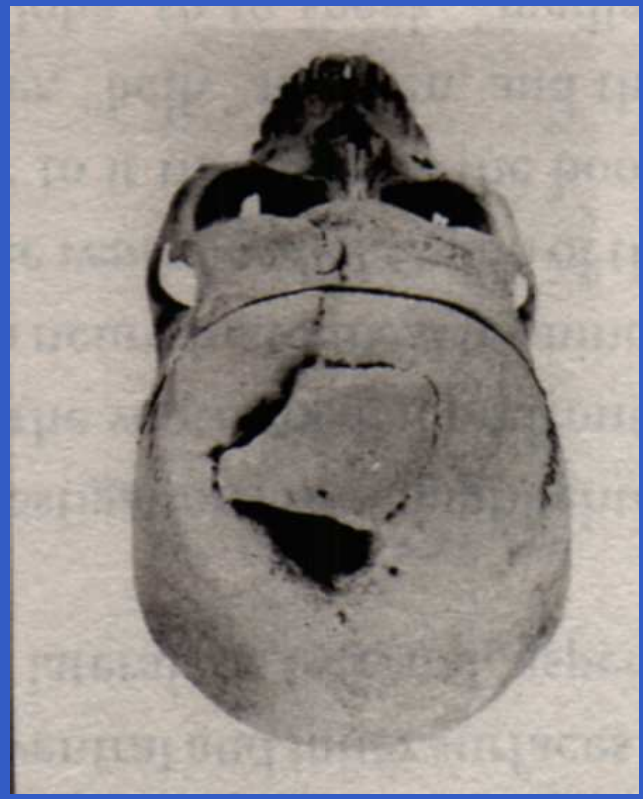
Emotion in Intelligent Agents

Peter-Paul van Maanen

November 27th, 2003

ppmaanen@cs.uu.nl

CNR (Rome) & UU (Utrecht)



Overview

- **1. Motivation**
- EMOTION THEORY
  - **2. What is emotion?**
  - **3. Why should machines experience emotion?**
  - **4. (How) can machines experience emotion?**
- TOWARD AI MODELS OF EMOTION
  - **5. The antecedents of emotion**
  - **6. The behaviors of emotion**
- **7. Conclusion**

1. Motivation

- Overcome: Emotion would interfere rationality

1. Motivation

- Overcome: Emotion would interfere rationality
- Evolution of the brain suggests differently

1. Motivation

- Overcome: Emotion would interfere rationality
- Evolution of the brain suggests differently
- Neurological research too

1. Motivation

- Overcome: Emotion would interfere rationality
- Evolution of the brain suggests differently
- Neurological research too

★ Doyle (1991):
"AI is the discipline aimed at understanding intelligent beings by constructing intelligent systems".

1. Motivation

- Overcome: Emotion would interfere rationality
- Evolution of the brain suggests differently
- Neurological research too

★ Doyle (1991):
"AI is the discipline aimed at understanding intelligent beings by constructing intelligent systems".

⇒ So construct emotionally intelligent systems

EMOTION THEORY

2. What is emotion?

- Unclear: as many as 92 *different* definitions in literature (Kleinginna & Kleinginna 1981)

2. What is emotion?

- Unclear: as many as 92 *different* definitions in literature (Kleinginna & Kleinginna 1981)

★ LeDoux (1994):

"Emotions are [. . .] conscious states".

2. What is emotion?

- Unclear: as many as 92 *different* definitions in literature (Kleinginna & Kleinginna 1981)

★ LeDoux (1994):
"Emotions are [. . .] conscious states".

★ Damasio (1994):
"[. . .] designate a collection of responses triggered from parts of the brain to the body, and from parts of the brain to other parts of the brain".

2. What is emotion?

- Unclear: as many as 92 *different* definitions in literature (Kleinginna & Kleinginna 1981)

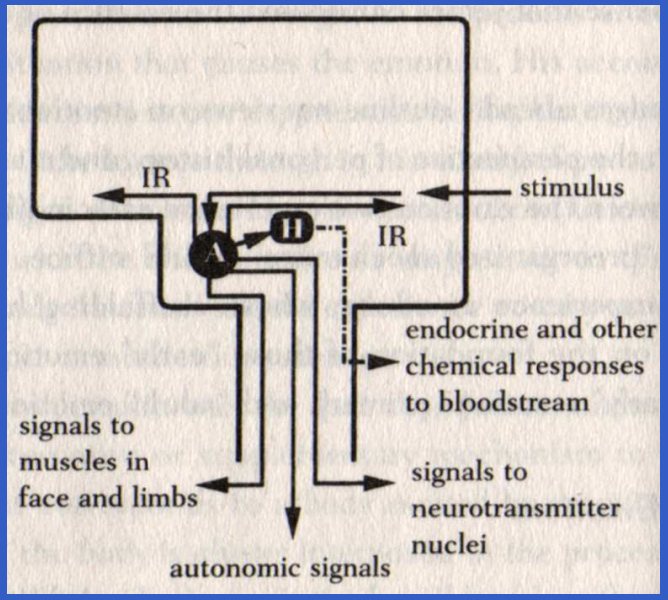
★ LeDoux (1994):
"Emotions are [. . .] conscious states".

★ Damasio (1994):
"[. . .] designate a collection of responses triggered from parts of the brain to the body, and from parts of the brain to other parts of the brain".

⇒ What LeDoux calls emotion is what Damasio calls feeling.

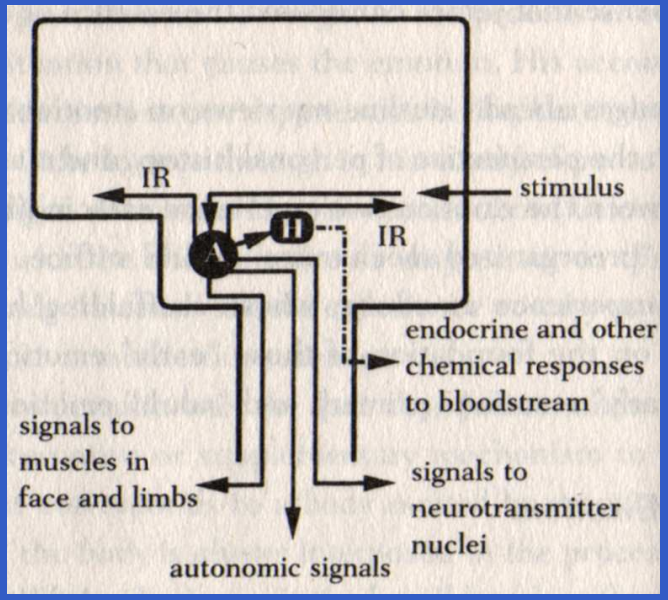
Two kinds of emotion: Primary

- PRIMARY EMOTIONS are innate emotional reactions to bodily stimuli



Two kinds of emotion: Primary

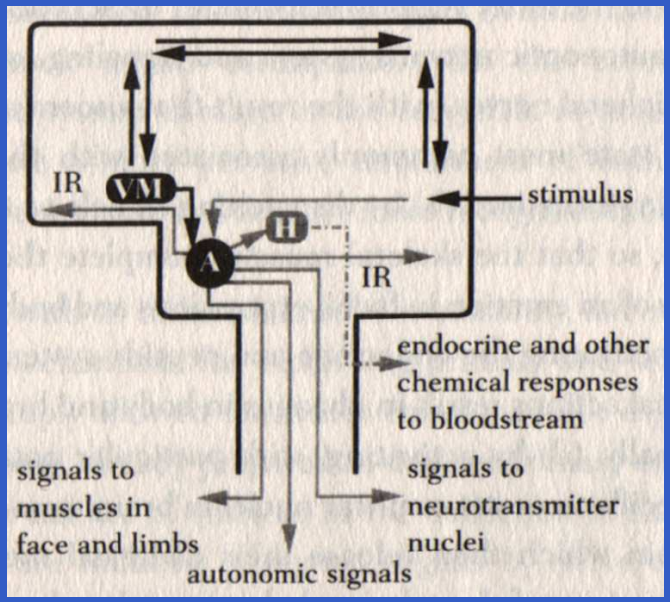
- PRIMARY EMOTIONS are innate emotional reactions to bodily stimuli



- Sometimes acquired, but mostly pre-wired

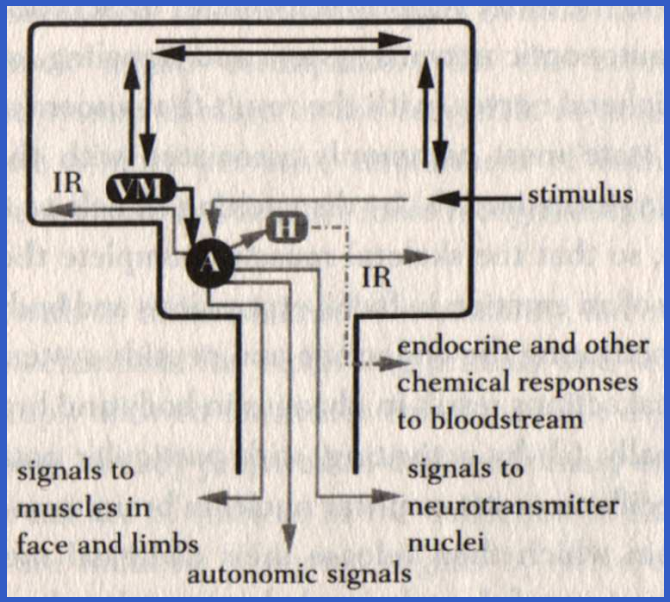
Two kinds of emotion: Secondary

- SECONDARY EMOTIONS are subconsciously created connections between cognitive constructs and primary emotions



Two kinds of emotion: Secondary

- SECONDARY EMOTIONS are subconsciously created connections between cognitive constructs and primary emotions
- Probably always acquired



Feeling

- FEELING is becoming conscious of what is happening in the body

Feeling

- FEELING is becoming conscious of what is happening in the body

? Body-based, why don't emotions trigger feeling directly?

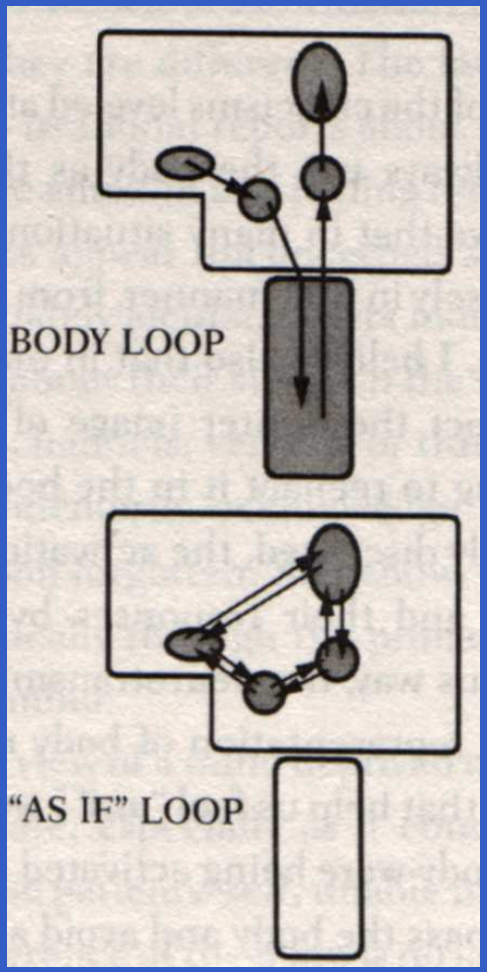
Feeling

- FEELING is becoming conscious of what is happening in the body

? Body-based, why don't emotions trigger feeling directly?

$\Rightarrow$ Emotions felt like this constitute the 'AS IF' LOOP

Feeling: body vs ‘as if’

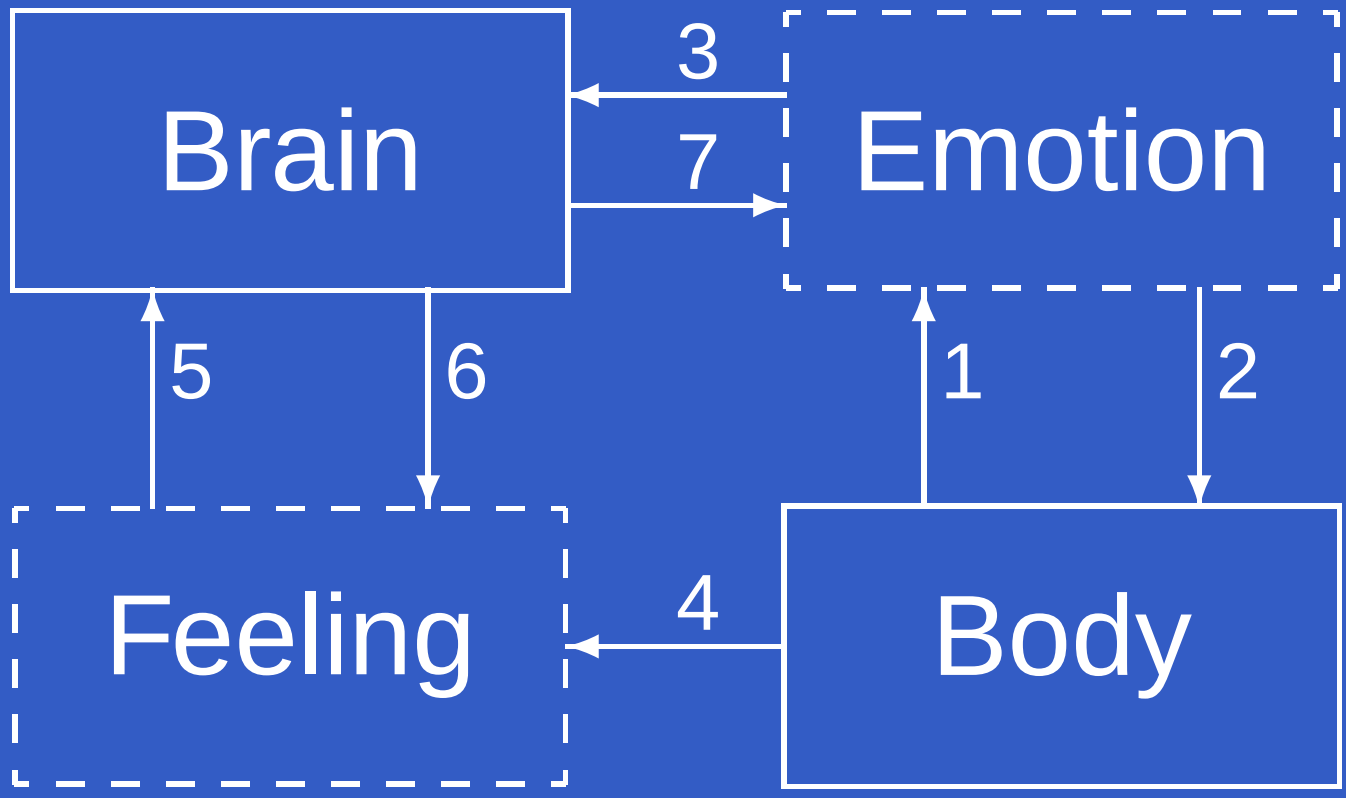


Definitions for agents

DEFINITION 1 (EMOTION). Emotion is a collection of physical or mental UNCONSCIOUS responses UNCONSCIOUSLY triggered by the mind.

DEFINITION 2 (FEELING). Feeling is a collection of mental CONSCIOUS responses UNCONSCIOUSLY triggered by the body or the mind.

Definitions for agents: Diagram



3. Why should machines experience emotion?

? Emotion important for natural agents, why for artificial agents?

3. Why should machines experience emotion?

? Emotion important for natural agents, why for artificial agents?

$\Rightarrow$ Maybe there are problems that can only be solved by emotion-like processes

Why should machines experience emotion?: Examples

- Reality is very detailed and constitutes many properties

Why should machines experience emotion?: Examples

- Reality is very detailed and constitutes many properties

Problem: How does an agent image the relevant part of reality?

Why should machines experience emotion?: Examples

- Reality is very detailed and constitutes many properties

Problem: How does an agent image the relevant part of reality?

Example: Recalling this room after we went to the corridor

Why should machines experience emotion?: Examples

- Reality is very detailed and constitutes many properties

Problem: How does an agent image the relevant part of reality?

Example: Recalling this room after we went to the corridor

Problem: How does an agent reason with limited information of reality?

Why should machines experience emotion?: Examples

- Reality is very detailed and constitutes many properties

Problem: How does an agent image the relevant part of reality?

Example: Recalling this room after we went to the corridor

Problem: How does an agent reason with limited information of reality?

Example: Do you want to marry me?

Why should machines experience emotion?: Examples

- Say $\varphi \equiv$‘marry Peter-Paul’, $\psi \equiv$‘like my life’, and a goal-generation rule (Castelfranchi et al. 2000):

$$G_i\psi \wedge B_i(\mathsf{GOODFOR}_i\ \varphi\ \psi) \Rightarrow G_i\varphi$$

Why should machines experience emotion?: Examples

- Say $\varphi \equiv$‘marry Peter-Paul’, $\psi \equiv$‘like my life’, and a goal-generation rule (Castelfranchi et al. 2000):

$$G_i\psi \wedge B_i(\mathsf{GOODFOR}_i\ \varphi\ \psi) \Rightarrow G_i\varphi$$

- We should add consistency:

$$G_i\chi \rightarrow \neg B_i(\mathsf{BADFOR}_i\ (\varphi \wedge \chi)\ \psi)$$

where we assume that $(G\varphi \wedge G\psi) \rightarrow G(\varphi \wedge \psi)$.

Why should machines experience emotion?: Examples

- We can now specify the operator GOODFOR$_i$ φ ψ:

$$(B_i\varphi \rightarrow \Diamond^n B_i\psi) \wedge (\Diamond^m B_i\psi) \wedge n > m$$

where $\Diamond^p\chi$ means that eventually χ is true with probability p.

Why should machines experience emotion?: Examples

- We can now specify the operator GOODFOR$_i$ φ ψ:

$$(B_i\varphi \rightarrow \Diamond^n B_i\psi) \wedge (\Diamond^m B_i\psi) \wedge n > m$$

  where $\Diamond^p\chi$ means that eventually χ is true with probability p.

- We should add the following rule to select the best:

$$(B_i\xi \rightarrow \Diamond^k B_i\psi) \rightarrow k \leq n$$

Why should machines experience emotion?: Examples

- We can now specify the operator GOODFOR$_i$ φ ψ:

$$(B_i\varphi \rightarrow \Diamond^n B_i\psi) \wedge (\Diamond^m B_i\psi) \wedge n > m$$

  where $\Diamond^p\chi$ means that eventually χ is true with probability p.

- We should add the following rule to select the best:

$$(B_i\xi \rightarrow \Diamond^k B_i\psi) \rightarrow k \leq n$$

$\Rightarrow$ To much effort!

Why should machines experience emotion?: Examples

$\Rightarrow$ Hence emotional appraisals would be advantageous if they would:

1. **drastically reduce the amount of possible goals to evaluate for acceptance, by means of precognitive filtering**
2. **drastically shorten the cognitive chain in evaluations, by means of providing a threshold when to stop**

Why should machines experience emotion?: Examples

$\Rightarrow$ Hence emotional appraisals would be advantageous if they would:

1. **drastically reduce the amount of possible goals to evaluate for acceptance, by means of precognitive filtering**
2. **drastically shorten the cognitive chain in evaluations, by means of providing a threshold when to stop**

$\Rightarrow$ The same can be said of DERIVABLEFROM φ ψ

Advantages of the emotional way

⇒ Body and mind are pre-cognitively tuned to cope with the current situation either by *interrupt* or *heuristic* with limited recources and time:

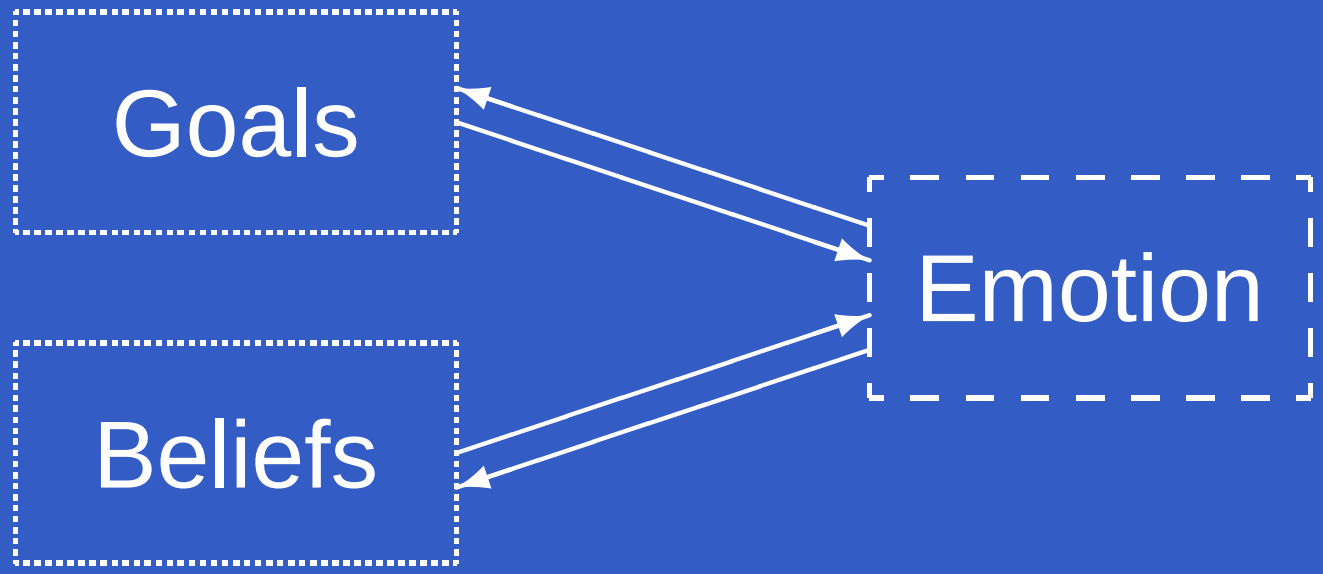
Advantages of the emotional way

⇒ Body and mind are pre-cognitively tuned to cope with the current situation either by *interrupt* or *heuristic* with limited recources and time:

1. **Introduces new urgent goals or alter the priorities of existing goals**
2. **Attention is demanded to important or relevant information**
3. **Keeps track of the higher goals, i.e. desires**

Advantages of the emotional way: Diagram

$\Rightarrow$ Emotional processes thus idealy relate to artificial reasoning in terms of beliefs and goals:



4. (How) can machines experience emotion?

? Would a machine built like us *experience* emotion like us?

4. (How) can machines experience emotion?

? Would a machine built like us *experience* emotion like us?

★ First step:

"Why don't we just build an embodied agent, able to have emotions and to react to those? This is not so hard" (Castelfranchi, 2003).

4. (How) can machines experience emotion?

? Would a machine built like us *experience* emotion like us?

★ First step:

"Why don't we just build an embodied agent, able to have emotions and to react to those? This is not so hard" (Castelfranchi, 2003).

⇒ For this we can propose agent-architectures, describing both having emotion and feeling.

Proposed agent-architecture requirements

1. **Rich enough body signals (primary input)**
2. **Rich enough mental constructs (secondary input)**
3. **`if-then`-like rules that identify typical structures (emotion/feeling triggers)**
4. **These `if-then`-like rules need to be learned (emotional adaptivity)**
5. **Check each rule on the agent's input in each deliberation cycle**
6. **If rules match then resulting typical behavior**

TOWARD AI MODELS OF EMOTION

4. The antecedents of emotion

$\Rightarrow$ To describe emotion artificially we need to look at it from the engineering point of view:

4. The antecedents of emotion

$\Rightarrow$ To describe emotion artificially we need to look at it from the engineering point of view:

- Research is needed in the area of the development of programs for machines that have rules like:

IF antecedent $\varphi_0, \ldots, \varphi_n$ of emotion E hold
THEN (at a later timepoint) trigger behavior of $E(\varphi_0, \ldots, \varphi_n)$
AND learn to connect $\varphi_0, \ldots, \varphi_n$ to E.

The antecedents of emotion: Example

★ Frijda (1986):

> “Fear [. . .] is uncertain expectation of the presence of negative valence, or absence of positive valence, over which there is insuffi cient control, but which event is modifi able, its degree corresponds to the measure of closure, and urgency of the situation”.

4. The antecedents of emotion

⇒ Emotions can thus be uniquely cathegorized by means of *component profiles*:

	Positive character	*Negative character*	*Desire*	*Interest*	*Positive valence*	*Negative valence*	*Presence*	*Absence*	*Certainty*	*Uncertainty*	*Change*	*Open*	*Closed*	*Intentionality of other*	*Intentionality of self*	*Controllability*	*Uncontrollability*	*Modifiability*	*Finality*	*Object*	*Event*	*Focality*	*Globality*	*Strangeness*	*Familiarity*	*Value*
Joy	■				▲	▼	▲	▼			(■)	■						■			■	■				
Fear		■			▲	▼	▼	▲		■			■				■	■			■	■				
Hope	■				▼	▲	▼	▲		■		■									■	■				
Anger		■			▼	▲	▲	▼						■		■					■	■				

Profiling: Formalizing antecedents

If the agent evaluates GOODFOR(φ, ψ) for a situation φ and active goal ψ, then positive demand characer holds for φ, and
if BADFOR(φ, ψ), then negative demand character holds for φ.

If the agent believes the eliciting situation φ from moment t, then change of φ holds if t is close to the current moment.

Profiling: Eliciting rule of *fear*

IF antecedent
$\texttt{negative\_demand\_character}(\varphi, \psi)$
AND $\texttt{uncertainty}(\varphi)$ AND $\texttt{closed}(\varphi)$
AND $\texttt{uncontrollability}(\varphi)$ AND
$\texttt{modifiability}(\varphi)$ AND $\texttt{event}(\varphi)$ AND
$\texttt{focality}(\varphi)$ AND
(($\texttt{positive\_valence}(\varphi)$ AND
$\texttt{absence}(\varphi)$) XOR
($\texttt{negative\_valence}(\varphi)$ AND
$\texttt{presence}(\varphi)$)) holds
THEN trigger behavior of *fear*(φ, ψ)
AND learn...

6. The behaviors of emotion: BDI example

- Eliciting rule for DISLIKING:

$$B_i(B_j\neg \mathit{Sad}_i(\varphi) \rightarrow I_j(\mathit{Sad}_i(\varphi))) \rightarrow \mathit{Dislike}_i(j)$$

6. The behaviors of emotion: BDI example

- Eliciting rule for DISLIKING:

$$B_i(B_j\neg \mathit{Sad}_i(\varphi) \rightarrow I_j(\mathit{Sad}_i(\varphi))) \rightarrow \mathit{Dislike}_i(j)$$

- Consequentive behavior rule:

$$\mathit{Dislike}_i(j) \rightarrow (B_i\neg \mathit{Sad}_j(\varphi) \rightarrow I_i(\mathit{Sad}_j(\varphi)))$$

The behaviors of emotion: KARO example

- Agent i is RESPONSIBLE for performing π resulting in φ:

  $$Res_i(\pi, \varphi) \leftrightarrow I_i(\pi, \varphi) \wedge Com_i(\pi) \wedge [\pi]_i(\varphi \wedge B_i\varphi)$$

  for some non-empty sequence actions π.

The behaviors of emotion: KARO example

- Agent i is RESPONSIBLE for performing π resulting in φ:

  $$\mathit{Res}_i(\pi, \varphi) \leftrightarrow I_i(\pi, \varphi) \wedge Com_i(\pi) \wedge [\pi]_i(\varphi \wedge B_i\varphi)$$

  for some non-empty sequence actions π.
- Example of an eliciting rule of ANGER:

  $$B_i\mathit{Res}_j(\pi, \mathit{Dislike}_i(j)) \rightarrow \mathit{Angry}_i(\pi, \varphi, j)$$

6. Conclusion

- **1. Distinguished processes of primary and secondary emotion, feeling, and gave general definitions**

6. Conclusion

- 1. Distinguished processes of primary and secondary emotion, feeling, and gave general definitions
- 2. The experinece of emotion as interrupt, and heuristical tool

6. Conclusion

- 1. Distinguished processes of primary and secondary emotion, feeling, and gave general definitions
- 2. The experinece of emotion as interrupt, and heuristical tool
- 3. As a result: Possible architecture for having emotion and feeling

6. Conclusion

- 1. Distinguished processes of primary and secondary emotion, feeling, and gave general definitions
- 2. The experinece of emotion as interrupt, and heuristical tool
- 3. As a result: Possible architecture for having emotion and feeling
- 4. Specified eliciting rules for emotions by antecedents, and a proposal general eliciting rule

6. Conclusion

- 1. Distinguished processes of primary and secondary emotion, feeling, and gave general definitions
- 2. The experinece of emotion as interrupt, and heuristical tool
- 3. As a result: Possible architecture for having emotion and feeling
- 4. Specified eliciting rules for emotions by antecedents, and a proposal general eliciting rule
- 5. Current AI languages can be tailored for expressing amotional behavior

QUESTIONS?

EXTENSION

Emotion-rationality?

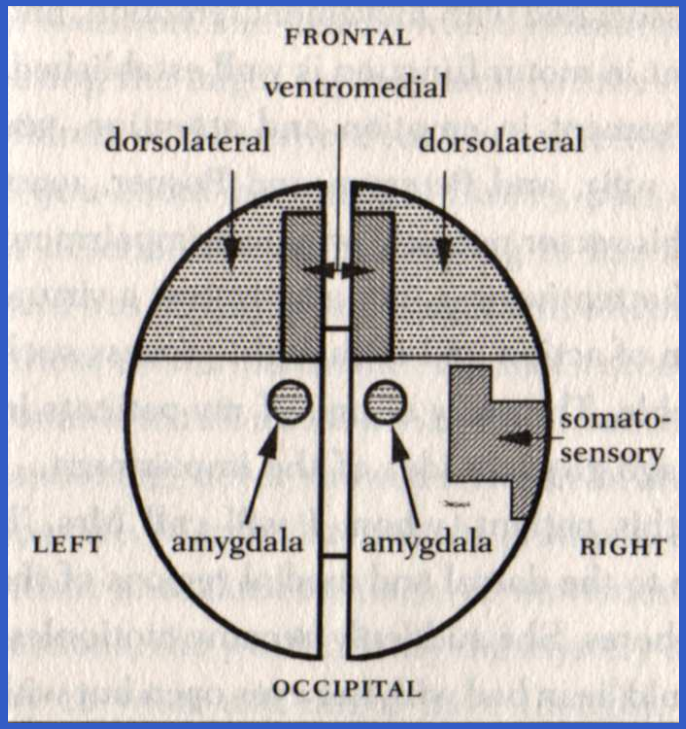


Fig. 1: A diagram representing the set of regions whose damage compromises both aspects of reasoning and processing of emotion (Damasio 1994).

Secondary emotion

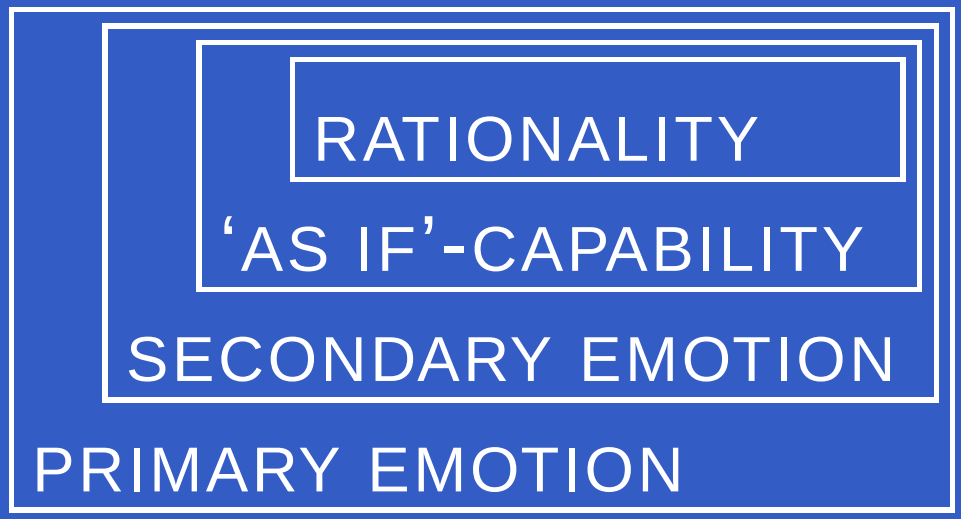


Fig. 2: Diagram of evolutionary development and dependency of cognitive processes. We probably inherited primary emotions, and merely the *ability* of the others.

Secondary emotion 2

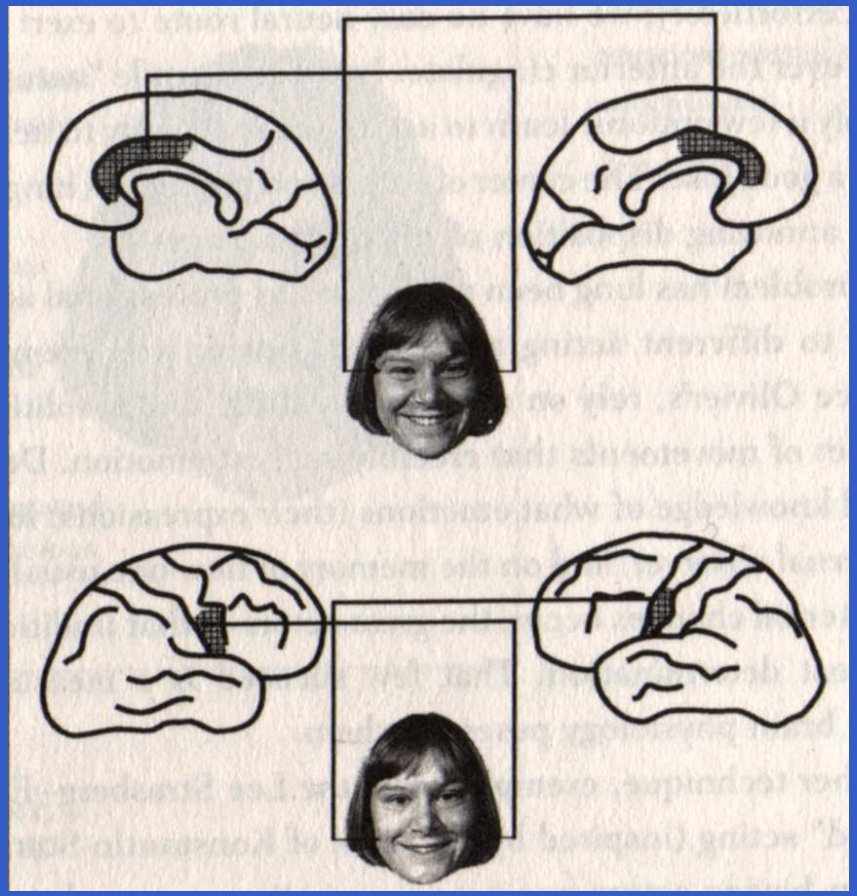


Fig. 3: The neural machinery for for the true smile is different from that of the voluntary (Damasio 1994).

Feeling

- On the moment feelings become cognitive, we probabably have images of them within our minds (from the somatosensory to the celebral cortices) (Shepard)

Feeling

- On the moment feelings become cognitive, we probabably have images of them within our minds (from the somatosensory to the celebral cortices) (Shepard)
- Images are multi-media and -dimensional conscious representations (Damasio 2004):

Feeling

- On the moment feelings become cognitive, we probabably have images of them within our minds (from the somatosensory to the celebral cortices) (Shepard)
- Images are multi-media and -dimensional conscious representations (Damasio 2004):

  “Thoughts are based directly on those neural representations, and only those, which are organised topographically and which occur in early sensory cortices”.

Feeling

- On the moment feelings become cognitive, we probabably have images of them within our minds (from the somatosensory to the celebral cortices) (Shepard)
- Images are multi-media and -dimensional conscious representations (Damasio 2004):

  > “Thoughts are based directly on those neural representations, and only those, which are organised topographically and which occur in early sensory cortices”.

- So even abstract thoughts would be imaged

Feeling 2

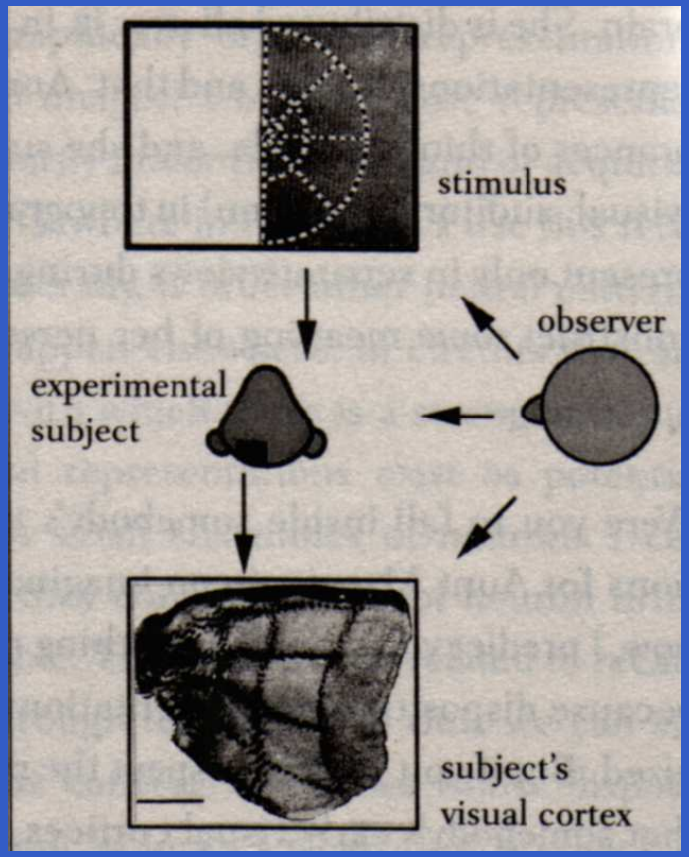
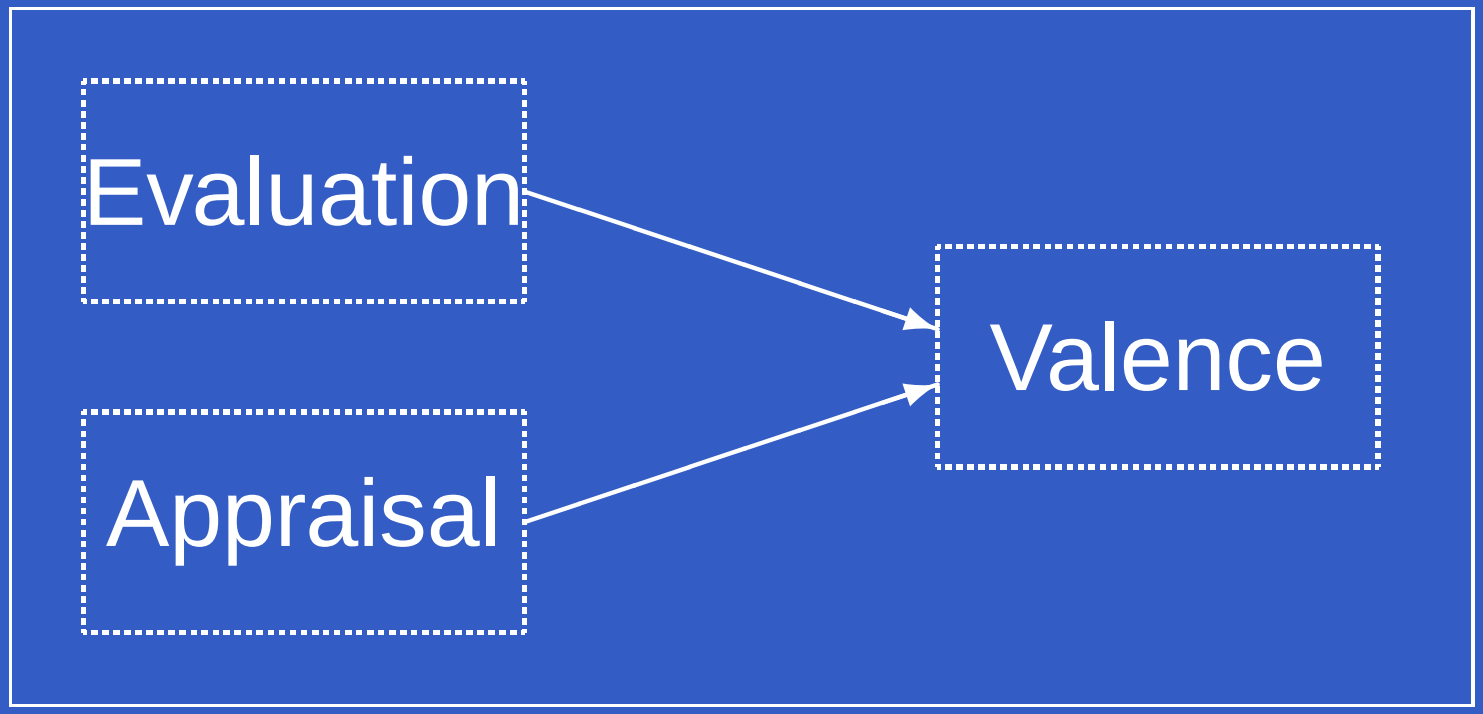
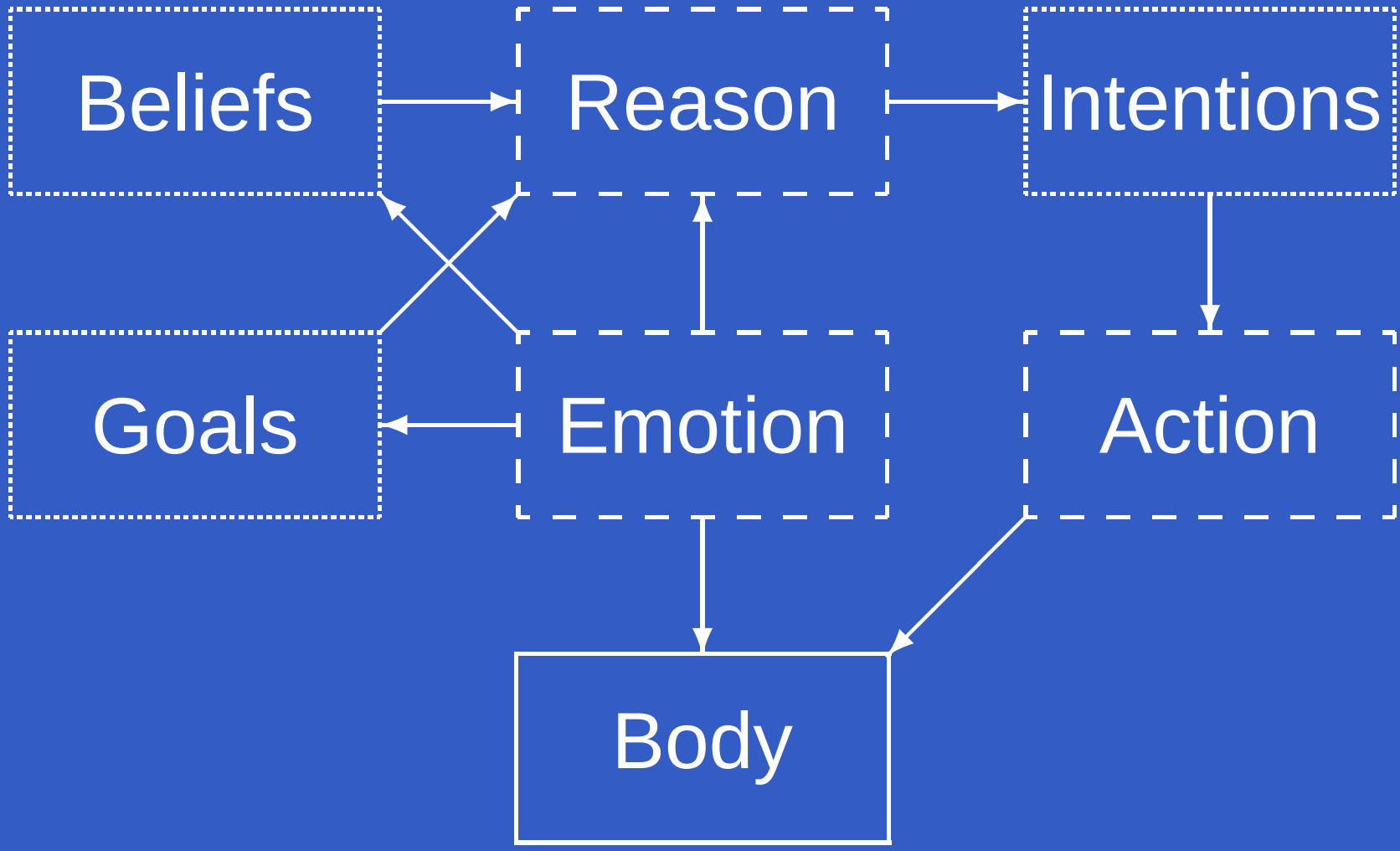


Fig. 4: Images are literally to be detected in brains (vision in macaques) (Tootell et al.).

Belief and goal generator



The position of emotion



Profiling by antecedents

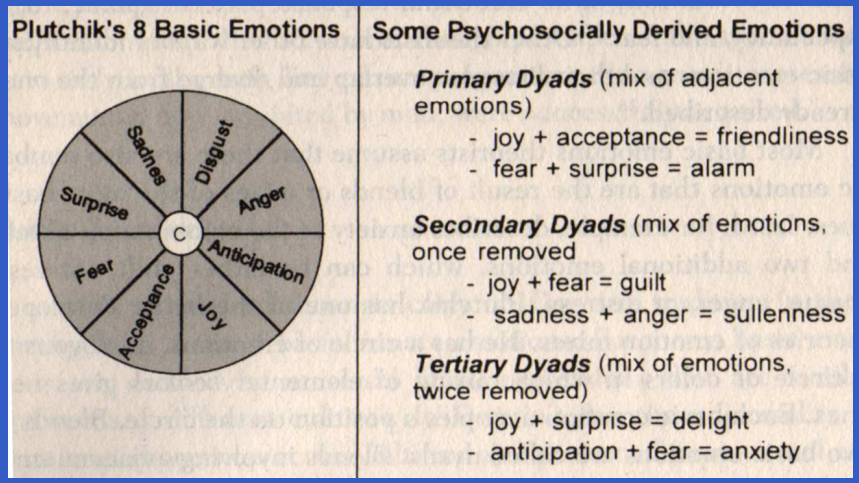


Fig. 5: Proposal of basic emotions seems not suffi cient to emotional subtleties (Plutchik).

KARO-based behavior: Explanation

- $I_i(\pi, \varphi)$: agent i has intention to do π for achieving φ
- $Com_i(\pi)$: i is committed to π, i.e. π is in i's agenda
- $B_i(\pi)$: i believes φ
- $\alpha \preceq \pi$: α is first action of sequence actions π
- $\pi \backslash \alpha = \pi'$ iff $\pi = \alpha; \pi'$
- $[\alpha]_i \varphi$: φ true after i performs action α

KARO-based behavior: Explanation 2

- $P_i(\pi, \varphi)$: i has practical possibility to perform π for achieving φ
- *angry*$_i(\pi, \varphi, j)$: i is angry with j with respect to remainder π of i's plan to achieve φ
- $Res_i(\pi, \varphi)$: i is responsible for doing π in order to achieve φ
- $Res_i(\pi, \varphi) \equiv I_i(\pi, \varphi) \wedge Com_i(\pi) \wedge [\pi]_i(\varphi \wedge B_i\varphi)$, for some non-empty sequence actions π and equals

Grouping by antecedents: Scheme

group	specification	intensity variables	some emotion types
Well-Being	appraisal of a situation as an event	desirability	**joy:** pleased about an event **distress:** displeased about an event
Fortunes-of-Others	presumed value of a situation as an event affecting another	desirability-for-other, liking, deservingness	**happy-for:** pleased about an event desirable for another **gloating:** pleased about an event undesirable for another **resentment:** displeased about an event desirable for another **jealousy:** resentment over a desired mutually exclusive goal **envy:** resentment over a desired non-exclusive goal **sorry-for:** displeased about an event undesirable for another
Prospect-based	appraisal of a situation as a prospective event	likelihood, effort, realization, desirability	**hope:** pleased about a prospective desirable event **fear:** displeased about a prospective undesirable event

Grouping by antecedents: Scheme 2

group	specification	intensity variables	some emotion types
Confirmation	appraisal of a situation as confirming or disconfirming an expectation	likelihood, effort, desirability	**satisfaction:** pleased about a confirmed desirable event **relief:** pleased about a disconfirmed undesirable event **fears-confirmed:** displeased about a confirmed undesirable event **disappointment:** displeased about a disconfirmed desirable event
Attribution	appraisal of a situation as an accountable act of some agent	praiseworthiness, strength of cognitive unit, expectation-deviation	**pride:** approving of one's own act **admiration:** approving of another's act **shame:** disapproving of one's own act **reproach:** disapproving of another's act

Grouping by antecedents: Scheme 3

group	specification	intensity variables	some emotion types
Attraction	appraisal of a situation as containing an attractive or unattractive object	familiarity appealingness	**liking:** finding an object appealing **disliking:** finding an object unappealing
Well-being/ Attribution	compound emotions		**gratitude:** admiration + joy **anger:** reproach + distress **gratification:** pride + joy **remorse:** shame + distress
Attraction/ Attribution	compound emotion extensions		**love:** admiration + liking **hate:** reproach + disliking **grief:** disappointment + liking

Profiling by antecedents: Scheme

	Positive character	*Negative character*	*Desire*	*Interest*	*Positive valence*	*Negative valence*	*Presence*	*Absence*	*Certainty*	*Uncertainty*	*Change*	*Open*	*Closed*	*Intentionality of other*	*Intentionality of self*	*Controllability*	*Uncontrollability*	*Modifiability*	*Finality*	*Object*	*Event*	*Focality*	*Globality*	*Strangeness*	*Familiarity*	*Value*
Joy	x				x		x				(x)	x						x			x	x				
	x					x		x			(x)	x						x			x	x				
Distress		x			x		x										x				x	x				
		x				x	x										x				x	x				
Desire			x		x			x													x	x				
Interest				x			x														x	x				
Grief		x			x			x			(x)								x		x	x				
Sorrow		x			x			x			(x)										x	x				
Fear		x				x	x			x			x				x	x			x	x				
		x			x			x		x			x				x	x			x	x				
Hope	x				x		x			x		x									x	x				
	x					x		x		x		x									x	x				
Anger		x			x			x						x		x					x	x				
		x				x	x							x		x					x	x				
Challenge	x					x	x			x																
Boredom		x		x				x								x					x	x				
Satisfaction	x		x		x							x									x		x			
Contentment	x				x		x														x	x			x	
Security	x					x		x													x	x				

Profiling by antecedents: Scheme 2

	Positive character	*Negative character*	*Desire*	*Interest*	*Positive valence*	*Negative valence*	*Presence*	*Absence*	*Certainty*	*Uncertainty*	*Change*	*Open*	*Closed*	*Intentionality of other*	*Intentionality of self*	*Controllability*	*Uncontrollability*	*Modifiability*	*Finality*	*Object*	*Event*	*Focality*	*Globality*	*Strangeness*	*Familiarity*	*Value*
Relief	x					x		x			x										x		x		x	
Anxiety		x				x	x			x											x	x				
Despair		x				x	x		x				x				x				x		x	x		
Disapp.		x			x			x			x		x				x				x	x				
Hate		x				x	x										x				x	x				
Frustration		x			x			x					x	x			x			x		x				
		x				x	x						x	x			x			x		x				
Guilt		x				x	x								x						x	x				x
Contempt		x				x	x													x		x				
Resignation						x	x												x		x	x				
Love	x				x		x													x		x				
Admiration	x				x		x							x						x		x				x
Pride	x				x		x								x					x		x				
Disgust		x				x	x								x					x		x				
Self-hatred		x				x	x														x		x			
Depression		x			x			x													x		x			
Bliss	x				x		x							x							x	x				x

Belief-based behavior

- Belief-based
  - Non-monotonicy
  - Possibility
  - Gradability
  - Non-logical omniscience
- Further extensions
  - Desirability
  - Interestingness
  - . . .

Belief-based behavior: KD45

(A1)	All propositional tautologies.	
(A2)	$(B\varphi \wedge B(\varphi \rightarrow \psi)) \rightarrow B\psi$	
(D)	$\neg B\bot$	*Not inconsistent.*
(A4)	$B\varphi \rightarrow BB\varphi$	$+$ *Introspection.*
(A5)	$\neg B\varphi \rightarrow B\neg B\varphi$	$-$ *Introspection.*
(R1)	$\varphi, \varphi \rightarrow \psi \vdash \psi$	*Modus ponens.*
(R2)	$\varphi \vdash B\varphi$	*Necessitation.*

Belief-based behavior: Logical omniscie

- Not all valid formulas are believed:

$$B\varphi \rightarrow \neg B \neg \varphi \quad \text{(A1 , A2, D, R1)}$$

- Belief of possibility to be wrong:

$$B(B\varphi \rightarrow \varphi) \quad \text{(A1, A2, D, A5, R1, R2)}$$

- Possible solution:

$$B_E\varphi \leftrightarrow ((B\varphi \vee P\varphi) \wedge A\varphi)$$

KARO-based behavior: Eliciting rule

- The eliciting rule of *anger*

$$I_i(\pi, \varphi) \wedge Com_i(\pi) \wedge B_i Res_j(\alpha, \neg P_i(\pi, \varphi)) \rightarrow$$

$$[\alpha]_j(B_i(\neg\varphi \wedge \neg P_i(\pi, \varphi)) \rightarrow \mathit{angry}_i(\pi, \varphi, j))$$

when $j \neq i$.