Master's Thesis

Neurosymbolic AI for Social Cognition - Literature Study

Jarne Demoen
Vrije Universiteit Brussel

29/10/2025

1 Neurosymbolic AI for Social Cognition

This thesis aims to combine neural and symbolic AI to improve how computers understand the term "social cognition". Social cognition is the way humans interpret others' emotions, intentions, and social situations.

- **Symbolic AI** uses explicit rules such as "people are happy at weddings"
- **Subsymbolic AI (deep learning)** uses pattern recognition from raw data (facial expression in images)
- **Neurosymbolic AI** merges both terms such that neural networks can extract features while logic modules can reason over structured knowledge

The research should determine whether symbolic knowledge improves recognition or reduces the need for more data when aiming for better performance. Additionally, it would also be interesting to find out whether the integration of symbolic knowledge enhances interpretability. I will mainly be using the FindingEmo dataset.

2 A Circumplex Model of Affect — James A. Russell

Russell's paper presents a psychological model of human emotions called the *Circumplex Model of Affect*, which proposes that emotions can be organized within a two-dimensional continuous space. Affect is the broadest and most general word psychologists use for how people feel emotionally or even physically. The continuous space consists of two axes:

- **Valence (Pleasure–Displeasure axis)**: how positive or negative a feeling is.
- **Arousal (Activation–Deactivation axis)**: how energized or calm the feeling is.

Together, these dimensions form a circular structure (*circumplex*) where emotions gradually blend into one another across valence and arousal, the same way as colors on a wheel. This circular arrangement provides a structured, interpretable framework for representing affect. In the context of neurosymbolic AI, it enables **symbolic encoding**

of affective knowledge (*anger* = unpleasant + high arousal) that can interact with neural perception modules, thus bridging psychological theory with AI reasoning.

In Schlosberg's early studies, participants categorized facial expressions of emotion. Errors in these categorizations revealed that similar emotions were frequently confused, indicating that they were **conceptually close**. From these confusion patterns, Schlosberg proposed a circular representation based on two bipolar dimensions:

- Pleasantness–Unpleasantness
- Attention–Rejection (or Activation–Deactivation)

Emotions, therefore, are better understood as positions along **continuous bipolar scales** rather than discrete, isolated categories. If one feels happy, one cannot simultaneously feel sad; if one feels tense, one cannot feel relaxed. Russell extended this reasoning, showing that any emotion word can be represented as a specific blend of valence and arousal, producing a full circle of affective experience.

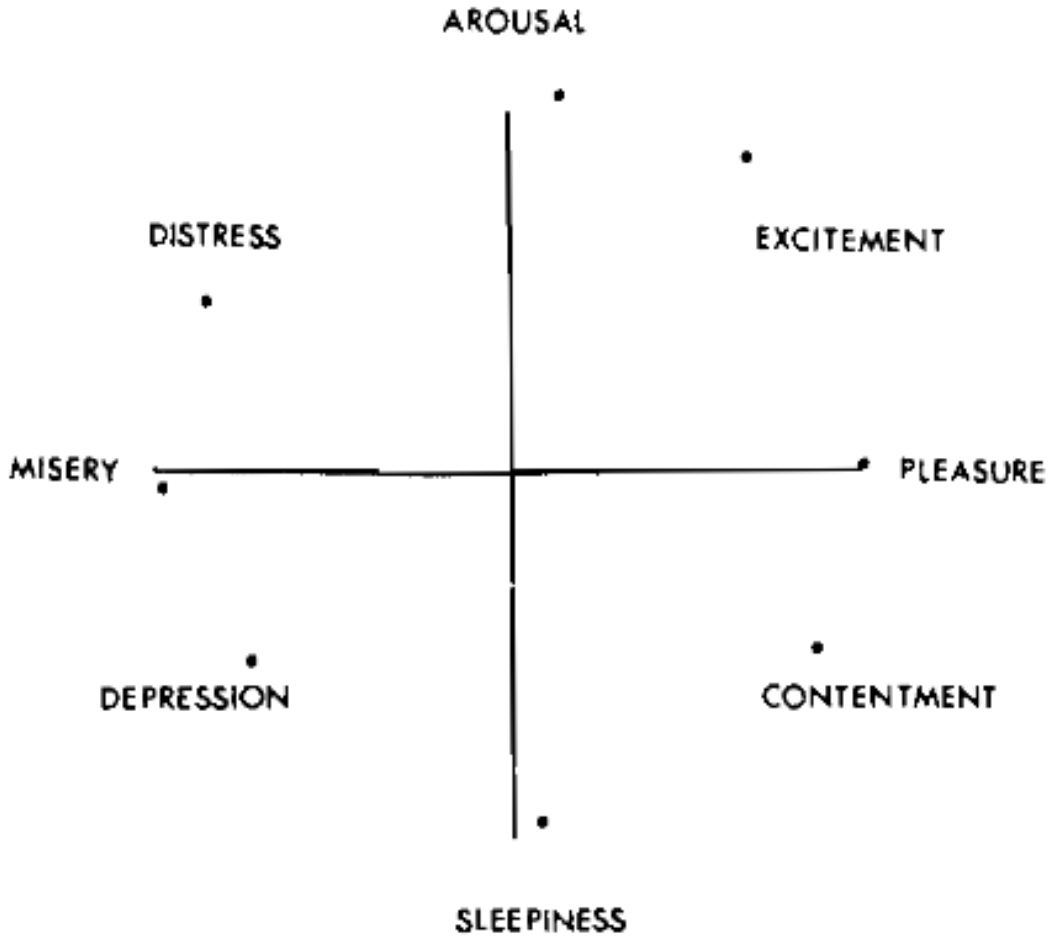

Figure 1. Eight affect concepts in a circular order.

Russell's experiment involved participants judging the relationships among the 28 emotion words. These served as the "points" later positioned within affective space using three techniques:

- **Ross' circular ordering:** Arranges variables along a circle, assuming circularity from the start.
- **Multidimensional scaling (MDS):** Derives spatial relationships purely from perceived similarity, without assuming circularity.
- **Unidimensional scaling:** Rates each term separately on pleasure–displeasure and arousal dimensions, producing two coordinates per emotion.

All three approaches yielded **remarkably consistent results**. Participants did not treat emotions as independent categories. Instead, emotion words such as *happy*, *excited*, and *content* had overlapping meanings, a property Russell called **fuzziness**. This fuzziness explains why affective terms distribute smoothly around a circle: emotions gradually blend into one another rather than having sharp boundaries. Participants' high consistency in word placement confirmed that people share a similar **mental map of emotions** structured around the valence and arousal dimensions.

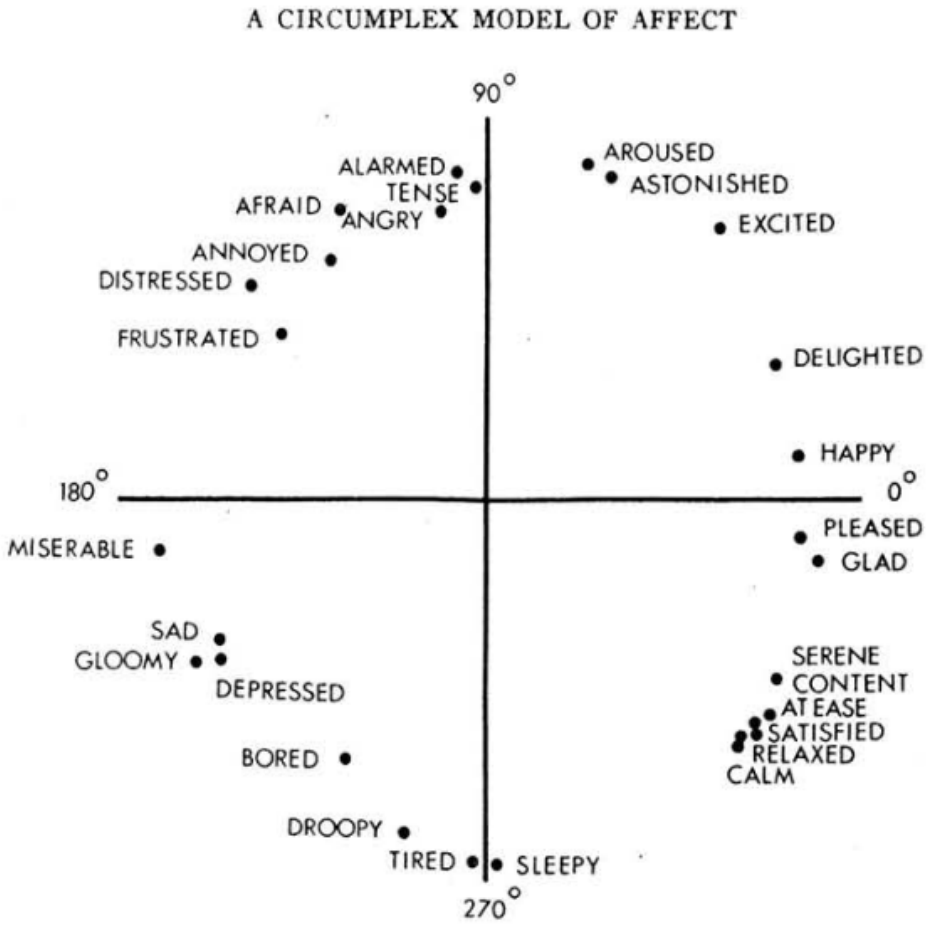

Figure 2. Direct circular scaling coordinates for 28 affect words.

Russell's results supported three major hypotheses about the cognitive structure of affect:

- **Bipolarity:** Opposite emotions (e.g., happiness–sadness) occur approximately 180° apart.
- **Two interpretable dimensions:** The horizontal and vertical axes correspond to pleasure–displeasure and arousal–sleepiness.

- **Continuity:** Rather than forming distinct clusters, emotion words spread continuously around the circle, creating a smooth emotional spectrum.

Earlier psychological models described emotions as 6–12 independent categories. Russell demonstrated instead that affective experience is **interconnected and continuous**, characterized by two bipolar dimensions that explain most of the variance in emotional life.

From feeling to cognition: Russell compared two types of emotion data: self-reports (how people feel) and judgment data (how people think about emotion terms). Traditionally, self-reports were believed to reveal genuine emotional experience, while judgment data were thought to capture only linguistic or semantic relationships. Yet, both data types yielded the same circular structure. This finding implies that **the way people experience emotions and the way they conceptualize them share the same underlying mental framework**.

Emotions are therefore not raw sensations; they are **interpreted experiences**. From a neurosymbolic perspective, this implies that the **symbolic layer of an AI system** can use this cognitive map as a structured knowledge base for affective reasoning. Neural networks may handle perception (detecting facial or scene context cues), while the symbolic layer interprets these signals within the valence–arousal framework, mirroring how humans conceptualize affect.