# Moral Crossroads: A Machine Learning Engineer's Dilemma

*Adnan Azmat (G2303265K)*
*MSc Artificial Intelligence, NTU Singapore*

"There is no such thing as a free lunch." This famous phrase, attributed to economist Milton Friedman, captures the essence of trade-offs: the idea that every choice has a cost and a benefit, and that one has to weigh them carefully before making a decision. Trade-offs are ubiquitous in computer science, where different factors often have to be balanced against each other. For example, an algorithm may have to trade time for space, a software project may have to trade quality for delivery speed, or an AI model may have to trade complexity for interpretability. However, in artificial intelligence, one of the most critical and difficult trade-offs an engineer faces is the one between integrity and mercenary motives.

Faiza, a machine learning engineer, who is the lead of her group, is well aware of the theoretical trade-off between bias and fairness in AI. She works in a rising social media company catering to a young audience. However, while working with ML models and diverse data, she has to balance a few conflicting aspects: obligation for good performance and metrics, timely delivery, model efficiency, data quality, ethical standards and social impact, giving rise to difficult choices. While working on a problem of content recommendation, she notices that the team's model can be potentially dangerous once in deployment. The recommendation engine recommends videos of the same type and more intensity, based on the user's behavior. Since the company is new, the guard rails on content and recommendation are few, and users may fall prey to harmful content. For example, users may incline towards extreme content, develop addictive tendencies and unrealistic expectations or even suffer from mental health issues. She estimates that the users potentially getting affected is likely to be a small percentage. She wonders whether she should report the problem and risk delaying the project or whether she should ignore the problem and risk harming the users. How can she resolve this trade-off and what are the consequences for her and the society?

The dilemma that the AI engineer faces is not a new one. It is a manifestation of the classic tension between utilitarianism and deontology, two normative ethical theories that prescribe how one should act in a given situation. Utilitarianism, as advocated by philosophers like Jeremy Bentham and John Stuart Mill, holds that the right action is the one that maximizes the overall well-being of the greatest number of people. Deontology, as proposed by philosophers like Immanuel Kant and John Rawls, maintains that the right action is the one that follows certain universal moral principles, regardless of the consequences. Both theories have their strengths and weaknesses, and both have been applied to various domains of human activity, including AI.

In the context of AI, utilitarianism and deontology can offer different perspectives. A utilitarian AI developer may argue that the content recommendation engine is justified, because it provides a positive experience for the majority of the users, who enjoy the personalized content. The utilitarian AI developer may also claim that the harm caused to the minority of the users is negligible and it can be compensated by several means like providing warnings or feedback. A deontological developer, on the other hand, may contend that the

engine is unethical, because it violates the rights and dignity of the individual users. The deontological developer may also assert that the harm caused to the minority of the users is unacceptable, or that it cannot be outweighed by the benefits for the majority of the users.

Faiza recognizes the merits and drawbacks of both the ideologies but is unsure which one to follow. She has been brought up with the idea that one should always do the right thing, even if it is difficult or unpopular. She believes that the company should invest time to give guardrails to save the few users who may be affected by the content recommendation engine. Faiza feels a moral responsibility to report the problem. However, she knows that the company is under pressure to deliver the project on time and to gain a competitive edge in the market. She fears that reporting the problem may delay the project, damage the reputation of the company and jeopardize her career.

After considerable deliberation, she decides to go ahead with shipping the model to production, without reporting it. She convinces herself that the users potentially getting affected is likely to be a small percentage who can always choose to opt out or switch to another platform. She tries to believe that she is doing the right thing for the company and the society, by providing a service that enhances the user experience. She refuses to see that she is contributing to the perpetuation of bias and unfairness in AI. The company values profit over ethics, and thus she feels herself accountable to comply with the expectations and deadlines of her superiors. Faiza feels powerless but tries to suppress her guilt .

However, a few days after the model goes into production, her conscience finally ignites and she reports the problem. She also adds a guardrail feature on the recommendations to the pile of backlog. She hopes that the product team will find this feature important enough to take up in the near future. She soon realizes that this is unlikely, and that the problem will persist for a long time. She decides to take matters into her own hands, and during the company hackathon, she fixes the issue with the help of two interns. She works tirelessly for a week, limiting the model's harmfulness and improving its fairness. There is no telemetry to tell if any user was affected during this period, but she feels a sense of relief and satisfaction. Her hard work does not win her any prizes, but she doesn't care. She is happy that she has done the right thing and restored her ethical integrity. She can finally go to sleep with contentment in her heart.