HAL open science

Contribution LNE au GT Intelligence Artificielle du CT3

Agnes Delaborde, David Madaschi

▶ **To cite this version:**

Agnes Delaborde, David Madaschi. Contribution LNE au GT Intelligence Artificielle du CT3. 2020. hal-03066948

HAL Id: hal-03066948

https://hal.science/hal-03066948v1

Preprint submitted on 15 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dr. Delaborde Agnes, ingénieur de recherche en évaluation IA & robotique, DEC 536 Evaluation des systèmes d'intelligence artificielle
M. Madaschi David, ingénieur expert directive machine, DEC 531 Sécurité électrique industrielle et marquage CE

13/02/2019

Contribution LNE au GT Intelligence Artificielle du CT3

Propos liminaires – statut et expertise du LNE

Le Laboratoire national de métrologie et d'essais (LNE) est un EPIC sous la tutelle du Ministère de l'Industrie. En tant qu'organisme indépendant et tiers de confiance, le LNE a pour missions l'évaluation et la certification des produits issus de la recherche et de l'industrie.

Expertise en Intelligence Artificielle (IA). Le LNE a organisé depuis 2008 plus de 800 évaluations de systèmes d'IA pour de nombreuses applications (traitement automatique des langues, robotique, véhicule autonome, traitement d'image et de vidéos, etc.). Ces évaluations sont menées : a) pour le compte des pouvoirs publics, pour promouvoir et maîtriser le développement de l'IA, b) pour les développeurs d'IA du secteur privé ou public, de sorte à faciliter et accompagner le développement, et c) pour fournir aux utilisateurs finaux la possibilité d'effectuer un choix éclairé lors de la sélection d'une solution parmi l'ensemble des solutions proposées sur le marché. Les évaluations concernent tant les performances des algorithmes, que la sécurité ou la qualité du système global. Le LNE est membre de la Commission Intelligence Artificielle AFNOR/CN JTC1/SC 42.

Expertise Directive « Machines ». Le LNE accompagne les fabricants de produits ou équipements électriques pour leur permettre de répondre aux exigences normatives relatives à l'application des directives européennes, dont la Directive « Machines ». A cet effet, le LNE dispose de bancs d'essai standardisés et de moyens techniques pour la réalisation des essais de conformité susceptibles de garantir la sécurité des produits électriques et électroniques. Le LNE propose également des prestations d'assistance technique, d'expertise et d'accompagnement des fabricants dans leur démarche de réponse aux exigences réglementaires.

Nature de la contribution LNE

Suite aux discussions initiées le 20 novembre 2018 dans la cadre de la réunion « IA » pilotée par le CT3, le LNE souhaite contribuer sur ces points :

- Définition et nature de l'IA dans la Directive « Machines »
- Qualifier les niveaux d'autonomie dans la Directive « Machines »
- Modularité du système autonome : faut-il considérer la machine ou ses modules
- Lecture de la Directive « Machines » à la lumière des nouvelles capacités apportées par l'IA

Définition et nature de l'IA dans la Directive « Machines »

Nous notons :

- qu'il est nécessaire que la définition de l'IA soit à-même de couvrir toute la gamme de systèmes dits « dotés d'intelligence »,
- qu'il est nécessaire de disposer d'une définition qui ne soit pas sujette à interprétation, ni ambiguë, et qu'elle fasse raisonnablement consensus parmi les communautés traitant de l'IA,
- qu'il est nécessaire de prévoir une façon de traiter de l'IA qui permette que les adaptations de la Directive « Machines » soit raisonnablement pérennes.

Cependant, nous remarquons :

- que nous ne sommes pas en mesure d'estimer la forme que prendra l'IA dans un avenir proche ; les fonds importants consentis pour la recherche et le développement en robotique et IA sont propices à l'émergence de technologies de rupture,
- qu'il n'existe pas à l'heure actuelle de définition de l'IA faisant consensus parmi les entités impliquées dans l'utilisation, la réglementation et le développement de l'IA.

Il semble nécessaire qu'une définition incluse dans la Directive « Machines » soit homogène avec la définition du comité technique de normalisation ISO/IEC JTC1 (Information technology), comité susceptible de produire des normes pouvant servir d'appui à l'application de la Directive « Machines », et ce afin de limiter les ambiguïtés lors de la vérification de conformité des machines dotées d'IA.

A l'heure actuelle, le sous-comité SC 42 « Artificial intelligence » de l'ISO/IEC JTC1, créé fin 2017, travaille à la définition d'un grand nombre de concepts liés à l'IA. Pour l'instant, aucune définition de l'IA n'a été proposée par le SC 42.

La norme ISO/IEC WD 23053:2018[1] en cours d'élaboration liste deux types d'implémentation des systèmes d'IA[2] : l'apprentissage automatique (système créant ses propres règles par méthode probabiliste à partir d'un jeu de données), et les systèmes expert (dont les algorithmes sont composés principalement de règles définies « en dur » par l'expert humain).

Nous trouvons une définition normée de l'IA par l'ISO/IEC JTC1, dans la norme ISO/IEC 2382:2015[3] : « *Capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning and learning. (SOURCE: ISO/IEC 2382-28:1995)* ». Nous ne savons pas si une telle définition tiendra face aux courants de pensée actuels qui contribuent à « démystifier » l'IA. Cette démystification vise à ramener à de justes proportions ce qui relève des capacités réelles de l'IA commercialisable (raisonnements « simples », très éloignés de la richesse et de la complexité des capacités humaines), et ce qui relève de l'IA de laboratoire (prototype ultra-performants capables de raisonnements complexes, mais fonctionnant en environnement extrêmement contrôlé et en contextes d'application très restreints). Cette volonté de démystification est particulièrement prégnante au sein de la communauté européenne, comme en atteste la présentation de madame Huet synthétisant en 2017 les initiatives de l'UE face à l'IA[4], insistant sur la nécessité de ne plus situer l'IA dans le domaine de la science-fiction, mais d'en éprouver scientifiquement ses limites, ses performances, sa sécurité – cette présentation n'est qu'un exemple parmi de nombreux autres.

Dans cette même idée de démystification, nous ne conseillons pas l'emploi des termes IA « forte » et IA « faible », que nous estimons issus de courants de pensées scientifiques, à visée notamment médiatique et explicative, mais qui ne constituent pas en eux-mêmes des définitions rigoureuses. En effet, cette distinction « faible » / « forte » est utilisée, principalement par dérive médiatique, tout autant pour faire référence au mode d'implémentation du système IA, qu'à la tâche du système IA. En effet, une IA peut se retrouver qualifiée de « forte » si elle s'appuie sur des algorithmes complexes, tels que

[1] ISO/IEC WD 23053:2018 « Framework for artificial intelligence systems using machine learning ». Document de travail.

[2] Le projet de norme cite également la programmation logique comme technique d'implémentation. Nous estimons toutefois qu'il s'agit d'une méthode spécifique de programmation d'un système expert ou d'un système par apprentissage automatique, et non d'une technique d'implémentation de l'IA à part entière.

[3] ISO/IEC 2382:2015 « Information technology -- Vocabulary ».

[4] *European Commission:* Cécile Huet, Deputy Head of Unit, Robotics and Artificial Intelligence, EC, DG CONNECT – The European Commission's initiatives on AI, conference Intelligent Machines, Smart Policies, Thursday 26 October 2017: AI developments and applications, OECD, Paris

des réseaux de neurones convolutifs – par opposition à un système expert par règles ou un automate à états finis. D'autre part, des tâches haut-niveau (i.e. proches de l'humain) telles qu'une tâche de détection automatique de signaux sociaux ou la simulation de processus cognitifs, peuvent être considérées comme relevant de l'IA « forte ». Nous notons que la notion d'IA « forte » fait originellement référence à la capacité d'un système à être doté d'une conscience propre. Cette dénomination fait donc plutôt référence à un futur développement hypothétique de l'IA, en lien par exemple avec les théories sur la singularité technologique, qu'à une réalité proche. La définition de la « force » de l'IA ne présente donc selon nous pas les caractéristiques attendues d'une définition utilisable en contexte réglementaire.

Dans le rapport « Draft Report of the SC 42 Study Group on Trustworthiness for AI » (31/07/2018), la définition d'un système d'IA est la suivante : « *any system that makes decisions, either based on statistical and machine learning algorithms (including but not limited to regressions, decision trees, support vector machines, neural networks and other algorithms that can learn from examples) or based on formal rule definitions* ». Cette définition présente l'avantage de se focaliser tout d'abord sur la capacité d'un système à prendre une décision, quel que soit le mode d'implémentation. Le rapport précise également : « *Artificial intelligence technologies are broad and include Deep learning, Knowledge Representation, Reasoning, Planning & Scheduling, Constraint Programming, Linear Programming and Optimization, Rules based systems, State Machines, Context free grammars, etc.* ». Cette description permet de mettre en avant la richesse des méthodes d'implémentation de l'IA, et l'impossibilité d'être exhaustif.

Le pré-rapport « Ethics Guidelines for Trustworthy AI »[5] de la Commission Européenne s'appuie sur une description des fonctionnalités systémiques, et non pas sur les méthodes d'implémentation du système : « *Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions.* ». Dans le même esprit, le rapport émis par le TNO[6] décrit l'autonomie d'un robot de cette façon : « *Autonomously' in this context means that they are programmed to be able to take decisions for themselves (using AI). These robots can 'see' their environment using sensors and anticipate it and respond to it.* ». Dans ces descriptions des systèmes, l'IA est juste abordée comme un élément permettant au système de jouer son rôle classique de perception – interprétation – décision – action. Une telle définition présente l'avantage de se focaliser sur les fonctionnalités du système, qui sont similaires d'un système à un autre, et non sur ses méthodes d'implémentation, qui elles, peuvent prendre des formes très variées et évoluer rapidement.

En conclusion, nous estimons qu'il serait plus adéquat de ne pas parler d'IA dans la Directive « Machines », car l'IA n'est qu'une méthode pour parvenir à l'automatisation d'un process, et la surexploitation du terme mène à une grande confusion sur la nature-même de l'IA. Nous rejoignons la tendance des instances normatives et des groupes de travaux européens à se focaliser sur les fonctionnalités des systèmes. Ceci semble plus en phase avec la Directive « Machines », qui ne s'intéresse pas à la façon dont la machine est conçue, mais à sa finalité. Il pourra alors s'agir soit d'une approche en termes de système (perception-décision-action, ou encore entrée-traitement-sortie), ou encore en termes de tâche (reconnaissance, optimisation, prédiction, etc.). Un aspect essentiel est d'insister sur le fait que la capacité de décision (ou de traitement) est réalisée de façon automatique.

Qualifier les niveaux d'autonomie dans la Directive « Machines »

Il est intéressant alors de qualifier, au sein de la Directive, le degré d'autonomie de la machine. En effet, la capacité d'un système à décider de ses propres actions peut avoir un impact sur l'émergence de situations à risques (perte de contrôle, mouvements erratiques, etc.), qu'il convient de pouvoir anticiper. Cette capacité décisionnelle, et les conséquences dommageables possibles, est liée à l'autonomie accordée au système, et la Directive doit en cela proposer une gradation de l'autonomie. La commission SAE définit par exemple, dans le cadre du véhicule autonome, six niveaux d'autonomie[7], allant de

[5] The European Commission's High-level Expert Group on Artificial Intelligence (2018) « Draft – Ethics Guidelines for Trustworthy AI, Working Document for stakeholders' consultation ». Brussels, 18 Dec. 2018.

[6] TNO (2018). « Emergent Risks to Workplace Safety; Working in the Same Space as a Cobot. ». Report for Ministry of Social Affairs and Employment. WG-2018.26. Netherlands, 28 August 2018.

[7] Society of Automotive Engineers. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. J3016_201806, 2018-06-15.

l'absence totale d'autonomie (le conducteur prend toutes les décisions, sans intervention de l'électronique), jusqu'à l'autonomie totale (le véhicule peut prendre toutes les décisions sans intervention humaine, quelles que soient les conditions). Dans le secteur industriel, des chercheurs suédois proposent en 2008, dans le cadre d'une définition taxonomique de l'automatisation industrielle particulièrement détaillée[8], une définition du niveau d'automatisation d'une machine comme « *[t]he allocation of physical and cognitive tasks between humans and technology, described as a continuum ranging from totally manual to totally automatic* ». La taxonomie propose une classification de l'autonomie sur sept niveaux, selon la fonction du dispositif (mécanique, structure de contrôle, dispositif d'information), et la nature des opérations réalisées tant par la machine que par l'opérateur. Cette taxonomie, exemplifiée, permet de qualifier le niveau d'autonomie des machines sur chaque poste de travail.

Les initiatives pour la qualification de l'autonomie sont nombreuses, et la Directive « Machines » doit à notre sens s'en inspirer dans sa définition, non pas de l'IA, mais des capacités de prise de décision automatique des machines.

Modularité du système autonome : faut-il considérer la machine ou ses modules

Les machines automatisées peuvent se présenter sous la forme d'ensembles complexes, c'est-à-dire composés de différents modules pouvant avoir chacun la capacité à « prendre une décision ». L'ensemble peut être composé par exemple de capteurs intelligents, prenant une décision « bas niveau » à partir des éléments perçus dans son environnement, et d'effecteurs (pour la préhension, la mobilité, etc.) s'appuyant chacun sur ses propres algorithmes de décision, mais dirigés par une unité de contrôle central supervisant et hiérarchisant le comportement général du système. Comme nous l'avons précédemment mentionné, il semble essentiel d'appréhender l'ensemble du point de vue des fonctions systémiques : entrée(s) – traitement(s) – sortie(s). Dans le cas d'une machine automatisée, même composée de différents modules embarquant de l'IA, il pourrait être envisageable de devoir évaluer la sécurité de son comportement comme un ensemble lorsque la machine est dotée d'un système permettant d'assurer un partage d'informations entre les différents modules. Ce serait le cas par exemple d'un système robotisé doté d'un superviseur capable de traiter les informations émanant de différentes sources et de réguler les comportements des différents modules ; mais là encore les choix terminologiques quant à la notion de « superviseur » peuvent amener à débat, et la notion de « partage d'informations » peut être suffisante pour caractériser les liens entre les différents modules.

Le groupe de travail de l'ISO/TC 299 signale notamment, dans le cadre de l'ISO Focus 2018 sur le partage de l'espace de travail avec un système robotisé[9], que la sécurité doit être évaluée à l'échelle du système complet et des éléments avec lesquels il interagit : « *"only a completed application can be described as safe". This means looking at the whole applied robot system, rather than just an individual part of it. Designing safety into the robots themselves, whilst looking at how they interact as part of a system (in particular where they interact with humans), is the next building block. "These two distinct aspects of addressing safety are covered by ISO 10218-1 and 10218-2 respectively, both of which are currently under revision"*».

Il pourrait être intéressant d'effectuer une distinction entre les robots composés de modules prenant des décisions et étant interconnectés (partageant de l'information dans différentes directions, qu'il s'agisse de données, de commandes, etc.), et les briques implantées sur une plateforme capables de prendre une décision, mais n'échangeant pas d'information avec le reste de la plateforme (il pourrait alors éventuellement s'agir d'une quasi-machine). Il pourrait être nécessaire d'imposer que les interactions entre les différents modules d'IA soient listées, afin de déterminer le périmètre de l'application autonome dont la sécurité devra être vérifiée.

[8] Frohm, Jörgen, Veronica Lindström, Mats Winroth, and Johan Stahre. "Levels of automation in manufacturing." Ergonomia(2008).

[9] ISO/TC 299 ISO Focus - Connecting with robot co-workers. 2018-11-15

Lecture de la DM à la lumière des nouvelles capacités apportées par l'IA

Face à la volonté de décartériser les robots, c'est-à-dire de laisser la possibilité à la machine d'effectuer ses opérations même lorsqu'un travailleur humain se trouve dans l'espace de travail, les procédures de sécurité sont modifiées. En effet, il n'est pas attendu du robot qu'il interrompe sa tâche à l'arrivée de l'opérateur, mais qu'il réduise sa vitesse, interrompe son mouvement en cas de contact, reçoive éventuellement des commandes par pression sur ses effecteurs, ou assiste l'opérateur dans la réalisation d'une tâche collaborative.

L'autonomie des machines peut également les amener à être mobiles dans l'usine, à déterminer de façon dynamique leur cadence, leurs activations et leurs pauses. Les robots actuellement implantés dans les usines sont principalement mus par automate, ce qui signifie que leur comportement est déterministe, et ne peut s'effectuer que dans un cadre extrêmement limité et connu du programmeur de l'automate (usuellement, un technicien de l'atelier). Dans le cas d'un système dirigé par une IA toutefois, la complexité des raisonnements n'est pas immédiatement explicable pour le personnel présent à proximité de la machine, c'est-à-dire que le comportement de la machine peut être moins anticipable que dans le cas d'un automate.

Il pourrait être nécessaire d'apporter des nuances à certaines exigences, lorsque celles-ci forcent à ce que la machine ne puisse effectuer une opération sans l'ordre de l'opérateur. Ces exigences peuvent être un frein à l'autonomie de la machine, et donc à son rendement attendu. Citons par exemple :

- « *les paramètres de la machine ne doivent pas changer sans qu'un ordre ait été donné à cet effet* » (Annexe I, 1.2.1) : un système doté d'IA peut définir ses propres paramètres (cadence, directions, etc.). Toutefois, la Directive précise que cela ne s'applique que lorsque le changement de paramètre peut amener à une situation dangereuse.
- S'il n'existe pas de moyen pour que la machine soit lancée sans pouvoir s'assurer qu'il n'y ait de personnes dans la zone, « *un signal sonore et/ou visuel doit être donné avant la mise en marche de la machine. Les personnes exposées doivent avoir le temps de quitter la zone dangereuse ou d'empêcher le démarrage de la machine* » (Annexe I, 1.2.2). Dans l'idée de la décartérisation, cette consigne n'est pas exactement applicable. Ceci ne devrait pas être attendu dans le cas des machines capables de détecter une présence et d'adapter leur fonctionnement en conséquence.

Bien que cette notion soit présente en filigrane dans l'intégralité du document, il pourrait être nécessaire que la Directive statue clairement sur la nécessité de pouvoir reprendre le contrôle de la machine lorsqu'elle est entièrement autonome. Cette remarque vaut particulièrement dans le cas des robots mobiles, pouvant se mettre hors de portée visuelle d'un opérateur susceptible de décréter la nécessité d'un arrêt d'urgence. D'autre part, dans ce cas, l'absence éventuelle de poste de commande pourrait rendre l'arrêt d'urgence impossible si la machine s'est placée en zone de danger, et des solutions palliatives doivent être proposées par le constructeur.

Dans le cas des machines pouvant choisir leur propre cadence, leurs moments d'intervention sur une ligne de production par exemple, il est également nécessaire de prévoir qu'elles signalent de façon non équivoque leur statut de veille, pour ne pas laisser croire aux personnes à proximité que la machine est éteinte. Une machine en fonctionnement doit donc signaler convenablement le fait qu'elle est susceptible de s'activer à tout moment (Annexe I, 1.7.1.2), qu'il s'agisse de témoins lumineux, ou de mouvements passifs.

A notre sens, la Directive « Machines » est suffisamment claire quant aux risques à supprimer ou réduire, ainsi qu'à l'évitement des situations dangereuses. Dans tous les cas étudiés à la lumière de l'IA, il apparaît toutefois que les capacités d'autonomie ne sont pas tout à fait gérées dans la Directive, dans la mesure où l'IA apportera notamment la possibilité de proposer des solutions de remédiation alternatives : adaptation à la présence de l'opérateur, capacité à gérer ses propres paramètres. La mobilité des machines nécessite également de prêter une attention particulière aux exigences traitant des fonctionnalités de reprise de contrôle.

De par notre statut d'évaluateurs, le point nous apparaissant le plus critique pour l'adoption des systèmes robotisés dotés d'IA repose sur la vérification de conformité. En effet, si la Directive est très claire sur la nécessité que l'action de la machine n'entraîne pas de situations dangereuses, l'estimation de la probabilité d'émergence du phénomène dangereux, ainsi que l'estimation de la couverture des comportements réalisables par les systèmes d'IA, sont à l'heure actuelle des questions non résolues.