Formal Models of Emotion

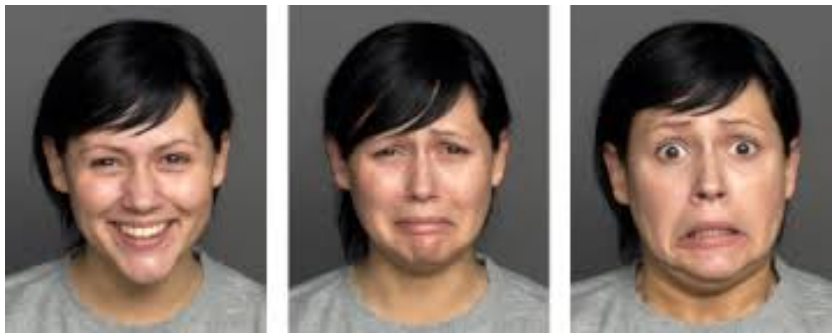
Mehdi Dastani

Utrecht University, The Netherlands

March 29, 2025

What is emotion?

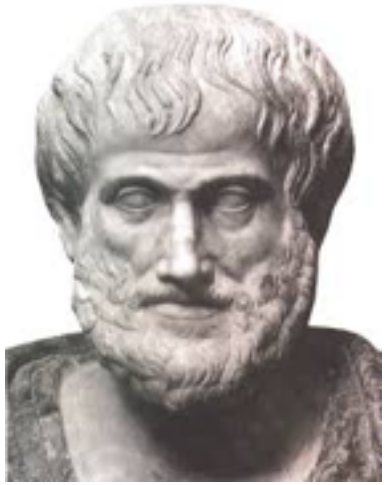
Everyone knows what an emotion is, until asked to give a definition. (Fehr & Russell 1984)



- What are emotions really?
- Surely they affect one's thinking and behaviour, but how?
- Are emotions all detrimental or do they serve some useful function?

`http://www.youtube.com/watch?v=A6Al6O9oATs`

Aristotle (*Rhetoric* and *Poetics*, 4th century BC)

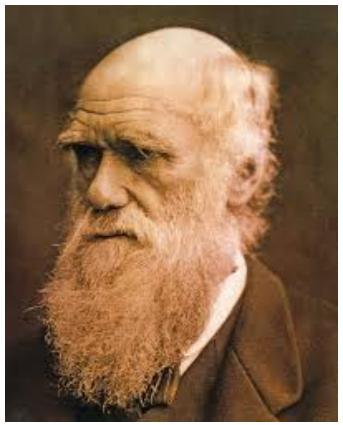


- Reason should rule our soul and monitor our emotional responses.
- Emotion elicitation can be used in audiences of public speaking (*Rhetoric*) and tragic drama (*Poetics*).
- Aristotle analyses several emotions in terms of
  - the beliefs they presuppose (e.g., anger requires the belief that oneself or one's friends are subject to wrongdoing),
  - their valence (e.g., anger is unpleasant),
  - their associated actions (e.g., anger gives an urge to take revenge), and
  - their cognitive effects (e.g., anger colours further judgments).

Descartes (*The Passions of the Soul*, 1649)

- Emotions are the passions of the soul, i.e., the things suffered by our thinking aspect.
- Descartes considers emotions as irrational and disruptive forces to reasoning and rationality.
- But, emotion was recognised as serving a useful function by directing ones thoughts and attentions to what is important and practical, and help to motivate decent behaviour and proper social life.

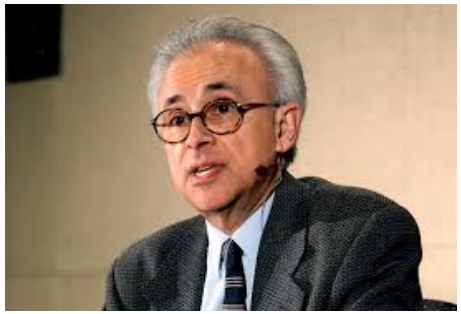
Darwin (*The Expression of the Emotions in Man and Animals*, 1872)



- Emotions are remnants of behaviours that were once useful in our evolutionary past.
- Darwin used emotions to show the continuity of adult human behaviour with the behaviour of infants and lower animals.
- But, Darwin recognised that certain emotions may facilitate social communication, his main idea was that emotions are involuntary and indicative of our primitive origins.

Damasio (*Descartes' Error: Emotion, Reason and the Human Brain*, 2005)

- When emotion is entirely left out of the reasoning picture, as happens in certain neurological conditions, reason turns out to be even more awed than when emotion plays bad tricks on our decisions.



- Damasio recognises that emotions can be detrimental to reasoning, but a life without emotion is a much worse fate.

- Emotions is a cognitive mechanism that directs/focuses attention to what is relevant and important.

Emotion vs. Rationality (Damasio 2005)

- Do people without emotions become “superrational”?
- This of course depends on what you regard as rational.
- Rationality: the ability to think deep and reason logical? The ability to maximise benefit?

“Doing a good job at succeeding in life” is quite rational too?

Human Concerns

- Humans have concerns but limited capabilities to meet them
- "Concerns" include goals, norms, preferences, interests, ideals, etc.
- Humans can never be certain whether concerns will be met
- But we can estimate how good/bad the situation looks like!

Emotions is a cognitive mechanism that directs/focuses attention to what is relevant and important.

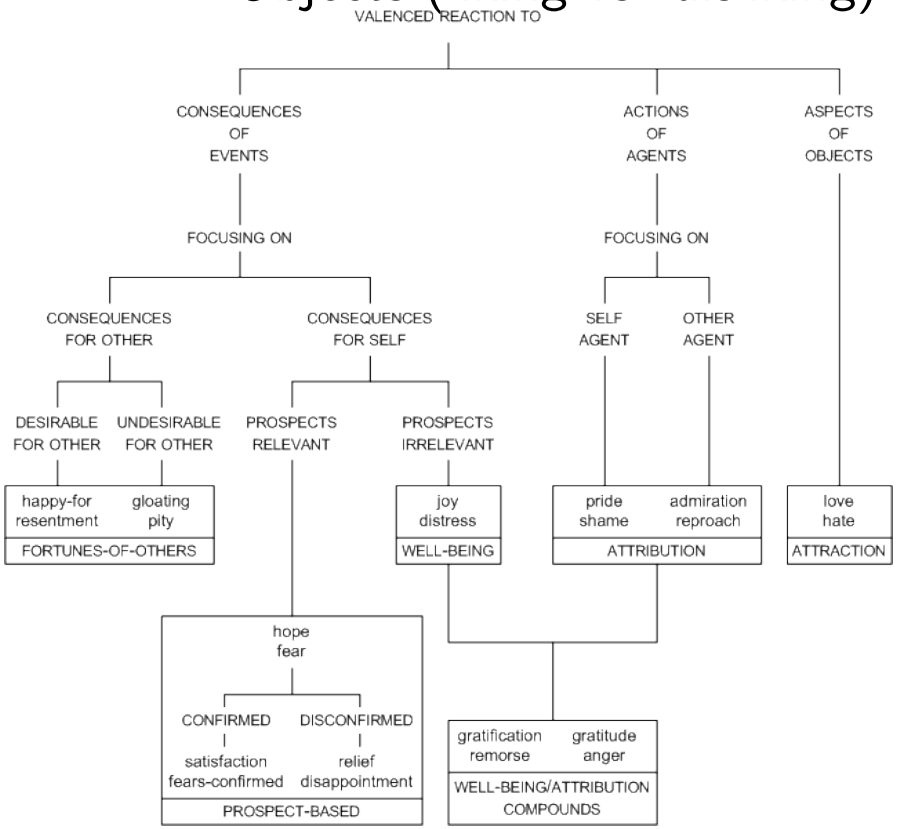
Affect (Scherer 1984)

- Different kinds of affective processes:
  - Impulses/feelings (e.g., hunger, lust, pain)
  - Emotions (e.g., happiness, sadness, anger, fear)
  - Mood (e.g., depression, euphoria)
- Note: no consensus on these distinctions.
- Here we will focus on "emotions" as distinguished above.

The Cognitive Structure of Emotions (OCC, 1998)

One can have a valenced reaction to:

- Consequences of events (pleased vs. displeased)
- Actions of agents (approving vs. disapproving)
- Objects (liking vs. disliking)



Fear Emotions

- TYPE SPECIFICATION: (displeased about) the prospect of an undesirable event
- TOKENS: apprehensive, anxious, cowering, dread, fear, fright, nervous, petrified, scared, terrified, timid, worried, etc.
- VARIABLES AFFECTING INTENSITY:
  1. the degree to which the event is undesirable
  2. the likelihood of the event
- EXAMPLE: The employee, suspecting he was no longer needed, feared that he would be fired.

Type Specifications

Joy: (pleased about) a desirable event

Distress: (displeased about) an undesirable event

Hope: (pleased about) the prospect of a desirable event

Fear: (displeased about) the prospect of an undesirable event

Pride: (approving of) one's own praiseworthy action

Shame: (disapproving of) one's own blameworthy action

Admiration: (approving of) someone else's praiseworthy action

Reproach: (disapproving of) someone else's blameworthy action

Happy-for: (pleased about) an event presumed to be desirable for someone else

Pity: (displeased about) an event presumed to be undesirable for someone else

Gloating: (pleased about) an event presumed to be undesirable for someone else

Resentment: (displeased about) an event presumed to be desirable for someone else

Logical Structure of Emotion (AAAI 2007)

$\mathcal{M} = \langle\langle \mathcal{W},\ \mathcal{R}_b,\ \mathcal{V}\rangle,\ \mathcal{R}_\alpha,\ \mathit{Aux},\ \mathit{Emo}\rangle$

- $\mathit{hope}_i(\varphi) \stackrel{def}{=} \mathit{desire}_i(\varphi) \wedge \mathit{prospective}_i(\varphi)$
- $\mathit{fear}_i(\varphi) \stackrel{def}{=} \mathit{undesire}_i(\varphi) \wedge \mathit{prospective}_i(\varphi)$
- $\mathit{joy}_i(\varphi) \stackrel{def}{=} \mathit{desire}_i(\varphi) \wedge \mathit{actual}_i(\varphi)$
- $\mathit{distress}_i(\varphi) \stackrel{def}{=} \mathit{undesire}_i(\varphi) \wedge \mathit{actual}_i(\varphi)$
- $\mathit{pride}_i(\alpha_j) \stackrel{def}{=} \mathit{approving}_i(\alpha_j) \wedge \mathit{cogUnit}_i(j)$
- $\mathit{admiration}_i(\alpha_j) \stackrel{def}{=} \mathit{approving}_i(\alpha_j) \wedge \neg \mathit{cogUnit}_i(j)$
- $\mathit{happy\_for}_{i,j}(\varphi) \stackrel{def}{=} \mathit{joy}_i(\varphi) \wedge \mathit{belief}_i\ \mathit{desire}_j(\varphi)$
- $\mathit{gloating}_{i,j}(\varphi) \stackrel{def}{=} \mathit{joy}_i(\varphi) \wedge \mathit{belief}_i\ \mathit{undesire}_j(\varphi)$

Where

$\mathit{prospective}_i(\varphi) \stackrel{def}{=} \mathit{new}\ \mathit{belief}_i(\neg\varphi \wedge \mathit{future}_i(\varphi))$

$\mathit{actual}_i(\varphi) \stackrel{def}{=} \mathit{new}\ \mathit{belief}_i(\varphi)$

Logical Structure of Emotion

Belief Updates can give rise to emotions

- $\models \mathit{goal}_i(\varphi) \wedge \psi \sqsubseteq \varphi \wedge \mathit{new\ belief}_i(\psi) \rightarrow \mathit{joy}_i(\psi)$
- $\models \mathit{goal}_i(\varphi) \wedge \psi \sqsubseteq \varphi \wedge \mathit{new\ belief}_i(\overline{\psi}) \rightarrow \mathit{distress}_i(\overline{\psi})$
- $\models \mathit{goal}_i(\varphi) \wedge \psi \sqsubseteq \varphi \wedge \mathit{new\ belief}_i(\neg\psi \wedge \mathit{future}_i\ \psi) \rightarrow \mathit{hope}_i(\psi)$
- $\models \mathit{goal}_i(\varphi) \wedge \psi \sqsubseteq \varphi \wedge \mathit{new\ belief}_i(\neg\overline{\psi} \wedge \mathit{future}_i\ \overline{\psi}) \rightarrow \mathit{fear}_i(\overline{\psi})$

Desirable consequence of an event can trigger joy, but a new desire for a known consequence does not trigger joy

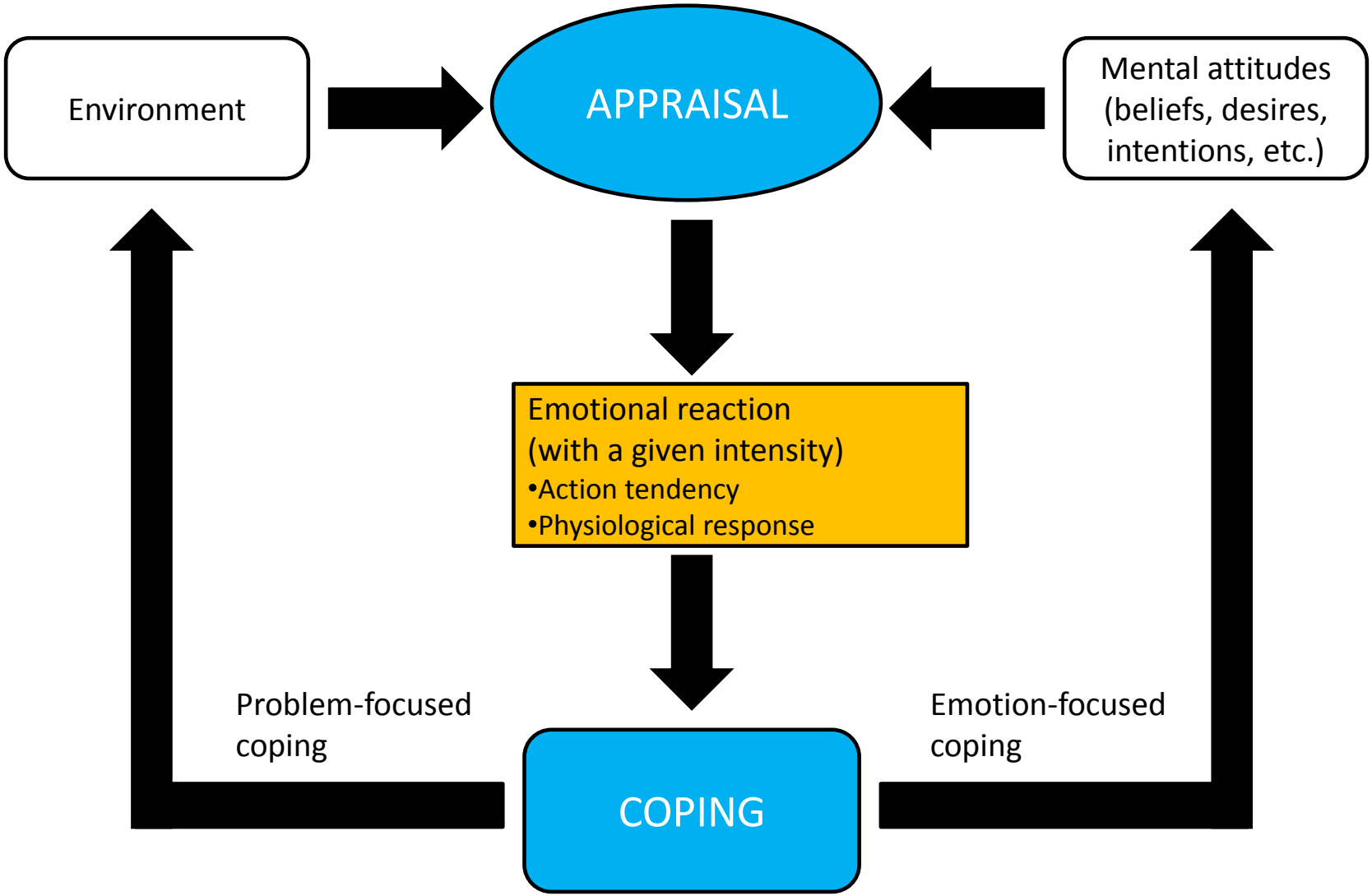
- $\models \mathit{new\ belief}_i\varphi \wedge \mathit{desire}_i(\varphi) \rightarrow \mathit{joy}_i(\varphi)$
- $\not\models \mathit{belief}_i\varphi \wedge \mathit{new\ desire}_i(\varphi) \rightarrow \mathit{joy}_i(\varphi)$

Relations between emotions; Intentions can give rise to emotion

- $\models \neg(\mathit{hope}_i(\varphi) \wedge \mathit{joy}_i(\varphi))$
- $\models \neg(\mathit{fear}_i(\varphi) \wedge \mathit{distress}_i(\varphi))$
- $\models \mathit{fear}_i(\varphi) \leftrightarrow \mathit{hope}_i(\varphi)$

Appraisal and coping circuit

(Lazarus, 1991; Gratch & Marsella, 2005)



Conclusion

- Emotion and reasoning are complementary.
- Emotion modelling using dynamic logic with graded beliefs and desires.
- Modeling Affective Reaction in Multi-agent Systems (van Kleef: *The interpersonal dynamics of emotion*).
- Emotion modelling requires modelling of complex phenomena like control, accountability/responsibility.

Appraisal and coping circuit (AAMAS 2012)

Example (Fear)

A robot has to transport some containers to their target positions. The robot can assess the state of its battery charge.

- **Appraisal**. The robot appraises the current situation: *low battery charge endangers the goal of having a container at its target position*.
- **Emotional experience**. Emotion is triggered: *fear of failing to place a container at its target position*.
- **Coping**. The triggered emotion affects the robot's behavior depending on its intensity: *fear leads the robot to reconsider its current intention to transport a container*.

Appraisal and coping circuit (ACM Transactions on Internet Technology)

Example (Anger)

A robot R_1 aims at picking up a container at a designated position.

- **Appraisal** (R_1 appraises the current situation)
  R_1 *believes that the container is removed and that robot* R_2 *is accountable for the removal*.
- **Emotional experience** (Emotion is triggered)
  R_1 *is angry because of not being able to pick up the container*.
- **Coping** (The triggered emotion affects R_1's behavior)
  anger leads R_1 *requests* R_2 *to return the container*.

Appraisal and coping circuit (PRIMA 2017)

Example (Social/Moral Anger)

A robot R_1 aims at picking up a container at a designated position. There is a manager agent M that coordinate the transportation.

- **Appraisal** (M appraises the current situation)
  M believes that the container of R_1 is removed, robot R_2 is accountable for the removal, and cooperation should be promoted.
- **Emotional experience** (Emotion is triggered)
  M is angry because R_2 frustrates R_1's goal and therefore cooperation is not promoted.
- **Coping** (The triggered emotion affects the M's behavior)
  anger leads M to request R_2 to return the container to R_1.

Appraisal and coping circuit (PRIMA 2017)

(Social/Moral) Anger

Displeasure from thwarting of a personal goal, or a social rule aimed at preserving the goal commitment of other agents, combined with attribution of **blame** for the goal-thwarting state of affairs to another agent, and an estimate of one's own coping potential as favouring *attack towards* the blameworthy agent.

Blame requires **Accountability** and **Control** (Lazarus)

Agent A is **blameworthy** of a state of affair φ iff A is **accountable** for φ and **could act differently**.

Dynamic Multi-Agent Logic of Graded Attitudes

The logical framework is built on previous work on Dastani & Lorini AAMAS 2012.

Definition (Language)

- $a \in Act = \{toggle(p) : p \in Atm\}$
- $e \in Evt_i = \{(i, a) : a \in Act\}$
- $Evt = \bigcup_{i \in Agt} Evt_i$
- $\epsilon \in Evt^*$

$$\begin{aligned} \varphi, \psi \quad ::= \quad & p \mid \neg\varphi \mid \varphi \wedge \psi \mid \\ & exc_i^h \mid Des_i^k I \mid K_i\varphi \mid Int_i(\epsilon, a) \mid \\ & Fut(\epsilon, e) \mid Past(\epsilon, e) \mid [e]\varphi \end{aligned}$$

Dynamic Multi-Agent Logic of Graded Attitudes

Definition (Language)

$$\begin{aligned} \varphi, \psi \quad ::= \quad & p \mid \neg\varphi \mid \varphi \wedge \psi \mid \\ & exc_i^h \mid Des_i^k I \mid K_i\varphi \mid Int_i(\epsilon, a) \mid \\ & Fut(\epsilon, e) \mid Past(\epsilon, e) \mid [e]\varphi \end{aligned}$$

- exc_i^h: the current world has a degree of exceptionality h for agent i
- $Des_i^k I$: the state of affairs I has a degree of desirability k for agent i
- $K_i\varphi$: agent i knows that φ is true
- $Int_i(\epsilon, a)$: after the sequence of events ϵ, agent i intends to perform action a
- $Fut(\epsilon, e)$: event e can possibly occur in the state reached by the sequence of events ϵ
- $Past(\epsilon, e)$: event e has just occurred in the state reached after the sequence of events ϵ
- $[e]\varphi$: the occurrence of event e leads to φ

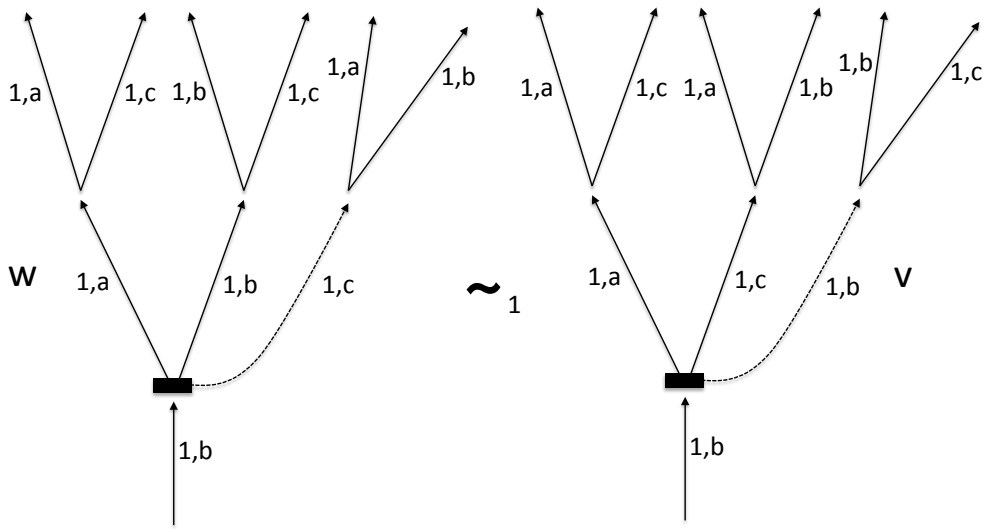
Dynamic Multi-Agent Logic of Graded Attitudes

Definition (Semantics)

$$\mathfrak{M} = \Big(W, (\sim_i)_{i\in Agt}, (\mathcal{E}_i)_{i\in Agt}, (\mathcal{D}_i)_{i\in Agt}, (\mathcal{I}_i)_{i\in Agt}, \mathcal{F}, \mathcal{P}, \mathcal{V}\Big)$$

- W is a nonempty set of worlds or states;
- $\sim_i \subseteq W \times W$ is an equivalence relation representing knowledge
- $\mathcal{E}_i : W \longrightarrow Num^+$ is a total function representing exceptionality degrees of states
- $\mathcal{D}_i : W \times Lit \longrightarrow Num$ is a total function representing desirability of facts
- $\mathcal{I}_i : W \times SeqEvt \longrightarrow 2^{Act}$ is a total function representing agents' intentions
- $\mathcal{F} : W \times SeqEvt \longrightarrow 2^{Evt}$ is a total function indicating future events;
- $\mathcal{P} : W \times SeqEvt \longrightarrow Evt$ is a partial function indicating past events;
- $\mathcal{V} : W \longrightarrow 2^{Atm}$ is a valuation function

Dynamic Multi-Agent Logic of Graded Attitudes



- $\mathcal{F}(w, nil) = \{(1, a), (1, b)\}$
- $\mathcal{P}(w, nil) = \{(1, b)\}$
- $\mathcal{F}(w, (1, b)) = \{(1, b), (1, c)\}$
- $\mathcal{F}(w, (1, c)) = \{(1, a), (1, b)\}$

Definition (Truth Conditions of formulae)

$$\mathfrak{M} = \Big(W, (\sim_i)_{i\in Agt}, (\mathcal{E}_i)_{i\in Agt}(\mathcal{D}_i)_{i\in Agt}, (\mathcal{I}_i)_{i\in Agt}, \mathcal{F}, \mathcal{P}, \mathcal{V}\Big)$$

- $\mathfrak{M}, w \models p$ *iff* $p \in \mathcal{V}(w)$;
- $\mathfrak{M}, w \models Des_i^k\, I$ *iff* $\mathcal{D}_i(w, I) = k$;
- $\mathfrak{M}, w \models exc_i^h$ *iff* $\mathcal{E}_i(w) = h$;
- $\mathfrak{M}, w \models Int_i(\epsilon, a)$ *iff* $a \in \mathcal{I}_i(w, \epsilon)$;
- $\mathfrak{M}, w \models Fut(\epsilon, e)$ *iff* $e \in \mathcal{F}(w, \epsilon)$;
- $\mathfrak{M}, w \models Past(\epsilon, e)$ *iff* $\mathcal{P}(w, \epsilon) = e$;
- $\mathfrak{M}, w \models \neg\varphi$ *iff* not $\mathfrak{M}, w \models \varphi$;
- $\mathfrak{M}, w \models \varphi \wedge \psi$ *iff* $\mathfrak{M}, w \models \varphi$ and $\mathfrak{M}, w \models \psi$;
- $\mathfrak{M}, w \models K_i\varphi$ *iff* $\mathfrak{M}, v \models \varphi$ for all $v \in W$ s.t. $v \sim_i w$;
- $\mathfrak{M}, w \models [e]\varphi$ *iff* $\mathfrak{M}^e, w \models \varphi$
  where $\mathfrak{M}^e$ is $\mathfrak{M}$ updated with event e.

Formalizing Controlability and Accountability

- $\langle\langle e \rangle\rangle \varphi \stackrel{def}{=} Fut(nil, e) \wedge [e]\varphi$
  event e possibly occur next and φ will be true afterwards.

- $\langle\langle -e \rangle\rangle \varphi \stackrel{def}{=} Past(nil, e) \wedge [-e]\varphi$
  event e has just occurred and φ was true before.

- Agent i has control over the state of affairs φ iff i is able to maintain its truth value, i.e.,

$$Control_i(\varphi) \stackrel{def}{=} (\varphi \wedge \bigvee_{e \in Evt_i} \langle\langle e \rangle\rangle \varphi) \vee (\neg\varphi \wedge \bigvee_{e \in Evt_i} \langle\langle e \rangle\rangle \neg\varphi)$$

  E.g., $Control_{R_2}(XatY)$

Formalizing Controlability and Accountability

- $\langle\langle e \rangle\rangle \varphi \stackrel{def}{=} Fut(nil, e) \wedge [e]\varphi$
  event e possibly occur next and φ will be true afterwards.

- $\langle\langle -e \rangle\rangle \varphi \stackrel{def}{=} Past(nil, e) \wedge [-e]\varphi$
  event e has just occurred and φ was true before.

- Agent i is accountable for φ by doing a iff φ is true now and was not true before event (i, a) occurred, i.e.,

$$Account_i(a, \varphi) \stackrel{def}{=} \varphi \wedge \langle\langle -(i, a) \rangle\rangle \neg\varphi$$

  E.g., $Account_{R_2}(pickXatY, \neg XatY)$

Formalizing Blame and Practical Possibility

- Agent i blames agent j for doing a and causing φ iff i believes that j is accountable for φ by doing a, and that before the event (j, a), j had control over φ, i.e.,

$$Blame^{k}_{i,j}(a, \varphi) \stackrel{def}{=} B^{k}_{i}(Account_{j}(a, \varphi) \wedge [-(j, a)]Control_{j}(\varphi))$$

E.g., $Blame^{k}_{R_1,R_2}(pickXatY, XatY)$, for some $k > 0$

- There is a practical possibility for agent i to make φ true, i.e.,

$$Pos_{i}(\varphi) \stackrel{def}{=} \bigvee_{e \in Evt_{i}} \langle\langle e \rangle\rangle \varphi$$

E.g., $Pos_{R_1}(R_1 holdsX)$

Formalizing Anger

Agent i is angry with intensity l at agent j for doing a and preventing i from achieving φ by doing b, i.e.,

$$Anger^{l}_{i,j}(a, \varphi, b) \stackrel{def}{=}$$

$$\bigvee_{l=merge(h,k)} (AchG^{k}_{i}(\varphi) \wedge Int_i\, b \wedge Blame^{h}_{i,j}(a, \neg \langle\langle (i, b) \rangle\rangle \varphi) \wedge B_i Pos_i(\varphi))$$

E.g., $Anger^{l}_{R_1,R_2}(pickXatY, R_1\, holdsX, pickXatY)$, for some $l > 0$

Validities

- After the occurrence of an event an agent is accountable for a state of affairs iff the state of affair does currently not hold, the state of affair is the case after the event, and the event creates the history.

$$\models [(i,a)]Account_i(a,\varphi) \leftrightarrow \neg\varphi \wedge [(i,a)]\varphi \wedge [(i,a)]Past(nil,(i,a))$$

- Blame requires choices in the direct past.

$$\models Blame^k_{i,j}(a,\varphi) \rightarrow B^k_i([-(j,a)] \bigvee_{e \in Evt_j} [e]\neg\varphi)$$

- No blame for unavoidable.

$$\models B^k_i(\varphi \wedge \bigwedge_{e \in Evt_j} [e]\neg\varphi) \rightarrow [(j,b)]\neg Blame^k_{i,j}(b,\neg\varphi) \quad \text{for} \quad (j,b) \in Evt_j$$

Validities

- No blame for trivialities and impossibilities.
  - $\models \neg Blame^{k}_{i,j}(a, \top)$
  - $\models \neg Blame^{k}_{i,j}(a, \bot)$

- Decomposition of accountability.
  - $\models Account_i(a, \neg\varphi) \leftrightarrow \neg Account_i(a, \varphi)$
  - $\models Account_i(a, \varphi \vee \psi) \leftrightarrow Account_i(a, \varphi) \vee Account_i(a, \psi)$
  - $\models Account_i(a, \varphi \wedge \psi) \rightarrow Account_i(a, \varphi) \vee Account_i(a, \psi)$
  - $\models Account_i(a, \varphi) \wedge Account_i(a, \psi) \rightarrow Account_i(a, \varphi \wedge \psi)$

- No anger at those who are not accountable for your disability or desired outcome of your choice.
  $\models$
  $(\neg B^{k}_{i} Account_j(a, \neg Fut(nil, (i, b))) \ \wedge \neg B^{k}_{i} Account_j(a, \neg[(i, b)]\varphi)) \rightarrow \neg Anger^{l}_{i,j}(a, \varphi, b)$
  for $l = merge(h, k)$ and $h \in Num^{+}$

Conclusion

- Emotion and reasoning are complementary.
- All ingredients required for emotion modelling: multi-agent dynamic logic with graded beliefs and desires.
- Emotion Stages: Appraisal, Experience, and Coping.
- Emotion modelling requires modelling of complex phenomena like control, accountability/responsibility.
- We need more complex actions and accountability for actions performed in arbitrary past.