AI blindspots

.AGORIA

Knowledge Centre
Data & Society

This AI Blindspots card set is inspired by AI Blindspot, which is available under a Creative Commons Attribution 4.0 International License.

The Knowledge Centre Data & Society, part of imec-SMIT-VUB, adapted the original card set to the Flemish context in order to support the development of trustworthy AI in Flanders. Agoria made the modifications to offer it to the whole Belgian eco-system.

AGORIA

Knowledge Centre Data & Society

What are AI blindspots and how can you detect them?

AI blindspots refer to oversights that can occur before, during, or after the development of an AI system. They originate from biases, prejudices and structural disparities in society.

It is challenging to predict the disadvantageous results of AI blindspots. But they can be mitigated by detecting them proactively and reacting accordingly.

This card set can help uncover potential AI blindspots by reflecting on decisions and actions beforehand.

Each card contains a set of questions to consider potential blindspots, a use case that illustrates the importance of this blindspot, and tools/tricks to help detect/mitigate blindspots.

The card set also includes a joker card to allow you to include other potential AI blindspots you or your team detect.

This card set is inspired by AI Blindspot of Ania Calderon, Dan Taber, Hong Qu, and Jeff Wen, developed during the Berkman Klein Center and MIT Media Lab's 2019 Assembly program.

The Knowledge Centre Data & Society, part of imec-SMIT-VUB, adapted the original card set to the Flemish context in order to support the development of trustworthy AI in Flanders. Agoria made the modifications to offer it to the whole Belgian eco-system.

The original card set identified three phases (i.e. planning, development and implementation). This card set only focuses on the first phase, namely planning. At a later stage, the card set will be expanded with cards of the other two phases.

PURPOSE

At the start of an AI project, determine the purpose of your AI system. Determining the purpose includes involving stakeholders, experts and your team to clearly delineate your purpose and the problem that will be solved with your AI system.

PHASE: PLANNING

HAVE YOU CONSIDERED?

A. Did you **clearly articulate** the problem and outcome you are optimizing for?
B. Is this **tool adequate** to obtain this outcome?
C. Do all involved and affected **stakeholders recognize** this as an important problem?
D. Did you consider the **advantages and disadvantages** of your AI system for each stakeholder?
E. How will you **guarantee to keep the state of purpose** of your AI system?

HOW NOT TO

A company introduced an AI system to speed up their production process, but as an indirect result, employees lost their bonuses. How could this have been avoided? Take the trade union as an involved stakeholder in your project and find a way to increase the speed without losing the bonus.

TOOLS & TRICKS

DATA BALANCE

Data balance means that you have checked your data on its representative quality. And that you have considered how you would mitigate unbalance.

PHASE: PLANNING

DATA BALANCE

PHASE: PLANNING

HAVE YOU CONSIDERED?

A. What is the **minimal viable data collection** you need according to domain experts?
B. Who/What might be **excluded in your data**?
C. How will **limitations** in your data impact the representative nature of your model and the actions your model supports?
D. If your **data is unbalanced**, can you mitigate this limitation?
E. Considering your data, can you describe the case or person where your **predictions will be most unreliable**?

HOW NOT TO

After the release of the massively popular Pokémon Go, several users noted that there were fewer Pokémon locations in primarily black neighborhoods. This came to be because the creators of the algorithms failed to provide a diverse training set, and didn't spend any time in these neighbourhoods.

TOOLS & TRICKS

DATA GOVERNANCE & PRIVACY

Questions with regard to data governance and the impact on the privacy of the data subjects whose personal data will be processed by the AI system, are all part of the preparation of your AI project. Determining the level of access to data and describing the flow of information will help you with protecting your data subject's rights.

PHASE: PLANNING

.AGORIA Knowledge Centre Data & Society

DATA GOVERNANCE & PRIVACY

PHASE: PLANNING

HAVE YOU CONSIDERED?

A. Can you **lawfully process or reuse the data**?
   - If you reuse the data, is the purpose the same?
   - Are appropriate contractual arrangements in place?
   - Can you process or reuse the data on the basis of consent or other grounds?

B. Do you gather **sensitive data** or not?

C. Are there **special regimes to protect your data**?

D. Who will have **access to the (collected) data**? (internally and externally)

E. Can you **comply with the data subject's rights** of the GDPR?

HOW NOT TO

A UK hospital together working together with Deepmind on a AI application detection and diagnosis of kidney injury was fined for violating the rules on personal data. It had transferred personal data on 1,6 million patients without their adequately informing them about this.

TOOLS & TRICKS

TEAM COMPOSITION

Know your team's unknown knowns. It is difficult to be aware of possible (ethical) issues if you are not aware of prejudice within your team. To avoid such blindspots, it is necessary to unveil them.

PHASE: PLANNING

.AGORIA Knowledge Centre Data & Society

TEAM COMPOSITION

PHASE: PLANNING

HAVE YOU CONSIDERED?

A. Did you consider **bias** in your team?
B. Is your **team diverse and multidisciplinary** or in touch with the problem area you try to solve?
C. Who you **should invite** to myth bust this wrong idea?

HOW NOT TO

Google's photo-categorization software has at times mistaken black people for gorillas. The chances of this occurring would decrease drastically if black team members tested the service.

TOOLS & TRICKS

CROSS BOUNDARY EXPERTISE

You may be an expert in machine learning but not in the field you apply machine learning to. This is fine if you have an expert to tell you what to look out for in terms of typical outliers, hugely important variables or common practices that may impact your data.

PHASE: PLANNING

.AGORIA Knowledge Centre Data & Society

CROSS BOUNDARY EXPERTISE

PHASE: PLANNING

HAVE YOU CONSIDERED?

A. Discussing with **domain experts** what the **minimal viable data collection** is that you need in order to allow your AI system to fulfill its purpose?
B. Using an expert to understand what the **impact** should be from your algorithm?
C. Which **variables are essential** for your problem?
D. An expert to help you **assess the results** of your algorithm?

HOW NOT TO

A new algorithm would help with diagnosing who needs to be assessed for pneumonia ASAP in the ER. According to the algorithm, people with asthma do not require immediare care. Experts did not agree with this estimation as asthma cases are treated with urgency in the ER. The experts stated that this was based on faulty assumptions by the AI system. According to the training data, asthma patients spent the least time in the ER. Therefore, the AI system deemed them to be unimportant for reaching efficiency in the ER.

TOOLS & TRICKS

ABUSABILITY

You want to create an AI system to improve something in the world. However, if you only focus on the good it does, you may overlook the ways in which it might cause harm. It is always better to prevent than to cure. So consider what a truly malevolent party could do to or with your application.

PHASE: PLANNING

HAVE YOU CONSIDERED?

A. How the AI system might be used **unethically**?
B. What the **consequences** would be if your AI system was used unethically?
C. Who you have involved to understand the **underlying social motivations and threat models**?
D. What your **mitigation strategy** is if your AI system is used unethically?
E. What to do if your algorithm develops **unethical behaviour**?
F. What the **key ethical principles** are that your AI system should exhibit?

HOW NOT TO

In 2016 Microsoft introduced Tay, a Twitter chatbot, to the world. Within 24 hours Tay was changed as she had learned to be a racist Twitter user based on the tweets addressed to her. Microsoft therefore decided to retire her.

TOOLS & TRICKS

JOKER CARD

PHASE: PLANNING

.AGORIA Knowledge Centre Data & Society

JOKER CARD

PHASE: PLANNING

HAVE YOU CONSIDERED?

HOW NOT TO

TOOLS & TRICKS