
AI|BLINDSP•T

CREATIVE LICENSES, TEAM, AND ACKNOWLEDGEMENTS

The AI Blindspot cards were developed by Ania Calderon, Dan Taber, Hong Qu, and Jeff Wen during the Berkman Klein Center and MIT Media Lab's 2019 Assembly program.

This project would not have been possible without the stimulating conversations we've had with the Assembly Cohort and Assembly Advisors. We are especially indebted to Shira Chung for designing the AI Blindspot cards, Lara Taber for designing the AI Blindspot logo, and Mariel Calderon for designing the AI Blindspot Discovery Process. We are grateful to Friederike Schuur, Walter Frick, and Erich Ludwig for insightful conversations provided in user feedback sessions, as well as John Bowers for his peerless edits. Finally, we want to thank Hilary Ross for her unwavering support and guidance during the Assembly program.

AI|BLINDSP•T

INTRO TO AI BLINDSPOT

AI Blindspots are oversights in a team's workflow that can generate harmful unintended consequences. They can arise from our unconscious biases or structural inequalities embedded in society. Blindspots can occur at any point before, during, or after the development of a model, from when the model is first conceptualized to when it is built to after it is deployed. The consequences of blindspots are challenging to foresee, but they tend to have adverse effects on historically marginalized communities. Like any blindspot, AI blindspots are universal—nobody is immune to them—but harm can be mitigated if we intentionally take action to guard against them.

WHAT DO WE MEAN BY AI?

Artificial intelligence has become a catch-all category of systems that derive patterns, insights, and predictions from big datasets. While they might aspire to emulate and automate intelligent human-like judgment, most algorithms referred to as AI are in fact simple, imperfect models susceptible to making erroneous inferences. The risk of delegating high-stakes social and commercial decisions to AI exposes everyone to unequal treatment because these seemingly impartial algorithms are produced by computer scientists, engineers, and companies whose data and practices may amplify historical biases in society. Fairness requires thoughtful vigilance across all sectors, especially from researchers inventing, engineers building, and organizations deploying AI systems. Above all, we need to safeguard and uplift people whose lives are affected by AI.

TABLE OF CONTENTS

Planning
- + Purpose
- + Representative Data
- + Abusability
- + Privacy

Building
- + Discrimination by Proxy
- + Explainability
- + Optimization Criteria

Deploying
- + Generalization Error
- + Right to Contest

HOW TO USE

It's hard to predict where inaccuracies and flaws may arise in the AI development process or know how to prevent them.

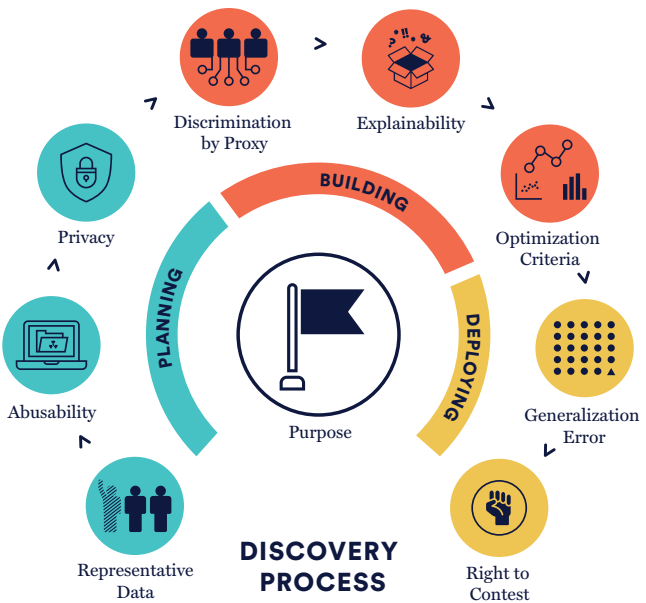
These cards encourage us to have conversations that can help uncover potential blindspots during planning, building, and deploying of AI systems. Each card contains:

+ a summary of a potential blindspot you may encounter
+ actions you can take to address it
+ an example of how it affects real-life scenarios
+ whom you should reach out to
+ a handy QR code that will take you to additional resources

There is no silver bullet, so we left room for the unknown unknowns in the form of a "joker card" to inspire you to think of other blindspots not included here. We encourage you to continue using and refining these cards over time.

AI BLINDSPOT DISCOVERY PROCESS

A 9 step discovery process that helps organizations spot and address unconscious biases and structural inequalities that can lead to unintended consequences when deploying AI systems.



STAGES

Planning

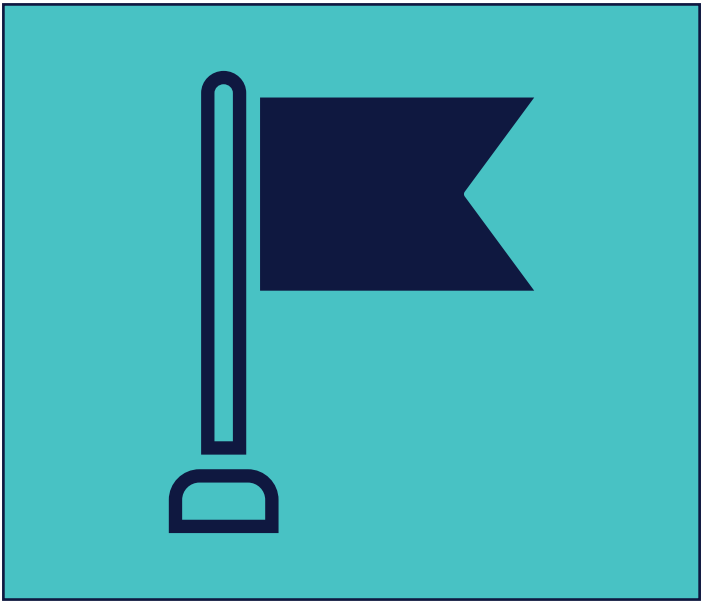
In the initial stages of your project, it is important to think critically about: why you want to use a particular technology (Purpose); how accurately your data reflects affected communities (Representative Data); what vulnerabilities your system might expose (Abusability); and how to safeguard personal identifiable information (Privacy).

Building

Vulnerable populations can be harmed due to the performance metric you choose (Optimization Criteria) or variables that act as proxies (Discrimination by Proxy). Depending on the sensitivity of the use case, you may need to understand and explain how the algorithm makes determinations (Explainability).

Deploying

You should be vigilant about monitoring for changes that might affect the performance and impact of your system (Generalization Error), and ensure that individuals have mechanisms to challenge decisions (Right to Contest).



Purpose

AI systems should make the world a better place. Defining a shared goal guides decisions across the lifecycle of an algorithmic decision-making system, promoting trust amongst individuals and the public.

PURPOSE

HAVE YOU CONSIDERED?

+ Clearly articulating the problem and outcome you are optimizing for
+ Assessing whether your tool is well-suited to this purpose
+ Working with individuals who may be directly affected to identify an appropriate way to measure success
+ Tracking and publishing why and how you are using your AI system

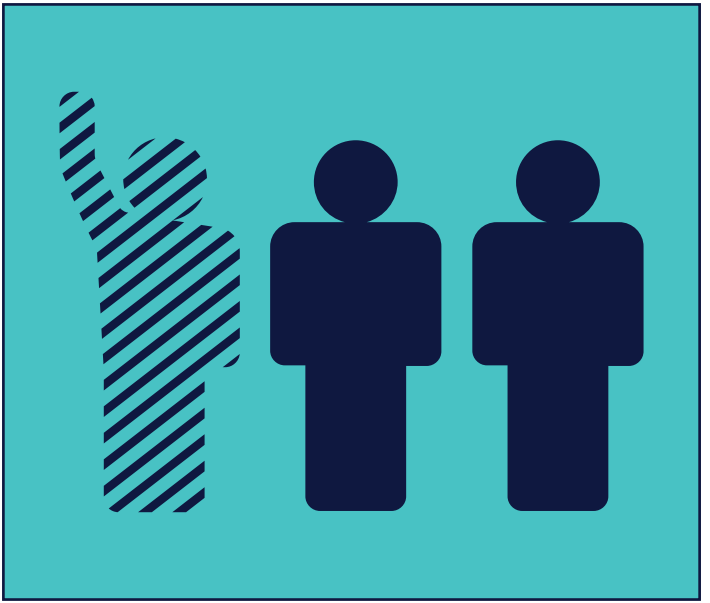
CASE STUDY

Epidemiologists pushed for access to anonymized call records to understand how populations moved during the 2014 Ebola outbreak in Sierra Leone. Yet the virus spreads through direct contact, so this did not help track Ebola. Evaluating benefits and potential harms would have focused attention on saving lives.

HAVE YOU ENGAGED WITH?

+ Affected communities
+ Subject matter experts
+ Policymakers

TAKE A LOOK



Representative Data

For an algorithm to be effective, its training data must be representative of the communities that it may impact. The way that you collect and organize data will benefit certain groups while excluding or harming others.

REPRESENTATIVE DATA

HAVE YOU CONSIDERED?

+ Exploring how your data might be incomplete or skewed, or encode historical biases
+ Including diverse voices in the data definition and collection process
+ Partnering with social service agencies for outreach to vulnerable groups

CASE STUDY

Researchers in Germany, the USA, and France developed an algorithm that detects skin cancer more accurately than dermatologists. The system finds 95% of melanomas, versus 89% by doctors. But it is effective only on light skin tones because a demographically diverse dataset has never been collected.

HAVE YOU ENGAGED WITH?

+ Affected communities
+ Subject matter experts
+ Civil rights organizations

TAKE A LOOK

PLANNING BUILDING DEPLOYING

Abusability

The designers of an AI system need to anticipate vulnerabilities and dual-use scenarios by modeling how bad actors might hijack and weaponize the system for malicious activity.

ABUSABILITY

HAVE YOU CONSIDERED?

+ Creating scenarios with hypothetical malicious and innocent bystander personas
+ Conducting “red team” exercises
+ Developing processes for long term mitigation and real-time damage control
+ Engaging sociologists, ethnographers, and political scientists to understand the motivations and incentives that underpin threat models
+ Conjuring up a worst-case scenario that might appear in tomorrow’s headline

CASE STUDY

Facebook’s ad network allows for political ads. They didn’t anticipate foreigners buying ads to influence elections. Facebook later required identity verification, but this inadvertently prevented nonpartisan news organizations from buying ads to promote their articles about politics.

HAVE YOU ENGAGED WITH?

+ Social scientists
+ Affected communities
+ Cybersecurity experts

TAKE A LOOK

PLANNING BUILDING DEPLOYING

Privacy

AI systems often gather personal information that can invade our privacy. Systems storing confidential data can also be vulnerable to cyberattacks that result in devastating data breaches to access personal information.

PRIVACY

HAVE YOU CONSIDERED?

+ Using privacy-enhancing technologies such as federated learning, differential privacy, de-identification, and secure data enclaves based on the level of risk
+ Conducting privacy and security risk assessments, and incorporating privacy by design measures in ethical review processes
+ Requiring affirmative, prospective consent from individuals of the intention to include data about them
+ Allowing individuals to object or withdraw their consent

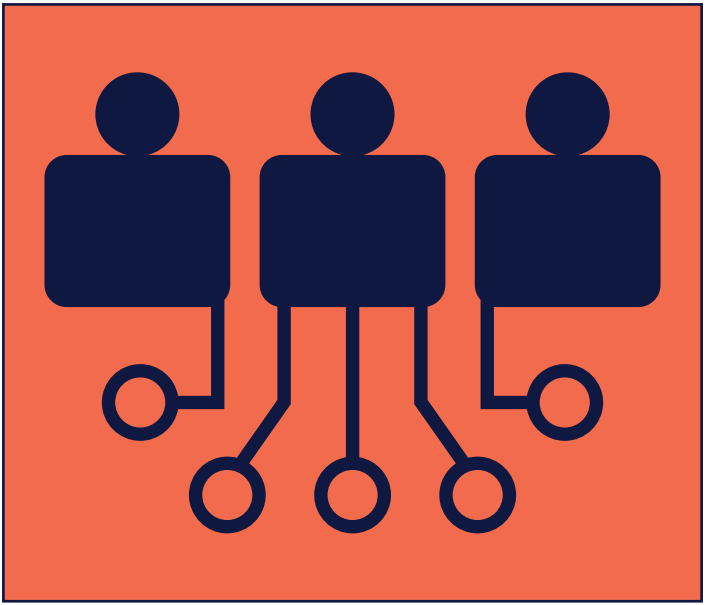
CASE STUDY

Facebook's ad network allows for political ads. They didn't anticipate foreigners buying ads to influence elections. Facebook later required identity verification, but this inadvertently prevented nonpartisan news organizations from buying ads to promote their articles about politics.

HAVE YOU ENGAGED WITH?

+ Privacy advocates
+ Legal counsel
+ Cybersecurity experts

TAKE A LOOK



Discrimination by Proxy

An algorithm can have an adverse effect on vulnerable populations even without explicitly including protected characteristics. This often occurs when a model includes features that are correlated with these characteristics.

DISCRIMINATION BY PROXY

HAVE YOU CONSIDERED?

+ How historical practices in the relevant field (employment, housing, etc.) might bias your data
+ Conducting participatory research and consulting with subject matter experts to help identify such historical biases
+ Identifying and removing features that are correlated with vulnerable populations
+ Developing a quantitative method to test for fairness with respect to different affected groups

CASE STUDY

Amazon's internal hiring algorithm was biased against women even though gender was not explicitly presented to the algorithm as a predictive feature. This occurred because of historical hiring biases that taught the algorithm to discriminate against candidates whose resume text referenced gender (e.g., "women's" teams).

HAVE YOU ENGAGED WITH?

+ Social scientists
+ Human rights advocates
+ Statisticians

TAKE A LOOK

PLANNING BUILDING DEPLOYING

Explainability

The technical logic of algorithms is complex, which make recommendations unclear. People involved in designing and deploying algorithmic systems have a responsibility to explain high-stakes decisions that affect individuals' well-being.

EXPLAINABILITY

HAVE YOU CONSIDERED?

+ Surveying individuals on whether they comprehend and trust the recommendations made by your model
+ Factoring in the stakes of decisions (e.g., recommending a movie vs. approving a home loan)
+ Choosing models that are easier to interpret (e.g., logistic regression, random forest)
+ Modeling counterfactual scenarios that would enable individuals to understand what would need to change to receive a desired result

CASE STUDY

The Fair Credit Reporting Act (FCRA) gives U.S. consumers the right to know and correct information about themselves. These rights may be violated if a complex model assigns credit scores but cannot justify them or provide information on how to improve them.

HAVE YOU ENGAGED WITH?

+ Affected communities
+ Legal counsel
+ UX researchers

TAKE A LOOK

PLANNING BUILDING DEPLOYING

Optimization Criteria

There are trade-offs and potential externalities when determining an AI system's metrics for success. It is important to balance performance metrics against the risk of negatively impacting vulnerable populations.

OPTIMIZATION CRITERIA

HAVE YOU CONSIDERED?

+ Whether optimizing against measurable key performance indicators (KPIs) will lead to a deviation from the original purpose of the system
+ Determining what KPIs are measurable and understanding the impacts on affected communities
+ Monitoring a set of metrics, such as F1 score, that balance performance across several dimensions

CASE STUDY

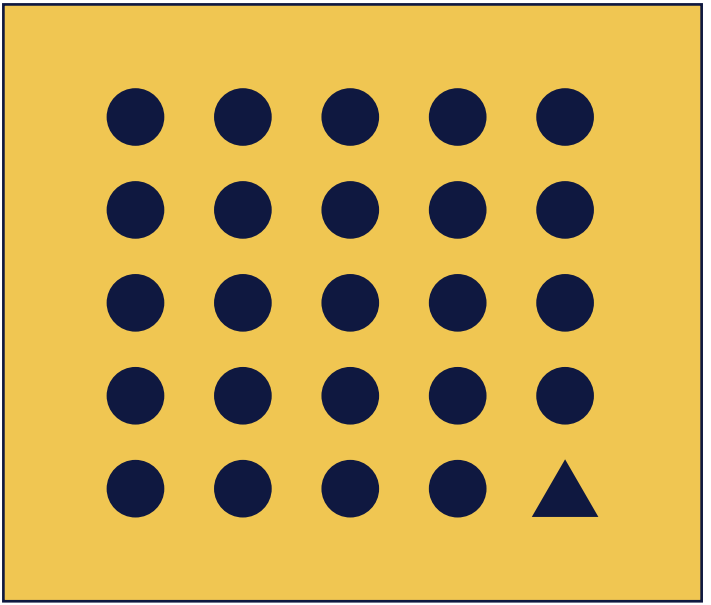
A hospital developed a cancer screening system and focused on maximizing accuracy as its performance metric. Optimizing for accuracy alone might lead to false positives that increase patient anxiety or false negatives that preclude early detection.

HAVE YOU ENGAGED WITH?

+ Subject matter experts
+ Statisticians
+ UX researchers

TAKE A LOOK

PLANNING BUILDING DEPLOYING



Generalization Error

Between building and deploying an AI system, conditions in the world may change or not reflect the context in which the system was designed, such that training data are no longer representative.

GENERALIZATION ERROR

HAVE YOU CONSIDERED?

+ Identifying possible shifts in demographics
+ Building a human review process for outliers
+ Determining if the input data and predicted values align with expectations
+ Formulating a plan for sunsetting the model when the model is obsolete or might cause harm

CASE STUDY

Facebook's video platform has been trying to automatically identify and remove mass shooting videos. However, it was unable to catch first-person shooter videos filmed on head-mounted cameras because the training data it had previously used was in the third-person perspective.

HAVE YOU ENGAGED WITH?

+ Social scientists
+ Statisticians
+ Affected communities

TAKE A LOOK

PLANNING BUILDING DEPLOYING

Right to Contest

Like any human process, AI systems carry biases that make them subjective and imperfect. The right to contest an algorithmic decision can surface inaccuracies and grant agency to people affected.

RIGHT TO CONTEST

HAVE YOU CONSIDERED?

+ Establishing transparency, accountability, and participatory mechanisms across the planning, building, and deploying process to incorporate diverse views
+ Offering individuals meaningful explanations for a given decision
+ Enabling people affected by an automated decision to identify potential grounds upon which it might be contested
+ Providing guidance on how people may change their data profile to achieve a desired result

CASE STUDY

A murder trial defendant contested using the STRmix algorithm to determine whether his DNA matched to the evidence at the murder scene. The defense challenged the tool's reliability as he was one of eight people who might match the sample. The judge rendered the sample inadmissible.

HAVE YOU ENGAGED WITH?

+ Affected communities
+ Legal counsel
+ Policymakers

TAKE A LOOK

PLANNING BUILDING DEPLOYING

PLANNING BUILDING DEPLOYING

HAVE YOU CONSIDERED?

CASE STUDY

HAVE YOU ENGAGED WITH?