AI blindspots

Knowledge Centre
Data & Society

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 taken during the planning, the development and the implementation phase of an AI technology.

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.

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

A&B: problem definition template, jobs-to-be-done insights
B: course on machine learning (Google)
C: translate other applications of your machine learning to your case: does it still makes sense?
D: stakeholder mapping and validation

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

Knowledge Centre Data & Society

DATA BALANCE

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

- A: interview with domain expert
- B, C & D: Data Collection Bias Assessment, Aequitas
- E: create a persona of the invisible man/woman

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

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

- Data Protection Impact Assessment
- Data flow mapping
- Engage with Data Privacy Specialists

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

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

- A: implicit association test
- B: site visit, empathy map, persona, ...

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

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

- Interview or focus group with expert(s)
- Workshop on technical and systems requirements

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

- Create scenarios to grasp your system's unethical practices, and map these consequences on innocent bystander personas.
- Involve experts from social sciences and law
- Thing-Centered Design

JOKER CARD

PHASE: PLANNING

Knowledge Centre
Data & Society

JOKER CARD

PHASE: PLANNING

HAVE YOU CONSIDERED?

HOW NOT TO

TOOLS & TRICKS

DISCRIMINATION BY PROXY

You are not allowed to discriminate against people on the following data categories: gender, ethnicity, religion, race, ... Most organisations avoid this by not collecting this type of data or using this data in feature selection. But have you considered how proxy-data categories can lead to the same discrimination? Shoe size is for example a proxy for gender.

PHASE: DEVELOPMENT

DISCRIMINATION BY PROXY

PHASE: DEVELOPMENT

HAVE YOU CONSIDERED?

A. Specific exceptions or practices in the context you are implementing your AI system?
B. Inviting affected stakeholders to stress test your system against historical biases?
C. Identifying and removing features that are correlated with vulnerable social groups?

HOW NOT TO

An AI system that predicts which patients would benefit from extra medical care prioritised healthier white patients instead of more at risk black patients. The algorithm was based upon how much a patient would cost to the healthcare system in the future, but did not consider that black patients spend less on medical care than white patients with the same chronic conditions.

TOOLS & TRICKS

- Involve domain experts
- A & C: check for unintentional correlations that might impact vulnerable groups
- B: contextmapping, workshop on participatory approaches to machine learning
- C: Aequitas

EXPLAINABILITY

Why is explainability important? The predictions or recommendations generated by your AI system can be unclear and may be surprising. When creating an AI system, you have the responsibility to clearly inform users about the underlying technical logic of the system and how predictions or recommendations are generated.

PHASE: DEVELOPMENT

EXPLAINABILITY

PHASE: DEVELOPMENT

HAVE YOU CONSIDERED?

A. If people **trust** the choices made by your system?
B. What the **impact** is of having an AI system generating a prediction versus a human?
C. How you can **interpret or explain the choices** of your AI system?

HOW NOT TO

A medical authority in the US used an AI system to determine reimbursements for disabled people. However, the court stated that these reimbursements were not possible because the decisions of the AI system were not explained.

TOOLS & TRICKS

- A & B: human-centered design methods
- C: Lime, WhatIf
- A, B & C: explain your AI system to a random person and check if the results of your solution are clear and comprehensible, AI Explainability 360, Value Proposition Design

PERFORMANCE BALANCE

When determining an AI system's metrics for success, trade-offs between optimal performance and negatively impacting vulnerable social groups must be made.

PHASE: DEVELOPMENT

PERFORMANCE BALANCE

PHASE: DEVELOPMENT

HAVE YOU CONSIDERED?

A. If the chosen **performance indicators** will not stray the AI system from its original purpose?
B. Which performance indicators are necessary and what the **impact of these indicators** will be on vulnerable social groups?
C. How **statistically accurate** your AI system is?

HOW NOT TO

AI can help by screening for cancer. However, if it is optimized to detect all potential persons with cancer, this will result in a higher amount of false positives. These can cause unnecessary anxiety with those persons without cancer.

TOOLS & TRICKS

- A & B: interviews with domain experts, IoT stress test (part of the Internet of Things Design Kit)
- C: check your method for statistical accuracy, intermediate/ prototype testing

INCLUSION/ OMISSION CHECK

Your AI system might be beneficial to most people but have you considered how specific people might be worse off? Consider if your system is inclusive for economically vulnerable persons, people with lower digital literacy or people with a disability.

PHASE: DEVELOPMENT

Knowledge Centre Data & Society

INCLUSION/OMISSION CHECK

PHASE: DEVELOPMENT

HAVE YOU CONSIDERED?

A. How your system might **exclude (vulnerable) people**?
B. How people might be **digitally excluded** with your system?
C. How to **minimize the number of affected people** by your AI system?

HOW NOT TO

AI systems are often perceived as enablers for digital inclusion. They can for example detect atypical browsing behaviours and thus identify people's difficulties when browsing the Internet. But what if things are the other way around and your AI system has a negative effect on (digitally) vulnerable people? How will you ensure your AI system is made for all and can be used by all?

TOOLS & TRICKS

- A: inclusion by design toolkit of SMIT (coming soon)
- B: 8 Profiles of Digital Inequalities: can all profiles make use of or benefit from your AI system?
- C: involve UI designers, organize a co-creation workshop with targeted end-users

DATASET SHIFT

A significant difference between your training datasets and testing datasets can result in what is called a 'dataset shift'. This can heavily impact the performance of your algorithms.

PHASE: DEVELOPMENT

DATASET SHIFT

HAVE YOU CONSIDERED?

A. A **systematic flaw** in the data collection or labeling process that causes a nonuniform selection of training examples from a population and results in biases during the training of an AI system?
B. If your data is (un)affected by **shifts in time and location**?

HOW NOT TO

If certain species are omitted in the training set of an image classification system for cats and dogs, the test set will reveal that not all images can be classified correctly.

TOOLS & TRICKS

- Identification and correction of dataset shifts

JOKER CARD

PHASE: DEVELOPMENT

Knowledge Centre
Data & Society

JOKER CARD

PHASE: DEVELOPMENT

HAVE YOU CONSIDERED?

HOW NOT TO

TOOLS & TRICKS

GENERALIZATION ERROR

Between conceiving, building and deploying an AI system, conditions in the world may have changed. Or they no longer reflect the development context. As a result training and testing data are no longer representative or adequate.

PHASE: IMPLEMENTATION

GENERALIZATION ERROR

HAVE YOU CONSIDERED?

A. Regularly checking your **training and testing data** against the current situation?
B. Ensuring a **human review process for outliers**?
C. Determining if the **input data and predicted values** align with expectations?
D. Planning how to ensure your **model can be retired**?

HOW NOT TO

Your email spam filtering system may fail to recognize spam that differs in form from the spam the automatic filter has been built on. As a result, the spam filtering system will not work properly.

TOOLS & TRICKS

- Involve statistical experts
- A & C: identification and correction of dataset shifts (see also card on dataset shift)
- B: outlier detection techniques

RIGHT TO EXPLANATION & OBJECTION

AI systems can carry biases that make them subjective or imperfect. The right to an explanation or objection to an algorithmic decision can mitigate inaccuracies and grant agency to people affected.

PHASE: IMPLEMENTATION

Knowledge Centre Data & Society

HAVE YOU CONSIDERED?

A. Ensuring **transparency and accountability** throughout the system and deployment?
B. Offering individuals **meaningful explanations for a given decision**?
C. Providing **guidance** on how to contest the AI decision?

HOW NOT TO

Based on algorithms, insurance rates can differ amongst clients, sometimes it is unclear where this price difference originates from. Clients should be informed about the reasoning of the algorithms in determining the premiums.

TOOLS & TRICKS

- A: Data Collection Bias Assessment
- B: stay up to date with explainable AI techniques (e.g. AI Explainability 360)
- C: ensure clear lines of communication towards end-users

USER MANUAL

The actual use and the context of use of an AI system determine if harmful unintended consequences will occur. The creation of a user manual to guide possible end users in a responsible use of your AI-system might help in this regard.

PHASE: IMPLEMENTATION

HAVE YOU CONSIDERED?

A. Have you considered the **different contexts of use** of your system?
B. If the **use of AI is not disproportionate** to the context?
C. Have you considered **informing users** on how to appropriately make use of the AI system?
D. How your AI system may be deployed in **a situation you did not envisage**?

HOW NOT TO

A camera surveillance might not be the best choice to increase security in a city, the same budget could perhaps be spent better on prevention or social outreach workers.

TOOLS & TRICKS

- A & C: write a user manual
- B: use participatory methods to determine proportionality
- C: consult legal expertise
- D: think of possible worst case scenarios and look for mitigation strategies

TRANSPARENCY

It is important to not only communicate about your AI system and the decisions it makes when your target audience asks for it. Gaining trust of your target audience starts with communicating proactively and transparently at all times.

PHASE: IMPLEMENTATION

HAVE YOU CONSIDERED?

A. Communicating and explaining your AI system and the decisions it makes to your target audience (the users of your AI system) and the outside world?

B. What moment would be the most ideal to communicate about your systems and the decision it makes to your target audience?

HOW NOT TO

You paid for an ethical audit and managed to mitigate all the uncovered challenges but you do not publicly communicate this in any way. As a result, your AI application is met with undeserved suspicion.

TOOLS & TRICKS

- A: a workshop on data-use notices, guidelines on people-centric approaches to notice, consent and disclosure, People + AI Guidebook
- B: develop a communication strategy together with the company's communication team

SERVICE RECON-SIDERATION

The implementation of your AI system might cause some changes in your current workflow and work profiles, as well as in your customer agreements. Being aware of possible changes, will help you to anticipate and act fast on these changes.

PHASE: IMPLEMENTATION

SERVICE RECONSIDERATION

PHASE: IMPLEMENTATION

HAVE YOU CONSIDERED?

A. Changes in your **current workflow and work profiles** due to the implementation of your AI system?
B. Evaluating the effect of your AI system on your **service's customer agreements**?

HOW NOT TO

Support specialists can talk to customers about more advanced and difficult topics, and they can easily solve such issues with the help of AI. But in some cases, customers receive a notification about problems with their accounts, before they knew something was wrong.

TOOLS & TRICKS

- A: customer journey mapping, Tarot Cards of Tech
- B: service design workshop

ACCOUNTABILITY & SIGN-OFF

As a company introducing or making use of an AI system, you must explain and justify the decisions and actions that were made to your partners, users and others who may interact with your AI system.

PHASE: IMPLEMENTATION

Knowledge Centre Data & Society

ACCOUNTABILITY & SIGN-OFF

PHASE: IMPLEMENTATION

HAVE YOU CONSIDERED?

A. Who will take the **final decision** if the algorithm can be released?
B. Who will be **held accountable** if something goes wrong?

HOW NOT TO

Learning platforms must be clear and explicit about the recommendation of certain learning paths or options over others. Trainees, course designers and teachers can review and update as they see fit.

TOOLS & TRICKS

- Trustable Technology Mark
- A: set up an independent ethical board that reviews the algorithm and the data on which it is built.
- B: Log the choices you have made during the collection of the data and during the design and development of the application (eg. by making use of the Data Collection Bias Assessment)

REVISITED PURPOSE

During and after the development of your AI system it is important to ask yourself if your AI system is still the best solution and means for the goal you have set.

PHASE: IMPLEMENTATION

HAVE YOU CONSIDERED?

A. Does your AI system still apply to the **purpose you had in mind**?
B. How long a **defined 'ruleset'** for your AI system remains in place? Or does it evolve constantly?

HOW NOT TO

The light sensor of a no-touch soap dispenser in a lavatory was solely trained with light-skinned persons. As a result, people of color could not make use of the soap dispenser.

TOOLS & TRICKS

- Revisit the problem definition template you had prepared

JOKER CARD

PHASE: IMPLEMENTATION

Knowledge Centre
Data & Society

JOKER CARD

PHASE: IMPLEMENTATION

HAVE YOU CONSIDERED?

HOW NOT TO

TOOLS & TRICKS