AI blindspots

.AGORIA

Knowledge Centre
Data & Society

Cette série de fiches AI Blindspots est inspirée des fiches AI Blindspot, disponibles sous licence Creative Commons Attribution 4.0 International.

Le Knowledge Centre Data & Society a adapté le jeu de cartes original au contexte flamand afin de soutenir le développement d'une IA digne de confiance en Flandre. Agoria a apporté les modifications nécessaires pour l'offrir à l'ensemble de l'écosystème belge.

AGORIA

Knowledge Centre Data & Society

Que sont les 'AI Blindspots' et comment les détecter ?

En Anglais, 'Blindspots' réfère aux «angles morts» ou faiblesses que nous pouvons rencontrer lors de la mise en place de projets IA.

Les faiblesses de l'IA peuvent résulter denégligences commises avant, pendant ou après le développement d'un système d'IA. Elles proviennent souvent de partis pris, de préjugés et de disparités structurelles au sein de la société.

Il est difficile de prédire leur conséquences mais elles peuvent cependant être atténuées en les détectant de manière proactive et en réagissant en conséquence.

Ces fiches aident à définir les éventuelles faiblesses et comment les détecter.

Chaque fiche contient une série de questions, un cas pratique et des outils/ astuces qui permettent de détecter et/ou atténuer les faiblesses.

La série de fiches comprend également une fiche joker pour vous permettre d'ajouter d'autres faiblesses potentielles de l'IA que vous ou votre équipe détectez.

Cette série de fiches est inspirée des fiches AI Blindspot d'Ania Calderon, Dan Taber, Hong Qu et Jeff Wen, développées lors du programme Assembly 2019 du MIT Media Lab et du Berkman Klein Center.

Le Knowledge Centre Data & Society a adapté le jeu de cartes original au contexte flamand afin de soutenir le développement d'une IA digne de confiance en Flandre. Agoria a apporté les modifications nécessaires pour l'offrir à l'ensemble de l'écosystème belge.

La série de fiches originale identifie trois phases (à savoir la planification, la construction et le déploiement). Elle se concentre uniquement sur la première phase, à savoir la planification. À un stade ultérieur, la série de fiches sera complétée avec des fiches des deux autres phases.

OBJECTIF

Au début d'un projet d'IA, déterminez l'objectif de votre système d'IA. Pour ce faire, vous devez impliquer les parties prenantes, des experts et votre équipe afin de définir clairement votre objectif et le problème qui sera résolu grâce à votre système d'IA.

PHASE: PLANIFICATION

Y AVEZ-VOUS PENSÉ ?

A. Avez-vous **clairement défini** le problème que vous cherchez à résoudre et le résultat que vous attendez ?
B. Cet **outil** est-il **adéquat** pour obtenir ce résultat ?
C. Toutes les **parties prenantes impliquées et affectées** reconnaissent-elles l'importance du problème ?
D. Avez-vous pris en considération les **avantages et inconvénients** de votre système d'IA pour chaque partie prenante ?
E. Comment comptez-vous **garantir le respect de la finalité de votre système d'IA** ?

À NE PAS FAIRE

Une entreprise a introduit un système d'IA pour accélérer son processus de production, mais cela a indirectement entraîné la perte des primes pour les employés. Comment cela aurait-il pu être évité ? Impliquez le syndicat dans votre projet et trouvez un moyen d'augmenter la vitesse sans pertes de primes.

OUTILS & ASTUCES

A&B : modèle de définition des problèmes
B : cours sur Machine Learning (Google)
C : transposez d'autres applications de votre Machine Learning à votre cas : est-ce toujours logique ?
D : cartographie des parties prenantes et validation

ÉQUILIBRE DES DONNÉES

L'équilibre des données signifie que vous avez vérifié la qualité représentative de vos données et que vous avez réfléchi à un moyen d'atténuer le déséquilibre.

PHASE: PLANIFICATION

.AGORIA Knowledge Centre Data & Society

ÉQUILIBRE DES DONNÉES

Y AVEZ-VOUS PENSÉ ?

A. Quelle est la **quantité minimale de données viables** dont vous avez besoin selon les experts dans le domaine ?
B. Qui/qu'est-ce qui pourrait être **exclu de vos données** ?
C. Comment les **limitations** de vos données affecteront-elles la nature représentative de votre modèle et les actions prises en charge par votre modèle ?
D. Si vos **données sont déséquilibrées**, pouvez-vous réduire ces limitations ?
E. Compte tenu de vos données, pouvez-vous décrire les personnes ou les cas pour lesquels vos **prévisions seront les moins fiables** ?

À NE PAS FAIRE

Après la sortie du très populaire Pokémon Go, plusieurs utilisateurs ont remarqué qu'il y avait moins de lieux Pokémon dans les quartiers principalement noirs. Cela s'explique par le fait que les créateurs des algorithmes n'ont pas réussi à fournir un ensemble de formation diversifié, et n'ont pas passé de temps dans ces quartiers.

OUTILS & ASTUCES

- A : entretien avec un expert du domaine
- B, C & D : Data Collection Bias Assessment, Aequitas
- E : créez un personnage d'homme/de femme invisible

GOUVERNANCE DES DONNÉES & VIE PRIVÉ

Les questions relatives à la gouvernance des données, et à l'impact sur la vie privée des personnes concernées, dont les données à caractère personnel seront traitées par le système d'IA, font partie de la phase de préparation de votre projet d'IA. Déterminer le niveau d'accès aux données et décrire le flux d'informations vous aideront à protéger les droits des personnes concernées.

PHASE: PLANIFICATION

.AGORIA Knowledge Centre Data & Society

Y AVEZ-VOUS PENSÉ ?

A. Pouvez-vous **traiter ou réutiliser légalement les données** ?
   - Si vous réutilisez les données, l'objectif poursuivi est-il le même ?
   - Les dispositions contractuelles appropriées ont-elles été établies ?
   - Pouvez-vous traiter ou réutiliser les données sur la base du consentement ou d'un autre fondement ?

B. Collectez-vous ou non des **données sensibles** ?

C. Existe-t-il des **régimes spéciaux pour protéger vos données** ?

D. Qui aura **accès** aux données (collectées) (en interne et en externe) ?

E. Pouvez-vous **respecter les droits de la personne** concernée repris dans le RGPD ?

À NE PAS FAIRE

Un hôpital britannique travaillant en collaboration avec Deepmind sur une application d'IA pour la détection et le diagnostic de lésions rénales a été condamné à une amende pour violation des règles sur les données personnelles. Il avait transféré des données personnelles sur 1,6 million de patients sans les en informer de manière adéquate.

OUTILS & ASTUCES

- 'Data Protection Impact Assessment'
- Cartographie des flux de données
- Entourez-vous de spécialistes en matière de confidentialité des données

COMPOSITION DE L'ÉQUIPE

Il est difficile d'être conscient d'éventuels problèmes (éthiques) si vous n'êtes pas conscient des préjugés existant au sein de votre équipe. Il est nécessaire de les mettre en lumière pour éviter de telles faiblesses. Connaissez les inconnues de votre équipe.

PHASE: PLANIFICATION

COMPOSITION DE L'ÉQUIPE

PHASE: PLANIFICATION

Y AVEZ-VOUS PENSÉ ?

A. Avez-vous pensé aux **partis pris existant** au sein de votre équipe ?
B. Votre **équipe** est-elle **diversifiée et multidisciplinaire** ou concernée par le problème que vous essayez de résoudre ?
C. Qui devriez-vous **inviter** pour démystifier cette idée reçue ?

À NE PAS FAIRE

Le logiciel de catégorisation des photos de Google a parfois confondu des personnes noires avec des gorilles. Le risque que cela se produise diminuerait considérablement si des membres noirs de l'équipe testaient le service.

OUTILS & ASTUCES

- A : test d'association implicite
- B : visite du site, carte d'empathie, personnage, ...

EXPERTISE TRANSVERSALE

Vous pouvez très bien être un expert en Machine Learning, mais ne pas l'être pas dans le domaine dans lequel vous utilisez cet apprentissage. Une solution est de se faire accompagner d'un expert qui attire votre attention sur les cas extrêmes typiques, les variables extrêmement importantes ou les pratiques courantes qui peuvent avoir un impact sur vos données.

PHASE: PLANIFICATION

Y AVEZ-VOUS PENSÉ ?

A. Avez-vous discuté avec des **experts du domaine** de la **quantité minimale de données viables** dont vous avez besoin pour permettre à votre système d'IA de remplir son objectif ?
B. Avez-vous fait appel à un expert pour comprendre quel **impact** votre algorithme devrait avoir ?
C. Quelles sont les **variables essentielles** pour votre problème ?
D. Avez-vous un expert à vos côtés pour vous aider à **évaluer les résultats** de votre algorithme ?

À NE PAS FAIRE

Un algorithme devait aider les urgentistes à déterminer les personnes devant être examinées pour une pneumonie au plus vite. Selon l'algorithme, les personnes asthmatiques n'avaient pas besoin de recevoir de soins immédiats. Les experts n'étaient pas d'accord avec cette conception car les cas d'asthme étaient toujours traités en priorité aux urgences. Les experts ont déclaré que cela était dû au fait que le système d'IA se basait sur des hypothèses erronées. Selon les données d'apprentissage, les patients asthmatiques passaient le moins de temps aux urgences. Par conséquent, le système ne leur avait accordé qu'une faible importance pour que le service des urgences soit plus efficace.

OUTILS & ASTUCES

- Entretien ou groupe de discussion avec un/des expert(s)
- Atelier sur les exigences techniques et celles des systèmes

USAGES ABUSIFS

Vous voulez créer un système d'IA pour améliorer quelque chose dans le monde. Cependant, en vous concentrant uniquement sur ses avantages, il est possible que vous négligiez ses nuisances éventuelles. Mieux vaut prévenir que guérir. Réfléchissez donc aux dommages qu'une personne vraiment malveillante pourrait causer à votre application ou à l'aide de celle-ci.

PHASE: PLANIFICATION

USAGES ABUSIFS

Y AVEZ-VOUS PENSÉ ?

A. Comment le système d'IA pourrait-il être utilisé de **manière contraire à l'éthique** ?
B. Quelles seraient les **conséquences** si votre système d'IA était utilisé de manière contraire à l'éthique ?
C. À qui avez-vous fait appel pour comprendre les **motivations sociales sous-jacentes et les modèles de menace** ?
D. Quelle est votre **stratégie d'atténuation** si votre système d'IA est utilisé de manière contraire à l'éthique ?
E. Que faire si votre **algorithme** développe un comportement **contraire à l'éthique** ?
F. Quels sont les **principes éthiques clés** que votre système d'IA devrait adopter ?

À NE PAS FAIRE

En 2016, Microsoft présentait Tay, un chatbot Twitter, au monde entier. Au bout de 24 heures, Tay avait adopté un tout autre comportement et était devenue utilisateur de Twitter raciste à cause des tweets qui lui avaient été adressés. Microsoft a donc décidé de la mettre hors ligne.

OUTILS & ASTUCES

- Créez des scénarios pour cerner les comportements malveillants et contraires à l'éthique de votre système, et cartographiez les conséquences de ces scénarios sur des spectateurs innocents.
- Faites-vous assister par des experts en sciences sociales et en droit

FICHE JOKER

PHASE: PLANIFICATION

.AGORIA Knowledge Centre Data & Society

FICHE JOKER **PHASE: PLANIFICATION**

Y AVEZ-VOUS PENSÉ ?

À NE PAS FAIRE

OUTILS & ASTUCES