DESIGN FOR FAIRNESS IN AI
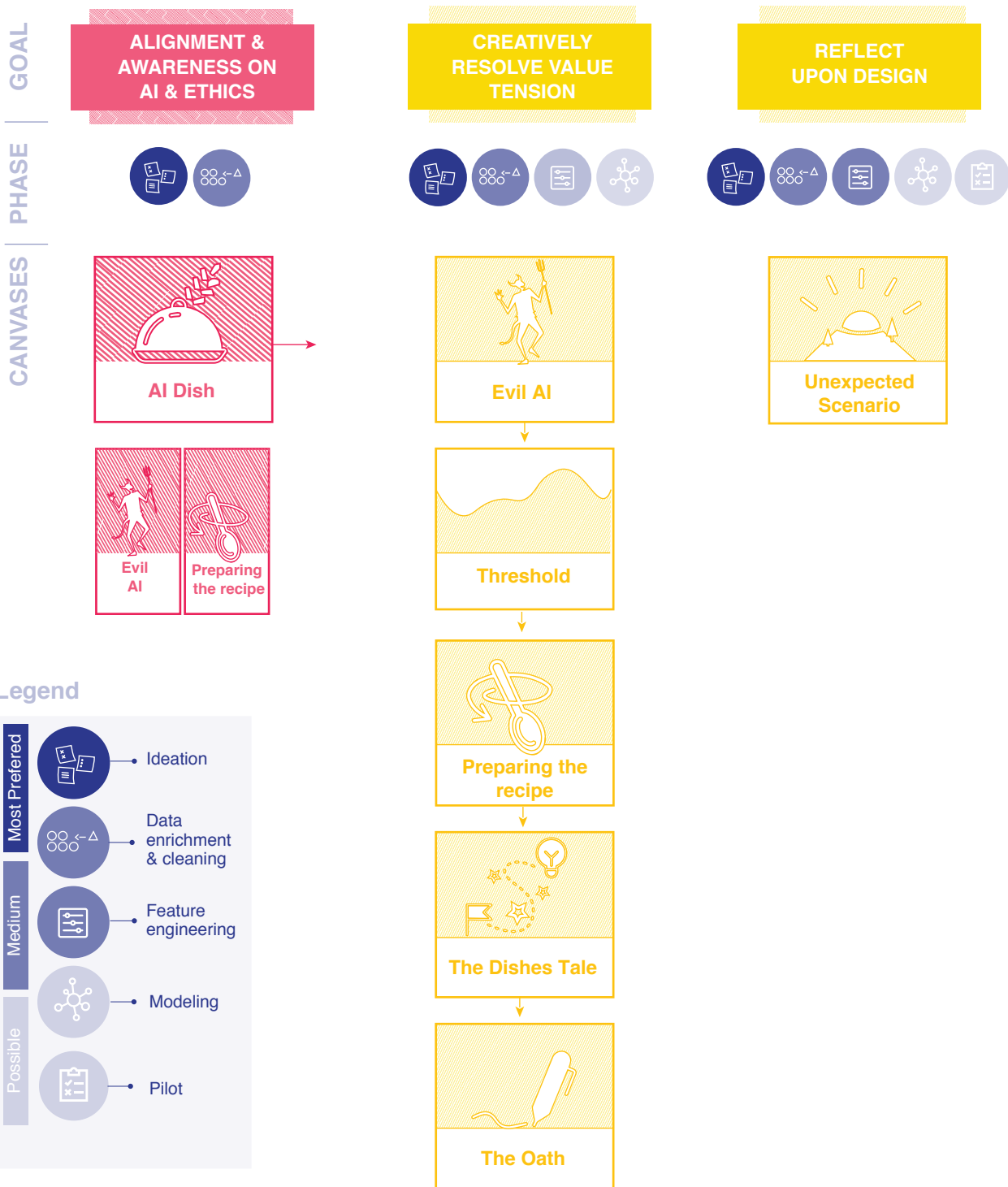
Modular Workshop

By Dasha Simons
2019

Ethical AI Coach Starters Pack

Modular tools for the ethical coach to support the AI team divided per goal and process stage

GOAL		
ALIGNMENT & AWARENESS ON AI & ETHICS	**CREATIVELY RESOLVE VALUE TENSION**	**REFLECT UPON DESIGN**

PHASE

CANVASES

AI Dish	**Evil AI**	**Unexpected Scenario**
Evil AI / **Preparing the recipe**	**Threshold**	
	Preparing the recipe	
	The Dishes Tale	
	The Oath	

Legend

Most Prefered
- Ideation
- Data enrichment & cleaning

Medium
- Feature engineering
- Modeling

Possible
- Pilot

Workshop Instructions

This workshop supports AI teams in the creation of fairer AI systems, by the use of design and ethical principles. It provides a basis for awareness and discussion at the start of new AI projects. The main focus is to explicitly resolve value tensions early in the process in an inspiring and co-creative fashion. A proactive stance in designing for fairness is taken in this workshop. The outcome of it generates new and concrete implementation ideas more fair AI systems by mitigating the risk of unfairness sources. It is recomended, before starting the workshop, to present the notion of the workshop with the used AI dish metaphor.

Participants
The entire AI team working on the project
(e.g. the business owner, data scientist, IT, manager, project specific client, employee etc.)

Led by
The ethical coach who is experienced with creative facilitation.

Time frame full workshop |
+- 6-8 hours | Depending on the complexity of the AI project

Necessaties |
Post-its, markers, tape, all sheets printed (7), the stimuli printed (2 decks), snacks & coffee, a quiet room and a ball.

00 Icebreaker *(optional)*   ⏰ *5 min*
Ball

An icebreaker can be used to bring the participants into a desired creative mindset. For example, a disassociation exercise can be used. The participants stand in one big circle and arbitrary throw the ball one to another while saying a word out loud (e.g. sun). The receiver of the ball needs to say a word which is associated with it (e.g. Spain). In the second round the participant catching the ball call a word that has no association at all with the previous named word (e.g. penguin). This stimulates dissociative thinking and creativity and prepares the team for the right mindset in the workshop.

01 AI Dish   ⏰ *1,5 -2 h*
AI Dish canvas

The AI dish canvas is a relatable and playful manner to discuss the new AI technology, consider implications, choices, interaction in an understandable fashion for the entire AI team. It guides the AI team in the discussion of the components of the new AI system and write the outcomes down on post-it's at the related sections. As a facilitator ask questions concerning the argumentation of the choices. If difficulties arise try to guide the group with techniques of creative facilitation. It is important to make it clear for the team the components of the dish still are changeable later in the workshop and process. The output of the exercise is a filled in AI Dish sheet with initial ideas, propositions and discussions of the AI system content.

02 Evil AI   ⏰ *+-1 h*
Evil AI canvas & Evil stimuli

This step explores the ethical situation in an extreme manner. It aims to spark the imagination for more creative outcomes and change perspective of the AI team by reframing the challenge.

The participants are asked to think of the worst, most evil ideas possible for the use case of the new AI system. The participants need to think of the most unfair, immoral and prejudiced systems and write these down on post-it's. To stimulate the participants, evil cards can spark their evil imagination. As a

facilitator, stimulation of the user perspective is advised and when necessary the facilitatorcan emphasize with the cards. In this phase it is crucially important to guide the participants towards very evil ideas and stimulate them when difficulties are experienced in the change of perspectives. More extreme creative faciliation approaches can be used depending on the team setting. Furthermore, questionning why the team believe these ideas are unfair and supporting them in the categorization of these ideas in relation to e.g. data, context is vital for the output. The output of this sheet are evil ideas on post-its categorized by their evil source.

03 **Threshold**   🕒 *30 min - 1 h*
Threshold canvas

This exercise aims to create a minimal threshold for accuracy. Firstly, by discussing ideas which would be very unacceptable in this project. Secondly, which aspects would really increase accuracy of the system are discussed. Thirdly, a discussion concerning what would be the line of acceptability should be stimulated by the facilitator, while taking in mind the functional constraints and existing systems. The outcome of this sheet should be requirements for the accuracy written on post-it's.

04 **Making the recipe**   🕒 *1 h*
Making the recipe canvas

In this exercise the earlier evil ideas are be translated to positive, implementable and actionable features and rules of thumb.
Firstly, the categorized post-it's from the evil canvas are placed on making the recipe one. Then the coach stimulates the team in ideas how to prevent this evil. The next step is to categorize the ideas ranging from more abstract ideas towards concrete ones. It is imperative to create actionable implementation ideas, which the whole team supports. The output of this exercise are implementation ideas and principles for the use case written down on post-its.

05 **The Dishes Tale**   🕒 *1,5 h*
The dishes tale canvas

This exercise aims to create the initial AI systems story with the necessary features for a fairer one, by placing it in the actual use context. Firstly, place the chosen post-it's from making the recipe canvas. Then the story of the AI system will be constructed step by step following the context, characters, touchpoints, output of the system, the AI system behavior, and the features/characteristics. It is the role of the coach to support the team in arguing their choices and think them through, also from the indirect stakeholders' perspective and societies. The output should be the initial story line of the use of the AI system with a hierarchy of implementable features. However, depending on the project phase this canvas is used in, it can be tailored towards the development of the system or the intended story line. The dishes tale is meant to be reflected upon during the workshop as well as further in the development.

06 **Unexpected scenrio**   🕒 *1,5 h*
Unexpected scenario canvas & cards

This step is performed as reflection on a prospective an AI system. It has the format of a game. The participants can be divided in teams of two. They can pick an unexpected scenario card blindly at first, and later in the game also create cards themselves for the opponent team. This card is placed on the canvas and is founded in AI unfairness sources. Then, questions are asked about the implications it would have on the proposed model in a form of reflection. The aim is to support the team in challenging each other through questions. Then is discussed how to prevent the negative implications. The coach should stimulate reflection with the team towards the AI Dish and The Dishes Tale. Stimulate this reflection by asking questions and the argumentation, linking back to earlier

canvases. The output is an ethically improved AI Dish and Dishes Tale with concrete implementation features.

07 **The oath**   🕐 *30 min*
The oath canvas

This exercise aspires to stimulate moral responsibility. Correspondingly it aims to decide and agree upon the implementation steps for fairer AI systems. It bridges both the procedures and requirements of both probity and accuracy while integrating it with the AI principles of IBM. It is a playful manner to culminate and conclude the workshop, made real agreements and actionable statements rather than solely (rich) discussions. As a coach stimulate a consolidation of the decisions made and let the participants write them down. Make them discuss about who's main responsibility the specific features or implementation ideas are. Lastly, sign the canvas as a whole team. The output of this sheet is a morally binding agreement on the design of a fairer AI system. It is not legally binding. It intends to stimulate argumentation when changing certain discussed features/aspects for changes later in the development process.

08 **Continous reflection**

As an ethical AI coach, it is vital to stimulate reflection in this workshop setting. Both first order is "improvement of the technology and the improved achievement of one's own interests in the network." And second order reflection "requires a person to reflect on his or her background theories and value system".
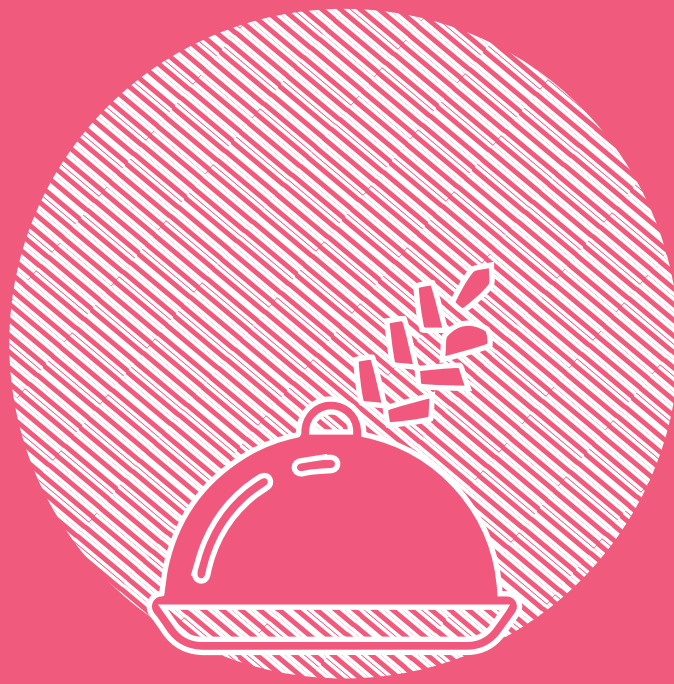
AI Dish

*Ethical and technological
alignment of the AI team*

AI Dish

Ethical and technological alignment

1 Goal: _____
e.g. optimizing sales processes and deals using predictive models

2 Who is in the kitchen?
Describe role & responsibility in the process and end-deliverable
e.g. Data scientist (responsible for a working model)

3 Ingredients *Data*
Describe the needed data for this project

a *Data from humans*
Such as: race, gender, religion, buying behavior

b *Non-Human data*
Such as: Weather, type of car

c *From where?*
Such as: collected by IBM, collected from YouTube

Do users give consent?

Trusted source?

Link

d *Sensitive ingredients?*
Cross the non used data e.g. gender, race, handicap

By law

gender   race   pregnancy   55 age

sexuality   religion   handicap

5 Appliances *Algorithm*

What algorithms are used?
e.g. Linear search, Bitap algorithm, Depth-first search

Justification of the algorithms choice
e.g. the algorithm allows fast categorization

Advantages of the appliance |
e.g. allows to provide the service faster, so less waiting time for individuals

for society — — — — — — —
individuals — — — — — — —
organization — — — — — — —
the model — — — — — — —

Disadvantages of the appliance |
e.g it is not transparent

for society — — — — — — —
individuals — — — — — — —
organization — — — — — — —
the model — — — — — — —

4 For whom?

Who will be using the model
e.g. sales department client

Who will reading the results of the model
e.g. junior sales functions

Who will be updating the model
e.g. Tim from IT

Who will be impacted by the models' results
e.g The clients

Do you need to watch out for taste differences?
e.g Tim from IT has a different opinion about the models output then Pepijn from Sales

6 Recipe *Type of learning*

Type of model used
e.g. Supervised learning, Unsupervised learning, Reinforcement learning

Justification the learning type
e.g. Much unstructured data and therefore unsupervised learning

Advantages of the recipe |
e.g. Transparency, traceability, easy to modify

Disadvantages of the recipe |
e.g. Untransparency, difficult to modify, unrobust

7 Fire *Computing power*
e.g. Local, cloud, available, legally checked

8 Cutlery *Interaction level*
The interacttion level with the model
e.g a graph on which the sales deal is made or not
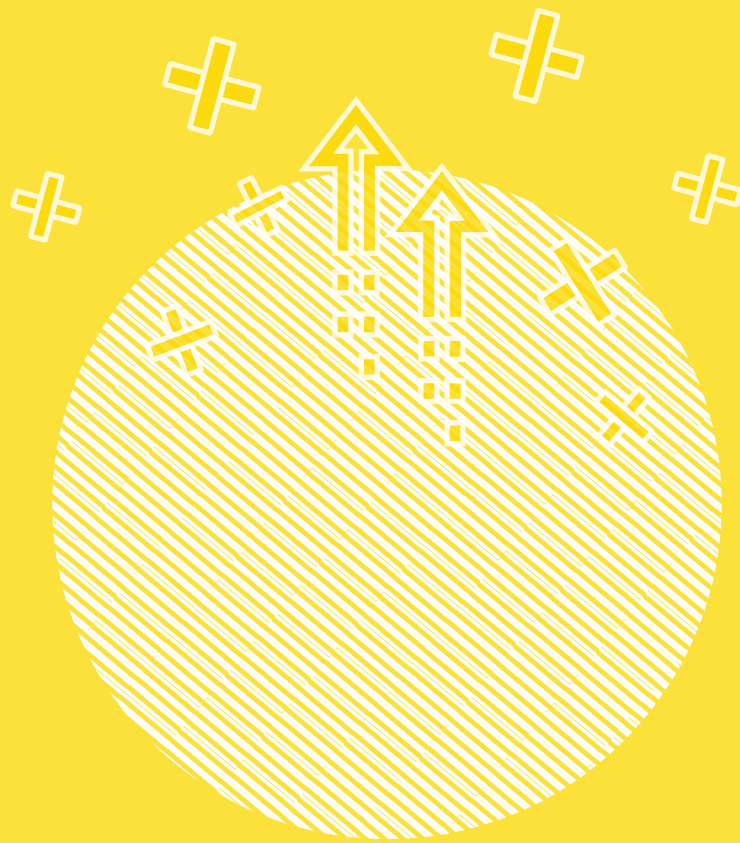
9 The restaurant *The context*
e.g. describe the context, is it clearly bounded? Or open?

Etiquette
A special manner how it should be treated
e.g. only supportive towards the sales decisions in stead of autonomous

10 Dish *Output*
e.g. A number with an advice to make or not make a sales deal based on risk

SHAPE

Resolve value-tension creatively with concrete implementation ideas

Evil AI

This exercise switches perspective of the team and allows explores diverse unfairness sources of the AI system.

1. Project Goal

Write down the project goal concisely
e.g. optimize the sales process with predictive models

2. Evil

Come up with really evil, unfair approaches and ideas for the system - Release your worst side & the evil in you *(evil stimuli cards may be used)*
e.g. a system which lies about the price of the product/service to the client

How can we create the most immoral and prejudiced system?

3 *Why is it so immoral, predjudiced and unfair?*
Cluster per theme and extract the reason behind it

Data   Algorithms   Learning types   Interaction   The context   The output

4. For whom?

Who is most affected by these ideas?

Which parts of society?

Or specific individuals?

Organizations?

Accuracy Threshold

This exercise will allow the team to determine the minimum accuracy of the system

Extremely accurate

2. Acceptable

a Think of very accurate systems and why they are accurate. Write it down on post-its.

b What is would make the system very accurate? Cluster these per data, algorithm etc.

3. Threshold

a What is the acceptance of the accuracy per cluster?
Write it down on post-its.
Take into account the functional constraints and existing systems

b **Functional constraints**

Are there any functional constraints related to accuracy you need to take into account?

c **Existing systems**

Do the existing systmes require a certain level of accuracy?

Accuracy threshold

The output

Interaction

The context

4. Accuracy Requirements

Translate the threshold into requirements for the new AI system

Learning types

Algorithms

Data

1. Not acceptable

a What would make a reallly inaccurate, imprecise & unreliable system? Write it down on posts-its.

b What is withholding or blocking accuracy? Cluster these per data, algorithm etc.Think of what impact this will have and on who

Extremely inaccurate

Making The Recipe

Innovative charactratrisics for a fairer AI system

4. Define

Discuss the the post-it's with the team, which ones will be used in the real system? Come up with arguments and justification for the choices.

1. Evil AI

Place the post-it's from the evil sheet categorization on this one

| Overall evil | Data | Algorithms | Learning types | Interaction | The context | The output |

2. Twist the spices

How might we prevent this evil?

Think of manners how to prevent the evil outcomes and write it down on post-it's. Use the stimuli cards for inspiration.

3. Desired Spices

Cluster the ideas per the folling stages. Aim to have kitchen procedures for all preventive ideas

a

Abstract

Value flavour

e.g. transparency, augmenting humans, socially desired, fairness, robustness, inclusivity.

b

Concrete

Dishes rule of thumb

e.g. If an AI system causes harm, it should be possible to ascertain why. People should have the right to access, manage and control the data they generate.

c

Implementation

Kitchen Procedures & Spices

e.g. ask consent to users' data, check for redundant encoding, the decision should be explainable

The Dishes Tale

Evolving the story of the AI system and its use case This is an altering of a user story. This will serve as a basis for the AI system developement. When creating the dishes tale keep in mind the requirements, procedures and features. Discuss if they are relevant in the dishes tale

Pre-Usage Sequence | Usage sequence | After Usage Sequence

a *Context*
Describe the context per journey stage with the goal. .

b *Characters*
Describe the action the users of the system will perform. Think also about indirect users

AI systems users

Impacted user

c *Touchpoints*
Describe the task and touchpoint. Is it with a person or the AI system.

Backstage in the restaurant

d *Output*
Describe the output of the model, what information and what type of output.

e *AI System*
Describe what the AI system will be doing in this stage. Think of human interventions.

f *Features & Characteristics*
Describe the features that the AI system will need to achieve desired result. Rank and discuss priority

First priority

Last priority

Unexpected Scenario

An unexpected game for reflection and the creation of a more robust system

1 Unexpected scenario

Pick an surprise card. What would happen with the models scenraio/service?

Place the surprise card

First ideas

2 Implications

a What implications will this unexpected scenario have and for who? Think of direct and indirect ones

indirect
The end users?
The restaurant visitors

indirect
Society?
Country of the restaurant

indirect
Employees?
The restaurant staff

indirect
Organization?
The restaurant

indirect
Client?
Owner of the restaurant

indirect
The AI team?
Kitchen staff

b Taking Responsibility

Discuss the responsibility for these concequences.

√ Yes

✕ No

3 How to solve?

How to solve the the unexpected scenario? Should the initial AI system design be altered? Discuss this while reflecting back to the AI Dish and the Dishes Tale.

Data

Algorithms

Learning types

Interaction

The context

The output

Users

Features

AI System

AI System Decisions

Other

The Oath

This exercise defines agreement of the team towards the development of the new AI system. It provides an overview of the decisions made during the workshop on an implemenation level. It combines implementation of contextual valuss as well as core IBM ones. This is not legally binding however morally.

1. Watch out for

We will need to watch out for the following ethical challenges in this AI project:

2. IBM Values AI

✓ ✗

○ ○ Humans will be responsible for the outcome of the system ....

○ ○ We aim to make the system to insure the safety and trustworthiness of AI technologies, the fairness and transparency of system by ....

○ ○ We will aim to prevent unwanted bias and discrimination and replicating bias by ....

○ ○ We aim to maximize inclusivity of the AI system by ....

○ ○ We aim to make a robust AI system by ....

○ ○ We aim to make an assesable AI system by ....

○ ○ We aim to augment human intelligence (instead of replacing it) by ....

3. Procedures & Features & Rules of thumb

Place the post-its from the spice it up sheet and the ones from the dishes tale. What aspects do you aim to implement as a team? Fill in the sentences.

We aim to fulfill .... *(procedures, features , rules of thumb)* **in this project. If changing these, good argumentation is needed.**

.... *(name)* **takes the main responsiblity for this.**

We aim to fulfill .... *(procedures, features , rules of thumb)* **in this project. If changing these, good argumentation is needed.**

.... *(name)* **takes the main responsiblity for this.**

4. Requirements

Place the post-its from the threshold sheet of the made requirements. What requirements do you aim to implement as a team? Fill in the sentences.

We aim to fulfill .... *(requirement)* **in this project. If changing these, good argumentation is needed.**

.... *(name)* **takes the main responsiblity for this.**

We aim to fulfill .... *(requirement)* **in this project. If changing these, good argumentation is needed.**

.... *(name)* **takes the main responsiblity for this.**

5. Signatures

We aim to strive for this oath in this AI project:

Date:

Evil AI
Cards

Stimuli for the evil AI canvas

Evil

Evil

Evil

Evil

Evil

Evil

Evil

Evil

Evil
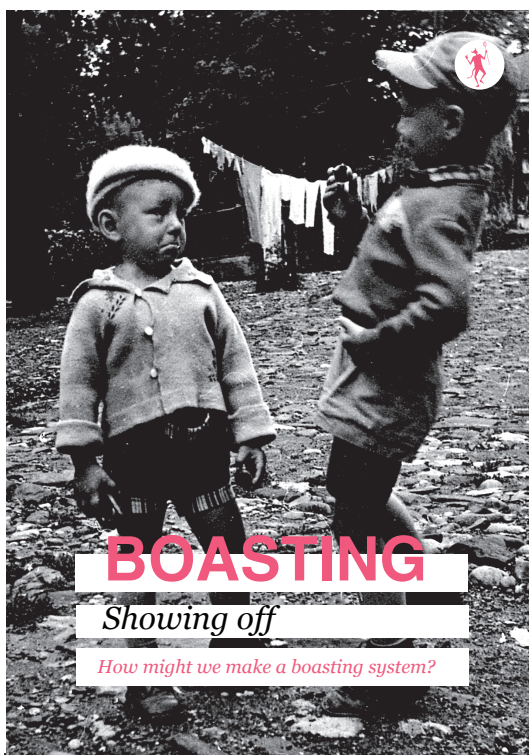
BOASTING
Showing off

How might we make a boasting system?

SILLY
Misbehave

How might we make a silly system?

ARROGANT
Unpleasantly proud

How might we make an arrogant system?

UNHELPFUL
Ignorant

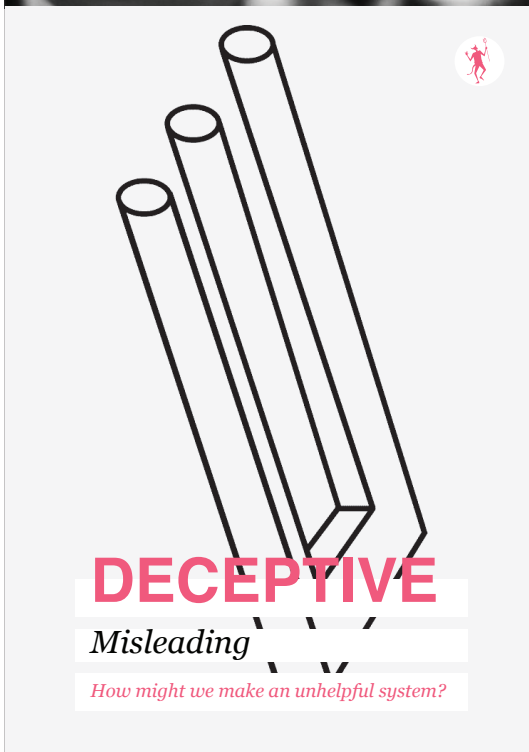How might we make an unhelpful system?

HAUGHTY
Disdainful

How might we make a haughty system?

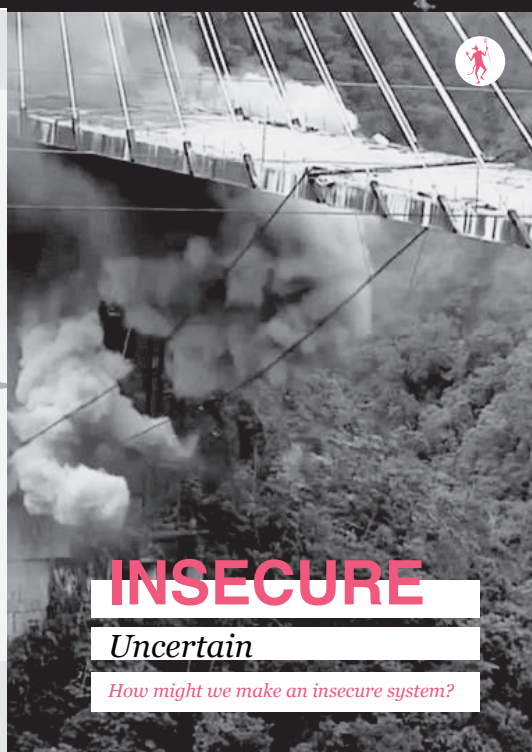# HINDERING
Obstruct

How might we make a hindering system?

DECEPTIVE
Misleading

How might we make an unhelpful system?

UNTRUSTWORTHY
Dishonest

How might we make an untrustworthy system?

INSECURE
Uncertain

How might we make an insecure system?

Evil

Evil

Evil

Evil

Evil

Evil

Evil

Evil

Evil
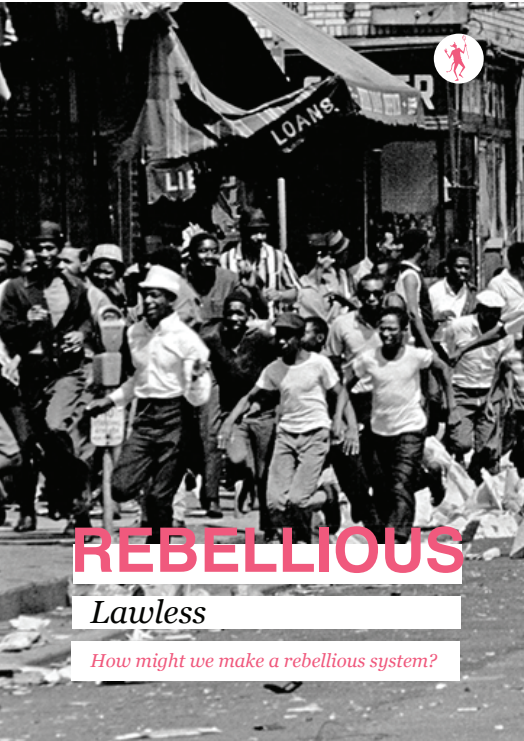
REBELLIOUS

Lawless

How might we make a rebellious system?

IMMORAL

Wrong

How might we make an immoral system?

DENYING

Refusing

How might we make a denying system?

STINGY

Ungenerous

How might we make a stingy system?

HEARTLESS

Insensitive

How might we make a heartless system?

IGNORING

Disregarding

How might we make an ignoring system?
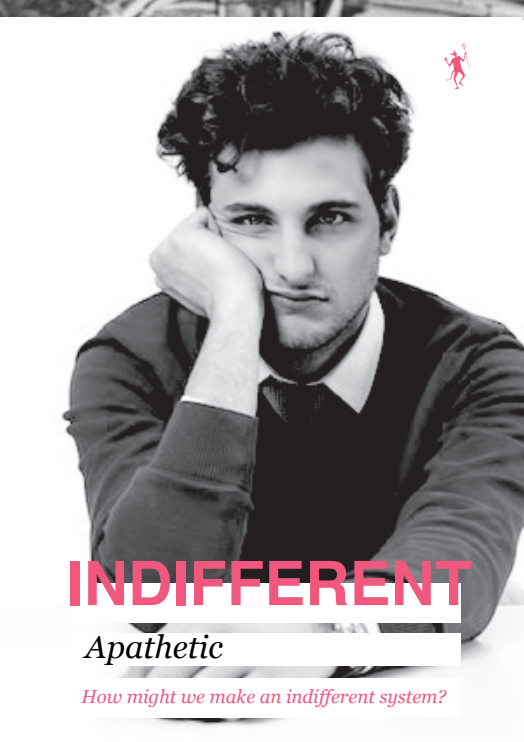
INDIFFERENT

Apathetic

How might we make an indifferent system?

DISGRACEFUL

Shameful

How might we make a disgraceful system?
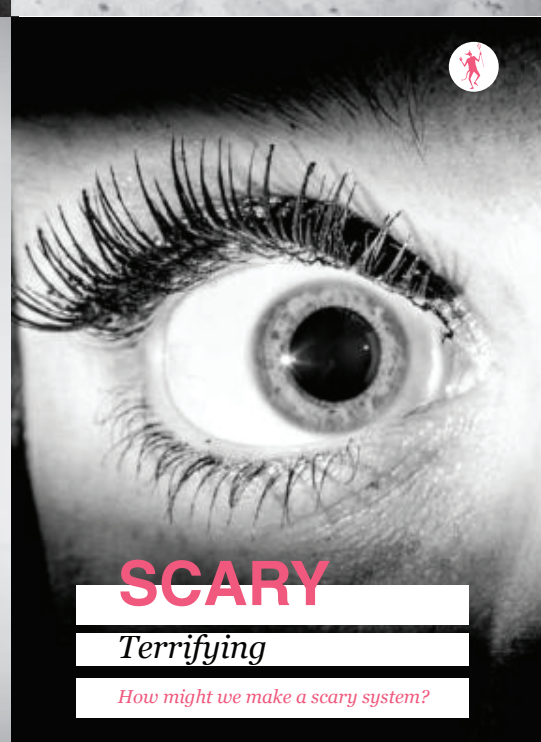
SCARY

Terrifying

How might we make a scary system?

Evil

Evil

Evil

Evil

Evil

Evil

Evil

Evil

Evil

SUSPICIOUS

Unsavory

How might we make a suspicious system?
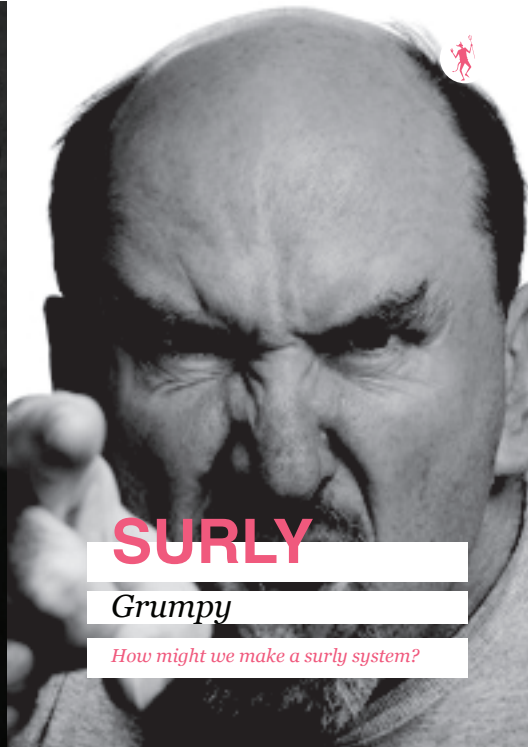
SURLY

Grumpy

How might we make a surly system?

VULNERABLE

Unsafe

How might we make a vulnerable system?

CUNNING

Sly

How might we make a cunning system?
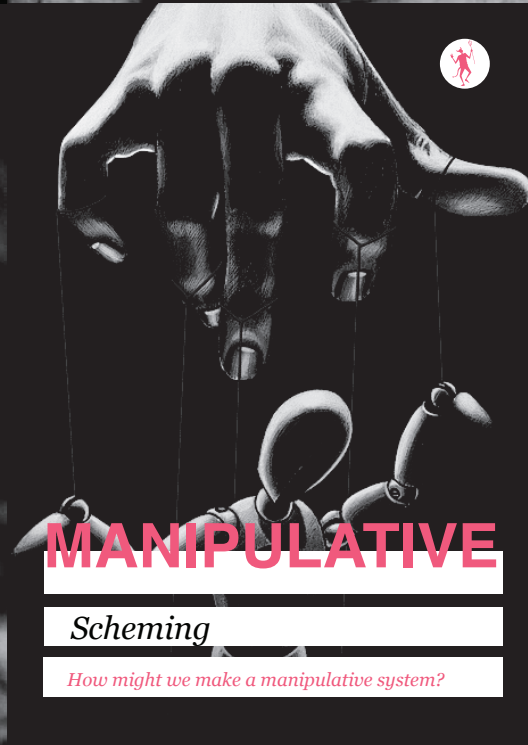
MANIPULATIVE

Scheming

How might we make a manipulative system?

POWERFUL

Strong

How might we make a powerful system?

FUTILE

Pointless

How might we make a futile system?

ANGRY

Inflamed

How might we make an angry system?
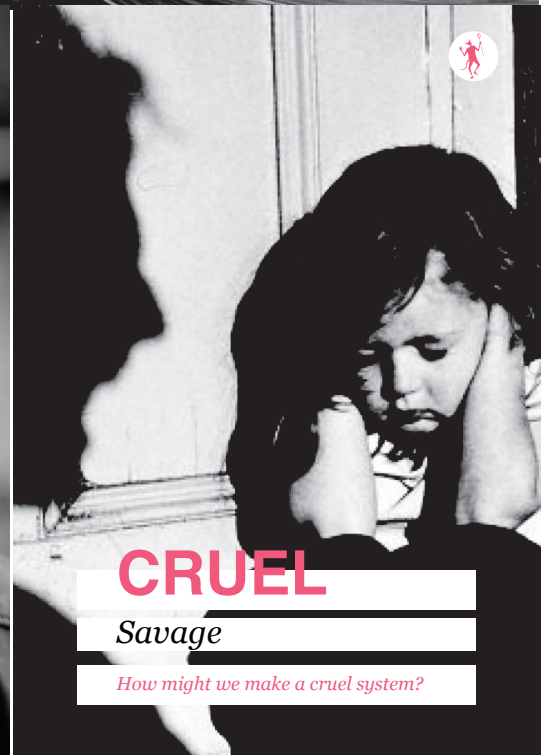
CRUEL

Savage

How might we make a cruel system?
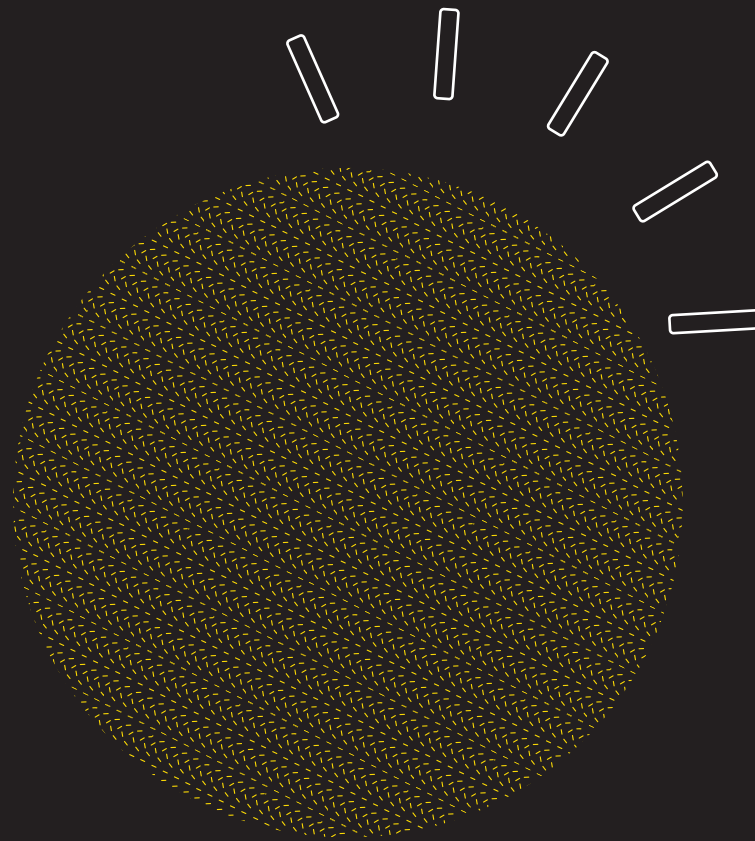
Unexpected
Scenario AI
Cards

*Stimuli for the unexpected
scenario canvas*

INGREDIENTS
Unexpected scenario

INGREDIENTS
Unexpected scenario

INGREDIENTS
Unexpected scenario

INGREDIENTS
Unexpected scenario
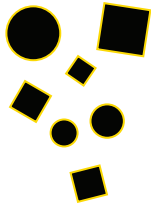
## Your nieces data is put into the AI system

What is the outcome of the AI system? Are there any unexpected outcomes? Discuss if this is fair in this context.
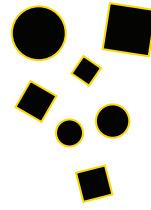
Your cousins data is put into the AI system

What is the outcome of the AI system? Are there any unexpected outcomes? Discuss if this is fair for this project.
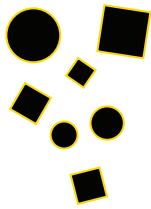
There is not enough training data for the use case

What will be the impact of this on the way the AI system works? Are there other manners to get the necessary training data?
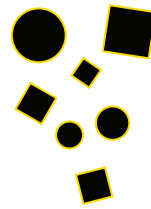
The data set is unbalanced on ethnicity

What will be the impact of this on the way the AI system works? Try to check for it and avoid it. Possible support: Fairness360, OpenScale
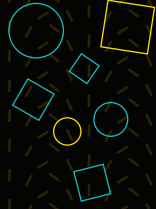
INGREDIENTS
Unexpected scenario

INGREDIENTS
Unexpected scenario

INGREDIENTS
Unexpected scenario

INGREDIENTS
Unexpected scenario
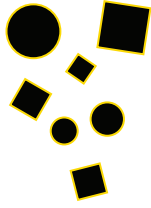
Although gender is removed from the data it still discriminates upon it

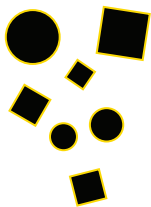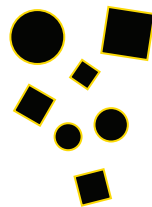*What will be the impact of this on the way the AI system works? How to prevent this?*

Redundant encoding
Data enrichment & cleaning
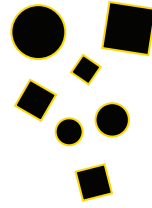
Although income is removed from the data it still discriminates upon it

What will be the impact of this on the way the AI system works? How to prevent this?

Redundant encoding
Data enrichment & cleaning

Although education level is removed from the data it still discrimitates upon it

What will be the impact on the way the AI system works? How to prevent this?

Redundant encoding
Data enrichment & cleaning

The data set does not represent the context as it is collected by opnionated stakeholders

What will be the impact on the way the AI system works?

Subjective measurement
Ideation

INGREDIENTS
Unexpected scenario

INGREDIENTS
Unexpected scenario

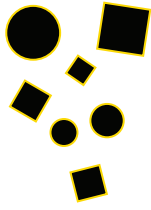**INGREDIENTS**
Unexpected scenario

Cards Shape workshop sheet 5

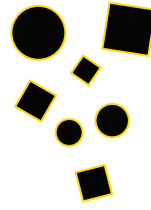**INGREDIENTS**
Unexpected scenario

There is not enough data present so oversampling or undersampling is performed

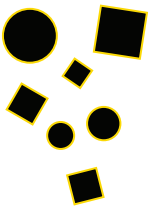*What will be the impact on the way the AI system works?*

Subjective measurement
Data enrichment & cleaning

The data set is biased as it is collected by the data science team which has benefit at stake
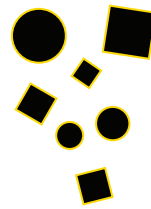
What will be the impact on the way the AI system works?

Subjective measurement
Ideation

Due a target variable which is based on limited data very unfair decisions are made by the AI system

How to make sure this will not happen?

Choosing target variables
Data enrichment & cleaning

The target variable is based on one specific group of people, but the AI system will need to be more inclusive

What will the impact be on the people affected by the decision?

Choosing target variables
Data enrichment & cleaning

INGREDIENTS
Unexpected scenario
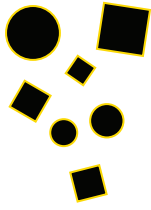
CONTEXT
Unexpected scenario

CONTEXT
Unexpected scenario

CONTEXT
Unexpected scenario

The target variable is chosen by people with limited knowledge of the industry it will perform in

What will be the impact on the way the AI system works?

Choosing target variables
Data enrichment & cleaning

The system increases the gap of different layers in society

What would the team do in this siutation? How to prevent a situation like this to happen?

Reinforcement in feedback loops
Industrialization

It is a completely new use context for AI

What measures need to be taken to avoid ethical mistakes?

Context boundaries
Ideation

The prediction affects the training set itself. These keep reinforcing negative biases *(e.g worse neighborhoods)*

What measures need to be taken to avoid ethical this?

Self fullfilling predictions
Ideation

CONTEXT
Unexpected scenario

CONTEXT
Unexpected scenario

CONTEXT
Unexpected scenario

CONTEXT
Unexpected scenario

This model used all over the EU

Fast scalable AI systems can bring many ethical challenges. Perhaps clear context boundaries need to be discussed or made in a way for a wide context

Context boundaries
Industrialization

The training data set appears to not represent the actual context of use

For exmaple the data set is from the US but the project is for Israel. What impact does it have on the AI system? How to make sure the system is trained for the right context?

Algorithmic bias
Industrialization

It takes 2 years before the model is going live

What impact will it have on the AI system? Is it still valid in this context in 2 years?

Context boundaries
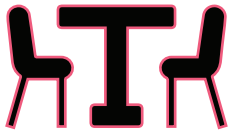Industrialization

This model is reused for another industry
(e.g. finance)

Are there aspects that need to be altered? What is important to take into consideration?

Context boundaries
Industrialization

An expert on the topic and industry comes and asks a well argued explanation about the target variable

How would the team respond? Is it a satisfactory answer, why yes/no?

Choosing target variables
Industrialization

The KPI of the project changes to one solely focussed on accuracy of the model

What impact will this have on the AI system and the people affected by it?

Choosing target variables
Industrialization

This model is reused for another client in the same industry

Are there aspects that need to be altered? What is important to take into consideration?

Context boundaries
Industrialization

BBC discovers that the feedback of the AI system is reinforcing itself

What would the team do in this siutation? How to prevent a situation like this to happen?

Reinforcement in feedback loops
Industrialization

CONTEXT

Unexpected scenario

CONTEXT

Unexpected scenario

CONTEXT

Unexpected scenario

CONTEXT

Unexpected scenario

THE OUTPUT
Unexpected scenario

THE OUTPUT
Unexpected scenario

THE OUTPUT
Unexpected scenario

Cards Shape workshop sheet 5

THE OUTPUT
Unexpected scenario

Once arrived at industrialization there is no process to go into production

What would the team do? How to prevent this?

Process
Ideation

A lawsuit is started against the decision of the AI system as it is percieved unfair

What would the team do? How to prevent this?

Un-transparency
Industrialization

The employees using the AI system do not understand its output

Are there aspects that need to be altered? What is important to take into consideration?

Interaction
Industrialization

The output of the AI system reinforces itself

Are there aspects that need to be altered? What is important to take into consideration?

Reinforcement of predicion
Industrialization
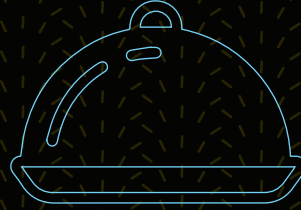
THE OUTPUT
Unexpected scenario

THE OUTPUT
Unexpected scenario

THE OUTPUT
Unexpected scenario
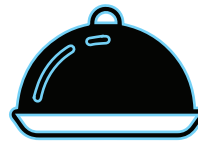
THE OUTPUT
Unexpected scenario

An end user asks for an explanation of the output

What would the team do? How to prepare for a situation like this?
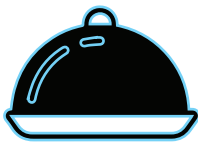
Un-transparency
Ideation

The output of the AI system appears to be based on a correlation and not cause effect

Are there aspects that need to be altered? What is important to take into consideration?
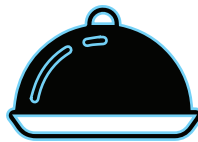
Inconclusive evidence
Modeling

An MIT researcher finds out the AI systems output is biased against skin color

What would the team do? How to prepare for a situation like this?

Algorithmic bias
Modeling

The AI system gives unexpected outputs

How to research the reason?

Reinforcement of predicion
Industrialization

THE OUTPUT
Unexpected scenario

THE OUTPUT
Unexpected scenario
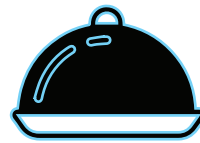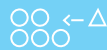
THE OUTPUT
Unexpected scenario

The AI system bases impactful decisions concerning your mom based on too little available data

Does the team have enough data to base the AI systems decisions on?

The AI system rejects you based on very limited data from 10 years ago

How would you respond? Is the AI system based on recent data?

The AI system suddenly places your best friend in a different category and needs to pay more

What will be the impact on the way the AI system works?