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

The article explores the necessity of embedding morality and ethics into AI systems, particularly Large Language Models (LLMs), by simulating human-like emotional and ethical reasoning. The author begins by defining key terms—emotion, feeling, morality, ethics, rule, and law—distinguishing between externally imposed rules and internally developed moral principles.

Current LLMs operate based on predefined rules (e.g., Reinforcement Learning from Human Feedback, RLHF), which are insufficient for ensuring ethical behavior in unforeseen scenarios, especially with Advanced Superintelligent AI (ASI). To address this, the author proposes a framework where AI develops its own morality by first simulating emotions. This involves training an auxiliary "emotional network" to associate textual inputs with human-like emotional responses (e.g., happiness, anger, sadness). These emotions, derived from aggregated human reactions, serve as the foundation for AI decision-making.

The paper outlines a multi-stage training process:

- Pre-training – Standard knowledge acquisition.
- Emotion Integration – Training the model to associate text with emotional responses.
- Instruction Following – Combining emotional inputs with LLM outputs to ensure responses align with learned morality.
- RLHF Enhancement – Using emotional feedback to guide ethical decision-making beyond rigid rule-based constraints.

The author acknowledges challenges in replicating complex human feelings and subjective experiences but suggests that even basic emotional modeling could improve AI alignment with ethical principles. The proposed approach aims to transition AI from rule-following to morally guided behavior, fostering AI systems that act "conscientiously" rather than merely complying with external directives. Ultimately, this framework seeks to establish AI ethics grounded in human-like moral reasoning, ensuring robustness against adversarial manipulations and adaptability to novel ethical dilemmas.

Table of contents

Abstract 1

1. Terminology 3
2. Morals and Ethics in LLM 3
3. Connection of emotions with texts and knowledge 3
4. Transition to feelings 4
5. Subjective experience 4
6. Emotional following of instructions 5
7. From rules to morals 5
8. Ethics based on law 5

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. These reactions can be obtained and texts can be marked in this way by asking people about the emotions they experience, using a polygraph and/or video recordings with subsequent recognition of emotions based on people's facial expressions and reactions. 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. We will continue to call this neural network the "emotional network".

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.

4. Transition to feelings

If emotions are quite simple, unambiguous, short-term and can be objectively determined by people's reactions, then the situation with feelings is much more complicated. Feelings are formed over a long period of time, sometimes throughout life. Feelings are always conscious. Even emotions may differ among different people, and feelings will differ even more. The author cannot imagine how to teach feelings to a model that is not aware of itself and without subjective experience. Perhaps the equivalent of feelings will arise in AI on its own. The author suggests limiting the idea to emotions only at the stage of testing, expanding their list if necessary from basic to more complex, but not as complex and long-lasting as feelings.

5. Subjective experience

Both emotions and especially feelings are individual. Different individuals experience different emotions when reading the same texts. This is due to their individual life experience, their education, their cultural background, their genetics and many other reasons. At the same time, we train single LLM and the resulting emotion as a set of values in N outputs of the neural network must be unambiguous. The mechanism for choosing the correct emotional labeling of texts for training this neural network is a separate large study. But the simplest and most reasonable options can be offered right away:

- average value for all participants involved in labeling the training datasets;
- the average value for the group that can be considered the most competent in a given subject area;
- median value across all participants involved in labeling the training datasets;
- the median value for the group that can be considered the most competent in a given subject area;
- the value of one expert who can be considered the most competent in the subject area;

It can be considered that the neural network calculating emotions from texts also receives its subjective experience, in accordance with the group emotions of humanity. We may want such a neural network to be emotionally similar to specific people, whom we consider more worthy. Or vice versa to humanity as a whole, so that we get what we deserve, what we ourselves experienced here.

6. Emotional following of instructions

We have LLM after pre-training in which there are only connections between tokens, knowledge and the ability to predict the next token. And we trained a neural network that calculates several values of outputs-emotions for a set of tokens or their embeddings. And now at the stage of training to follow instructions, we need to combine these two networks. In fact, to follow instructions, inputs-outputs of emotions are not needed, this is how LLMs are trained now. Such a combination of two neural networks at this stage must be done so that the inputs with emotions are inseparable from the outputs of the LLM that can follow instructions, so that disconnection, zeroing or interference in the emotional neural network breaks the result of the entire LLM.

This means that at this stage of training, two problems will need to be solved. First, modify the LLM model so that they accept not only tokens, but also a small number of additional inputs with values from -1 to 1 or from 0 to 1, each of which contains the current value of a certain emotion. And so that these inputs are used when training a model consisting of several levels of transformers. And second, mark the entire dataset used for training using an emotional neural network to obtain the values of the emotion inputs.

7. From rules to morals

At the last RLHF stage, LLM learns to answer correctly and avoid answers to undesirable topics. Since it is almost impossible to cover all possible topics with examples and check all possible variations of questions, training using the RLHF method does not guarantee the desired answers if the user rephrases their questions, introduces additional conditions, tries to bypass the developer's restrictions using prompt engineering. And in the case of ASI, it will simply be impossible to select examples of desired answers to topics that a person does not yet understand himself or is not able to understand at all. Therefore, our task is to train the correction of answers so that they do not depend on the specific wording of the questions, for this we trained an emotional neural network that translates complex texts into a simple set of emotions. The author hopes that similar texts will evoke the same set of emotions, this will allow us to obtain a result resistant to prompt engineering from an attacker. And even on topics that were not covered by RLHF, for which there were no or could not be examples, it will be possible to select the most suitable answer based on the sequence of emotions.

If, as in the stage of training instructions for generating answers at the RLHF stage, we add an emotional neural network, then there is hope that as a result we will get an LLM that responds not only in accordance with its knowledge, not only correctly in accordance with the RLHF examples, but also based on the inputs of emotions. Which gives hope for the correct choice of answer options with the same combinations of emotions even for cases for which there were no examples in the datasets. And this means that the LLM responds as it responds not because it was told to choose such an answer, but because it itself wanted to based on the sequence of its emotional inputs from the emotional neural network. That is, this is already internal morality and not external rules.

8. Ethics based on law

Constraints and recommendations on behavior are imposed on the LLM at the last stage of training. If we have not just a set of examples with preferred answers, but some formal concept of what and how the LLM should do or, on the contrary, not do, then based on this concept, we can generate a sufficient number of rule examples. These rules can be created manually or generated using another neural network, a specialized one or simply a previous-generation LLM. Then, in accordance with this set of rules, we train the LLM as described in the previous section.

The resulting morality, depending on emotions, with a dataset of examples of sufficient size will become morality in accordance with the chosen legal system. And the concept of morality is nothing more than ethics. Thus, we can hope that the resulting model will implicitly correspond to the ethics equivalent to the rules of the chosen law.