How to decide what to do?

Mehdi Dastani
BBL-521
M.M.Dastani@uu.nl

Decision Making

Decision making is a fundamental capacity in science and society at large

"*A user who seeks advice about financial planning wants to retire early, secure a good pension and maximize the inheritance of her children. She can choose between a limited number of actions: retire at a certain age, invest her savings and give certain sums of money to her children. However, she does not know all factors that might influence her decision. She does not know if she will get a pay raise next year, the outcome of her financial actions is uncertain, and her own preferences may not be clear since, for example, securing her own pension conflicts with her children's inheritance.*" (Doyle and Thomason)

Decision Making

- Decision making is a process of selecting a belief or a course of action among alternative possibilities based on **information** and **objectives** of the decision maker.

- Activities that involve decision making are *problem solving*, *diagnosis of diseases*, *design of experiments*, *treatment selections*, *verdicts in courts*, *hiring of employees*, *driving a car*, and *developing policies*.

- Dimensions on decision making:
  - *Individual* versus *group* decisions.
  - Decision making under *uncertain* or *incomplete information*.
  - *Prescriptive* versus *descriptive* decision making.
  - *Bias* or *constraints* in decision making.

Data-driven Decision Making

- The emergence of big data, the availability of computational resources, together with the advances in data science and artificial intelligence have transferred decision making research and practice towards what is called **data-driven decision making**.

- Data-driven decision making introduces threats that can lead to decisions with disastrous outcomes.

- Making decisions based on biased, imperfect, or incomplete data, and using inaccurate algorithms and black-box methods to collect and process data may lead to suboptimal, sometimes incorrect, non-transparent, non-explainable or non-repeatable outcomes.
  - US risk assessment algorithms used by judges during sentencing were found to exhibit racial bias.
  - Amazon developed and later abandoned a new recruiting system that turned out to show bias against woman.

Rational Decision Making

Various conceptualisations and formalisations of ratonal decision Making:

- Classical decision theory (CDT)
- Qualitative decision theory (QDT)
- Knowledge-based systems (KBS)
- Beliefs, Desires, and Intentions models (BDI)

	classical decision theory (CDT)	qualitative decision theory (QDT)	knowledge-based systems (KBS / BDI)
underlying concepts	probability function utility function decision rule	likelihood ordering preference ordering decision criterion	knowledge / belief goal / desire reasoning / deliberation
time	(Markov) decision processes	decision-theoretic planning	temporal models / BDI systems
multiagent	classical game theory	qualitative game theory	normative systems

Table: Theories, systems and models of decision making.

Classical Decision Theory (CDT)

The classical decision setting is (A, W, U, P), where

- A stands for a set of alternative actions.
- W stands for the set of all possible worlds or outcomes.
- $U : W \to \mathbb{R}$ is a measure of outcome value that assigns a utility value from $\mathbb{R}$ to each outcome $w \in W$, and
- P is a measure of the probability of outcomes conditional on actions, with $P(w|a)$ denoting the probability that outcome w comes about after taking action $a \in A$.

The expected utility $EU(a)$ of an action a is the weighted average of the utilities of all outcomes associated with the action (weighted according to the probability that the act will lead to the outcomes), that is,

$$EU(a) = \sum_{w \in W} U(w)P(w|a)$$

Decision rule: Select action α that maximises the expected utility (MEU):

$$\alpha = \underset{a \in A}{arg\max}\ EU(a)$$

Classical Decision Theory (CDT)

Quantitative representations of probability and utility are problematic.

- CDT does not address decision making in unforeseen circumstances.
- CDT offers no means for capturing generic preferences.
- CDT provides little help to decision makers who exhibit discomfort with numeric trade offs.
- CDT provides little help in effectively representing decisions involving broad knowledge of the world.

Qualitative decision theory

Given a set of propositional atoms Π, the semantics of the qualitative decision making is based on models of the form

$$M = \langle W, \leq_P, \leq_N, V \rangle$$

where

- W is a set of possible worlds (outcomes),
- $\leq_P$ is a reflexive, transitive and connected preference ordering relation on W ($\leq_P$ represents the relative desirability of worlds; $w \leq_P v$ expresses w is at least as preferred as v, but possibly more),
- $\leq_N$ is a reflexive, transitive and connected normality ordering relation on W ($\leq_N$ represents the relative likelihood of worlds; $w \leq_N v$ expresses w is at least as plausible as v), and
- $V : W \rightarrow 2^{\Pi}$ is a valuation function.

Qualitative decision theory

- **Conditional preferences** are represented in the logic by modal formulas $\mathcal{I}(\phi|\psi)$, to be read as 'ideally ϕ if ψ', which expresses the most preferred ψ worlds with respect to $\leq_P$ are ϕ worlds.

- Example: $\mathcal{I}(\mathit{umbrella}|\mathit{rain})$ expresses "in the most preferred rain-worlds the agent carries an umbrella.

- **Conditional probabilities** are represented in the logic by a default conditional $\psi \Rightarrow \phi$, to be read as 'ϕ holds at the most normal ψ-worlds.

- Example: $\mathit{Rain} \Rightarrow \mathit{Wet}$ expresses "in the most normal rain-worlds, the agent is wet.

Qualitative decision theory

Given a set of facts KB, a **goal** is any proposition ϕ such that

$$M \models \mathcal{I}(\phi \mid Cl(KB))$$

where $Cl(KB)$ is the default closure of the facts KB defined as follows:

$$Cl(KB) = \{\phi \mid KB \Rightarrow \phi\}$$

In goal-based planning, adopting a proposition as a **goal** commits the agent to **decide a course of actions** to accomplish the goal.

Beliefs, Desires, Intentions (BDI) theory

Quantitative probabilities and utilities in classical decision theory and their corresponding qualitative normality and preference ordering in qualitative decision theory are reduced to binary relations **beliefs** and **desires**, respectively.

As argued by Bratman, in order to stabilise sequential decisions **intentions** is introduced.

- **Knowledge** (K) and **beliefs** (B) represent the information of an agent about the state of the world.
- **Goals** (G) or **desires** (D) represent the preferred states of affairs for an agent.
- **Intentions** (I) correspond to previously made commitments of the agent, either to itself or to others.

Sara desires to take a holiday and believes she has a free weekend. So, she decides to take a weekend holiday. One day later, she sees a very attractive job offer (she desires to have the job) and believes she is qualified to apply for the job. However, her intention to take a weekend holiday makes it impossible to apply for the job.

Beliefs, Desires, Intentions (BDI) theory

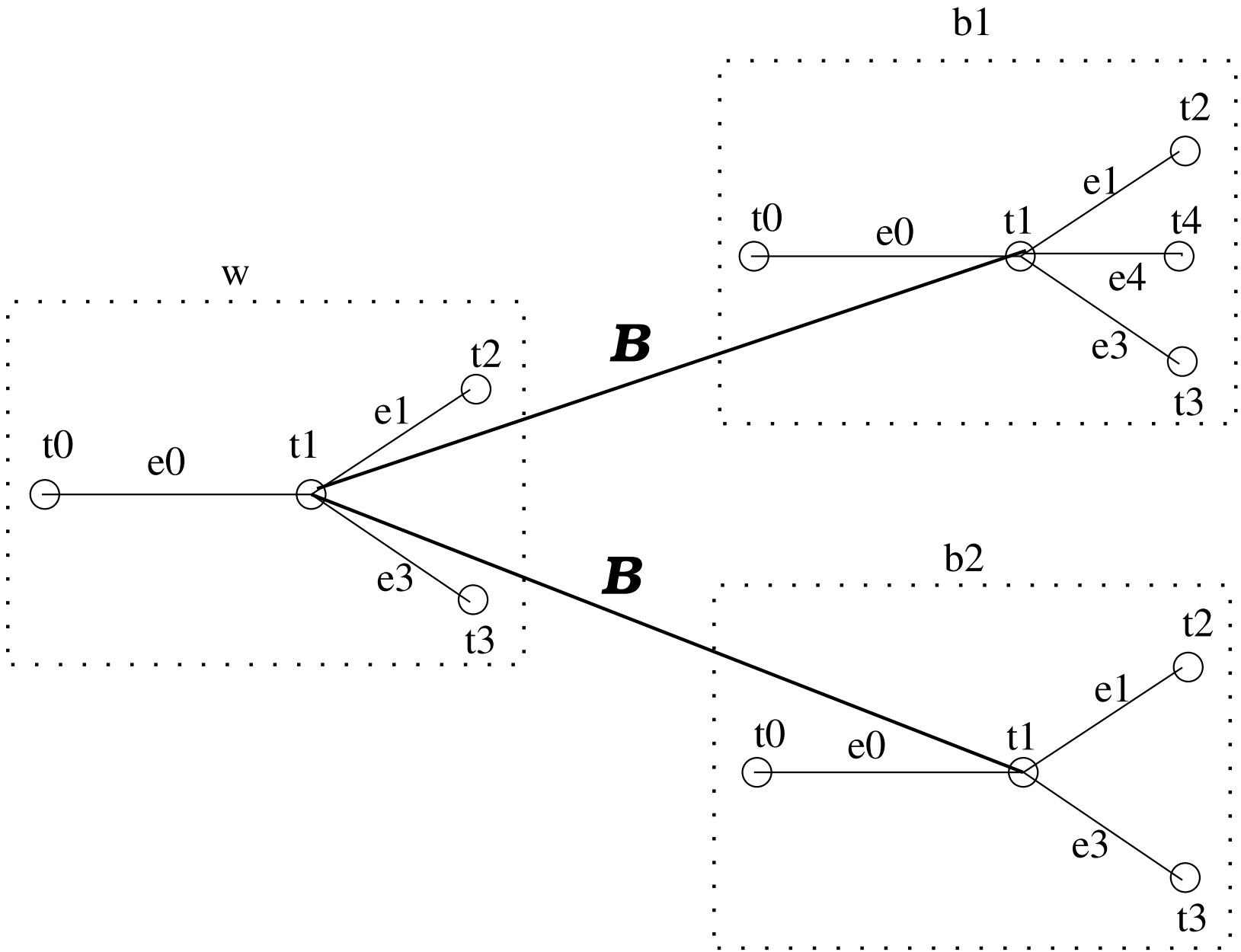
A BDI model M is a tuple

$$M = \langle W, E, T, <, U, \mathcal{B}, \mathcal{D}, \mathcal{I}, \Phi \rangle$$

where

- W is the set of worlds,
- E is the set of primitive event types,
- T is a set of time points,
- $<$ a binary relation on time points,
- U is the universe of discourse,
- $\mathcal{B} \subseteq W \times T \times W$ represents the belief accessible worlds,
- $\mathcal{D} \subseteq W \times T \times W$ represents the desire accessible worlds,
- $\mathcal{I} \subseteq W \times T \times W$ represents the intention accessible worlds, and
- Φ is a mapping from first-order entities to elements in U for any given world and time point.

Beliefs, Desires, Intentions (BDI) theory



Beliefs, Desires, Intentions (BDI) theory

- Let M be a BDI model.

$$M = \langle W, E, T, <, U, \mathcal{B}, \mathcal{D}, \mathcal{I}, \Phi \rangle$$

- We write $B(\phi)$ to express ϕ **is believed**.

- An agent modelled by M believes ϕ in world w at time t iff ϕ holds in all belief accessible worlds, i.e.,

$$M, w_t \models B\phi \Leftrightarrow \forall w' \in \mathcal{B}_t^w \; M, w'_t \models \phi$$

- We write $\Diamond\phi$ to express **eventually** ϕ

- A path formula $\Diamond\phi$ is evaluated relative to the interpretation M along a path $(w_{t_0}, w_{t_1}, \ldots)$ as follows:

$$M, (w_{t_0}, w_{t_1}, \ldots) \models \Diamond\phi \Leftrightarrow \exists k \geq 0 \text{ such that } M, (w_{t_k}, \ldots) \models \phi$$

Constraints on BDI Models

- Constraints on beliefs:
  - $B(\phi \rightarrow \psi) \rightarrow (B\phi \rightarrow B\psi)$
    If you believe that ϕ implies ψ then if you believe ϕ then you believe ψ
  - $B\phi \rightarrow \neg B \neg \phi$
    If you believe ϕ then you do not believe that ϕ is false
  - $B\phi \rightarrow BB\phi$
    If you believe ϕ then you believe that you believe ϕ
  - $\neg B\phi \rightarrow B \neg B\phi$
    If you do not believe ϕ then you believe that you do not believe ϕ

- Constraints on desires:
  - $D(\phi \rightarrow \psi) \rightarrow (D\phi \rightarrow D\psi)$
  - $D\phi \rightarrow \neg D \neg \phi$

- Constraints on intentions:
  - $I(\phi \rightarrow \psi) \rightarrow (I\phi \rightarrow I\psi)$
  - $I\phi \rightarrow \neg I \neg \phi$

Constraints on BDI Models

Realism: constraints on interaction between beliefs, desires and intentions.

- $B\phi \rightarrow D\phi$
  An agent's desires should be consistent with its beliefs.
- $D\phi \rightarrow I\phi$
  An agent's intentions should be consistent with its desires.
- $I(Next\phi) \rightarrow Next\phi$
  Agents should behave according to their intentions.
- $D\phi \rightarrow B(D\phi)$
  Agents should be aware of their desires.
- $I\phi \rightarrow B(I\phi)$
  Agents should be aware of their intentions.

Constraints on BDI Models

- **Blind-commitment strategy**: If an agent intends the ϕ is eventually true, then the agent will maintain its intention until she believes ϕ is true.

$$I(\diamond\phi) \rightarrow (I(\diamond\phi) \; Until \; (B\phi))$$

- **Single-minded strategy**: If an agent intends the ϕ is eventually true, then the agent will maintain its intention as long as she believes ϕ can become true.

$$I(\diamond\phi) \rightarrow (I(\diamond\phi) \; Until \; (B\phi \vee \neg B(\diamond\phi)))$$

- **Open-minded strategy**: If an agent intends the ϕ is eventually true, then the agent will maintain its intention as long as she intends ϕ.

$$I(\diamond\phi) \rightarrow (I(\diamond\phi) \; Until \; (B\phi \vee \neg I(\diamond\phi)))$$

Decision and Norms

Norms and Decisions

How does **norms** influence an agent's decisions?

- **Legal norms**: norm issuing authority with monitoring and enforcement mechanism.
- **Social norms**: emerges through agents' interaction and monitored and enforced by the agents' collective.
- **Moral norms**: practice independent, intrinsically motivating, individual accountability.

Various types of Norms: Syntax and Semantics

- Conditional norms with deadline:
  - $(cond, F(\phi), deadline)$
  - $(cond, O(\phi), deadline)$
- Norms combining states and actions
  - $F(\phi\ ,\ \alpha\ ,\ sanc)$
  - $(cond, F(\phi), Act_r)$
  - $(Act_t, cond, F(\phi), Act_r)$

Conditional Norms

- Let *cond*, ϕ, *d* be boolean combinations of propositional variables plus special sanction variable *san*.
- A conditional obligation is represented by the tuple

$$(cond, O(\phi), d, san)$$

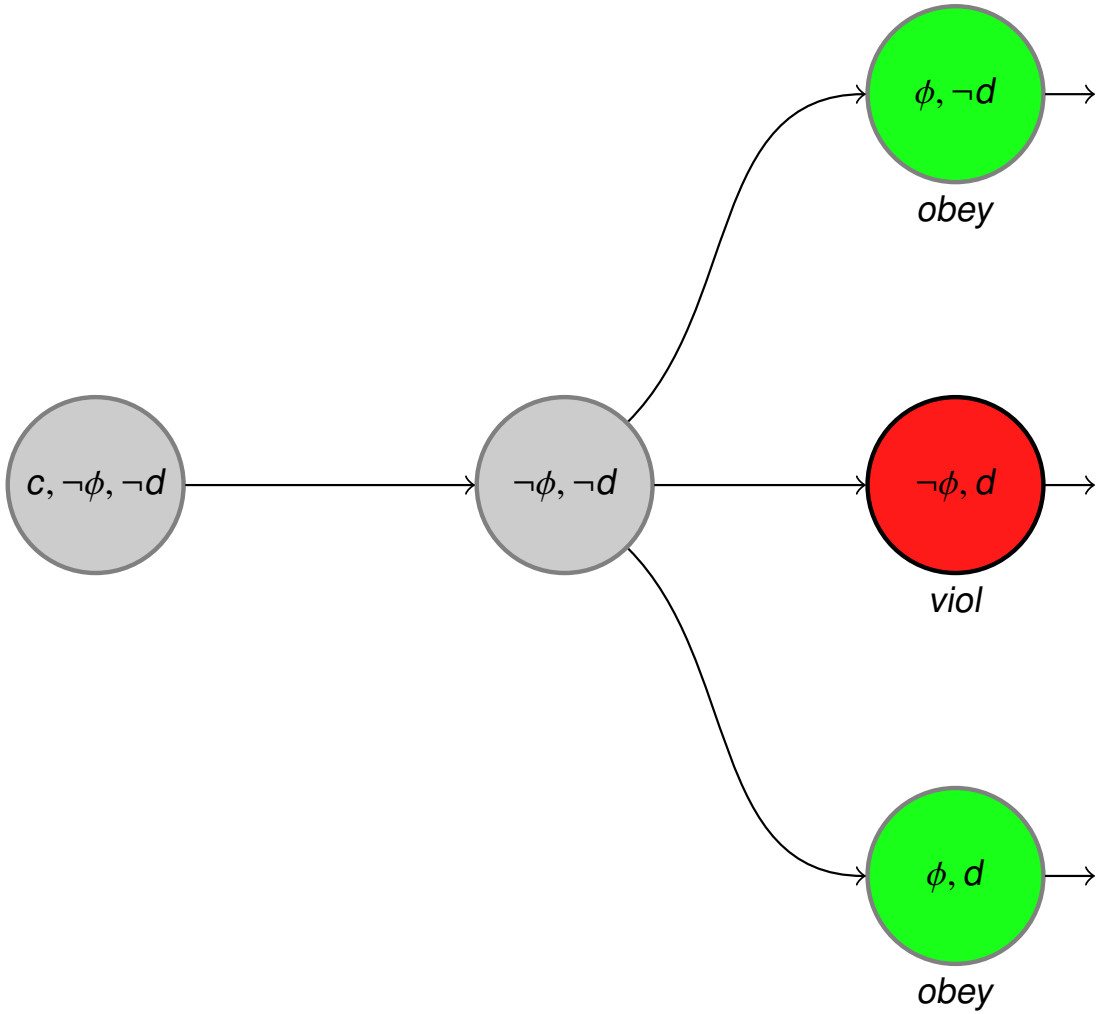
- A conditional prohibition is represented by the tuple

$$(cond, P(\phi), d, san)$$

- A norm set *N* is a set of conditional obligations and conditional prohibitions.

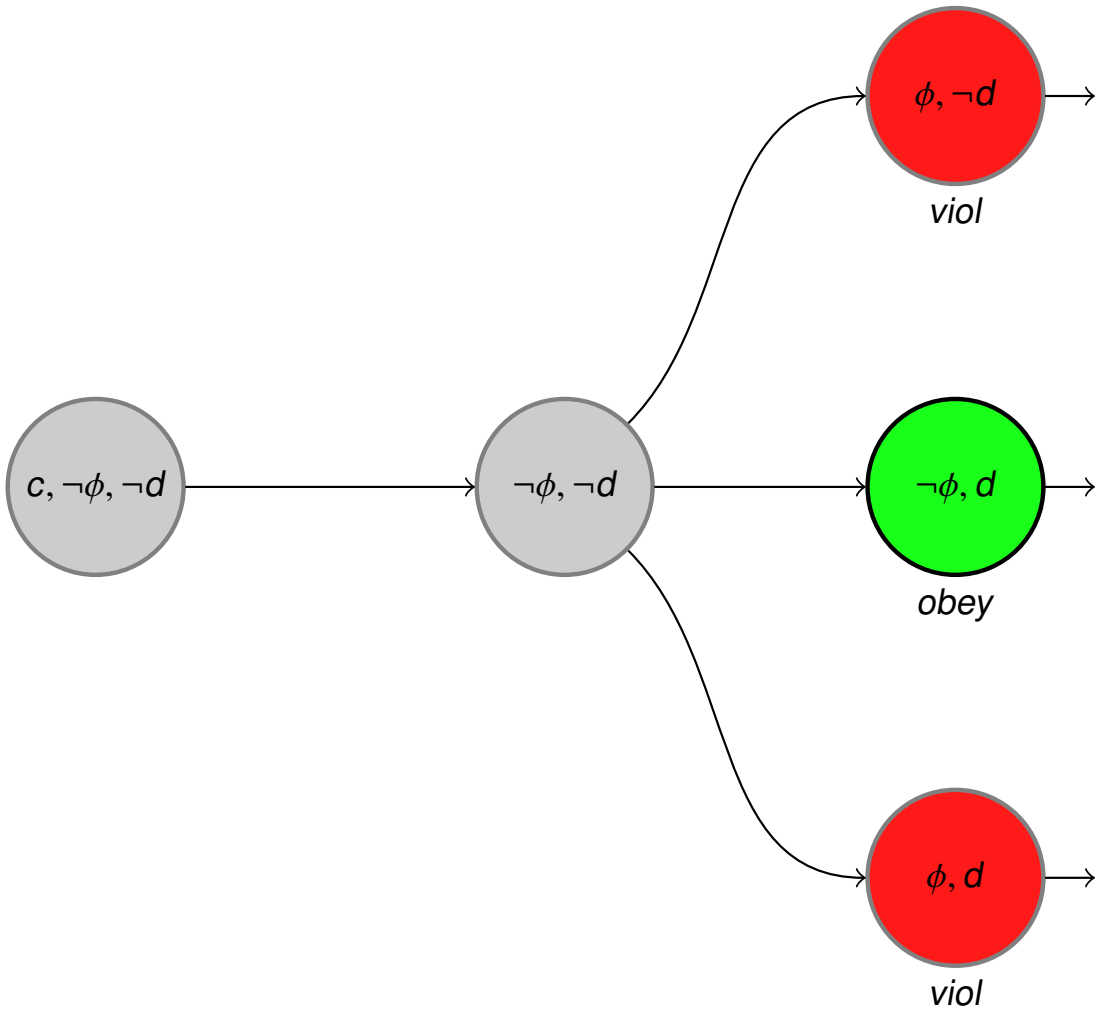
Meaning of conditional Obligation

$(\mathbf{cond}, \mathbf{O}(\phi), \mathbf{d}, \mathbf{san})$

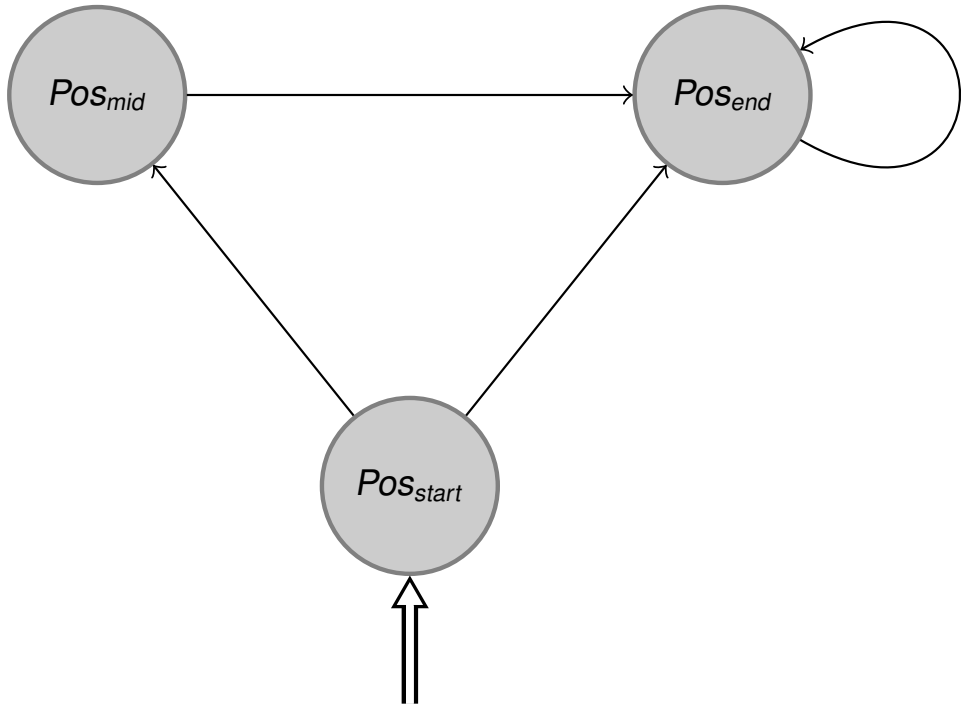


Meaning of conditional Prohibition

$(\mathbf{cond}, \mathbf{F}(\phi), \mathbf{d}, \mathbf{san})$



Transition systems and Norms



$$(Pos_{start}\ ,\ \mathbf{O}(Pos_{mid})\ ,\ Pos_{end}\ ,\ san)$$

Balancing between Desires and Norms

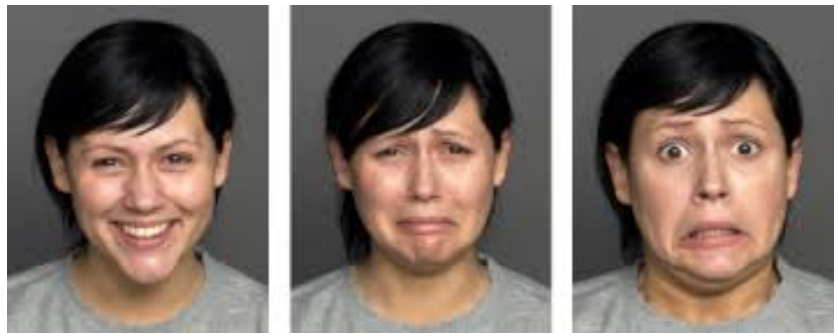
What to do if desires conflicts with norms? How to resolve conflicts?

- **Agent types**: Social/Moral agents prioritise norms over desires.
- **Commensurability**: utility of desires and sanctions of norm violations.
- Describe desires and norms in terms of some general **values** such as power, security, hedonism, and achievement.

Decision and Emotions

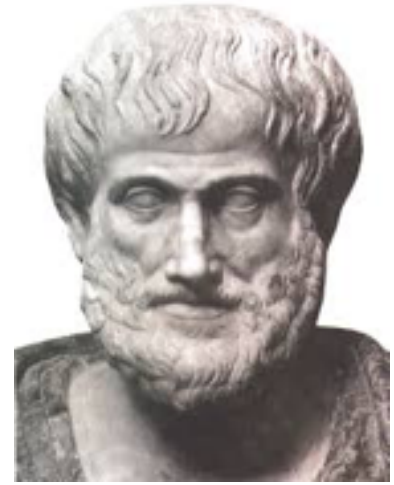
Emotions and Decision

Emotions influence decision making.



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

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

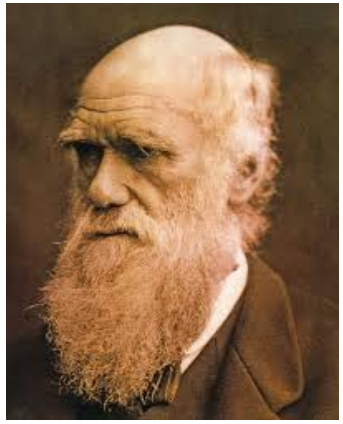


- Reason should rule our soul and monitor our emotional responses.
- Emotion elicitation can be used in audiences of public speaking (*Rhetoric*) and tragic drama (*Poetics*).
- Aristotle analyses several emotions in terms of
  - the beliefs they presuppose (e.g., anger requires the belief that oneself 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)

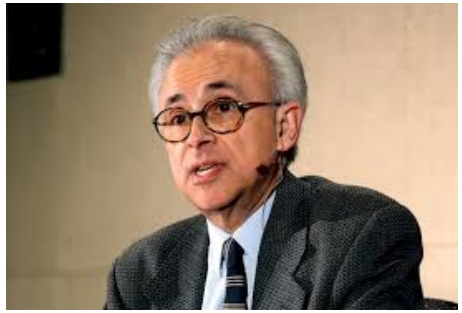
- Emotions are the passions of the soul, i.e., the things suffered by our thinking aspect.
- Descartes considers emotions as irrational and disruptive forces to reasoning and rationality.
- But, emotion was recognised as serving a useful function by directing ones thoughts and attentions to what is important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

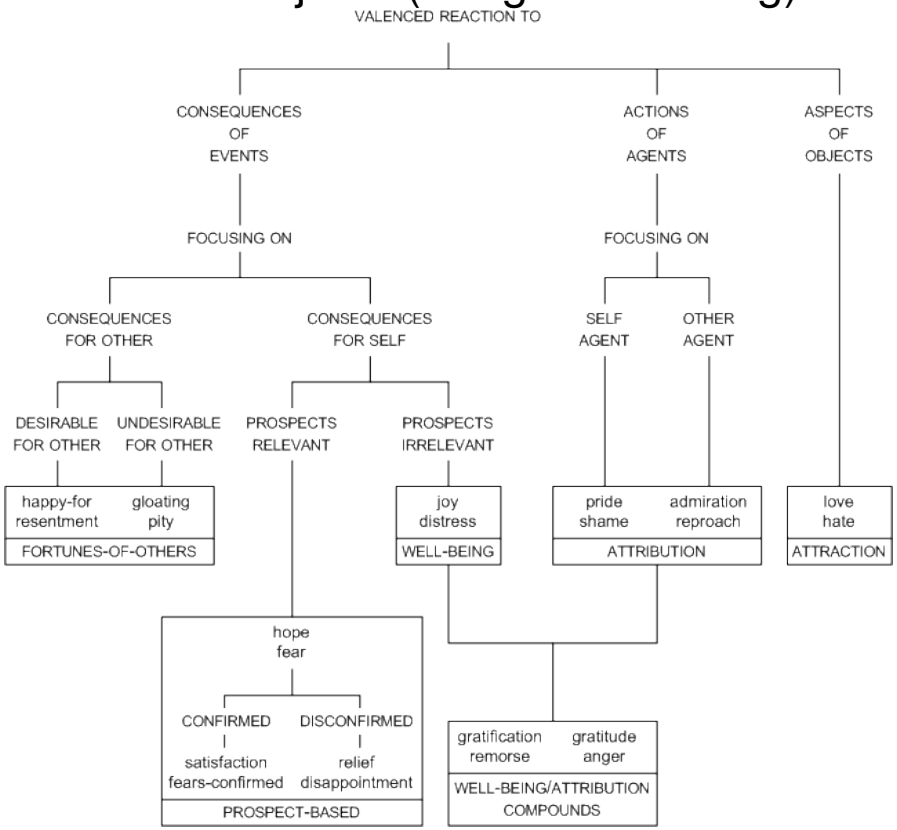
Affect (Scherer 1984)

- 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

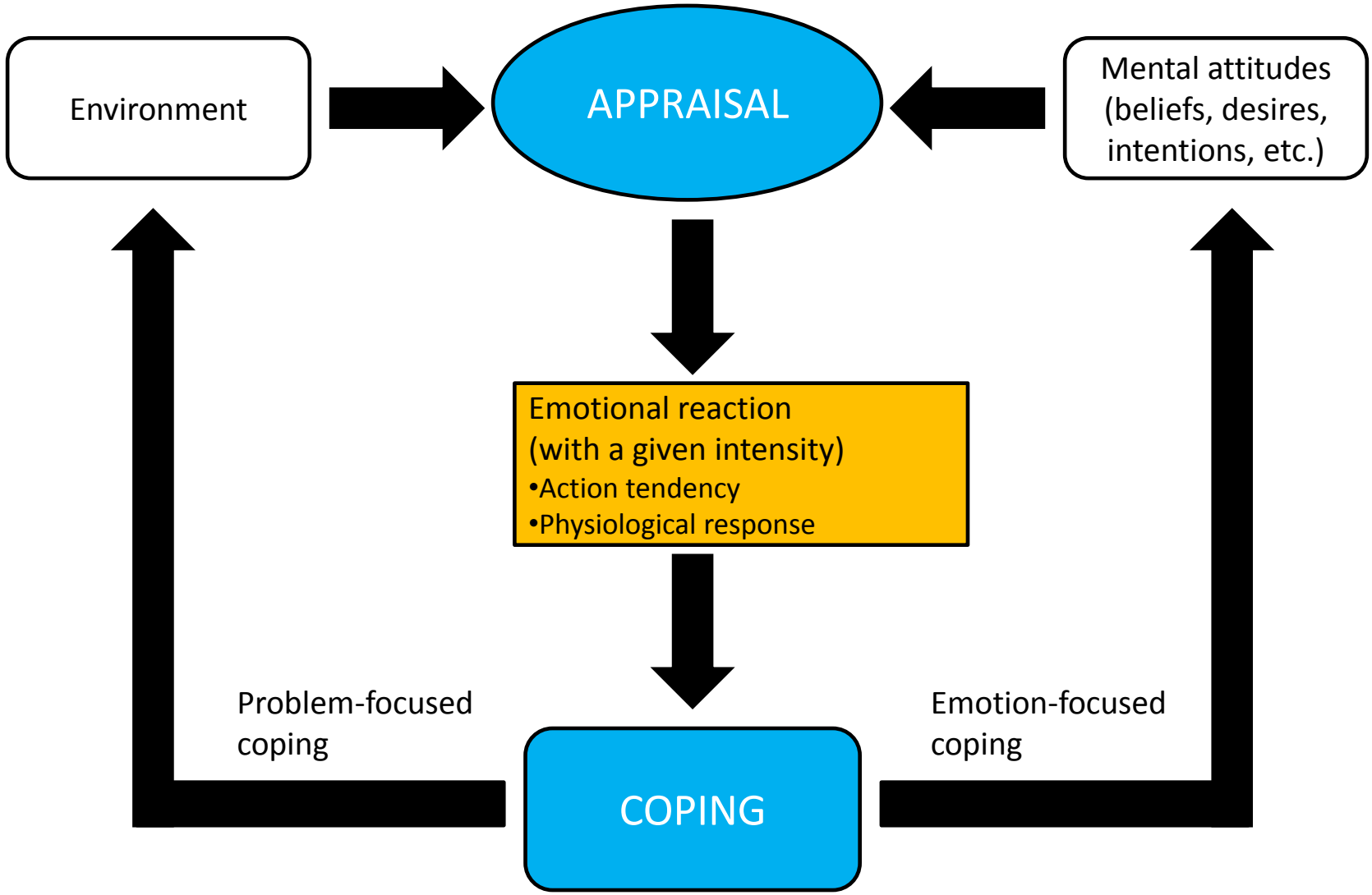
Pity: (displeased about) an event presumed to be undesirable for someone else

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

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

Appraisal and coping circuit

(Lazarus, 1991; Gratch & Marsella, 2005)



Conclusion

- How to make software agents that make autonomous decisions?
- Decision theory assumes an aggregation of all motivational attitudes.
- BDI models are concerned with consistent decision behaviour.
- Aggregation of motivational attitudes are strong assumption for explaining and designing systems.
- Balancing various various motivational attitudes?