Dr. Humera Noor Minhas

Director of Engineering | eyeo

Implementing Ethical AI

Lots of questions and a few answers

★ W3C TPAC 2023, Seville, 11–15 September ★

My e-bike accident

We have the ML guidelines available:
https://www.w3.org/TR/webmachinelearning-ethics/

It starts with our values

SENSE
PRINCIPLES
WORTHINESS
FAITH
MOTIVATION
RESPONSIBILITY
INTEGRITY
HONESTY
VALUES
THEORY
SOCIAL
EQUALITY
RESPECTABILITY
TREATMENT
TRUTH
HYPOCRISY
ETHICS
CONVENTIONALITIES
PHILISOPHY
PERFORMANCE
MORALS
TRUST
RULES
INNOCENCE
AUTHENTICITY
CHARACTER
ETIQUETTE
RESPECT
HONOR
FAIRNESS
CRITERIA
DECENCY
VIRTUE
STANDARDS
PURPOSE
ATTITUDE
ACCEPTANCE
FAIRNESS
CONDUCT

Putting theory into practice

THEORY
PRACTICE
THEORY INTO PRACTICE

W3C Operationalization Guidelines

Putting the Principles into Practice

- ICO* AI and data protection risk toolkit
- Risk assessment
- Workshop Templates

ICO: UK’s Information Commissioner’s Office

Open Questions

- Workshop(s)?
  - How many? Who? When?
- AI Ethics: Engineering or a Product problem?
- Ethics by design: how-to?
- Data collection: consent?
- Sustainability?
- Open source vs closed-software?
- Compliance with Current regulations? Upcoming regulations?
- Privacy and security?
- Plan for a “kill switch”?

Discovery

Marty Cagan's product risk dimensions

Impact Mapping

Team Analysis

DEI and Ethics Trainings / Workshops

Is this problem best solved with AI?

How does AI add value to our users?

Do we have the skills to build an AI product?
- Machine Learning Engineer
- Machine Learning Researcher
- Applied ML Scientist
- Software Engineers
- Data Engineer
- Data Scientist
- Product Manager
- Product Designer

Did my team get training on bias reduction, fairness and ethical principles?

What kind of impact are we creating?

How might we:
- build trust?
- foster mutual respect?
- build connection?
- affirm life?
- create wellbeing?
- create belonging?
- nurture collaboration?
- be regenerative?
- strengthen humanity?
- center unity?
- add joy?
- nurture resilience?
- build empathy and compassion?
- encourage vulnerability?
- be ok with paradox?

Analysis

AI Business Model Canvas

Code of Ethics

Data Sheets for DataSets

Is there non-biased training data?

Did creators consent to the use of their data for training purposes?

Does the output compete with or economically harm the creators of the training content? If so, did they consent to the use of their data?

What is the environmental impact of building the model / product?

Are there compliance aspects of the envisioned AI solution? e.g. EU AI Act

What social and societal impact might this product have?

What might "it goes horribly wrong" look like?

What are the ethical guardrails we want to have around AI in product?
(e.g. no harm to human life, no harm to mental health, no unfair bias, no social profiling, no cruelty, etc...)

Development

Model Cards

Input / Output Validation

What is the main use case of my product and will this involve decisions impacting people's lifes?

Can this be used to restrict access to opportunities?

Can this be used to cause harm to human life?

Did we plan for a "kill switch" in case we need to roll back? What is our Plan B feature set in case we need to roll back the AI solution?

How can we restrict use of the product for unintended use cases?

What part of our code can be open-source? Can we think of ways this product can cause harm in the wrong hands?

What disclaimers do we need to consider for the UX of our product?

Is our output clearly labelled as AI generated to limit user confusion?

How can we make the data usage transparent?

What updates to our terms of use become necessary as a result of using AI in the product?

Introduction

Review Meeting every 4-6 weeks including Customer success and compliance officer

Monitoring of Input / Output Validations

Feedback mechanisms

How is our product actually being used?

Are there unintended use cases that cause harm?

How do we positively contribute and add values to our users? How might we build on that?

Can we now see use cases of our product that:
- reinforce bias?
- cause harm to human life?
- restrict access to opportunities?
- contribute to social profiling?

Is our product still in compliance with all current regulation around AI in products?

What new regulations will take effect in the coming 6-12 months that we have to consider?

How do we solicit input from the general public on their concerns and ideas around the use of our product?

What is a seven-generation impact scenario of the use of our product?

Do we need to roll back?

Do we need to restrict access?

Do we need to update user agreements?

Can users and team members safely raise red flags?

Growth & Maturity & Decline

Quarterly Review of Ethical questions

Update of Impact Mapping

How is this contributing to the positive impact we want to see, and can we build on that?

What new Technologies have entered the market that could cause harm in combination with our product?

Do we need to further restrict access?

Is there evidence of usage of our product that violates our ethical guardrails? (If so, what can we do?)

Do we need to roll back or block access?

Is it still viable to solve the user problem with AI, or are there simpler and more cost effective ways to solve this problem?

Do we need to update usage agreements?

Post Mortem

Post Mortem Analysis

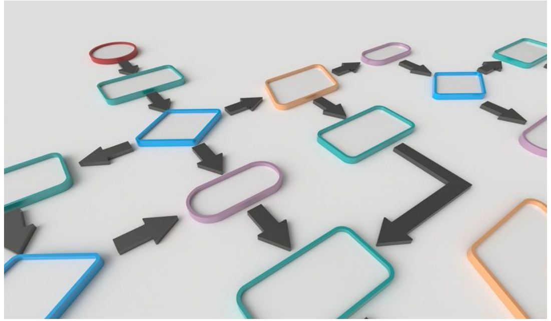
Assignment of post-mortem custodian

Is there value in keeping the model accessible, or is it better for humanity to delete the code?

Recommendations

Share Knowledge

Define Ownership



Establish Processes

What's Next?

- Breakout session:
  - Ethical AI in a Rapidly Evolving Landscape - deep dive
  - Ecija - 1st floor
  - Tomorrow, 12:15–13:15 CET
- Chat:
  - https://irc.w3.org/?channels=%23ethical-ai
- Zoom:
  - https://w3c.zoom.us/j/89661535795?pwd=a3poTktTeFk5YkJTVTlNZlo0OXkyQT09
- Beyond TPAC:
  - h.minhas@eyeo.com
  - https://www.linkedin.com/in/humeranoor/