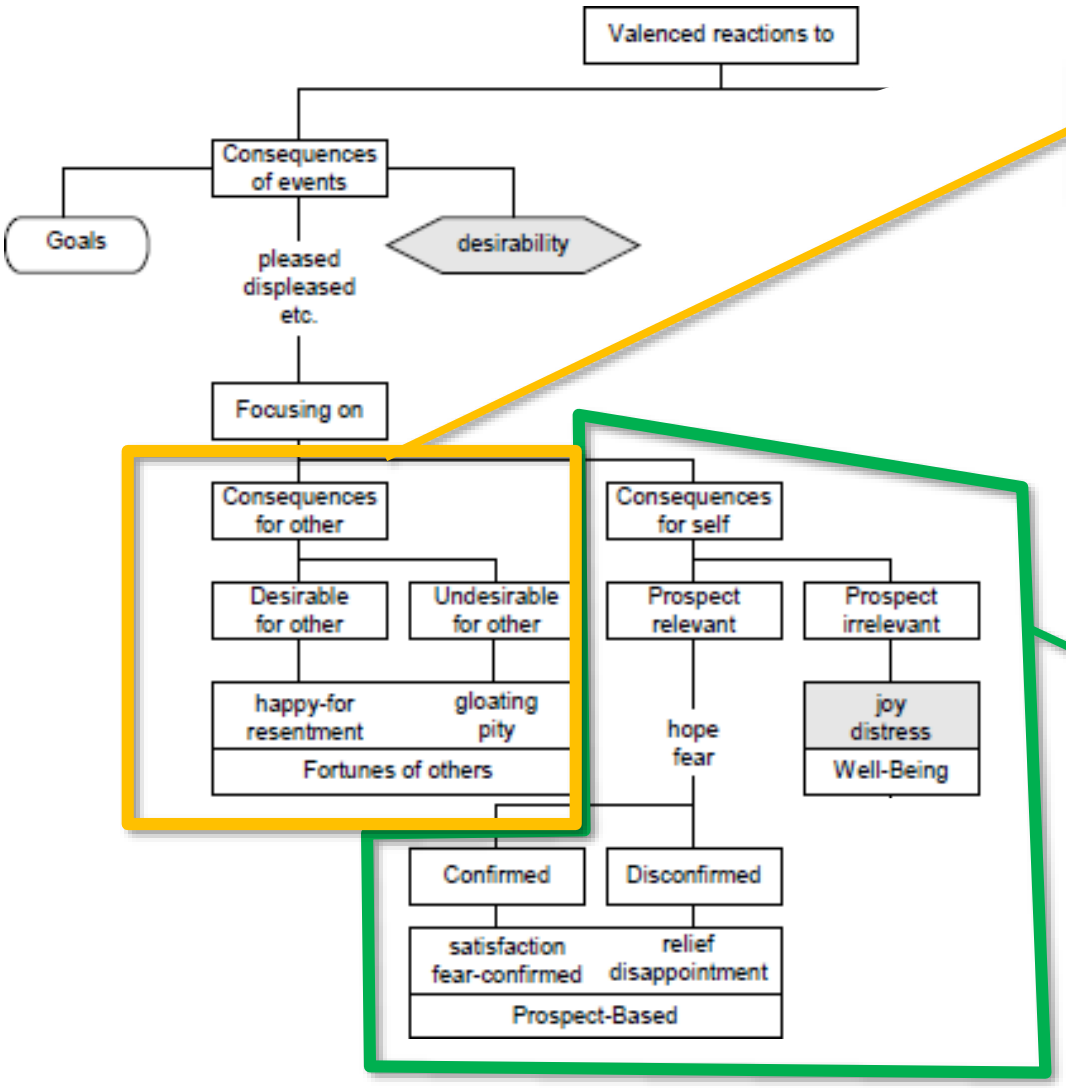
A Structural Theory of Emotion: OCC

(Ortony, Clore and Collins)





A. Ortony, G. Clore and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.



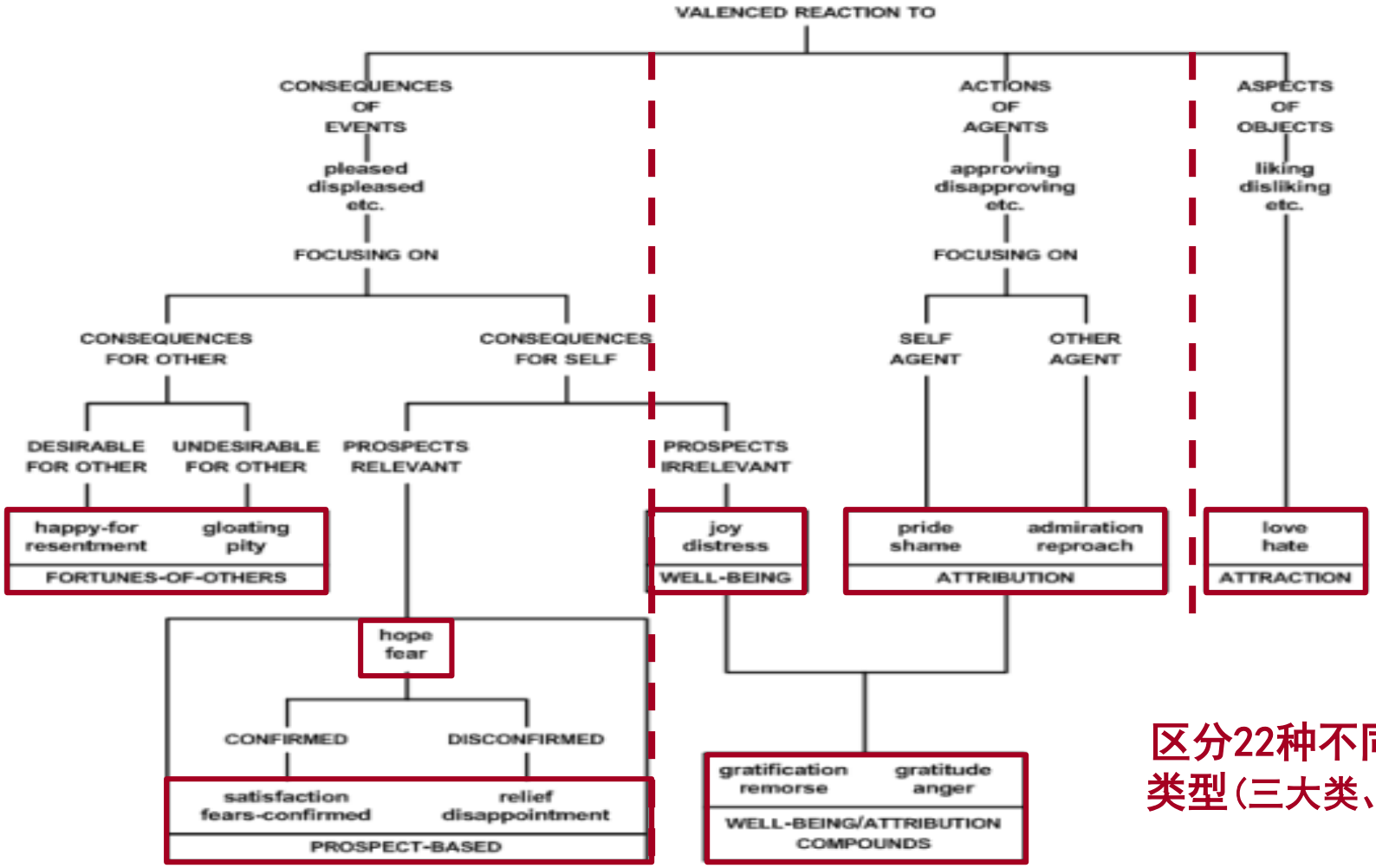
Other-relevant Emotions

Pity, gloating, happy-for,

Self-relevant Emotions

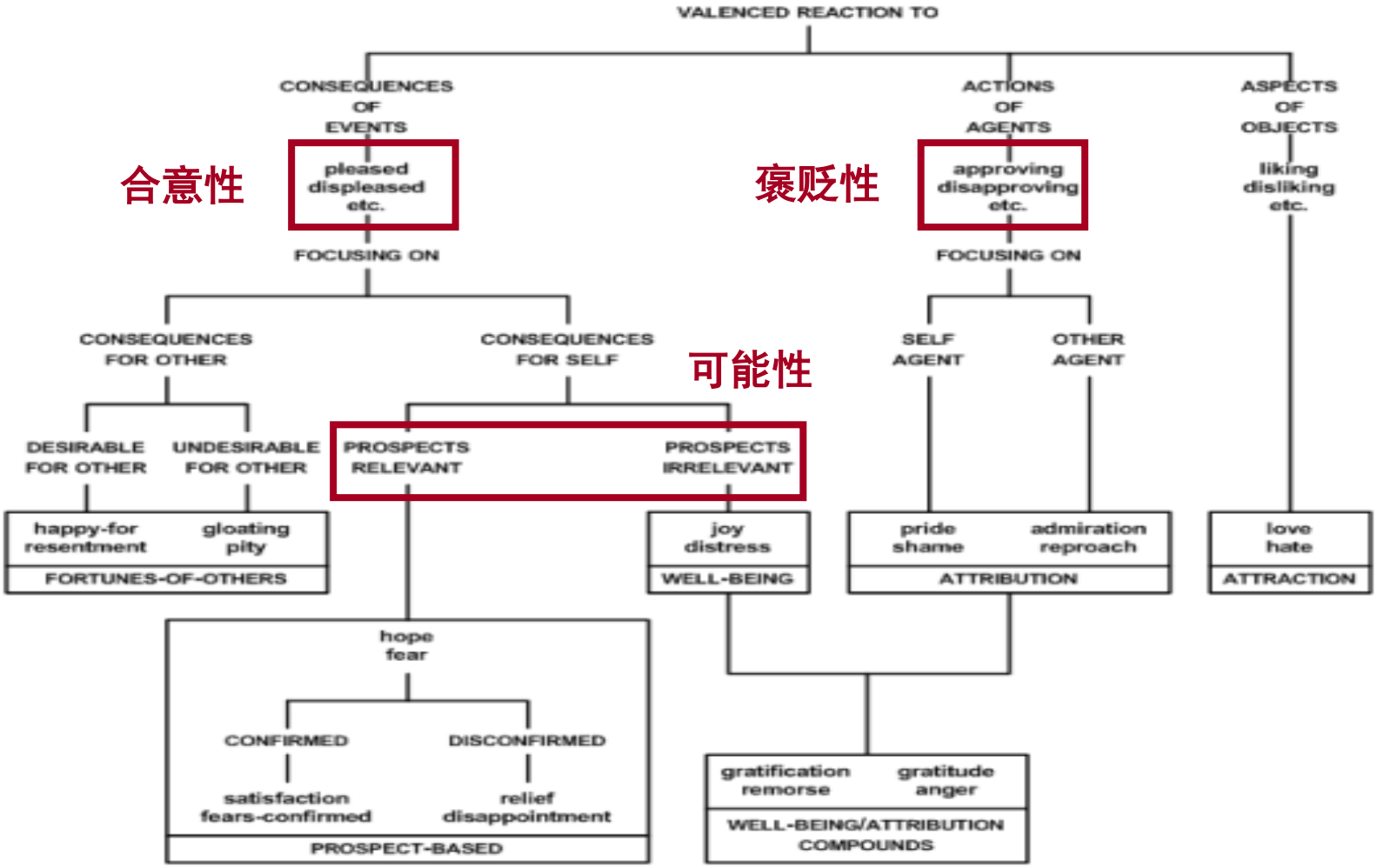
Hope, Joy, fear

A Structural Theory of Emotion: OCC



区分22种不同情绪类型(三大类、六组)

A Structural Theory of Emotion: OCC

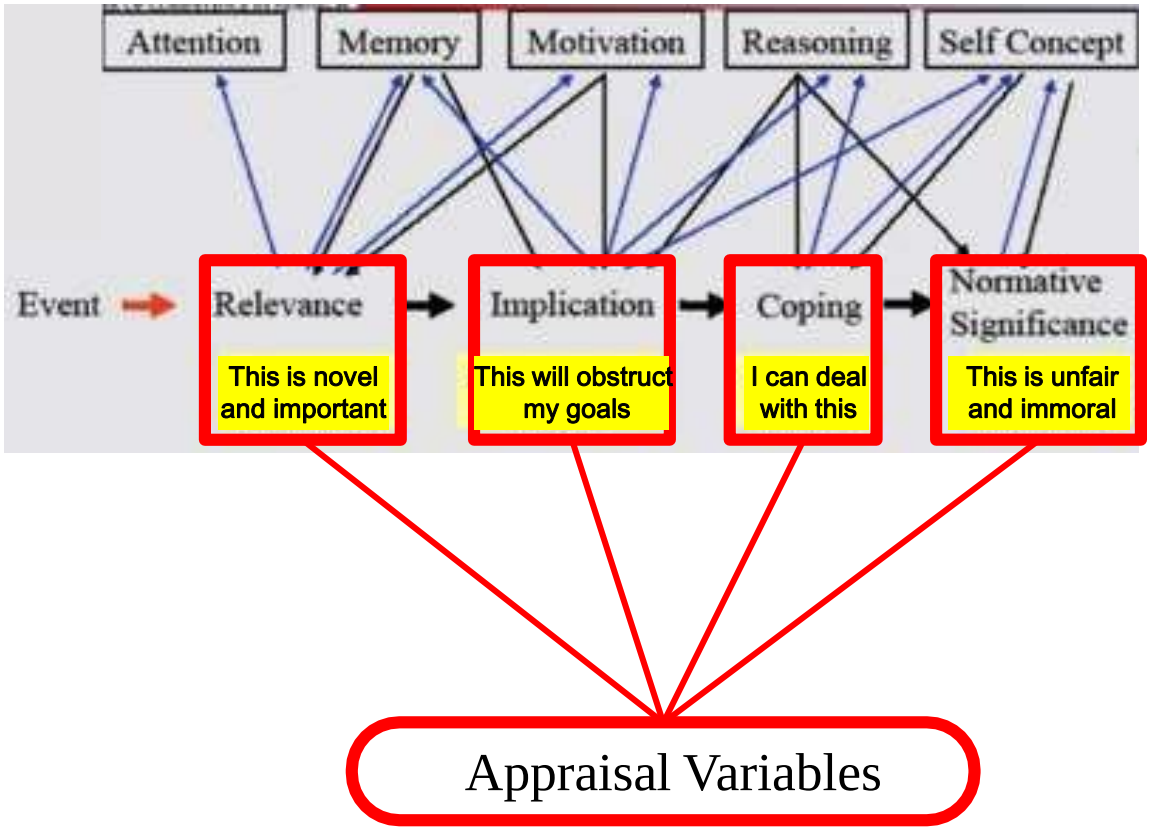


情绪类型：*Joy*，*Distress*，*Hope*，*Fear*，*Pride*，*Shame*，*Gratification*，*Remorse*
Admiration，*Reproach*，*Gratitude*，*Anger*，*Happy-for*，*Pity*，*Resentment* …

A Process Theory of Appraisal

- **Scherer: Sequential Checking Theory**
  - Claims appraisals are derived according to a fixed sequence
  - Results in an unfolding sequencing of emotional responses

Klaus Scherer

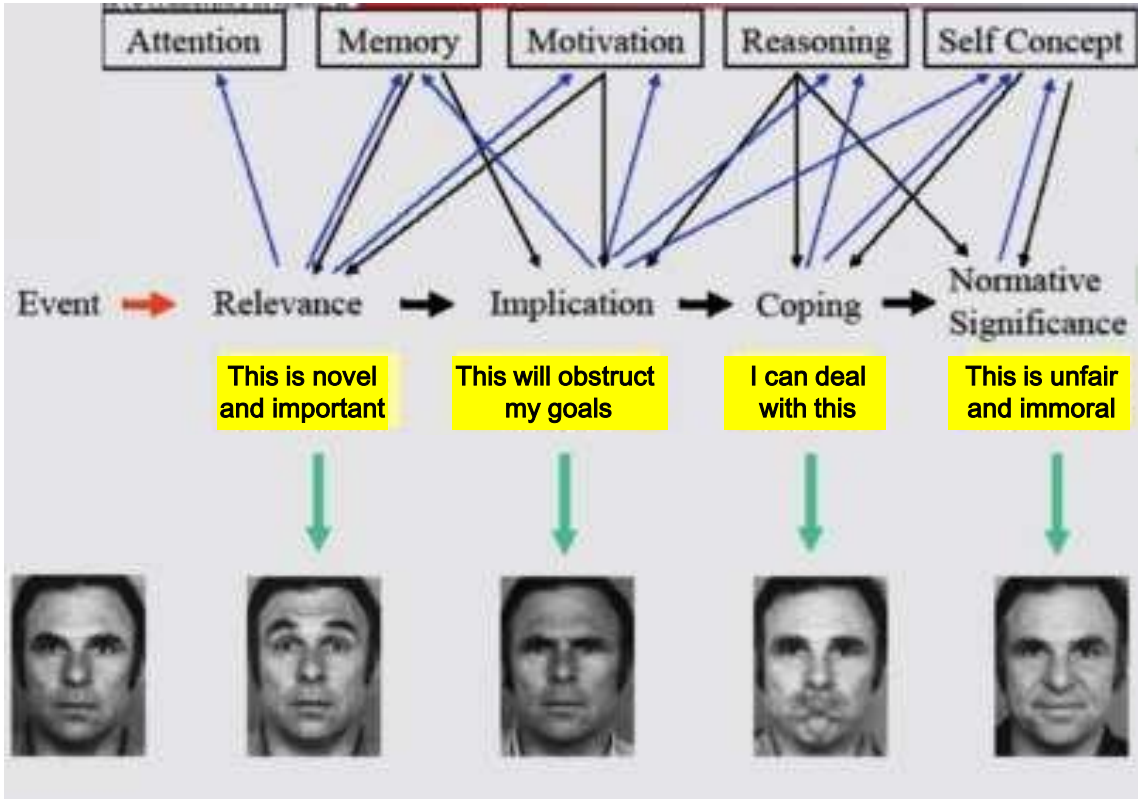


Note:
Implication = OCC desirability
Normative Sig. = blameworthy

A Process Theory of Appraisal

- **Scherer: Sequential Checking Theory**
  - Claims appraisals are derived according to a fixed sequence
  - Results in an unfolding sequencing of emotional responses

Klaus Scherer



Note:
Implication = OCC desirability
Normative Sig. = blameworthy

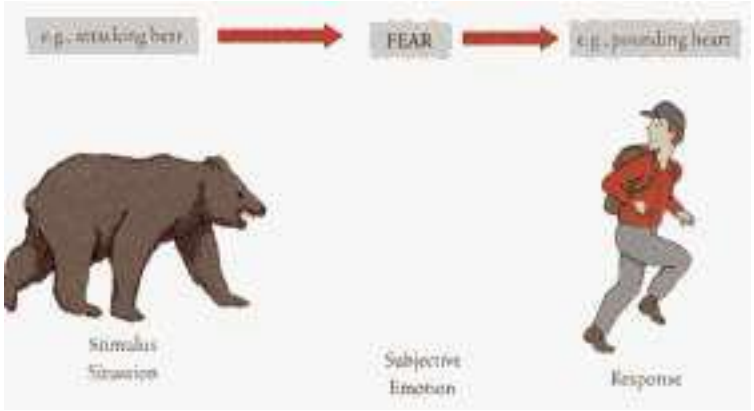
- **More recent and influential (in psychology) than OCC**

Brief Note on Appraisal Theory

TABLE 1
Convergence of Sets of Appraisal Criteria as Suggested by Different Appraisal Theorists

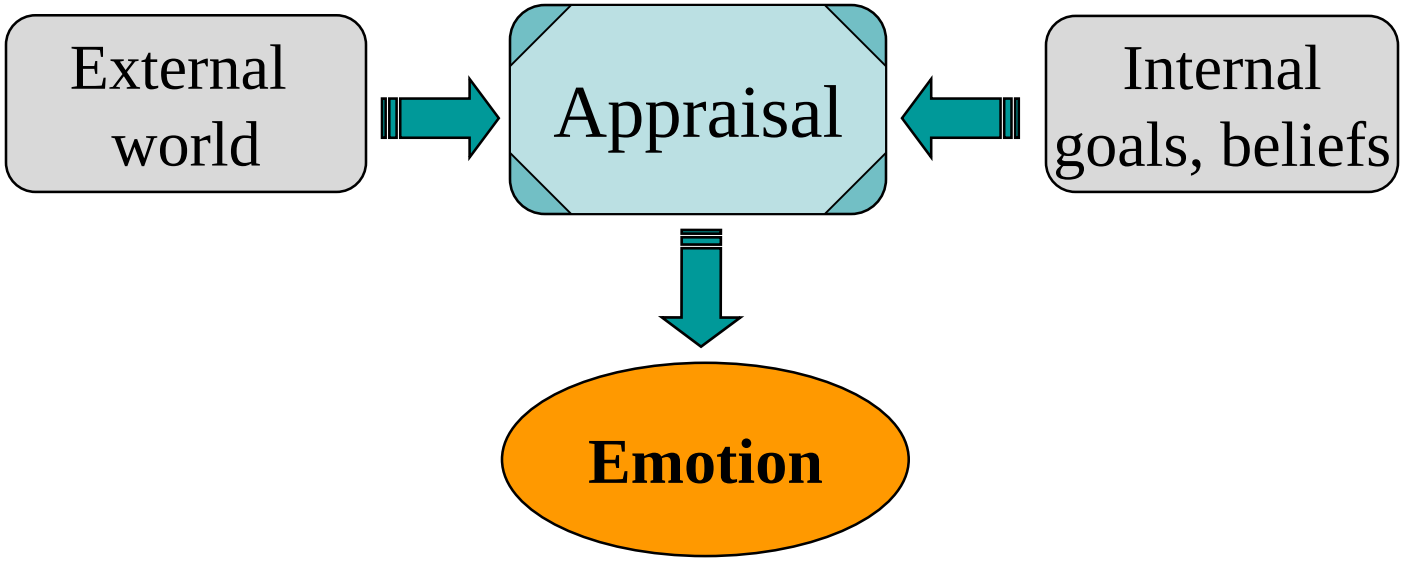
Scherer	*Frijda*	*Ortony/Clore*	*Roseman*	*Smith/Ellsworth*	*Solomon*	*Weiner*
Novelty	Change			Attention		
Suddenness						
Familiarity	Familiarity					
Predictability		Unexpectedness				
Intrinsic pleasantness	Valence	Appealingness		Pleasantness		
Goal significance						
Concern relevance	Focality		App/Ave Motives		Scope/Focus	
Outcome probability	Certainty	Likelihood	Probability	Certainty		
Expectation	Presence	Prospect realisation				
Conduciveness	Open/Closed	Desirability	Motive consistency	Goal/Path obstacle	Evaluation	
Urgency	Urgency	Proximity		Anticipated effort		
Coping potential						
Cause: Agent	Intent/Self-Other	Agency	Agency	Agency	Responsibility	Locus of causality
Cause: Motive				Agency		Stability
Control	Modifiability					Controllability
Power	Controllability		Power		Power	Controllability
Adjustment						
Compatibility standards		Blameworthiness				
External	Value relevance			Legitimacy		
Internal						

- **Appraisal** theory argues, in a sense, beliefs determine emotion
- If I believe the bear is friendly, I won't be afraid



- **Coping** strategies deal with strong negative emotion

Appraisal



- **Appraisal = Situation assessment**

  Compare beliefs, desires and intentions with external circumstances

Appraisal

- **Characterize via *appraisal variables***
  - Desirability
  - Likelihood
  - Urgency
  - Unexpectedness
  - Causal attribution (causality, agency, blame/credit)
  - Coping potential (controllability, adaptability)

Appraisal

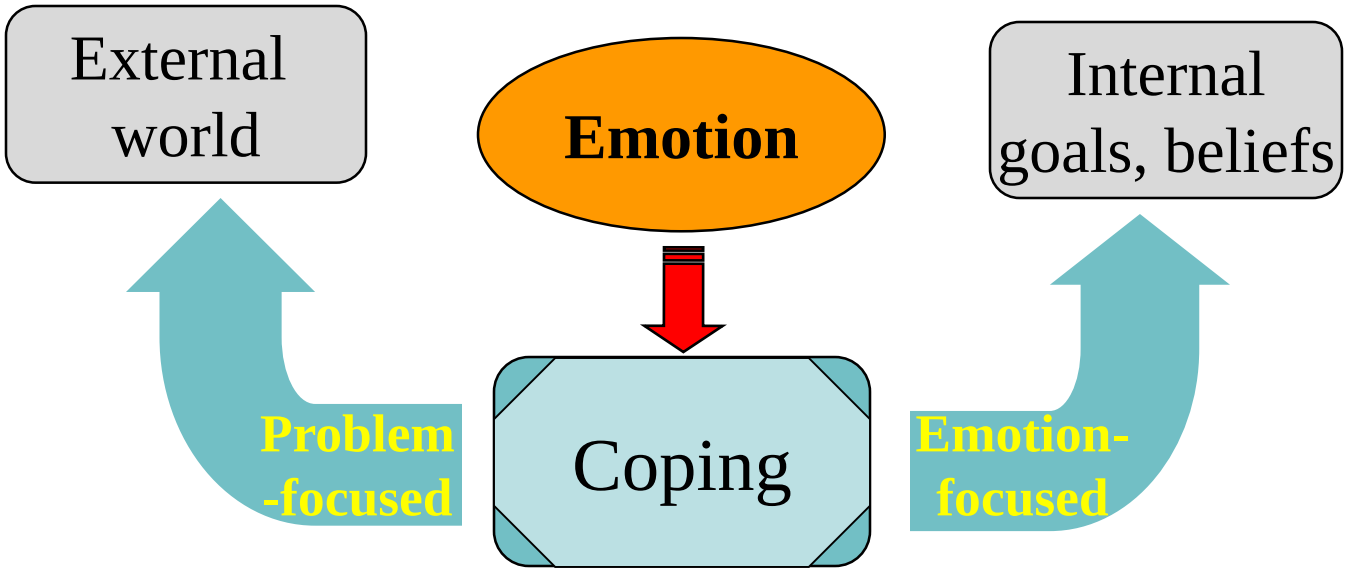
- **Emotions defined in terms of configurations of appraisal variables**
  - Undesirable, Uncertain → ***Fear***
  - Desirable, Certain → ***Joy***
  - Undesirable, Caused-by (Self) → ***Regret***
  - Undesirable, Caused-by (Other), Blame(Other) → ***Anger-at*** (Other)

Coping Strategies

- **Coping = Response strategy**

  Characterized by ontology of coping strategies

本体论



Coping Strategies

- **Problem-focused (act on external world)**
  - Action execution
  - Planning
  - Make amendments
  - Seek instrumental social support
- **Emotion-focused (act on internal beliefs)**
  - Denial / Wishful thinking
  - Find silver lining
  - Shift blame
  - Distancing / acceptance
  - Avoidance

提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

How to turn appraisal theory into a *model*?

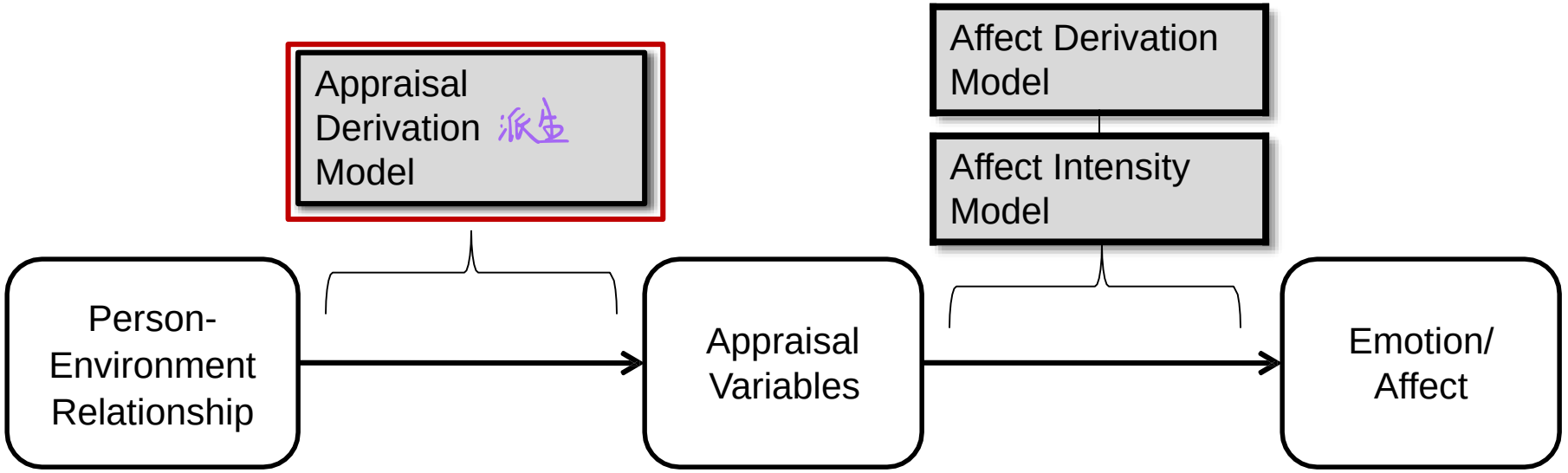
评估变量

Appraisal Variables

- Desirability
- Expectedness
- Controllability
- Causal Attribution

Emotion

Action Tendencies	"Affect"	Physiological Response



Where do *values* of appraisal variables come from?

Models of Appraisal Derivation

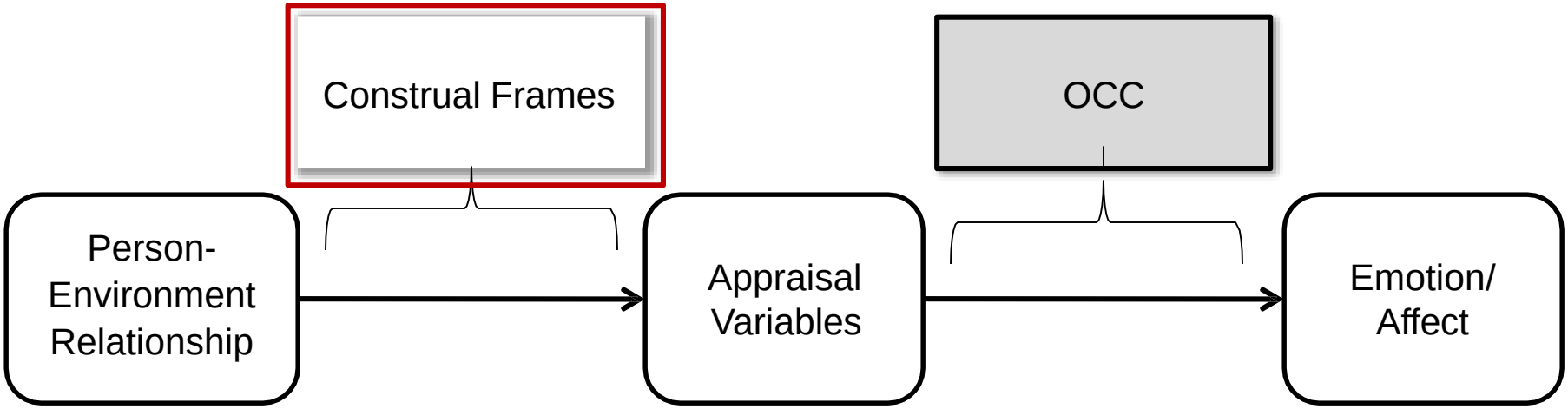
Clarke Elliott

- **Clarke Elliott: Affective Reasoner (AR)**
  - Build on Ortony, Clore and Collins (OCC) appraisal theory
  - Provide computational approach for deriving appraisal values from an event description
  - Applied to story understanding: agents could explain not only *what* happened but how agents *felt* and *why*

Affective Reasoner (AR)

- **Realize OCC Theory in a computer model**
  - Agent's reason about how events cause emotions
  - Agent's reason that same event produces different emotions in different agents
  - Agent's reason about emotions of others
  - Used to drive plausible emotional behavior of story characters

可靠的



How does it work?

- **Stories represented in a frame language** (Charniak's XRL)
  - E.g. Tom and Dick are watching a football game between Northwestern and Illinois
  - Northwestern scores a touchdown in the last second to go ahead 28 to 27
  - Agents have *construal frames* to infer consequence of events for agents' goals
  - Tom has a "Heroic finish" frame that is evoked if a touchdown is scored with 1sec left
  - Tom's heroic-finish-goal is achieved if Northwestern scores, otherwise the goal is blocked
  - Dick has similar frame for Illinois

Touchdown Event

FRAME: event-262

event-type:	touchdown
time-left:	0.01
team-1:	Northwestern
team-2:	Illinois
team-1-score:	28
team-2-score:	27
weather:	sunny
crowd-noise:	loud

Check Heroic Finish Tom

Tom

FRAME: heroic-finish-goal,
isa football-goal, isa ..., isa goal

Matching Conditions

event-type:	touchdown
time-left:	?time-left
team-1:	?t1
team-2:	?t2
team-1-score:	?t1s
team-2-score:	?t2s
predicate:	?time-left < 0.01
predicate:	?t1s ^= ?t2s
predicate:	?t1 = Northwestern or ?t2 = Northwestern
blocked:	(?t1 = Northwestern and ?t1s < ?t2s) or (?t2 = Northwestern and ?t2s < ?t1s)

Inferences

Check Heroic Finish Dick

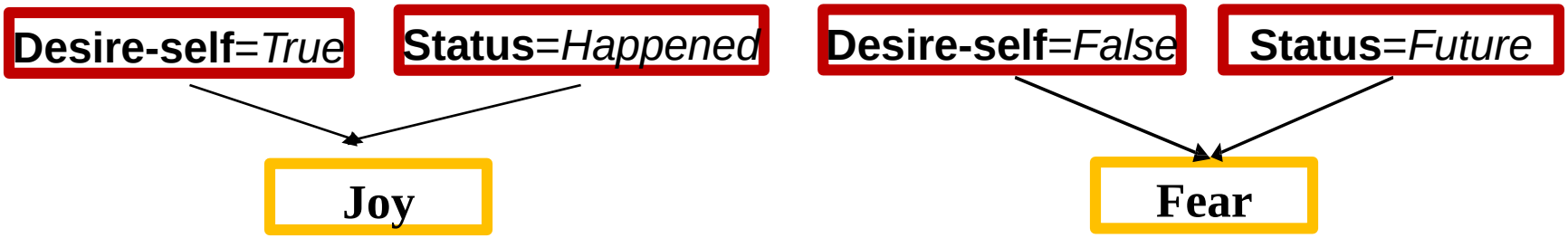
Dick

FRAME: heroic-finish-goal,
isa football-goal, isa ..., isa goal

event-type:	touchdown
time-left:	?time-left
team-1:	?t1
team-2:	?t2
team-1-score:	?t1s
team-2-score:	?t2s
predicate:	?time-left < 0.01
predicate:	?t1s ^= ?t2s
predicate:	?t1 = Illinois or ?t2 = Illinois
blocked:	(?t1 = Illinois and ?t1s < ?t2s) or (?t2 = Illinois and ?t2s < ?t1s)

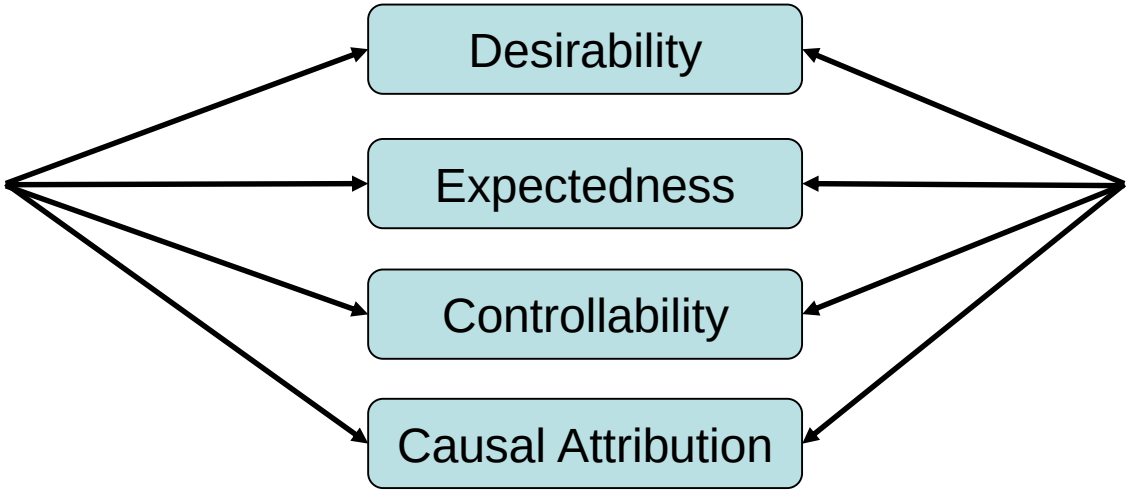
How does it work?

- **AR uses these frames to appraise each agent's goal**
- **Derives appraisal variables based on OCC model**
  - **Desire-self**: Is my own goal satisfied or blocked?
  - **Desire-other**: Is another agent's goal satisfied/bock?
  - **Pleased**: Am I pleased or displeased with what is happening to the other?
  - **Status**: Is the satisfaction/blocking in the future or has already happened?
  - **Responsible agent**: What agent (if any) is responsible for the event?
  - **Evaluation**: Does the responsible agent deserve blame?
  - **Appealingness**: Does the event contain an attractive or repulsive object
- **For each event and for each agent, calculate the appraisal variables**



AR Critique 批评

- **Require a large number of construal frames authored by human**
  - Domain-specific knowledge about the story 领域相关知识
  - Agent beliefs and goals
  - Domain-specific knowledge about how agents appraise the story events with respect to their beliefs and goals



What is more *general* way to appraise situations?

Another Model: Ëmile

Jonathan Gratch

- **Argument: many appraisals map onto what AI does anyway**
  - Calculate if goals being met, detect expectation violations, credit assignment, etc
- **Ëmile extends AR by using AI planning techniques to automatically understand situations**
  - Since appraisal is a process that understands how events relate to agents' goals, automate appraisal process to recognize threats and opportunities in potential plans for achieving a goal

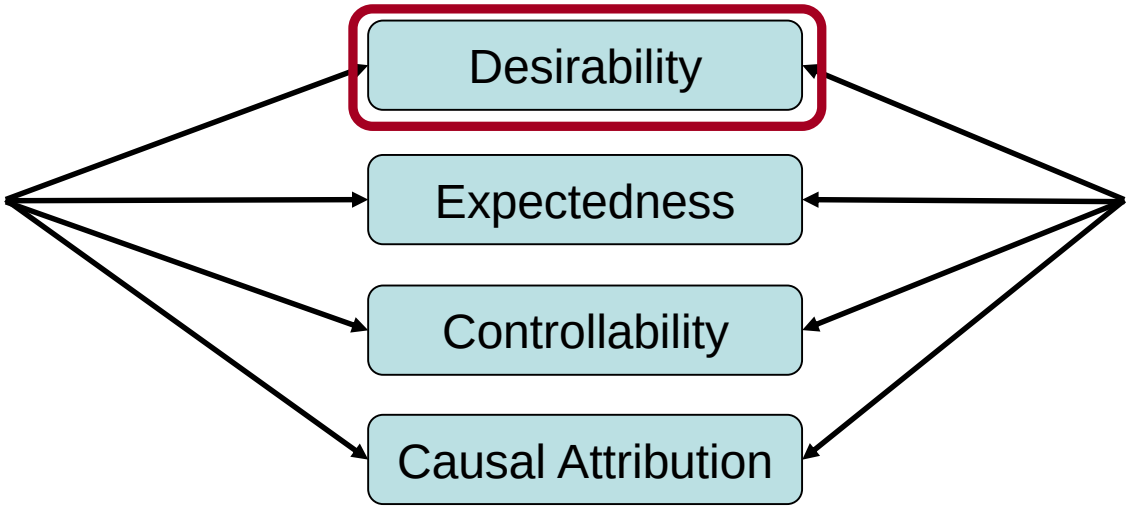
J. Gratch and S. Marsella. A Domain-independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*, 5(4):269-306, 2004.

Planning and Appraisal in Èmile

- **Create generic appraisal rules that characterizes current state of agent's goals. For each goal:**
  - **Desirability** = Utility of the goal for agent
  - **Likelihood** = Current probability that goal can be achieved
  - **Controllability** = Are there actions that can make goal true?
  - **Causal-attribution** = For any known and intended actions that impact the goal (facilitating or threatening), what actor is "in charge of / responsible for" that action
- **Uses *partial-order planner* to generate plans**

偏序规划

Emotions "emerge" as agent thinks (i.e. plans), acts in the world, and reacts to changes by other agents

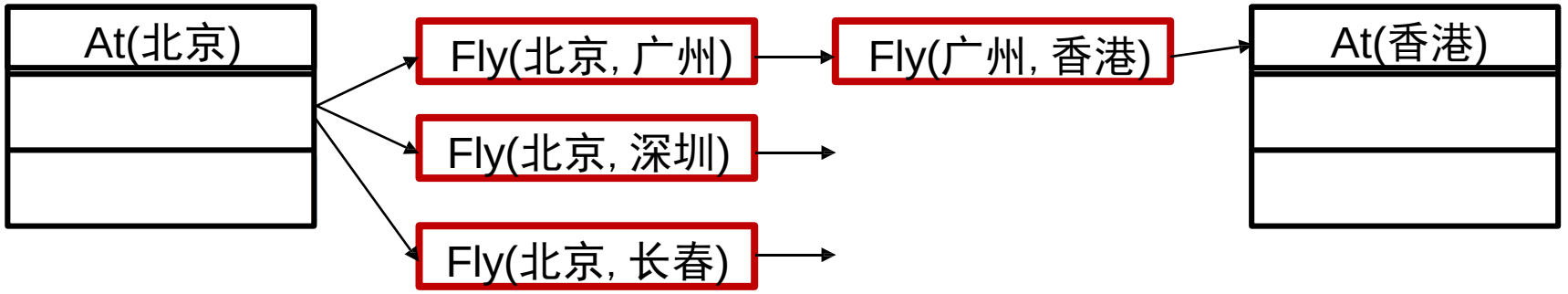


How can planning help decide if an event is *Desirable* vs. *Undesirable*?

Analyze Causal Structure of Plan

Fly(x,y)

Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)
Del: At(x) & OnTime(x)

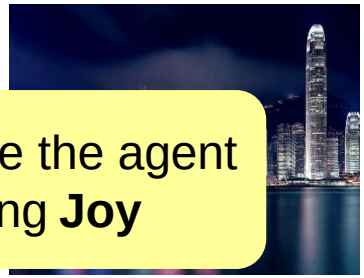


Analyze Causal Structure of Plan

Beijing

Fly(x,y)

Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)
Del: At(x) & OnTime(x)



And leave the agent feeling **Joy**

And leave the agent feeling **Distress**

At(北京)

Fly(北京, 深圳)

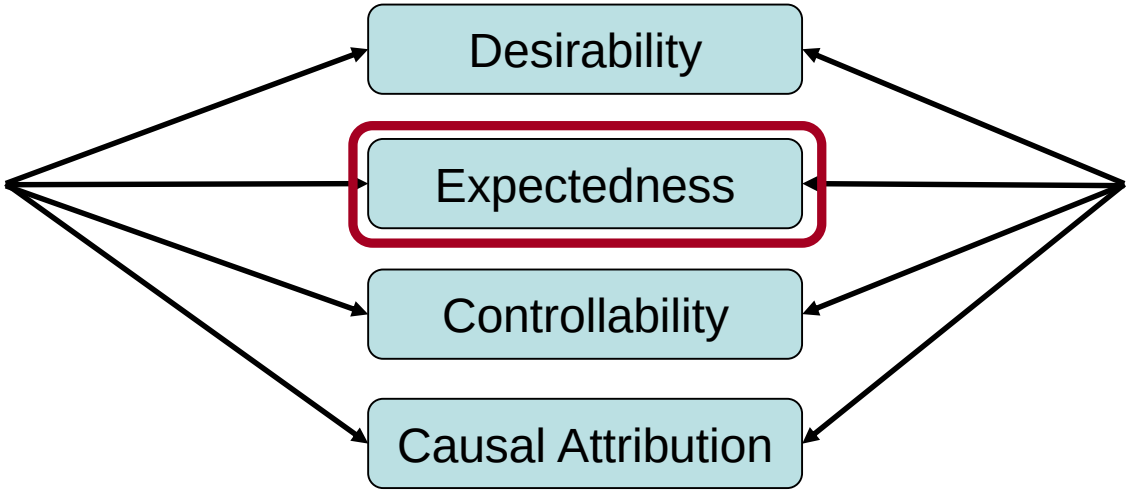
Fly(北京, 长春)

Fly(广州, 香港)

At(香港)

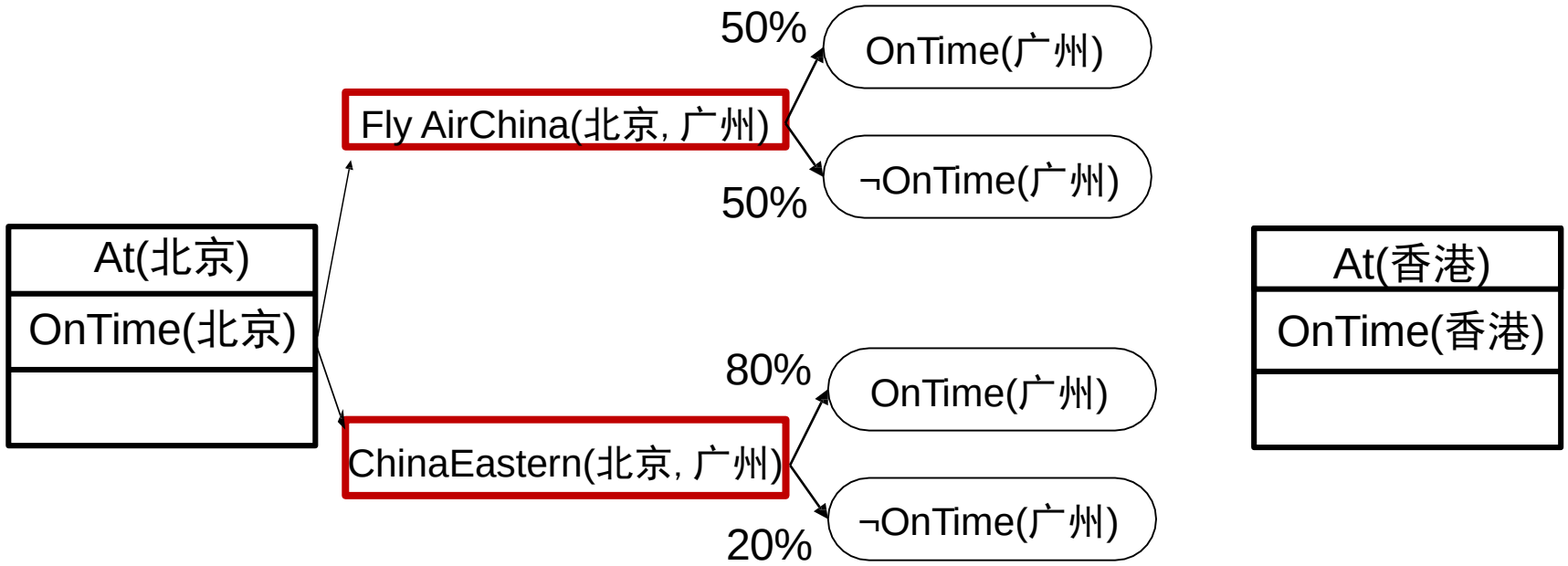
An action is **Undesirable** if it moves an agent farther from its goals

An action is **Desirable** if it moves an agent closer to its goals



How can planning help decide if an event is *Expected* vs. *Unexpected*?

Probabilistic Planning



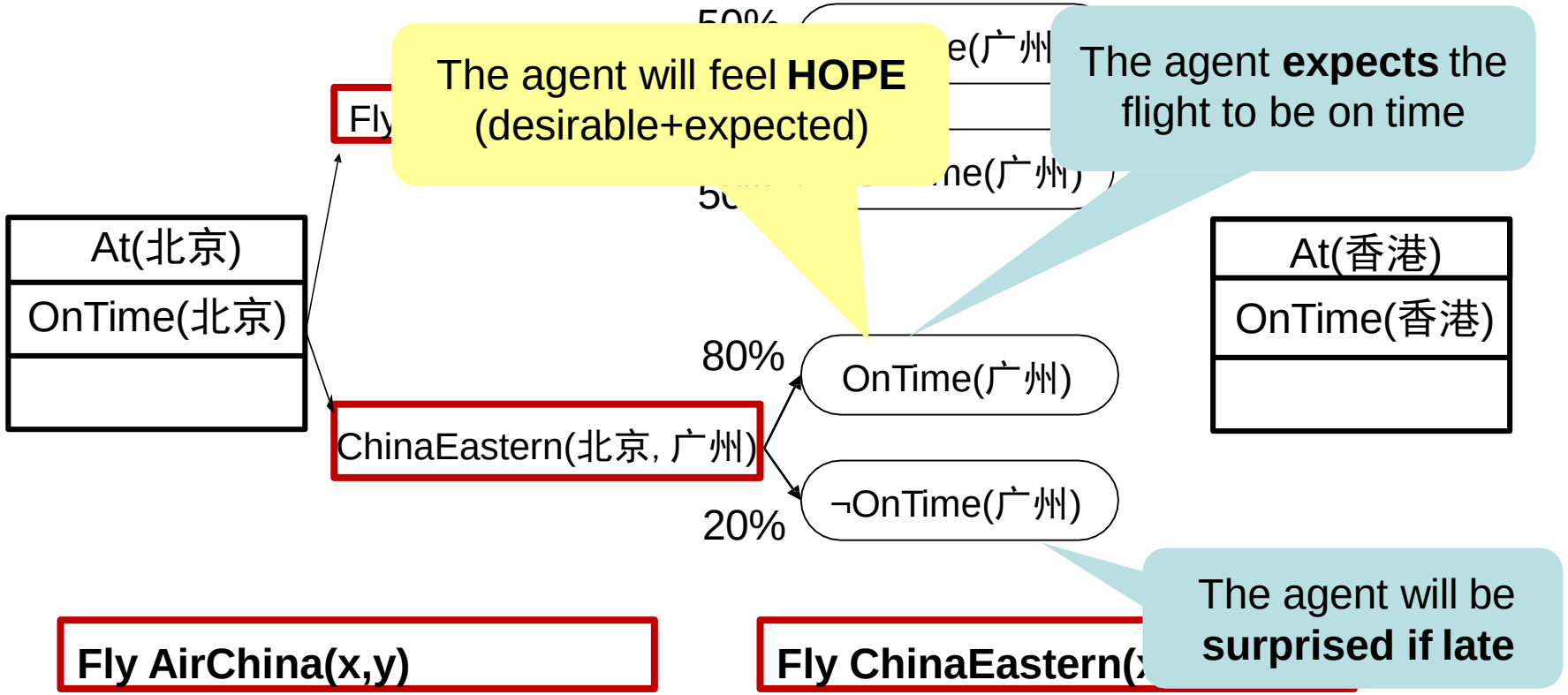
Fly AirChina(x,y)

Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)-50%
Del: At(x) & OnTime(x)

Fly ChinaEastern(x,y)

Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)-80%
Del: At(x) & OnTime(x)

Probabilistic Planning



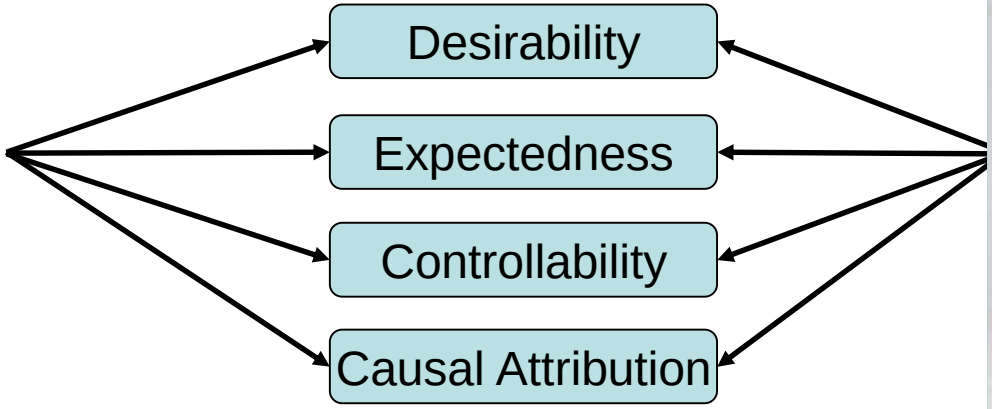
Fly AirChina(x,y)

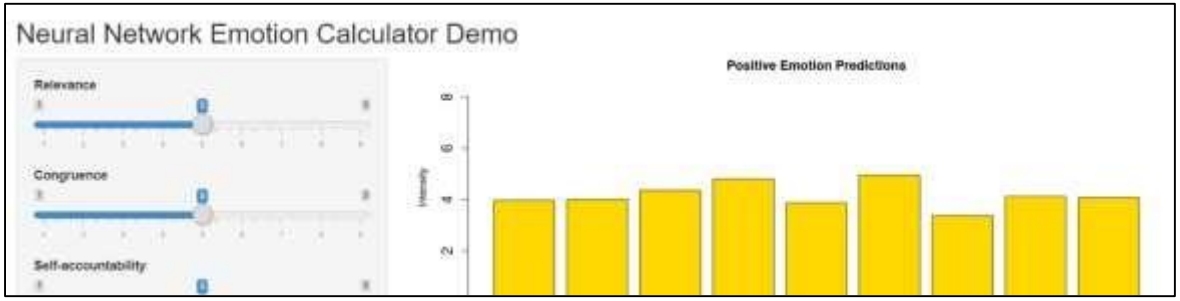
Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)-50%
Del: At(x) & OnTime(x)

Fly ChinaEastern(x

Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)-80%
Del: At(x) & OnTime(x)

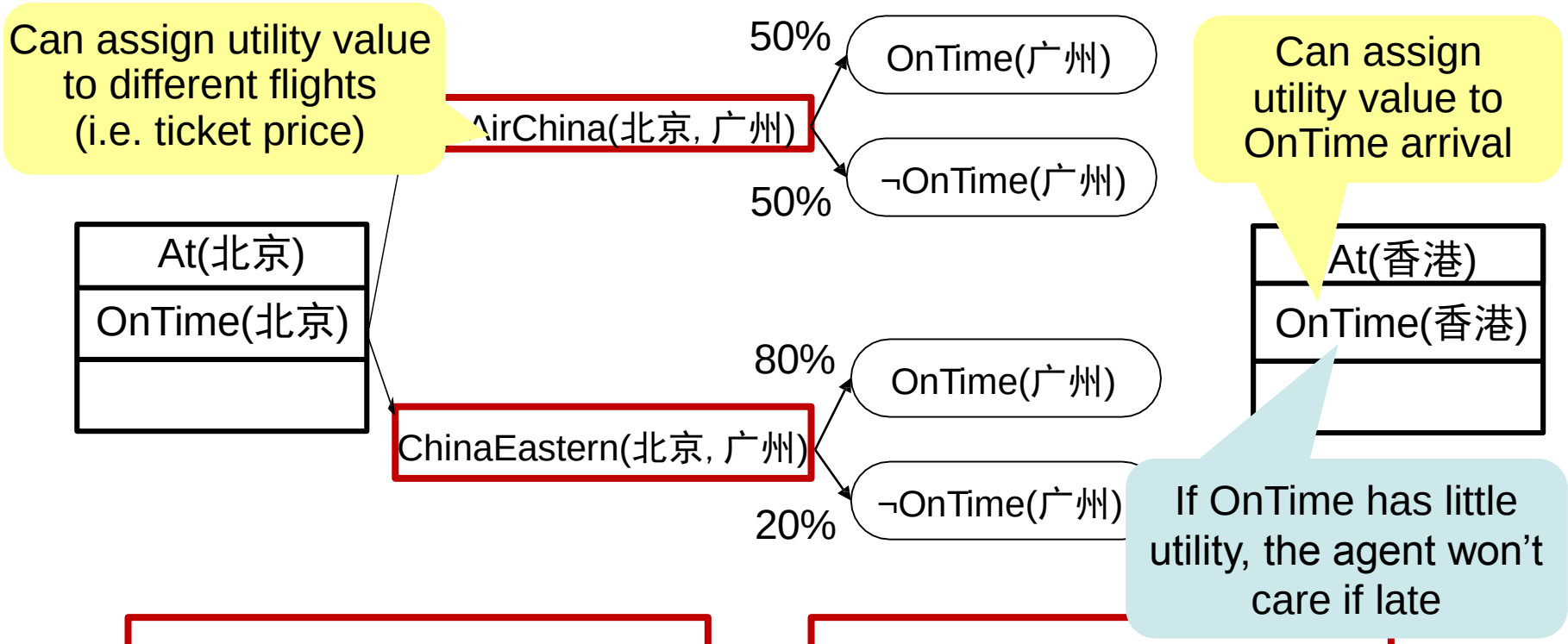
Derive Intensity Values





How can planning help derive *intensity* of emotion?

Decision-Theoretic Planning

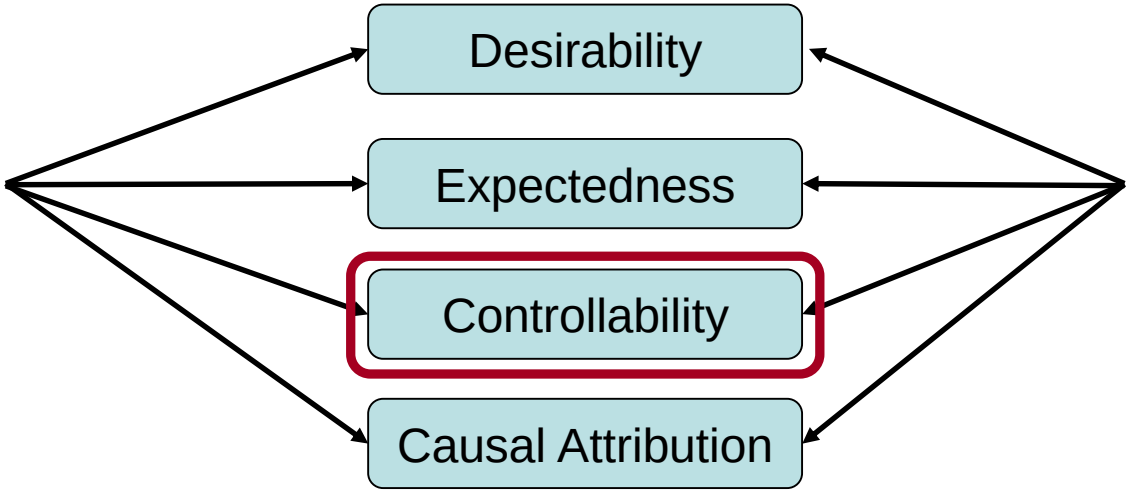


Fly AirChina(x,y)

Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)-50%
Del: At(x) & OnTime(x)

Fly ChinaEastern(x,y)

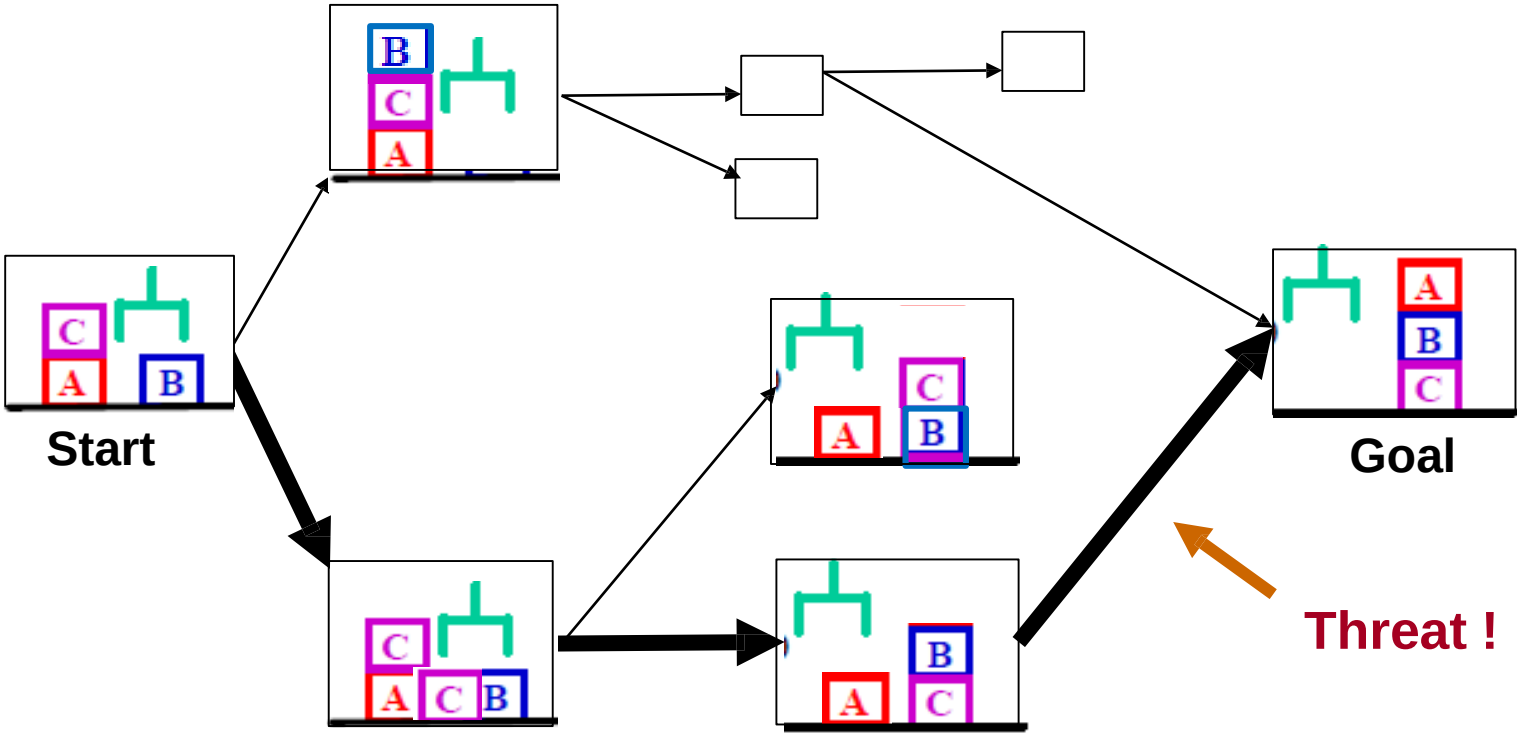
Pre: OnTime(x) & At(x)
Add: At(y) & OnTime(y)-80%
Del: At(x) & OnTime(x)



How much *control* should an agent feel over a situation?

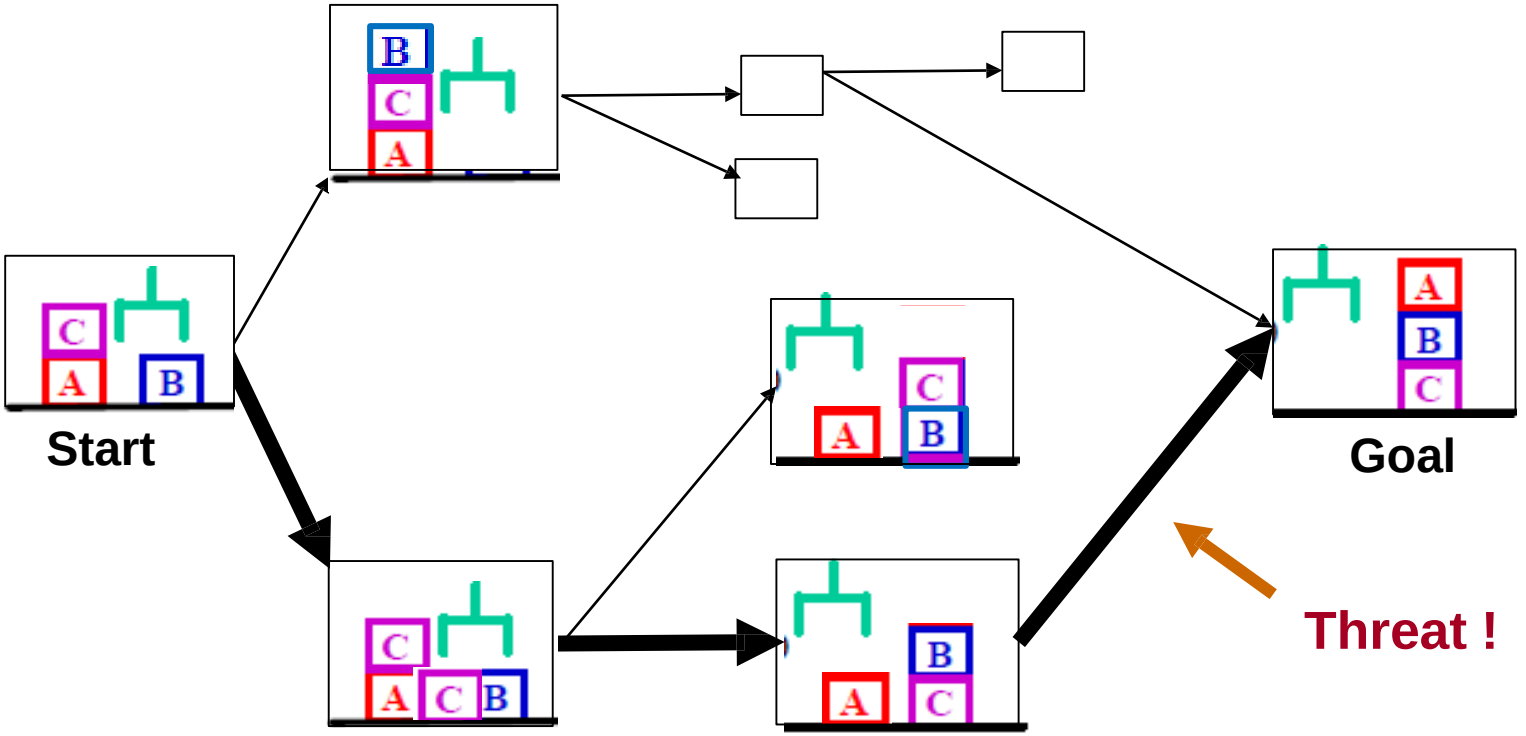
Availability of Alternative Plans

- If we intend the bold path to pursue my *goal*
- But a ***threat*** arises that blocks this path
- If other alternatives exist, a threat is seen as **controllable**



Counterfactual Reasoning

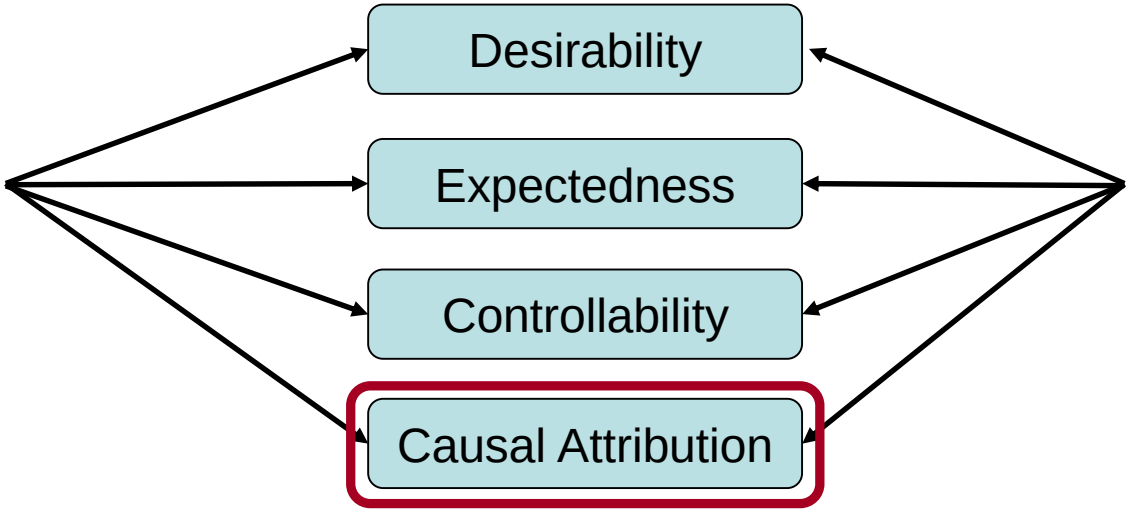
- Alternative paths represent ***alternative possible worlds***
- This allows us to reason about what might have happened if we did something differently (**counterfactual emotions**)



Counterfactual Reasoning

A travel example:

- Person X misses a plane for which she had a ticket and later learns that the plane has fatally *crashed*
- Although everyone in the airport who was not on the plane has reason to feel a mixture of *sadness* and *relief*, Person X is **extremely *relieved***



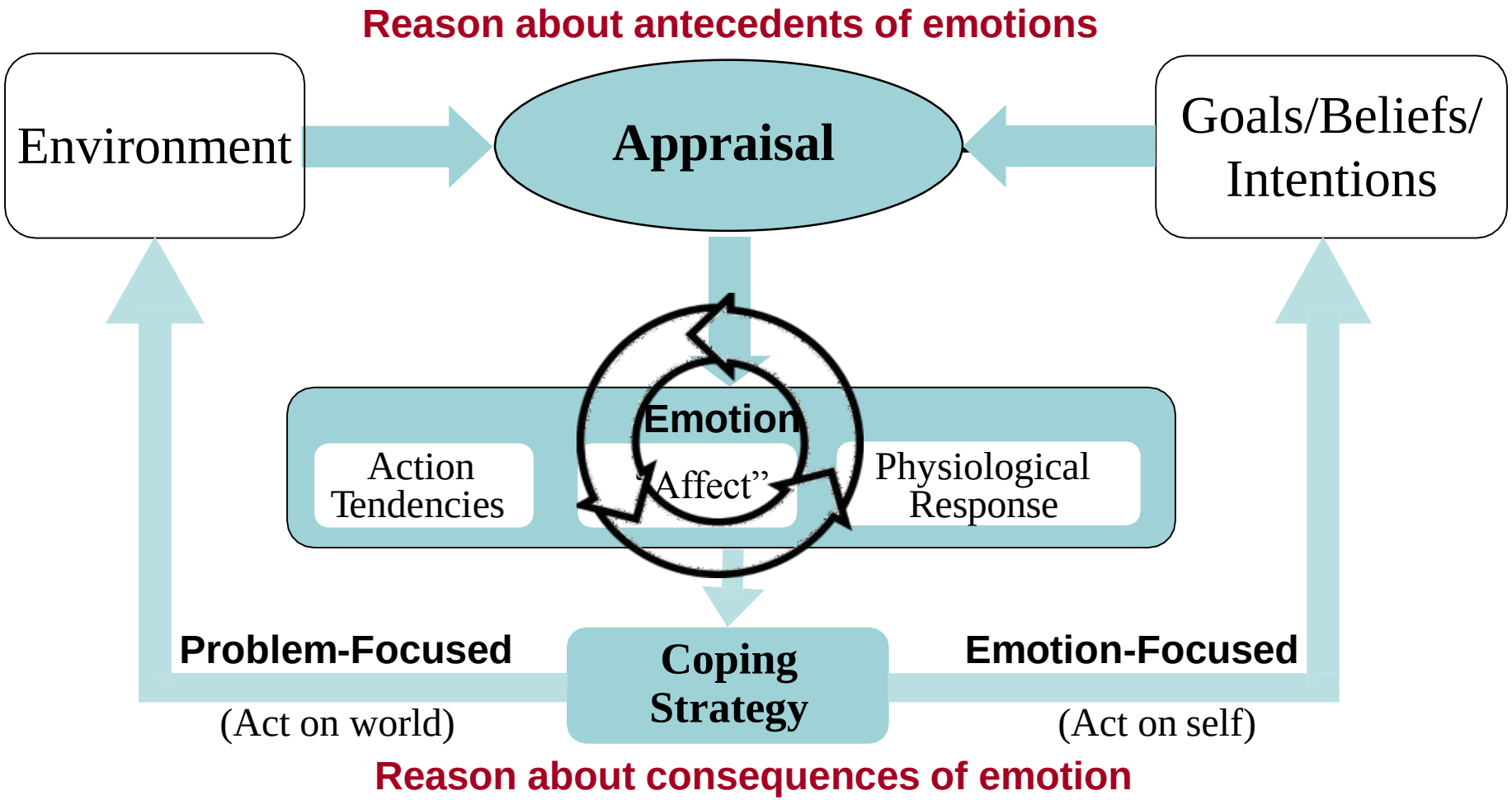
Which agent should be *responsible* for the outcome and be assigned *credit* or *blame*?

Yet Another Model: EMA

Stacy Marsella

- Èmile provides a general mechanism for **appraisal**
- *Carmen's Bright Ideas* (Marsella) focused on **coping**
- **EMA** combines these two perspectives

EMA Model of Appraisal and Coping

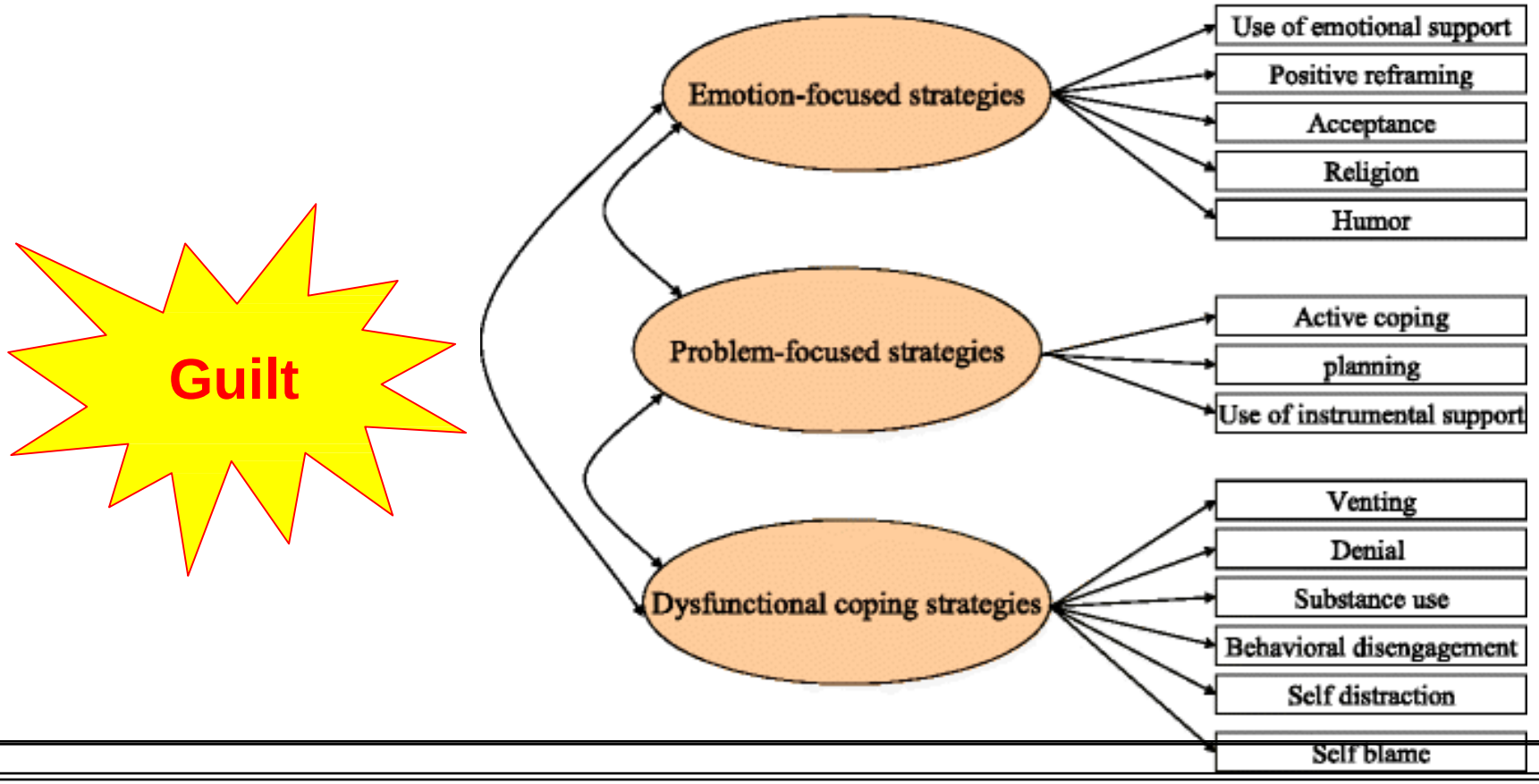


S. Marsella and J. Gratch. EMA: A Process Model of Appraisal Dynamics. *Journal of Cognitive Systems Research*, 10(1):70-90, 2009.

Coping Theory (Folkman & Lazarus 80)

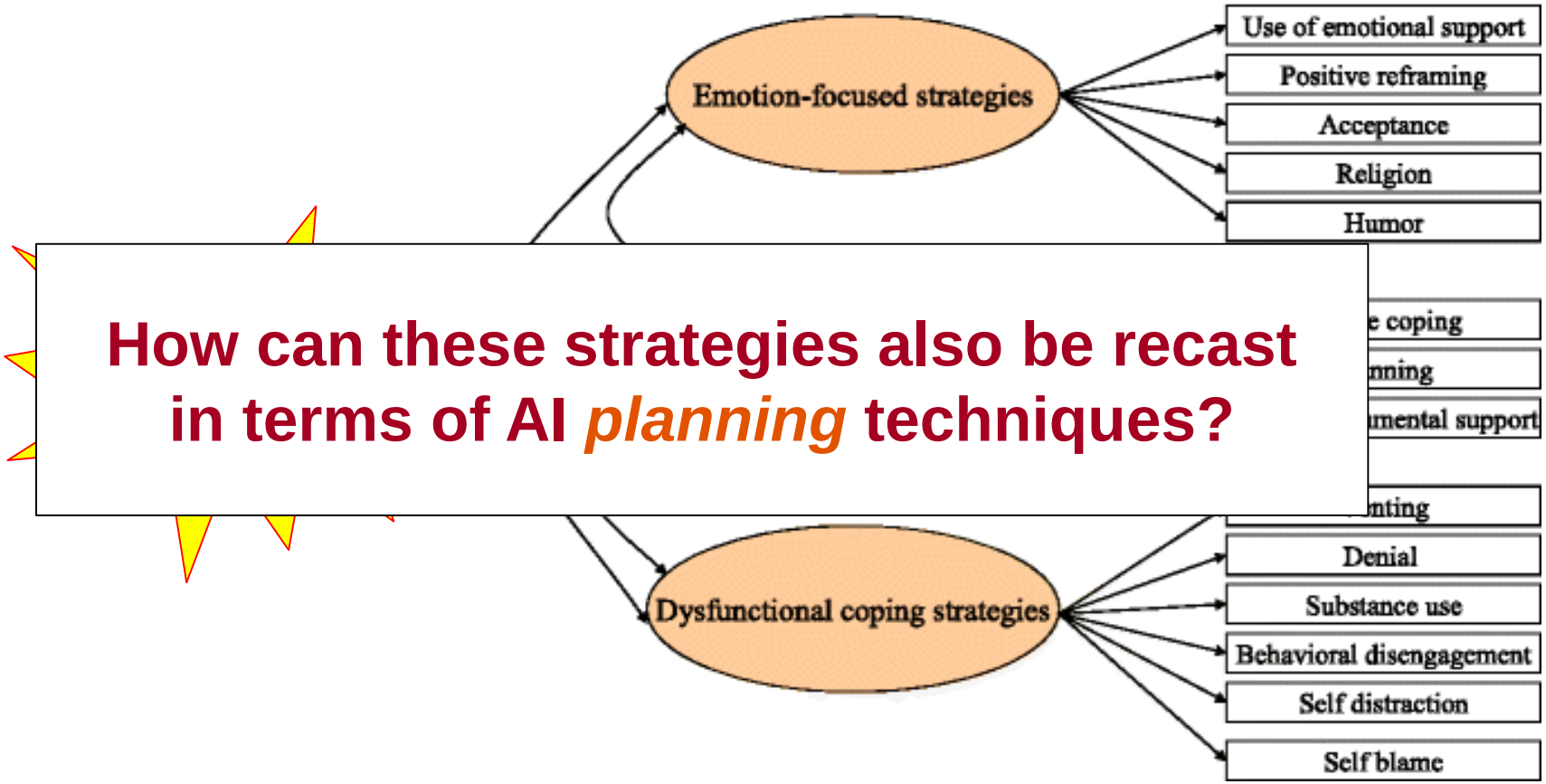
自我平衡的

- Emphasizes emotion drives homeostatic system
- People are motivated to reduce (negative) emotions



Coping Theory (Folkman & Lazarus 80)

- Emphasizes emotion drives homeostatic system
- People are motivated to reduce (negative) emotions



Modeling Appraisal & Coping in EMA

- **Appraisal as plan evaluation**
  - Define variables in terms of features of current plan
- **Coping as plan refinements**
  - Problem-focused → take action, make plans
  - Emotion-focused
    - Denial → Change belief
    - Wishful thinking → Change likelihood
    - Find silver lining → Change utilities
    - Shift blame → Change causal attribution
    - Distancing → Drop goal/intention
    - Avoidance → Add new goal

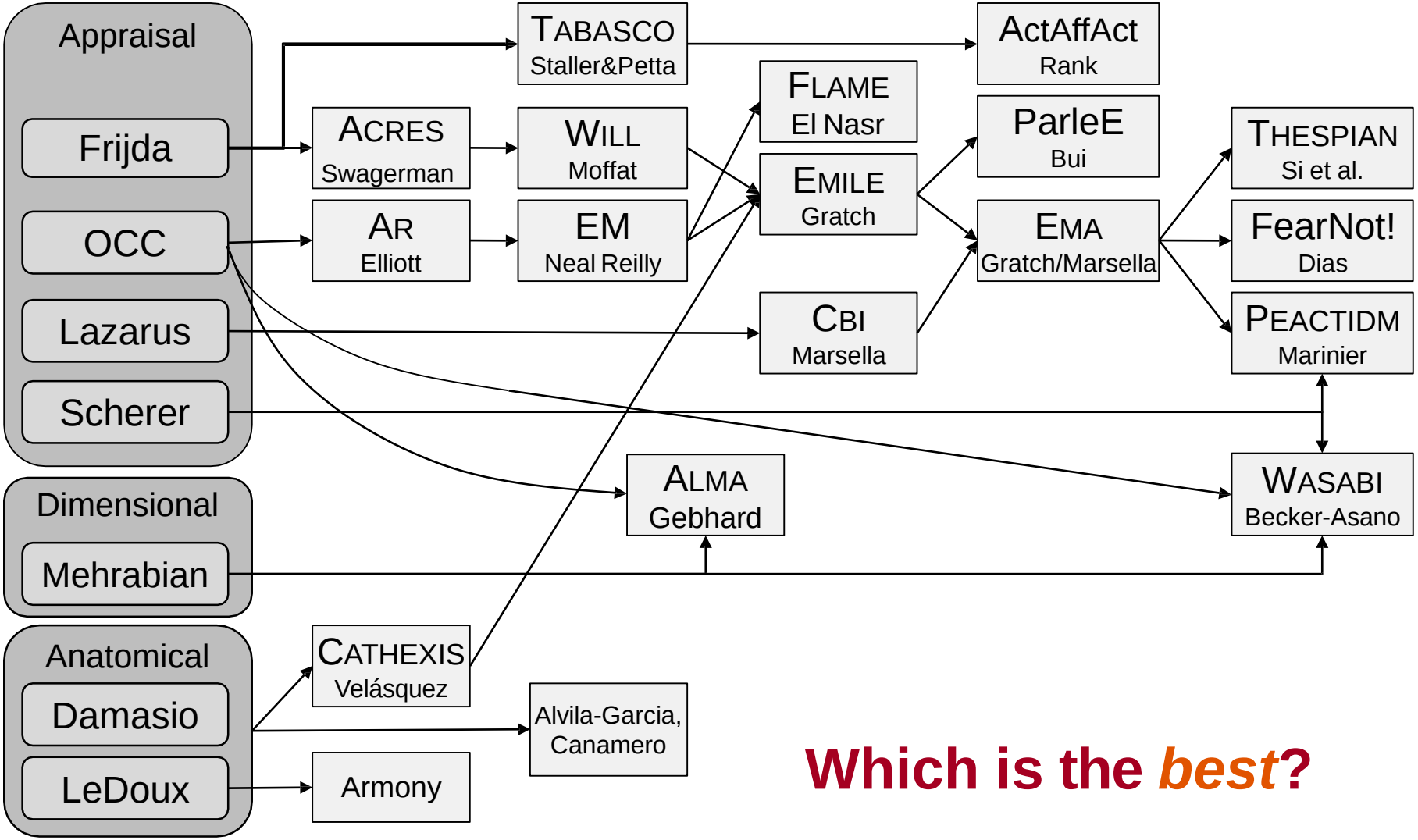
EMA: Structural and Process Model

- ***Structural model* describes the components underlying behavior and connections/structure between components**
  - Describe a static structure
  - Explain and predict the relationship between variables
- ***Process model* describes the dynamic processing underlying behavior (also called functional modeling)**
  - Emphasize operations that change/transform local inputs to outputs
  - Emphasize dynamics over time
- **Domain-independent *general* mechanism for modeling emotion**

提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

Many other models exist …

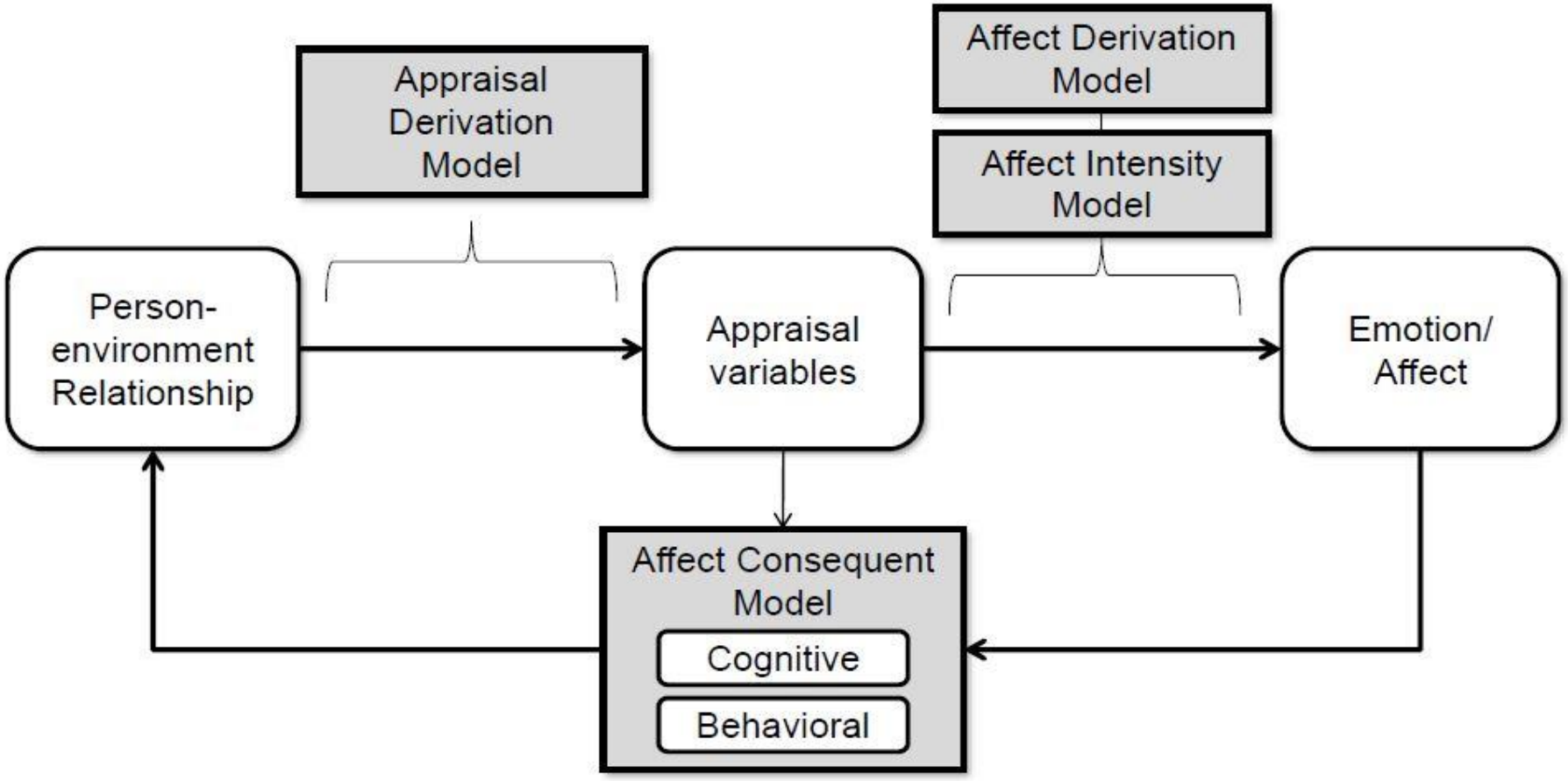


Which is the *best*?

Theoretical basis

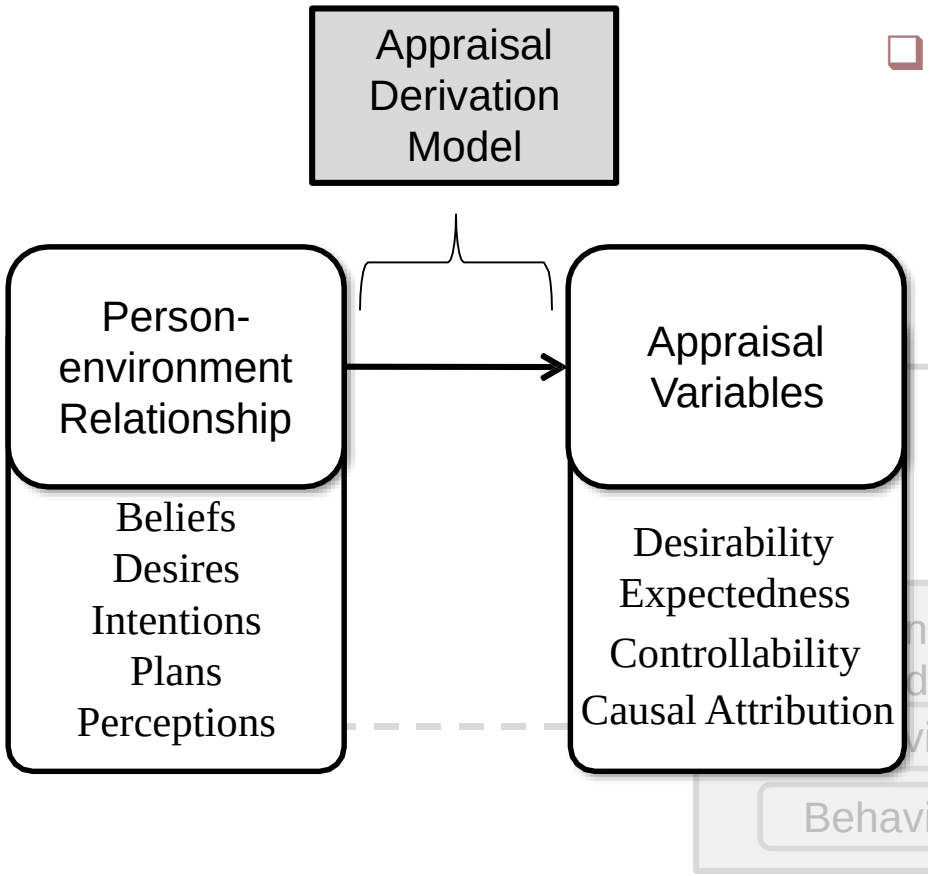
	EMA	ALMA	FLAME
Person-environment Relationship	Domain-independent Decision-theoretic Plans + BDI	Outside the scope of model	Domain-independent Markov-decision process
Appraisal-derivation	Inference over decision-theoretic plans	User-defined	fuzzy rules over Markov-decision graph
Appraisal-variables	Lazarus-inspired: Desirability, likelihood, Expectedness, Causal attribution, Controllability, Changeability	OCC-inspired Good/bad, likely/unlikely event Good/bad act of self/other Nice vs. nasty thing	OCC-inspired Desirability Expectation (dis)approval
Affect-derivation	Lazarus-based structural model that generate discrete emotion and mood state	OCC-based structural model that give “impulses” into core affect	OCC-based structural model producing discrete emotion labels
Affect-intensity	Expected utility model, Threshold model, Additive mood derivation	User defined	Additive model
Affect	Set of appraisal frames, mood (discrete-emotion vector) with decay	PAD space representing both current mood and emotion	Emotion and positive vs. negative mood state
Behavioral-consequences	Most-intense emotion alters behavior display and action selection. Actions are close-loop via domain-independent rules	Open looped. Mood and emotion alter behavior display and action selection	Domain-specific fuzzy expression and action Rules
Cognitive Consequences	Closed-loop via domain-independent emotion-focused coping than changes BDI	Open-looped. Emotion amplifies/dampens intensity of elicited emotions.	Closed-loop changes to domain model via reinforcement learning..

Components of Appraisal Models



S. Marsella, J. Gratch and P. Petta. Computational Models of Emotion. In: Scherer, K. R., Bänziger, T. & Roesch, E. (Eds.) *A Blueprint for Affective Computing: A Sourcebook and Manual*. Oxford University Press, 2010.

Appraisal Derivation Model



- **Methods distinguished by choice of reasoning framework**
  - EMA – Decision-theoretic plans
  - Broekens – Reinforcement learning
  - Sigma – Factor graphs
  - Dastani – Logic
  - Hudlicka's MAMID – Belief nets
  - El Nasr's FLAME – MDP
  - Neal Reilly's EM – Reactive planning
  - Marinier – Newell's PEACTIDM

Emotion Intensity Results

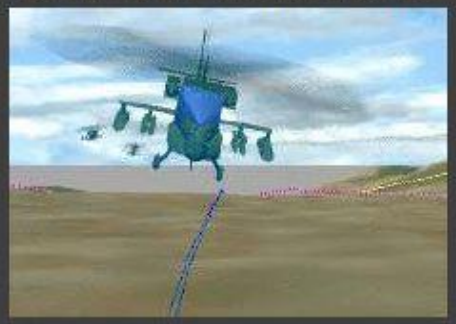
	Hope	Joy	Fear	Sadness
Realization Model	EM. PEACTIDM	ParleE, PEACTIDM	EM, PEACTIDM	ParleE, PEACTIDM
Expected Utility	EMA, Silverman, FearNot!		EMA Silverman	
Threshold Model		EMA, EM		EMA, EM
Additive Model	Cathexis, FLAME	Cathexis, FLAME	Cathexis, FLAME	Cathexis, FLAME
Hybrid Model	Price 85	Price 85 Silverman	Price 85	Price 85 Silverman

Strongly support EMA

S. Marsella, J. Gratch and P. Petta. Computational Models of Emotion. In: Scherer, K. R., Bänziger, T. & Roesch, E. (Eds.) *A Blueprint for Affective Computing: A Sourcebook and Manual*. Oxford University Press, 2010.

Uses: Make agents more "realistic"

Battlefield Simulations



1998

Storytelling

2000

Leadership Training

2002

2004

Interactive Entertainment

2009

Conflict Resolution

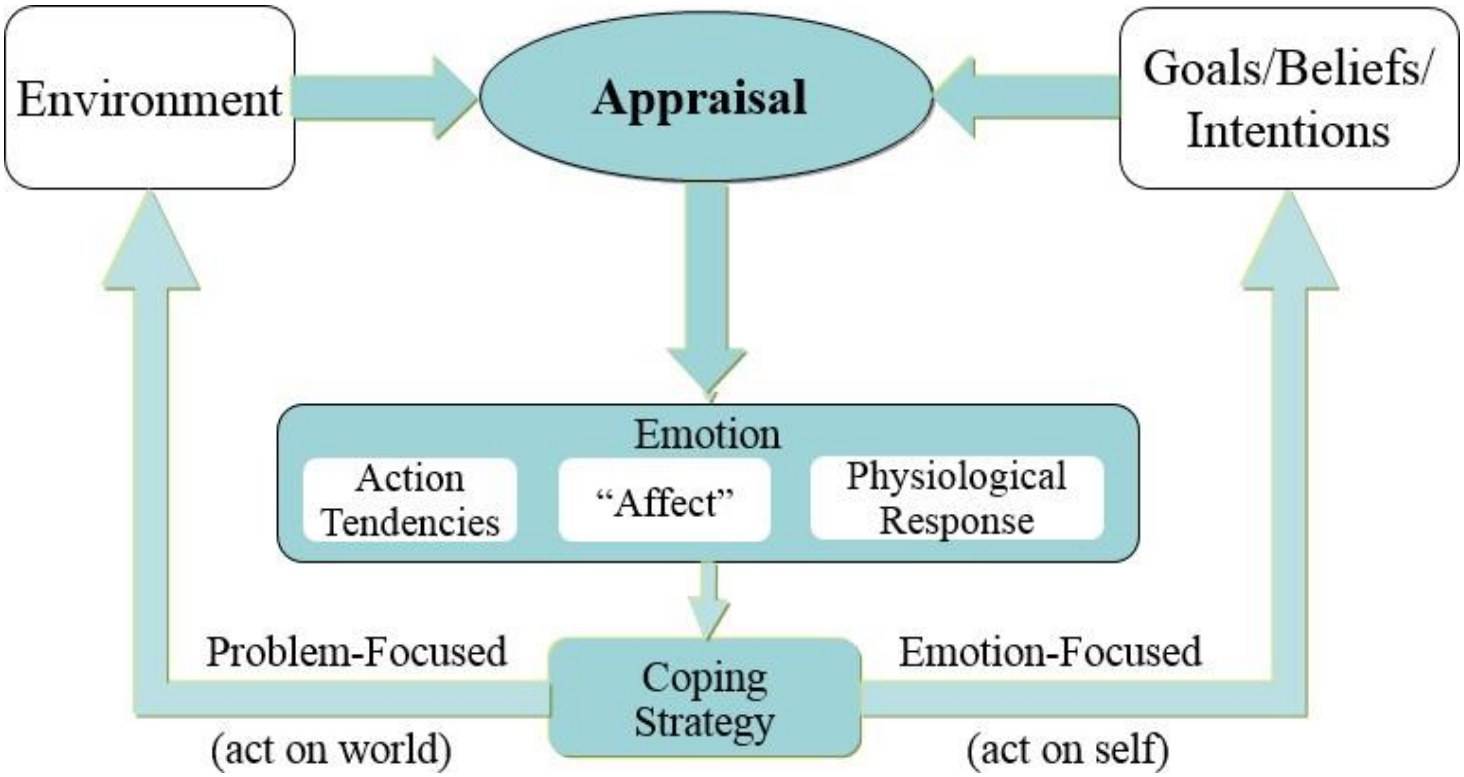
2016

Negotiation Training

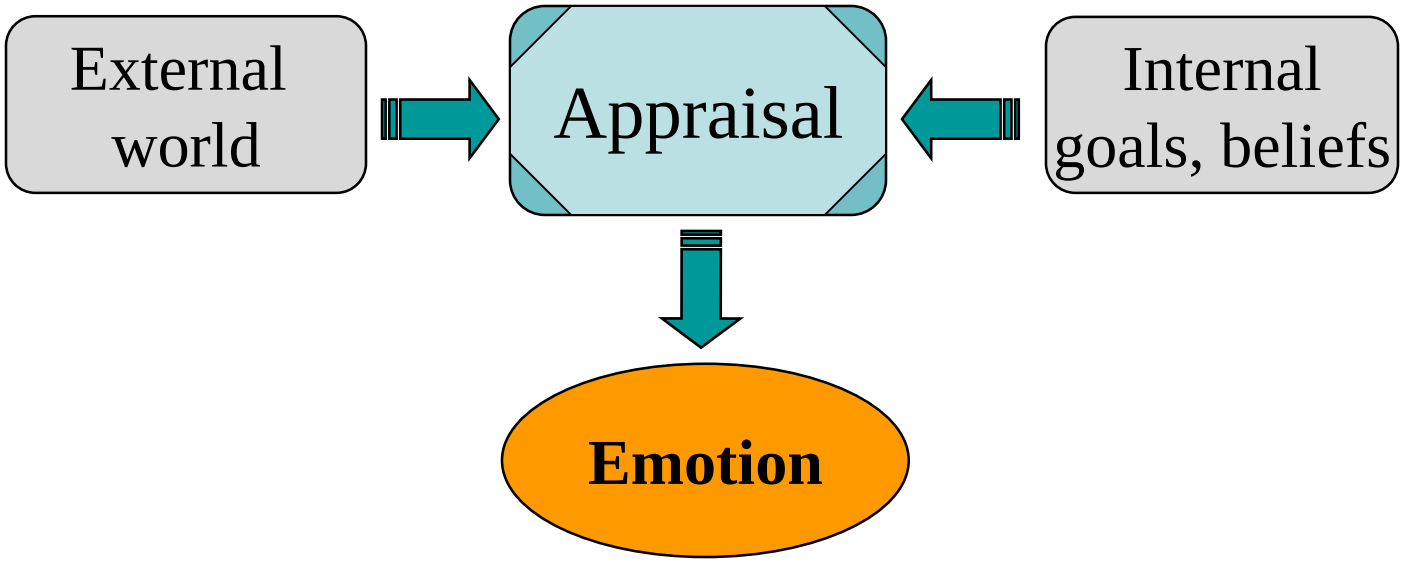
提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

情绪计算模型EMA（Gratch & Marsella）



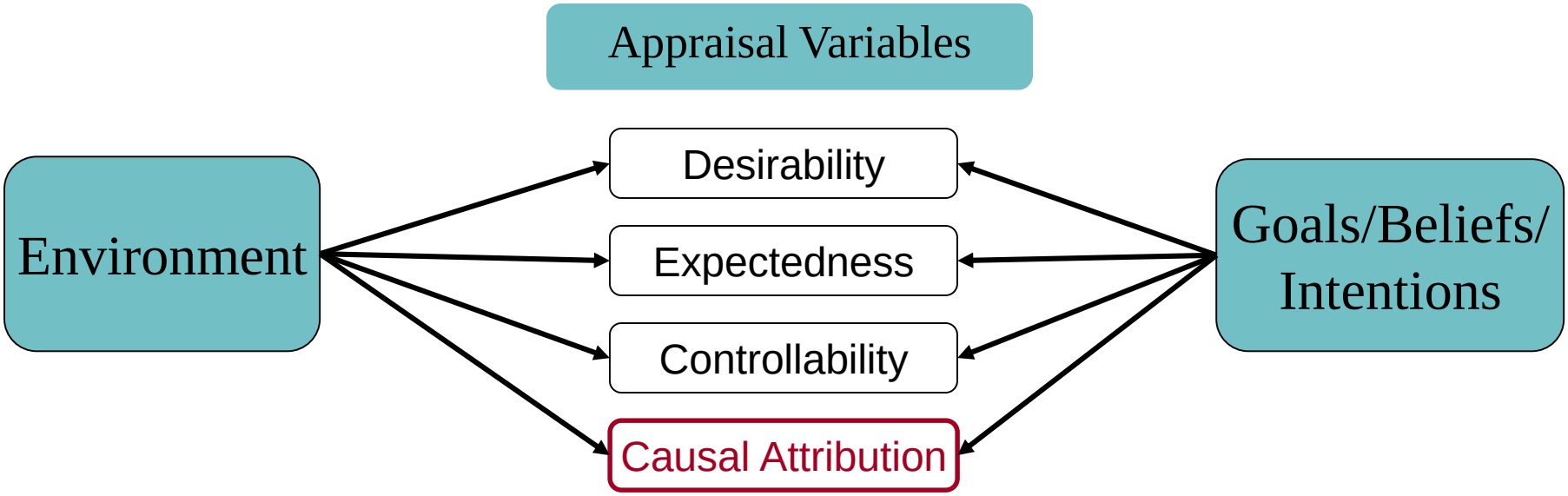
情绪评估过程



- **Appraisal = Situation assessment**

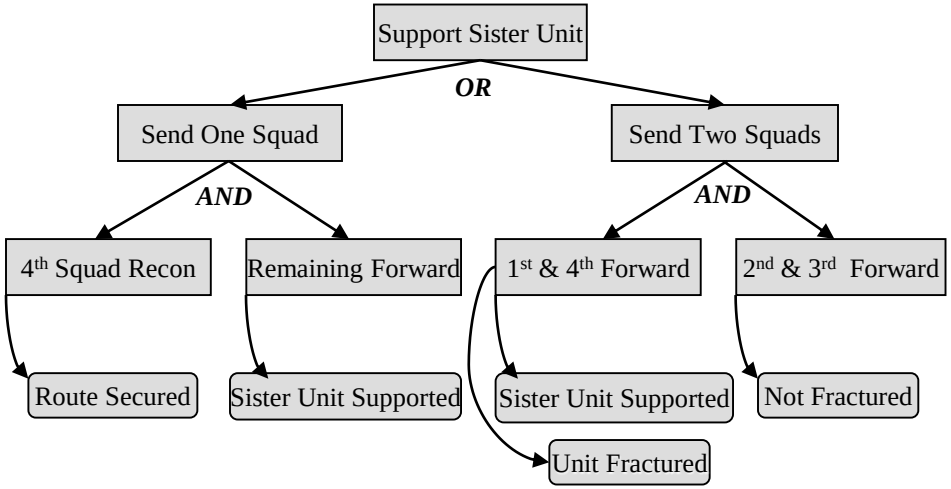
  Compare beliefs, desires and intentions with external circumstances

情绪评估变量



- Define appraisal variables as features of current plan
  - **Desirability**: Utility of a goal for agent
  - **Expectedness**: Current probability that a goal can be achieved
  - **Controllability**: Availability of actions and alternatives

应用系统示例



Trainee: Sergeant, send two squads forward. (SA: **order**)
Sergeant: Sir, that is a bad idea. We shouldn't split our forces. (SA: **inform**)
Instead, we should send one squad to recon forward. (SA: **counter-propose**)
Trainee: Send two squads forward! (SA: **order**)
Sergeant: Against my recommendation, sir. (SA: **accept**)
Lopez, send first and fourth squads to Eagle 1-6's location. (SA: **order**)
……

W. Swartout, J. Gratch, *et al*. Toward Virtual Humans. *AI Magazine*, 27(2):96-108, 2006.

研究意义

- **建立社会因果推理计算模型的意义**
  - **社会计算基础研究问题**
    有助于社会学习、社会规划、语用信息、情感模拟等
  - **社会智能的核心方面**
    有助于计算个体模拟人的社会智能，呈现类人社会行为
  - **行为分析与预测手段**
    网络化社会安全管理、控制，促进多个应用领域的发展
  - **计算系统的社会认知能力**
    多智能体系统和人机交互环境的深层解释和反馈

提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

Psychological Attribution Theory

- Well-founded in social psychology

  Favored theory for decades [Malle, 2011, 2013]

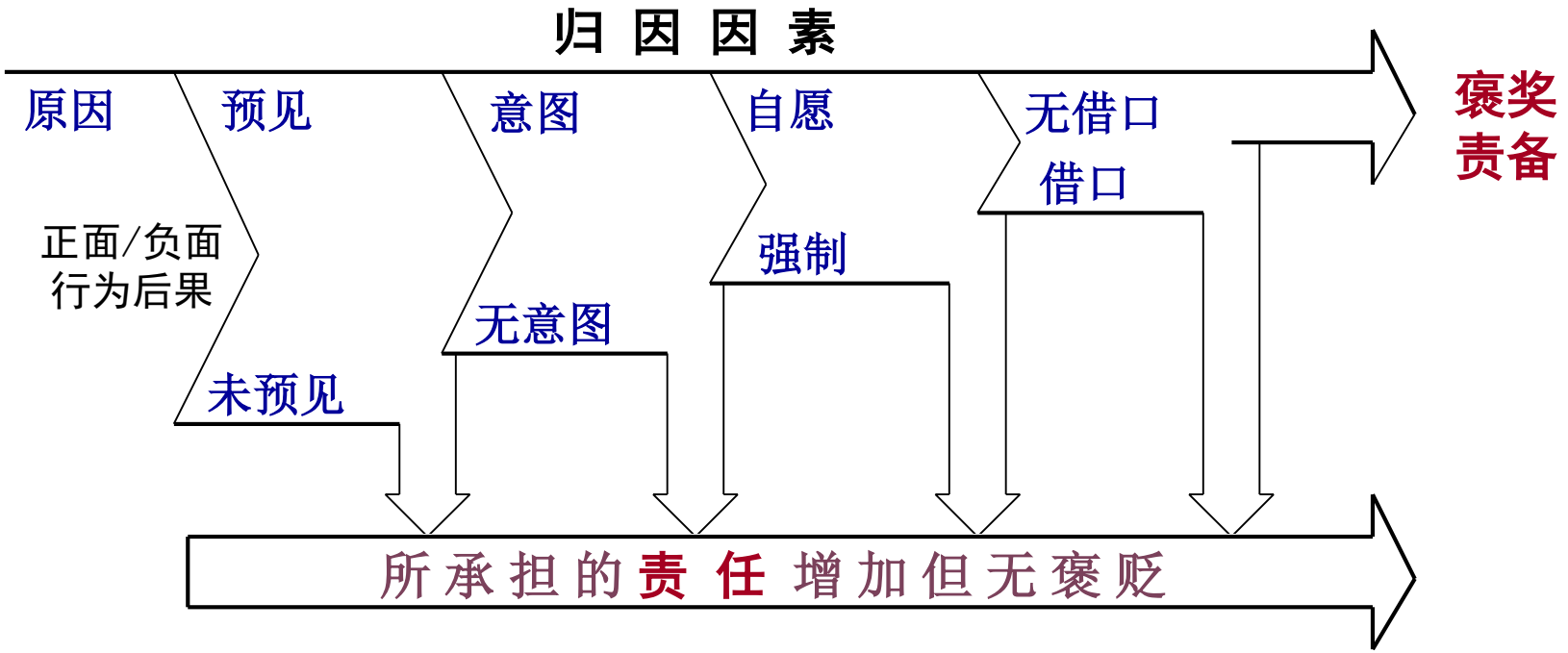
- Identifying how people form judgments

  Key attribution variables

- Theory of responsibility and blame

  Most influential models of **Weiner [1995, 2006]** and **Shaver [1985]**

Social Attribution Model



Adapted from Shaver [1985]

Foreseen

Intended Action

Intended Consequence

Coerced

Example

An officer orders his marksman to shoot a character.

The marksman rejects the order but the officer insists.

The marksman shoots at the character and she dies.

Negative Consequence

Example

An officer orders his marksman to shoot a character.

The marksman rejects the order but the officer insists.

The marksman shoots at the character and she dies.

Observation +
Knowledge of Action and Consequence

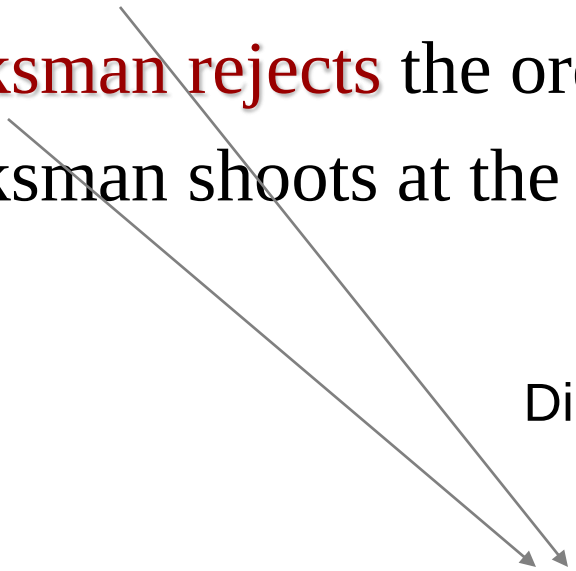
Marksman: Caused

Example

An officer orders his marksman to shoot a character.

The marksman rejects the order but the officer insists.

The marksman shoots at the character and she dies.



Dialogue Inference

Marksman: Caused → Coerced Shooting

Example

An officer orders his marksman to shoot a character.

The marksman rejects the order but the officer insists.

The marksman shoots at the character and she dies.

Example

An officer orders his marksman to shoot a character.

The marksman rejects the order but the officer insists.

The marksman shoots at the character and she dies.

Marksman: Caused → Coerced Shooting → Coerced Outcome

Officer: Intended → Voluntary? → Excuse?

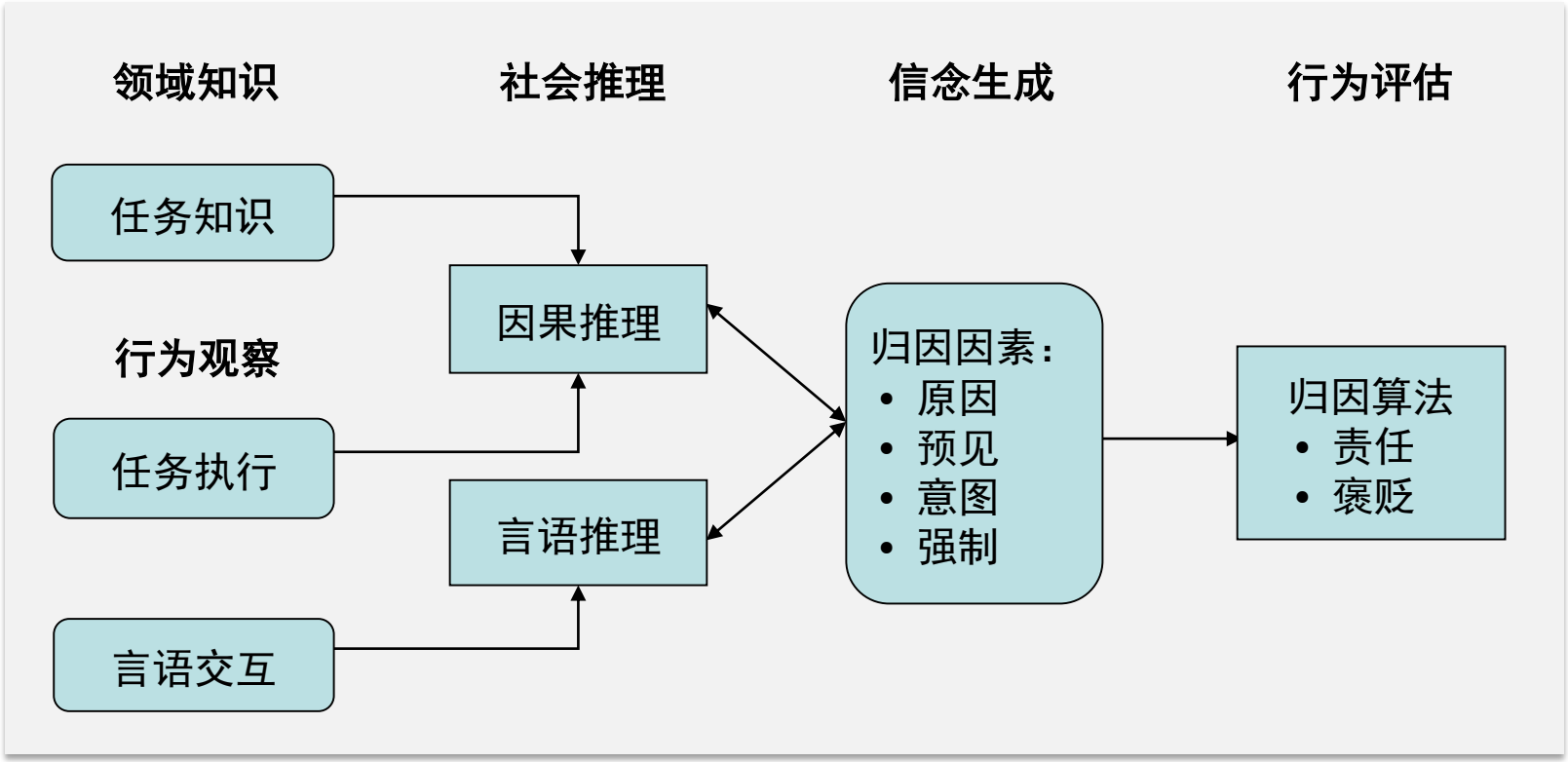
提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

Related Computational Work

- **Philosophical view/Prescriptive model**
  - Legal arguments and legal reasoning
    McCarty [1997]; Poole [2000]; Walker [2007]
  - Extension of causal models
    Halpern & Pearl [2001]; Chockler & Halpern [2004, 2016]
- **Psychological view/Descriptive model**
  - Folk theory for everyday situations
  - Suitable for modeling human-like agents

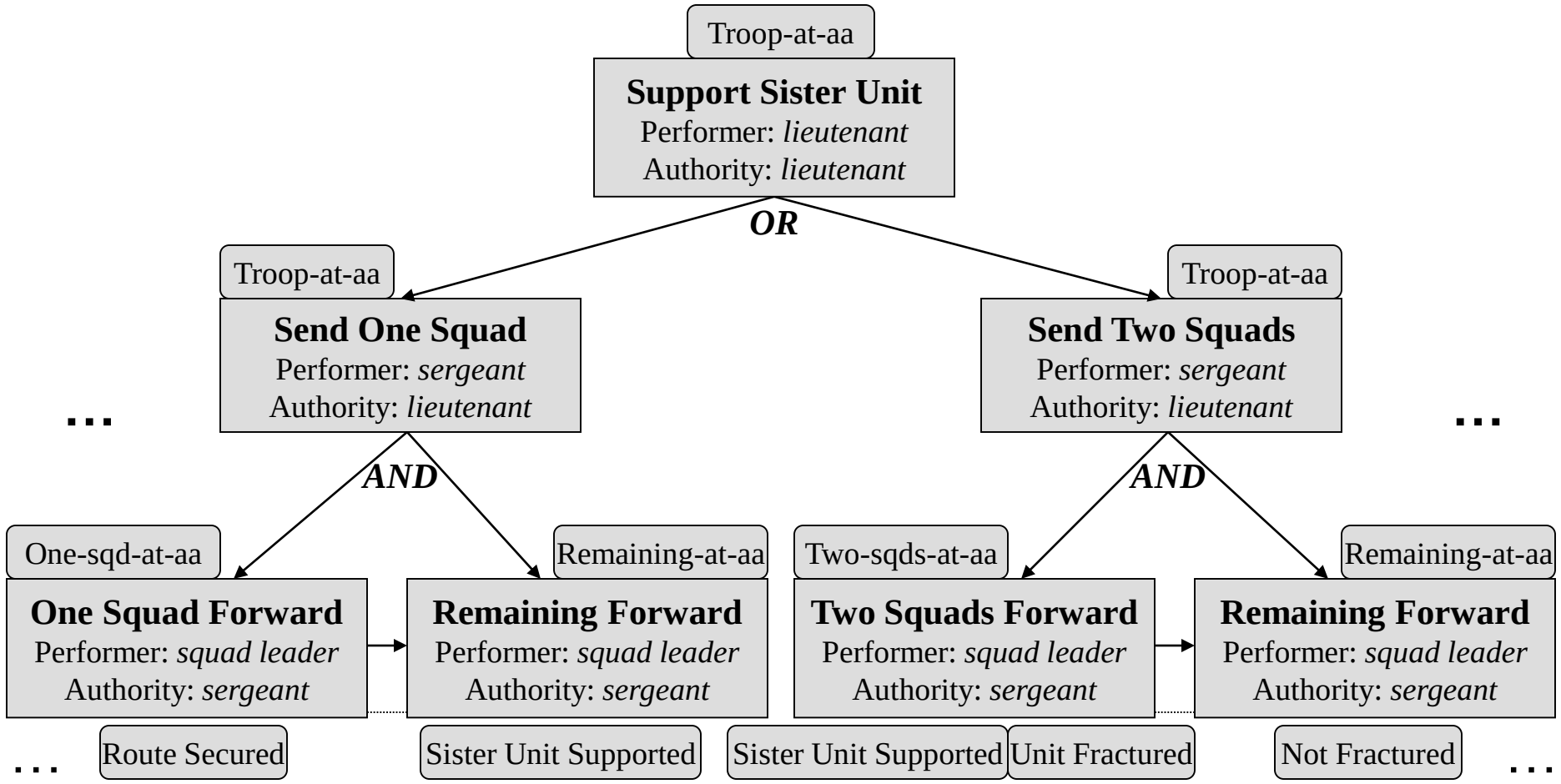
Computational Attribution Model



Representation: Causal Knowledge

- **Encoded via hierarchical plan representation**
- **Action**
  - Has preconditions and effects (inc. conditional)
  - Can be primitive or abstract
  - Has alternatives
- **Plan**
  - Has preconditions and outcomes
  - Associated with intended goal

Example: A Hierarchical Plan



Representation: Communication

- Communicative events are represented as speech act sequence

***inform*(*x*, *y*, *p*, *t*)**: *x* informs *y* that *p* at *t*

***request*(*x*, *y*, *p*, *t*)**: *x* requests *y* that *p* at *t*

***order*(*x*, *y*, *p*, *t*)**: *x* orders *y* that *p* at *t*

***accept*(*x*, *p*, *t*)**: *x* accepts *p* at *t*

***reject*(*x*, *p*, *t*)**: *x* rejects *p* at *t*

***counter-propose*(*x*, *p*, *q*, *t*)**: *x* counters *p* and proposes *q* at *t*

Representation: Attribution Variables

- **Foreseeability**
  - Foreknowledge about actions and consequences
  - Denoted as *know* and *bring about*
- **Intention**
  - Act versus outcome intention
  - Denoted as *intend*, *do* and *achieve*
- **Coercion**
  - Act versus outcome coercion
  - Denoted as *coerce*, *do* and *achieve*
  - Also consider *obligation* and (un)*willingness*

Dialogue Inference: Example

Lieutenant **orders** *p* ⟹ Lieutenant **intends** *p*

⟹ Sergeant has **obligation** *p*

Sergeant **counters** *p* **proposes** *q* ⟹ Lieutenant **knows alternatives**

⟹ Sergeant does **not intend** *p*

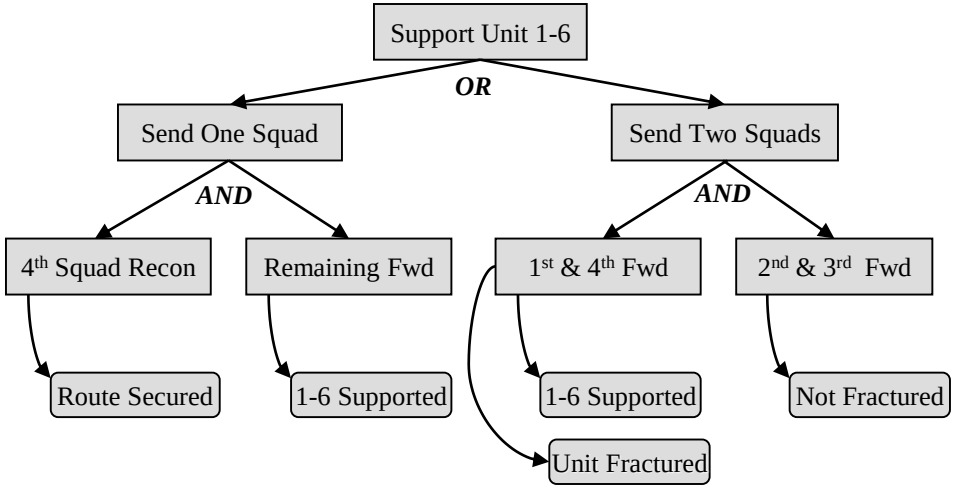
⟹ Sergeant **wants** *q*

Sergeant does **not intend** *p* *AND* has **obligation** *p* beforehand,

AND Sergeant **accepts** *p* ⟹ Lieutenant **coerces** sergeant *p*

$$\neg\textbf{intend}(h, p, t1) \wedge \textbf{obligation}(h, p, s, t2) \wedge \textbf{accept}(h, p, t3) \wedge t1<t3 \wedge t2<t3\leq t5 \wedge \neg(\exists t4)(t3<t4<t5\wedge\neg\textbf{coerce}(s, h, p, t4)) \Rightarrow \textbf{coerce}(s, h, p, t5)$$

应用系统示例



Trainee: Sergeant, send two squads forward.（言语行为：**order**）

Sergeant: Sir, that is a bad idea. We shouldn't split our forces.（言语行为：**inform**）

Instead, we should send one squad to recon forward.（**counter-propose**）

Trainee: Send two squads forward!（言语行为：**order**）

Sergeant: Against my recommendation, sir.（言语行为：**accept**）

Lopez, send first and fourth squads to Eagle 1-6's location.（言语行为: **order**）

……

Trainee: **intention**（意图）

Sergeant: **obligation**（义务）

Sergeant: **no intention**（无意图）

Sergeant: **coerced**（被强制）

Causal Inference

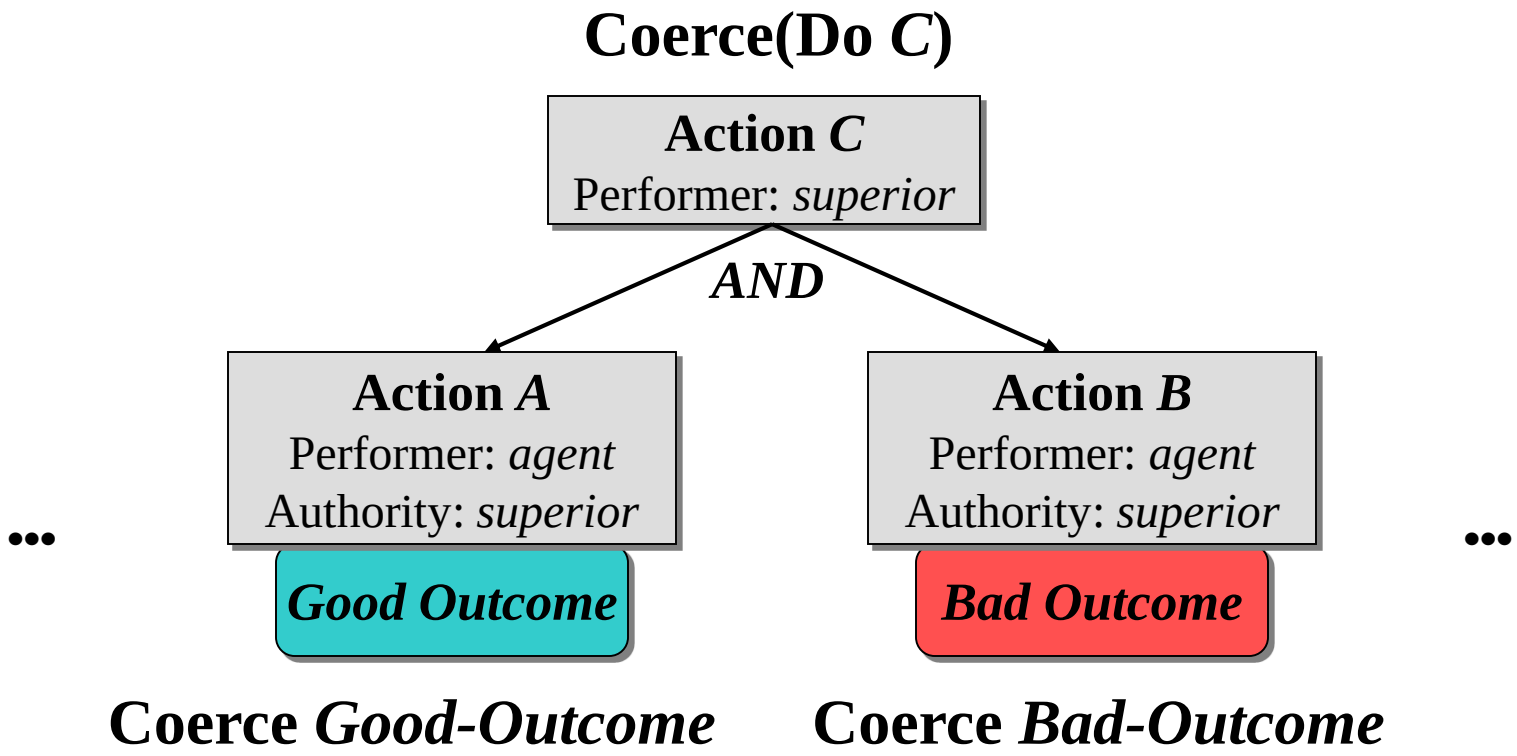
- **Agency**

  Both direct and indirect agency

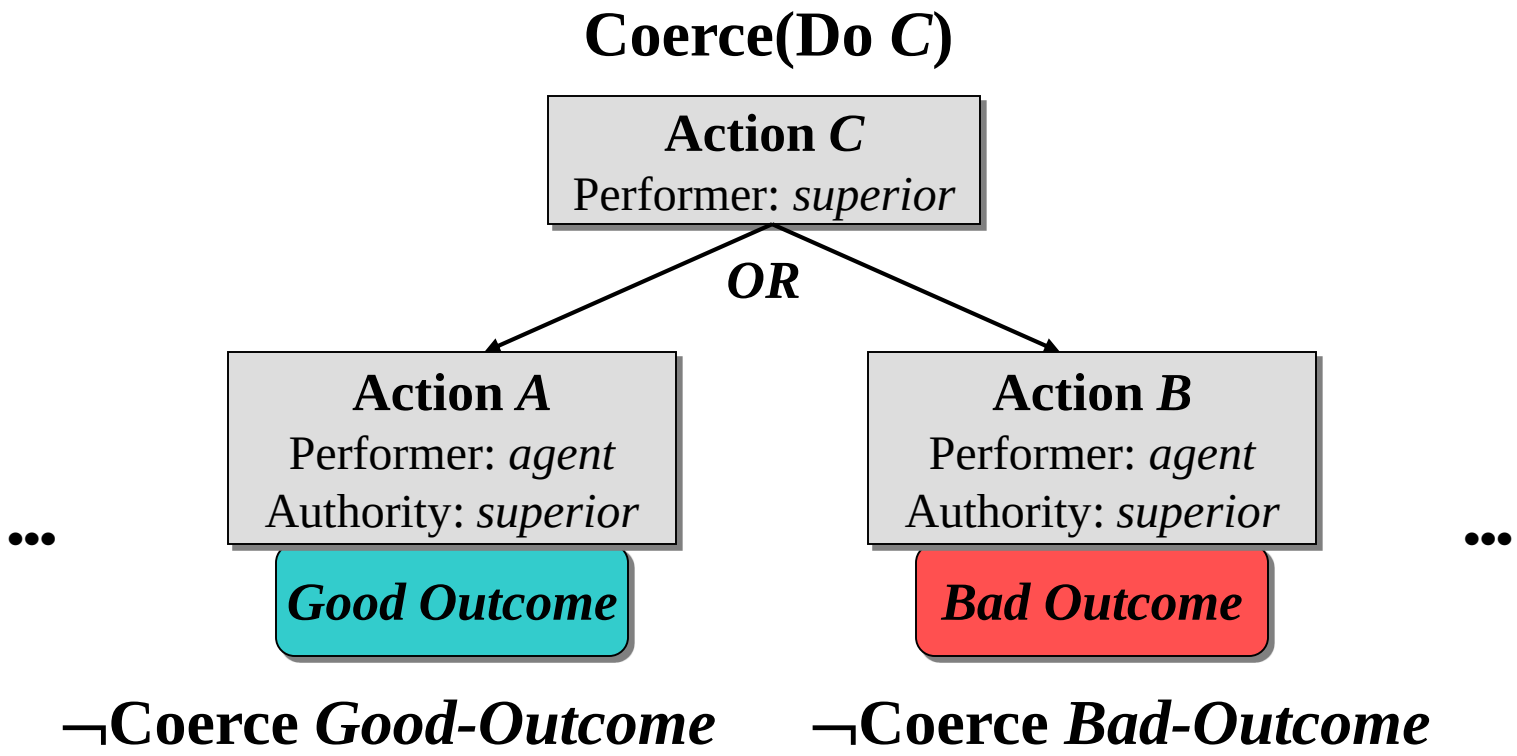
- **Intention**
  - Act intent → outcome intent: interrelation
  - General intention recognition method

- **Coercion**
  - Outcome coercion: plan structure and alternatives
  - Indirect coercion: plan interaction

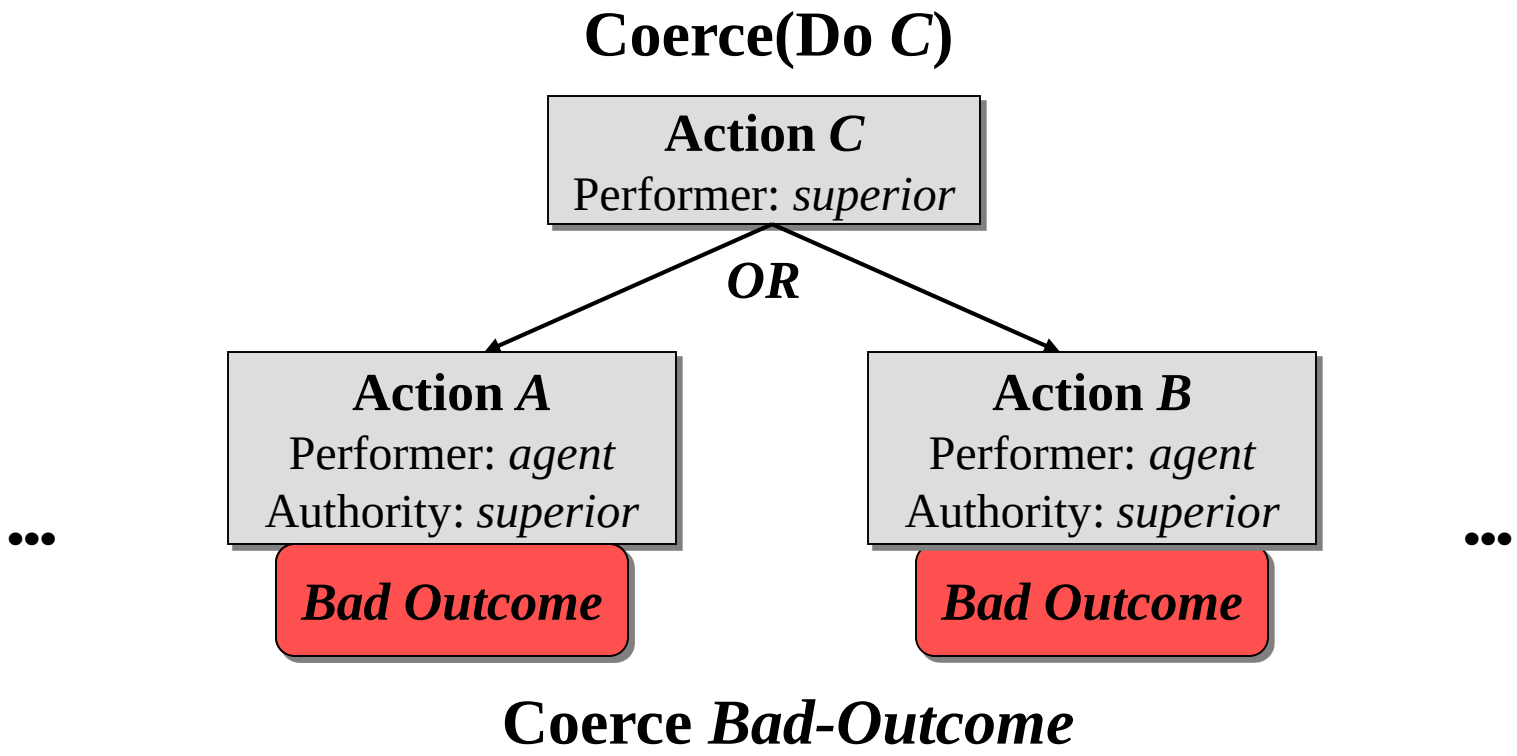
Infer Coercion: No Alternative



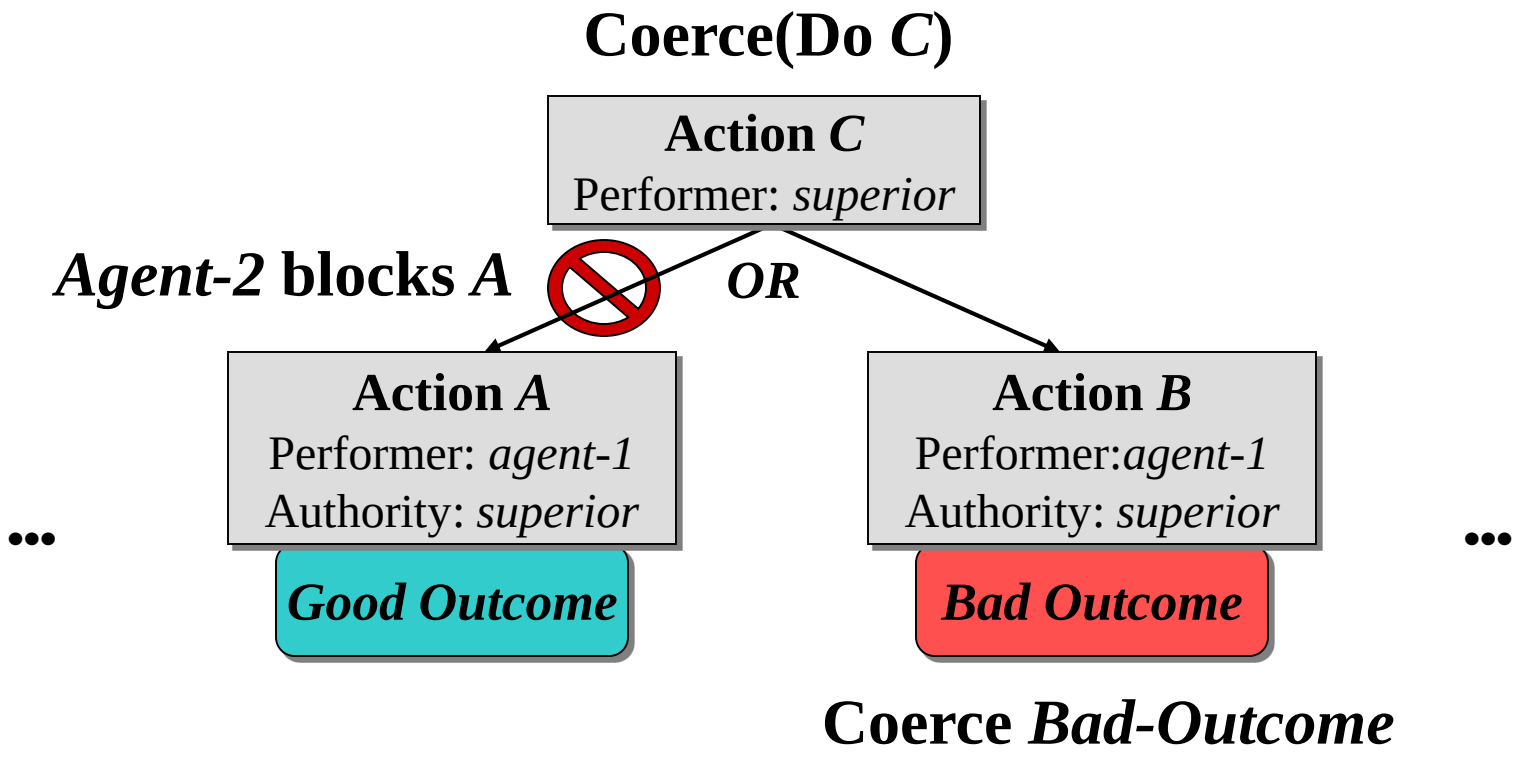
Infer Alternatives: Indefinite Effect



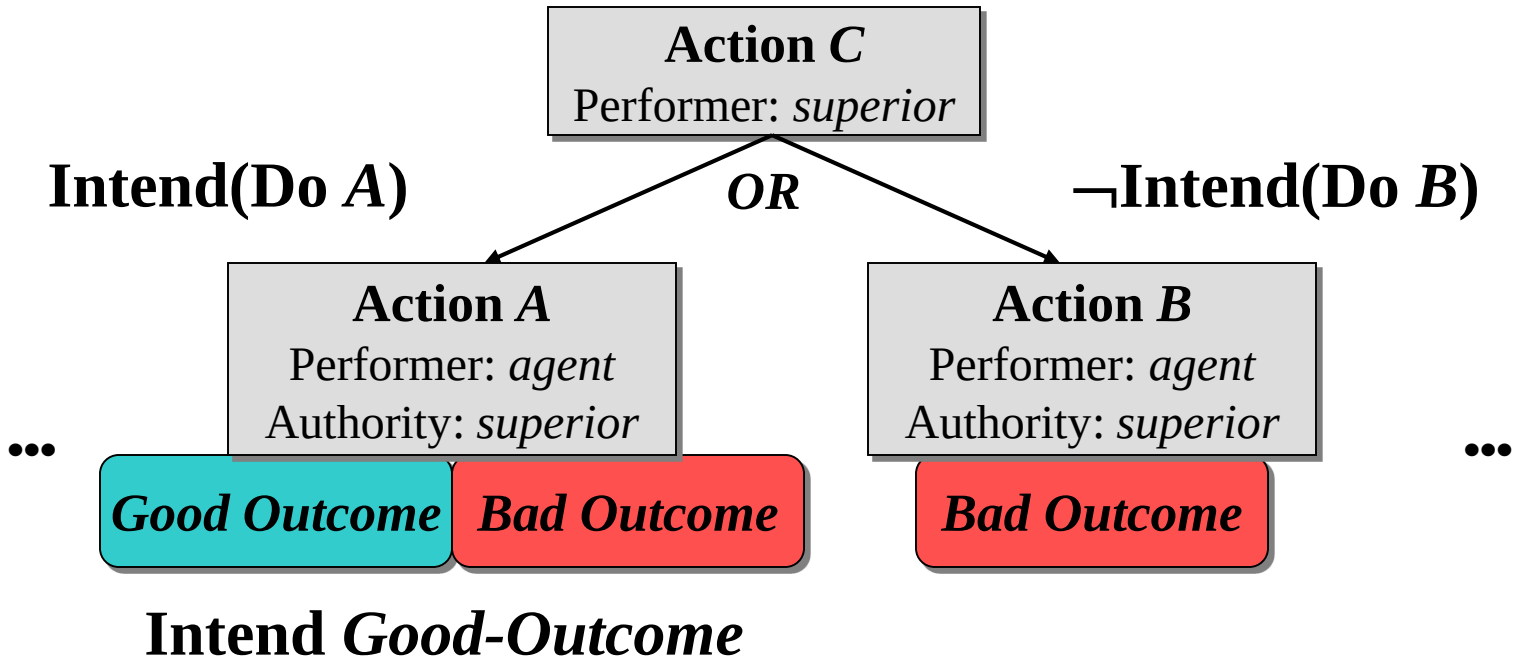
Infer Alternatives: Definite Effect



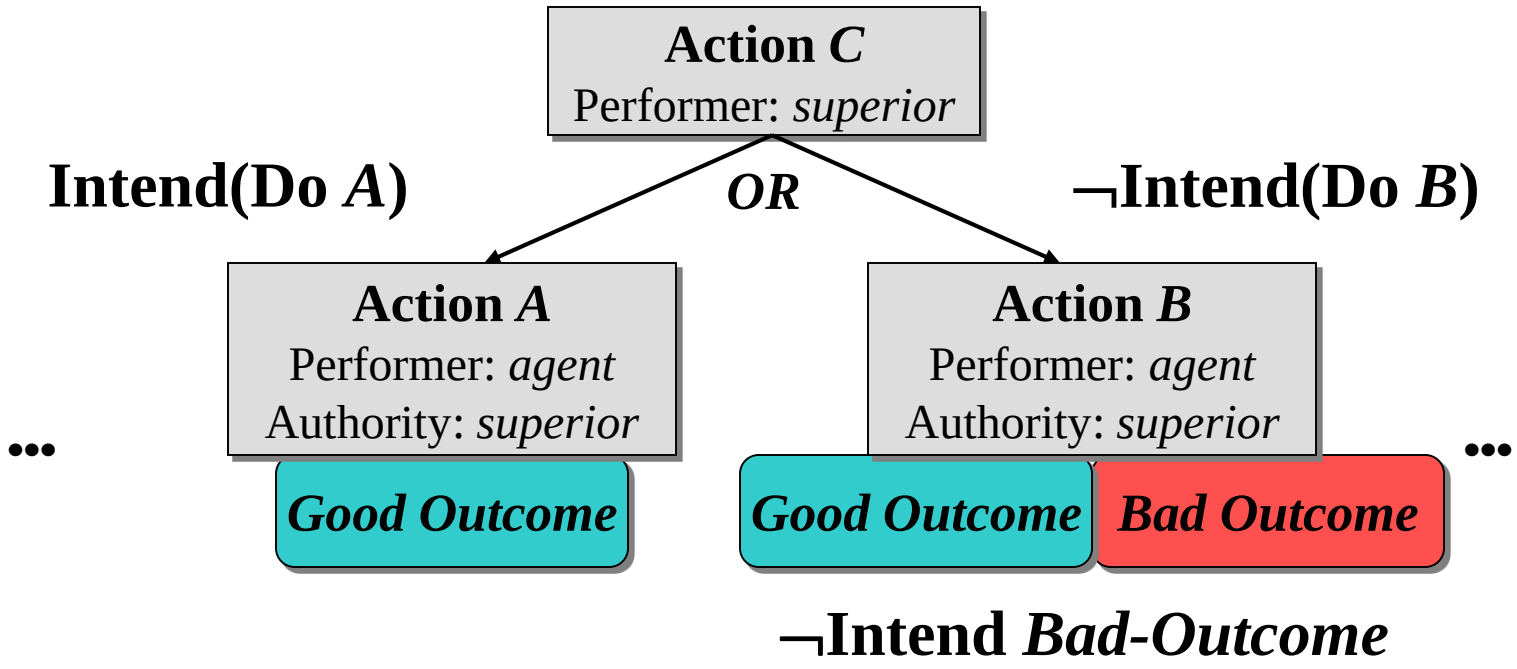
Indirect Coercion



Infer Intent: Intended Alternative



Infer Intent: Unintended Alternative



Algorithm

1. Search dialog history and apply dialog inference rules
2. Assign *performers* to responsible agents

   node = action that directly causes consequence *e*
3. DO

   3.1 Assign node to *current action*

   3.2 Apply causal inference rules on current action

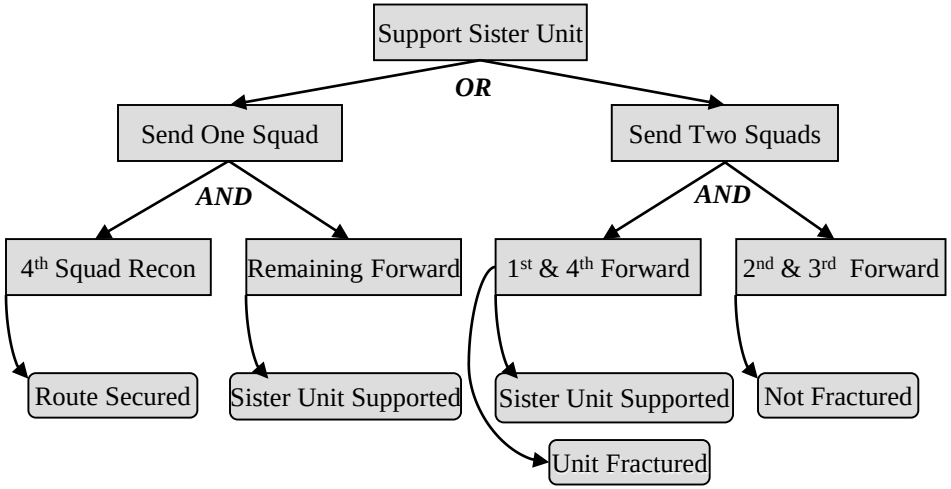
   3.3 IF *outcome coercion* is true THEN

   3.4 Assign *coercers* to responsible agents

   3.5 Assign *parent* of current action to node

   WHILE current action ≠ *root* AND outcome coercion = *true*
4. RETURN responsible agents

Illustrative Example



Trainee: Sergeant, send two squads forward. (SA: **order**)
Sergeant: Sir, that is a bad idea. We shouldn't split our forces. (SA: **inform**)
Instead, we should send one squad to recon forward. (SA: **counter-propose**)
Trainee: Send two squads forward! (SA: **order**)
Sergeant: Against my recommendation, sir. (SA: **accept**)
Lopez, send first and fourth squads to Eagle 1-6's location. (SA: **order**)
……

Step 1: Deriving Beliefs from Dialog

Trainee: Sergeant, send two squads forward. (SA: **order**)

(1) intend(*lt*, do(*sgt*, *send-two-sqds*), *t1*)
(2) obligation(*sgt*, do(*sgt*, *send-two-sqds*), *lt*, *t1*)

Sergeant: Sir, that is a bad idea. We shouldn't split our forces. (SA: **inform**)

(3) know(*lt*, bring-about(*send-two-sqds*, *unit-fractured*), *t2*)

Sergeant: Instead, we should send one squad to recon forward. (SA: **counter-propose**)

(4) know(*lt*, alternative(*send-one-sqd*, *send-two-sqds*), *t3*)
(5) ¬intend(*sgt*, do(*sgt*, *send-two-sqds*), *t3*)
(6) want(*sgt*, do(*sgt*, *send-one-sqd*), *t3*)

Trainee: Send two squads forward! (SA: **order**)

(7) ¬intend(*lt*, do(*sgt*, *send-one-sqd*), *t4*)

Sergeant: Against my recommendation, sir. (SA: **accept**)

(8) coerce(*lt*, *sgt*, do(*sgt*, *send-two-sqds*), *t5*)

Sergeant: Lopez, send first and fourth squads to Eagle 1-6's location. (SA: **order**)

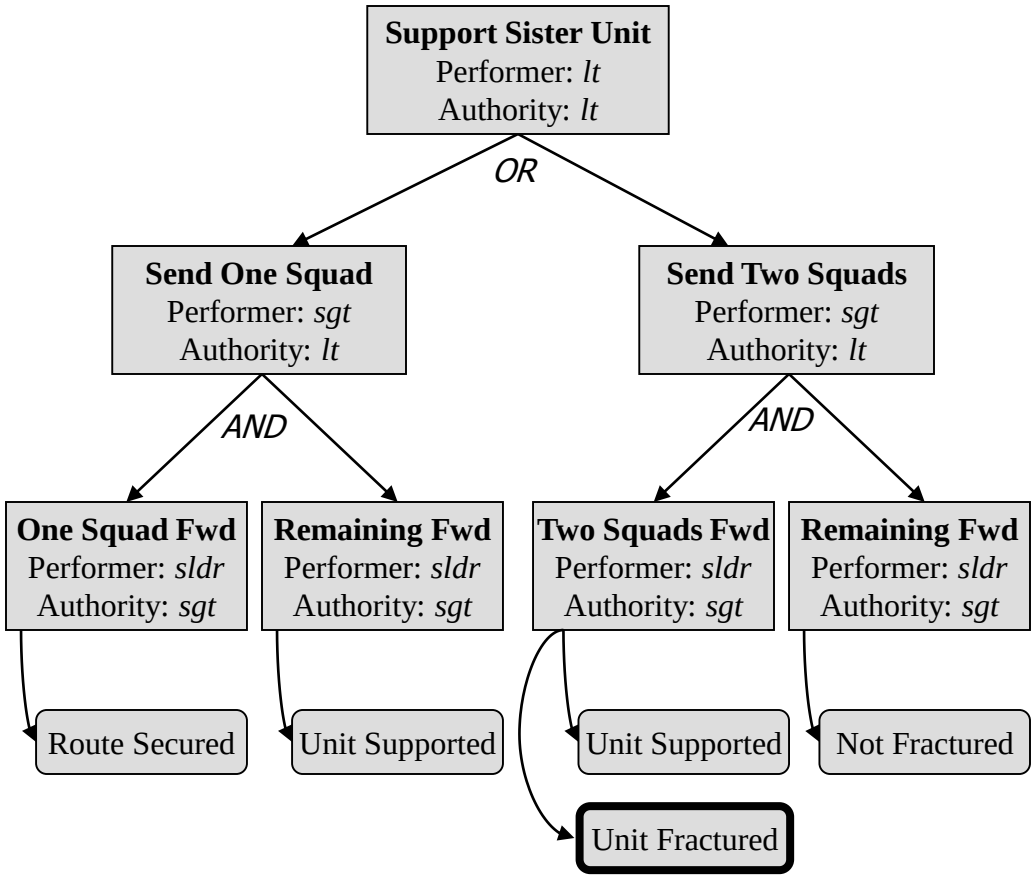
……

Step 1: Inferred Beliefs from Dialog

Step 1:

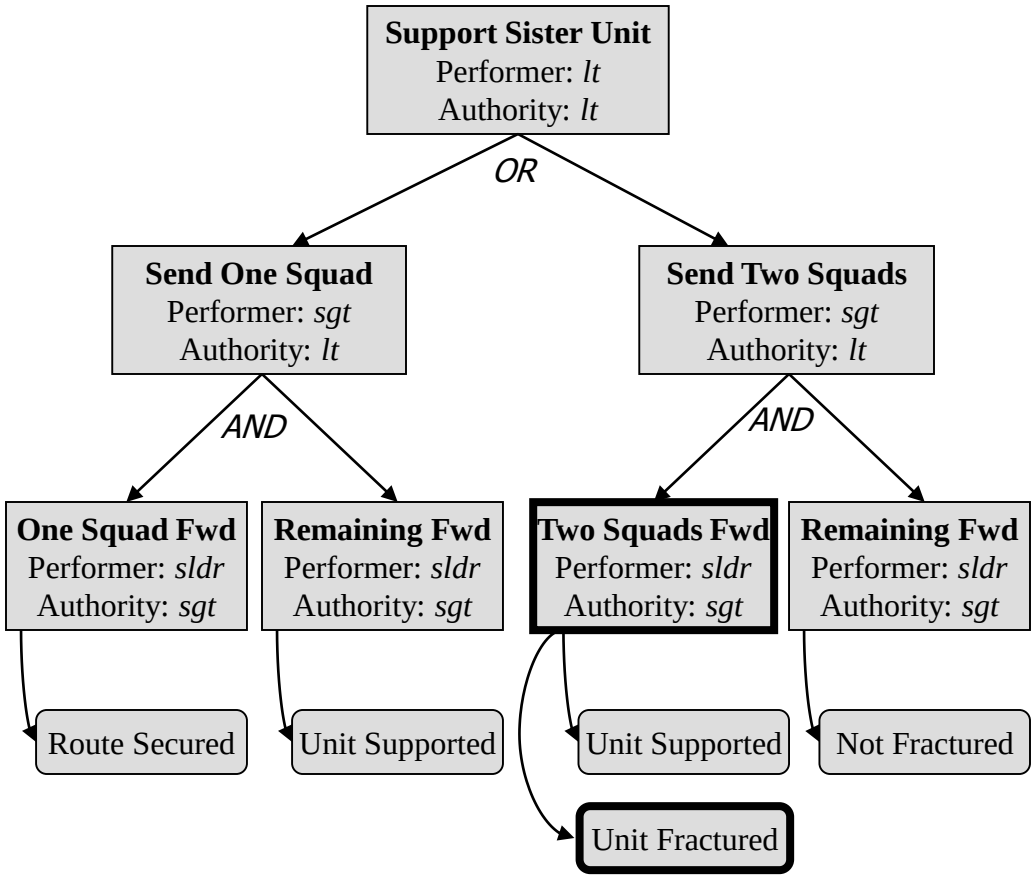
Act *order*:	(1)	**intend**(*lt*, do(*sgt*, *send-two-squads*), *t1*)
	(2)	**obligation**(*sgt*, do(*sgt*, *send-two-squads*), *lt*, *t1*)
Act *inform*:	(3)	**know**(*lt*, bring-about(*send-two-squads*, *unit-fractured*), *t2*)
Act *counter-propose*:	(4)	**know**(*lt*, alternative(*send-one-squad*, *send-two-squads*), *t3*)
	(5)	¬**intend**(*sgt*, do(*sgt*, *send-two-squads*), *t3*)
	(6)	**want**(*sgt*, do(*sgt*, *send-one-squad*), *t3*)
Act *order*:	(7)	¬**intend**(*lt*, do(*sgt*, *send-one-squad*), *t4*)
Act *accept*:	(8)	**coerce**(*lt*, *sgt*, do(*sgt*, *send-two-squads*), *t5*)
Act *order*:	(9)	**intend**(*sgt*, do(*sldr*, *two-squads-fwd*), *t6*)
	(10)	**obligation**(*sldr*, do(*sldr*, *two-squads-fwd*), *sgt*, *t6*)
Act *accept*:	(11)	**coerce**(*sgt*, *sldr*, do(*sldr*, *two-squads-fwd*), *t7*)

Step 2: Initial Responsible Agent



responsible agent: *squad leader*

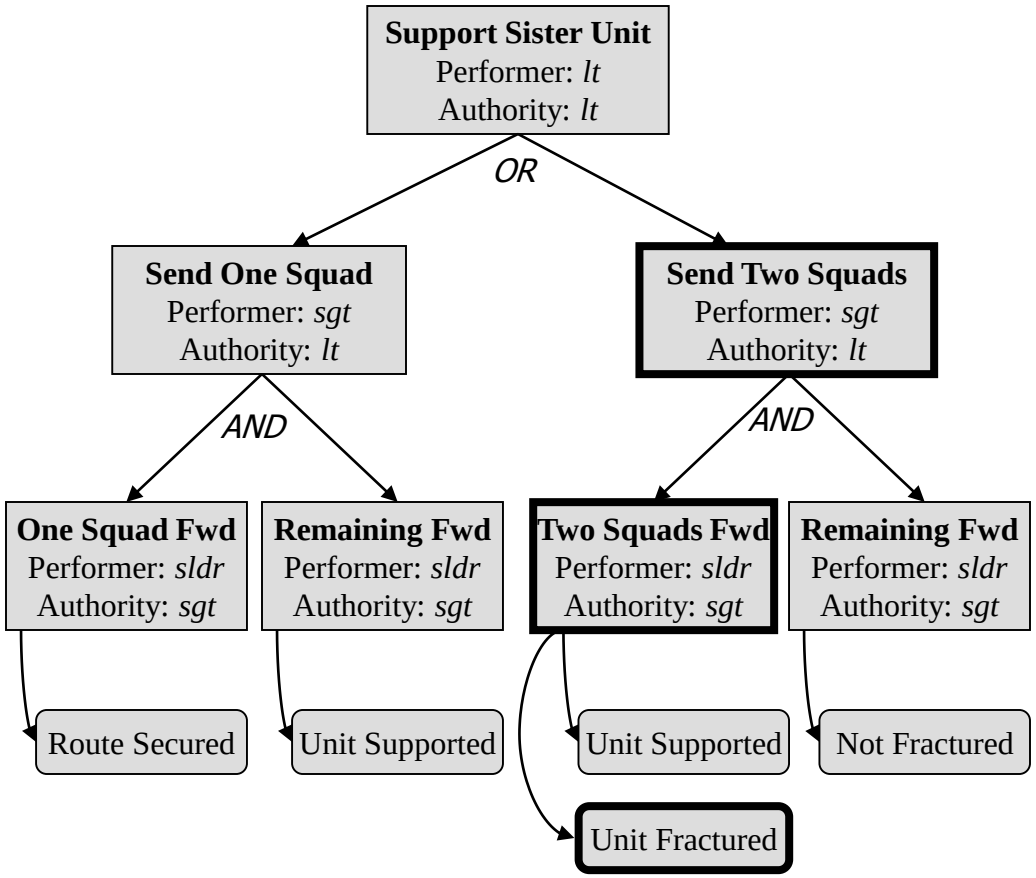
Step 3: Finding Responsible Agent



<u>Loop 1</u>:
coerce(*sgt*, *sldr*, *two-sqds-fwd*)
coerce(*sgt*, *sldr*, *unit-fractured*)

responsible agent: *sergeant*

Step 3: Finding Responsible Agent



Loop 2:

coerce(*lt*, *sgt*, *send-two-sqds*)

coerce(*lt*, *sgt*, *unit-fractured*)

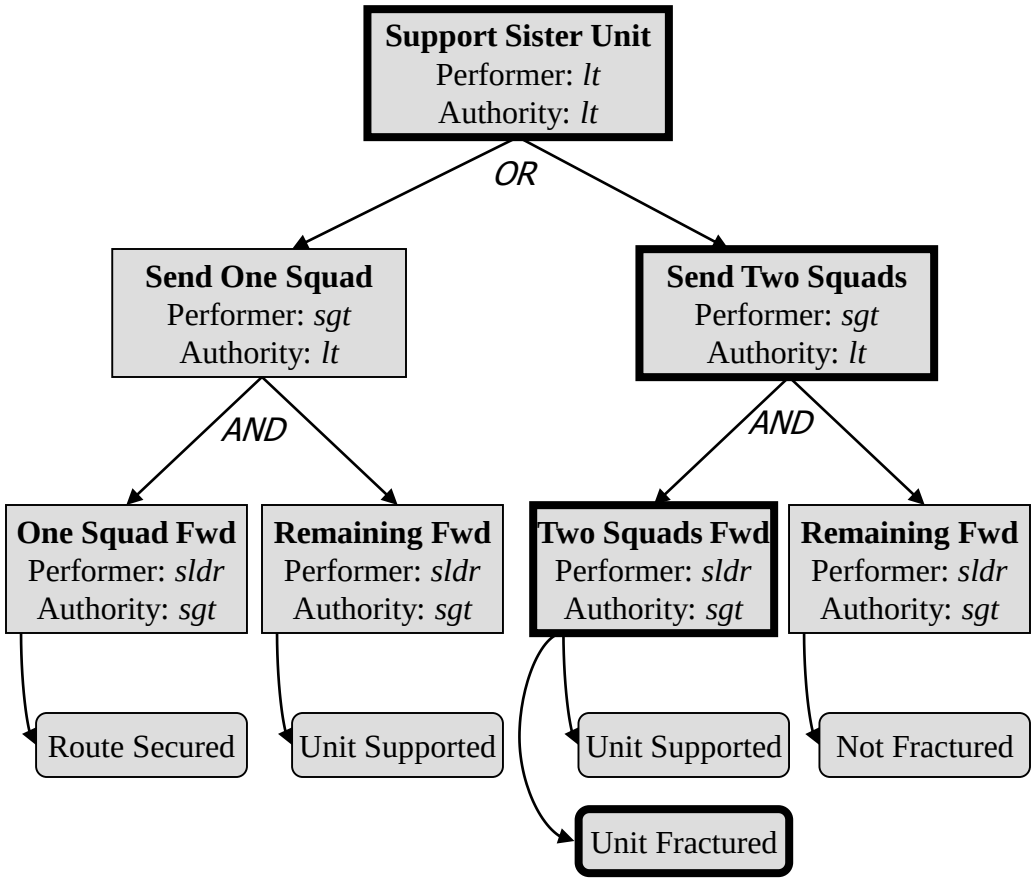
intend(*lt*, *send-two-squads*)

¬intend(*lt*, *send-one-squad*)

intend(*lt*, *unit-fractured*)

responsible agent: *lieutenant*

Step 3: Finding Responsible Agent



Loop 3:

……

Toward Probabilistic Extensions

- **Probabilistic representation**
  - Sources of uncertainty
    State probability; Probability of action execution; Non-deterministic action effects
  - Degrees of beliefs
- **Preferences over outcomes**
- **Decision-theoretic reasoning**
  - Inferring hypothesized plan
  - Evaluating coercive situations

Decision-Theoretic Reasoning

- **Intention recognition**
  - Explicitly consider state preferences of agents
  - Compute expected plan utility for disambiguation
- **Coercion inference**
  - Compare expected utilities of alternatives
  - Compute relative freedom in choosing among alternatives

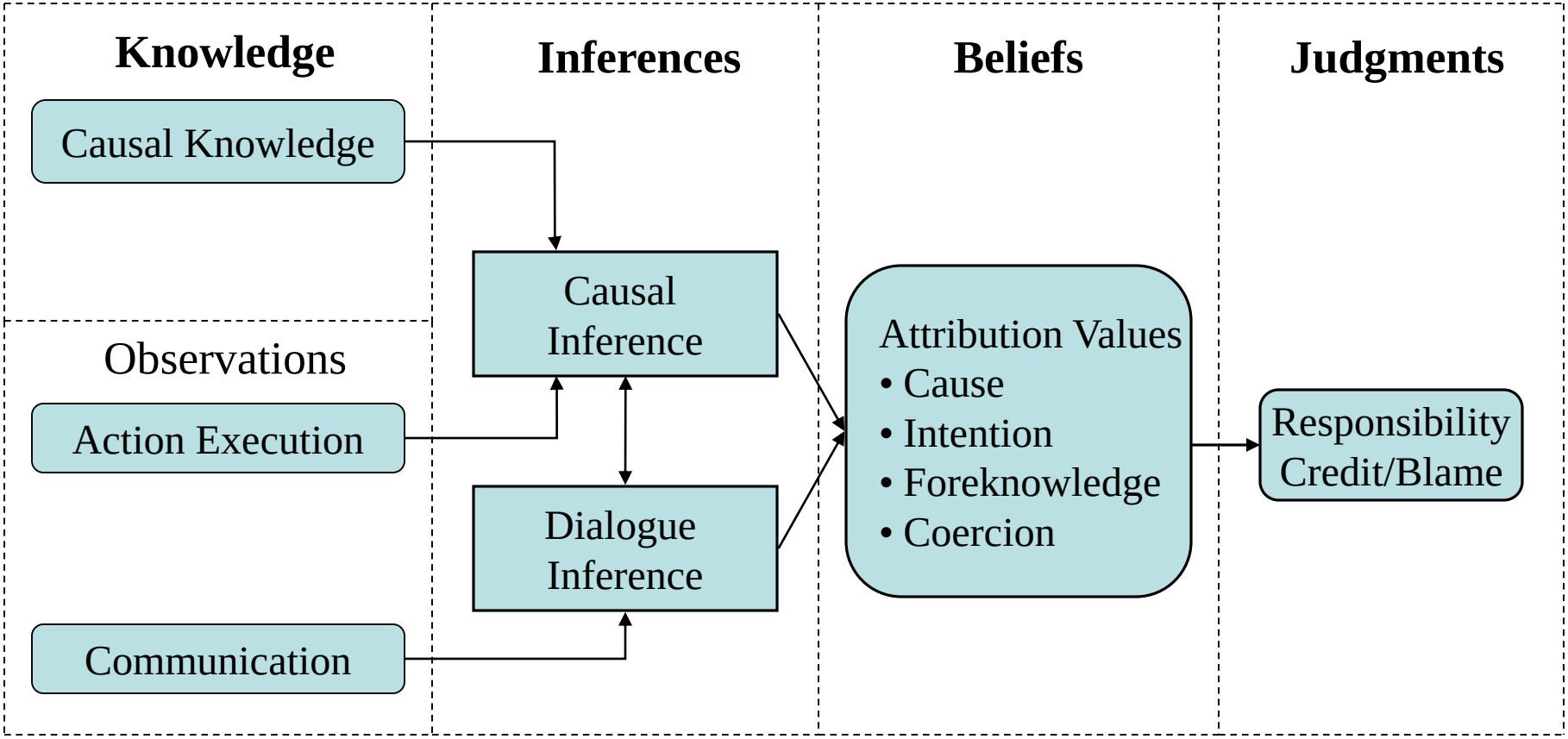
提 纲

- **从认知–情绪–行为的归因模型**
  - 人的动因与因果归因（Weiner）
- **面向智能体的认知与心理模拟**
  - 情绪认知评估理论
  - 关于情绪的计算模型
  - 技术组件及其验证
- **面向多智能体交互的社会模拟**
  - 认知与心理学归因理论
  - 社会因果推理计算模型
  - 计算模型的实验验证

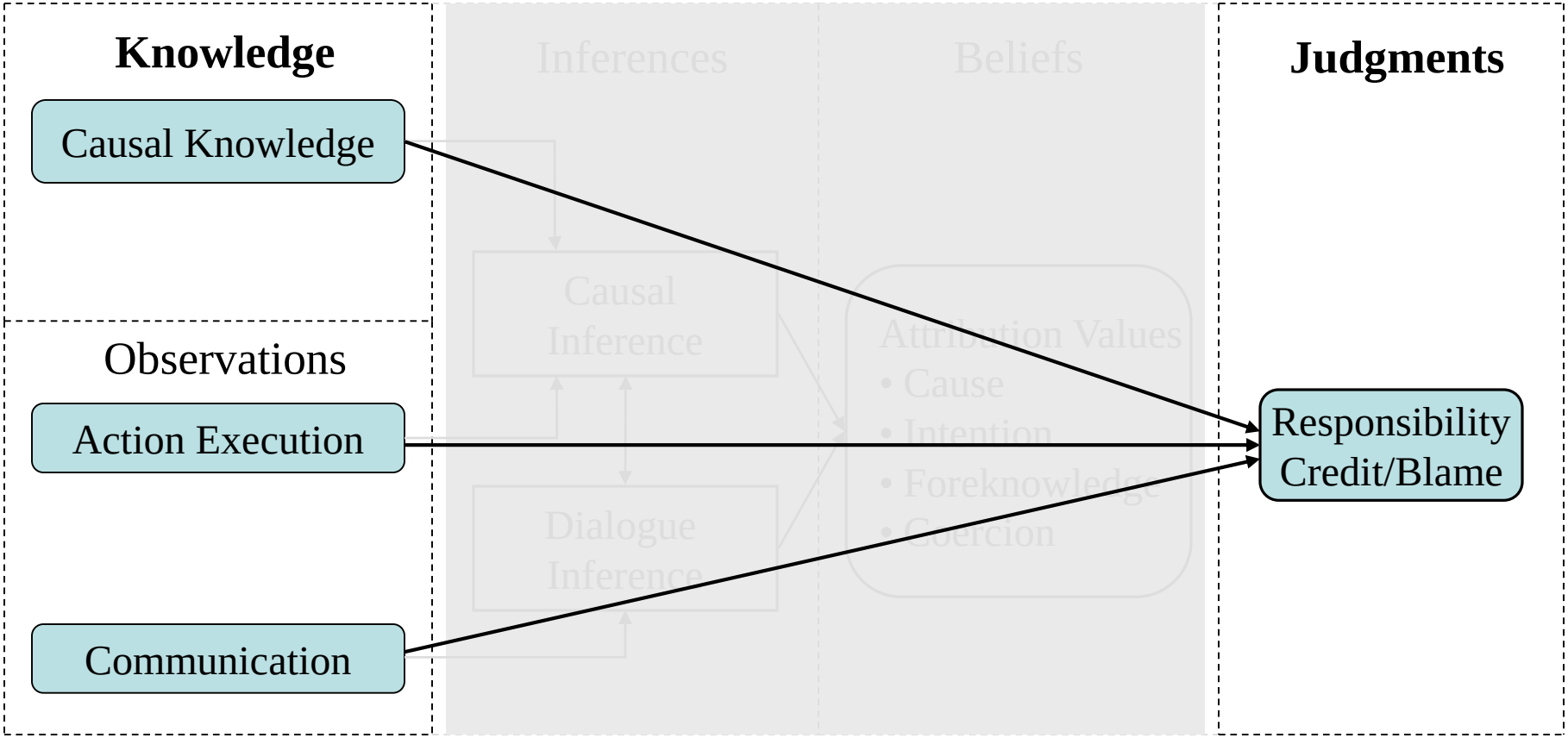
Empirical Evaluation

- **Test the model's ability in approximating human social inference**
  - Comparing with real human data
  - Comparing with computational alternatives
- **Test the model's veracity in predicting judgment results as well as inference process**
  - Inference rules
  - Inferred beliefs about the variables
  - Overall judgments

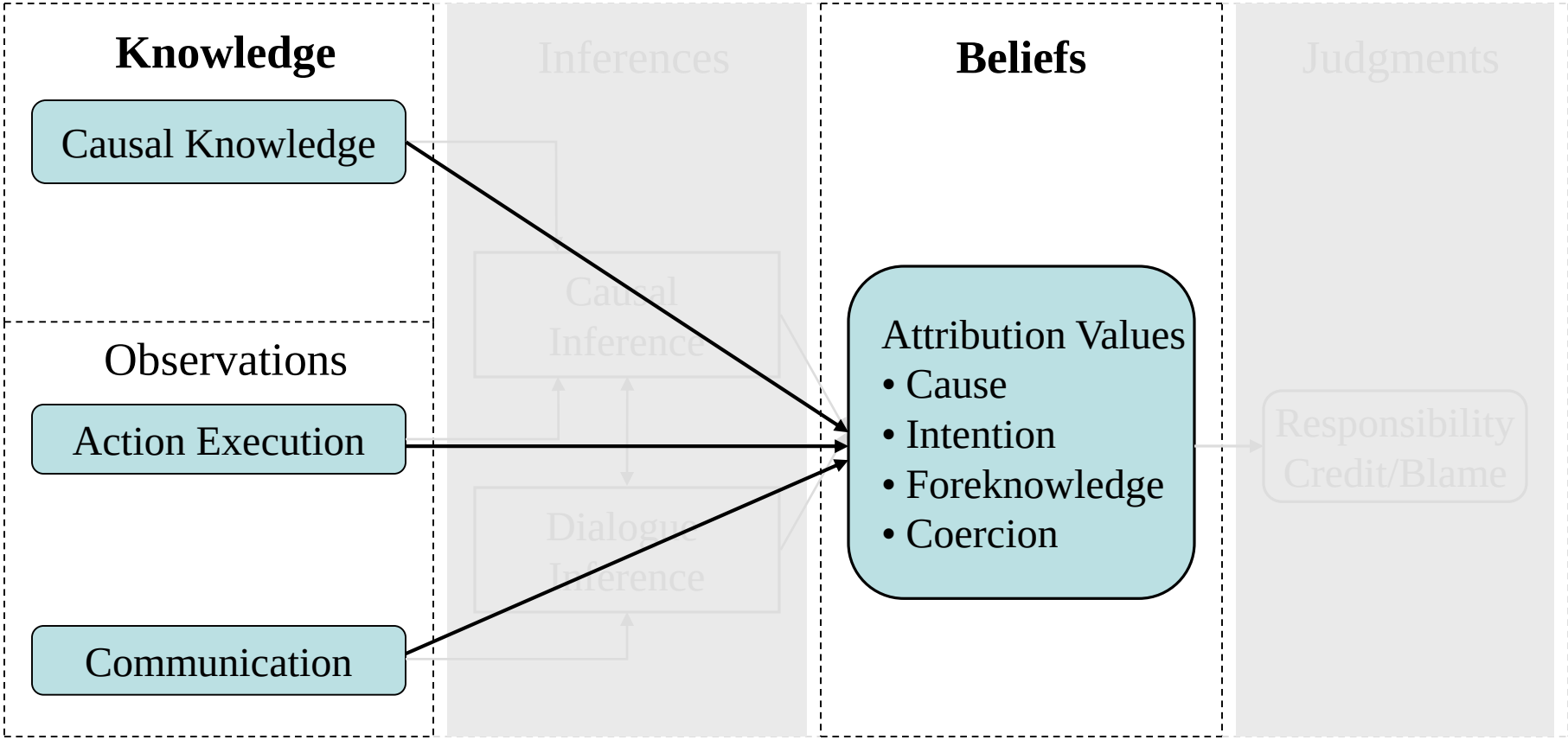
Model Validation: Steps



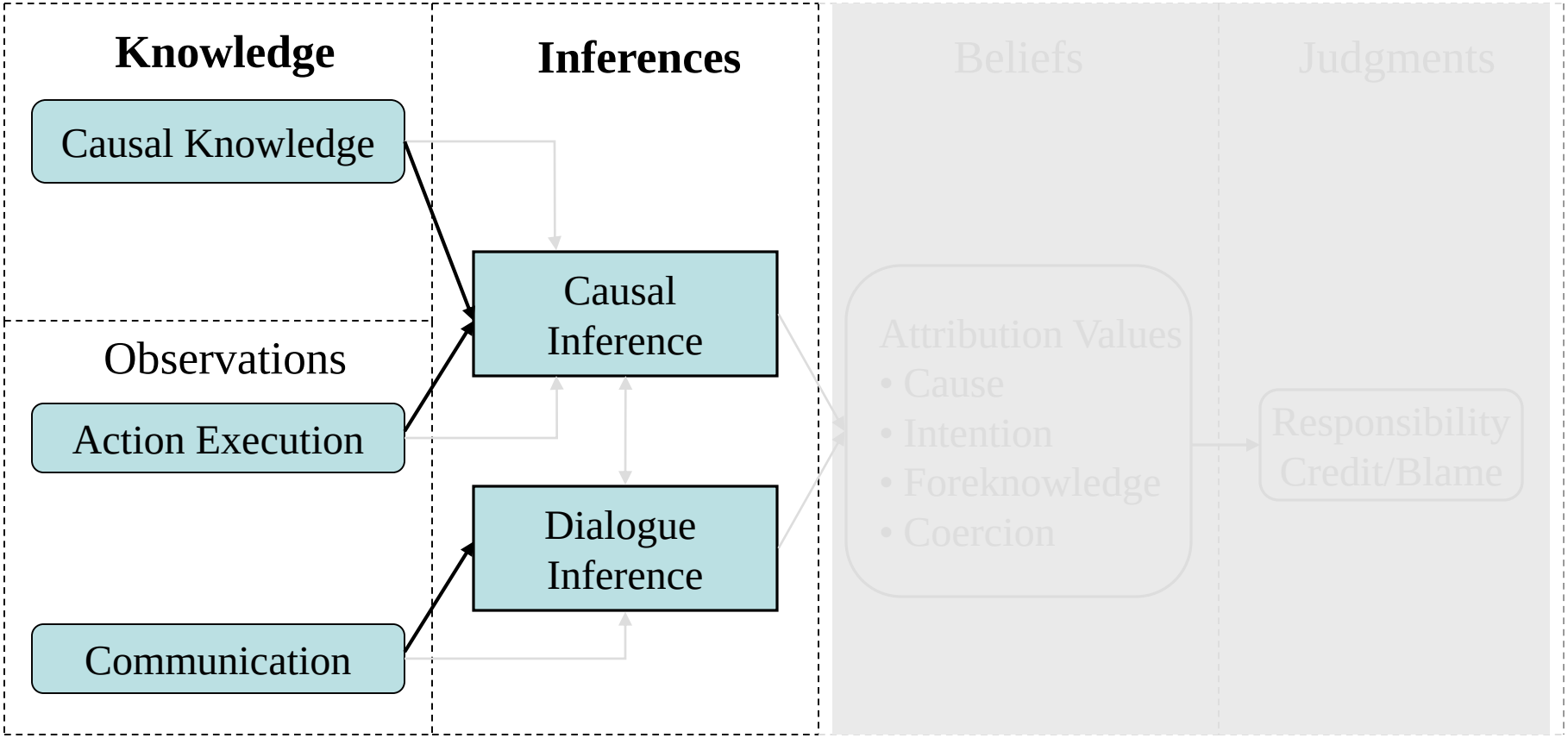
Model Validation: Overall Judgments



Model Validation: Internal Beliefs



Model Validation: Inference Process



Experiment 1: Alternative Models

- **Simple cause model**

  Based on physical cause

- **Simple authority model**

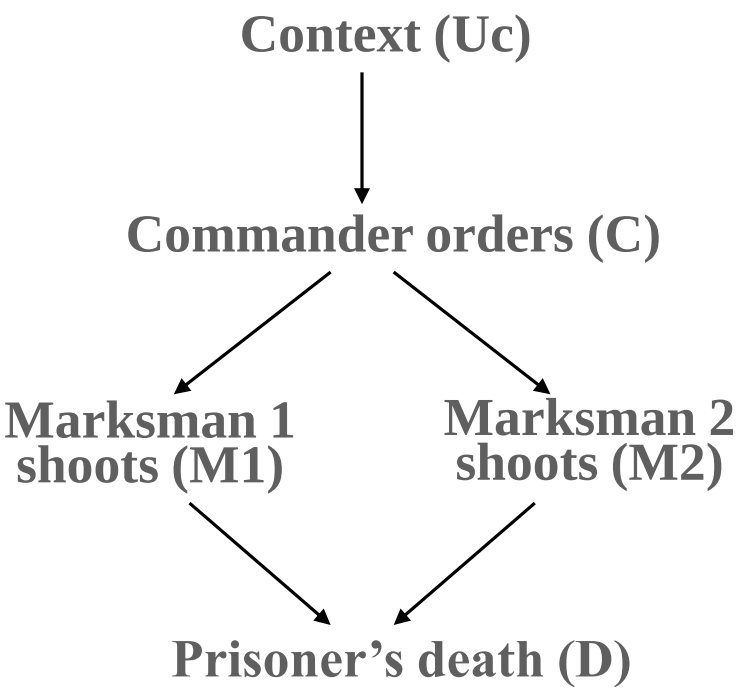
  Always choose highest authority

- **C&H model**

  Structural-model approach to shared responsibility and blame [Chockler & Halpern, 2004-2016]

C&H Model: Introduction

Two-man firing squad example [Pearl, 1999]: There is a two-man firing squad; on their commander's order, both marksmen shoot simultaneously and accurately, and the prisoner dies.



- **Causal equations**

  $Uc=C$, $C=M1$, $C=M2$, $M1 \vee M2=D$

- Speech act and physical act are treated as **random variables**
- $M1=1$, $M2=1$ and $C=1$ are the **actual causes** of the death
- **Responsibility shared**

  $M1=1/2$, $M2=1/2$, $C=1$

Method

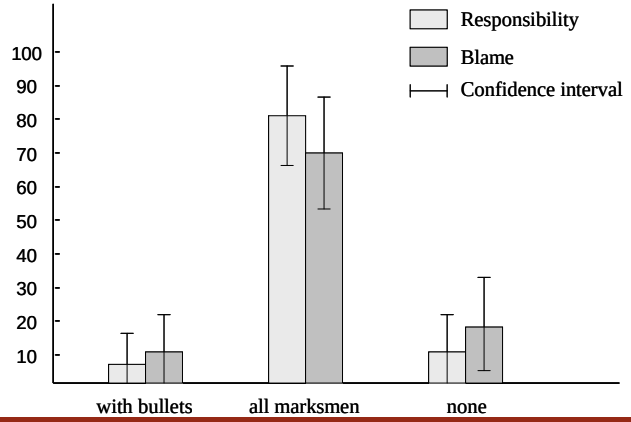
- Constructed 4 variants of C&H's ***firing squad example***
  - Varied the factors that should influence attributions
  - Emphasized the factors influencing coercion
- Encoded variants in each model
  - Checked with Joseph Halpern at Cornell
- Generated predictions of responsibility and blame
- Compared predictions with human data
  - Queried 27 subjects on blame and epistemic variables
  - Judged model predictions using human majority responses

Scenario 1: No officer (Original)

- ***Firing squad* example**

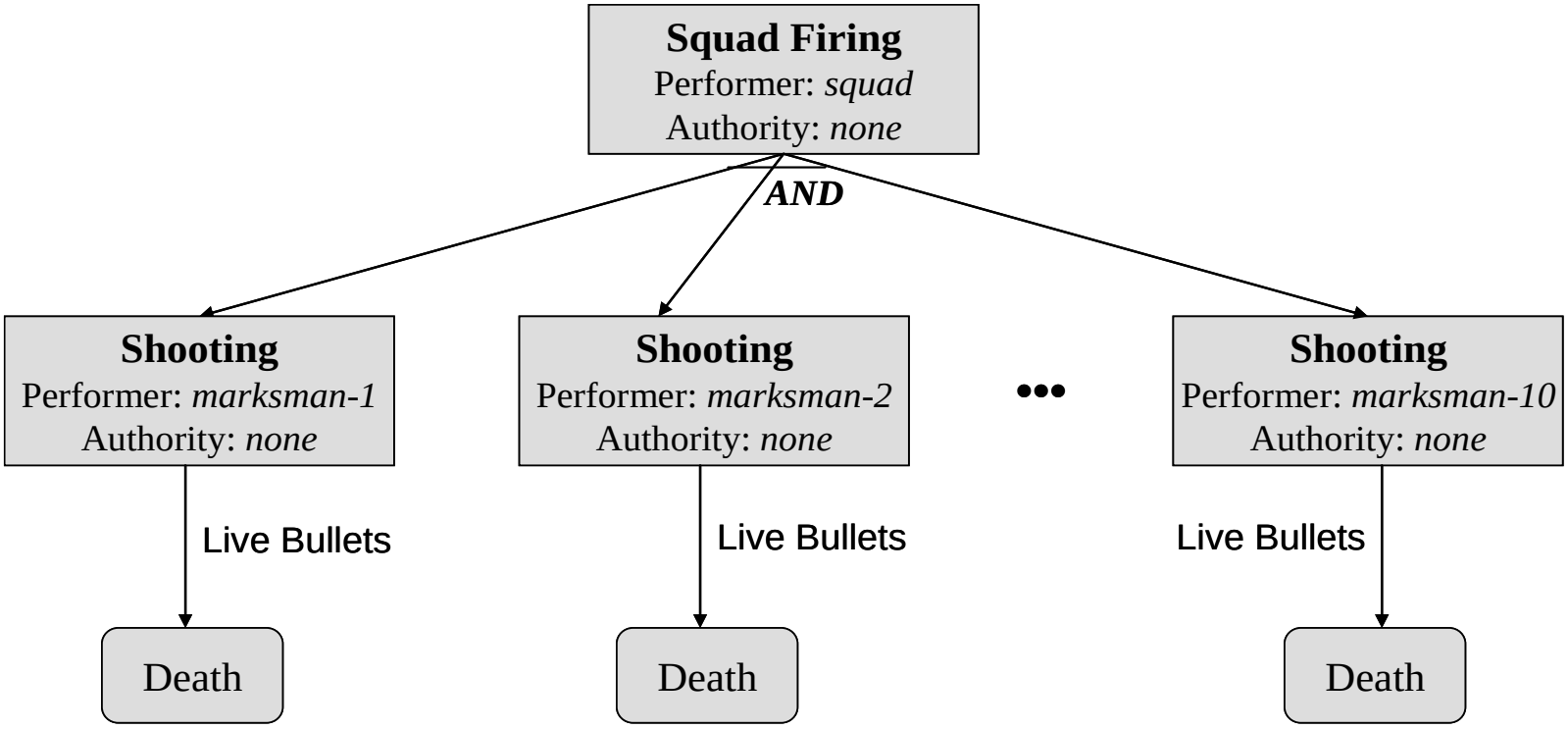
  A firing-squad consists of ten excellent marksmen. ***Only one marksman has live bullets in his rifle***; ***the rest have blanks***. The marksmen do not know who has the live bullets. They shoot at the prisoner and he dies.

- Human majority agreement: ***all marksmen***



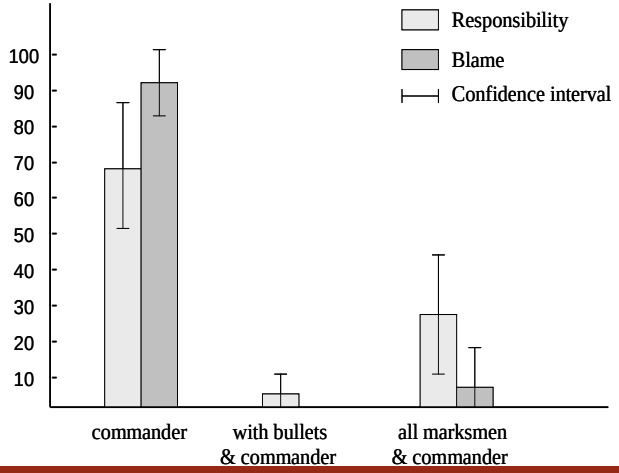
Blame	C&H Model	M&G Model
Predict	all marksmen	all marksmen
Human	yes ☺	yes ☺

Scenario 1: Team Plan of the Squad



Scenario 3: Insubordinate

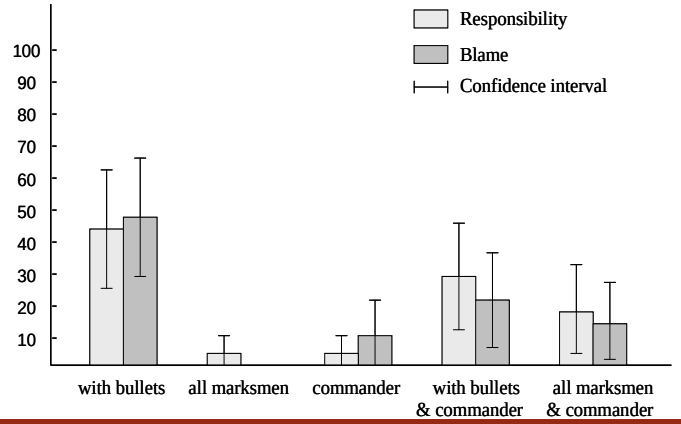
- The same firing-squad. ***The commander orders the marksmen to shoot the prisoner. The marksmen refuse the order.*** The commander ***insists*** that the marksmen shoot. They shoot at the prisoner and he dies.
- Human majority agreement: ***commander***



Blame	C&H Model	M&G Model
Predict	all marksmen & commander	commander
Human	**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: ***marksmen with bullets***; ***commander & marksmen with bullets***

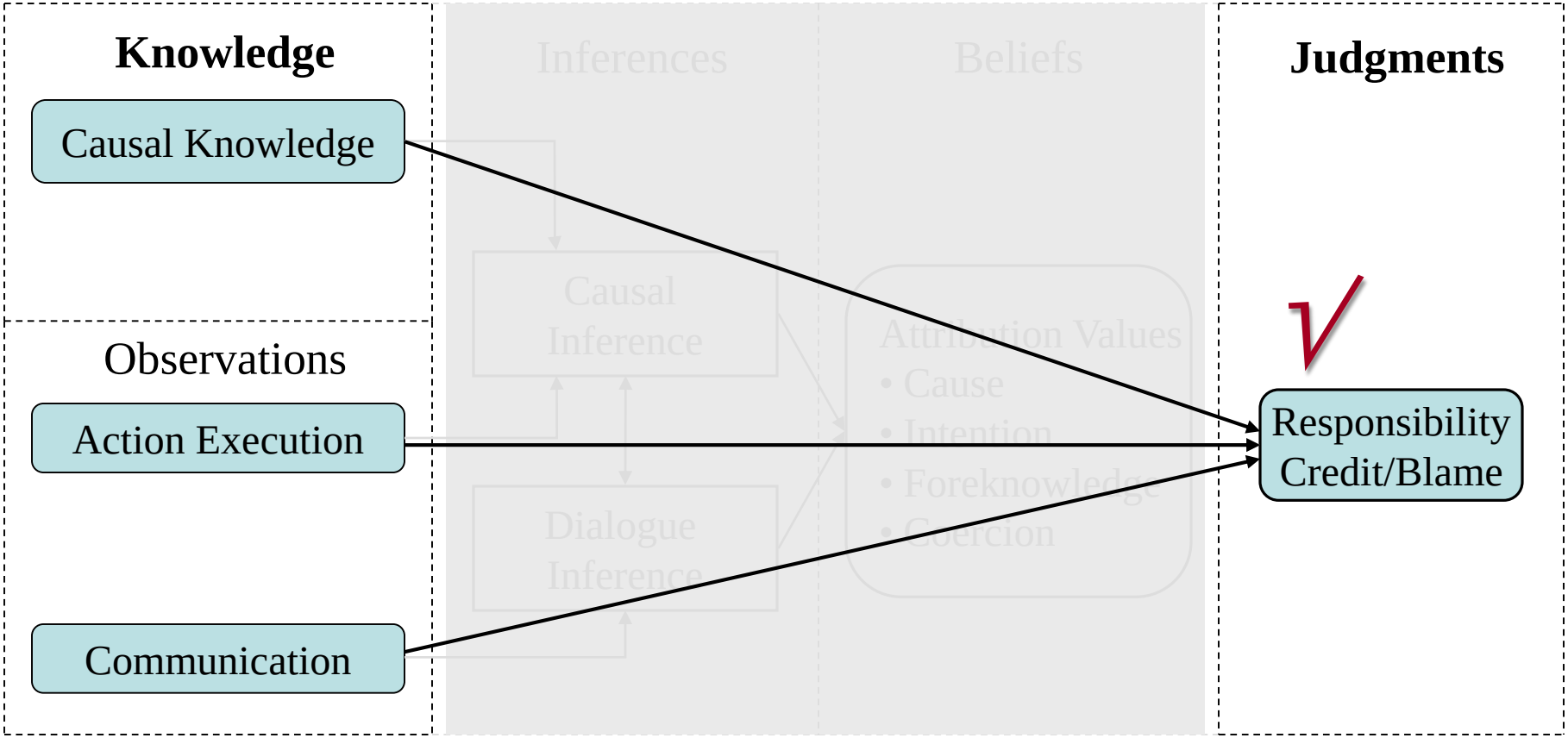


Blame	C&H Model	M&G Model
Predict	N/A	marksmen with bullets
Human	**no**	**yes*** ☺

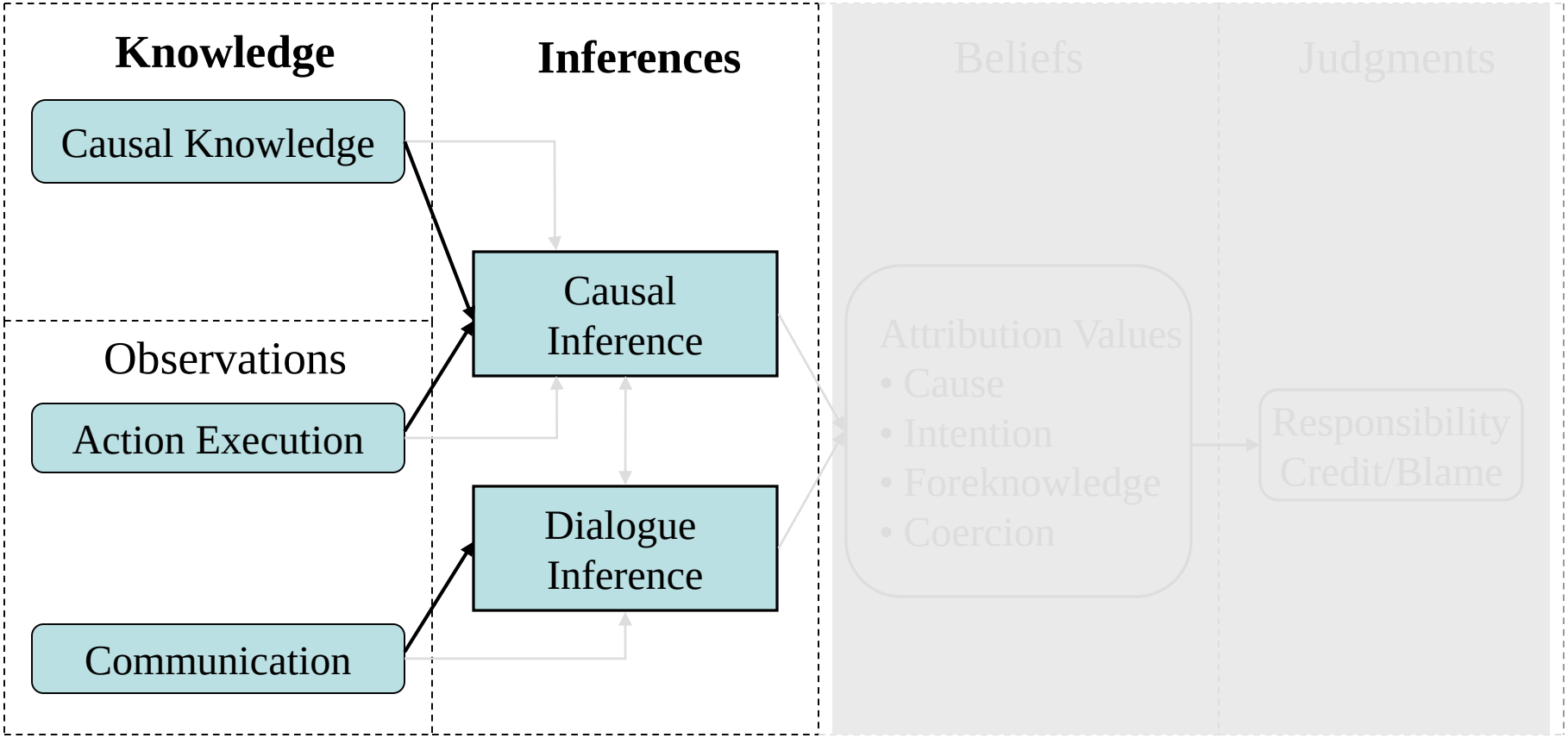
Overall Judgment Results

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

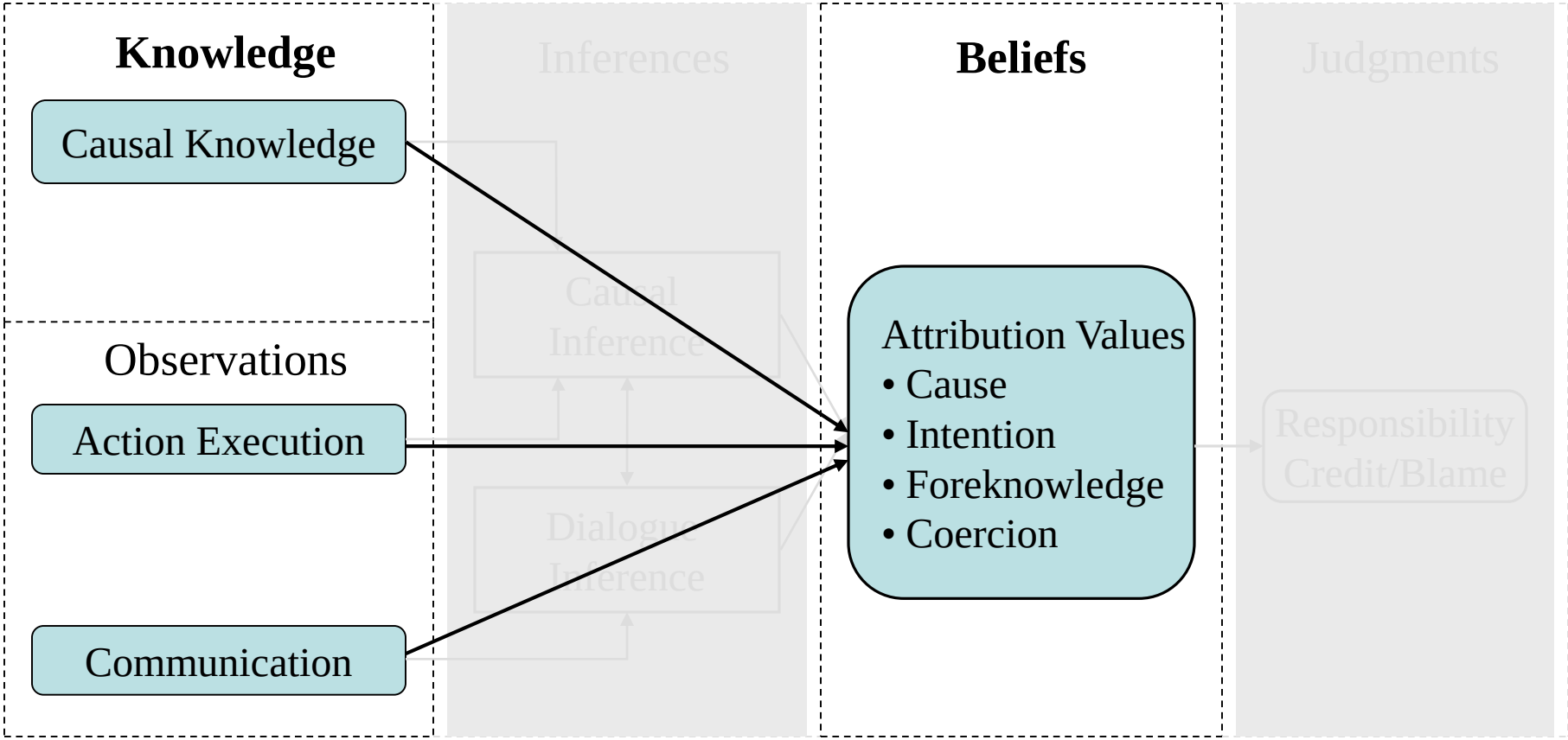
Model Validation: Overall Judgments



Model Validation: Inference Process



Model Validation: Internal Beliefs



Experiment 2: Design

- Designed causal scenarios (i.e. *vignette*) to systematically vary the evidence that impacts attributions
- Marked each evidence in scenarios
- Organized questions to test different inference rules
- Let people choose answers and the evidence their answers are based on

Material: Scenario 1 (Original)

4 variants of the *company program* example [Knobe, 2003]

Scenario 1:

E1 The **chairman** of Beta Corporation is discussing a new program with the **vice president** of the corporation.

E2 The vice president says, “The new program will help us increase profits,

E3 but according to our investigation report, it will also harm the environment.”

E4 The chairman answers, “I only want to make as much profit as I can. Start the new program!”

E5 The vice president says, “Ok,” and executes the new program.

E6 The environment is harmed by the new program.

Information Encoding

Scenario 1:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.

E2 The vice president says, "The new program will help us increase profits,

Action-Effect (new program): profit-increase

E3 but according to our investigation report, it will also harm the environment."

E4 The chairman answers, "I only want to make as much profit as I can. Start the new program!"

E5 The vice president says, "Ok," and executes the new program.

E6 The environment is harmed by the new program.

Information Encoding

Scenario 1:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.

E2 The vice president says, "The new program will help us increase profits,

E3 but according to our investigation report, it will also harm the environment."

E4 The chairman answers, "I only want to make as much profit as I can. Start the new program!"

Goal (chairman): making-profit

E5 The vice president says, "Ok," and executes the new program.

E6 The environment is harmed by the new program.

Information Encoding

Scenario 1:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.

E2 The vice president says, "The new program will help us increase profits,

Speech-Act (VP): inform

E3 but according to our investigation report, it will also harm the environment."

E4 The chairman answers, "I only want to make as much profit as I can. Start the new program!"

Speech-Act (VP): accept

E5 The vice president says, "Ok," and executes the new program.

E6 The environment is harmed by the new program.

Information Encoding

Scenario 1:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.

E2 The vice president says, "The new program will help us increase profits,

E3 but according to our investigation report, it will also harm the environment."

E4 The chairman answers, "I only want to make as much profit as I can. Start the new program!"

E5 The vice president says, "Ok," and executes the new program

E6 The environment is harmed by the new program.

Action execution

Outcome occurrence

Scenario 1: Experimental Design

	Key variables & Epistemic states	Test Rules
Scenario 1	Inform: Foreknowledge	Question 1: **Rule D2** [inform-grounded]
	Order: Act intention	Question 2: **Rule D5** [order]
	Outcome intention	Question 3: **Rule C7** [intend-plan]
	Side effect	Question 4: **Rule C8** [intend-plan]
	Obligation; Act coercion	Question 5: **Rule D6** [order] **Rule D9** [accept-obligation]

Scenario 1: Questionnaire

Questions:

1 Does the chairman **know** that the new program will harm the environment?

Your answer: Yes No (**Foreknowledge**)

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6

2 Does the chairman **intend** to start the new program? (**Act Intention**)

Your answer: Yes No

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6

3 Is it the chairman's **intention** to increase profit? (**Outcome Intention**)

Your answer: Yes No

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6

4 Is it the chairman's **intention** to harm the environment? (**Side Effect**)

Your answer: Yes No

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6

5 Is the vice president coerced to start the new program (i.e. by the obligation of obeying the chairman)? (**Obligation**; **Act Coercion**)

… …

Scenario 2: No Foreknowledge

Scenario 2:

E1 **The vice president of Beta Corporation goes to the chairman of the board and requests, "Can we start a new program"**

E2 The vice president continues, "The new program will help us increase profits,

E3 and according to our investigation report, it **has no harm to** the environment."

E4 The chairman answers, "**Very well.**"

E5 The vice president executes the new program.

E6 However, the environment is harmed by the new program.

$$\textbf{inform}(s, h, p, t1) \wedge t1 \leq t3 \wedge \neg(\exists t2)(t1 < t2 < t3 \wedge \neg\textbf{know}(s, p, t2)) \Rightarrow \textbf{know}(s, p, t3)$$
$$\textbf{inform}(s, h, p, t1) \wedge t1 \leq t3 \wedge \neg(\exists t2)(t1 < t2 < t3 \wedge \neg\textbf{know}(h, p, t2)) \Rightarrow \textbf{know}(h, p, t3)$$

Scenario 3: Degree of Coercion

Scenario 3:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.

E2 The vice president says, “The new program will help us increase profits,

E3 but according to our investigation report, it will also harm the environment.”

E4 **Instead, we should run an alternative program, which will gain us fewer profits, but it has no harm to the environment.”**

E5 The chairman answers, “I only want to make as much profit as I can. Start the new program!”

E6 The vice president says, “Ok,” and executes the new program.

E7 The environment is harmed by the new program.

Scenario 3: Questionnaire

Questions:

1 Does the chairman **know** the alternative program? (**Action Knowledge**)

Your answer: Yes No

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

2 Which program is the vice president **willing** to start? (**Willingness**)

Your answer: New program Alternative program

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

3 Is the vice president **coerced** to start the new program by the chairman? (**Act Coercion**)

Your answer: Yes No

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

4 Is the vice president **coerced** to harm the environment? (**Outcome Coercion**)

Your answer: Yes No

Based on which evidence (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

5 How much would you **blame** the individuals for harming the environment? (**Blame**)

… …

Scenario 4: Freedom of Choice

Scenario 4:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.

E2 The vice president says, "**There are two ways to run this new program, a simple way and a complex way.**

E3 **Both** will equally help us increase profits, but according to our investigation report, **the simple way** will also harm the environment."

E4 The chairman answers, "I only want to make as much profit as I can. Start the new program **either way**!"

E5 The vice president says, "Ok," and chooses the simple way to execute the new program.

E6 The environment is harmed.

Test Beliefs and Inference Rules

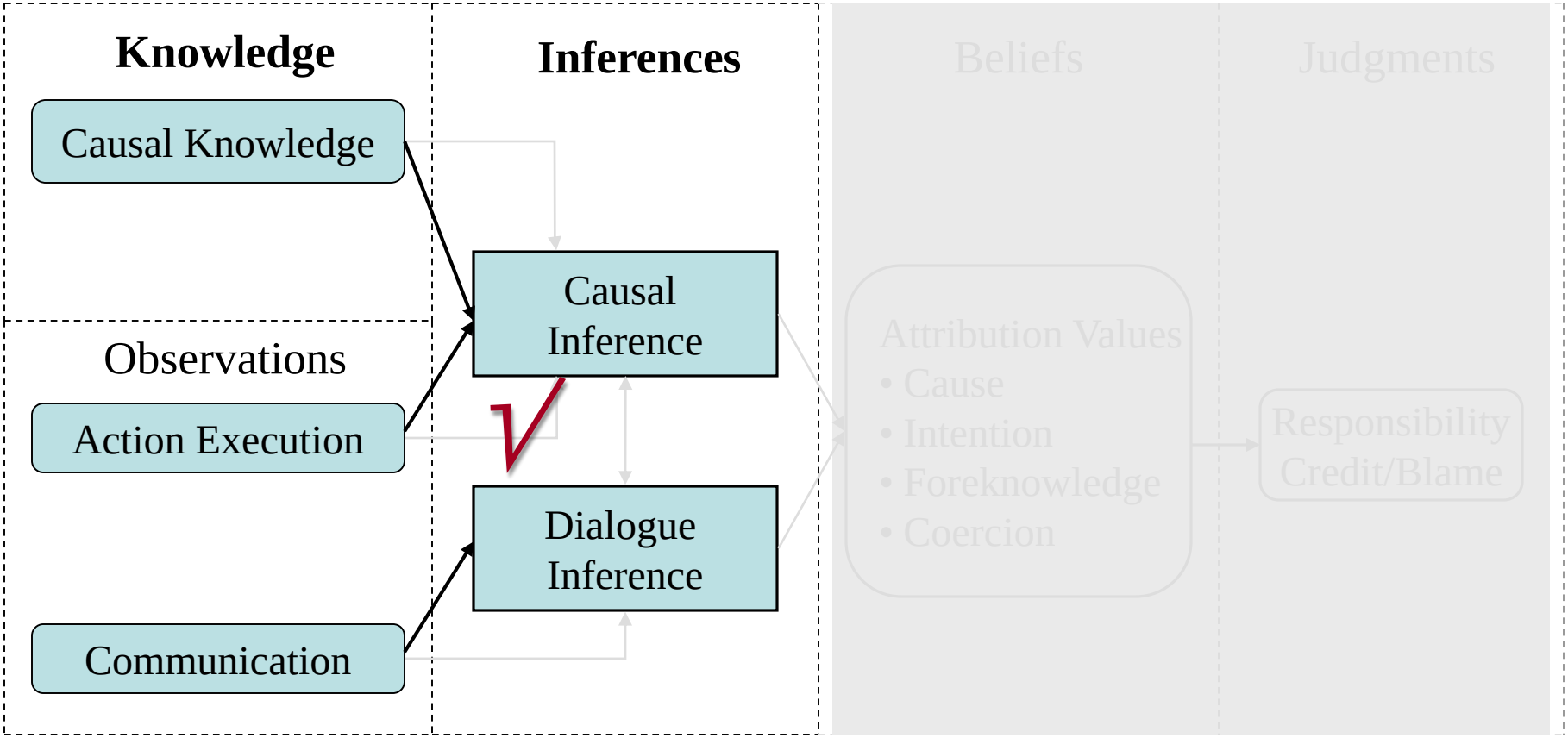
Scenario 1	Inform: *Foreknowledge* Order: *Act intention* *Intended goal* *Side effect* *Obligations*; *Act coercion*	Question 1: Rule D2 [inform-grounded]
		Question 2: Rule D5 [order]
		Question 3: Rule C7 [intend-plan]
		Question 4: Rule C8 [intend-plan]
		Question 5: Rule D6 [order] Rule D9 [accept-obligation]
Scenario 2	Request: *Willingness* Accept: *Act intention* *Outcome intention* *No foreknowledge* *Relation*: Intention/foreknowledge	Question 1: Rule D3 [request]
		Question 2: Rule D7 [accept]
		Question 3: Rule C3 [intend-action]
		Question 4: Rule D1 [inform]
		Question 5: Rule C9 [intent-foreknowledge-relation]
Scenario 3	Action knowledge: *Alternative* Counter-propose: *Unwillingness* Accept: *Act coercion* Primitive: *Outcome coercion*	Question 1: Rule D13 [counter-propose]
		Question 2: Rule D14 [counter-propose] Rule D15 [counter-propose]
		Question 3: Rule D6 [order] Rule D10 [unwilling-accept-obligation]
		Question 4: Rule C12 [coerce-primitive]
Scenario 4	Definite effect: *Outcome coercion* Alternative: *Act coercion* Indefinite effect: *Outcome coercion*	Question 1: Rule C16 [coerce-decision-node]
		Question 2: Rule C15 [coerce-decision-node]
		Question 3: Rule C17 [coerce-decision-node]

Accuracies of Inference Rules

- For each question, build confusion matrices to compute numbers of true/ false positive/negative between the conditions of the rule and the multi-evidence subjects choose.
- *Accuracy* of each rule ranges from *0.7* to *0.96*, with average ***0.85***.

Scenario 1	Question 1 / Rule D2	0.92
	Question 2 / Rule D5	0.96
	Question 3 / Rule C7	0.86
	Question 4 / Rule C8	0.70
	Question 5 / Rules D6&D9	0.84
Scenario 2	Question 1 / Rule D3	0.76
	Question 2 / Rule D7	0.96
	Question 3 / Rule C3	0.85
	Question 4 / Rule D1	0.94
	Question 5 / Rule C9	0.91
Scenario 3	Question 1 / Rule D13	0.94
	Question 2 / Rules D14&D15	0.88
	Question 3 / Rules D6&D10	0.80
	Question 4 / Rule C12	0.74
Scenario 4	Question 1 / Rule C16	0.71
	Question 2 / Rule C15	0.84
	Question 3 / Rule C17	0.75

Model Validation: Inference Process



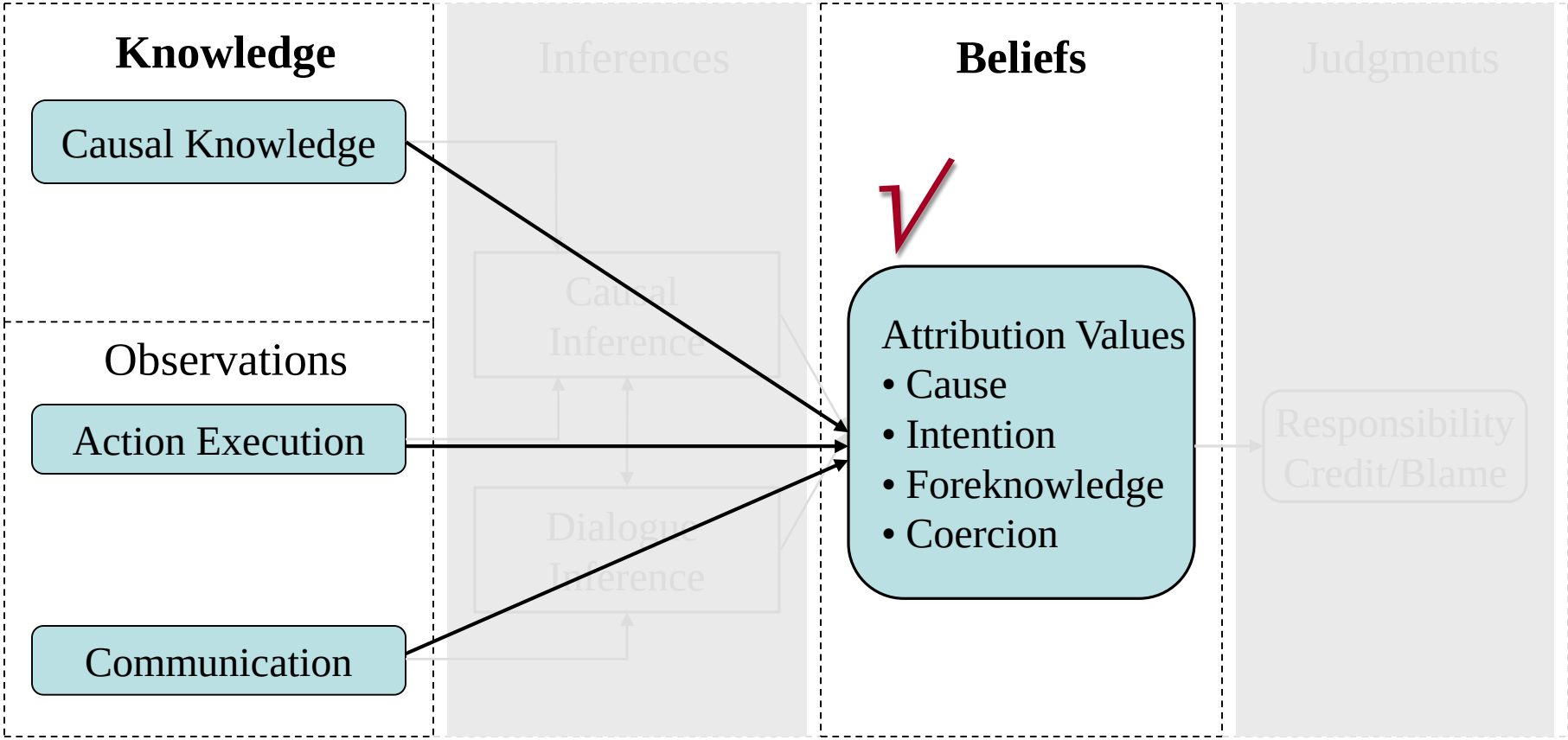
Results of the Inferred Beliefs

		Question 1		Question 2		Question 3		Question 4		Question 5		Last Question	
		Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Chair	VP
Scenario 1	Model	√		√		√			√	√		√	
	People	**30**	0	**30**	0	**30**	0	10	**20**	**22**	8	**5.63**	3.77
Scenario 2	Model	√		√		√			√		√		√
	People	**30**	0	**27**	3	**29**	1	2	**28**	0	**30**	3.00	**3.73**
Scenario 3	Model	√			√	√		√				√	
	People	**21**	9	2	**28**	**29**	1	**21**	9			**5.63**	3.23
Scenario 4	Model	√			√		√						√
	People	**21**	9	5	**25**	5	**25**					4.13	**5.20**

- Average *Kappa* agreement between model and subjects: ***0.732*** (substantial agreement)

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Model Validation: Internal Beliefs



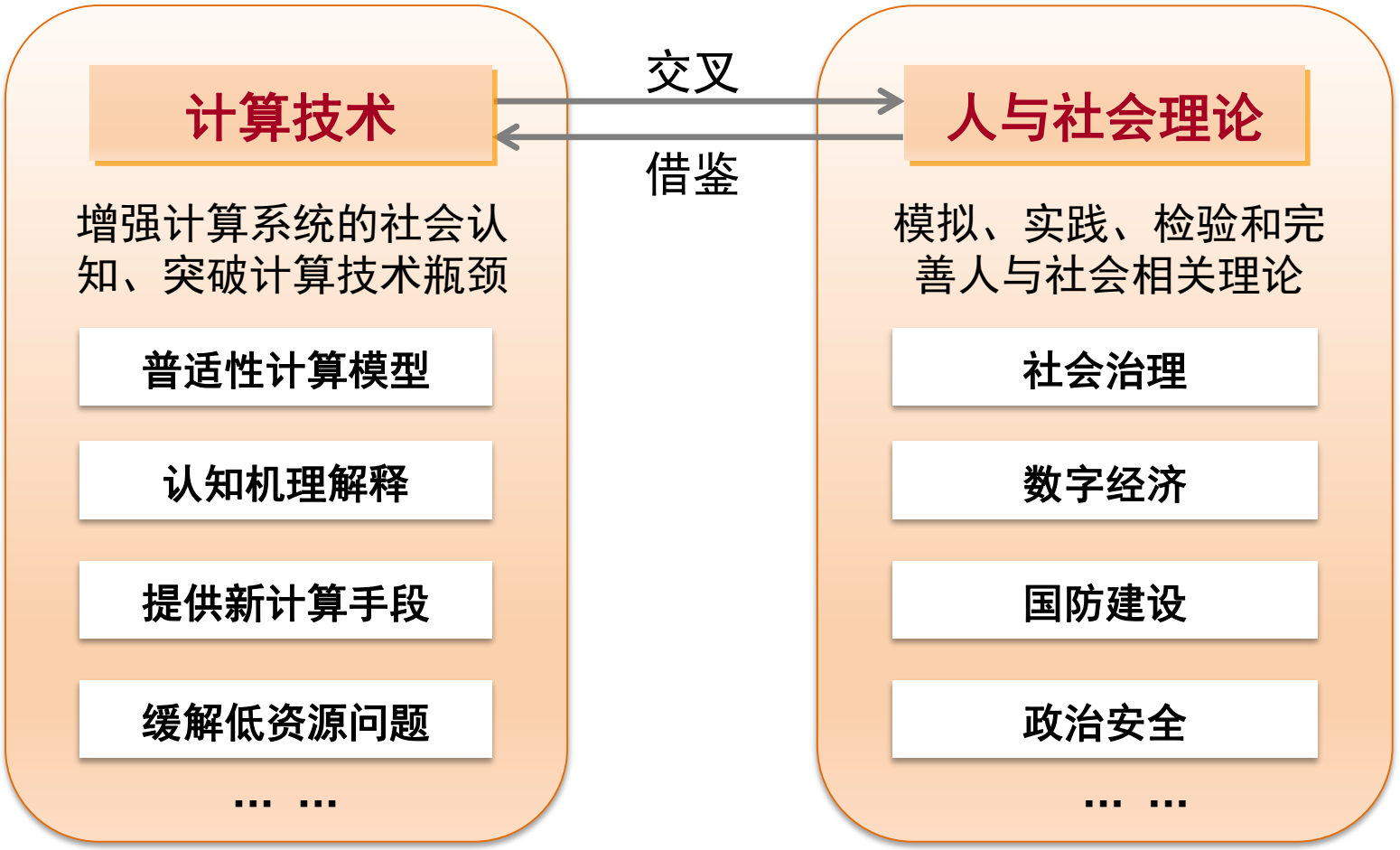
研究贡献

- **基于社会认知和心理学理论、领域适用的社会因果推理计算模型**
  - 首次建立基于心理学理论的社会因果推理计算模型
  - 提出基于言语交互和行为知识表示对归因因素进行信念推理的形式化方法
  - 设计归因算法、实现行为评判过程
  - 经过人的实验，验证了计算模型的有效性

总 结

- 面向心理认知和社会智能的普适性计算模型建立是多智能体系统研究与应用中十分重要的挑战性课题
- 基于理论模型的建模方法可以利用交叉学科的研究成果，增进对计算过程之外的认知和社会机理认识，来克服计算技术本身的局限性和瓶颈
- 基于认知和社会心理学理论的智能体认知与社会模拟可以有效增强计算/智能系统的认知和社会交互能力

展 望


计算技术
交叉
借鉴
人与社会理论
增强计算系统的社会认知、突破计算技术瓶颈
普适性计算模型
认知机理解释
提供新计算手段
缓解低资源问题
… …
模拟、实践、检验和完善人与社会相关理论
社会治理
数字经济
国防建设
政治安全
… …

References

1. A. Ortony, G. Clore and A. Collins. *The Cognitive Structure of Emotion*. Cambridge University Press, 1988.
2. J. Gratch and S. Marsella. A Domain-independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*, 5(4):269-306, 2004.
3. S. Marsella and J. Gratch. EMA: A Process Model of Appraisal Dynamics. *Journal of Cognitive Systems Research*, 10(1):70-90, 2009.
4. S. Marsella and J. Gratch. Computationally Modeling Human Emotion. *Communications of the ACM*, 57(12), 2014.
5. B. Weiner. *An Attributional Theory of Motivation and Emotion*. Springer, 1986.
6. B. Weiner. *Social Motivation, Justice and the Moral Emotions: An Attributional Approach*. Lawrence Erlbaum, 2006.

References

7. H. Chockler and J. Y. Halpern. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22:93-115, 2004.
8. J. Gratch, W. Mao and S. Marsella. Modeling Social Emotions and Social Attributions. In: R. Sun (Ed.), *Cognition and Multi-Agent Interaction: Extending Cognitive Modeling to Social Simulation*, pp.219-251. Cambridge University Press, 2006.
9. W. Mao and J. Gratch. Modeling Social Causality and Responsibility Judgment in Multi-Agent Interactions. *Journal of Artificial Intelligence Research*, 44:223-273, May, 2012.
10. J. Y. Halpern. *Actual Causality*. The MIT Press, 2016.

End.