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

Marin Mato

2024-10-25

"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. Sometimes the algorithm might suggest new topics to try to keep you scrolling for long, but they might not always end up being your cup of tea. 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.

This paper will analyze the relevant terms and experiments introduced in "Coarse Ethics: How to Ethically Assess Explainable Artificial Intelligence" by Takashi Izumo and Yueh-Hsuan Weng, with the purpose of coming to a deeper understanding of the AI practices in the industry that might not always be ethical. Moreover, I will discuss the process and the results that the two researchers came up with, outlining their statistical approach and methodology (Izumo, T. & Weng, Y.-H, 2022). Due to the rapid development of the AI industry, algorithms do not only decide the next video or post that you will see – they also make decisions that can impact someone's access to health care, financial services like loans, and even legal rights. The higher the stakes of a certain decision, the more detailed the explanation to the affected user should be.

One of the main concepts introduced in the paper is XAI or Explainable Artificial Intelligence. Just by looking at this label, one can think that it has to do with the user completely understanding the details of a certain algorithm. That is not quite the case here, as this term defines the explanation provided to the affected group that is based on the algorithm's decision. For instance, XAI provides transparency for a person whose request got rejected for a loan from a bank that utilizes algorithms to speed up their processes. In a perfect world, the general public would want the developers to only work on models that are extremely transparent and provide simplified explanations on their decision-making, championing the concept of fairness and not uncertainty. However, this is where the limitations of XAI come into the picture, as in many cases, simplifying an algorithm to provide clear explanations on the outcomes might lead to compromises in its accuracy. To assist this claim, the study mentions self-driving cars, where accuracy has to be a priority over simpliciation, in order for the passengers to have a safe trip to the destination.

But why should the developers care about this matter? Why can't they just continue to increase accuracy without worrying about transparency? The reason is pretty straightforward – the majority of the Americans are more concerned than excited about incorporating AI into their daily lives (Faverio, M. & Tyson, A. 2023). If you want people to trust you on something, you have to be transparent throughout the process. The fundamental challenge of balancing accuracy and interpretability is the main issue that Izumo and Weng are worried about, and to tackle this, they introduced the concept of Coarse Ethics or CE. They believe that CE is the correct approach towards ethically simplifying algorithms to the point that the outcomes are not compromised by a significant amount.

According to the researchers, we evaluate the simplification based on two requirements, providing adequately high coverage and preserving the order of importance. Firstly, adequate high coverage relates to the framework of including every key factor used during the decision making process in the explanation. The user has to know all the important details that were assessed by the algorithm to arrive at a certain outcome. For instance, a user was denied a loan by a bank that utilizes algorithms to make more precise decisions. If the explanation for the rejection tells the user that their financial status was the reason that they were not selected, the individual will think that the algorithm is biased and unfair. However, if the explanation depicts in detail the factors that led to the rejection, such as insufficient income or credit score, the user will understand the problem better. In other words, simplifying is the main goal, but over-simplifying leads to misleading results that can raise concerns on the fairness of a certain algorithm.

The other requirement for an adequate explanation for the affected user is order preservation, meaning that the relative ranking or preference order of the evaluated options must remain consistent with the original AI model. For instance, if Option A has been ranked higher than Option B in the original algorithm, their orders cannot be reversed when trying to coarse correctly. The study provided a great example to make this requirement more understandable. The experiment involved students taking an exam that was graded out of 100 points. To evaluate the scores, they provided three different methods, with E1 ranking precisely on how many points were earned, E2 ranking with categories like 'fair', 'good', 'very good' and 'excellent' and E3 ranking with 'Pass/Fail'. In our terms, E1 is the complex algorithm that we are trying to simplify into two options, E2 and E3. If student A scores 83 and student B scores 82, A will be ranked higher in E1. However, when we switch to E2, they will both be under 'very good' and with E3, they both pass. Here is where order preservation is crucial, as the simplification of E1 cannot lead to B being ranked higher than A in E2 or E3, just because they fall under the same category. It is permissible for them to be considered at coarsely equivalent 'A approx. B', but claiming 'B>A' in the simplified version will result in violation of order of preservation. All in all, simplifying is encouraged under these two conditions that must always be met in order to ensure accuracy in the explanation of the decision, according to the researchers.

After providing a clear outline of coarse ethics, a reader might still fail to understand the real-world complications of oversimplifying. The consequences might be as simple as just reversing the order of two students that took the same exam, like above, but they can also involve much more harmful situations. To further explain how important it is to ethically assess every simplified algorithm before it is implemented in the real world, I will rely on another experiment conducted by Izumo and Weng. Today, even though self-driving cars are not available to public use yet, several newly produced cars contain features that assist a driver in certain situations. The researchers used the GenEth system, which evaluates whether the car's AI algorithm should take control of the car or leave it up to the driver in certain situations. It is pretty understandable that keeping the driver and their surroundings safe is the main goal of every self-driving system. They are implemented with the sole purpose of aiding the driver in crucial moments, sometimes even determining whether the car is involved in an accident or not. As mentioned above, a high stakes decision needs a clear and convincing explanation. Thus, these systems are programmed to only step in when it is needed the most in order to avoid a harmful situation. With this in mind, the researchers are wondering if simplifying a complex algorithm, such as a self -driving system, might be unnecessary at times.

In other words, while simplifying AI decision-making processes enhances explainability, it can also lead to unethical outcomes if not handled carefully. For instance, a self-driving car facing an imminent collision with a pedestrian must choose between swerving abruptly (Action A) or maintaining its course (Action B). During the experiment, the system assigns a higher ethical priority to 'preventing harm to humans' over 'passenger comfort,' leading it to select Action A. However, if developers simplify the AI's reasoning by reducing the complexity of its ethical evaluations, the relative importance of these factors might shift. This oversimplification can elevate 'passenger comfort' above 'preventing harm,' causing the system to choose Action B instead. Such a reversal violates the order-preservation requirement of CE and results in an unethical decision that endangers lives.

Furthermore, this simplification may also violate the requirement of adequately high coverage. By reducing the complex evaluation to just a few factors, the algorithm might exclude essential duties, such as 'obeying to traffic laws' or 'minimizing overall harm,' from its decision-making process. This lack of comprehensive coverage means the algorithm's simplified reasoning does not fully represent all the significant ethical factors

originally considered, leading to inaccurate decisions. This experiment underscores the critical need to preserve the ethical priorities of the complex algorithm when simplifying, ensuring that explainability does not come at the expense of ethical responsibility. Adhering to both the adequately high coverage and order-preservation requirements of Coarse Ethics, outlined by the researchers, is essential to maintain ethical integrity and prevent harmful outcomes in real-world applications.

In response to the title, which is the question that was raised before starting to write this paper, the challenge of balancing ethics while simplifying the explanation for users is much more difficult than it might sound at first. One can argue against simplifying, as it might be considered as a tool to halt or slow down the rapid advancement of the AI industry. This train of thought revolves around the idea that developers should strictly worry about accuracy, as every person who is disadvantaged by the algorithm would argue that it is unfair. However, as mentioned above, the main goal down the line is for the general public to incorporate AI into their daily lives out of necessity, without being concerned at all. Millions of people already use ChatGPT, but imagine all of us using other forms of AI to make life easier with full trust in the algorithms. Thus, advancements in the field are important, but transparency will make AI more acceptable globally, which would aid with expanding the industry even further.

In conclusion, the ethical assessment of Explainable Artificial Intelligence is a fundamental need in today's technological landscape that keeps advancing rapidly. By defining and exploring Coarse Ethics, Izumo and Weng have highlighted the delicate balance between simplifying complex AI algorithms for human interpretability while maintaining ethical integrity and accuracy. The two fundamental requirements—adequately high coverage and order preservation—serve as essential guidelines to ensure that simplification does not compromise the ethical standards of AI decision-making processes. The examples discussed, from the ranking of students' exam scores to the critical decisions made by self-driving cars, illustrate the real-world implications of neglecting these requirements. Oversimplification can lead to misleading outcomes which not only undermines the accuracy and fairness of AI systems but can also result in harmful consequences, endangering lives and eroding public trust in technology. Therefore, it is crucial for developers and stakeholders to assess the ethical dimensions of AI models, especially when simplifying them for explainability to users.

Adhering to the principles of Coarse Ethics ensures that while AI becomes more accessible and understandable to humans, it does not do so at the expense of ethical responsibility. As AI continues to integrate into various facets of our lives, maintaining this balance will be essential in fostering technologies that are not only innovative but also transparent and trustworthy.

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-data-says-about-americans-views-of-artificial-intelligence/