Coarse Ethics: How to ethically tackle the challenge between accuracy and human interpretability

Marin Mato

2024-12-12

Introduction

"Why am I seeing this?" is a question that crosses everyone's mind when we see something deemed as inappropriate for the moment on any type of social media app. Nowadays, platforms like Instagram, TikTok, Facebook, X and more, depend on your recent likes or views to continue to suggest content matching your preferences. The algorithm is tailored to the user's needs, and the more time an individual spends on the app, the more accurate the prediction for the upcoming content will be. That is not unique to these algorithms, as models make better predictions as they are continuously trained. Sometimes the algorithm might suggest new topics to try to keep you scrolling for long, but they might not always end up being exactly what you had in mind. Therefore, the user can easily determine in their head that the assumption from the computer turned out to be wrong and may even click "Not Interested" to ensure that it will not happen again. This decision from AI is low stakes, and there are not really any complications, other than you exiting the app if the algorithm keeps suggesting you the wrong videos. In this particular case, the algorithm does not owe the user an explanation on their wrong decision, as the social media apps are quite transparent on how their models work – stating that all the suggestions are influenced by the individual's behavior on the platform. But what about algorithms that decide much more than your next video, do they owe you an explanation? Do the developers of an algorithm that decides who gets treated in a hospital owe the disadvantaged users an explanation? Even if you are not concerned about the ethics of machine learning, your answer should still be a resounding Yes. Moreover, the explanation of the outcome has to be simple enough for the user to fully understand the rejection while not losing trust in the decision making process of the algorithm. This paper aims to verify and provide an in-depth analysis of the findings from the study conducted by Izumo and Weng, titled Coarse ethics: how to ethically assess explainable artificial intelligence. (Izumo, T., & Weng, Y.-H., 2022)

Analysis of methods

To put it in the shortest way possible, the study concluded by Izumo and Weng revolves around one big question, which may sound simple, but is one of the most challenging topics in the AI industry: "How to avoid oversimplification". To answer this ambiguous matter, the two researchers introduced a new concept titled "Coarse Ethics" that provides two guidelines to ethically simplify the results of an algorithm. The first one, order preservation, states that the ranking of the options included in the original algorithm must not be changed or reversed in the simplified version. For instance, if Option A is ranked higher than Option B in the complex algorithm, their order must remain the same for the simplification process to ethically work out. The second guideline, adequate coverage, is centered around providing a robust inclusion of all the important factors that led to the final decision by the algorithm. In other words, as complex as it might get to explain to the affected user, the developers must include every original factor that was inputted in the AI model to arrive at the outcome. As stated above, this paper fundamentally agrees with the methodology and the theoretical insights of the study. However, to take it a step further, I will test the two Coarse Ethics requirements in order to evaluate their ease of application in my own experiment.

To verify the findings of the study, I implemented a simulation that focuses on reflecting the complexity of autonomous vehicle decisions. Specifically, I modeled an "ethical score" E, that evaluates how ethical a

certain action carried out by an autonomous vehicle is, based on a number of factors, such as legal compliance, passenger comfort, average speed of the car and the pedestrian density of the place. We define the ethical score E as follows:

$$E = \alpha \cdot legal\_compliance + \beta \cdot passenger\_comfort - \gamma \cdot (pedestrian\_density \cdot avg\_speed) + \varepsilon$$

In this equation, the parameters α, β, γ are weighted in relative to their importance in the decision making process of an autonomous vehicle. Out of the 4 factors, legal compliance and passenger comfort are ranked the highest, as they are fundamental in ethically assessing a certain action that the car makes. On the other hand, pedestrian density and average speed are introduced as an interaction variable, with their product being negatively weighted due to the nature of these parameters. For instance, driving slow in a very populated area will not reduce the ethical score that much. However, speeding in that same dense area will result in a much more substantial reduction of the final score. Lastly, the ε is needed to add noise to the final result of the model, with the intention of recreating a real-world scenario by adding a fluctuating variable. The ranking of the variables based on their importance to the ethics calculation is detailed below.

Table 1: Ranking of Variables by Importance in the Ethical Score Calculation

Variable	Reasoning	Relative_Importance
Legal Compliance	Fundamental for lawful and ethical action	Highest
Passenger Comfort	Crucial to ensuring a positive experience for passengers	High
Pedestrian Density	Multiplied by speed to represent risk in crowded areas	Moderate
Average Speed	Influences risk in dense pedestrian contexts	Moderate/Contextual

After computing the ethical score for a set of hypothetical autonomous vehicle decisions, we ranked these decisions from best (highest E) to worst (lowest E). As outlined in the study, explaining the outcome is the part where Coarse Ethics becomes useful. The simplification process for my algorithm involved coarsening these raw scores into different categories, with the intention of making it easier for the readers to understand the results. The first method was splitting the outcomes into three labels, with "High" containing the top 1/3 of the outcomes, "Moderate" containing the next 1/3 and "Low" containing the last 1/3. The second method focused on a more arbitrary way of splitting the outcomes, with "High" having the top 10%, "Moderate" having the next 10% and the "Low" containing the remaining 80%. Lastly, I provided a simpler ranking method, by assigning "Pass" to the top half and "Fail" to the bottom half of the outcomes. In order to assess the efficiency of the simplification of the results, the simulation introduces the concept of inversions – which occur when the order of preservation is violated – either when the decision with higher E is ranked lower or even on par with a decision with lower E due to oversimplification. To address the adequate high coverage requirement, the outcomes were reviewed in the different simplification methods to ensure that all the factors were highlighted in the final ranking. For example, the final rankings must differentiate between a decision taken in a high pedestrian density compared to one in low density.

The results of the simulation align closely with the conclusions taken from the study conducted by Izumo and Weng. The method of dividing the outcomes into even tertiles concluded with the least amount of inversions, and by doing so, it maintained the original ethical hierarchy. This coarsening technique not only preserved the original order, it also provided adequate high coverage, reflected in the rankings. For instance, the influence of speeding in a dense pedestrian area compared to driving at a moderate speed was reflected in the categories, as no such ethical scores were categorized together. As the coarsening techniques changed, the number of inversions started increasing. The second method, which arbitrarily assigned the top 10% to "High", the next 10% to "Moderate" and the rest to "Low", was associated with a noticeably larger number of inversions. Even though the best scores were still distinguishable from the rest, the larger part of the outcomes – 80% – were categorized together regardless of their stark ethical difference. In addition to violating the order of preservation, these categories failed to highlight the importance of each factor when calculating the score. By utilizing this coarsening technique, the decision to speed in a high pedestrian density area was grouped

together with ones that were going slightly over the speed limit. Lastly, the third method took simplification to another level, and so did the violations of the Coarse Ethics. This method was the easiest to understand but the furthest one from the truth. By just grouping the outcomes into two groups, a decision slightly above average ended up in the same category as a decision that maximized passenger comfort, obeyed the traffic laws and slowed down in a high pedestrian dense area. To conclude, simplification, when carried out carefully and by following the Coarse Ethics requirements, results in explanations that the affected users can understand with ease.

Analysis of Normative Consideration

The developers of a certain algorithm can view the ethical shift of the AI industry as a process that is slowing down the advancements in the field. They can feel like the ethical standpoint is blocking their creativity of incorporating machine learning models in every industry out there. However, we must remember the importance of following ethical standards when developing an algorithm that might provide life-altering outcomes for the affected users. Before the in-depth analysis of the philosophical appeal of Coarse Ethics, a developer should consider the public trust as a motivation behind ensuring that the explanations offered by their algorithm are convincing and simple enough for everyone to understand. The majority of the American public still remains uncertain towards the use of AI in everyday life (Faverio, M., & Tyson, A. 2023). However, the only way to make people trust algorithms is by focusing on the explanation of the outcomes. Therefore, Coarse ethics not only provides a framework to tackle the oversimplification issue, but it also outlines a new strategy to further increase the AI industry by incorporating it into our daily lives. That way, when developers ensure that their work follows the ethical standards, they are directly creating more opportunities that need AI solutions in the future.

The rationale behind Coarse Ethics can be seen as an extension of our moral duty to treat individuals with respect and dignity, which would closely appeal to Kantian ethics. Kant's ideology suggests that we must never use people merely as means to an end, but rather always treat them as ends in themselves. When an algorithm's explanatory process follows the Coarse Ethics requirements, by preserving the original order of outcomes and including all critical factors that shaped the decision, it respects users by granting them access to the moral reasoning embedded in the model. Instead of presenting an oversimplified version of the outcome with an inadequate explanation, the developers abide by a moral duty to be honest, thorough, and fair. This faithful representation ensures that those affected by the AI's decisions are not misled or undervalued, but are given the moral courtesy of transparency and understanding.

By adhering to Coarse Ethics, the model's developers acknowledge that the explanations they provide carry ethical weight. Reducing complex considerations to extremely simple categories or ignoring key factors would violate the essence of a Kantian framework, which demands that moral agents do not sidestep their obligations, even if doing so might make communication simpler. In other words, Coarse Ethics is not just a technical guideline, but a moral commitment: it prevents the commodification of moral decisions into mere data points, ensuring that every step of the simplification process remains answerable to the moral laws outlined by us.

Even if simplifying decisions makes them easier to understand, such simplifications must never come at the expense of moral duty. In my simulation, if the original algorithm ranked one option as more ethical than another, the simplified version must preserve that order. Due to the simplified explanation, the user is not only better informed but also treated as a moral stakeholder whose capacity to reason should never be undermined. Thus, through the Kantian lens, Coarse Ethics ensures the AI's simplified explanations remain aligned with our binding moral duties, upholding the inherent worth every individual affected by the algorithm's decisions.

Conclusion

The question presented at the top of this paper now has a response that should be used as a starting point for the ethical shift of the AI industry. Coarse Ethics, a concept introduced by Izumo and Weng in their study, is an effective yet responsible method to avoid oversimplification of an algorithm's result with the intention of providing an adequate explanation to the affected users. The consequences of oversimplifying are clear –

the public's trust will slowly fade away due to the decreasing accuracy of algorithms. Therefore, simplifying outcomes responsibly and carefully is a moral obligation for every developer. By insisting on preserving the original order of outcomes and including all critical factors in the explanation, Coarse Ethics reaffirms the moral duty of developers to treat users as fully informed stakeholders rather than passive recipients of non-transparent decisions. Coarsening techniques do not hinder innovation; rather, they lay the groundwork for building sustainable public trust and ultimately advancing the AI industry's integration into everyday life.

References

Izumo, T., & Weng, Y.-H. (2022). Coarse ethics: How to ethically assess explainable artificial intelligence. AI and Ethics, 2, 449–461. https://link.springer.com/article/10.1007/s43681-021-00091-y#Sec18

Faverio, M., & Tyson, A. (2023, November 21). What the data says about Americans' views of artificial intelligence. Pew Research Center. https://www.pewresearch.org/short-reads/2023/11/21/what-the-datasays-about-americans-views-of-artificial-intelligence/