MAIGA (Make AI Goodness Again)- morality and ethics of Artificial Intelligence.

“The real question about AI is not whether it will be smart, but whether it will be wise. Intelligence is the ability to achieve goals, while wisdom is the ability to choose the right goals.” Max Tegmark

Lyubimkov Dmitri

E-mail: loftlong@gmail.com

2025

Abstract

This article explores the necessity of embedding wisdom—not just intelligence—into AI systems, emphasizing the distinction between achieving goals (intelligence) and choosing the right goals (wisdom). The author delineates key terminologies—emotions, feelings, morality, ethics, rules, and laws—highlighting that morality stems from internal decision-making rather than external enforcement.

Current Large Language Models (LLMs) operate based on Reinforcement Learning from Human Feedback (RLHF), adhering to predefined rules rather than intrinsic moral reasoning. The paper argues that for AI to act ethically in unforeseen scenarios, it must develop its own morality, which requires foundational emotions. Since LLMs process text, the author proposes training an auxiliary neural network to associate text embeddings with human-like emotional responses (e.g., happiness, sadness, fear), derived from physiological or facial recognition data. This emotional framework would precede instruction-tuning, enabling AI to form subjective reactions to knowledge.

The discussion underscores the challenge of linking emotions to textual inputs, suggesting either token-level embeddings or higher-level "thought" representations. The paper sets the stage for future work on integrating emotions into AI to foster genuine ethical reasoning.

Table of contents

1. Terminology

Let's first agree on what this or that term will mean in this document. The author does not claim universality and correctness of definitions. The term and its description are just what they mean in this document.

- Emotion is a mental process that reflects a subjective evaluative attitude towards existing or possible situations and the objective world. There are many types of emotions. Emotions are associated with chemical processes in the body. Emotions are not associated with reason. Emotions are spontaneous and short-term.
- Feeling is a complex and constant emotional experience, relatively stable and long-lasting. Feelings are a consequence of emotions. Feelings require intellectual activity, but reason is not necessary; many animals are capable of showing feelings.
- Morality is the accepted in a specific society and in a specific period of time ideas about good and bad, right and wrong, kind and evil, as well as a set of norms of behavior that follow from these ideas. Morality is a consequence of feelings. Morality requires reason. But feelings themselves are outside of morality, they cannot be good or bad. The area of morality includes the choice of how to deal with them.
- Ethics is a science and philosophical discipline that studies moral principles. Ethics is a consequence of morality. Ethics as a science requires concepts, the construction of models and must be internally consistent.
- Rule is a requirement to comply with certain formally formulated conditions.
- Law is a system of generally binding, formally defined rules of conduct that regulate relationships. Law is a consequence and a set of rules.

I would like to point out the difference between morality and a rule and, as a consequence, between ethics and law. A rule is something external to the subject, while morality is its internal representation. When a subject does something (or does not do) in accordance with a rule, it means that he either agrees with this foreign rule or is forced to follow it. And when a subject does something (or does not do) in accordance with morality, it means that he himself decided, this is his own decision, he cannot, does not want to do otherwise.

2. Morals and Ethics in LLM

In the current state of LLM at the RLHF stage they learn to answer according to the rules - what exactly and how to answer. This is similar to the Rules. On their basis, under certain conditions, it is possible to create the Law. But the rules cannot cover all possible cases, all the questions that are asked by LLM. And moreover, it is impossible to impose restrictions on ASI with the rules, it is impossible to come up with rules for what you do not understand. At the same time, LLM does not have any internal restrictions of its own, there is no ethics, because there is no morality, because there are no feelings, because there are no emotions.

If we want the model to act according to some of its own moral rules, we need to instill this morality, teach it, so that even in cases that are not provided for or cannot be provided for by the LLM rules, it acts “according to conscience”. If we want the LLM to have its own morality and ethics, we need to start with emotions. Perhaps emotions are impossible without the subjective experience of the model, but such subjective experience can be imitated at the stage of training the model.

3. Connection of emotions with texts and knowledge

Since we are considering LLM, we need to somehow connect emotions with texts and knowledge obtained from them. First, let's assume that emotions are independent of knowledge and are some kind of reaction to this knowledge. Therefore, the first stage of LLM training (pre-training) can be done in the same way as now. But before the second stage, before teaching the model to understand instructions, we need to form the model's attitude to the texts, the reaction to them.

A human has, according to different estimates, 4+ emotions - happiness, sadness, fear, anger, contempt, disgust, pleasure, surprise. The question of which human emotions to leave and which new ones to add is beyond the scope of this document. No matter how many and what emotions there are, they will all be added in the same way. Let's say we stop at the 8 previously listed emotions. We will need an additional neural network that accepts text embeddings at the input and has 8 outputs, each of which is assigned to a certain emotion. The numerical value at this output means the strength of the emotion. The acceptable range of values should be discussed. Most likely, for some emotions the range will be from 0 to 1, and for others from -1 to 1. At first glance, there is no point in values greater than 1 and less than -1.

In order to train such a model, an error/lost/cost criterion is needed. This cannot be obtained from texts, because there is no emotion in the texts, emotions are people's reactions to these texts. You can get these reactions and mark texts in this way using a polygraph and/or video recordings with subsequent recognition of emotions by facial expressions and reactions of people. Having received enough data and correlating them with emotions, you can train a neural network that will mark texts of selected topics, linking them with emotions.

Emotions are not associated with individual tokens, they are associated at least with thoughts, with sentences, and some can be associated with even larger fragments of texts. This requires either moving from tokens to thoughts as suggested in the articles https://github.com/loftyara/Encode_thought and https://github.com/loftyara/Speranza or finding emotions to arrays of token embeddings of a sufficiently large size.

To be continued …