UCLouvain

**COMMUNIQUÉ DE PRESSE**

3 juin 2024

## Recherche UCLouvain

# Comment lutter contre l'intelligence artificielle extrémiste ?

**EN BREF :**

- Si **l'intelligence artificielle** est source de progrès, son double usage par **des groupes extrémistes et terroristes** peut être **dévastateur**.
- Stéphane Baele, professeur à l'UCLouvain FUCaM Mons, a analysé les menaces dans un rapport et donne quelques **recommandations** pour les atténuer.

**CONTACT PRESSE :**

**Stéphane Baele**, professeur de sciences politiques à l'UCLouvain : stephane.baele@uclouvain.be

Ces dernières années, l'**intelligence artificielle** (IA) connait un **développement exponentiel** et s'est imposée dans tous les domaines de notre vie. Elle permet des progrès considérables en médecine, en sport et dans bien d'autres secteurs, si elle est utilisée à bon escient. Or, parallèlement à l'explosion de l'IA, le domaine numérique a vu un autre phénomène s'amplifier : l'**extrémisme**. Et la combinaison IA + extrémisme peut avoir **des effets dévastateurs**, comme l'analyse un rapport que vient de publier le professeur Stéphane Baele.

*« Comme chaque percée technologique passée, l'IA sera, et est déjà, utilisée de diverses manières par des entrepreneurs de la haine et de la violence pour renforcer des agendas extrémistes »*, affirme le professeur de science politique sur le campus FUCaM de l'UCLouvain. En examinant « **l'extrémisme IA** », soit la rencontre toxique et inéluctable de ces deux évolutions, il **comble un vide dans la littérature scientifique** qui s'était jusqu'à présent plutôt penchée sur les dimensions éthiques et légales de l'IA.

Dans son rapport, le chercheur s'attelle à **identifier les différentes technologies liées à l'IA et les tactiques** que des acteurs extrémistes pourraient mettre, voire mettent déjà en œuvre au service de leur cause.

**Trois technologies d'IA identifiées**

- Les **générateurs de contenu** (type ChatGPT, Midjourney…) peuvent déjà produire des photos, vidéos, audios synthétiques (les **deepfakes**) impossibles à distinguer d'un « vrai » contenu. *« Citons par exemple le deepfake audio imitant la voix de Joe Biden appelant à ne pas voter à la primaire démocrate, ou celui d'un candidat à l'élection présidentielle slovaque complotant un trucage des élections. »* Ces logiciels, dont l'usage se banalise, ont un potentiel de propagande et de désinformation énorme.
- La **« pattern-recognition AI »**, basée sur la reconnaissance des formes, a notamment permis de booster la recherche contre le cancer. Mais un double-usage de cette technologie, au lieu de fabriquer des remèdes, pourrait servir à **générer des poisons**. *« En 2022, en inversant leurs instructions, des chercheurs utilisant un modèle d'IA pour la découverte de médicaments ont fait produire au modèle des milliers d'agents neurotoxiques pouvant être utilisés comme des armes chimiques hautement léthales. »*
- Enfin, les **modèles de décisions stratégiques**, qui sont passés du champ des jeux de stratégie au domaine militaire, présentent des risques par le biais des **armes autonomes** (drones, véhicules blindés…). *« Il est illusoire de penser que ces armes resteront dans le champ des armées nationales et ne percoleront pas au sein d'organisations terroristes. »* Pensons par exemple à la proximité d'une organisation comme le Hezbollah avec l'Iran, grand fabricant et fournisseur de drones…

Les progrès de l'IA étant actuellement exponentiels et le cadre légal n'étant qu'embryonnaire, les modèles d'IA générative sont déjà sujets aux abus, tandis que les autres modèles ne sont pas à l'abri de détournements. La publication de ce rapport entend poursuivre un double objectif : informer sur l'émergence de ce nouveau problème de sécurité, mais formuler également certaines **recommandations**.

**Limiter l'open source ?**
*« Les plateformes de médias sociaux doivent intensifier leur développement d'outils de détection de contenus générés par IA et clairement les identifier. Les gouvernements devraient encourager les principaux acteurs de l'IA à intégrer des garde-fous plus solides dans l'usage de leurs technologies. »*

Stéphane Baele pointe aussi une question plus existentielle, touchant aux fondements d'internet : l'open source. *« Tant les gouvernements que le secteur académique et les acteurs privés doivent se questionner et faire la balance bénéfices/risques de **l'accès public aux informations liées à l'IA**. Des directives collectives doivent être imposées pour éviter que des modèles et données sensibles ne tombent dans le domaine public, mais aussi pour protéger les chercheurs à l'origine de ces modèles. »*

**Un guide anti-terroriste IA**
Enfin, le politologue en appelle à une **approche proactive et offensive des autorités**. *« Entre les démocraties libérales et les régimes autoritaires, le rapport de force est asymétrique. L'usage de l'IA par les extrémistes, tel que décrit dans ce rapport, devrait être exploité par nos services de renseignement et de sécurité pour contrer les dynamiques extrémistes. »* Un peu comme la cybersécurité a appris des hackers…

*AI Extremism. Technologies, tactics, actors* est publié et disponible en libre accès sur voxpol.eu. Coordonné par le professeur Stuart MacDonald du Centre de recherche sur les cybermenaces (CYTREC) de l'université de Swansea, VOX-Pol est un réseau de recherche de premier plan au niveau mondial sur l'extrémisme et le terrorisme en ligne. Depuis 2022, Stéphane Baele fait partie de son comité de direction.

Stéphane Baele est professeur de sciences politiques à l'UCLouvain FUCaM Mons. Il a travaillé durant 9 ans à l'université d'Exeter en tant qu'*associate professor*. Ses recherches se focalisent sur la communication des acteurs politiques violents et extrémistes.


**Service presse** presse@uclouvain.be | LLN et institutionnel : **Isabelle Decoster** +32 486 42 62 20 et **Céline Gransard** +32 479 92 16 65 | Bruxelles : **Dominique Hoebeke** +32 496 115 177 | Hainaut : **Ugo Petropoulos** +32 498 42 15 13