Master's Thesis

Neurosymbolic AI for Social Cognition

Jarne Demoen
Vrije Universiteit Brussel

4/11/2025

1 Introduction

Humans effortlessly integrate facial expressions, body language, and situational cues to infer emotions and intentions. A raised eyebrow conveys doubt, while a smile can convey happiness or trust. Machines, however, have a harder time interpreting body language, context, and multi-person group structure to derive the overall meaning of a situation. Modern techniques such as machine learning and deep learning can recognize visual patterns with high accuracy, but they do so in a purely subsymbolic manner: neural networks map inputs to outputs without providing insight into how decisions are made. As a result, even if a system predicts the correct emotion of a social scene, it remains unclear *why* the model reached that conclusion. This lack of transparency motivates the need for learning frameworks that support both perception and reasoning.

This thesis explores how neurosymbolic AI can help machines take their next step towards genuine social cognition. Social cognition is the human ability to interpret emotions, intentions, and the subtle dynamics that shape everyday social interactions.

Machines are increasingly expected to navigate human environments. For example, they support people at home, assisting clinicians, mediating online interactions, or working together with humans in shared physical spaces. In all of these settings, an accurate recognition of emotions is essential for safe, trustworthy, and socially aware AI systems. Yet today's emotion-recognition models remain largely pattern-driven. They could detect a smile, but they often miss whether it is a joyful smile at a wedding, a polite smile in a meeting, or a strained smile in a stressful situation. This motivates the need for approaches that integrate perceptual cues with structured, human-like knowledge.

Neurosymbolic AI provides a promising path forward by combining two complementary paradigms:

- **Symbolic AI**, which represents knowledge through a human-readable structure, such as logical rules, for example, "people are usually happy at weddings" or relationships between events.
- **Subsymbolic AI** (deep learning), which excels at uncovering patterns in raw data, such as recognizing facial expression in images.

- **Neurosymbolic AI** which unifies both approaches by using neural networks for perception and symbolic reasoning modules to interpret structured relationships.

In this work, symbolic knowledge refers to contextual cues that describe the social setting, such as the environment type and information derived from multiple people's facial expressions. Rather than solely relying on pixel-level features, the system incorporates a structured representation of what is actually happening in the scene and who is expressing which emotion. This allows us to investigate how relationships between context and group emotional configuration may influence the final emotion prediction.

This thesis addresses the following research questions:

1. Can explicit symbolic knowledge about context and group composition improve the accuracy of emotion recognition in social scenes?
2. Can such knowledge reduce the amount of training data required compared to a purely subsymbolic baseline?
3. Does a neurosymbolic setup make it easier to interpret and explain the model's decisions, for example by inspecting which rules were used?

To study these questions, the experiments use the FindingEmo dataset, which contains naturalistic, multi-person social scenes annotated for valence, arousal, and discrete emotion categories. Its emphasis on contextual and group-based emotional understanding makes it well-suited for evaluating neurosymbolic approaches.

In this thesis, a neurosymbolic model was created using DeepProbLog. Neural networks are used to predict emotions from individual faces, while logical rules capture contextual information and the relations between faces and scene type. DeepProbLog combines these components in a single probabilistic logic framework, which allows us to empirically assess how integrating perception and symbolic structure affects performance and interpretability in emotion recognition.

2 A Circumplex Model of Affect — James A. Russell

Russell's *Circumplex Model of Affect* provides a psychological foundation for modeling human emotions. It organizes emotions in a continuous two-dimensional space with the following axes:

- **Valence** (Pleasure–Displeasure): how positive or negative an emotion is.
- **Arousal** (Activation–Deactivation): how energetic or calm it feels.

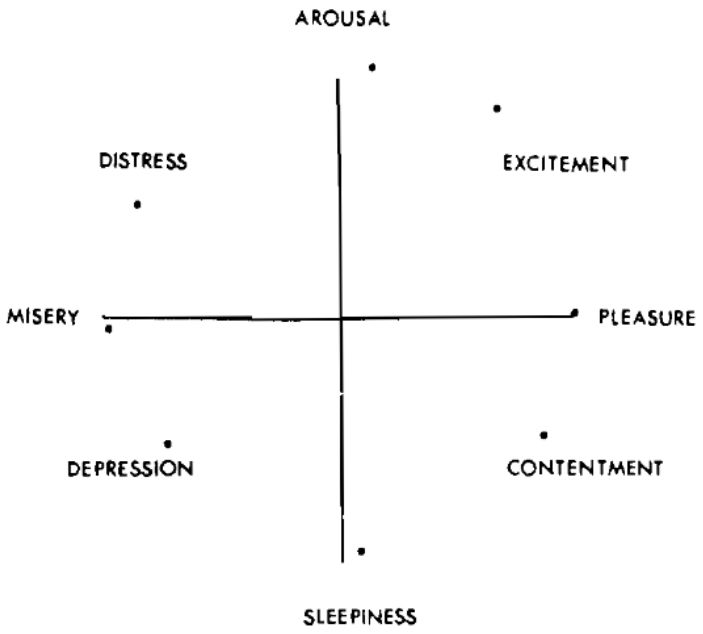

Figure 1: Eight affect concepts in a circular order

Each emotion corresponds to a point on this circle, where emotions gradually blend rather than forming sharp boundaries. Happiness, excitement, and contentment occupy nearby positions, while opposite emotions like depression and excitement lie about 180° apart.

Russell's findings demonstrated that people share a consistent "mental map" of emotions structured along these dimensions. Both people's reported feelings and their judgments about emotion-related words followed the same circular pattern, suggesting that how we feel emotions and how we think about them share the same underlying structure.

2.1 Relevance to Neurosymbolic AI.

This continuous, interpretable model provides a symbolic foundation for affective reasoning. Neural components can predict perceptual cues such as faces, postures and scenes, while the symbolic layer represents these within the valence–arousal framework, allowing the system to reason about emotions the way humans do (anger = unpleasant + high arousal).

3 FindingEmo: An Image Dataset for Emotion Recognition in the Wild (Mertens et al., 2024)

3.1 Overview and Motivation

FindingEmo is a dataset designed for emotion recognition in complex, real-world social scenes rather than isolated faces. Each image contains multiple people engaged in social contexts such as weddings, protests, or family gatherings. Labels include both continuous (Valence, Arousal) and discrete emotion categories based on Plutchik's Wheel of Emotions.

This dataset represents a more realistic view of emotion understanding as a *context-dependent* process: tears at a wedding indicate joy, while tears at a funeral imply sadness. Recognizing emotions, therefore, requires reasoning about context and relationships, which are the key aspects of social cognition.

3.2 Annotation Methodology and Challenges

Annotations combine the Circumplex Model of Affect (for Valence and Arousal) and Plutchik's emotion categories (for discrete labels). Plutchik's wheel of emotions contains 24 primary emotions grouped into eight families, each with three levels of intensity varying from mild to strong. Emotions that are psychological opposites, such as joy and sadness, are placed on opposite sides of the wheel. This structure makes it possible to label emotions at different levels of detail, using either the broader eight categories (Emo8) or all 24 specific emotions (Emo24).

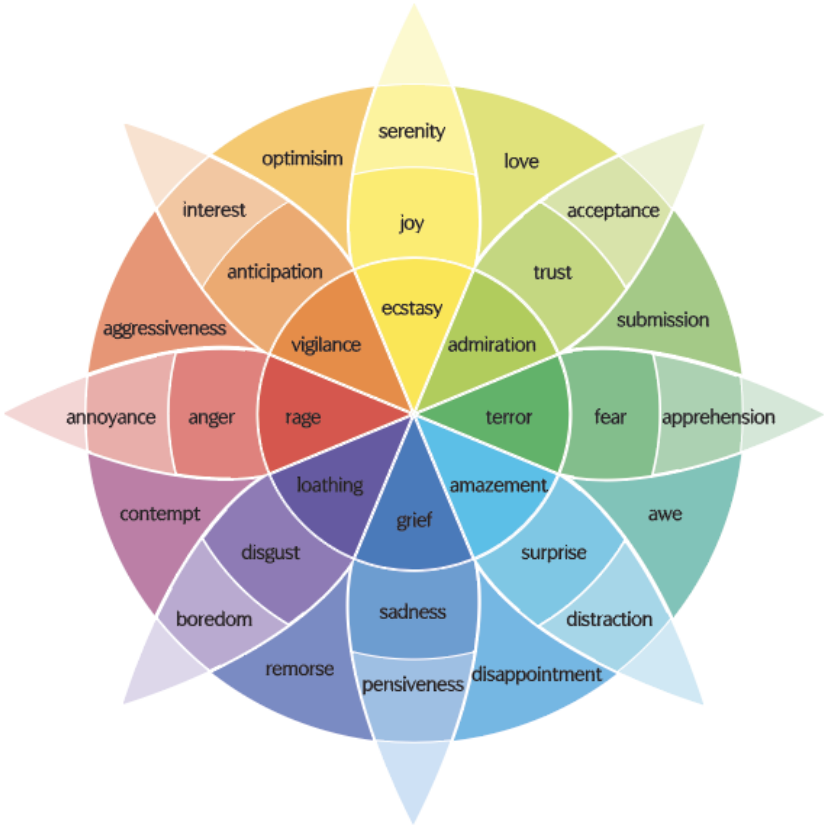

Figure 2: Plutchik's Wheel of Emotions (PWoE)

Reliability studies show stronger agreement on Valence than Arousal, confirming that emotional intensity is more subjective. Such variability is valuable for neurosymbolic approaches, as probabilistic reasoning can explicitly handle uncertainty in human perception.

The dataset exhibits moderate class imbalance: joy and anticipation are frequent, while disgust and surprise are rare. Symbolic priors could compensate for this. For instance, disgust is likely in scenes involving unpleasant stimuli (spoiled food or injury), and fear often appears in threatening contexts (danger, darkness, aggressive postures). By integrating such rules into the reasoning process, the symbolic component could guide

or adjust the neural predictions, improving recognition of these low-frequency emotions even when few training examples exist.

3.3 Baseline Performance and Insights

Transfer learning on CNNs (VGG16, ResNet) and transformers (CLIP, DINOv2) showed that Valence prediction is more reliable than Arousal. Errors frequently occurred between semantically similar emotions such as anger–disgust and joy–trust. This pattern supports the idea that symbolic reasoning could encode adjacency relationships from Plutchik's wheel to refine neural predictions.

Combining facial and contextual features through late fusion yielded only minor gains, revealing that simple feature concatenation does not produce real reasoning. A neurosymbolic system could instead use logical rules to integrate these cues transparently, such as:

```
wedding_scene + smiling_faces → joy.
```

3.4 Relevance to the Thesis.

FindingEmo provides both the perceptual complexity and the affective structure needed to test neurosymbolic models. Neural networks handle visual interpretation, while symbolic logic connects context, emotion, and social conventions, enabling the model to move toward a human-like understanding of collective emotion.

4 DeepProbLog: Neural Probabilistic Logic Programming (Manhaeve et al.)

4.1 Core Concept

DeepProbLog integrates neural networks into the probabilistic logic programming framework *ProbLog*, enabling joint learning of perception and reasoning. In ProbLog, each fact $p :: f$ has an associated probability. DeepProbLog extends this by allowing some probabilities to come from neural networks, known as *neural predicates*. For example, a CNN detecting whether an image shows a cat with confidence 0.9 produces the fact `0.9::cat(image1).` The system thus combines:

- Symbolic reasoning (logical rules and probabilistic inference)
- Neural perception (data-driven probability estimates)

while remaining fully differentiable and trainable end-to-end.

4.2 Indirect Learning and Latent Representations

DeepProbLog supports learning from indirect supervision. In the MNIST addition task, the program encodes:

```
addition(X,Y,Z) :- digit(X,DX), digit(Y,DY), Z is DX + DY.
```

The network is never told the digit labels but learns to represent them correctly so that the logical rule produces the right sum. The logic provides structure; the neural module learns the perceptual mapping that satisfies it. This mechanism of *weak supervision* is particularly relevant to social cognition, where understanding an emotion or social situation often depends on several subtle interacting cues rather than a single, clearly defined label.

4.3 Annotated and Neural Annotated Disjunctions

An **Annotated Disjunction (AD)** encodes probabilistic choices among mutually exclusive outcomes:

$$0.4 :: earthquake(none);\ 0.4 :: earthquake(mild);\ 0.2 :: earthquake(severe).$$

DeepProbLog extends annotated disjunctions to **Neural Annotated Disjunctions (nADs)**, in which the probabilities of the alternative outcomes are provided by a neural network rather than being fixed values. Each nAD connects the continuous, data-driven predictions of a neural model with the discrete reasoning structure of the logic program.

In the MNIST experiment, for example, a neural network m_{digit} takes an image as input and outputs a probability distribution over the digits 0–9:

```
nn(m_digit, Img, [0..9]) :: digit(Img,0); ... ; digit(Img,9).
```

Here, each network output corresponds to a probabilistic fact such as `digit(Img,3)` with probability p_3. The logic engine then treats these probabilities like any other probabilistic facts, allowing the system to reason symbolically while naturally accounting for the uncertainty that comes from neural perception.

4.4 Learning and Differentiability

To train the hybrid system, DeepProbLog employs *Algebraic ProbLog (aProbLog)* with the *gradient semiring*. Each probabilistic fact carries both its probability and the derivative of that probability with respect to learnable parameters. During inference, these are propagated through the Sentential Decision Diagram (SDD). This allows gradients from the final loss to flow backward through both the logic and neural layers, enabling standard gradient descent optimization.

4.5 Application to Social Cognition

In this thesis, DeepProbLog will serve as the reasoning backbone for modeling social emotions. Neural components perform scene classification and facial emotion recognition:

```
nn(scene_net, Img, [wedding, funeral, meeting]) ::
    scene(Img,wedding); scene(Img,funeral); scene(Img,meeting).

nn(emotion_net, Face, [happy, sad, neutral]) ::
    emotion(Face,happy); emotion(Face,sad); emotion(Face,neutral).
```

Logical rules combine these predictions into scene-level emotional reasoning:

```
positive_valence(Img) :-
    scene(Img,wedding), face_in(Face,Img), emotion(Face,happy).

negative_valence(Img) :-
    scene(Img,funeral), face_in(Face,Img), emotion(Face,sad).
```

The system is trained on high-level labels such as `final_valence(Img, V)`, where $V \in \{\text{positive}, \text{neutral}, \text{negative}\}$. Gradients propagate from this final target through the reasoning structure to refine both the neural and probabilistic components.

4.6 Advantages for the Thesis

- Provides an interpretable bridge between neural perception and symbolic reasoning.
- Enables learning from indirect supervision (weakly labeled social scenes).
- Supports probabilistic reasoning under uncertainty (essential for human emotion interpretation).