Minds and Machines (Lecture 7): Emotion and AI

Prof. Ioannis Votsis

Philosophy Discipline
ioannis.votsis@nulondon.ac.uk
www.votsis.org

Introduction

A little historical context

- Influential treatments of the emotion can be found in Aristotle, Descartes and James.

- As for the concept, a variety of cognate concepts have been used over the years, including:

* passion	* desire	* sentiment
* affection	* appetite	* upheaval

- The English term/concept was imported from the French ‘émotion’ sometime between the 16th-18th centuries.

- Apparently, it is only in the 18th century that the term begins to denote a class of mental states (Dixon 2003; 2012).

Basic vs. complex emotions

- Emotions are often distinguished on the basis of their complexity. Roughly speaking, they fall into two categories:

Basic Emotions	Complex Emotions
anger	dread
disgust	envy
fear	guilt
happiness	narcissism
sadness	pride
surprise	schadenfreude

NB: Some examples are more contentious than others w.r.t. whether they are: (i) emotions, (ii) simple or complex.

Plan

- In what follows, we consider:

  * some contemporary accounts of emotions

  * whether these are applicable to artificial agents

- A number of pertinent questions arise, including:

  * What is the relationship between emotions and motivations?

  * What is the relationship between emotions and rationality?

Theories of Emotion

Feeling theories

- This family of views has its roots in Aristotle, Descartes and James. As Scarantino and de Sousa (2018) note:

  "The Feeling Tradition takes the way emotions feel to be their *most* essential characteristic, and defines emotions as distinctive conscious experiences" (emphasis added).

  NB: A.k.a. 'non-cognitive' theories, they are advocated by, among others, James (1884) and Lange (1885).

  Example: Fear is the feeling of our body's reactions to a dangerous situation.

  Problem: Feelings may not be sufficient to individuate emotions (Cannon 1929; Schachter and Singer 1962).

Evaluative theories

- This family of views has its roots in Aristotle and the Stoics. As Scarantino and de Sousa (2018) note:

  "The Evaluative Tradition… defines emotions as being (or involving) distinctive evaluations [or judgments] of the eliciting circumstances."

  NB: A.k.a. 'cognitive' theories, they are advocated by, among others, C.D. Broad (1954) and M. Nussbaum (2001; 2004).

  Example: Anger is directed (e.g. at a given person) and is our evaluation/judgment that this person has wronged us.

  Problem: Such judgments can occur regardless of any accompanying emotion (Solomon 2004).

Hume's theory of motivation

- It's worth considering Hume's view on passion and action as it throws light on another objection to evaluative theories.

- According to Hume, reasoning w/ideas is not sufficient for action. We also need some desire/affection to motivate us.

  "Where the objects themselves do not affect us, their [presumed causal] connexion can never give them any influence; and 'tis plain, that as reason is nothing but the discovery of this connexion, it cannot be by its means that the objects are able to affect us (Treatise 2.3.3.3/414).

  Example: The recognition of a connection between exercise and slimming is not sufficient motive for exercise.

Motivational theories

- As Scarantino and de Sousa (2018) note:

  "The Motivational Tradition defines emotions as distinctive motivational states… where a motivational state broadly understood is an internal cause of behaviors aimed at satisfying a goal"

  NB: Advocated by, among others, Dewey (1894; 1895).

  Example: When we feel angry, that is itself a motivational state that leads to action, e.g. physically lashing out.

  Problem: Motivations may not be sufficient to individuate emotions.

Hybrid theories

- More recently, the leading theories combine ideas from the feeling, the evaluative and the motivational traditions:

  "… a gradual convergence of the Evaluative and Feeling Traditions, with the former now identifying emotions as evaluative perceptions… and the latter identifying emotions as evaluative feelings… the distinction between evaluative (or cognitivist) theories and feeling theories is increasingly blurred" (Scarantino and de Sousa 2018)

  NB: Advocated by, among others, Brady (2013), Prinz (2004) and Roberts (2003).

  Example: When we feel fear we perceive the galloping of our heart but also evaluate the situation and take action.

Emotion and Rationality

The changing landscape

- In the past, theorising about the emotions was likely to render them devoid of any (links to) rationality.

  Examples:
  * The Stoics thought of the emotions as false judgments.
  * The Romans described anger as a short bout of madness (*ira furor brevis est*).

- More recently, thanks in part to cognitivist/hybrid theories, the relation to rationality is revisited (e.g. Solomon 1973).

- For example, fear *can be* rationally apt when directed towards danger as it allows us to avoid it/take evasive action.

  NB: Clearly, emotions are not always rationally apt.

Coherence and strategic alignment

- One relevant dimension of rationality, viz. coherence, has been explored in some detail by scholars, e.g. Helm (2009).

  External Coherence:
  Anger may arise because of my *belief* I've been wronged. If my belief changes, then probably so does my emotion.

  Internal Coherence:
  Fear of war is likely to be accompanied with *sadness* at the loss of life and potentially empathy or sympathy.

- Emotions can also help us act in ways that are strategically aligned w/our interests, e.g. guilt prevents us from cheating.

  NB: Again, not all instances of emotions will be helpful.

Emotions and AI

- The increasing reconciliation of emotion and rationality and the rise of cognitivist/hybrid theories bodes well for AI.

- That's because rationality and cognition seem easier (not easy) to mechanise and therefore to algorithmically produce.

- Emotions are becoming more prominent in AI design:

  * emotions as interrupters (Sloman and Croucher 1981)
  * reading emotions from speech (Mirsamadi et al. 2017)
  * reading emotions from facial expressions (Kim et al. 2019)

  "An analysis of more than 6 million YouTube videos finds that people around the world make similar facial expressions in similar social contexts" (L.F. Barrett 2021: 200).

DeLancey’s Passionate Engines

A naturalistic turn

- DeLancey (2002) provides a rare monograph on AI and emotions from a philosophical perspective.

- Unlike many philosophers, he has one eye on AI engineering and makes two methodological commitments.

- One of these is that the emotions need to be understood through the naturalistic lens.

  Naturalism: Employing scientific/empirical methods with the aim of discovering natural kinds.

- He contrasts naturalism with conceptual analysis and social constructivism and rejects both.

A non-cognitivist turn

- Another commitment is that the cognitivist/rationality-first approach is not likely to help us understand autonomy.

  Cognitivism: Prioritising cognitive processing (e.g. reasoning) over abilities like perceptions and sensorimotor reflexes.

  Autonomy: 'the capacity to be goal-directed' (Scarantino 2004: 229).

- Given their ties, he also thinks that the symbol-manipulation approach to AI is not likely to produce autonomous systems.

- He thus calls for "a revised [non-cognitivist] view of the mind".

The biomorphic argument

- DeLancey wants AI researchers to draw inspiration from nature and particularly biology (the biomorphic argument).
- After all, nature has designed various autonomous systems more efficient, e.g. more adaptive, than any GOFAI system.
- Such systems almost always lack symbol manipulation but possess affective mechanisms that enable autonomy.
- He thus urges AI engineering to follow biology:

  "[We must think] of an intelligent system not as a symbol manipulator, but as foremost a passionate engine… artificial autonomous systems should recapitulate… examples of phylogeny" (208).

The End