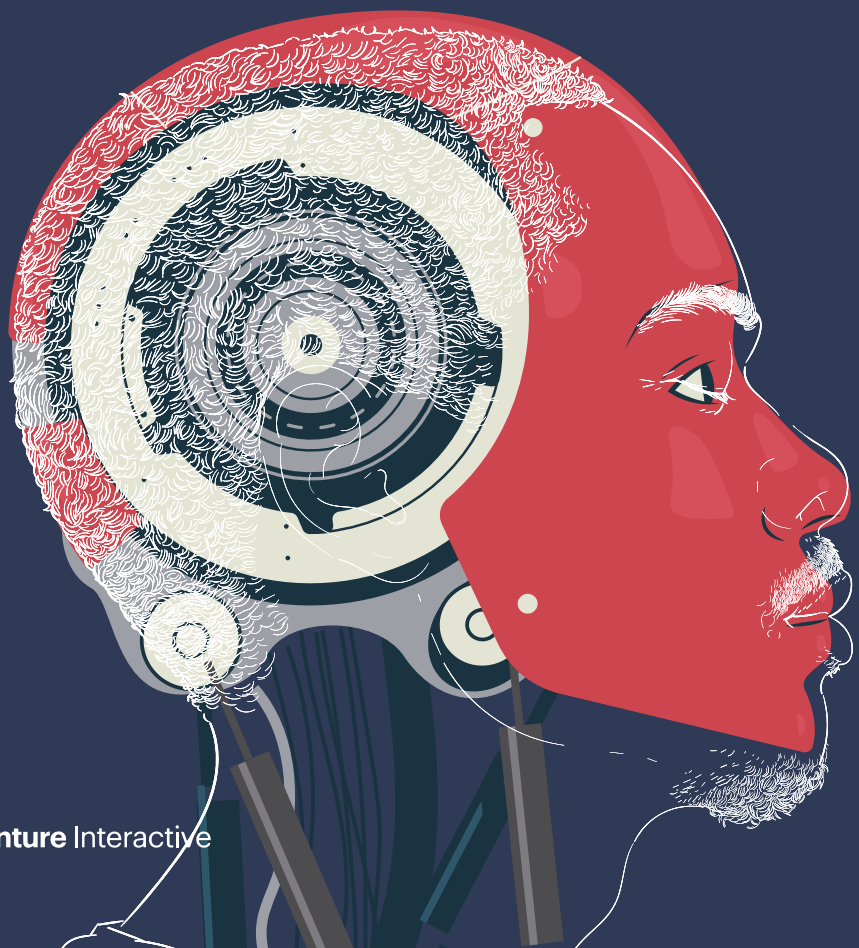
Ethical Toolkit for the Development of AI Applications

Toolkit Instructions

Mario Alberto Sosa Hidalgo

2019



MOBGEN | Accenture Interactive

TU Delft

Author

Mario Alberto Sosa Hidalgo
mariososadi@gmail.com

August, 2019

** The cover was designed using illustrations from Freepik.com

Table of Contents

Introduction & How to Use the Strategic Blueprint

This ethical toolkit, in the form of a "full-day" workshop, assists in the generation of ideas and supports dialogue for an ethical development of AI applications. It provides a basis for discussion at the start of any project that integrates AI into a digital application. The main idea of this toolkit is to trigger solutions and communicate the topic of AI ethics to development teams and clients in a creative and collaborative fashion. The outcome would generate the implementation of responsible ideas for the creation of more ethical AI applications.

1. How to use the Ethical Strategic Blueprint

In order to make the ethical toolkit tangible and useful for the company, a full-day workshop setting was chosen. This workshop, which is presented in the toolkit as a Strategic Blueprint, features two important ethical stages.

First, it is important to create an uniform understanding of ethics and its main approaches. Because of this, an "Ethics & [A]I" mini-workshop is proposed. This workshop concludes with the implementation of a game called "The Evil in [A]I" that aims to trigger ethical alignment within the development team and, at the same time, enhance discussion around the topic of ethical consequences of AI. This first stage is recommended to be performed alongside the client, making everybody aware of their ethical views and enhance an alignment on the ethical understanding of the project.

In addition to this, a second stage where the ethical aspects regarding the development of the project are explored and addressed. This stage begins with the creation of a project overview checklist by the development team. If is followed by the ideation of ethical risks and opportunities by a set of trigger cards. It ends with the identification and mapping of the risks for its evaluation on an Ethical Axis. Finally, as part of the second stage, the strategic blueprint ends with away of translating the main insights of this analysis to ethical specifications that should be followed during the rest of the project (Project's Moral Code). Each stage is described in a deeper level in the next sections, including the instructions to follow to take the most out of them.

STRATEGIC BLUEPRINT

Ethical Toolkit for the Development of AI Applications

Educating Phase (Team & Client)

Executing Phase (Project Team)

1 Ethical Alignment

Ethics & [A]I Mini-workshop

The ethical alignment workshop is a strategic feature of the toolkit that aims to educate and ethically align both the client and the team involved in the AI application project. The deck explores the topic of ethics and its relevance in an AI context.

The Evil in [A]I Game

This module, which has the format of a game, explores the current ethical situation of AI in an extreme way. It aims to detonate a change of perspective around AI and to strengthen the knowledge acquired during the "Ethics & AI" workshop.

2 Project Vision & Values

AI Project General Checklist

The AI Project Checklist is a detailed and pictorial way of discussing the new AI application, thinking about the stakeholders involved, the type of AI models used, the ethical principles to be considered for the development team and the client.

The Responsible Artificial Intelligence Deck

The Responsible Artificial Intelligence Deck, is a deck of cards intended to trigger ideas regarding the ethical risks that could occur with the development of an AI app. Furthermore, this deck is designed to help AI Applications designers and developers to think about the Ethics of AI and the impacted stakeholders in a more visual way.

Ethical Risks Cards

The Ethical Risks Cards objective is to write down the possible ethical risks ideated using the Responsible AI Deck that could be evaluated later on using the Ethical Evaluation Axes canvas.

Ethical Evaluation Matrix

Based on the scale of the axes, or on areas of the canvas that are important for the client, it would be decided which specific ethical risks should include in the decision-making processes of the project. This gives a a holistic view of the ethical risks of the implementation and the principal ethical areas to take into consideration.

3 Moral Code or Best Practices

Moral Code of the Project

By defining the ethics of the project, the team will integrate the most relevant risks in a reflection manner for an AI application. This is an structured way of to culminating the ethical strategy by making concise moral agreements and tangible statements.

Facilitating an "Ethics & [A]I" Mini-workshop

1 - 1.5 hrs

1. Introduction

As we are entering deeply into a completely digitalized era, smart systems are becoming part of our lives. Many commercial intelligent applications are becoming ubiquitous and some are getting implemented in higher levels of decision making (i.e. AI judges), therefore, it is important to understand and discuss the ethical consequences of this "digital transformation" process.

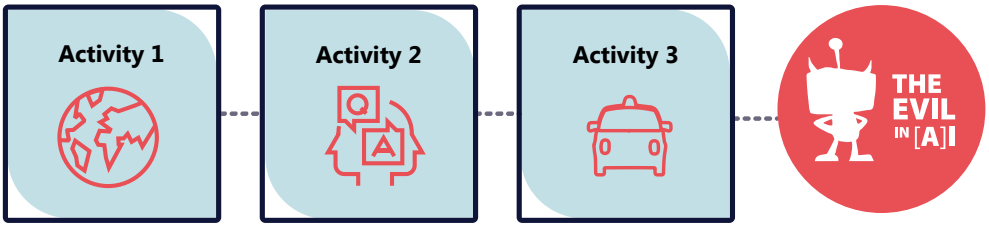
The "Ethics & [A]I" workshop is an educative effort to trigger dialogue and support critical innovation culture at MOBGEN | Accenture Interactive. With the information provided in this workshop you can facilitate a session to teach a client, designers, developers, and managers about ethics and its different approaches. This workshop is intended to encourage the team working an AI related project to think critically and to integrate ethical considerations into their process.

This workshop is part of the Ethical Toolkit created by Mario Sosa as a part of his graduation project from the Masters Degree of Strategic Product Design at the Industrial Design Engineering Faculty of the TU Delft. For more information, please visit the TU Delft repository to obtain a copy of the original thesis.

2. List of Materials

- Presentation Deck
- Ethical Approaches Card (print slide #24 of the Presentation Deck)

3. Workshop Timeline



4. Details of the Workshop

What is this?

This is a workshop on ethics explained by using an AI context. It is intended to support teams that are currently working on an AI-related project. The workshop takes a total of 60-90 minutes which includes the introduction of an energizer in the form of a game in the end.

What are the Learning Goals

1. Learn the three classic normative ethical approaches (deontological, consequentialist, virtue)
2. Understanding the importance of ethical decision making.
3. Reflect on the topic of the ethical considerations of the development of AI applications.

A piece of advice

Dare to reach out and invite others to the workshop. This way, more people would be involved in the ethical decision making, which brings diversity and ethical robustness to AI projects.

Recommended Setup

In order to have a successful experience with the workshop, make sure you can get a suitable room, for example, by using a beamer to project the Presentation Deck.

How to facilitate?

Just go through the presentation deck and try to lookout for terms and concepts that you might find complicated. The workshop doesn't need to be perfect, just good enough so others could understand the concepts and the importance of the topic.

5. Activity 1: Worldview

Time:

15 minutes

Material:

Post-Its, Markers

Purpose:

Use this activity as an opportunity to introduce yourself and the rest of the participants. This icebreaker will also make your participants to familiarize with Artificial Intelligence in an empathic manner (imagining they are the autonomous car).

Instructions

1. Tell the participants to answer the question on the Presentation Deck. "**If YOU were an autonomous car, where in the world would you like to go?"**
2. Ask the participants to introduce themselves and explain the reason why they selected their answer.

6. Activity 2: Quick Ethical Dilemmas

Time:

15 minutes

Material

None-needed

Purpose

This activity lets the participants familiarize themselves with the concept and the emotions related to the ethical decision making process by using Ethical Dilemmas.

Instructions

1. Make groups of 2.
2. Tell the participating teams to decide upon one of the two solutions (A or B) for the first Ethical Dilemma proposed. The solutions are not open to discussion, each team should chose only one with the amount of information supplied. The workshop facilitator must be sure that each team select their preferred option by giving some time for discussion among its members.
3. After that, ask each team to tell their answers and explain the reason of the selection (give some time for discussion)
4. Do steps 2 and 3 again but using the second Ethical Dilemma (use the same teams).
5. Explain the reasons why Ethical Dilemmas are extremely hard.
6. Present the definition of Ethics.

Best Friend's Big Day Dilemma

Your best friend is getting married in an hours time. You are already all fancy dressed up for the wedding when you discover that yours best friend's partner has been having an affair. You have concrete evidence to prove their guilt.
What would you do?

A. Tell your best friend, even if you ruin the wedding, to avoid them marrying to a cheater.

B. Say nothing since your role there is to be supportive with your friend's happiness.

6. Activity 3: Trolley Problem

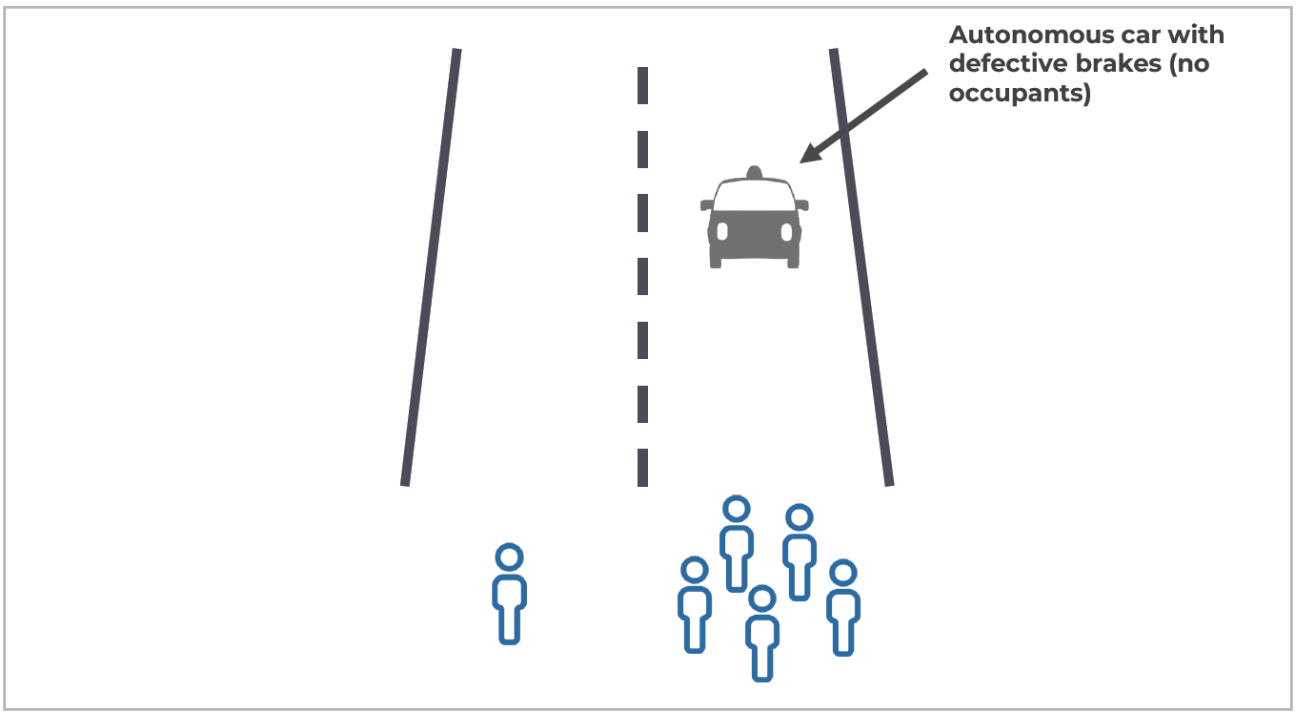
Time
20 minutes

Material
Ethical Approaches Card (print slide #24 of the Deck)

Purpose
This activity lets the participants familiarize themselves with the three main normative ethical approaches.

Instructions
1. Make groups of 2.
2. Read the first variation of the trolley problem
3. Ask each team to discuss which would be the most moral action according to them and then make them share it with the rest of the teams.
4. Repeat step 3 for each of the remaining trolley problem cases.
5. After the final discussion session, introduce the main ethical approaches by using the cards.

Tips
It is important to acknowledge and encourage the different ethical views that people might have during this exercise. Moreover, stimulating discussion around the topic would enhance an ethical mindset. Don't worry if you don't know some of the things that participants might ask. The whole idea of this exercise is to enhance ethical discussion and argumentation.



The “Evil in [A]I” Game

0.5 - 1 hrs

Players:

2-5

Time:

30 minutes

Material:

AI Cards, Evil Cards.

Instructions

1. Shuffle the “AI Cards” and randomly select one per player.
2. Then, each participant should randomly pick a card in the same way, from the “Evil cards” deck.
3. After this, all the players have to show and loudly read the combination they’ve got (even if it seems like a “non-compatible” combination, for example, a sexist facial recognition software).
4. Afterwards, everybody should vote about which participant has the most unethical example of all (from the combinations).
5. After the discussion is done, the winner selected would gain one “Evil point” and should (see Score Card next). The first participant to collect 4 Evil points is the winner. All the card configurations should be mentioned and written down on post-its after each round has passed.

SCORE CARD

THE EVIL IN [A]I

AI App Project Checklist Instructions

1 - 1.5 hrs

The AI project checklist is an exercise intended to make the team involved in the development of an AI application to discuss the implications of the project. The team should fill in the features that the AI application would have, as well as the ethical principles that should be taken into consideration for the project. This can be done by using sticky notes before write down in the actual canvas. It is important to mention that the selected characteristics of the project could be changed later on in the process. The output is a project checklist sheet with the initial specifications and ethical considerations of the AI application.

Section 1 - Goal

The team would write down the goal of the AI application project

Section 2 - Team Members

In order to create the feeling of accountability, this checklist includes this section where all the people involved in the development of the AI application should write their names.

Section 3 - Context of the Project

In this section the context of the project should be discussed and stated.

Section 4 - Stakeholders Impacted

The stakeholders impacted in the project are explored in this section. A “primary” stakeholder is indicated in the first column, as it is expected that the team could define an “unintended” stakeholder related to the former one (i.e. Children - Parents) and should be written down in the second column. It is advised to use the Responsible Artificial Intelligence Deck to get inspired about the possible stakeholders that can be impacted with the applications.

Section 5 - Type of Impact

In order to generate empathy and to assess the ethics of the project, in this section the team needs to select the type of impact that the application might generate for the stakeholders involved.

Section 6 - Type of AI model and Algorithm

In this section, the team can discuss the AI technology used for the application.

Section 7 - Data Type

Data is a vital part of an AI application as well as for data regulations, hence, it is important to know the type of data that the application would use.

Section 8 - Data Source

It is well known that a big source of bias in AI systems come from the data. Because of this, it is important to elaborate on the sources of the data sets that are going to be used for the AI application. It is important to also know if the data sources are GDPR compliant or are universally trustworthy.

Section 9 - Ethical Principles Checklist

This section establishes the ethical considerations to take into account at the beginning of the project. The section features the main ethical principles around AI and it is intended to let the development team sort these principles depending on the level of importance they have for the team. It is advised to use the Responsible Artificial Intelligence Deck to get inspired.

AI APP PROJECT CHECKLIST

Name of the Poject

1 Goal

2 Team Accountable for App

Client	*MOBGEN | Accenture Interactive*

3 Context of the Project

e.g. AI application for healthcare

4 Stakeholders Impacted

Primary	*Unexpected/Unintended*
E.g. Business People	*Eg. Partners, Customers, other businesses*

5 Type of Impact

The AI application will have one or more areas of impact. e.g. AI-poered restaurant reviews impact reputation and financial health.

- ☐ *Financial*
- ☐ *Emotional*
- ☐ *Access to goods*
- ☐ *Property*
- ☐ *Reputation*
- ☐ *Life / Safety*
- ☐ *Privacy*
- ☐ *Liberty / Freedom*
- ☐ *Rights / IP*

6 AI Algorithm & Learning Model used

7 Data Type

Human Data

Non-Human Data

8 Data Source

	GDPR compliant	*Trustworthy*
	☐	☐
	☐	☐
	☐	☐
	☐	☐

9 Ethical Principles Checklist (Order by relevance for project **High(+) / Med(0) / Low(-)**)

- Data Privacy
- Data Safety
- Explainability
- Transparency
- Honest Communication
- Accountability
- Fairness
- Value Alignment
- Human Well-Being
- Governance
- User Safety

The Responsible AI Deck

1 - 2 hrs

1. Introduction

The Responsible Artificial Intelligence Deck, is a deck of cards intended to trigger ideas regarding the ethical risks that could occur with the development of an AI app. Furthermore, this deck is designed to help AI Applications designers and developers to think about the Ethics of AI and the impacted stakeholders in a more visual way. It also features a "Risk/Opportunities" card, which is intended to trigger ideas regarding how to make an ethical principle an opportunity instead of only looking ethics as "compliance" in case something goes wrong. This relatable and playful manner to face the responsible development of AI is expected to prompt discussion and alignment among the members of the development team. Hereby there is a suggested way of using the deck, however, due to its versatility it is expected that more ways of use are discovered in the future.

2. Suggested way of using The Responsible AI Deck

1. Look through all the cards in the deck carefully in order to get familiar with the four different types of cards and its content.
2. Be sure that you count with an adequate space to place the cards like, for example, on a big table.
3. Select the specific AI Context card that adequates to the context of the project. Use the empty cards if necessary to write the context down if it is not included in the deck. Place this card in the middle of the table.
4. Discuss and select all the stakeholders that the team determines are going to be impacted by the AI application. Make sure to include the stakeholders that were indicated in the Project Checklist. Place the selected

stakeholders on the left of the AI Context card.

5. Organize the AI Ethical Principles cards considering the sorting criteria used in the Project Checklist and place them on the right of the AI Context card.
6. Individually, try to identify any ethical problems using this graphical overview about the project and write them down into the Ethical Risk Cards.
7. Afterwards, organize the generated risks in the Ethical Evaluation Axes.
8. In case you need more inspiration, select any card from the Human Values cards that the team considers interesting or necessary.
9. Use the empty cards to include a stakeholder, ethical principle, or human value that hasn't been considered in the Deck.

3. Alternative way of using the Deck

1. Look through all the cards in the deck carefully in order to get familiar with the four different types of cards and its content.
2. Make the participants make groups of 3 or 4 participants
3. Each group will randomly select one card of each type and place all 4 of them on the table.
4. Each group should start a conversation regarding possible triggered ethical risks that might occur.
5. Fill in the Ethical Risk Cards in order to have a record of the ideas generated.
6. This mode is intended for sparking the conversation around the topic with a client for example.

4. Risks/Opportunities mode

1. Look through all the cards in the deck carefully in order to get familiar with the four different types of cards and its content.
2. Organize the AI Ethical Principles cards considering the sorting criteria used in the Project Checklist and place them on the right of the AI Context card.
3. Use the Risk/Opportunities card to find the ethical risks associated to the selected ethical principles, and also the ethical opportunities that these could trigger. For example. In a context of AI applied to the Workplace, a big risk regarding fairness is the possible bias caused by gender gap issues. On the other hand, an opportunity to solve this issue would be to promote a more equitative culture within the workplace using the AI application. Another example would be if the application being developed could also help to predict and prevent harassment inside the workplace environment. You can use the Ethical Risk Cards to describe the Risks and Opportunities discovered through this method.


Stakeholders
Elders
AI Context
24
Customer Service
AI assists with this in a number of areas: by providing answers to simple customer questions 24/7, sorting messaging, prioritising queries to a human agent quickly where necessary, and by freeing up human agents to facilitate higher value calls.
AI Ethical Principles
Accountability
Who is responsible of a bad output from the AI system?
Companies should be accountable for the possible mistakes their AI applications make, as well as to provide strategies for such failovers.
Human Values
Support
To help someone emotionally or in a practical way.


AI Ethical Principles
Data Privacy
Is the AI dataset compliant with GDPR?
AI companies should address the privacy considerations regarding user data as an input or output of any smart system. Some regulations might apply (i.e. GDPR)
AI Ethical Principles
GDPR
Data Safety
Is the AI dataset GDPR compliant?
Developers should ensure that the data used to train the AI model won't produce harmful consequences for any of the stakeholders involved in the project. Some regulations might apply (i.e. GDPR)
AI Ethical Principles
Explainability
Can it be explained why the AI system did certain action?
Every AI system has to be explainable enough for people to understand the reason why it made certain action.
AI Ethical Principles
Human Well-Being
Is this AI system intended to enhance humans or automate jobs?
To what extent an AI application is made to help humans improve their actions and not only to automate processes.
AI Ethical Principles
Honest Communication
Are you clear, and honest enough about communicating the real capabilities of the AI system?
AI companies should be honest, straightforward, and clear about the capabilities and restrictions of their AI applications
AI Ethical Principles
User Safety
Here is this AI system safe for humans?
Every AI application should be safe enough for all users, stakeholders involved in their context of use, or simply any person interacting with it.
AI Ethical Principles
Value Alignment
How aligned are the developers' ethical views with those of the stakeholders impacted?
Developers should align their ethical views with those of the stakeholders impacted with the AI system.
AI Ethical Principles
Transparency
How does the system works?
AI systems should be transparent enough to show users the way they work and how they perform to make certain actions.
Risks
Opportunities
AI Ethical Principles
Accountability
Who is responsible of a bad output from the AI system?
Companies should be accountable for the possible mistakes their AI applications make, as well as to provide strategies for such failovers.
AI Ethical Principles
Fairness
Is this system ethically fair? (bias)
Developers must ensure that their AI applications are as unbiased as possible. Strategies for fairness failures should be addressed during the development of the AI applications.
AI Ethical Principles
Governance
What is the level of autonomy of the AI system?
Developers must address the level of autonomy that an AI application would have and which bodies would control (or monitor) it when it is delivered.

Instructions for the Ethical Risk Cards & the Ethical Evaluation Axes

1 - 2 hrs

1. How to use the Ethical Risk Cards

The Ethical Risk Cards should be filled individually by all the members of the team depending on the amount of ethical risks that have been identified using the Responsible AI Deck. The instructions are as follows:

1. Using the Responsible AI Deck, identify and write them down in the card the AI context of the project, the impacted stakeholders, and the ethical principles that are taken in consideration.
2. In the last section, describe the identified ethical risk as clear as possible using either words or sketches.
3. Write as many different ethical risk cards as possible. It is also possible to write the identified risk in sticky notes before using the Ethical Risk Cards.

2. Suggested way of using The Responsible AI Deck

The evaluation canvas features a couple of axes where the "y" axis is for the level of "Impact" the risk could have for all the stakeholders involved. The "x" axis is for the "Likelihood" of the risk occurring. To perform the ethical evaluation follow the next steps.

1. Place the ethical risks cards on the "x" axis first by qualitatively assess the likelihood of the risk to occur.
2. It is important to enhance some discussion regarding the reason of the positioning.
3. After this was done, he other axis should be discussed and positioned as well based on the impact that could represent for the stakeholders defined in the Project Checklist.
4. Define the ethical "red areas", where each one or both the likelihood of occurring and the impact for the stakeholders are high.
5. Define the ethical "gray areas", where there is a medium level of likelihood or impact and/or where there is no consensus towards how to ethically evaluate the risk.

Based on the scale of the axes, or on areas of the canvas that are important for the client, the decide which specific ethical risks should be included in the decision-making processes of the project. With this exercise the team can get a holistic view of the ethical risks of the implementation and the principal ethical areas to take into consideration. It is important to mention that the evaluation axis can be referenced and modified throughout the project.

ETHICAL RISK CARD

AI Context: REAL STATE

Stakeholder(s) Affected

POOR PEOPLE

AI Ethical Principle(s)

FAIRNESS

Risk Idenitified

THE SMART ALGORITHM COULD DISCRIMINATE PEOPLE BASED ON THEIR SOCIAL AND FINANCIAL SITUATION, THIS MIGHT CREATE A BIAS TOWARDS SOCIAL HOUSING

ETHICAL EVALUATION AXES

Project's Moral Code Instructions

1 - 2 hrs

This module is intended to stimulate an ethical responsibility for the project. By defining the ethics of the project, the team will integrate the most relevant risks in a reflection manner for an AI application. This is an structured way of to culminating the ethical strategy by making concise moral agreements and tangible statements. The output of this module is a moral agreement that could be translated to implementable product specifications for the AI application. It is of vital importance to mention that this Moral Code should not be used as a justification for unethical behavior. The ethics of the project should align with the ethics of the context where the project is being developed. Remember that according to Popper, "in order to maintain a tolerant society, the society must be intolerant of intolerance".

1. How to use it?

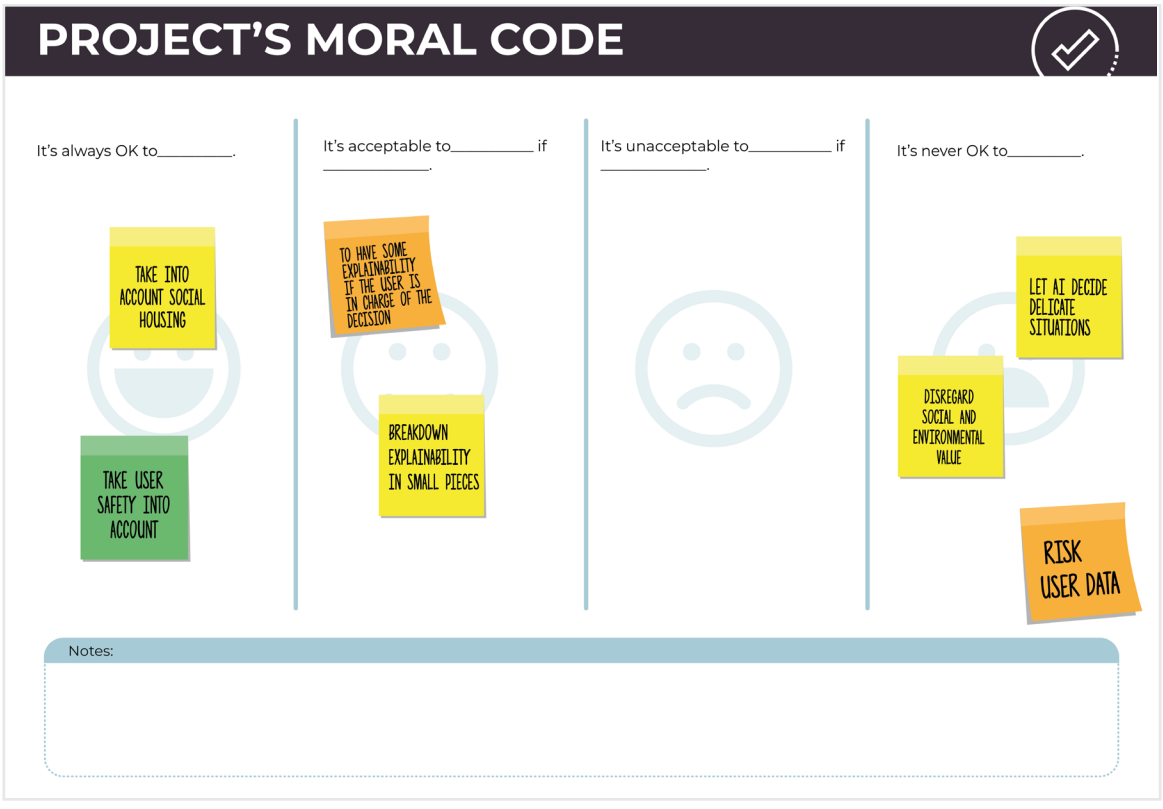
The moral code establishes a scale of moral acceptance for the project. The team will define which actions are always or never adequate, as well as the middle gray area of what is acceptable and unacceptable if a condition is met, according to the ethical evaluation performed in previous stages.

2. It is always/never OK

The extreme sections are intended to be filled by the team with the moral considerations that they consider to be always OK and never OK. The never OK section is meant to contain the extreme "red areas" of ethical considerations that were evaluated using the Ethical Evaluation Axes.

3. It is acceptable to/if

The team can fill this section with the "grey area" ethical implications defined in the Ethical Evaluation Axes.



MOBGEN | Accenture Interactive

TU Delft