# AI ethicist's dilemma

We present what we refer to as the **AI ethicist's dilemma**, which emerges when an AI ethicist has to consider how their **own success in communicating an identified problem is associated with a high risk of decreasing the chances of successfully remedying the problem**.

We want to examine how the ethicist can resolve the dilemma and arrive at ethically sound paths of action through combining three ethical theories: virtue ethics, deontological ethics and consequentialist ethics.

> **Assume that a researcher uncovers major problems related to the increased usage of AI and AI based social media. When researchers discover such problems, what action should they take?**

This situation routinely occurs as AI ethicists navigate the choices involved in deciding how to effect change and remedy the problems they uncover.

*For example, if their research emphasised the grave and unfortunate consequences of Facebook, should they promote this research by building communities on said networks?*

The dilemma also extends beyond social media, and entails the ethicist's relation to the major technology companies in general.

*If they write a book about the dangers of Amazon, for example, should they promote the book by posting links to the book on* `amazon.com` *?*

*And should they seek work opportunities at, for example, Google or OpenAI if they are deeply concerned about the negative environmental and social implications of large-scale language models?*

These examples relate to what we refer to as the ethicist's dilemma, which emerges when an ethicist has to consider how their own success in communicating an identified challenge is associated with a high risk of decreasing the chances of successfully facing the challenge. **This dilemma occurs in situations in which the means to achieve one's goals are seemingly best achieved by supporting that which one wishes to correct and/or practicing the opposite of that which one preaches**. It is thus a proper moral dilemma.

One way to approach the problem is to follow an ethical decision making process. This process consists of the following steps of:

1. formulating the moral problem,

2. analysing the problem,
3. considering one's options,
4. ethically evaluating these options, and finally
5. reflecting and arriving at a morally acceptable action.

We can base the analysis on a **fundamental moral pluralism** and the dilemma in question is examined in light of all three major ethical theories: virtue ethics, deontological ethics, and consequentialist ethics.

## The moral problem

The problem area here discussed encompasses the negative consequences of new technologies, with a particular emphasis on artificial intelligence and Big Data. AI based social media will be a case given particular attention, but the dilemma in question also applies to other AI based technologies.

> Q1: Make a list of issues related to the phenomena here discussed (AI based social media).

Assuming that the issues above either are real, or that they are at least perceived to be real by a hypothetical AI ethicist warning against them, the question becomes: What should they do about it—if anything?

In the following attempt to answer this question, two assumptions are introduced to narrow the scope of the analysis and clarify what sort of situations and potential ethicists are covered by the dilemma.

- The first assumption is that the AI ethicist in question has a desire to effect change, and feels a certain duty to do so, once they discover non-trivial challenges caused by AI and its applications.
- The second assumption is that the challenges in question are non-trivial.

To sum up and simplify the moral problem:

> AI, and AI based social media in particular, negatively impacts individuals and society in non-trivial ways, and the AI ethicist feels obligated take action to counter these impacts.

## The ethicist's strategy

The next step in the ethical cycle is to consider the avenues of action available to the concerned AI ethicist.

For clarity and to facilitate a straight-forward analysis, we reduce the strategies available to two: (a) working from within the system and (b) working from without.

- **Strategy 1**: Seeking change by allying with Big Tech or actively using their technology and infrastructure.
- **Strategy 2**: Seeking change through distancing oneself from and marginalizing the sources of the problems and allying with other sources of power to effect change.

## Evaluating the ethicist's options

The purpose of an ethical cycle is to arrive at an answer to the question of what someone should do (i.e., which of the two strategies outlined above should they adopt).

Answering what the ethicist should do, without implicitly smuggling in our own ethical inclinations and philosophical underpinnings, thus necessitates the explicit use of the mainstream ethical theories most often used to guide and evaluate ethical behaviour: virtue ethics, consequentialist ethics, and deontological ethics. Seeing the dilemma in light of these three ethical theories helps us disentangle some of the considerations involved in deciding how to act when facing the dilemma.

This approach is, consequently, based on a **moral pluralism** premised on the idea that **insight from analyses based on all three theories are required to understand the ethics of any action**.

### Virtue Ethics

Virtue ethics involves focusing on what characterizes the ethical person. Rather than focusing on the consequences of actions, or a set of duties or rules, the virtue ethicist emphasises the virtues associated with a moral character, as seen in relation to what is conducive to ethos. Rather than focusing directly on what to do, the virtue ethicist considers what sort of person one should be. Certain character traits, such as courage, honesty, and benevolence are considered virtues conducive to moral flourishing. Basic AI virtues could be: justice (that corresponds to Algorithmic fairness, non-discrimination, bias mitigation, inclusion, equality, diversity), hobesty (that corresponds to organizational transparency, openness, explainability, interpretability, technological disclosure, open source, acknowledge errors and mistakes), responsability (that corresponds to liability, accountability, replicability, legality, accuracy, considering (long term) technological consequences), and care (that corresponds to non-maleficence, harm, security, safety, privacy, protection, precaution, hidden costs, beneficence, well-being, sustainability, peace, common good, solidarity, social cohesion, freedom, autonomy, liberty, consent).

From a virtue ethics standpoint, a crucial consideration for the AI ethicist relates to the problem of not practicing what one preaches. Doing so might both (a) be wrong in itself as it is not conducive to a good life, and (b) decrease the chances of successfully effecting

change, because one is perceived to be hypocritical (a moral vice; not a virtue). The latter points towards the consequences of not being virtuous, and can thus be argued to belong to the domain of consequentialism, discussed below. However, all virtue ethical evaluations have a tinge of consequentialism embedded in them, as one is concerned with discovering what results in a good life.

### Consequentialist ethics (utilitarianism)

As the name implies, consequentialist ethics is focused on the consequences of our actions. According to a popular version of consequentialism—utilitarianism—what should guide our actions is an evaluation of what will create the greatest amount of happiness for most people. The utilitarian can live with the dilemma here discussed if the means (use of the system) provided the chance to achieve the ends (change of the system). However, they will also have to grapple with the possibility that using the system will, as argued below, end up strengthening the system.

### Deontology

Deontology is often associated with Kant and the idea that we should check whether the actions we consider could be turned into universal maxims that guide the actions of all—not just ourselves in a particular situation. A deontologically inclined ethicist who has just written a book about the moral rules we ought to follow to challenge the power of Big Tech, for example, should in theory abide by the same rules themselves, unless some universalizable version of the rules in question justify other avenues of action.

## What should the ethicist do?

This section addresses the use the ethical theories to arrive at a choice between the two available strategies.

Q2: What are the potential and limitations of strategy 1 when considering it from a moral pluralistic point of view?

Q3: What are the potential and limitations of strategy 2 when considering it from a moral pluralistic point of view?

Q4: What is the optimal choice (Strategy 1, Strategy 2, Other)?

Q5: How did you scored different options?

Reference: Sætra, H.S., Coeckelbergh, M. & Danaher, J. The AI ethicist's dilemma: fighting Big Tech by supporting Big Tech. AI Ethics (2021).