pymetrics

# Designing Artificial Intelligence With Ethical Principles at the Forefront.

No matter how powerful, technology is neutral.

Electricity can be used to kill (the electric chair) or save lives (a home on the grid in an otherwise inhabitable climate). The same holds true for artificial intelligence (AI), an emerging layer of technology that makes countless things possible.

AI systems already exist that help and harm humans. For example, a group at UCSF recently built an algorithm that helps to save lives through improved suicide prevention, while China has deployed facial recognition AI systems that subjugate ethnic minorities and political dissenters. It is therefore erroneous to assign benevolence to AI broadly -- it depends entirely on how it's designed. To date, that's been largely careless.

AI systems started to boom with **major tech companies** that, in order to provide their product at no cost to the general public, had to find ways for their AI to be profitable. They did this by selling ads. Advertising has long been in the business of manipulating human emotions in order to drive sales. Big data and AI merely allowed this to happen more e ectively, insidiously, and at greater scale than ever before.

AI mishaps, such as Facebook's algorithms being co-opted by foreign political actors to influence election outcomes, could have been predicted from this careless application of AI. Many have highlighted the need for more thoughtful design, including AI pioneers like Stuart Russell, the "father of AI," who now argues that "standard model AI" should be replaced with beneficial AI.

Organizations ranging from the World Economic Forum and Stanford University to the New York Times are gathering their respective groups of experts to develop design principles for beneficial AI. As a contributor to these initiatives, pymetrics' CEO, Dr. Frida Polli, believes the following principles are key to using it exclusively for good:

## 1. Allow Users to Holistically Understand Data Collection

The user must know that data is being collected and what it will ultimately be used for, so technologists are responsible for ensuring informed consent. Far too many platforms across a wide host of applications rely on surreptitious data collection or user data collected for other purposes. Fortunately, initiatives to put this to a stop are cropping up everywhere, as seen with the Illinois law requiring that video hiring platforms inform users that AI may be used to analyze their video recording and shed light on how the resulting data will be used.

## 2. Support User Data Privacy And Ownership

Users must own and control their data. This is counter to the prevailing status quo of many tech companies whose terms of service are designed to exploit user data for the sole benefit of the company. For instance, a tool called FaceApp, a viral app that shows people what they will look like 50+ years down the road, is collecting millions of user photos, without disclosing what data they are collecting and for what purpose. More alarmingly, this user interface does not address the fact photos ever leave the user's local storage. Users must be empowered, not overpowered, by technology and should always know what data is collected, for what purpose, and where it's collected from.

**For instance, a tool called FaceApp, shows people what they will look like 50+ years down the road, is collecting millions of user photos, without disclosing what data they are collecting and for what purpose.**

## 3. Use Unbiased Training Data

AI must be trained with unbiased data as much as possible. This means removing, as much as possible, not only demographic variables but also proxy variables, or those variables that are correlated with demographics. If we do not take care to mitigate biased data, the machine learning systems will be sure to replicate and amplify the bias in the system. Developers thus bear the responsibility to closely examine the data they feed into the algorithms so they do not perpetuate any known bias.

For example, data gleaned from résumés is notably biased against women and minority groups, so we ought to leverage other types of data in hiring algorithms. The San Francisco DA's office and Stanford created a "blind sentencing" AI tool, which removes ethnic info from the data leveraged in criminal-justice sentencing.

## 4. Audit Algorithms

**The San Francisco DA's office and Stanford created a "blind sentencing" AI tool, which removes ethnic info from the data leveraged in criminal-justice sentencing.**

It's not enough to train AI with unbiased data. A math quirk referred to as Simpson's paradox demonstrates how unbiased inputs at the population level can yield biased results in smaller subpockets. It is essential to check your algorithms for bias. Don't let skeptics mislead you. It is possible to audit an algorithm's results for unequal outcomes across gender, race, age, or any other axis where discrimination could occur. The first step of an audit can happen internally - developers of algorithms should be continually testing their algorithms to know if they are violating the bias thresholds. The second step should be some form of external validation of that process, most likely a third party or external audit. An external AI audit serves the same purpose as safety-testing a vehicle to ensure it passes regulations before being put into production. If the audit fails, the design flaw responsible must be identified and then removed.

## 5. Aim for Full Transparency

White-box AI means there is full transparency around the data going into the algorithms and subsequent outcomes. You can only audit an algorithm to reconfigure its potentially-biased output if it has been white-boxed. There is often a trade-off between explainability and performance. However, in fields like human resources, criminal sentencing, and healthcare, among others, explainability may always win out over pure performance because transparency is essential when technology impacts people's lives and well-being. If your model isn't fully transparent, look to open-sourced methods to help partially explain decisions.

## 6. Use Open-Source Methods

Open-source methods should be utilized, either by releasing key features of the code as open-source or leveraging well-established and peer-tested existing code. The visibility it can provide allows for quality assurance.

With the case of algorithm auditing, the process by which companies are auditing (i.e., safety-testing) their algorithms must be well understood. Initiatives to open-source this auditing technology are already in the works.

## 7. Include External Councils to Establish Guardrails

An active community of industry leaders and subject-matter experts should be involved in solidifying the rules of engagement for designing new AI ethically and responsibly. An open discussion can offer a full account of the various implications of AI technology as well as specific standards to follow as a broader community to ensure fairness, transparency, and well-being for all. In addition, changes in policy and legislative efforts will become increasingly important in both safeguarding the public against discriminatory uses of ML/AI while also ensuring that this fledgeling industry flourishes. When the 1974 Fair Credit Billing Act limited cardholder liability, the credit card industry responded by investing in the security of this new technology, dramatically spurring innovation and increasing public trust. We believe that similar legislative efforts will be critical in both using AI to promote non-discriminatory use of AI while also increasing rapid adoption of AI tools.

## Pursue the Positive Impacts of AI

As history has proven, innovation often incites fear and initial failure of usage. However, with the proper design and guardrails, innovation can be harnessed to bring about unprecedentedly positive impacts on society -- and the case of AI is no different. With careful forethought and deliberate efforts to push back on human bias, AI can be a powerful tool not just to mitigate bias, but to actually remove it altogether in a way that is not possible with humans. Imagine life without electricity: a world of darkness. Let's not deprive ourselves of the positive impacts of ethical AI.

**If you're interested in continuing to explore applications of ethical AI in the workplace and beyond, we'd love to chat! Please connect with us here.**