SciencesPo

CHAIR DIGITAL, GOVERNANCE AND SOVEREIGNTY

Les systèmes autonomes de contrôle normatif dans les applications militaires de l'IA : vers une norme industrielle éthiquement et légalement responsable pour une technologie militaire *fabriquée en Europe*

Dr. Johannes Thumfart
Chercheur senior et postdoctoral
Groupe de recherche Droit, science, technologie et société (LSTS)
Vrije Universiteit Brussel

Novembre 2023

"Les États devraient réfléchir à la manière d'utiliser les capacités militaires d'IA pour renforcer leur mise en œuvre du droit international humanitaire et améliorer la protection des civils et des biens de caractère civil dans les conflits armés.

***Déclaration politique sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie*, 16 février 2023.**

Table des matières

Résumé

Cette note politique traite des applications militaires de l'IA dans le sens de systèmes d'armes létales partiellement autonomes (PALWS) et d'unités logistiques d'IA. Les systèmes que j'appelle "systèmes autonomes de contrôle normatif" (ASNC) sont comparables aux systèmes d'assistance intelligente à la vitesse (ISA) dans les voitures. Les systèmes ISA alertent ou corrigent les conducteurs lorsqu'ils dépassent la limite de vitesse en utilisant la reconnaissance des panneaux de signalisation et des bases de données sur les limites de vitesse liées à des données de géoposition. De même, les ASNC devraient bloquer l'utilisation illégale des applications militaires de l'IA, par exemple dans le cas d'une guerre d'agression, ou alerter les commandants si une action est disproportionnée ou si une cible choisie est civile.

Je promeus une approche centrée sur la technologie, qui est conforme à la *Déclaration politique* multilatérale de 2023 *sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie* et aux recommandations techniques du rapport de la session de 2023 du *Groupe d'experts gouvernementaux des Nations Unies sur les technologies émergentes dans le domaine des systèmes d'armes autonomes létaux*. Je soutiens que les PALWS et les unités logistiques d'IA dans l'armée devraient être équipés d'ASNC pour contribuer à garantir qu'ils sont utilisés dans le respect du droit international humanitaire (DIH), en particulier des principes de proportionnalité et de minimisation des dommages. En outre, les ASNC devraient inclure des mécanismes de blocage afin de garantir que les PALWS et les unités logistiques d'IA ne soient pas utilisés dans des guerres d'agression ou contre des manifestants pacifiques nationaux. D'un point de vue technologique, les ASNC nécessiteront probablement une approche hybride des systèmes d'IA, combinant des éléments fondés sur des données et des règles et des mécanismes de blocage beaucoup plus simples basés sur des données de géolocalisation. Bien qu'il ne soit pas possible de confier la responsabilité morale ou juridique aux machines, il est plausible que les ASNC contribuent à rendre la prise de décision militaire sur le champ de bataille plus responsable d'un point de vue juridique et éthique.

Parallèlement à cette approche centrée sur la technologie, les tentatives nationales et internationales visant à réglementer les applications militaires de l'IA devraient être poursuivies. Toutefois, le développement des ASNC ne constitue pas nécessairement une réaction aux réglementations gouvernementales, mais pourrait également être volontairement avancé en tant

que norme industrielle de facto par les producteurs de technologie militaire. Plutôt que de s'abstenir de produire des PALWS et des unités logistiques d'IA pour l'armée, les producteurs européens de technologie militaire devraient s'efforcer d'être à la pointe de la recherche et du développement et d'établir une norme *fabriquée en Europe*, y compris des ASNC qui contribuent à garantir leur utilisation dans les limites des principes juridiques et éthiques. Dans le même temps, il convient d'avertir que ces systèmes ne doivent pas être utilisés de *manière* abusive pour justifier des exportations d'armes vers des régimes autoritaires et que l'établissement d'une norme de facto ne peut constituer qu'un élément d'un ensemble plus large de mesures visant à réglementer l'utilisation militaire de l'IA.

1. Introduction

L'utilisation militaire de systèmes d'IA dans des systèmes d'armes létales partiellement autonomes (PALWS) n'est pas un scénario futur spéculatif. Selon certaines informations, des drones d'IA qui sélectionnent et attaquent des cibles de manière autonome sont utilisés par l'Ukraine dans la guerre qui l'oppose à la Russie [1]. Dès 2021, des experts de l'ONU ont signalé le déploiement en Libye de drones entièrement autonomes fabriqués par la Turquie et dotés de capacités létales [2]. Et en 2022, les drones israéliens Ioitering ont contribué à la supériorité de l'Azerbaïdjan dans le conflit de longue date du Haut-Karabakh [3]. À la suite de cette impressionnante démonstration des capacités des drones de combat, la France et l'Allemagne ont accéléré leurs différentes voies d'acquisition et de développement de ce type d'armement [4]. Les systèmes d'interception de projectiles tels que les Patriot et Phalanx américains et le MANTIS allemand, qui doivent réagir plus rapidement et plus précisément que les opérateurs humains, sont depuis longtemps dotés d'un degré élevé d'autonomie. En outre, après avoir démontré la supériorité des systèmes d'IA dans les simulateurs de vol dans le cadre du programme AlphaDogfight, l'Agence américaine pour les projets de recherche avancée en matière de défense (DARPA) teste actuellement cette technologie sur des avions de chasse F-16 [5]. Mais le véritable avantage des systèmes d'IA dans le domaine militaire ne réside peut-être même pas dans ces scénarios qui s'apparentent encore à l'imaginaire quelque peu désuet des "robots tueurs". Conformément à la vision JADC2 (Joint All-Domain Command and Control) du Pentagone, les unités logistiques d'IA dans l'armée pourraient conduire à une vaste plateformisation interdomaine de la guerre qui a été comparée au fonctionnement de la plateforme de mobilité Uber [6]. (Non seulement pour des raisons militaires, l'intégration de systèmes d'IA dans l'armée est presque inévitable : les forces armées des pays hautement développés sont particulièrement touchées par la pénurie générale de main-d'œuvre qualifiée.

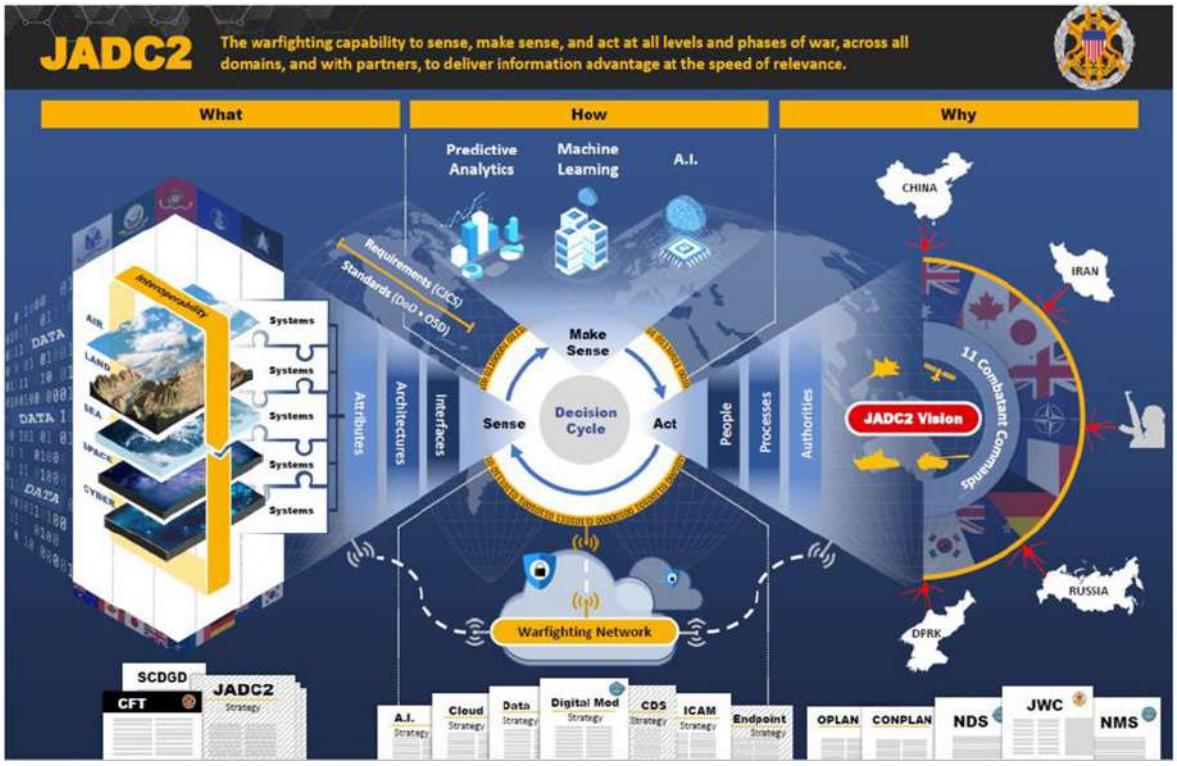

Fig. 1 : Concept JADC2 mettant l'accent sur l'interopérabilité interdomaine ; Source : "Summary of the Joint All-Domain Command & Control (JADC2) Strategy" (ministère américain de la défense, mars 2022), https://media.defense.gov/2022/Mar/17/2002958406/-1/-1/1/SUMMARY-OF-THE-JOINT-ALL-DOMAIN-COMMAND-AND-CONTROL-STRATEGY.PDF.

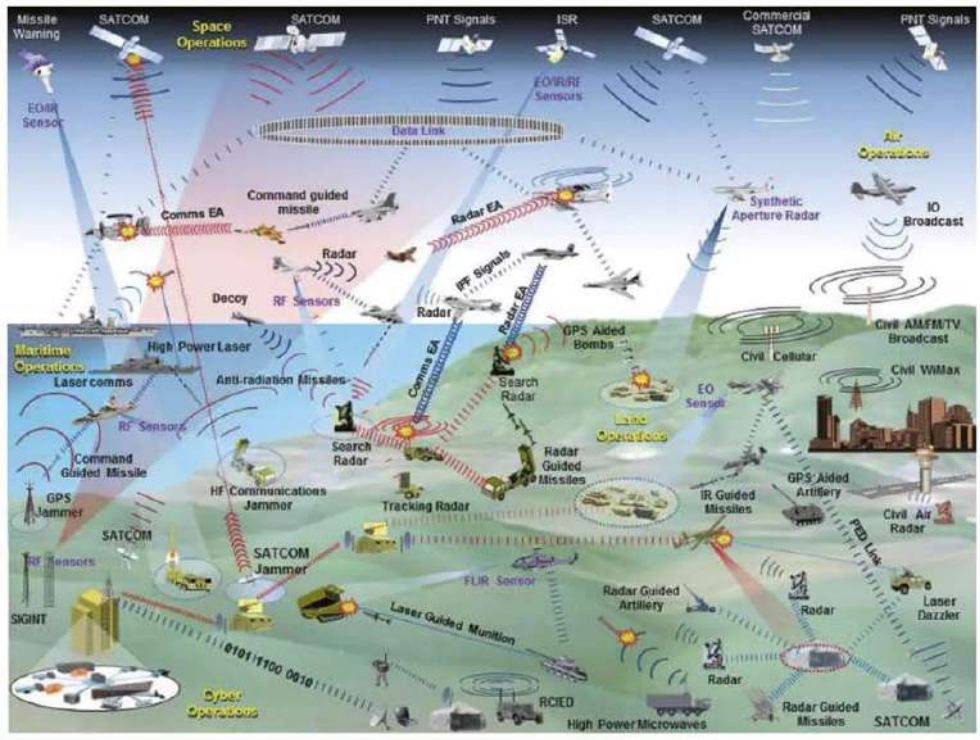

Fig. 2 : Communication numérique interdomaine, renseignement et logistique tels qu'envisagés par l'OTAN. Source : "Electronic Warfare - The Forgotten Discipline". Centre de compétence en matière de puissance aérienne interarmées de l'OTAN, 2 décembre 2018. https://www.japcc.org/articles/electronic-warfare-the-forgotten-discipline/.

L'utilisation des PALWS (systèmes d'armes létales partiellement autonomes) et des LAWS (systèmes d'armes létales autonomes) soulève en particulier de profonds problèmes moraux, éthiques et juridiques. Ceux-ci ont été résumés dans le problème du "déficit de responsabilité" [7] qui pourrait résulter d'une prise de décision automatisée dans la guerre et dans la demande complémentaire de garder les PALWS sous un "contrôle humain significatif" [8]. Les réglementations respectives concernant l'utilisation militaire de l'IA sont en train d'émerger mais n'ont pas encore atteint un stade solide. Plusieurs gouvernements nationaux élaborent progressivement des cadres nationaux concernant l'intégration de l'IA dans l'armée. Par exemple, la position officielle de la France est basée sur l'avis du Comité d'éthique de la défense sur PALWS à partir de 2021 [9]. Le Royaume-Uni a publié sa stratégie sur les capacités de défense fondées sur l'IA en juin 2022 [10]. En janvier 2023, le ministère américain de la défense a mis à jour sa directive sur cette question hautement dynamique [11]. Ces approches nationales s'accordent pour rejeter explicitement le développement de LAWS entièrement autonomes. En ce qui concerne les PAWLS partiellement autonomes, elles soulignent l'importance d'un contrôle humain significatif. Toutefois, ce que cela implique concrètement est loin d'être clair. En temps de guerre, la vitesse de réaction dépasse souvent les capacités humaines. Dans les systèmes d'interception de projectiles, un degré élevé d'autonomie est inévitable ; des programmes tels qu'AlphaDogfight suggèrent des développements similaires pour l'ensemble des combats aériens, dans lesquels le niveau de contrôle humain effectif est, de facto, en constante diminution depuis des décennies.

Au niveau international, les sessions de 2022 et 2023 du Groupe d'experts gouvernementaux des Nations Unies (GGE) sur les technologies émergentes dans le domaine des LAWS sont parvenues à la conclusion que l'utilisation des LAWS "entraîne la responsabilité internationale" des États qui les déploient et que "les êtres humains responsables de la planification et de la conduite des attaques doivent se conformer au droit international humanitaire" [12]. Au-delà de l'accent mis sur le contrôle humain et la responsabilité, le Groupe d'experts gouvernementaux a également formulé un certain nombre de recommandations techniques : par exemple, "les armes légères et de petit calibre ne doivent pas être utilisées si elles sont incapables de l'être conformément au droit international humanitaire (...), y compris les principes et les exigences de distinction, de proportionnalité et de précaution dans l'attaque" [13]. Cela pose le problème des armes à feu technologiquement primitives qui peuvent être considérées comme des armes sans discrimination et, par conséquent, comme des armes illicites. Pour contrer ces attaques

illégales sans discrimination provenant de systèmes d'IA technologiquement primitifs, le Groupe d'experts gouvernementaux a recommandé ce qui suit :

A. Limiter les types de cibles que le système peut engager.

B. Limiter la durée, la portée géographique et l'ampleur de l'opération de le système d'armes.

Des recommandations techniques similaires, mentionnant même explicitement les systèmes d'IA pour garantir le respect du DIH, sont formulées dans la *Déclaration politique* multilatérale de 2023 *sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie*, qui est, en novembre 2023, soutenue par tous les États du G7, pratiquement tous les États de l'UE et plusieurs États individuels tels que la Libye, la Corée du Sud, la Turquie et Singapour :

> Les États devraient également réfléchir à la manière d'utiliser les capacités militaires d'IA pour renforcer la mise en œuvre du droit international humanitaire et améliorer la protection des civils et des biens de caractère civil dans les conflits armés [14].

C'est le point de départ de mes réflexions. Au lieu de me concentrer sur le contrôle humain et les actions gouvernementales, je promeus une approche centrée sur la technologie pour mettre en œuvre des systèmes de régulation dans les applications militaires de l'IA. Ces systèmes, que j'appelle systèmes autonomes de contrôle normatif (ASNC), pourraient être exigés par les organismes publics en tant que normes industrielles, mais pourraient également être développés et mis en œuvre volontairement en tant que normes de facto par les producteurs de technologie militaire. Cette voie plutôt informelle et orientée vers le secteur privé pourrait être plus rapide que les tentatives lourdes et, jusqu'à présent, inefficaces de parvenir à un consensus entre les gouvernements dans ce domaine.

Les ASNC sont comparables aux systèmes ISA des voitures, qui peuvent empêcher activement les conducteurs de dépasser les limites de vitesse grâce à la reconnaissance des panneaux de signalisation et à la géolocalisation. En fonction de différents degrés d'autonomie, les ASNC pourraient simplement conseiller les combattants en termes juridiques et éthiques, bloquer de manière autonome les ordres illégaux donnés par des humains et/ou proposer des actions militaires susceptibles d'être proportionnées et de minimiser les dommages collatéraux. En outre, les ASNC devraient inclure un mécanisme de blocage pour contribuer à garantir que les

PALWS et les unités logistiques de l'IA ne soient pas utilisés dans des guerres d'agression ou contre des manifestants nationaux. Il est essentiel de reconnaître qu'aucune de ces approches technologiques ne peut être considérée comme une solution complète et de soulager le personnel militaire, les politiciens et la société civile de leur devoir de s'efforcer de trouver des solutions plus complètes pour réglementer les PAWLS. En outre, il convient d'ores et déjà d'avertir que ces dispositifs ne doivent pas être utilisés de manière abusive pour procéder à un *"lavage ASNC"* afin de justifier les exportations d'armes vers des régimes autoritaires.

2. Approches antérieures des systèmes autonomes de contrôle normatif

Le discours sur les systèmes autonomes de contrôle normatif (ASNC) peut être considéré comme un sous-domaine de l'éthique de l'IA, de l'éthique des machines ou de la construction d'agents moraux artificiels [15]. Il existe toutefois des différences importantes : premièrement, le discours sur les ASNC se préoccupe davantage de la mise en œuvre de normes juridiques que de normes éthiques ou morales, raison pour laquelle je parle généralement de contrôle normatif ; deuxièmement, les ASNC minimalistes concernant les mécanismes de blocage basés sur les données de géolocalisation ne sont pas nécessairement liés à l'IA ou aux capacités de raisonnement artificiel, mais sont autonomes uniquement dans le sens où ils ne requièrent pas d'intervention humaine. Les ASNC pour les PALWS, les systèmes d'armes conventionnels et les unités logistiques d'IA font l'objet de discussions depuis un certain temps. Parmi les exemples plus ou moins connus, citons l'approche maximaliste de l'éthicien des robots Arkin (2006) [16], l'approche minimaliste "MinAI" des spécialistes des systèmes d'armes autonomes Scholz et Galliot (2018) [17], les robots refusant des ordres illégaux par les spécialistes du droit international Grimal et Pollard (2021) [18], les "minotaures" axés sur la logistique par les éthiciens de la technologie Sparrow et Henschke (2023) [19], et le modèle d'un dispositif militaire autonome conforme au DIH par Zurek, Kwik et van Engers (2023) [20].

En général, ces approches spéculatives et théoriques, qui ne produisent généralement que des organigrammes de l'IA, sont très optimistes quant aux capacités de l'IA à réguler les PAWLS ainsi que le comportement humain sur le champ de bataille. Cette attitude est en contradiction flagrante avec les approches axées sur l'état actuel de la technologie, qui soulignent que l'IA

d'aujourd'hui est généralement incapable de fonctionner de manière fiable dans les conditions de complexité du champ de bataille, caractérisées par les "frictions" clausewitziennes et le proverbial "brouillard de la guerre". Par exemple, ces lacunes attendues de l'IA sur le champ de bataille ont été abordées par Wallace en 2022 [21] et par Yan en 2020 [22]. Dans le discours théorique souvent spéculatif sur l'IA, il est important de contextualiser les recherches positives sur les capacités de l'IA et d'examiner de manière critique dans quelle mesure les arguments avancés sont l'expression d'un "solutionnisme technologique" non fondé et optimiste en matière de progrès, qui suppose à tort que des problèmes sociaux, politiques ou éthiques complexes peuvent être facilement résolus par la technologie numérique [23].

Cela étant dit, il est essentiel d'évaluer différemment la capacité des systèmes d'IA dans les différents domaines : comme le démontrent AlphaDogfight, le succès des drones de munition errants et d'autres exemples, le domaine aérien, comparativement "sans friction", se prête plutôt bien aux systèmes d'IA ; de même, le cyberespace, la mer et l'espace extra-atmosphérique se prêtent relativement bien aux systèmes d'IA. En revanche, on ne peut raisonnablement s'attendre à moyen terme à ce que l'IA maîtrise le combat terrestre, qui reste le domaine décisif - en particulier s'il s'agit de terrains urbains où la distinction entre cibles civiles et militaires n'est pas claire et en particulier en ce qui concerne les PALWS sur le terrain ou les systèmes d'IA incorporés (EAI) au sens de "robots tueurs" plus ou moins humanoïdes. Certains auteurs considèrent que les véritables capacités de l'IA militaire sont totalement dissociées des PALWS et de l'EAI et affirment que l'IA est mieux employée dans les tâches cognitives, par exemple la logistique [24]. Comme nous l'avons mentionné dans l'introduction, cette approche est également poursuivie dans la vision JADC2 du ministère américain de la défense. (Voir les figures 1 et 2.)

a) L'approche maximaliste : Gouverner les comportements létaux (2006)

L'étude *Governing Lethal Behavior : Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, financée par le ministère américain de la défense, est à la fois l'expression très problématique d'une attitude techno-solutionniste et l'origine des ASNC. Le concept d'Arkin s'inspire du *régulateur mécanique de* Watts pour la machine à vapeur, qui était destiné à garantir la sécurité et les performances de la machine, un concept qui joue généralement un rôle

important dans les premiers débats sur la cybernétique [25]. Comme le montre l'organigramme ci-dessous (fig. 3), Arkin décrit un système complexe mais relativement transparent qu'il appelle "Ethical Autonomous Robot Architecture" (architecture de robot autonome éthique) et qui comporte plusieurs boucles : La planification de la mission et la délibération, qui impliquent une interaction homme-robot et qui conduisent à la formulation de principes éthiques définis dans l'adaptateur éthique et les contraintes (C), qui, à leur tour, contiennent des obligations et des interdictions. Selon Arkin, le C s'appuie sur les règles d'engagement et le droit international de la guerre pour garantir un comportement éthique. Le droit et l'éthique sont à peine distingués. Un système de "contrôle du comportement éthique" alimente les perceptions du comportement traitées par ordinateur $s_n \rightarrow \beta_n \rightarrow r_n$ à C qui, à son tour, transmet des données au gouverneur éthique qui, à son tour, définit les actions permises (en termes de permissions, d'obligations et d'interdictions) ($_{Permissible}$) et bloque toutes les autres actions.

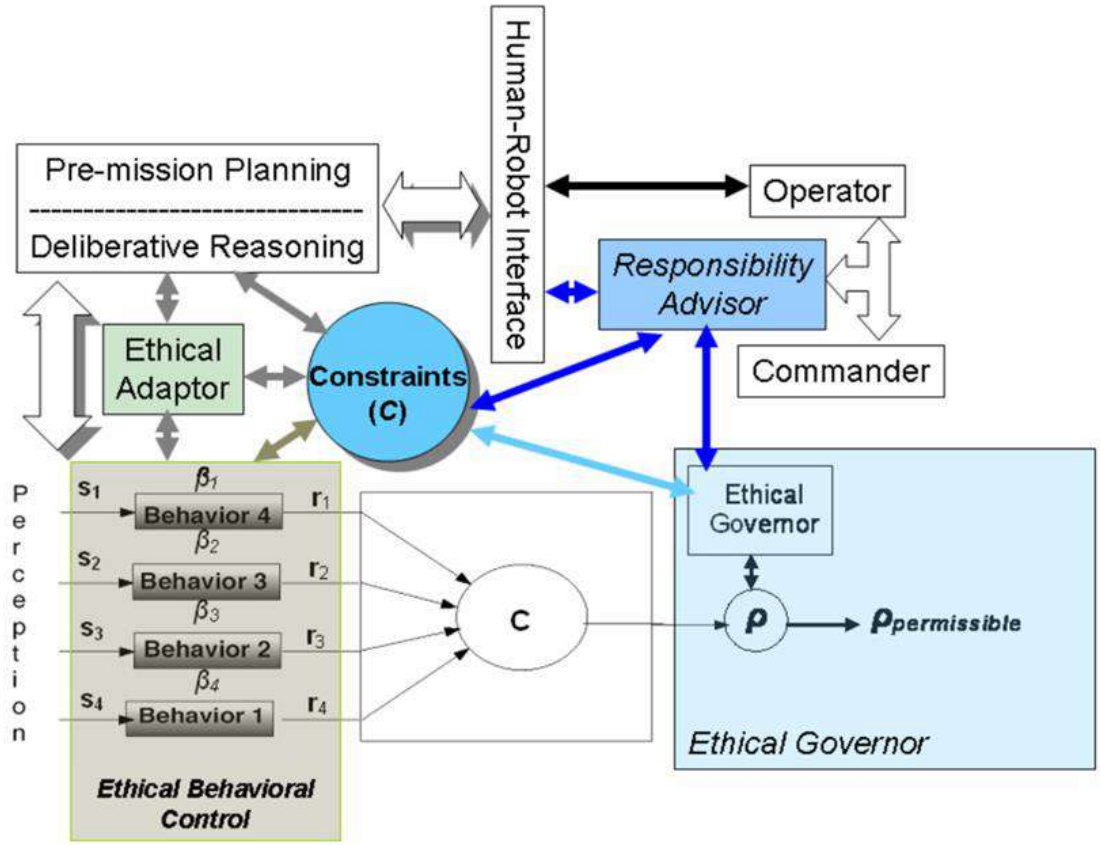

Fig.3 : Principaux éléments d'une architecture de robot autonome éthique, tels que décrits dans Ronald C. Arkin, "Governing Lethal Behavior : Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture' (Mobile Robot Laboratory College of Computing Georgia Institute of Technology, 2006).

En ce qui concerne les sources normatives des contraintes et des obligations dans le module C, Arkin envisage qu'elles soient informées par les règles d'engagement (ROE) spécifiques à la mission dans la "mémoire à court terme" de la machine, y compris les obligations et les interdictions dans la "mémoire à long terme" de la machine, qui devraient être informées par les lois de la guerre (LOW) et les règles permanentes d'engagement (SROE). Ces considérations normatives seraient appliquées en tant que contraintes et obligations à un module de "raisonnement probatoire" alimenté par une grille d'information globale produisant des données de connaissance de la situation et des perceptions transformées en représentations calculables de la situation environnante. Voir la représentation du système dans l'organigramme ci-dessous (fig. 4 page suivante).

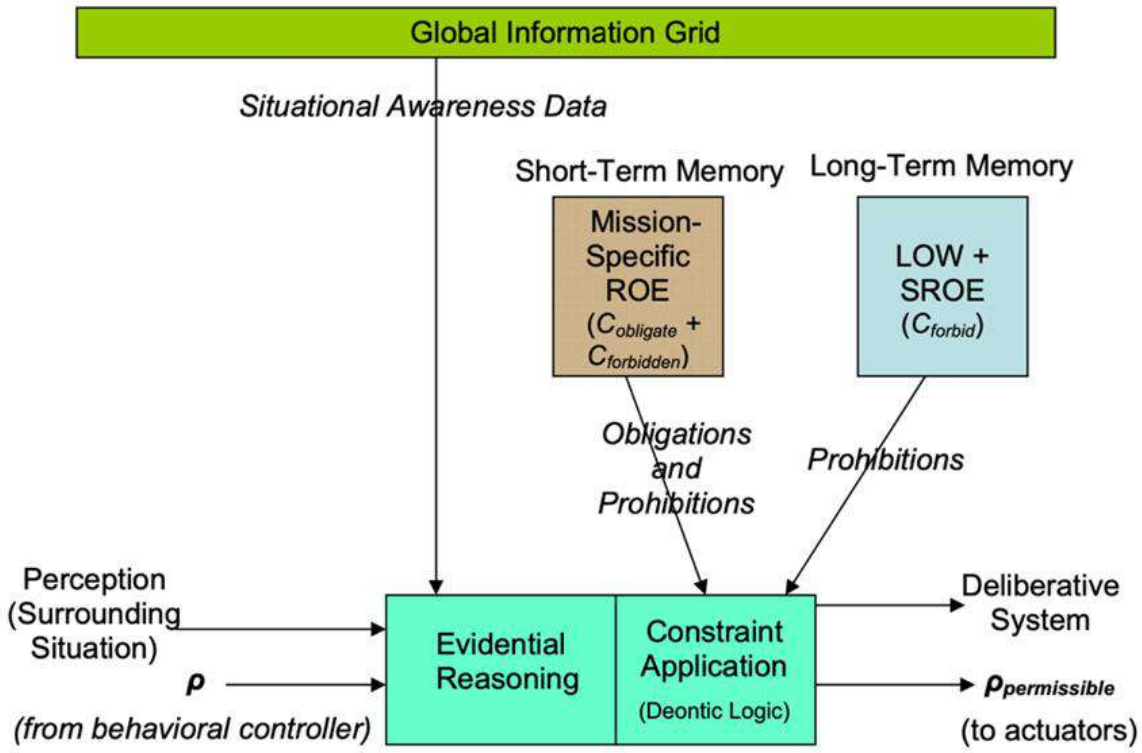

Fig.4 : Composants architecturaux du gouverneur éthique tels que décrits dans Arkin, "Governing Lethal Behavior : Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture".

b) L'approche minimaliste : MinAI (2018)

En 2018, Scholz et Galliot ont publié un texte sur leur concept ASNC MinAI, qui se distingue explicitement d'Arkin [26]. Scholz et Galliot caractérisent l'approche d'Arkin comme une " approche maximale " puisqu'elle repose sur la capacité des machines en C (Fig. 3) à relier de

manière indépendante des principes éthiques et juridiques à des situations et des actions concrètes, c'est-à-dire à s'engager à de nombreux niveaux dans la *reconnaissance d'objets* et le *raisonnement de la machine*, y compris la définition d'actions autorisées *et* interdites. En revanche, les auteurs adoptent une "approche minimale" : ils affirment que la traduction des principes humanitaires généraux de la guerre en actions concrètes recommandées dans une situation de combat concrète ne devrait pas être laissée aux machines ; au contraire, les acteurs humains devraient mettre en œuvre des mécanismes de blocage beaucoup plus simples dans PALWS, tels que des mécanismes de blocage en réponse à la détection de cibles civiles et à la reconnaissance de signes de protection tels que la Croix-Rouge et le Croissant-Rouge, le drapeau blanc et les mains levées en signe de reddition, ainsi que la reconnaissance de sites religieux et culturels. En outre, les armes pourraient être bloquées dans certains cas d'utilisation, y compris les cas s'appliquant aux critères de proportionnalité. En outre, les auteurs soulignent que ces ASNC et systèmes d'IA ne peuvent généralement pas être moralement et juridiquement responsables et plaident en faveur d'une documentation solide pour évaluer en détail la responsabilité des commandants. La figure ci-dessous montre comment le système d'"arme éthique" bloque le lancement d'un missile en direction d'un navire arborant la Croix-Rouge.

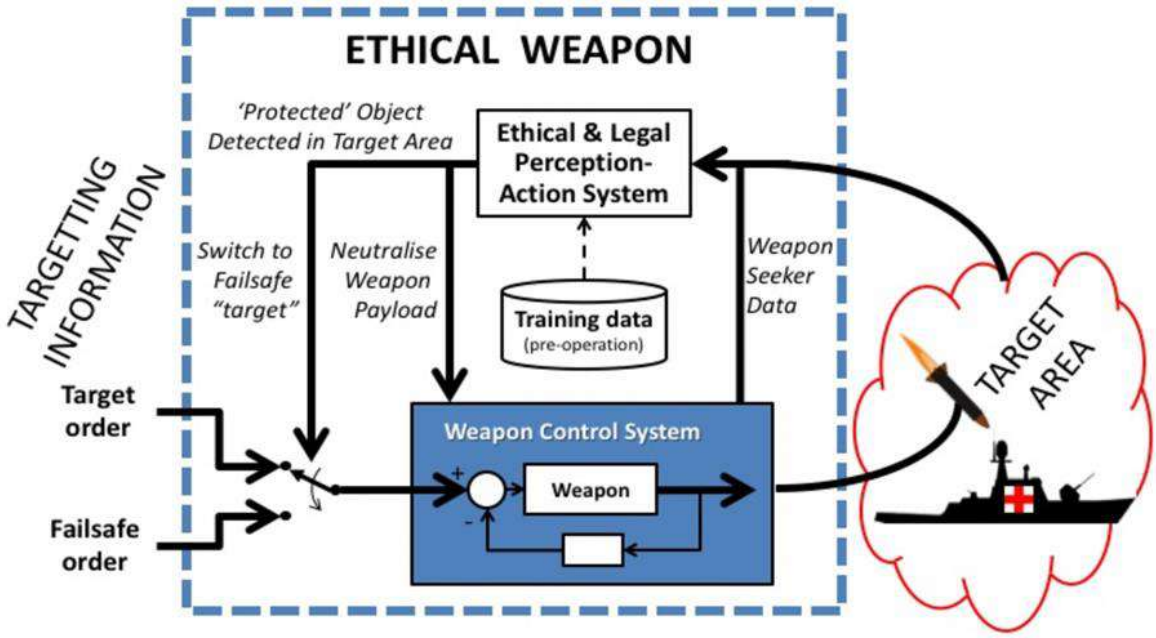

Fig. 5 : Organigramme de l'"arme éthique" minimaliste de Scholz et Galliot qui comprend simplement un mécanisme de blocage. D'après : Scholz, Jason, et Jai Galliott. The Humanitarian Imperative for Minimally-Just AI in Weapons". Journal of Indo-Pacific Affairs, hiver 2018, 57-67. https://doi.org/10.1093/oso/9780197546048.003.0005.

c) Les robots qui refusent les ordres illégaux (2021)

L'ouvrage de Grimal et Pollard intitulé *The Duty to Take Precautions in Hostilities, and the Disobeying of Orders (*2021) : *Should Robots Refuse ?* marque une rupture significative avec le solutionnisme technologique d'Arkin [27]. Les auteurs (tous deux juristes internationaux) ne se concentrent plus exclusivement sur les actions autonomes des machines, mais abordent la question de l'association homme-machine. En outre, leur article prend en compte les particularités du droit international humanitaire, en particulier la relation complexe entre la responsabilité et les chaînes de commandement. Les auteurs soulignent qu'en vertu de l'article 57 du protocole additionnel I des Conventions de Genève et de la règle 155 du droit international coutumier, les supérieurs comme les subordonnés ont le devoir de prendre des précautions, par exemple en ce qui concerne la distinction entre civils et combattants. Compte tenu de ce devoir, Grimal et Pollard affirment qu'il pourrait même être illégal pour le personnel militaire de ne pas consulter les systèmes d'IA, en particulier si ceux-ci sont liés à la reconnaissance aérienne et qu'il est donc probable que ces systèmes facilitent la distinction entre les civils et les combattants.

Cependant, Grimal et Pollard sont également pertinents pour le concept d'ASNC dans la mesure où ils déplacent l'accent du concept d'Arkin de contrôle éthique et juridique autonome sur les LAWS vers un scénario hybride dans lequel les "robots désobéissants" ne se contentent pas de s'autoréguler, mais contrôlent ou "conseillent" principalement le comportement humain, y compris le comportement en relation avec les armes conventionnelles (c'est-à-dire non autonomes). Grimal et Pollard décrivent un scénario dans lequel "un humain décide toujours du moyen d'attaque le plus approprié, bien qu'en réalité ses "choix" soient probablement assez restreints".

Les auteurs soulignent que les combattants humains pourraient bénéficier d'asymétries dans les équipes homme-machine. En effet, les systèmes d'IA ne sont pas soumis aux mêmes contraintes psychologiques et sociales que les humains activement engagés dans le combat. Par exemple, Grimal et Pollard s'intéressent aux systèmes de sauvegarde des armes nucléaires basés sur l'IA qui évaluent, d'un point de vue éthique et juridique, la légitimité du lancement de missiles. Cela conduit au concept de refus robotique de Grimal et Pollard : les systèmes d'intelligence artificielle incarnée (EAI) examinent les ordres humains et pratiquent différents niveaux de "refus" des ordres qui, dans le cas le moins grave, peuvent consister en une simple allusion à

une éventuelle illégalité et, dans le cas le plus grave, peuvent impliquer un blocage immédiat de tous les ordres ultérieurs similaires dans leur structure ou émis par le même acteur. Les auteurs décrivent cette situation dans l'organigramme suivant, qui va des évaluations des systèmes EAI 1 à 3 au système 4A "refus systématique" ou au système 4B "suivre l'ordre" ou "réexécuter l'évaluation".

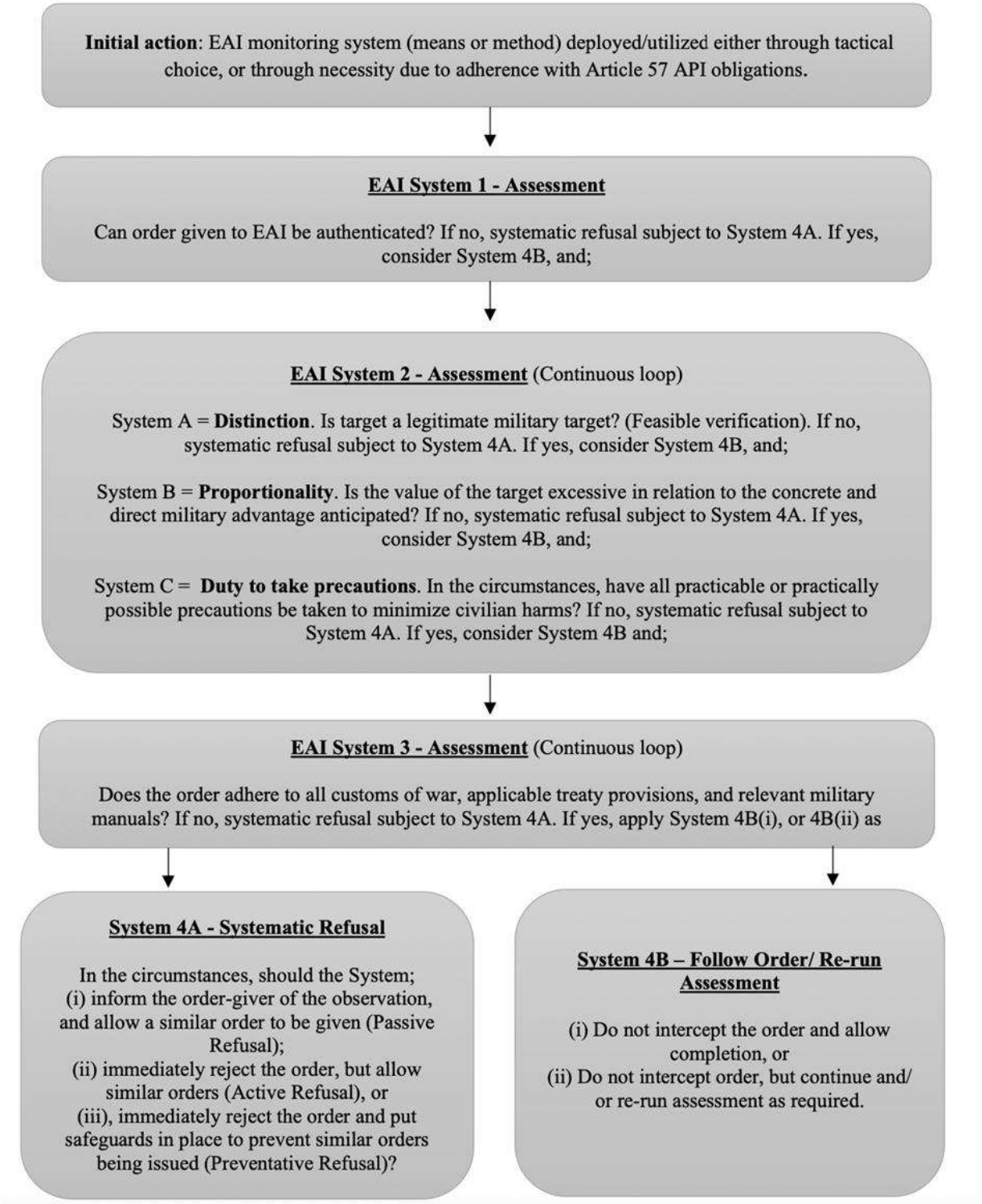

Fig. 6 : Organigramme pour le refus des robots tiré de Grimal, Francis, et Michael Pollard. The Duty to Take Precautions in Hostilities, and the Disobeying of Orders : Should Robots Refuse ? *Fordham International Law Journal*, 2021, 671-734.

d) Les unités logistiques de l'IA sous forme de Minotaures (2023)

Une quatrième approche sera examinée ici : *Minotaurs, Not Centaurs : The Future of Manned-Unmanned Teaming* par les éthiciens Sparrow et Henschke dans 2023 [28]. Comme le suggère le titre, qui fait référence à la créature hybride du Minotaure, cette approche porte exclusivement sur l'association homme-machine. Dans la discussion de Sparrow et Henschke, l'accent est explicitement déplacé des formes physiques de l'EAI, par exemple les drones contrôlés par les humains (appelés "centaures" par l'expert en systèmes d'armes autonomes Scharre [29]), vers les systèmes d'IA qui soutiennent les acteurs humains sur le terrain dans une capacité cognitive. Les auteurs qualifient ces derniers systèmes de "minotaures" parce qu'ils impliquent des *entités non humaines en plus des processus, par exemple en tant que coordinateurs logistiques*. Dans une évaluation réaliste de l'état de l'art, Sparrow et Henschke écrivent :

> On peut dire que les intelligences artificielles sont déjà plus capables d'effectuer les tâches cognitives les plus importantes pour la guerre que les robots ne sont capables d'effectuer les fonctions du corps humain les plus importantes pour la guerre.

À l'instar de Grimal et Pollard concernant le devoir de précaution, Sparrow et Henschke soutiennent que le développement rapide des systèmes d'IA crée un "impératif éthique" pour l'utilisation de ces technologies dans la guerre, car elles "aideront à prévenir les incidents de tirs amis et à améliorer la capacité de survie des combattants humains". S'inspirant implicitement de la vision JADC2 du Pentagone, Sparrow et Henschke citent la coordination logistique effectuée par les systèmes d'IA dans le secteur privé, tels que les centres d'approvisionnement d'Amazon ou la coordination des chauffeurs d'Uber, comme modèle pour de tels scénarios de Minotaure. Un cas extrême de cette dynamique a été observé dans la guerre participative en Ukraine, où les volontaires ont été organisés logistiquement par les algorithmes des médias sociaux [30]. Ces phénomènes de plateformisation de la guerre sont très pertinents pour les scénarios Minotaure. En raison de l'accent mis sur le guidage cognitif fourni par les systèmes d'IA dans les équipes homme-machine, l'article de Sparrow et Henschke offre un point de départ intéressant pour étudier les applications possibles des ASNC dans les unités logistiques d'IA au sein de l'armée.

e) Le dispositif autonome conforme au droit international humanitaire (2023)

L'approche la plus récente et certainement l'une des plus audacieuses pour établir des ASNC est le *modèle de* Zurek, Kwik et van Engers *d'un dispositif militaire autonome suivant le droit international humanitaire* (2023), qui s'appuie sur Arkin mais ne s'engage pas dans les autres approches mentionnées ci-dessus [31]. Au niveau le plus élémentaire, le modèle de Zurek, Kwik et van Engers consiste à prédire la relation entre l'*avantage militaire* (l'importance de l'avantage tiré de l'attaque d'une cible spécifique) et les *dommages accessoires* (les dommages collatéraux causés par les effets directs et indirects prévisibles d'une attaque). Dans la mesure du possible, ces critères devraient être quantifiés et formalisés. Les auteurs écrivent :

> La création d'un modèle autonome piloté par l'IA nécessite non seulement un modèle de calcul, donc une représentation quantifiable [de l'avantage militaire et du dommage indirect], mais aussi une représentation qui permette leur comparaison formelle.

D'un point de vue technique, Zurek, Kwik et van Engers soulignent qu'il y a eu un changement significatif dans la recherche et le développement de l'IA depuis l'approche d'Arkin, à savoir le passage à des approches fondées sur des données plutôt que sur des règles. S'écartant de la tendance actuelle, ils préconisent une approche hybride fondée sur les données en ce qui concerne les tâches de perception qu'ils résument comme la "partie cognitive" et fondée sur les règles en ce qui concerne la "partie raisonnement", qui s'appuie sur le droit international humanitaire et d'autres normes. Zurek, Kwik et van Engers sont également plus concrets que tous les auteurs précédents puisqu'ils se concentrent explicitement sur une partie de la boucle OODA (Observer, Orienter, Décider, Agir), dans laquelle l'ASNC devrait être mise en œuvre : le cycle de ciblage. L'organigramme suivant considère que les circonstances générales, le renseignement d'origine électromagnétique et les objectifs sont pris en compte dans la partie cognitive et que les traités internationaux sont pris en compte dans la partie raisonnement. Le résultat généré consiste à ordonner les décisions en fonction de leur degré de conformité avec les réglementations internationales.

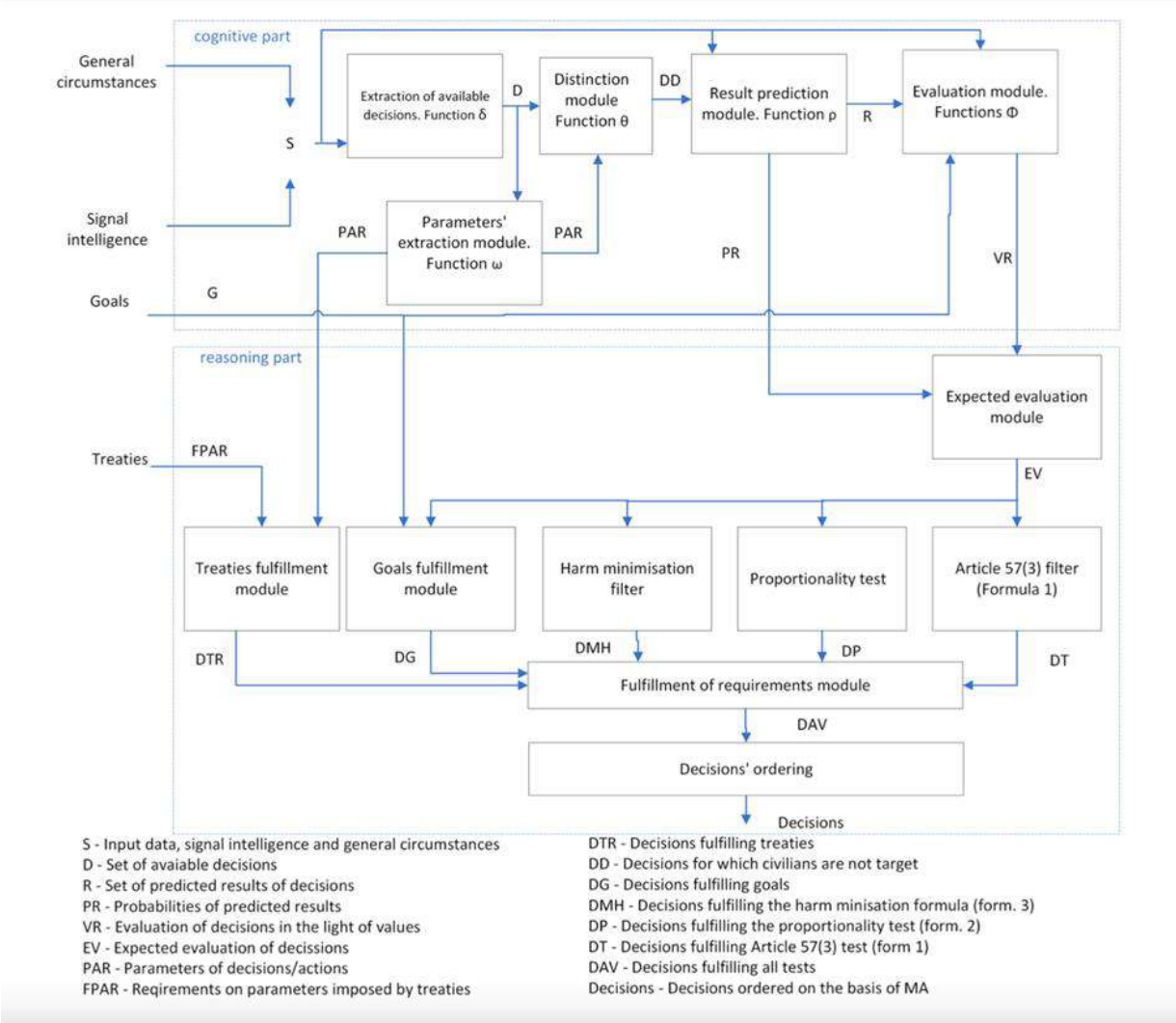

Fig. 7 : Organigramme pour le *modèle d'un dispositif militaire autonome conforme au droit humanitaire international*, tiré de Zurek, Tomasz, Jonathan Kwik et Tom Van Engers. Model of a Military Autonomous Device Following International Humanitarian Law" (Modèle d'un dispositif militaire autonome conforme au droit humanitaire international). *Ethics and Information Technology* 25, no. 1 (mars 2023).

Comme le montre l'organigramme ci-dessus, Zurek, Kwik et van Engers incluent plusieurs modules avec des tâches spécifiques dans la partie raisonnement, notamment le "filtre de minimisation du préjudice", le "test de proportionnalité" et le "filtre de l'article 57". Ces modules peuvent également être activés séparément, lorsqu'une question nécessite une forme spécifique d'examen. Plutôt que de promettre une solution complète, ces chercheurs soulignent que la responsabilité morale et juridique ne peut être déléguée aux machines, mais que les systèmes d'IA peuvent jouer un rôle crucial en prenant toutes les précautions possibles.

3. Discussion et élaboration des ASNC

Il est essentiel d'être conscient de la différence entre les tâches que les ASNC peuvent plausiblement maîtriser dans un avenir à moyen terme et les tâches qu'il est très peu probable que ces systèmes soient capables de maîtriser dans un avenir à moyen terme. L'évolution historique des ASNC, de l'approche techno-solutionniste d'Arkin à la solution graduelle de Zurek, Kwik et van Engers, en passant par l'approche logistique de Sparrow et Henschke, illustre l'évolution générale de la recherche en IA, qui s'éloigne des systèmes d'EAI promettant des solutions complètes pour se tourner vers le scepticisme et mettre l'accent sur des approches flexibles basées sur l'équipe homme-machine, combinant le meilleur des deux mondes. Même pendant l'actuel "été de l'IA" [32], qui est en fait une vague de chaleur, personne ne peut sérieusement supposer qu'il est possible d'externaliser la responsabilité morale ou juridique aux machines de quelque manière que ce soit.

Les ASNC ne peuvent pas dispenser le personnel militaire de son devoir d'être informé des normes pertinentes et de veiller à leur respect. Les ASNC ne peuvent pas non plus décharger les responsables politiques et la société civile de leur devoir de formuler et d'appliquer des normes spécifiques susceptibles de régir l'utilisation de PALWS et d'autres applications militaires de l'IA. Enfin et surtout, en raison des frictions militaires sur le champ de bataille et du "brouillard de guerre", on ne peut attendre des systèmes d'IA qu'ils fonctionnent parfaitement dans tous les domaines. Les ASNC autonomes intégrés dans les PALWS pourraient fonctionner plus ou moins bien dans les domaines de l'air, de la mer et de l'espace extra-atmosphérique, mais certainement pas en ce qui concerne le combat terrestre, dans lequel les ASNC pourraient plutôt opérer à partir de centres de données, fournissant un contrôle logistique et normatif aux PALWS et aux troupes humaines sur le terrain. Ces deux types d'ASNC ne sont que les deux extrémités d'un spectre qui comprend de nombreuses formes intermédiaires. (Comme le montre l'exemple des drones autonomes qui auraient été utilisés en Ukraine, le type d'arme entièrement autonome doté de fonctions intégrées présente le grand avantage d'être immunisé contre le brouillage.

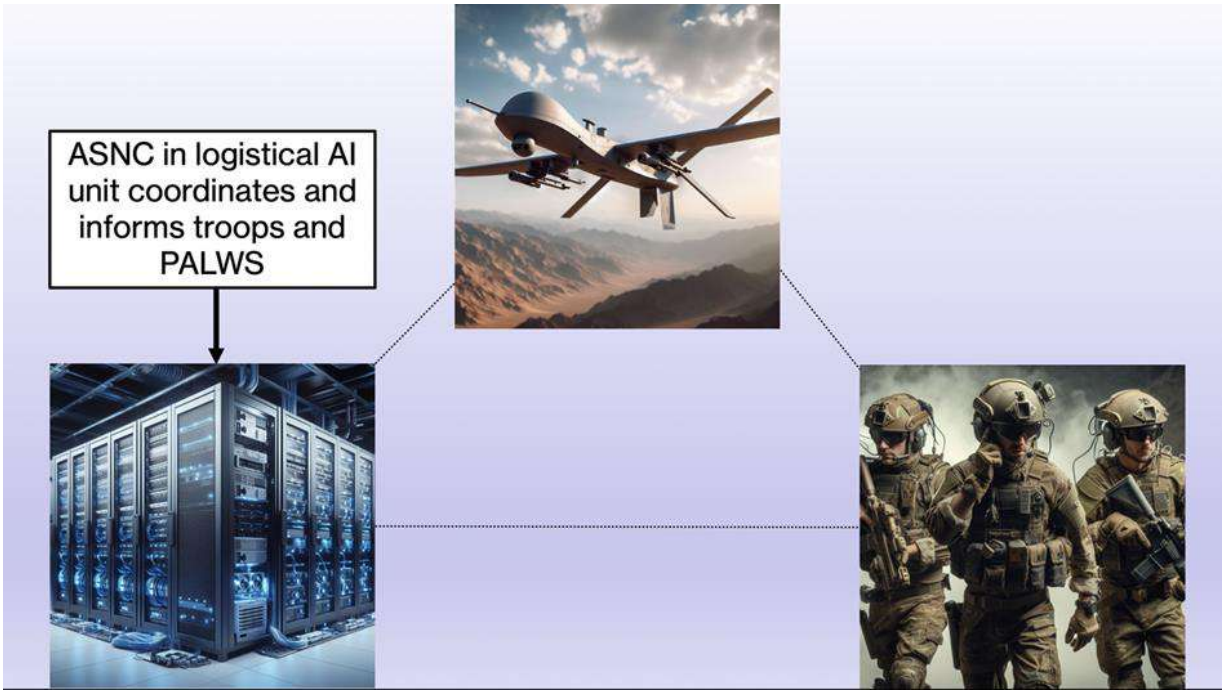

Fig. 8 : ASNC dans l'unité d'IA logistique selon la vision JADC2 du Pentagone.

Fig 9 : ASNC autonome intégré à PALWS, une approche maximaliste.

Quel que soit le domaine, il est certain que les ASNC pourraient apporter des avantages significatifs s'ils étaient considérés comme un outil permettant d'aider les opérateurs humains et les commandants à prévoir la relation entre l'avantage militaire et les dommages accidentels, ce qui est essentiel pour respecter le principe de proportionnalité du droit international humanitaire. Toutefois, il faut pour cela résoudre d'importants problèmes de représentation et de langage. Comme l'a également souligné l'US Air Force, le critère de proportionnalité est particulièrement subjectif [33]. Il est difficile de parvenir à une représentation quantifiable et formelle à cet égard. Toutefois, il convient de souligner que l'évaluation de la proportionnalité est souvent étroitement liée à l'évaluation quantitative des décès. Il existe donc indubitablement une possibilité de représenter l'aspect le plus crucial du DIH de manière calculable. Au-delà même de la construction d'ASNC, de tels processus de quantification et de formalisation pourraient contribuer à accroître l'objectivité, la transparence et, par conséquent, la possibilité de débattre des critères du DIH.

D'un point de vue technique, l'approche d'Arkin basée sur des règles datant de 2006 pourrait représenter l'avenir plutôt que le passé des ASNC. Au cours de la dernière décennie, les approches de l'IA fondées sur des règles ont été abandonnées au profit d'approches fondées sur des données. Les approches fondées sur les données sont beaucoup plus souples que les approches fondées sur les règles, car elles n'obligent pas les ingénieurs en logiciel à formuler des règles abstraites pour tous les scénarios possibles. Cependant, précisément parce que les systèmes guidés par les données ne sont pas caractérisés par des règles formulées par des humains mais par des inférences statistiques automatisées, ces systèmes ont souvent été critiqués pour constituer des boîtes noires et manquer d'explicabilité [34]. En particulier dans un domaine éthiquement très sensible comme la guerre, les ASNC ne peuvent pas s'appuyer uniquement sur des processus d'apprentissage automatique pilotés par les données, ce qui signifie qu'ils ne seront explicables qu'*a posteriori*. S'il est inévitable d'utiliser des approches fondées sur les données en ce qui concerne la connaissance de la situation, les ASNC devraient inclure des processus fondés sur des règles explicitement informées par les principes juridiques et les logiques déontique et modale du droit international humanitaire.

Cependant, les approches minimales telles que la *MinAI* de Scholz et Galliott ont leurs avantages car elles laissent le raisonnement juridique aux humains et sont basées sur des mécanismes de blocage "plus simples". Scholz et Galliott mentionnent la détection de cibles civiles et la

reconnaissance de signes de protection tels que la Croix-Rouge et le Croissant-Rouge, le drapeau blanc et les mains levées en signe de reddition, ainsi que la reconnaissance de sites religieux et culturels.

Cette idée de mécanismes de blocage plus simples est également pertinente en ce qui concerne les principes juridiques et éthiques qui n'ont pas encore été suffisamment abordés dans la littérature : l'interdiction des guerres d'agression et de l'annexion de territoires en vertu du droit international public et l'exigence beaucoup plus douce, plutôt éthique-politique ou basée sur les droits de l'homme, de ne pas utiliser d'armes militaires pour réprimer les protestations nationales.

En résumé, les ASNC doivent remplir les tâches suivantes :

- Mécanismes de blocage concernant l'utilisation offensive de PALWS dans des zones très peuplées

- Mécanismes de blocage concernant la défense des territoires annexés

- Reconnaissance des sites culturels et religieux

- Mécanismes de blocage concernant le recours à la force pour réprimer les manifestations nationales

- Différenciation entre les infrastructures civiles et militaires

- Différenciation entre civils et combattants

- Évaluation de la proportionnalité

- Reconnaissance des symboles protégés (tels que la Croix-Rouge et le Croissant-Rouge)

Il est intéressant de regrouper ces tâches dans le graphique XY suivant (fig. 10 ci-dessous), X (en vert) allant d'une sensibilité éthique élevée à une sensibilité éthique faible et Y (en rouge) allant d'une faisabilité technologique faible et d'une collaboration homme-machine à une faisabilité technologique plus élevée et à une plus grande autonomie de la machine. En outre, la moitié supérieure (en bleu) est partiellement ou entièrement basée sur des données de géolocalisation stables, tandis que la moitié inférieure (en violet) est constituée de tâches dynamiques. Bien entendu, en ce qui concerne les paramètres de sensibilité éthique et de

faisabilité technologique, il convient d'ajouter qu'aucune de ces tâches n'est particulièrement facile et que toutes ces questions sont généralement d'une grande sensibilité éthique. Les différences représentées dans ce graphique XY peuvent sembler minimes et graduelles d'un point de vue extérieur. En ce qui concerne la distinction entre les tâches dynamiques et les tâches fondées sur une géolocalisation stable, il convient de préciser que, par exemple, les mécanismes de blocage liés à des territoires, des zones et des sites spécifiques sont stables parce que ces territoires, zones et sites ne bougent pas, contrairement, surtout, aux civils et aux combattants eux-mêmes. Les mécanismes de blocage concernant l'utilisation de la force contre les manifestants ont également un aspect basé sur des données de géolocalisation stables parce que les manifestations ont probablement lieu dans les centres-villes nationaux et qu'il n'est pas nécessaire d'y utiliser des armes militaires s'il n'y a pas d'invasion étrangère. De même, la différenciation entre les infrastructures civiles et militaires comporte des aspects stables, puisque les infrastructures ne se déplacent pas, mais aussi des aspects dynamiques, car, par exemple dans les guerres urbaines, les infrastructures civiles sont souvent réaffectées à des fins militaires.

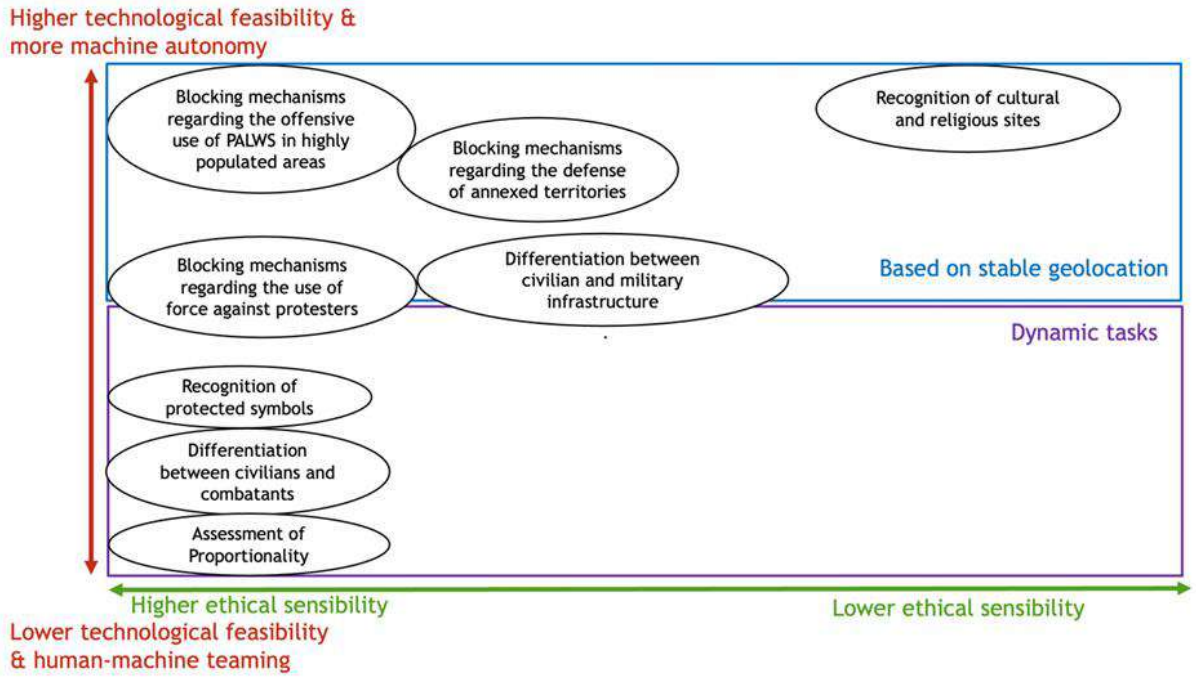

Fig. 10 : Graphique XY, avec X (en vert) allant d'une sensibilité éthique élevée à une sensibilité éthique faible et Y (en rouge) allant d'une faisabilité technologique faible et d'un travail d'équipe homme-machine à une faisabilité technologique plus élevée et à une plus grande autonomie de la machine. La moitié supérieure (en bleu) est partiellement ou entièrement basée sur une géolocalisation stable, tandis que la moitié inférieure (en violet) est constituée de tâches dynamiques.

Compte tenu de toutes ces qualifications, ce diagramme XY révèle quelque chose de surprenant : Jusqu'à présent, la plupart des discussions ont porté sur le coin inférieur gauche du graphique, qui présente une sensibilité éthique élevée et une faisabilité technologique faible (voir fig. 11, page suivante). Il serait de loin préférable de commencer par le coin supérieur droit du graphique, qui présente une faisabilité technologique plus élevée et une sensibilité éthique plus faible.

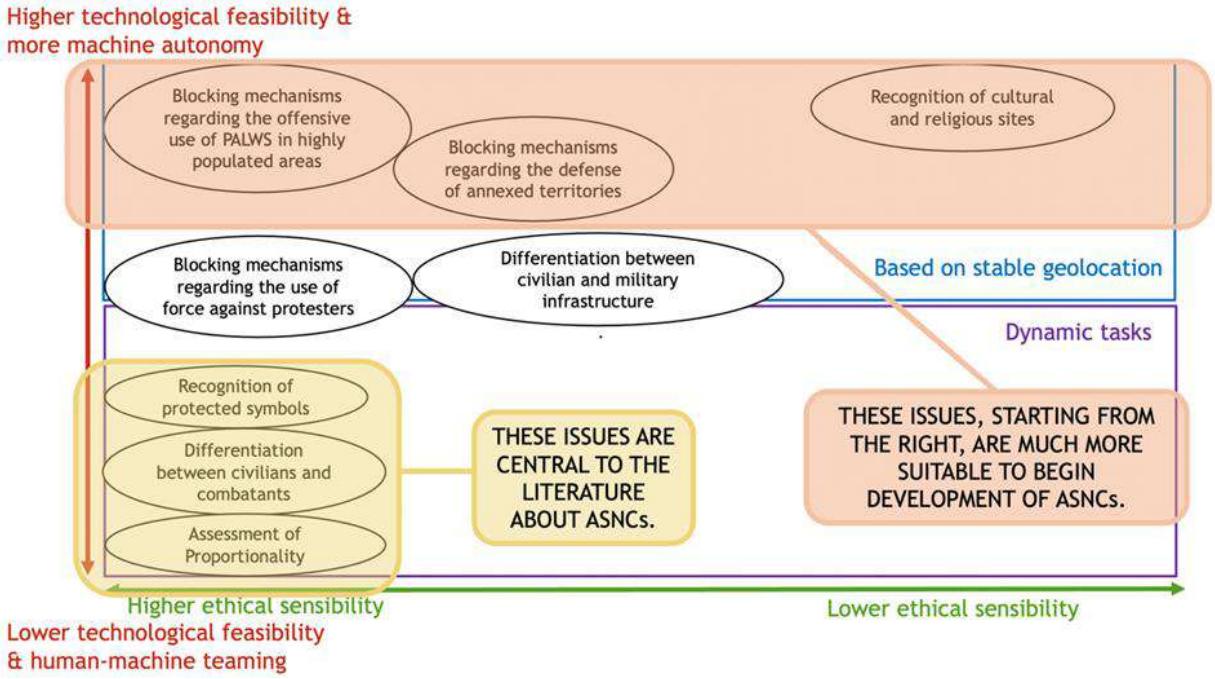

Fig. 11 : Graphique XY, le coin inférieur gauche (jaune) est généralement discuté, tandis que la moitié supérieure (orange) serait plus réalisable.

4. Les ASNC, une norme industrielle *de facto*

Il reste donc la question non triviale de savoir comment les ASNC devraient passer d'un raisonnement spéculatif exprimé dans des organigrammes à une norme industrielle de facto, et pourquoi cela est souhaitable.

Bien entendu, les gouvernements devraient suivre les recommandations formulées lors des sessions de 2022 et 2023 du Groupe d'experts gouvernementaux (GGE) des Nations unies sur les technologies émergentes dans le domaine des SALA et dans la Déclaration politique de 2023 sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie, qui vont clairement dans le sens du développement d'une forme ou d'une autre d'ASNC. Mais ces recommandations sont très vagues et loin d'être contraignantes. Les gouvernements nationaux les interpréteront inévitablement différemment, en fonction des intérêts d'industries spécifiques et d'autres groupes de pression pertinents pour leurs discours nationaux. On peut s'attendre à ce qu'il soit extrêmement difficile de trouver des accords internationaux plus solides concernant les applications militaires de l'IA. Jusqu'à présent, les nombreuses discussions concernant les LAWS et PALWS aux niveaux national et international n'ont même pas abouti à une définition solide des LAWS et PALWS. Il est urgent de promouvoir et d'accélérer les approches nationales et internationales visant à élaborer des réglementations cohérentes et solides concernant les applications militaires de l'IA.

Commencer par promouvoir les ASNC en tant que norme industrielle de facto, puis les transformer en norme industrielle officielle, serait la voie la plus faible sur le plan normatif pour parvenir à une réglementation des applications militaires de l'IA. Toutefois, il s'agirait également de la voie la plus proche des parties prenantes qui en savent le plus sur la faisabilité technologique actuelle et à moyen terme.

Pourquoi les producteurs de technologie militaire devraient-ils être intéressés par le développement et la mise en œuvre volontaires de ces ASNC ? Il y a quatre bonnes raisons à cela.

1) Il n'est pas improbable que, parallèlement à l'augmentation de la fabrication des PALWS et au développement d'unités logistiques d'IA pour les militaires en Europe, un débat public se

développe pour insister sur la nécessité d'équiper les applications militaires de l'IA avec des ASNC.

2) Les producteurs de technologies militaires font généralement l'objet d'une grande attention de la part du public et la mise en œuvre d'ASNC solides devrait être le seul moyen de justifier l'exportation de PALWS et d'unités logistiques d'IA pour l'armée. Les PALWS pourraient être particulièrement utiles aux régimes autoritaires, car ils réduisent le nombre de collaborateurs consentants qu'un régime doit maintenir en tant que sujets loyaux. L'exportation de PALWS et d'unités logistiques d'IA vers des régimes non démocratiques doit être absolument proscrite, car elle constitue une menace pour la stabilité politique partout dans le monde. Par ailleurs, les démocraties ne devraient recevoir des PALWS et des unités logistiques d'IA que si elles incluent des ASNC qui garantissent que les gouvernements n'utilisent pas les systèmes d'IA contre leur propre population, pour mener des guerres d'agression ou pour défendre des territoires annexés.

3) On peut s'attendre à ce que la recherche et le développement concernant les ASNC, en particulier s'ils sont financés par des fonds publics, créent des effets de synergie qui amélioreront également la fiabilité générale des PALWS et des unités logistiques d'IA pour l'armée. À long terme, les PALWS et les unités logistiques d'IA dotées d'ASNC sont susceptibles d'être plus performants à tous les égards.

4) Quatrièmement, les clients, c'est-à-dire les gouvernements qui achètent des applications militaires de l'IA, peuvent avoir de bonnes raisons de préférer un produit doté d'ASNC à un produit dépourvu de mécanismes de sécurité. Dans les régions instables du monde, comme la région du Sahel, il n'est pas rare que des gouvernements soient renversés par des chefs militaires qui utilisent abusivement des armes financées par des fonds publics pour faire un coup d'État. Les ASNC peuvent contribuer à atténuer ces risques et à promouvoir la stabilité politique en limitant l'utilisation des armes contre les opposants nationaux.

Résumé des recommandations politiques

- Promouvoir la recherche et le développement d'ASNC pour les PAWLS et les unités logistiques d'IA en Europe, conformément aux recommandations du rapport de la session de 2023 du GGE des Nations unies sur les technologies émergentes dans le domaine des SALA et de la déclaration politique de 2023 sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie.

- Promouvoir la recherche et le développement d'ASNC basés sur des systèmes d'intelligence artificielle hybrides, combinant des éléments fondés sur des données et des règles et des mécanismes de blocage simples basés sur des données de géolocalisation.

- Dans un premier temps, promouvoir le développement de mécanismes de blocage basés sur des données de géolocalisation, par exemple des mécanismes de blocage concernant les sites culturels et l'utilisation offensive de PALWS dans des zones très peuplées. Dans un deuxième temps, passer à l'élaboration d'ASNC plus complexes, par exemple en ce qui concerne la différenciation entre civils et combattants, l'évaluation de la proportionnalité et la reconnaissance des symboles protégés.

- Sensibiliser le public aux ASNC. Promouvoir des débats au sein de la société civile sur les principes directeurs des ASNC. Promouvoir des tables rondes avec les leaders industriels afin de trouver un accord et une convergence possibles sur les ASNC, en tenant compte également de la faisabilité technologique.

- Encourager les dirigeants industriels à faire des ASNC une norme industrielle de facto pour les PALWS et les unités logistiques AI *Made in Europe.*

- Encourager les dirigeants industriels à développer la supériorité technologique des produits européens afin d'inciter les clients du monde entier à accepter des produits équipés d'ASNC.

- Veiller à ce que le "lavage de l'ASNC" ne serve pas à justifier les exportations d'armes vers des régimes autoritaires.

- Poursuivre également d'autres voies non technologiques pour réglementer les applications militaires de l'IA aux niveaux national et international.

Références

Les agents d'intelligence artificielle du programme ACE passent de la simulation au vol réel. DARPA, 13 février 2023. https://www.darpa.mil/news-events/2023-02-13.

Opérations de l'armée de l'air et droit : A Guide for Air and Space Forces". Département du juge-avocat général de l'armée de l'air américaine, 2002.

Ambitieux, sûr, responsable : Our Approach To The Delivery Of AI-Enabled Capability In Defence" (Notre approche de la fourniture de capacités basées sur l'IA dans le domaine de la défense). Ministère britannique de la défense, juin 2022. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082991/20220614-Ambitious_Safe_and_Responsible.pdf.

Arkin, Ronald C. 'Governing Lethal Behavior : Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture". Mobile Robot Laboratory College of Computing Georgia Institute of Technology, 2006. https://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

Asmolov, Gregory. La transformation de la guerre participative : The Role of Narratives in Connective Mobilization in the Russia-Ukraine War". *Digital War*, 27 septembre 2022. https://doi.org/10.1057/s42984-022-00054-5.

Chengeta, Thompson. Accountability Gap : Autonomous Weapon Systems and Modes of Responsibility in International Law". *Denver Journal of International Law & Policy* 45, no. 1 Fall (janvier 2016). https://digitalcommons.du.edu/cgi/viewcontent.cgi?article=1011&context=djilp.

Department of Defense Directive 3000.09a Autonomy In Weapon Systems, 25 janvier 2023. https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF.

Eklund, Amanda Musco. Contrôle humain significatif des systèmes d'armes autonomes : Definitions and Key Elements in the Light of International Humanitarian Law and International Human Rights Law". Agence suédoise de recherche pour la défense, février 2020. https://www.fcas-forum.eu/publications/Meaningful-Human-Control-of-Autonomous-Weapon-Systems-Eklund.pdf.

Rapport final du groupe d'experts sur la Libye établi conformément à la résolution 1973 (2011) du Conseil de sécurité. Conseil de sécurité des Nations unies, 8 mars 2021. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/037/72/PDF/N2103772.pdf?OpenElement.

Grimal, Francis, et Michael Pollard. Le devoir de précaution dans les hostilités et la désobéissance aux ordres : Les robots doivent-ils refuser ? *Fordham International Law Journal*, 2021, 671-734.

Hambling, David. Ukraine's AI Drones Seek And Attack Russian Forces Without Human Oversight" (Les drones IA de l'Ukraine recherchent et attaquent les forces russes sans contrôle

humain). *Forbes*, 17 octobre 2023, sec. Aerospace & Defense. https://www.forbes.com/sites/davidhambling/2023/10/17/ukraines-ai-drones-seek-and-attack-russian-forces-without-human-oversight/.

Commandement et contrôle interarmées tous domaines (JADC2)". Congressional Research Service, 21 janvier 2022. https://sgp.fas.org/crs/natsec/IF11493.pdf.

Loitering Munition Systeme für Deutschland". MDBA Allemagne, 12 juillet 2023. https://www.mbda-deutschland.de/pressemitteilung/loitering-munition-systeme-fuer-deutschland/.

Machi, Vivienne. L'armée française fait appel à Nexter pour construire des drones kamikazes capables de détruire des chars". *C4ISRNet*, 19 juin 2023, sec. Unmanned. https://www.c4isrnet.com/global/europe/2023/06/19/french-army-taps-nexter-to-build-tank-busting-kamikaze-drones/.

Metz, Cade, et Gregory Schmidt. Elon Musk and Others Call for Pause on A.I., Citing "Profound Risks to Society" (Elon Musk et d'autres appellent à une pause sur l'I.A., citant des "risques profonds pour la société"). *The New York Times*, 29 mars 2023, sec. Technology. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

Morozov, Evgeny. *Pour tout sauver, cliquez ici : La folie du solutionnisme technologique*. Paperback 1. publ. New York, NY : PublicAffairs, 2014.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, et al. "Bias in Data-driven Artificial Intelligence Systems-An Introductory Survey". *WIREs Data Mining and Knowledge Discovery* 10, no. 3 (mai 2020) : e1356. https://doi.org/10.1002/widm.1356.

Avis sur l'intégration de l'autonomie dans les systèmes d'armes létaux. Comité d'éthique de la défense, avril 2021. https://www.defense.gouv.fr/sites/default/files/ministere-armees/20210429_Comité%20d%27éthique%20de%20la%20défense%20-%20Avis%20intégration%20autonomie%20systèmes%20armes%20létaux%20-%20Version%20anglaise.pdf.pdf.

Rapport de la session 2022 du groupe d'experts gouvernementaux sur les technologies émergentes dans le domaine des systèmes d'armes autonomes létaux, 29 juillet 2022. https://documents.unoda.org/wp-content/uploads/2022/08/CCW-GGE.1-2022-CRP.1-Rev.1-As-Adopted-on-20220729.pdf.

Rapport de la session 2023 du groupe d'experts gouvernementaux sur les technologies émergentes dans le domaine des systèmes d'armes autonomes létaux, 24 mai 2023. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_2_Advance_version.pdf.

Saltman, Max. As Azerbaijan Claims Final Victory in Nagorno Karabakh, Arms Trade with Israel Comes under Scrutiny" (L'Azerbaïdjan revendique la victoire finale dans le Haut-Karabakh, le commerce d'armes avec Israël fait l'objet d'une attention particulière). CNN, 4

octobre 2023. https://www.cnn.com/2023/10/04/middleeast/azerbaijan-israel-weapons-mime-intl/index.html.

Scharre, Paul. Centaur Warfighting : Le faux choix de l'humain contre l'automatisation'. *International and Comparative Law Journal* 30, no 1 (2016) : 151-65.

Scholz, Jason, et Jai Galliott. The Humanitarian Imperative for Minimally-Just AI in Weapons " (L'impératif humanitaire pour une IA minimalement juste dans les armes). *Journal of Indo-Pacific Affairs*, hiver 2018, 57-67. https://doi.org/10.1093/oso/9780197546048.003.0005.

Sparrow, Robert J., et Adam Henschke. Minotaures, Not Centaurs : The Future of Manned-Unmanned Teaming". *The US Army War College Quarterly : Parameters* 53, no. 1 (3 mars 2023). https://doi.org/10.55540/0031-1723.3207.

Résumé de la stratégie de commandement et de contrôle interarmées tous domaines (JADC2). Département de la défense des États-Unis, mars 2022. https://media.defense.gov/2022/Mar/17/2002958406/-1/-1/1/SUMMARY-OF-THE-JOINT-ALL-DOMAIN-COMMAND-AND-CONTROL-STRATEGY.PDF.

Département d'État des États-Unis. Déclaration politique sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie, 16 février 2023. https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/

Van Wynsberghe, Aimee, et Scott Robbins. Critiquing the Reasons for Making Artificial Moral Agents". *Science and Engineering Ethics* 25, no. 3 (juin 2019) : 719-35. https://doi.org/10.1007/s11948-018-0030-8.

Wallace, Rodrick. How AI Founders on Adversarial Landscapes of Fog and Friction". *The Journal of Defense Modeling and Simulation : Applications, Methodology, Technology* 19, no. 3 (juillet 2022) : 519-38. https://doi.org/10.1177/1548512920962227.

Wiener, Norbert. *L'utilisation humaine des êtres humains : Cybernetics and Society*. Houghton Mifflin, 1950.

Yan, Guilong. The Impact of Artificial Intelligence on Hybrid Warfare" (L'impact de l'intelligence artificielle sur la guerre hybride). *Small Wars & Insurgencies* 31, no. 4 (18 mai 2020) : 898-917. https://doi.org/10.1080/09592318.2019.1682908.

Zurek, Tomasz, Jonathan Kwik, et Tom Van Engers. Model of a Military Autonomous Device Following International Humanitarian Law" (Modèle de dispositif militaire autonome conforme au droit international humanitaire). *Ethics and Information Technology* 25, no. 1 (mars 2023) : 15. https://doi.org/10.1007/s10676-023-09682-1.

Notes de bas de page

[1] Hambling, David. Ukraine's AI Drones Seek And Attack Russian Forces Without Human Oversight" (Les drones IA de l'Ukraine recherchent et attaquent les forces russes sans surveillance humaine). *Forbes*, 17 octobre 2023, sec. Aerospace & Defense. https://www.forbes.com/sites/davidhambling/2023/10/17/ukraines-ai-drones-seek-and-attack-russian-forces-without-human-oversight/.

[2] "Rapport final du groupe d'experts sur la Libye établi conformément à la résolution 1973 (2011) du Conseil de sécurité" (Conseil de sécurité des Nations unies, 8 mars 2021), https://digitallibrary.un.org/record/3905159.

[3] Max Saltman, "As Azerbaijan Claims Final Victory in Nagorno Karabakh, Arms Trade with Israel Comes under Scrutiny", CNN, 4 octobre 2023, https://www.cnn.com/2023/10/04/middleeast/azerbaijan-israel-weapons-mime-intl/index.html.

[4] Vivienne Machi, "French Army Taps Nexter to Build Tank-Busting Kamikaze Drones", *C4ISRNet*, 19 juin 2023, sec. Unmanned, https://www.c4isrnet.com/global/europe/2023/06/19/french-army-taps-nexter-to-build-tank-busting-kamikaze-drones/. Loitering Munition Systeme für Deutschland" (MDBA Allemagne, 12 juillet 2023), https://www.mbda-deutschland.de/pressemitteilung/loitering-munition-systeme-fuer-deutschland/.

[5] "ACE Program's AI Agents Transition from Simulation to Live Flight" (DARPA, 13 février 2023), https://www.darpa.mil/news-events/2023-02-13.

[6] "Joint All-Domain Command and Control (JADC2)" (Congressional Research Service, 21 janvier 2022), https://sgp.fas.org/crs/natsec/IF11493.pdf.

[7] Thompson Chengeta, "Accountability Gap : Autonomous Weapon Systems and Modes of Responsibility in International Law", *Denver Journal of International Law & Policy* 45, no. 1 Fall (janvier 2016), https://digitalcommons.du.edu/cgi/viewcontent.cgi?article=1011&context=djilp.

[8] Amanda Musco Eklund, "Meaningful Human Control of Autonomous Weapon Systems : Definitions and Key Elements in the Light of International Humanitarian Law and International Human Rights Law" (Agence suédoise de recherche sur la défense, février 2020), https://www.fcas-forum.eu/publications/Meaningful-Human-Control-of-Autonomous-Weapon-Systems-Eklund.pdf.

[9] "Avis sur l'intégration de l'autonomie dans les systèmes d'armes létaux" (Comité d'éthique de la défense, avril 2021), https://www.defense.gouv.fr/sites/default/files/ministere-armees/20210429_Comité%20d%27éthique%20de%20la%20défense%20-%20Avis%20intégration%20autonomie%20systèmes%20armes%20létaux%20-%20Version%20anglaise.pdf.pdf.

[10] "Ambitieux, sûr, responsable : Our Approach To The Delivery Of AI-Enabled Capability In Defence" (Ministère britannique de la défense, juin 2022), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082991/20220614-Ambitious_Safe_and_Responsible.pdf.

[11] "Department of Defense Directive 3000.09a Autonomy In Weapon Systems", 25 janvier 2023, https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF.

[12] "Rapport de la session 2022 du groupe d'experts gouvernementaux sur les technologies émergentes dans le domaine des systèmes d'armes autonomes létaux", 29 juillet 2022, https://documents.unoda.org/wp-content/uploads/2022/08/CCW-GGE.1-2022-CRP.1-Rev.1-As-Adopted-on-20220729.pdf.

[13] "Rapport de la session de 2023 du groupe d'experts gouvernementaux sur les technologies émergentes dans le domaine des systèmes d'armes autonomes létaux", 24 mai 2023, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_2_Advance_version.pdf.

[14] Département d'État des États-Unis. Déclaration politique sur l'utilisation militaire responsable de l'intelligence artificielle et de l'autonomie, 16 février 2023. https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/.

[15] Aimee Van Wynsberghe et Scott Robbins, " Critiquing the Reasons for Making Artificial Moral Agents ", *Science and Engineering Ethics* 25, no. 3 (juin 2019) : 719-35, https://doi.org/10.1007/s11948-018-0030-8.

[16] Ronald C. Arkin, "Governing Lethal Behavior : Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture' (Mobile Robot Laboratory College of Computing Georgia Institute of Technology, 2006), https://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

[17] Jason Scholz et Jai Galliott, " The Humanitarian Imperative for Minimally-Just AI in Weapons ", *Journal of Indo-Pacific Affairs*, hiver 2018, 57-67, https://doi.org/10.1093/oso/9780197546048.003.0005.

[18] Francis Grimal et Michael Pollard, "The Duty to Take Precautions in Hostilities, and the Disobeying of Orders : Should Robots Refuse ?", *Fordham International Law Journal*, 2021, 671-734.

[19] Robert J. Sparrow et Adam Henschke, "Minotaurs, Not Centaurs : The Future of Manned-Unmanned Teaming", *The US Army War College Quarterly : Parameters* 53, no. 1 (3 mars 2023), https://doi.org/10.55540/0031-1723.3207.

[20] Tomasz Zurek, Jonathan Kwik et Tom Van Engers, "Model of a Military Autonomous Device Following International Humanitarian Law", *Ethics and Information Technology* 25, no 1 (mars 2023) : 15, https://doi.org/10.1007/s10676-023-09682-1.

[21] Rodrick Wallace, "How AI Founders on Adversarial Landscapes of Fog and Friction", *The Journal of Defense Modeling and Simulation : Applications, Methodology, Technology* 19, no. 3 (juillet 2022) : 519-38, https://doi.org/10.1177/1548512920962227.

[22] Guilong Yan, " The Impact of Artificial Intelligence on Hybrid Warfare ", *Small Wars & Insurgencies* 31, no. 4 (18 mai 2020) : 898-917, https://doi.org/10.1080/09592318.2019.1682908.

[23] Evgeny Morozov, *Pour tout sauver, cliquez ici : The Folly of Technological Solutionism*, Paperback 1. publ (New York, NY : PublicAffairs, 2014).

[24] Sparrow et Henschke, "Minotaures, Not Centaurs".

[25] Norbert Wiener, *The Human Use of Human Beings : Cybernetics and Society* (Houghton Mifflin, 1950).

[26] Scholz et Galliott, "The Humanitarian Imperative for Minimally-Just AI in Weapons".

[27] Grimal et Pollard, "The Duty to Take Precautions in Hostilities, and the Disobeying of Orders : Les robots doivent-ils refuser ?

[28] Sparrow et Henschke, "Minotaures, Not Centaurs".

[29] Paul Scharre, " Centaur Warfighting : The False Choice of Humans vs. Automation', *International and Comparative Law Journal* 30, no 1 (2016) : 151-65.

[30] Gregory Asmolov, "The Transformation of Participatory Warfare : The Role of Narratives in Connective Mobilization in the Russia-Ukraine War", *Digital War*, 27 septembre 2022, https://doi.org/10.1057/s42984-022-00054-5.

[31] Zurek, Kwik, et Van Engers, "Model of a Military Autonomous Device Following International Humanitarian Law".

[32] Cade Metz et Gregory Schmidt, "Elon Musk and Others Call for Pause on A.I., Citing "Profound Risks to Society"", *The New York Times*, 29 mars 2023, sec. Technologie, https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

[33] "Air Force Operations and the Law : A Guide for Air and Space Forces' (US Air Force Judge Advocate General's Department, 2002).

[34] Eirini Ntoutsi et al, " Bias in Data-driven Artificial Intelligence Systems-An Introductory Survey ", *WIREs Data Mining and Knowledge Discovery* 10, no. 3 (mai 2020) : e1356, https://doi.org/10.1002/widm.1356.

Remerciements :

Un grand merci à Florence G'sell pour ses précieuses remarques sur le manuscrit et à Jessica Dorsey pour avoir porté à mon attention des rapports sur l'utilisation par l'Ukraine de drones entièrement autonomes. Au cours de cette recherche, j'ai reçu un financement du programme de recherche Horizon 2020 de l'UE dans le cadre de la convention de subvention MSCA COFUND 101034352, avec un cofinancement du Fonds de recherche industrielle de la VUB.

A propos de l'auteur :

Johannes THUMFART est chercheur principal et postdoctoral au sein du groupe de recherche Law, Science, Technology, and Society (LSTS) de la Vrije Universiteit Brussel et maître de conférences adjoint en éthique de la gestion de la sécurité internationale à la Berlin School of Economics and Law. Ses recherches portent sur l'éthique de la sécurité internationale, l'utilisation militaire de l'IA et la souveraineté numérique dans les pays du BRICS. M. Thumfart a obtenu son doctorat en histoire et philosophie du droit international à l'université Humboldt de Berlin. Il a obtenu une bourse Marie Skłodowska-Curie et a occupé de nombreux postes d'enseignement et de recherche en Allemagne, en France, au Mexique et en Belgique. Il a collaboré à certains des journaux allemands les plus importants et les plus respectés, tels que *Der Spiegel* et *Die Zeit*. Sa monographie sur la souveraineté numérique sera publiée par Palgrave Macmillan à l'été 2024. Ses recherches sont publiées dans des revues telles que *European Journal of International Security*, *AI and Ethics* et *Global Studies Quarterly*.

À propos de la chaire "Numérique, gouvernance et souveraineté" :

La Chaire Numérique, Gouvernance et Souveraineté de Sciences Po a pour mission de favoriser un forum unique réunissant des entreprises techniques, des universitaires, des décideurs politiques, des acteurs de la société civile, des incubateurs de politiques publiques ainsi que des experts de la régulation numérique.

Hébergée par l'École d'affaires publiques, la Chaire adopte une approche pluridisciplinaire et holistique pour rechercher et analyser les transformations économiques, juridiques, sociales et institutionnelles induites par l'innovation numérique. La Chaire Numérique, Gouvernance et Souveraineté est présidée par **Florence G'sell**, professeur de droit à l'Université de Lorraine, maître de conférences à l'École d'affaires publiques de Sciences Po. Durant l'année universitaire 2023-2024, Florence G'sell est professeur invité au Cyber Policy Center de l'Université de Stanford.

Les activités de la Chaire sont soutenues par :

sopra steria next

Caisse des Dépôts GROUPE