ÉCOLE
CENTRALE LYON

Nouvelles Technologies de l'informatique
et de la communication

Mos 4.4

Etude de veille : Les failles et faiblesse de l'Intelligence Artificielle

Laetitia HAYE

Encadrants : Daniel MULLER
Nicolas JARDIN
Aranud DUBOS

11 mars 2021

Introduction

Ce rapport décrit la méthodologie utilisée pour mettre en place une veille efficace dans le cadre du MOS 4.4 Nouvelles Technologies de l'Informatique et de la Communication. L'objectif de la veille technologique a été de surveiller sur une période de deux mois les innovations et les connaissances émergentes sur le sujet des « failles et faiblesses de l'IA ». Le but était ainsi d'obtenir une vision d'ensemble des différents problèmes auxquels sont confrontés les systèmes d'intelligence artificielle et d'étudier les techniques récentes pour exploiter ou combler ces failles.

Pour mener une veille technologique pertinente, il faut suivre plusieurs étapes et exploiter les techniques d'acquisition, de tri et d'organisation des informations disponibles qui facilitent le travail d'état de l'art. Ce rapport est ainsi divisé en trois parties - Ciblage, Collecte, puis Traitement & Diffusion - et les outils utilisés tout au long de la veille sont présentés pour chaque étape.

Ciblage

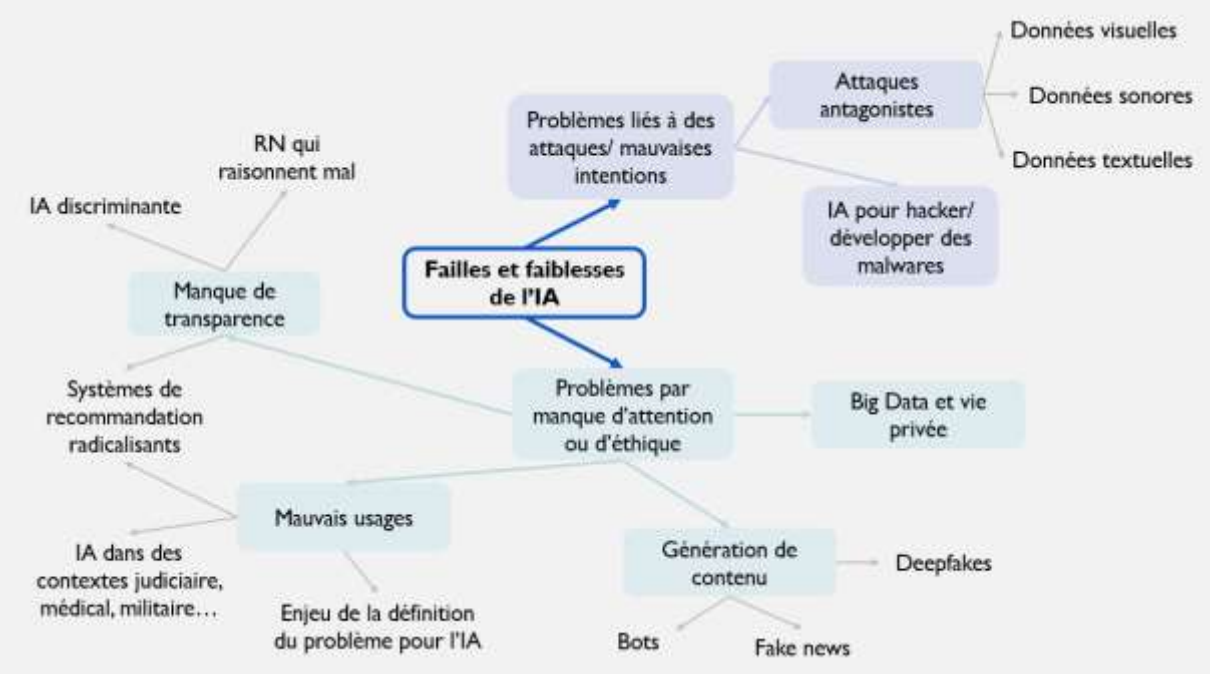
Avant de mettre en place le dispositif de veille, il a fallu délimiter clairement le sujet et expliciter les questions auxquelles je voulais répondre. Le sujet que j'ai choisi étant très large, j'ai défini les objectifs suivants :

- déterminer les failles techniques exploitables par des personnes malveillantes de systèmes artificiels à priori fonctionnels ;
- étudier les problèmes liés à des dysfonctionnements de système d'intelligence artificielle ;
- regarder les enjeux éthiques liés à l'utilisation de l'intelligence artificielle dans différents contextes.

A l'inverse, j'ai décidé d'éviter les éléments suivants :

- entrer dans le détail des deepfakes, des GANs ou de l'IA utilisée par les réseaux sociaux (cela aurait été trop long et ces sujets sont déjà traités par des camarades) ;
- étudier les problèmes juridiques liés à l'IA, notamment en ce qui concerne les algorithmes discriminatoires et la protection des données personnelles ;
- envisager des questions « philosophiques » sur les éventuels problèmes de l'IA à long terme (superintelligence, l'impact sur le marché du travail…).

Après avoir réalisé des premières recherches sur les problèmes liés à l'intelligence artificielle, j'ai pu distinguer différents sujets d'intérêts et affiner mes objectifs de veille. J'ai commencé à dresser une mindmap pour organiser la recherche et je l'ai enrichie au fur et à mesure des informations trouvées.



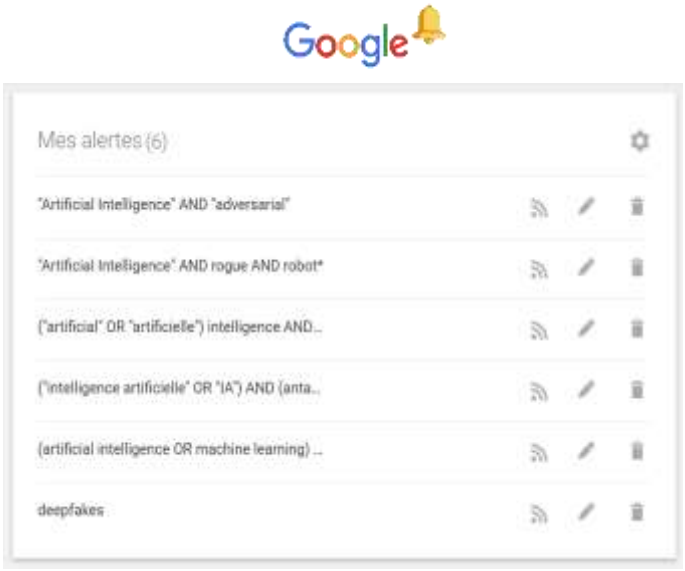
L'étape suivante a été d'identifier des sources d'informations fiables et adaptées au sujet. J'ai utilisé principalement :

- arxiv pour des publications scientifiques
-Youtube pour des exemples de deepfakes, des interviews de spécialistes et des conférences ou TED talks sur les dangers liés à l'IA.
- des articles et des blogs pour être informée de nouveautés.
- des revues spécialisées telles que MIT Technology Review pour aborder certains aspects techniques sans trop entrer dans les détails.

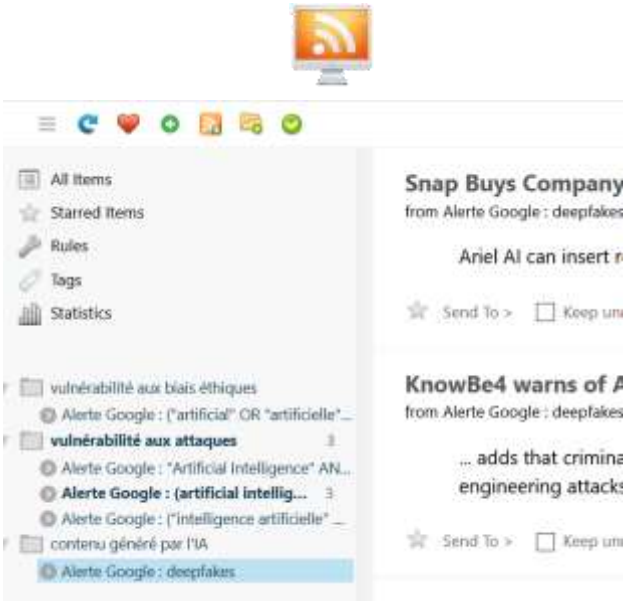
Collecte

La collecte des informations consiste à rassembler et stocker les documents et articles trouvés par le biais de différentes sources. J'ai utilisé Google Alerts pour être informée de la parution de nouveau contenu en rapport avec mon sujet. Les thèmes identifiés lors du premier tour d'horizon m'ont permis de fixer des mots-clés tels que « adversarial », « reverse engineering », « machine learning hacking ». Etant donné mon sujet, il a été nécessaire de combiner plusieurs mots-clés (par exemple « "Artificial Intelligence" AND rogue AND robot* »).

J'ai écrit toutes les requêtes en français et en anglais (en faisant attention aux différentes traductions possibles, par exemple « ("Artificial Intelligence" OR AI) AND "adversarial" » a donné « ("intelligence artificielle" OR "IA") AND (antagoniste* OR contradictoire*) ») pour avoir un spectre de résultats plus large, mais en pratique, la quasi-totalité du contenu trouvé était en anglais. Pour certaines alertes, je me suis rapidement rendu compte que j'avais mis des termes trop génériques qui englobent des éléments en dehors du sujet (notamment avec les mots « deepfakes » et « hack »). J'ai utilisé des termes plus spécifiques par la suite.



J'ai ensuite utilisé Feedbro pour rassembler en un même endroit tous les flux RSS générés. J'ai créé des dossiers par type de problème pour automatiser le tri des résultats. En plus de me faire gagner beaucoup de temps, cet outil a facilité l'analyse et la suppression des articles qui n'étaient pas adaptés au sujet.



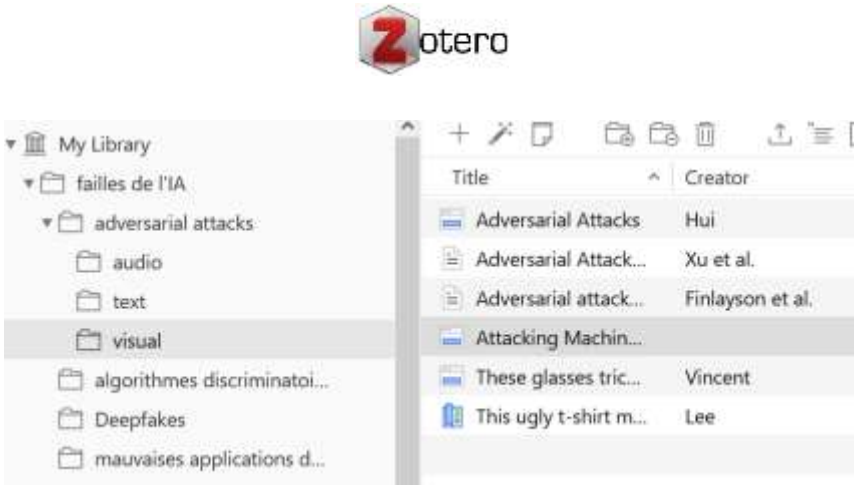
J'ai également recherché des informations manuellement, d'une part pour me familiariser avec certaines technologies de machine learning (nécessaire pour comprendre leurs failles) et d'autre part, parce que les deux mois de veille ont été un peu court pour voir apparaître une multitude de nouveaux problèmes liés à l'IA ou de solutions pour y remédier.

Une difficulté que j'ai rencontrée à été de trouver un équilibre entre les publications scientifiques très poussées (sujet trop vaste pour entrer dans les détails de chaque thème) et les articles un peu superficiels et sensationnels.

Traitement et partage des informations

Afin de pouvoir stocker et organiser les différentes sources d'informations quel que soit leur type (articles, PDF, vidéos youtube, etc.), j'ai utilisé l'outil de curation Wakelet. C'est un outil très simple à utiliser et le fait de pouvoir déplacer facilement une source par rapport à une autre m'a permis d'organiser mes idées en triant les ressources par thème.

J'ai également utilisé le logiciel de gestion bibliographique Zotero. Il est gratuit et facile à prendre en main, permet de gérer les références de différentes sources de données et j'ai apprécié pouvoir créer des dossiers parce que cela m'a permis de retrouver plus facilement les ressources au moment de créer la synthèse de veille.



Pour le partage d'informations, je n'ai pas créé de compte Twitter mais j'ai utilisé messenger pour communiquer avec des amis qui étudient également l'IA et leur faire part de mon sujet. Cela m'a permis de découvrir d'autres articles.

J'ai également rendu ma page Wakelet consultable en ligne : la collection créée pour cette veille est disponible via le lien suivant : https://wakelet.com/wake/Zrn5f4D6s9Ayd_M7fWheq/edit.

Enfin, j'ai partagé publiquement une page web HTML hébergée sur Github.io qui synthétise les informations trouvées au cours de la veille.

Conclusion

Ce rapport résume les différentes étapes qui ont guidé ma veille et les outils utilisés pour chacune d'entre elles (cf schéma ci-dessous pour une vision synthétique de l'ensemble). Ce MOS m'a permis de découvrir des outils que je ne connaissais pas qui font gagner beaucoup de temps et qui pourront m'être utiles par la suite. Il a également été l'occasion d'appliquer à un cas concret la méthodologie d'une veille technologique et d'apprendre de nouvelles choses sur un sujet qui m'intéresse.

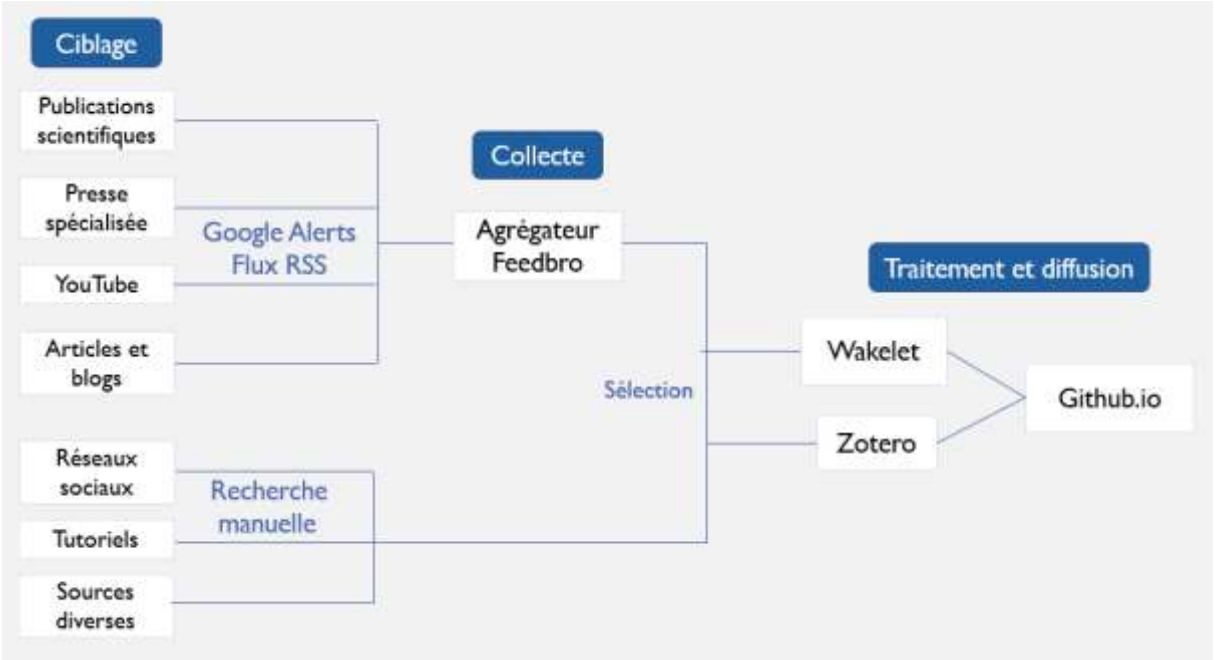


Figure 1 - Dispositif de veille mis en place