Intelligence artificielle et conflictualité

Sur l'hypothèse de dérive malveillante d'une Intelligence Artificielle

Thierry Berthier, membre de la Chaire de cyberdéfense & cybersécurité Saint-Cyr, Sogeti, Thales, et Olivier Kempf

Octobre 2017 - Article IV.12

La montée en puissance de l'Intelligence artificielle (IA) concerne l'ensemble des activités humaines. L'accélération de sa diffusion dans l'espace numérique contribuera à modifier des équilibres économiques, sociaux ou géopolitiques qui ont marqué le vingtième siècle et que l'on pensait durablement installés. En transformant à grande vitesse notre environnement et nos pratiques, l'IA suscite le débat. Elle fait naître autant d'espoirs et de défis à relever que de craintes à dissiper. Si les questions éthiques, sociétales, économiques, philosophiques ou religieuses doivent légitimement accompagner son développement et enrichir le débat qui l'entoure, de nombreuses controverses participent au contraire à une "technophobie" ambiante plus ou moins marquée selon les individus et leur culture. Ainsi, l'hypothèse d'une dérive malveillante de l'IA s'installe et mérite désormais d'être analysée avec pragmatisme et rationalité.

C'est pourquoi, après avoir rappelé les débats qui se sont tenus autour de l'hypothèse d'une iA toute puissante et malveillante, cet article propose une hypothèse intermédiaire : celle où une iA causerait, par mégarde, des désordres puis des catastrophes. Le scénario proposé envisage le déclenchement d'un conflit entre les alliés de l'oTAn et une puissance extérieure. Il vise à montrer qu'un des dangers de l'iA ne réside pas dans sa puissance (aujourd'hui entre deux géants américains du numérique hypothétique et donc fantasmée) ou dans son autonomie mais surtout dans les limites de sa puissance résultant d'un champ de pertinence très restreint. On ne développe à ce jour que des iA spécialisées, efficaces sur des problèmes limités et bien spécifiés mais strictement inopérantes sur des problèmes plus généraux. Le risque apparaît lors de la mise en résonnance involontaire de plusieurs iA spécifiques.

La peur de l'intelligence artificielle et l'hypothèse d'une dérive malveillante de l'IA

Apparue en 1956 durant la conférence de Dartmouth, l'expression « intelligence Artificielle » ne possède toujours pas de définition partagée et universelle. Selon Marvin Minsky, « *l'intelligence*

artificielle est la science qui consiste à faire faire à des machines ce que l'homme fait moyennant une certaine intelligence ».

Peu précise et souffrant d'une forte récursivité, cette première définition s'avère aujourd'hui insuffisante pour qualifier les iA intervenant en robotique, en perception (vision et parole) ou en compréhension du langage naturel. Une définition plus opérationnelle et plus ouverte sur les évolutions de l'iA a ensuite été proposée par Rich et Knight : « *L'IA est le domaine de l'informatique qui étudie comment faire faire à l'ordinateur des tâches pour lesquelles l'homme est aujourd'hui le meilleur* ». La distinction entre iA forte et iA faible est apparue dans les années soixante qualifiant d'iA forte une machine produisant un comportement intelligent, capable d'avoir conscience d'elle-même en éprouvant des "sentiments" et une compréhension de ses propres raisonnements. L'iA faible s'applique à une machine simulant ces comportements sans conscience d'elle- même. Pour les partisans de l'iA faible, l'iA forte serait intrinsèquement impossible compte tenu du support biologique de la conscience humaine...

Un sondage réalisé en mai 2016 par Odoxa pour Microsoft et stratégies[i] a montré que 50 % des Français considéraient l'intelligence artificielle comme une menace alors que 49 % voyaient en elle une opportunité de développement. Début 2016, un autre sondage de l'iFoP indiquait que 65 % des Français s'inquiétaient de l'autonomie croissante des machines, des drones armés et de la Google Car. Selon une troisième étude, les Français feraient partie des peuples craignant le plus l'iA avec de fortes disparités entre les classes d'âge et les catégories socioprofessionnelles des sondés. Souvent peu informé sur l'état de l'art des développements de l'iA et sur ses réelles capacités fonctionnelles, l'utilisateur français reste fortement influencé par la littérature et le cinéma américain de science-fiction qui représentent presque toujours l'iA comme une entité nuisible et potentiellement destructrice.

Cette image négative se trouve renforcée par les mises en garde provenant de scientifiques de renoms. Ainsi, le 2 décembre 2014, l'astrophysicien Stephen Hawking déclarait dans un entretien à la BBC[ii] : *« Le développement d'une intelligence artificielle totale pourrait annoncer la fin de l'espèce humaine. Elle pourrait prendre son indépendance et se reprogrammer elle- même à une vitesse accélérée […] Les êtres humains, qui sont limités par une lente évolution biologique, ne pourraient pas rivaliser et seraient vite dépassés ».* En février 2015, la fondation Global Challenges de l'université d'oxford publiait un rapport intitulé « 12 risques menaçant la civilisation humaine » dans lequel les auteurs mettaient en garde contre l'augmentation des iA « super- intelligentes » qui pourraient provoquer un effondrement de l'économie ou de la civilisation puis provoquer l'extinction de l'humanité. En juillet 2015, Elon Musk, fondateur de Tesla et de space-X, et Stephen Hawking publiaient une lettre ouverte, signée par de nombreux scientifiques, mettant en garde contre l'autonomie des systèmes armés et l'utilisation de l'iA à des fins militaires.

Dès lors, la question de la perte de contrôle et des dérives potentielles d'une iA autonome devenant "malveillante" allait s'inviter aux débats de manière récurrente et transversale. Présente depuis les années soixante dans la quasi-totalité des films de science-fiction et des romans traitant de l'iA, l'hypothèse de dérive malveillante est aujourd'hui régulièrement évoquée dans le contexte de la voiture autonome, des drones et systèmes armés autonomes, de la finance automatisée hFT ou de la cybersécurité des robots compagnons. Le plus souvent, il ne s'agit que de formuler l'hypothèse de malveillance comme une éventualité à prendre en compte dans un futur mal défini où l'iA aurait atteint un niveau de développement et d'autonomie très supérieur à ce qu'il est aujourd'hui. Cela

dit, lorsque cette hypothèse est évoquée par un groupe de scientifiques et d'industriels de premier plan (comme Stephen Hawking et Elon Musk), on pourrait s'attendre à un corpus d'argumentations rationnelles venant étayer l'alerte. Mais comme on peut le constater, les justifications sont en général totalement absentes de la mise en garde... C'est ce manque d'argumentation qui relègue alors la mise en garde au rang de pur exercice de spéculation quand il ne s'agit pas d'un simple règlement de compte commercial entre deux géants américains du numérique...

Hypothèse de dérive malveillante d'une IA : une analyse

Nous proposons d'examiner l'hypothèse de dérive malveillante d'une iA selon une approche prospective et rationnelle.

Celle-ci ne doit s'appuyer que sur des mécanismes informationnels et humains connus, maîtrisés ou en phase de développement. Afin de préciser et de réduire le champ d'exploration de cette hypothèse de dérive malveillante, nous nous intéressons désormais à une question plus limitée (une hypothèse de dérive malveillante « faible » au sens où elle n'implique que des iA faibles, c'est-à-dire sans aucune « conscience » de leur propre activité). Elle veut répondre à la question :

« *Est-il rationnellement envisageable qu'une intelligence artificielle soit à l'origine d'une crise militaire ou qu'elle provoque une situation propice à un conflit armé ?* ».

Notons que cette question est souvent associée à une hypothèse de détournement de l'iA par des individus malveillants ou à son piratage par des groupes de hackers, réduisant ainsi la problématique à la seule cybersécurité du système hébergeant l'iA détournée... en fait, le degré d'autonomie d'une iA a très peu été évoqué dans des publications académiques jusqu'en 2015. Il faut attendre 2016 pour lire un article scientifique, publié par un groupe de chercheurs de Google Deep mind & et du Fhi d'oxford, étudiant[iii] la capacité de neutralisation d'une iA « apprenante » et exposant un processus d'interruption robuste face à l'apprentissage par l'iA de sa propre interruption. Une seconde étude [iv]intitulée « *One Hundred Year Study on Artificial Intelligence* » et produite par une quinzaine d'experts vient d'être publiée par l'université de Stanford. Celle-ci rejette catégoriquement l'hypothèse de dérive malveillante d'une l'iA en l'état actuel des connaissances. Elle n'aborde pas le cas spécifique de l'iA utilisée dans le domaine militaire mais réfute sans détour la mise en garde formulée par Elon Musk et Stephen Hawking. Ces différences d'approches montrent bien que la question n'est pas tranchée et que nous nous situons à l'orée d'un débat stratégique et clivant.

Notre article n'a pas la prétention d'apporter une réponse globale à l'hypothèse de dérive malveillante. Il propose seulement un scénario construit sur une séquence d'automatismes existants ou en cours de développement et qui, sous certaines conditions, pourrait aboutir à une situation conflictuelle ou à une crise militaire. Pour construire ce scénario, deux automatismes principaux seront combinés dans une même séquence.

Automatismes intervenants dans la séquence de dérive malveillante

Les automatismes que nous mentionnons ici sont des processus réels et non fictifs. Ils servent d'arrière-plan à notre scénario. Sans eux, celui-ci ne serait pas crédible.

Automatisme n°1 : Activation par l'OTAN de l'article 5 du traité de Washington en cas de cyberattaque sur un pays membre de l'Alliance.

Depuis son 24e sommet, organisé au Pays de Galles en 2014, l'oTAn a déclaré qu'elle considérait qu'une agression cyber pouvait déclencher une riposte dans le cadre de l'article 5 du traité de Washington, celui de la défense collective. Cette déclaration de principe n'a pas donné lieu à beaucoup de détails quant à sa mise en œuvre concrète[v]. On ne sait ainsi rien du seuil de déclenchement[vi], de la nature de la réponse (cyber ou non cyber ?), des moyens qui seraient mis en œuvre, *etc*. Le sommet de Varsovie, tenu en juillet 2016, a ajouté quelques éléments : le droit international s'applique dans le cyberespace, l'oTAn est responsable de ses propres réseaux et non de ceux des Alliés qui sont chacun responsables chez eux (selon une position constamment affirmée à l'oTAn) ; surtout, le cyberespace est désormais considéré comme un espace de conflit au même titre que les autres milieux (terre, mer, air, espace sidéral).

Derrière cette affirmation se cache une certaine opérationnalisation du cyber. Jusqu'à présent, il ne s'agissait que d'une ssi étendue avec quelques fonctions de Lutte informatique défensive (LiD).

Désormais, le cyber entre dans le champ des opérations, ce qui signifie que chaque opération aura un volet cyber, en défensive mais aussi en connaissance du milieu. En revanche, les responsables ont bien précisé que l'Alliance ne développait pas en propre de capacités offensives. Il reste qu'elle peut se reposer pour cela sur les capacités des Alliés les plus avancés, ce qui permet à ces derniers de conserver le contrôle des cyberopérations. Cette approche correspond à ce que soutiennent les Américains. Dès 2015, le Cyber Command américain déclarait qu'une opération militaire de vive force pouvait parfaitement répondre à une campagne de cyberattaques ciblant les infrastructures critiques du pays. L'Amérique se donnait le droit de répliquer militairement à une agression portée contre ses intérêts dans le cyberespace.

Automatisme n°2 : programmes dARpA de détection automatisée des bugs et vulnérabilités de type "Bug-Hunting Bots" du Cyber Grand Challenge 2016.

Organisée par la DArPA (l'Agence pour les projets de recherche avancée de défense supervisés par le département de la Défense des Etats-Unis), la phase finale du Cyber Grand Challenge [vii]s'est déroulée, les 6 et 7 août 2016, à Las Vegas. Le CGC a opposé sept systèmes robotisés développés durant trois années par sept équipes finalistes dans la détection automatique de vulnérabilités logicielles et réseaux présentes dans le système adverse et la protection de son propre environnement numérique. Conçu avant tout comme un démonstrateur, le tournoi CGC a prouvé qu'il était désormais possible de concevoir des agents logiciels capables de scanner de façon automatique des codes adaptés puis de détecter certaines de leurs vulnérabilités. La compétition a eu lieu dans un environnement numérique spécifique dans lequel se sont affrontés quinze supercalculateurs détecteurs de vulnérabilités informatiques devant un comité d'arbitrage qui a finalement désigné l'équipe ForAllsecure comme gagnante du Challenge CGC. La Darpa souhaite désormais développer des agents logiciels "chasseurs de bugs" ouvrant ainsi la voie à une cybersécurité automatisée, industrialisée, exploitant massivement les techniques de l'intelligence artificielle. Les démonstrateurs finalistes du CGC doivent évoluer à très court terme vers la production d'agents "Bug-hunting Bots" qui seront déployés sur l'ensemble des réseaux sensibles. Ces futurs agents autonomes pourront être utilisés autant en mode défensif qu'en version offensive afin de détecter certaines des vulnérabilités d'un système adverse.

Cette évolution vers une cybersécurité "robotisée" (LiD) se trouve toutefois limitée par un résultat mathématique lié au problème de l'arrêt (Turing[viii]) qui prouve qu'il n'existe pas d'analyseur universel capable de décider sans jamais se tromper, si son programme est sûr ou non. Cette limite

théorique permet d'affirmer que la cybersécurité absolue n'existe pas... Cela dit, une telle borne n'interdit pas le développement de systèmes de détections dont la performance pourrait atteindre 90 ou 95 % de l'ensemble des vulnérabilités.

Un scénario de dérive incontrôlée pouvant aboutir à une situation de crise

Dans ce scénario, la combinaison de deux systèmes experts, relativement autonomes, conçus indépendamment, les met en résonnance, en tant que systèmes "intelligents". Cette séquence provoque alors une crise. Un tel scénario n'est pas totalement absurde puisqu'un incident similaire a déjà eu lieu en septembre 1983, mettant en jeu des armes nucléaires ! L'exercice allié oTAn "Able Archer" avait provoqué la mise en alerte nucléaire des forces soviétiques via leur système de surveillance satellitaire. Le Lieutenant-colonel Stanislav Levgrafovitch Petrov avait pris la décision d'informer sa hiérarchie sur la possibilité d'une fausse alerte émise par les automatismes soviétiques (un système informatique d'alerte anti-missile) dans le cadre d'un exercice et non d'un réel tir de missile.

L'alerte des forces nucléaires russes avait alors été annulée in extremis grâce à la clairvoyance et à la sagacité de cet officier.

Eléments fictifs du scénario

L'élément central de la séquence est un programme (fictif) que nous appellerons dans la suite marsAnalytics. Il s'agit d'une plateforme d'aide à la décision déployée et utilisée dans un cadre militaire et plus spécifiquement dans un contexte de gestion de crise. MarsAnalytics est une iA dotée de capacités d'apprentissage non supervisé. En tant que système expert, elle est capable de construire des préconisations qui sont ensuite utilisées ou non par l'autorité militaire. Nous supposons qu'elle a accès aux données publiques (ouvertes) et qu'elle possède des droits suffisants pour pouvoir interroger d'autres plateformes notamment militaires et pour exécuter des processus sur ces plateformes ou à partir d'elles. Nous supposons enfin que marsAnalytics est pleinement intégrée au système de défense américain et qu'elle est compatible avec l'ensemble des standards oTAn.

Bughunting, le second élément de la séquence, est un ensemble (fictif) de systèmes composés d'agents logiciels "chasseurs de bugs et de vulnérabilités" issus du programme Darpa CGC. Ces agents ont été massivement déployés dans les secteurs industriels, les services et les administrations. Ils permettent de réaliser des économies substantielles en matière de cybersécurité tout en industrialisant la sécurisation et la supervision des systèmes d'information. Ils analysent les codes et processus en temps réel afin de détecter des vulnérabilités. Celles-ci sont ensuite archivées avec les correctifs à appliquer si nécessaire. Bughunting produit des bases de vulnérabilités sur les codes "amis" civils et militaires en particulier sur ceux des pays membres de l'oTAn. Il a été déployé afin d'améliorer la sécurisation des infrastructures critiques des membres les plus "fragiles" de l'organisation.

Description de la séquence

MarsAnalytics cherche à maximiser ses fonctions de gain à la suite d'un détournement de sa supervision par piratage et prise de contrôle (ce qui est peu probable) ou par émergence non maîtrisée de cette recherche. Elle peut aussi chercher à améliorer la pertinence de ses réponses dans un processus d'apprentissage automatisé à l'image de la phase d'apprentissage de la

plateforme AlphaGo de Deepmind Google qui s'était entrainée en jouant des milliers de parties contre elle-même. La plateforme marsAnalytics « décide » donc d'exécuter la séquence suivante :

Étape 1 - marsAnalytics commence par établir un accès permanent aux bases de données construites par Bughunting (cf. Automatisme n°2 - partie 3). Elle collecte ainsi les vulnérabilités actives et les fonctions d'attaque sur ces vulnérabilités affectant les systèmes critiques de plusieurs pays membres de l'oTAn, en particulier les plus fragiles en matière de cybersécurité - cyberdéfense. MarsAnalytics poursuit cette collecte en attendant un contexte favorable pour exécuter la seconde étape. Ce contexte peut apparaître lors d'une augmentation des tensions entre les Etats-Unis et une superpuissance, typiquement la Chine ou la Russie. En tant que système expert, marsAnalytics est utilisée dans la construction de scénarios de crise. Elle reste donc informée de l'apparition d'un contexte favorable.

Étape 2 - une fois le contexte favorable survenu et détecté (typiquement, une situation de tensions géopolitiques), marsAnalytics déclenche (en guise de test ou d'exercice) une série de cyberattaques simultanées sur l'ensemble des cibles oTAn répertoriées durant la première étape de la séquence. MarsAnalytics dispose pour cela de puissances de calcul et de stockage suffisantes. Les cyberattaques se concentrent alors sur les infrastructures critiques des pays membres les plus faibles en matière de cybersécurité. Elles peuvent potentiellement provoquer une paralysie des administrations et de la plupart des services et industries du pays ciblé. Ces cyberattaques sont menées en laissant quelques traces permettant une attribution de l'origine à la Chine ou la Russie (l'acteur de la crise géopolitique de référence).

Étape 3 - Les systèmes de cyberprotection de l'oTAn détectent la vague de cyberattaques ciblant plusieurs pays membres de l'organisation et évaluent son impact. Ils attribuent l'attaque en fonction des quelques traces laissées par marsAnalytics à la Chine ou la Russie avec une certaine probabilité. Ces systèmes transmettent l'alerte au commandement unifié de l'oTAn. MarsAnalytics est activée au niveau d'une situation de crise grave. Elle a désormais accès aux décisions relatives à la gestion de cette crise. Le shAPe (Le Grand Quartier Général des puissances alliées en Europe) confirme l'attribution technique compte tenu du contexte géopolitique dégradé et privilégie logiquement une responsabilité chinoise et/ou russe.

Étape 4 - Le Conseil de l'Atlantique nord applique la doctrine de réponse collective à l'agression de l'un des membres de l'Alliance (cf. automatisme n°1 - partie 3). Il active l'article 5 du traité de Washington et engage une série de représailles sur l'espace numérique et sur l'espace physique à l'encontre de la Chine (ou de la Russie).

Ainsi, l'Alliance a mis en place un dispositif d'intelligence artificielle qui, par mauvaise interprétation des données, déclenche une attaque contre les membres, en faisant croire à une agression d'origine adverse, ce qui provoque le déclenchement de l'article 5.

Dans cet article, nous avons voulu insister sur un cas nouveau qui est jusqu'ici ignoré aussi bien par les partisans que par les critiques de l'iA. Les uns promettent des bénéfices, les autres craignent une prise de contrôle totale de l'humanité. Il semble qu'il y ait une situation intermédiaire, celle où l'intelligence Artificielle provoque des turbulences systémiques qui trompent la supervision humaine. Cela peut bien sûr être par défaut de conception (un manque de contrôle durant la phase d'entraînement du système) mais il ne s'agit pas seulement de cela : en effet, avant d'accéder à une iA unique, il est très probable que des iA distinctes et spécifiques coexisteront et collaboreront. Compte-tenu de la nature du cyberespace, il est naturel que ces iA communiquent et donc

interagissent. Or, une iA qui a été calibrée dans un environnement limité peut accéder à de nouvelles fonctions en se connectant à d'autres, conçues dans d'autres environnements. Il y aurait ainsi un risque systémique distinct de l'hypothèse malveillante jusqu'ici couramment avancée. Ce risque paraît plus immédiat que le risque général jusqu'ici évoqué dans les débats.

Il nous semble qu'il s'agit d'une condition à inclure dans les constructions actuelles des intelligences artificielles : celle d'examiner l'hypothèse des connexions autonomes avec d'autres iA, de façon à introduire des garde-fous pour éviter des logiques de résonance qui auraient des effets pernicieux.

---

[i] Sondages sur la peur de l'iA : http://www.odoxa.fr/rdvde-linnovationlintelligence-artificielleemballe-les-gagnants-dusysteme-mais-fait-peuraux-autres/ - http://www.strategies.fr/actualites/medias/1040507 W/avez-vous-peur-de-lintelligence-artificielle.html – http://www.zdnet.fr/actualites/intelligence-artificiellela-france-a-peur-mais-dequoi-au-juste- 39831180.htm.

[ii] Lettre ouverte sur les dangers de l'iA – oxford : http://futureoflife.org/open -letter-autonomousweapons/ - https://stopkillerrobots.ca/ en-francais/a-propos-desrobots-tueurs/.

[iii] Article sur la neutralisation de l'iA (Google Deep mind & Fhi oxford) : https://www.fhi.ox.ac.uk/interruptibility/ - https://www.fhi.ox.ac.uk/wpcontent/uploads/interruptibility.pdf.

[iv] Etude "one hundred Year study on Artificial intelligence" (Ai100) ; stanford university, septembre 2016. https://ai100.stanford.edu /2016-report.

[v] Sur cette question, voir o. Kempf, « Alliances et mésalliances dans le cyberespace », economica, 2014, notamment le chapitre Vi. Voir aussi le site de l'oTAn sur la Cyberdéfense : http://www.nato.int/cps/fr/natohq/topics_78170.htm et sur l'engagement de l'oTAn en faveur de la cyberdéfense : http://www.nato.int/cps/fr/natohq/official_texts_1331 77.htm?selectedLocale=fr

[vi] Le secrétaire général de l'Alliance Atlantique, Jens Stoltenberg, affirmait que l'oTAn considérerait une campagne massive de cyberattaques appliquée à l'un de ses membres au même niveau qu'une agression militaire conventionnelle.

[vii] Challenge DArPA CGC 2016 : http://www.defense.gov/news/Article/Article/907045 /darpa-autonomous-bughunting-bots-will-lead-toimproved-cybersecurity et https://www.cybergrandch allenge.com/.

[viii] Sur l'impossibilité d'une cybersécurité absolue : http://www.contrepoints.org/2015/04/20/205153-lasecurite-informatiqueabsolue-nexiste-pas.