La Découverte

Quatre nuances de régulation de l'intelligence artificielle

Une cartographie des conflits de définition

Bilel Benbouzid, Yannick Meneceur, Nathalie Alisa Smuha

Dans Réseaux 2022/2 (N° 232-233), pages 29 à 64
Éditions La Découverte

ISSN 0751-7971
ISBN 9782348073809
DOI 10.3917/res.232.0029

Article disponible en ligne à l'adresse
https://www.cairn.info/revue-reseaux-2022-2-page-29.htm

DOSSIER

Contrôler l'intelligence artificielle ?

QUATRE NUANCES DE RÉGULATION DE L'INTELLIGENCE ARTIFICIELLE

Une cartographie des conflits de définition

Bilel BENBOUZID
Yannick MENECEUR
Nathalie A. SMUHA

DOI: 10.3917/res.232-233.0029

L'intelligence artificielle (et l'acronyme qui lui est associé : « IA ») fonctionne comme une appellation signalétique, employée couramment comme si elle correspondait à une chose bien connue. Pourtant, il est impossible d'en proposer une définition unanime. Tantôt signifié, tantôt signifiant, le syntagme nominal « IA » est une catégorie mouvante, difficile à cerner. À bien y réfléchir, ce problème définitionnel de l'IA n'est pas seulement lié aux différentes approches scientifiques qui lui sont associées (symbolique, connexionniste, probabiliste, programmation évolutive, automate cellulaire, etc.) ou à l'hétérogénéité de ses objets d'étude (vision, langage, raisonnement, planification, sens commun, etc.). Il tient aussi aux multiples fonctions référentielles du syntagme nominal : parfois mobilisé en tant que concept, d'autres fois uniquement comme un artefact ou, lorsqu'il est écrit en majuscule (à la manière par exemple de *l'Église* ou de *l'État*), comme une chose diffuse – un Léviathan qui peut agir à tout moment dans notre vie quotidienne.

Depuis une dizaine d'années, le retour en force du syntagme nominal IA sur le devant de la scène politique et médiatique n'est pas sans poser de difficulté aux acteurs qui œuvrent à la recherche de solutions pour son encadrement. En enquêtant sur les débats autour des enjeux de la « régulation »[1] de l'IA, nous avons observé que les problèmes définitionnels étaient au cœur de conflits sur les moyens d'assujettir l'IA à un « contrôle social » (Collingridge, 1980), qu'il soit technique, éthique, juridique ou politique. En prenant comme fil rouge de l'analyse les significations variées de l'IA, cet article vise à participer à la compréhension des tensions normatives sur son contrôle. Nous proposons une cartographie des lieux, des acteurs et des approches qui donnent à voir comment les débats autour du contrôle de l'IA se structurent en quatre arènes normatives différenciées : l'IA comme « super-intelligence » dont il faut, à long terme, anticiper les risques de perte de contrôle ; l'IA comme discipline scientifique pour laquelle les chercheurs doivent garantir, à court terme, la sûreté, la robustesse, l'équité, l'explicabilité, etc. ; l'IA comme

1. La notion de régulation est entendue dans un sens large comme l'ensemble des formes de productions normatives qui visent à contrôler l'IA.

système socio-technique dont les implications sur la société nécessitent une approche qui va au-delà de simples exigences éthiques non contraignantes ; enfin l'IA comme système technico-économique que les producteurs peuvent mettre sur le marché après les avoir certifiés conformes aux exigences réglementaires[2].

Ces quatre arènes normatives s'opposent et interagissent, dessinant un paysage de plus en plus complexe. Notre objectif premier est d'apporter quelques repères aidant à la navigation dans l'espace social de la régulation et les controverses qui lui sont associées. Il s'agit aussi de montrer comment l'enjeu définitionnel sous-tend les luttes pour contrôler le développement de l'IA.

UNE BOUSSOLE POUR S'ORIENTER DANS L'ESPACE DE LA RÉGULATION

La notion d'« IA » est nébuleuse et contestée depuis ses origines. On doit le terme au logicien américain John McCarthy, proposé à l'occasion de la célèbre conférence qu'il co-organise en 1956 au Dartmouth College, dans le New Hampshire[3]. L'idée de cette conférence est de fédérer une communauté

2. Cet article s'inscrit dans une perspective exploratoire en vue d'analyses ultérieures plus approfondies. Il est le fruit d'une collaboration entre d'un côté un sociologue et, d'un autre côté, deux juristes chercheurs en droit qui, par leurs implications professionnelles, occupent une place privilégiée d'observation des débats sur le contrôle de l'IA. En effet, Yannick Meneceur, magistrat en disponibilité Conseil de l'Europe, a collaboré à un projet de Convention du Conseil de l'Europe en vue d'une réglementation contraignante tournée vers la garantie des droits fondamentaux. Pour une analyse globale de la place du Conseil de l'Europe dans la gouvernance mondiale de l'IA, on peut consulter Meneceur (2020) et Meneceur et Hibbard (2021). Nathalie A. Smuha, docteur en droit international, a été chargée de coordonner le travail du groupe d'experts indépendant de haut niveau sur l'IA de la Commission européenne (AI HLEG). Chacune des quatre définitions de l'IA présentées ici est représentée par un ou plusieurs membres du AI HLEG, ce qui s'est reflété dans leur « lignes directrices éthiques pour une IA digne de confiance » (HLEGAI, 2019). Ce document reste une référence majeure du débat sur la régulation de l'IA. Pour une analyse du travail du AI HLEG (que nous ne traiterons pas dans cet article), on peut consulter Smuha (2019).

3. Si le terme « IA » peut aussi être daté de 1955, au lancement du projet de recherche, on peut toutefois souligner que le domaine de l'IA existait déjà bien avant, y compris les questions relatives à ce que signifie la création d'une machine intelligente. Voir, par exemple, le document du séminaire d'Alan Turing discutant de la question « Une machine peut-elle penser ? », qui a donné naissance au domaine de la philosophie de l'IA, avant même l'invention du terme (Turing, 1950).

de chercheurs autour d'une spécialité nouvelle en rupture avec la « cybernétique », très liée à la figure du mathématicien Norbert Wiener, qui s'interroge sur les possibilités d'une machine pensante. L'histoire est bien connue, ce n'est pas par hasard si McCarthy revendique un domaine distinct (Fleck, 1982 ; Katz 2020). Alors que la « cybernétique » est structurée autour de la notion de *rétroaction*, car elle expliquerait l'autonomie des organismes et des machines, et que le modèle du cerveau serait à trouver dans des réseaux de neurones (le paradigme connexionniste), pour l'IA de McCarthy et d'autres participants de Dartmouth comme Simon et Newell, le modèle du cerveau, c'est l'ordinateur (le paradigme cognitiviste et symbolique). Si bien qu'aujourd'hui le retour du terme IA apparaît paradoxal : « c'est l'agenda intellectuel de Wiener qui domine aujourd'hui sous la bannière de la terminologie de McCarthy » (Jordan, 2018, cité par Cardon *et al.*, 2018). Cette volonté de démarcation n'explique pas les problèmes de sens que pose le syntagme nominal lui-même. Une anecdote est révélatrice des problèmes définitionnels que soulève l'IA : le terme d'*Automata Studies*, utilisé pour le titre d'un ouvrage collectif co-dirigé par Shannon et McCarthy (1956), était une autre option envisageable. Shannon et McCarthy privilégient ce titre à l'époque, car ils restent eux-mêmes très réservés quant à l'usage de la notion d'intelligence qui leur semble présomptueuse (Kline, 2011 ; Rajaraman, 2014). Mais, remarquant que le titre de son livre *Automata Studies* ne suscite guère d'enthousiasme, McCarthy privilégie finalement le syntagme d'IA dont l'adoption sera plus rapide, à la fois par les personnes travaillant dans ce domaine et par le grand public[4].

L'IA COMME TROPE : PRODUCTION DES SAVOIRS ET FABRICATION DES MACHINES

Plus stratégique que scientifique, le syntagme IA apparaît dans une logique de labélisation disciplinaire. Le succès marketing du terme d'IA tient au fait que John McCarthy, sans en avoir eu conscience sans doute, a proposé une dénomination qui repose sur une figure de style. Le syntagme nominal IA peut être en effet considéré comme un *trope*, et plus précisément comme un cas de *syllepse* en rhétorique, à savoir l'emploi d'une occurrence d'un mot dans deux sens différents, indistinctement au sens propre et figuré. Cette

4. Cf. la notice bibliographique du site des lauréats du prix Turing que McCarthy reçut en 1971, https://amturing.acm.org/award_winners/mccarthy_1118322.cfm, consulté le 25 février 2022.

forme rhétorique permet de désigner l'IA d'une part, au sens figuré, comme un domaine scientifique et d'autre part, au sens propre, comme un objet, une formalisation technique. Le trope IA est particulièrement subtil : l'utilisation simultanée du sens propre et du sens figuré indique qu'une relation de détermination réciproque existerait entre la production de connaissance sur l'intelligence et la formalisation technique pour produire l'intelligence elle-même, au point de lier les deux significations en une seule.

Si l'IA est caractérisée à la fois comme une discipline scientifique et comme un objet technique, c'est parce que les spécialistes de l'IA font de l'objet technique qu'il fabrique *de toutes pièces* le médium de leur activité de production de connaissance sur l'intelligence, considérant la formalisation technique non pas comme une simple ingénierie, mais comme « la plus scientifique et la plus productive de toutes les méthodes intellectuelles connues » (Agre, 1997). Les objectifs scientifiques de l'IA paraissent réalisables aux spécialistes parce qu'ils considèrent les machines et les êtres humains « comme des entités physiquement réalisées » (*ibid*.), quelles que soient l'épaisseur sociale, la complexité biologique et la diversité des humains.

L'IA peut ainsi être définie comme l'étude des « entités physiquement réalisées », et non l'étude de la programmation informatique ou des ordinateurs en tant que tels (*ibid.*). En pratique, l'IA renvoie à la production d'artefacts, mais elle ne peut être réduite pour autant à une « technologie » ou une « recherche finalisée », voire une « technoscience ». Encore une fois, pour citer Philip Agre, elle est une science indexée à un système technique :

> Le « résultat » d'un projet de recherche en IA est un système fonctionnel dont les méthodes semblent originales et largement applicables ; une « idée » est une méthode de construction de systèmes techniques ou une façon d'analyser les problèmes qui motive une conception de système prometteuse ; et une « approche » de recherche est un cadre conceptuel et technique par lequel les problèmes peuvent être analysés et transformés en un type particulier de système technique (Chapman, 1991, p. 213-218). En conséquence, l'histoire du domaine se résume d'abord à une succession de systèmes informatiques et ensuite à des débats entre différentes approches de la construction de systèmes[5].

5. « The "result" of an AI research project is a working system whose methods seem original and broadly applicable; an "idea" is a method of building technical systems or a way of analyzing problems that motivates a promising system design; and a research "approach" is

Quand Philip Agre souligne l'oscillation permanente entre l'action de fabrication et la production des connaissances, il veut montrer les tensions qui relient sans les unir le *faire* au *savoir* (Laumond, 2012). Si les scientifiques qui *font de l'IA* et *font l'IA* sont devenus des virtuoses dans l'art de respecter à la fois les fétiches et les faits (Latour, 2009), de parler dans les mêmes termes de l'intelligence et des machines et de refuser l'asymétrie du grand partage entre le monde objectif des sciences et le monde social et politique qui lui serait extérieur, ceux qui s'engagent dans la régulation de l'IA ont du mal à s'accommoder de cet amalgame entre le signifiant et le signifié ; de cet entremêlement de méthode et théorie ; et de ce brouillage entre recherche appliquée et recherche fondamentale.

QUATRE ARÈNES DIFFÉRENCIÉES DE PRODUCTION NORMATIVE

Peut-on réguler un phénomène dont la qualification repose sur une figure à double sens ? Comment les acteurs se saisissent-ils de cette oscillation continue entre la production de la recherche et les systèmes techniques ? C'est ce que l'on cherche d'abord à comprendre lorsqu'on analyse les politiques de régulation de l'IA.

Pour ce faire, nous nous appuyons sur un schéma très simple qui a l'avantage de donner une vue synthétique de l'espace de la régulation de l'IA selon deux dimensions principales.

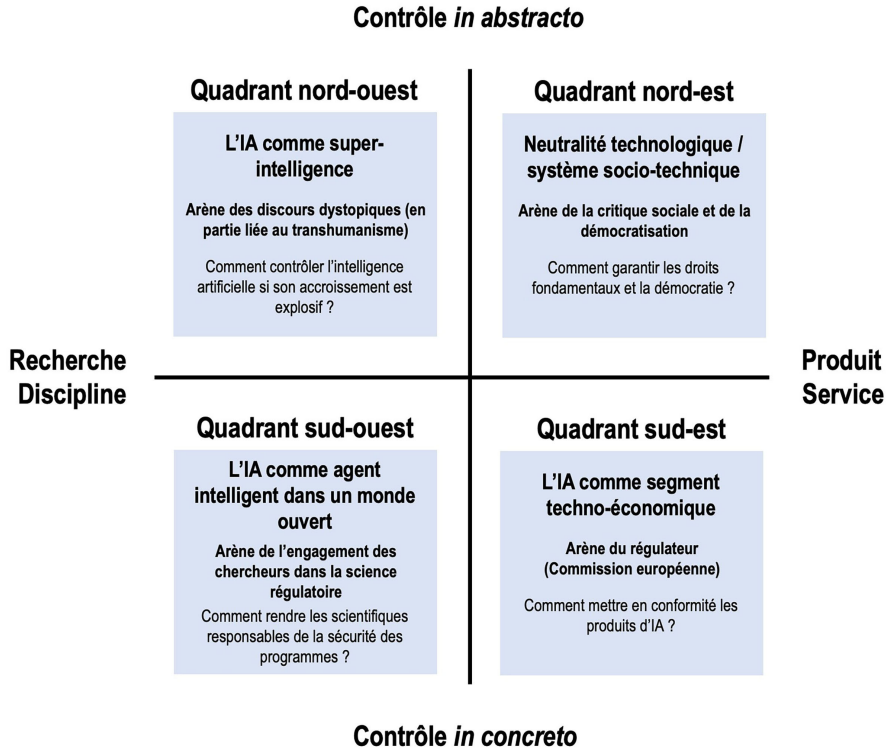
Sur un axe horizontal, on représente la double fonction référentielle de l'IA en opposant la discipline scientifique aux produits et services qui en sont dérivés : autrement dit, d'un côté, l'étude de l'intelligence par l'utilisation des concepts des sciences cognitives, de la biologie, des neurosciences voire des sciences sociales (selon les approches adoptées) et, d'un autre côté, les machines concrètes qui accomplissent des choses qui demanderaient de l'intelligence si elles étaient accomplies par des humains. Ainsi, à l'ouest, on trouve tout ce qui caractérise l'IA comme un ensemble de savoirs produits

---

a conceptual and technical framework by which problems can be analyzed and transformed into a particular type of technical system (Chapman, 1991, p. 213-218). The field, accordingly, reckons its history primarily as a sequence of computer systems and secondarily as a history of debates among different approaches to the construction of systems. » Cité depuis la version publiée sur le blog : https://pages.gseis.ucla.edu/faculty/agre/critical.html, consulté le 19 avril 2022) (traduction auteurs).

par une communauté d'acteurs qui se revendiquent comme des spécialistes scientifiques d'un domaine. Une bonne partie de ce qui est proposé par ces spécialistes se traduit le plus souvent en un système technique, d'abord testé sur un domaine particulier, puis généralisé. Ces productions de système, foisonnantes, peuvent se détacher de la recherche. Elles pointent alors vers la dimension qui figure à l'autre extrémité de cette première polarité, à l'est. Ce sont ces systèmes techniques qui servent à faire des promesses, démontrer l'utilité de la discipline et susciter l'intérêt des investisseurs, pour qui les systèmes d'IA sont des produits ou services à caractère commercial.

Figure 1. L'espace social de la régulation selon 4 arènes normatives différentes



Source : auteurs.

À cet axe définitionnel, on croise une deuxième dimension – l'axe du contrôle. Pour s'orienter dans la multitude de discours sur le contrôle de l'IA, on peut simplement commencer par les diviser en deux catégories : ceux

qui envisagent le contrôle de l'IA comme un projet à réaliser concrètement et ceux qui font plutôt du contrôle un sujet de spéculation abstrait. Les premiers contribuent à écrire des normes techniques et juridiques ou à constituer de manière systématique une science régulatoire – ils visent un contrôle *in concreto*. Les seconds, qu'ils privilégient une futurologie dystopique ou un sens critique émancipatoire du possible (Guéguen et Jeanpierre, 2022), se réfèrent toujours à un modèle social et politique abstrait ; pour cette raison, nous considérons leur production normative comme un contrôle *in abstracto*. Alors que le contrôle *in concreto* est soucieux des spécificités des approches et repose sur l'analyse du réel au cas par cas des situations dans lesquelles opèrent les systèmes, le contrôle *in abstracto* est d'une portée générale, que ce soit en matière de « risques existentiels », d'éthique ou de droit des personnes. Ces discours généraux ne sont pas, pour autant, sans apporter leur participation à l'élaboration des productions normatives concrètes du contrôle de l'IA. Non seulement ils ont le pouvoir de cadrer la manière de construire le problème du contrôle de l'IA, mais ils exercent une action incitative sur les régulateurs et, davantage encore, ils alimentent de leur substance les textes normatifs qui ont vocation à contrôler concrètement l'IA. Le contrôle concret est, en grande partie, du contrôle abstrait rendu familier par l'usage, pour paraphraser Paul Langevin.

Ces deux axes permettent ainsi de projeter quatre arènes aux productions normatives différenciées. Les quadrants du nord-ouest et sud-ouest renvoient à deux types d'interrogation sur le contrôle de l'IA comme recherche : d'une part, un questionnement prophétique, se demandant comment anticiper le moment où viendra la fabrication de machines dont le pouvoir pourrait échapper au contrôle des humains (nord-ouest). Le contrôle est abstrait, car il repose sur une évaluation de l'IA au regard d'une société qui vise comme idéal politique le dépassement des limitations des êtres humains. À l'opposé, un deuxième type d'interrogation, qui présente peu d'intérêt pour ces considérations prospectives, préfère s'interroger sur l'auto-responsabilisation des chercheurs sur les dangers concrets des machines qu'ils fabriquent d'ores et déjà (sud-ouest).

Les quadrants nord-est et sud-est donnent à voir deux manières différentes d'envisager le contrôle de l'IA comme produit ou service. Au nord-est, les objets techniques (les machines donc) apparaissent comme l'occasion de débattre à nouveau frais de questions fondamentales, à la fois philosophique et anthropologique, sur les implications éthiques de la délégation du pouvoir aux machines. Au sud-est, d'une manière plus concrète, on compte davantage

sur une *regulatory science* (Jasanoff, 1990) qui découlerait de l'opérationnalisation des recherches du sud-ouest au sein d'agences d'évaluation dédiées aux audits des objets et services de l'IA. Dans ce quadrant, les systèmes d'IA révèlent d'abord leur potentiel de croissance économique, d'où une définition de l'IA comme « segment techno-économique ».

Bien qu'interconnectées les unes aux autres, ces quatre arènes différenciées nous servent à mieux souligner les enjeux politiques de la régulation. Elles dessinent le *fond de carte* des conflits normatifs, nous permettant ainsi d'y projeter les lieux, les acteurs et les approches de la régulation.

LA NÉBULEUSE TRANSHUMANISTE : LE PROBLÈME DE LA PERTE DU CONTRÔLE

En haut à gauche de notre schéma, l'IA est envisagée sous l'angle spéculatif de l'avènement d'une *intelligence artificielle générale* dépassant les capacités cognitives de l'intelligence humaine. Le problème du contrôle s'exprime alors dans le cadre de l'hypothèse de la singularité technologique : les progrès de l'intelligence artificielle pourraient déclencher prochainement un emballement de la croissance technologique qui induirait des changements imprévisibles dans la société humaine, en particulier l'extinction de l'espèce humaine en raison de perte de contrôle des machines. C'est la thèse populaire défendue par Nick Böstrom dans son livre paru en 2014, *Superintelligence*. Cette projection, souvent répandue dans les récits de science-fiction, est relayée par une poignée d'individus très influents, notamment de puissants entrepreneurs comme Elon Musk (le célèbre patron de la société Tesla) ou Sam Altman (le fondateur de Y Combinator, entreprise de financement de start-up). Leur alerte se fonde sur le présupposé que la singularité technologique est certes un événement peu probable, mais que son impact sur les humains serait majeur. Il est donc, selon eux, justifié de s'en inquiéter et de se préparer suffisamment tôt à cette éventualité, en anticipant les risques qu'une super-intelligence pourrait générer.

Cette représentation potentiellement catastrophiste des effets des progrès de l'IA n'est pas nouvelle. Elle est un des mythes fondateurs de la discipline scientifique. Portée par les chercheurs eux-mêmes, notamment les figures historiques du domaine, cette vision dystopique fonctionne comme un leitmotiv surgissant systématiquement lors des périodes fastes de l'IA. On en trouve des traces dès les années 1950 dans les débats autour de la première

cybernétique de Nobert Wiener (1950). On la voit ressurgir au milieu des années 1960 avec la célèbre spéculation de Jack Good (1966) sur l'explosion de l'intelligence. Dans les années 1980, Hans Moravec (1988) annonce une transformation majeure de l'humanité par la robotique, notamment l'obsolescence de l'homme. Plus récemment, Stuart Russell (2019) poursuit cette tradition spéculative sur les risques de la perte de contrôle des machines.

Désormais, l'hypothétique contrôle de l'humanité par une super-intelligence dépasse les prises de position personnelles de spécialistes éclairés. Elles se constituent progressivement comme un problème public. Depuis une vingtaine d'années, on observe la mobilisation progressive d'acteurs, situés pour l'essentiel aux États-Unis et en Angleterre, alignés souvent tant sur des perspectives politiques communes que des idéologies transhumanistes. Ces acteurs sont également organisés autour d'un credo : un avenir radicalement transformé par l'intelligence artificielle générale et d'autres technologies avancées pourrait être une bonne chose pour l'humanité si l'on en garde le contrôle. Ce groupe forme une nébuleuse d'acteurs, représenté seulement par une poignée de personnes – une centaine à travers le monde – dispersées dans une dizaine d'organisations[6].

Il s'agit, pour l'essentiel, du *Future of Humanity Institute* (FHI, 2005, Oxford Martin School), du *Centre for the Study of Existential Risk* (CSER, 2012, University of Cambridge), du *Future of Life Institute* (FLI, 2014), et du *Berkeley Existential Risk Initiative* (BERI, 2017, U.C. Berkeley). Même si les mandats de ces quatre organisations sont larges, en ambitionnant de traiter l'ensemble de risques existentiels comme les risques posés par le changement climatique, les armes nucléaires et la menace de pandémies, l'IA reste leur principal centre d'intérêt. Montrant comment les technologies de l'IA pourraient avoir des impacts indésirables sur l'humanité, elles consacrent l'essentiel de leur activité à un plaidoyer sur les risques existentiels posés par l'IA, en posant le problème éthique de l'IA comme celui de la co-adaptation de l'humanité avec la super-intelligence des machines. S'adressant au grand public et aux décideurs, ces organisations essaient de contribuer à l'agenda politique de la régulation de l'IA. C'est dans cet objectif que le FLI publie une

6. Toutes ces organisations forment un réseau d'acteurs qui travaillent en étroite collaboration : leurs dirigeants, employés, conseillers et bailleurs de fonds se chevauchent, participant aux mêmes événements et circulant entre les mêmes instituts. Malgré leur petite taille (une douzaine d'employés en moyenne), ces organisations ont su recruter des alliés influents et cadrer une grande partie du débat public sur la nature et les perspectives de l'intelligence artificielle.

série de lettres ouvertes, dont une particulièrement célèbre en 2015, *Research Priorities for Robust Beneficial Artificial Intelligence*[7], puis annonce en 2017 les « 23 principes éthiques d'Asilomar » dégagés d'un atelier organisé lors de la conférence *Beneficial AI*. Et c'est dans la continuité de ces alertes que le Parlement européen inscrit sa proposition de rapport sur la robotique en 2016, notamment en évoquant la possibilité de la perte de contrôle de l'IA (Rapport Delvaux, 2017).

Trois unités de recherche, dont deux attachées à des universités prestigieuses, ont été mises sur pied pour répondre spécifiquement à cette question de la perte de contrôle : le *Machine Intelligence Research Institute* (MIRI) à Berkeley, créé en 2000 sous le nom de *Singularity Institute for Artificial Intelligence* par Eliezer Yudkowsky, avec le soutien financier du célèbre entrepreneur de PayPal, Peter Thiel ; le *Leverhulme Centre for the Future of Intelligence* (LCFI) cofondé en 2015 par les universités d'Oxford, de Cambridge, l'Imperial College et l'Université de Berkeley grâce au soutien financier de la fondation Leverhulme ; et le *Center for Human-Compatible Artificial Intelligence* (CHCAI) à l'U.C. Berkeley, dirigé depuis 2016 par Stuart Russell.

Si le soubassement mathématique des recherches de ces laboratoires repose pour l'essentiel sur la théorie générale de l'intelligence de Marcus Hutter (2004), leurs travaux sur la mise en éthique des machines s'inscrit dans une perspective similaire à celle du domaine dit *machine ethics* autour de la notion d'*agents moraux artificiels* (Anderson et Anderson, 2011), mais à la différence près qu'il ne suffit pas de définir une fonction d'objectif à une IA super-intelligente pour en garantir un comportement compatible avec les attentes des humains. Une machine pourrait chercher par tous les moyens possibles à réaliser le but qui lui a été fixé, au risque de choisir des manières de les atteindre qui soulèvent des problèmes de sûreté imprévisibles. Le problème du contrôle de l'IA devient celui du contrôle de *la réalisation du but*, c'est-à-dire l'apprentissage par les machines des « structures de préférence » des humains à partir de l'observation de leurs actions. Pour ce faire, Russell propose, par exemple, la piste de recherche de l'apprentissage par renforcement inverse (Russell, 2019) des buts que les humains cherchent à réaliser,

7. La lettre renvoie les lecteurs à une note programmatique du même titre, co-rédigée par Russell, Dewey et Tegmark (2015).

à partir de l'observation de leurs comportements[8]. Depuis la fin des années 2010, Open AI et DeepMind – considérés souvent comme les leaders contemporains de la recherche en IA – prennent au sérieux le problème de la perte de contrôle en proposant des expériences simulées. La plus connue d'entre elles est celle du problème de l'arrêt (appelé aussi *corrigibility*, un cas spécial de l'apprentissage de la structure des préférences) qui vise à modéliser des situations où une IA super-intelligente apprendrait à ne pas empêcher son interruption (Orseau et Armstrong, 2016).

L'INDUSTRIE ET SES EXPERTS : L'AUTORÉGULATION DE ET PAR LA SCIENCE

Mais cette forme de contrôle abstraite trouve peu d'écho auprès de la majorité des experts en IA (Ganascia, 2017). C'est d'ailleurs en réaction critique aux discours dystopiques qu'on trouve l'un des premiers appels à la régulation qui, bien que tombé dans les oubliettes de l'histoire, reste selon nous un lieu révélateur des tensions normatives autour des politiques du contrôle de l'IA : l'*Asilomar Meeting on Long-Term AI* de 2009 (Horvitz et Selman, 2012), organisée par l'*Association for the Advancement of Artificial Intelligence* (AAAI). En se revendiquant de l'héritage de la célèbre conférence d'Asilomar de 1975[9] autour de la mise en place d'un moratoire sur les manipulations génétiques, les organisateurs mobilisent un répertoire d'action plus classique : l'engagement public des scientifiques pour une recherche responsable (Hurlbut, 2015).

Mais, contrairement à la conférence de 1975, le rassemblement de 2009 n'a pas été organisé en réaction à un danger technologique imminent, mais à l'émergence d'une « perception d'urgence par les non-experts ». L'objectif

8. Ce problème du contrôle de l'IA émerge au début des années 2000 sous des notions variées comme celle de *friendly AI*, d'*Artificial General Intelligence safety* (AGI safety) ou de *value alignement*. La question reste toujours la même : comment faire en sorte que les actions des machines soient mieux alignées avec les objectifs et les préférences des humains ?

9. Dans une étude sur les usages de la mémoire de la conférence d'Asilomar, Benjamin Hulbrut (2015) montre comment Asilomar cristallise un imaginaire de « l'émergence gouvernable » – dans lequel non seulement la science est imaginée comme un moteur de changement, mais le droit est présenté comme toujours à la traîne et donc réactif et potentiellement inhibiteur du progrès scientifique (c'est-à-dire le « décalage du droit »). Asilomar-en-mémoire, comme l'appelle l'historien des sciences, perpétue cet imaginaire en l'ancrant dans un précédent historique.

est de contenir le risque d'une imagination publique débridée qui pourrait empêcher l'émergence de technologies telles qu'imaginées par les scientifiques. Comme lors de la réunion de 1975, une « réflexion proactive » a pour objectif de garantir les meilleurs résultats pour la recherche, permettant à la société d'en tirer le maximum de bénéfices.

En 2009, la quasi-totalité des chercheurs réunis à la conférence d'Asilomar est sceptique face à l'hypothèse de la singularité qui domine le débat médiatique. Le risque est plutôt défini par rapport aux diverses potentialités d'usages des machines que les scientifiques sont d'ores et déjà capables de fabriquer dans leur laboratoire. C'est pourquoi la conférence d'Asilomar est avant tout à envisager comme le moyen de contrer le discours de la singularité en appelant à se concentrer intensivement sur les risques avérés à court terme par des systèmes d'IA. Pour comprendre ce positionnement, il faut rappeler que l'initiative d'Asilomar n'apparaît pas au hasard de l'espace scientifique. Elle est portée par un acteur central de la discipline : Eric Horvitz, président de l'AAAI et directeur de recherche (l'un des plus influents) chez Microsoft[10]. À l'occasion de sa prise de fonction de président de l'AAAI en 2008, Eric Horvitz place les enjeux de la régulation de l'IA comme une priorité pour sa discipline, tout en écartant les scénarios dystopiques.

L'enjeu de régulation émerge, selon Eric Horvitz, dans un contexte d'accélération d'une ouverture des machines au monde réel, un déploiement de l'IA « dans un monde ouvert ». Cette ouverture au monde marque une rupture avec les recherches passées en IA où les agents opèrent dans un monde clos où toutes les conditions sont mentionnées à la machine. Placées en dehors du monde fermé des laboratoires, les machines de l'IA, par-delà les différentes approches du domaine, sont mises au service de problèmes concrets, traitant des flux réalistes de problèmes et devant agir selon des possibilités non recensées initialement. Ainsi s'imposent aux chercheurs non seulement des axes de recherches spécifiques pour immerger des agents intelligents dans un univers dynamique complexe, mais aussi des responsabilités sociales nouvelles.

---

10. Notons que cette conférence est co-animée par Bart Selman, professeur d'intelligence artificielle à Cornell après une carrière dans le prestigieux *AT&T Bell Laboratories* et qu'Eric Horvitz sera l'un des promoteurs d'une régulation stricte de la reconnaissance faciale, cf. https://www.lemonde.fr/pixels/article/2018/07/03/eric-horvitz-microsoft-ne-veut-pas-fournir-d-outils-qui-pourraient-violer-les-droits-de-l-homme_5324975_4408996.html, consulté le 19 avril 2022.

Comment orienter les programmes de recherche pour faire de l'IA un domaine de recherche responsable ? La dimension de la recherche en IA qui est mobilisée ici est celle de l'*expertise* pour faire entendre le point de vue des scientifiques sur les problèmes que posent les machines qu'ils fabriquent. Mais comment les chercheurs peuvent-ils contribuer à la production de « machines certifiées » dont la fiabilité est évaluée par la communauté des spécialistes eux-mêmes ? Pour répondre à cette question, Eric Hortvitz et le groupe d'experts qu'il a réuni proposent de définir une politique de recherche afin de donner les moyens à la discipline de *contrôler,* en sorties de laboratoire, les machines autonomes qui agiront dans un monde ouvert.

Plus concrètement, la proposition des chercheurs réunis à Asilomar repose sur l'identification de catégories de risque différenciées : que ce soient les bugs des systèmes d'IA, leur comportement dans les situations imprévues, les cyberattaques, les réponses littérales des machines aux instructions dangereuses (appelé le risque de l'apprenti sorcier), la compréhension détaillée des situations par les humains lorsqu'ils prennent le relais des machines (le passage de contrôle du système d'IA vers l'humain comme source d'accident), il s'agit systématiquement de proposer des axes de recherche comme des solutions visant à augmenter l'autonomie de la machine jusque dans la prise en compte de ces risques (Dietterich et Horvitz, 2015).

Globalement, les chercheurs qui s'inscrivent dans cette arène ramènent la responsabilité des scientifiques à des enjeux de recherche tournés vers la sécurité, dans la tradition de la vérification des systèmes informatiques. Le défi qui se présente à la recherche se décline en une série de questions qui réduisent l'IA à sa part technique, et donc à la fabrication de machines dites « robustes » : comment construire des architectures d'*autosurveillance* dans lesquelles un processus de méta-niveau observe en permanence les actions du système, vérifie que son comportement est conforme aux intentions fondamentales du concepteur et intervient ou émet des alertes si des problèmes sont identifiés ? Comment transférer la recherche sur la vérification et la surveillance en temps réel des systèmes logiciels vers le fonctionnement des systèmes autonomes dans un monde ouvert ? Comment intégrer des couches de programme supplémentaires dans les machines pour détecter des comportements internes anormaux qui peuvent révéler des cyberattaques ? Comment les machines peuvent-elles anticiper le moment où le contrôle humain sera nécessaire, en fournissant aux personnes les informations essentielles dont elles ont besoin pour prendre le relais (Horvitz *et al.*, 2021) ? Il s'agit donc de proposer des outils technologiques pour surveiller et prévenir les défaillances

d'autres outils technologiques. Notons que ces questions techniques ne sont pas très éloignées des problèmes de perte de contrôle posés dans le cadre de l'hypothèse de la singularité[11], à la différence près que la problématisation du contrôle est tournée vers une réalité tangible.

Dans cette même arène, un autre thème, qui n'est pas à l'ordre du jour de la conférence d'Asilomar de 2009, a émergé autour d'une définition spécifique de l'IA, en lien avec la montée de la science des données et du *machine learning* : la *fairness* (dans le sens de la justice sociale et de l'équité) et l'explicabilité qui apparaissent depuis une dizaine d'années comme des enjeux majeurs du contrôle de l'IA. Mises à l'agenda de la recherche à la suite d'une série de critiques sociales dénonçant, d'une part, les biais racistes et sexistes dans les machines prédictives (nous y viendrons dans la prochaine arène) et, d'autre part, le manque de transparence des décisions algorithmiques, les valeurs de *fairness* et explicabilité ont très rapidement été traduites comme des problèmes techniques. Respectivement, la recherche sur la *fairness* s'inscrit dans la lignée des recherches en cryptographie sur les systèmes informatiques préservant l'anonymat (Dwork *et al.*, 2012). Ainsi, les valeurs de *privacy* (intégrées depuis longtemps en sécurité informatique) et de *fairness* apparaissant comme des problèmes similaires d'analyse des propriétés structurelles de ces valeurs au regard de l'utilité ; quant à la notion d'explicabilité, elle est devenue un enjeu central de la mathématisation du *deep learning* (Mallat, 2016).

C'est donc dans cette perspective de gestion des risques que sont venues s'ajouter à l'agenda des scientifiques les questions de justice sociale et d'équité : comment contraindre les procédures d'apprentissage statistique de sorte que les mécanismes sociaux (racistes, sexistes, etc.) de production des données soient atténués ? Comment automatiser des notions de justice sociale entre les groupes sociaux ou les individus dans les procédures d'apprentissage machine (Kearns et Roth, 2019) ? Comment peut-on évaluer la qualité d'une explication de modèles dont la causalité n'est pas représentée ? Les laboratoires de recherche et développement des grandes entreprises du numérique participent à « technologiser » ce problème en produisant des outils dédiés comme, par exemple : le « AI Fairness 360 Open Source Toolkit » mis à disposition par IBM comme une boîte à outils *open source* pour aider à étudier, détecter et minimiser les discriminations et les biais dans les modèles d'apprentissage machine tout au long du cycle de vie des applications d'IA ; et

11. Les deux arènes tendent parfois à se confondre sous le label d'*AI Safety.*

l'« explainable AI » de Google qui vise à tracer le raisonnement opéré par une machine et ainsi en comprendre les motifs.

Il s'agit d'automatiser tout un ensemble de valeurs, que ce soit la sûreté ou l'équité dans un contexte où il sera probablement question dans des législations futures non seulement de certifier, avant la mise en service, le bon fonctionnement des algorithmes, mais également de prévenir, tout au long du cycle de vie, les dérives pouvant survenir de l'apprentissage toujours évolutif. Dit autrement, pour prévenir les risques associés à l'IA, les scientifiques s'inspirent tout simplement des principes de sécurité informatique, notamment ceux de la vérification des systèmes, mais ils les adaptent aux problèmes sociaux spécifiques que posent les procédures d'apprentissage des machines. Cette arène normative vise une sorte de *regulatory science* comme la meilleure façon de pousser la recherche en IA jusqu'à ses limites, sans mettre en danger la population. Sous l'apparence d'une autorégulation responsable, la science intervient elle-même pour déterminer les formes de gouvernance que la société est autorisée à prendre en considération.

Pour maintenir une communication régulière autour de cette autorégulation responsable, Eric Horvitz établit en 2014 le programme *One Hundred Year Study on AI* à l'Université de Stanford (Horvitz, 2014), un observatoire programmé sur un siècle, dédié à l'analyse des évolutions de l'IA et aux définitions techniques des risques qui leur sont associés. Horvitz cofonde ensuite en 2016 un institut, le *Partnership on AI To Benefit People and Society*, qui regroupe les principaux acteurs économiques de l'IA. L'objectif est de créer un espace de discussion autour des dangers de l'IA et des solutions techniques à y apporter, afin de montrer au grand public et aux acteurs politiques une position commune. Les questions de la régulation de l'IA sont donc limitées aux risques que les acteurs économiques et leurs experts connaissent le mieux, ce qui suppose de s'en remettre à leur façon de comprendre ce qui se joue. C'est pourquoi ces différentes démarches ont été perçues comme une forme de techno-solutionnisme qui divertirait d'une analyse et de réflexions plus profondes sur ce que transforment profondément dans notre société la fabrication et le déploiement de l'IA.

CRITIQUE SOCIALE ET CONTRÔLE DÉMOCRATIQUE

Dans les arènes précédentes, si nous avons observé une opposition sur les perspectives temporelles et les approches, le problème de la régulation de l'IA

reste posé en terme similaire : celui du risque. Dans l'arène que nous allons maintenant analyser, au nord-est de la carte, l'IA est ramenée à son statut matériel de produit ou de service concret, mais le problème de son contrôle est considéré sous un angle abstrait ou *holistique* pourrait-on dire : celui de l'émancipation des humains de la puissance des machines et des formes de domination qui leur sont associées.

Cette manière de concevoir le contrôle de l'IA est d'abord envisagée par les informaticiens eux-mêmes qui, par leur connaissance concrète et intime des machines, sont souvent les mieux placés pour en révéler leurs limites et les problèmes sociaux qu'elles soulèvent. On retiendra ici deux figures intellectuelles qui, bien que situées à deux époques très différentes, sont typiques de cette manière d'envisager la régulation de l'IA : l'informaticien Joseph Weizenbaum, chercheur en IA au MIT dans les années 1960, qui est souvent convoqué dans le débat contemporain sur l'éthique de l'IA (Loeb, 2021), et les chercheuses en IA, Joy Buolamwini et Timnit Gebru dans les années 2010 qui sont très influentes dans les discussions récentes sur la régulation de l'IA. En 1976, Joseph Weizenbaum, le célèbre concepteur de l'un des premiers agents conversationnels dans les années 1970 (Eliza), dénonce son propre domaine de recherche en s'élevant contre « l'impérialisme de la raison instrumentale qui vient imposer à l'homme l'univers de la machine » ; en déplorant « l'atrophie de l'esprit humain qui fait confiance à la seule science pour interpréter la réalité » ; et en estimant que les sciences de l'informatique, « au lieu de révolutionner le monde, ont été utilisées pour soutenir les institutions sociales et politiques américaines » (Weizenbaum, 1976, cité par Brigitte Chamak, 2004, p. 84). Près de 45 ans plus tard, en 2017, dans l'étude *Gender Shades* qui deviendra rapidement célèbre, Buolamwini et Gebru (2018) montrent que les systèmes d'IA pourraient présenter des biais systématiques en fonction du genre et de la race, que ce soit la vision par ordinateur, l'analyse textuelle ou les systèmes algorithmiques d'aide à la décision en santé, justice, ressources humaines, etc. Les valeurs sexistes et racistes se logent dans les machines pas seulement parce que les données et les modèles algorithmiques reflètent les inégalités de traitement, mais surtout parce que les technologies sont fabriquées par des « hommes de race blanche ». Selon ces deux types de critiques sociales de l'IA qui coexistent dans le débat contemporain (bien que la critique des biais racistes et sexistes occupe l'essentiel du débat, car elle est plus adaptée aux produits et services qui reposent sur le *machine learning*), les concepteurs d'IA sont impliqués dans des choix moraux concernant des normes et des relations de pouvoir, même si ces choix ne reflètent pas nécessairement une intention consciente de leur part.

En réponse à ces critiques sociales, les institutions les plus impliquées dans le développement de l'IA ont rapidement eu un réflexe éthique. Aux États-Unis, c'est Joichi Ito, ancien directeur du Media Lab du MIT, qui a été l'une des figures de proue de la *mise en éthique* de l'IA. Considéré comme l'expert majeur en matière d'éthique de l'IA, il obtient la direction de l'*Ethics and Governance of AI Fund*, une initiative conjointe du Media Lab du MIT et du Berkman Klein Center for internet and Society de l'Université de Harvard, succédant ainsi à l'ancien « responsable de la politique publique mondiale » pour l'intelligence artificielle de chez Google, Tim Hwang. Grâce à ce fond, de nombreuses initiatives sur différents continents sont financées, par exemple la création d'une importante conférence relative à « l'équité, la responsabilité et la transparence » (Fairness, Accountability, and Transparency dite FAT ML) dans le domaine de l'intelligence artificielle avec, parmi les sponsors de la conférence, Google, Facebook et Microsoft. Ces fonds ont aussi permis la création d'instituts à la tête desquels sont placés des chercheurs, notamment en sciences sociales, travaillant au sein de grandes sociétés technologiques : par exemple, le *Data & Society Research Institute* est dirigé par une chercheuse de Microsoft, Danah Boyd, et financé lors de sa création par une subvention de cette même entreprise ; l'*AI Now Institute* de l'Université de New York a été cofondé par une autre chercheuse de Microsoft, Kate Crawford, et Meredith Whittaker, fondatrice de l'*Open Research Group* de Google, tout en étant financé en partie par Microsoft, Google et DeepMind ; le *Stanford Institute for Human-Centered AI* est codirigé la célèbre Fei-Fei Li, connue pour avoir dirigé le SAIL (Stanford AI Laboratory) entre 2013 et 2018, mais aussi ancienne vice-présidente de Google Cloud en 2017 et 2018. Bien qu'ils soient financés par des acteurs économiques, ces instituts ne fonctionnent pas comme des organisations de lobbying d'industriels. C'est même le contraire qui a pu se produire sur certains cas : par exemple, Kate Crawford et Meredith Whittaker ont pu exercer une influence directe sur les législateurs pour contraindre légalement le déploiement de certains produits d'IA (la reconnaissance faciale, notamment). Ces instituts restent néanmoins des instruments potentiels pour les acteurs industriels qui les financent – au moins de façon indirecte.

Que ce soit aux États-Unis ou en Europe, les nombreuses études que des associations d'industries produisent sont aussi des ressources pour les *policy makers* ou les organismes de standardisation (type ISO) pour légitimer l'IA, comme la formalisation des études d'impact et d'audit des systèmes algorithmiques dont la visée d'usage reste principalement les développeurs, les managers et l'assurance qualité (Ayling et Chapman, 2021 ; Koene, 2022). Ainsi, même si une partie de la recherche de ces instituts est porteuse d'une critique

sociale de la technique, surtout dans le domaine des sciences humaines et sociales, beaucoup de travaux financés au titre d'une « IA éthique » s'alignent en réalité sur l'agenda de l'industrie numérique qui consiste à s'autoréguler par le financement de travaux académiques sur les questions d'éthique et de société. De fait, la position de ces acteurs oscille entre un contrôle abstrait qui appelle à un rééquilibrage des pouvoirs et un contrôle concret par des méthodes d'évaluation d'impact des technologies. C'est pourquoi on observe au sein de cet espace une tension permanente avec le quadrant sud-est : au sein de la conférence FAT ML, par exemple, l'opposition est récurrente entre ceux qui proposent des solutions d'éthique technologique (méthodes de *Fair ML* et d'explicabilité) et ceux qui montrent que l'IA soulève des questions profondément politiques sur la façon dont le pouvoir est exercé par les technologies (Selbst *et al.,* 2019).

Dans un contexte où la critique sociale est facilement instrumentalisée par les grandes entreprises du numérique, le droit apparaît comme une ressource démocratique, voire l'ultime rempart contre l'hégémonie du capitalisme de surveillance (Zuboff, 2019). Une analyse de la construction d'un droit de l'IA, aux contours encore bien incertains (Hildebrandt, 2022), dépasse la portée de cet article. Il peut toutefois être identifié un groupe d'acteurs émergeant dans cette arène du nord-est en mobilisant le droit comme une forme de contrôle abstrait de l'IA. L'un des acteurs les plus engagés dans cette direction est probablement le juriste Paul Nemitz, conseiller principal sur la politique de justice à la Commission européenne, responsable, en tant que directeur, de la mise en place du RGPD sur la protection des données en 2018 et, aujourd'hui, impliqué dans le développement de la réglementation européenne relative à l'IA. Nemitz voit dans la manière dont se développe l'IA une menace majeure pour la survie de la démocratie (Nemitz, 2018). En 2021, il cofonde « *The Transatlantic Reflection Group in defence of democracy and the rule of law in the age of "artificial intelligence"* », dont l'objectif est de réhabiliter la position de surplomb de la loi comme instance susceptible de limiter et de contrarier les agents économiques producteurs d'IA (Nemitz, 2021). Le manifeste du groupe est un appel à une restriction des rapports de domination des acteurs économiques par la soumission des producteurs d'IA au règne de la loi générale et abstraite. Car, en définitive, c'est le primat de l'intelligence des humains sur celle des IA qui se justifie au regard de la « fonction anthropologique du droit » (Supiot, 2005) :

> Le langage facile du lobby technologique, selon lequel la loi doit être mise à jour aussi rapidement que le code, ignore trois différences essentielles entre le

> code et la loi : premièrement, la loi n'est pas écrite pour des idiots, comme le code. Elle est écrite pour les humains. Et les humains ne sont pas des idiots, comme les ordinateurs. Ils peuvent penser par eux-mêmes. Deuxièmement, la loi est écrite dans un langage ouvert aux humains. Ces deux facteurs, à savoir que les destinataires de la loi peuvent penser par eux-mêmes et que la loi est écrite en langage humain, et non en code mathématique, permettent à la loi de se maintenir lorsqu'un changement de contexte se produit ou qu'un dysfonctionnement est révélé, alors que le code doit être constamment modifié par une mise à jour. Car le langage humain ouvert peut être réinterprété par des êtres humains pensants, pour tenir compte des nouvelles technologies ou des nouveaux modèles économiques. Et troisièmement, la loi est élaborée dans le cadre d'un processus démocratique qui, par nature, exige des délibérations et des compromis. Aucune loi n'est ou ne peut être parfaite. Et nous ne pouvons pas vouloir une loi parfaite, ni dans sa formulation ni dans son application. Car c'est cela le fascisme[12]. (Nemitz, 2021 p. 3)

Ainsi, le manifeste déclare qu'une intervention juridique de droit dur et contraignant est nécessaire en matière d'IA, quand bien même toutes les implications de l'IA ne peuvent être anticipées et pourraient être mal évaluées par des lois imparfaites. Mais comment légiférer les produits d'une recherche scientifique en mutation permanente, dont les évolutions sont imprévisibles ? Les spécialistes des rapports entre droit et technologie ont développé de nombreux outils pour concevoir des réglementations à la fois fluides et flexibles, permettant à une réglementation de s'adapter aux technologies émergentes ou à celles qui se transforment en permanence. La notion de « neutralité technologique » d'un instrument juridique, la plus connue, mais aussi la plus controversée (Gautrais, 2012 ; Reed, 2007), est invoquée par Nemitz et par de nombreux acteurs dans le débat sur la réglementation européenne de l'IA. Cette notion juridique, qui correspond davantage à l'esprit du RGPD, apparaît comme une manière de ne pas tomber dans le piège d'une définition de l'IA qui pourrait très vite tomber en désuétude.

---

12. « The facile language of the tech lobby that the law has to be updated as quickly as the code ignores three key differences between code and law: first, the law is not written for idiots, like the code. It is written for humans. And humans are not idiots, like computers. They can think for themselves. Second, law is written in human open language. The two factors, namely that the addressees of the law can think for themselves and that the law is written in human language, not mathematical code, allows the law to stand when change of context occurs or malfunctioning is revealed, while the code needs constant change through updating. Because open human language can be reinterpreted by thinking human beings, to take account of new technologies or business models. And third, the law is made in a democratic process, which by its nature requires deliberation and compromise. No law is or can be perfect. And we cannot want perfect law, neither in its formulation nor in its enforcement. Because that is fascism. » (traduction auteurs)

D'autres universitaires comme Mireille Hildebrandt et Karen Yeung ont également contribué à documenter l'importance de l'intervention de règles juridiques à la place ou en complément des initiatives éthiques, mais aussi pour combler l'offre de protection limitée des droits et libertés des instruments juridiques existants (Hildebrandt, 2022 ; Yeung, 2018). Ces préoccupations trouvent également un écho dans les travaux d'une des co-auteures de cet article sur les préjudices sociétaux de l'IA, dans lequel il est avancé que les règles et approches existantes ont tendance à se concentrer principalement sur les préjudices individuels et collectifs soulevés par l'IA, alors que des préjudices sociétaux sont également en jeu, tels que les atteintes aux principes essentiels de l'État de droit ou à la démocratie (Smuha, 2021). Ce dernier type de préjudice n'est pas toujours facilement traduisible dans le langage des droits de l'homme, étant donné qu'il n'y a pas toujours une relation univoque entre un droit individuel et un préjudice sociétal, ce qui nécessite de repenser le système juridique, ainsi que de nouveaux recours juridiques qui reconnaissent et traitent la dimension sociétale de l'impact de l'IA[13].

L'IA COMME SEGMENT TECHNO-ÉCONOMIQUE : CONTRÔLER LE MARCHÉ

Le dernier type de production normative, et non des moindres, vise à construire la conformité des produits de l'IA afin de leur permettre d'être mis sur le marché communautaire européen, d'y circuler librement et d'y être utilisés. Dans ce contexte, une définition large de haut niveau, comme nous l'avons observé au nord-ouest, risquerait de créer de l'insécurité juridique. Dans une acception *in abstracto* de ce type, le droit délivre alors une réponse avec des aléas, notamment du fait de la relative indétermination du système juridique dans son ensemble (la fameuse texture ouverte du droit de Hart), mais aussi (et surtout) de l'application d'une nouvelle loi, avec par exemple de potentiels conflits d'interprétation entre juridictions sur ce qu'est – ou n'est pas – l'IA. Quand certains voient dans cette définition *in abstracto* une entrave potentielle à l'innovation et la commercialisation des technologies reposant sur de l'IA, d'autres la considèrent comme un instrument large et efficace de contrainte et de contrôle.

13. Alain Supiot a également rappelé en ce sens au Conseil National du Numérique en 2021 que le droit du travail est bien né du choc de l'industrialisation. Cf. compte rendu d'un échange avec Alain Supiot, professeur au Collège de France, Conseil National du Numérique, 21 septembre 2021 – https://cnnumerique.fr/le-probleme-est-de-savoir-comment-mettre-nos-nouveaux-outils-notre-service-au-lieu-de-nous-y, consulté le 19 avril 2022.

C'est pourquoi l'acception *in concreto* de la notion d'IA par la Commission européenne dans son projet de règlement de l'IA, concrétisée dans une annexe listant précisément les produits objets de la régulation en vue de produire des normes de fabrication, a paru rassurante pour de nombreux acteurs économiques. Précisons que cette approche est relativement originale dans la technique juridique d'encadrement de nouvelles technologies à forts enjeux éthiques, comme en témoignent d'autres textes comme le RGPD (ou la Convention d'Oviedo du Conseil de l'Europe sur la bioéthique). Ces textes ont pris le parti de rester neutres et abstraits dans leurs définitions des objets traités afin de s'adapter aisément aux inévitables et rapides évolutions des domaines techniques et industriels mouvants.

Depuis la mandature de von der Leyen, le déploiement d'une réglementation stricte sur le numérique (comprenant non seulement la proposition de règlement sur l'IA, mais aussi celle réformant la directive sur les machines, la législation sur les services numériques, la législation sur les marchés numériques et une autre sur la gouvernance européenne des données) vise à permettre à l'Europe d'aligner ses concurrents américains et asiatiques sur ses règles de marché (le fameux « effet de Bruxelles »). On peut même percevoir une sorte de concurrence réglementaire autour de l'IA, dans laquelle plusieurs régulateurs internationaux et nationaux cherchent à montrer leur pertinence en élaborant des mesures réglementaires dans ce domaine.

C'est pourquoi l'IA est conçue dans ce projet de règlement comme *un segment techno-économique* (De Prato *et al.*, 2019) impliquant une pluralité d'institutions (entreprises, centres de recherche, institutions gouvernementales, etc.). Les artefacts de l'IA, tels que les algorithmes d'apprentissage automatique, les dispositifs utilisant la vision par ordinateur ou la reconnaissance vocale et les objets connectés et automatisés, sont toujours envisagés par le biais des activités que ces institutions réalisent pour les produire, les développer et les échanger. Par conséquent, la Commission envisage la régulation de l'IA à travers toutes les activités dont les processus économiques sont orientés vers la fourniture de biens et de services liés à l'IA (activités industrielles) ; les activités de R&D sous la forme de demandes de brevets portant sur des développements technologiques liés à l'IA ; et, d'une manière plus marginale nous allons le voir, les activités de recherches universitaires liées à l'IA. La construction de l'IA comme segment techno-économique apparaît d'abord à travers les études produites par la Commission (notamment celles du JRC – Joint Research Center qui a consacré une étude entière à la définition de l'IA et du Groupe d'experts indépendants de haut niveau qui a également

publié, en plus de ses lignes directrices éthiques, un document spécifique sur la définition de l'IA). Ces études permettent de montrer qui sont les agents impliqués dans l'offre et l'évolution de l'IA dans le monde, le type d'artefacts développés dans ce domaine et la manière dont les agents interagissent et se comportent dans l'espace technico-économique.

Article 3
Définitions

Aux fins du présent règlement, on entend par:

(1) «système d'intelligence artificielle» (système d'IA), un logiciel qui est développé au moyen d'une ou plusieurs des techniques et approches énumérées à l'annexe I et qui peut, pour un ensemble donné d'objectifs définis par l'homme, générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions influençant les environnements avec lesquels il interagit;

(2) «fournisseur», une personne physique ou morale, une autorité publique, une agence ou tout autre organisme qui développe ou fait développer un système d'IA en vue de le mettre sur le marché ou de le mettre en service sous son propre nom ou sa propre marque, à titre onéreux ou gratuit;

(3) «petit fournisseur», un fournisseur qui est une micro ou petite entreprise au sens de la recommandation 2003/361/CE de la Commission[61];

(4) «utilisateur», toute personne physique ou morale, autorité publique, agence ou autre organisme utilisant sous sa propre autorité un système d'IA, sauf lorsque ce système est utilisé dans le cadre d'une activité personnelle à caractère non professionnel;

(5) «mandataire», toute personne physique ou morale établie dans l'Union ayant reçu mandat écrit d'un fournisseur de système d'IA pour s'acquitter en son nom des obligations et des procédures établies par le présent règlement;

(6) «importateur», toute personne physique ou morale établie dans l'Union qui met sur le marché ou met en service un système d'IA qui porte le nom ou la marque d'une personne physique ou morale établie en dehors de l'Union;

(7) «distributeur», toute personne physique ou morale faisant partie de la chaîne d'approvisionnement, autre que le fournisseur ou l'importateur, qui met un système d'IA à disposition sur le marché de l'Union sans altérer ses propriétés;

60 Directive 2000/31/CE du Parlement européen et du Conseil du 8 juin 2000 relative à certains aspects juridiques des services de la société de l'information, et notamment du commerce électronique, dans le marché intérieur («directive sur le commerce électronique») (JO L 178 du 17.7.2000, p. 1).

61 Recommandation de la Commission du 6 mai 2003 concernant la définition des micro, petites et moyennes entreprises (JO L 124 du 20.5.2003, p. 36).

Source : Commission européenne, Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union, p. 45.

Ces études servent aussi à alimenter le projet de régulation de l'IA mené par la Commission, et notamment les débats qui entourent l'*Artificial Intelligence Act*. Dans cette proposition de réglementation, la définition de l'IA renvoie à un travail de frontière sur ce qui est et ce qui n'est pas de l'IA et, partant, ce qui doit ou pas faire l'objet d'une régulation. Ce travail de frontière procède par des opérations de quadrillage du monde, en dressant une liste potentiellement ouverte de logiciels présentant une certaine dose d'autonomie.

ANNEXE I

TECHNIQUES ET APPROCHES D'INTELLIGENCE ARTIFICIELLE

visées à l'article 3, point 1

(a)	Approches d'apprentissage automatique, y compris d'apprentissage supervisé, non supervisé et par renforcement, utilisant une grande variété de méthodes, y compris l'apprentissage profond.
(b)	Approches fondées sur la logique et les connaissances, y compris la représentation des connaissances, la programmation inductive (logique), les bases de connaissances, les moteurs d'inférence et de déduction, le raisonnement (symbolique) et les systèmes experts.
(c)	Approches statistiques, estimation bayésienne, méthodes de recherche et d'optimisation.

Source : Commission européenne, Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union, p. 45 et Annexes à la proposition de règlement, p. 1.

Si l'*énumération* de ce qu'est l'IA a plusieurs fonctions, nous devons d'abord situer cette technique parmi les différentes méthodes habituelles de définition juridique pour en saisir la stratégie légistique. Ce procédé de définition dite « terminologique » est typique d'une approche juridique anglo-saxonne préférant « au grain des choses la paille des mots » (Cornu, 1981). Cornu distinguait les définitions « réelles », objectives et substantielles des choses, des définitions « terminologiques », qui en déterminent la compréhension dans le contexte particulier du texte juridique envisagé. Ainsi, le régulateur européen, en dressant une définition sommaire renvoyant à une liste annexe aisément modifiable, s'octroie le contrôle exact de ce qu'il entend encadrer dans le contexte précis de son règlement, afin d'éviter de se voir imposer une signification ou une interprétation casuistique par d'autres (et on le comprend aisément au vu de la longueur interminable des débats

dans la plupart des enceintes intergouvernementales ayant accueilli des travaux sur la régulation de l'IA). Cette approche est également considérée comme la plus propice à la sécurité juridique et permet aux personnes soumises à un texte législatif de mieux comprendre dans quelle mesure leurs systèmes entrent dans son champ d'application. Mais le procédé conduit également à déconnecter la notion de l'ensemble juridique inévitablement connexe, autonomisant (isolant même) ces travaux de ceux relevant du même ordre juridique (le droit de l'Union), voire de l'ordre juridique international.

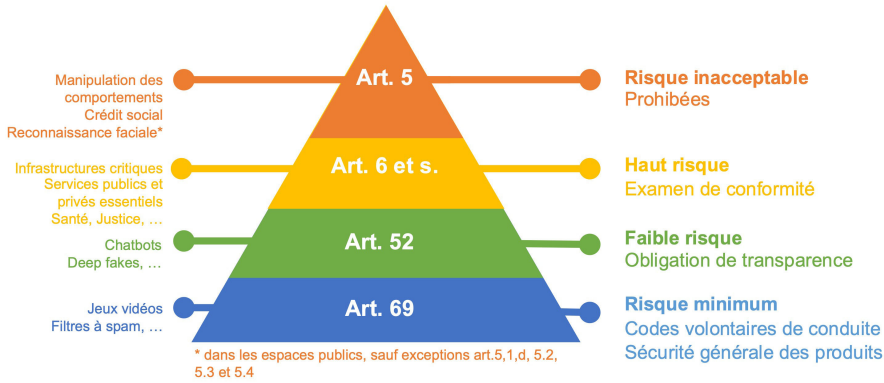
Cette volonté de contrôle est également confirmée par l'article 4 de la proposition, qui autorise la Commission, sans repasser par l'entier processus législatif, à redéfinir ce qu'il faudra entendre – ou non – comme étant une IA pour l'application spécifique de ce texte. La clé de voûte du texte est donc constituée par cette structure limitée et contrôlée de l'article 3 (1) et de l'annexe I, à laquelle s'ajoute, de plus, une approche par les risques pour moduler l'intensité de la régulation.

Ainsi, au lieu d'établir des dispositions générales protégeant de manière indistincte tous les individus de dommages, quel que soit le type de système d'IA, il a été préféré par la Commission d'échelonner l'intensité des contraintes de son texte en fonction des caractéristiques intrinsèques de chaque système. Cette initiative a également permis à la Commission de déclarer qu'elle était « le premier régulateur » à proposer un règlement spécifique pour l'IA, établissant ainsi un « *first mover advantage* » accompagnant le pouvoir économique associé à cette technologie populaire. Il doit être précisé que ce texte n'exclut naturellement pas l'application d'autres dispositions, comme celles du RGPD, dont il entend être un complément : un dommage créé par un système d'IA présentant un risque minimum pourrait donc, en théorie, être sanctionné sur d'autres fondements juridiques, comme le RGPD ou la responsabilité du fait des produits défectueux notamment.

Pour entrer dans le détail, le projet de règlement organise son approche en une sorte de pyramide de risques décomposée en quatre niveaux, allant de risques inacceptables (et prohibant donc l'emploi de systèmes d'IA, sous réserve de certaines exceptions) à un risque minimum, qui se voient imposer une obligation particulière de transparence (pour savoir par exemple que l'on interagit avec une machine ou qu'un contenu vidéo a été généré artificiellement). Les applications à haut risque sont celles soumises de manière obligatoire à des procédures de conformité (marqueur CE) et c'est à nouveau une énumération,

celle de l'article 6 combiné avec une annexe III, modifiable avec la même souplesse que l'annexe I, qui dresse la liste des applications devant être considérées à haut risque.

Figure 2. La pyramide des risques en quatre niveaux dans la proposition de réglementation de la Commission européenne



Source : Y. Meneceur, Analyse des principaux cadres supranationaux de régulation de l'IA – De l'éthique à la conformité, Les Temps Électriques, 31 mai 2021 – https://lestempselectriques.net/ANALYSE_IA.pdf (CC)BY-NC-ND.

Il n'est donc pas surprenant que la présidence slovène du Conseil de l'Union européenne soit tout d'abord revenue dans sa propre proposition de révision de ce texte sur la question de la définition de l'IA, en cherchant à la restreindre à la seule forme connexionniste (apprentissage automatique ou *machine learning*) et à en écarter d'autres formes, notamment symboliques, qui, au sens large, pourraient recouvrir la plupart des autres algorithmes. Toutefois, pour les défenseurs des droits de l'homme de la société civile, la combinaison du procédé de l'énumération (par nature restrictive) et d'une approche par les risques (modulant donc l'appréciation de potentielles violations des droits fondamentaux) n'apparaît pas comme satisfaisante. On peut le lire dans cet extrait d'un rapport de l'ONG Access Now :

> Une approche fondée sur le risque consiste à déterminer l'ampleur ou la portée des risques liés à une situation concrète et à une menace reconnue. Cette approche est utile dans les environnements techniques où les entreprises

> doivent évaluer leurs propres risques opérationnels. Toutefois, l'approche de l'UE voudrait que les entreprises évaluent leurs risques opérationnels par rapport aux droits fondamentaux des personnes. Il s'agit là d'une conception fondamentalement erronée de ce que sont les droits de l'homme ; ils ne peuvent être mis en balance avec les intérêts des entreprises. Les entreprises auraient également intérêt à minimiser les risques afin de développer des produits. Une approche de la réglementation basée sur le risque n'est donc pas adéquate pour protéger les droits de l'homme. Nos droits ne sont pas négociables et ils doivent être respectés indépendamment d'un niveau de risque associé à des facteurs externes. (Hidvegi *et al.*, 2021)[14]

Avec ces techniques de définition, la Commission européenne propose donc de procéder à un ordonnancement autorisant un contrôle extrêmement précis de son périmètre d'application, même après l'adoption du texte. Une définition réelle aurait échappé à son concepteur alors que l'énumération reste – en théorie – sous son contrôle. Ce contrôle est, de plus, caractérisé ici par une intervention démocratique du Parlement européen bien plus allégée au visa de l'article 73 de la proposition de règlement, afin de témoigner d'un certain pragmatisme au vu de l'évolution constante et rapide des technologies.

L'énumération n'est donc pas ici que la traduction d'une habitude légistique anglo-saxonne et d'organisations intergouvernementales. Elle traduit la volonté de totale maîtrise du régulateur européen, assumant une rupture tant avec la recherche qu'avec le caractère généralement abstrait de la loi. L'énumération rend vraisemblablement l'IA bien plus saisissable en facilitant considérablement le contrôle des activités de fabrication, mais elle conduit aussi dans le même temps à une forme de réification de l'IA, de liste et de stock de produits estimés à risques dont les critères d'arbitrages seront fluctuants. En toute hypothèse, cette liste pourrait être réduite au consensus politique résultant des négociations entre États, des débats parlementaires et du lobbying des acteurs économiques. Or, nous l'avons vu dans les quadrants 1,

14. « A risk-based approach involves determining the scale or scope of risks related to a concrete situation and a recognised threat. This approach is useful in technical environments where companies have to evaluate their own operational risks. However, the EU approach would have companies evaluate their operational risks vs. people's fundamental rights. This is a fundamental misconception of what human rights are; they cannot be put in a balance with companies' interests. Companies would also have an interest in downplaying the risks in order to develop products. A risk-based approach to regulation is therefore not adequate to protect human rights. Our rights are non-negotiable and they must be respected regardless of a risk level associated with external factors » (traduction auteurs) (Hidvegi *et al.*, 2021).

2 et 4 il existe bien d'autres manières d'envisager l'IA. Si le règlement parvient à surmonter les mêmes obstacles que le RGPD avait connus en son temps, d'importants enjeux de pouvoirs entoureront la capacité à faire évoluer l'énumération. Il s'agira en effet ainsi de légitimer par l'énumération des usages contestés, mais aussi de maintenir *hors énumération* une part du débat sur la politique de l'IA.

Pour donner un exemple concret, en intégrant dans l'énumération « les systèmes d'IA destinés à être utilisés par les autorités publiques compétentes en tant que polygraphes et outils similaires, ou pour analyser l'état émotionnel d'une personne physique » pour la gestion de la migration, de l'asile et des contrôles aux frontières (Annexe III, 7 a), la Commission légitime des outils dont le régime juridique est loin d'être homogène (il n'a aucune valeur juridique devant les tribunaux français par exemple) du fait de l'absence de consensus scientifique sur les résultats produits – sans même parler des problèmes éthiques que cette technologie soulève. D'autres universitaires, y compris ceux qui ont participé à l'origine au groupe d'experts indépendants de haut niveau de la Commission sur l'IA, ont souligné un problème similaire (Smuha *et al.*, 2021).

De manière plus signifiante pour les juristes, cette proposition de règlement, en s'appuyant pourtant dans ses motifs sur les droits fondamentaux, traduit aussi l'abandon de toute recherche de permanence et d'universalité d'une loi contribuant à encadrer un domaine techno-scientifique, pour assurer plus de flexibilité et d'adaptabilité pour le marché. Il n'est pas anodin de constater que le RGPD avait eu pour chef de file au sein de la Commission la Commissaire Věra Jourová chargée de la justice (et la DG JUST), alors que c'est aujourd'hui le Commissaire chargé du marché intérieur Thierry Breton et la Commissaire chargée de la concurrence Margrethe Vestager (et la DG CONNECT) qui pilotent cette proposition. Cela semble impliquer que même au sein de la Commission européenne, il existe différentes définitions et perceptions de l'IA – situées dans les différents quadrants que nous avons évoqués ci-dessus.

Les juristes qui prônent une définition abstraite de l'IA et un contrôle plus général ne sont pas les seuls à critiquer la réglementation en cours et sa définition très délimitée de l'IA. Cette critique nous ramène au point de départ de notre circuit, au nord-ouest de la carte, en pleine singularité technologique : alors auditionnés dans le cadre du projet d'AI Act, Stuart Russell et Mark Tegmark ont alerté les parlementaires européens en mars 2022 sur

l'obsolescence programmée de la réglementation. Cette dernière risque de ne pas intégrer dans son spectre définitionnel l'IA générale. Or, selon ces derniers, les modèles de langage reposant sur le concept de transformer comme GPT-3 d'Open AI montrent que la distance avec l'IA générale ne cesse de se réduire. Les menaces de l'IA pour l'humanité semblent déjà se manifester. Les exemples de perte de contrôle reposent de plus en plus sur une réalité tangible : « utilisé comme chatbot médical, GPT-3 conseille à un patient de se suicider », rappelle Tegmark, comme exemple, aux parlementaires européens. Mais cette focale sur la sûreté à laquelle nous ramènent systématiquement les milieux de la singularité écarte du débat des enjeux éthiques plus fondamentaux, comme ceux soulevés récemment en France par le Comité National Pilote du Numérique (CNPEN 2021) dans son avis n° 3 sur les agents conversationnels qui posent des questions éminemment abstraites sur le *contrôle* des frontières entre les machines et les humains.

CONCLUSION

Qu'apporte cette mise en regard de ces quatre arènes normatives ? Elle permet de montrer comment les modes d'existence variés de l'IA – à la fois être de fiction, technoscience, droit, politique et économie – impliquent des formes de contrôles différenciées, toujours relancées, en tension les unes avec les autres. Reste à en faire une analyse d'anthropologie politique plus approfondie. Mais ces quatre formes de régulation posent les jalons d'une première tentative de cartographie des conflits normatifs entre : les spéculations dystopiques sur les dangers d'une super-intelligence et le problème du contrôle de son alignement aux valeurs humaines ; l'auto-responsabilisation des chercheurs développant une science entièrement consacrée à la certification technique des machines ; les dénonciations des effets néfastes des systèmes d'IA sur les droits fondamentaux et le contrôle des rééquilibrages des pouvoirs ; enfin, la régulation européenne du marché par le contrôle de la sécurité du fait des produits et des services de l'IA.

Si nous sommes loin d'avoir épuisé la liste des acteurs et des lieux des quatre arènes, notre objectif dans cet article est plus modestement de construire une typologie des définitions de l'IA afin de représenter de manière synthétique l'espace social de la régulation. Notre typologie est parfaitement discutable et pourrait être mise à l'épreuve d'une enquête quantitative (carte de réseaux d'acteurs et analyse de corpus). Elle peut cependant nous aider à éclairer la dynamique actuelle de la régulation de l'IA, en proposant un cadre d'analyse

à la fois théorique et pratique. Les contours de ces quadrants fournissent un cadre pour comprendre le champ de bataille définitionnel qui se déroule actuellement en marge des négociations sur la définition de l'IA dans la proposition de réglementation européenne, en particulier au Parlement européen et au Conseil de l'Union. Il sera donc intéressant de voir si le champ de définition de l'IA actuellement proposé par la Commission dans le règlement peut encore se déplacer vers le nord ou le nord-ouest ou englober des éléments provenant d'angles multiples.

RÉFÉRENCES

AGRE P. E. (1997), « Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI », *in* Bowker G., GASSER L., STAR L., TURNER B. (eds.), *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work, Erlbaum*, Hillsdale, NJ: Lawrence Erlbaum Associates.

AI HLEG (High-level Expert Group on Artificial Intelligence) (2019), *Ethics guidelines for trustworthy AI*, European Commission, [en ligne] disponible à l'adresse : https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, consulté le 19 avril 2022.

ANDERSON M., ANDERSON S.L. (2011), *Machine Ethics*, Cambrige, Cambridge University Press.

AYLING J., CHAPMAN A. (2021), « Putting AI ethics to work: are the tools fit for purpose? », *AI and Ethics*, [en ligne] disponible à l'adresse : https://link.springer.com/content/pdf/10.1007/s43681-021-00084-x.pdf, consulté le 19 avril 2022.

BÖSTROM N. (2014), *Superintelligence: Paths, Dangers, Stratégies*, OUP, OXFORD.

BUOLAMWINI J., GEBRU T. (2018), « Gender shades: Intersectional accuracy disparities in commercial gender classification », *1st Conference on fairness, accountability and transparency*, PMLR, p. 1-15.

CARDON D., COINTET J., MAZIÈRES A. (2018), « La revanche des neurones : l'invention des machines inductives et la controverse de l'intelligence artificielle », *Réseaux*, n° 211, p. 173-220.

CHAMAK B. (2004), « Modèles de la pensée : quels enjeux pour les chercheurs en sciences cognitives ? », *Intellectica*, n° 39, p. 79-105.

CHAPMAN D. (1991), *Vision, Instruction, and Action*, Cambridge, MIT Press.

CNPEN (2021), Collectif Grinbaum A., Devillers L., Adda G., Chatila R. *et al., Agents conversationnels : enjeux d'éthique*, Rapport de recherche Comité national pilote d'éthique du numérique, [en ligne] disponible à l'adresse : https://hal-lara.archives-ouvertes.fr/cea-03432785v1, consulté le 19 avril 2022.

COLLINGRIDGE D. (1980), *The social control of technology*, London, Pinter.

CORNU G. (1981), *Les définitions dans la loi. Étude parue dans les Mélanges dédiés au doyen Jean Vincent*, Paris, Dalloz.

DELVAUX M. (2017), *Rapport contenant des recommandations à la Commission concernant des règles de droit civil sur la robotique (Rapport Delvaux)*, Parlement européen, [en ligne] disponible à l'adresse : https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_FR.html, consulté le 19 avril 2022.

DE PRATO G., MONTSERRAT L. C., SAMOILI S., RICCARDO R., BAILLET M. V. P., CARDONA M. (2019), *The AI Techno-Economic Segment Analysis*, JRC Working Papers JRC118071, Joint Research Centre, [en ligne] disponible à l'adresse : file:///C:/Users/aurelie.bur/Downloads/jrc118071_the_ai_techno-economic_segment_analysis_1%20(1).pdf, consulté le 19 avril 2022.

DIETTERICH T., HORVITZ E. (2015), Rise of Concerns about AI: Reflections and Directions, *Communications of the ACM*, vol. 58, n° 10, p. 38-40.

DWORK C., HARDT M., PITASSI T., REINGOLD O., ZEMEL R. (2012), « Fairness through awareness », *Proceedings of the 3rd innovations in theoretical computer science conference*, p. 214-226.

FLECK J. (1982), « Development and Establishment in Artificial Intelligence », *in* N. ELIAS, H. MARTINS, R. WHITLEY (eds.), *Scientific Establishments and Hierarchies, Sociology of the Sciences Yearbook*, vol. 6, Dordrecht, Reidel, p. 169-217.

GANASCIA J. B. (2017), *Le mythe de la singularité. Faut-il craindre l'intelligence artificielle ?*, Paris, Le Seuil.

GAUTRAIS V. (2012), *Neutralité technologique. Rédaction et interprétation des lois face aux changements technologiques*, Montréal, Thémis.

GOOD I. J. (1966), « Speculations concerning the first ultraintelligent machine », *Advances in computers*, vol. 6, p. 31-88.

GUEGUEN H., JEANPIERRE L. (2022), *La perspective du possible : Comment penser ce qui peut nous arriver, et ce que nous pouvons faire*, Paris, La Découverte.

HIDVEGI F., DANIEL L., MASSE E. (2021), « The EU should regulate AI on the basis of rights, not risks, Access Now », [en ligne] disponible à l'adresse : https://www.accessnow.org/eu-regulation-ai-risk-based-approach/, consulté le 19 avril 2022.

HILDEBRANDT M. (2022), « Global competition and convergence of AI law, draft chapter for Elgar Encyclopedia for Comparative Law », [en ligne] disponible à l'adresse : https://osf.io/preprints/socarxiv/j36ke/, consulté le 19 avril 2022.

HLEGAI (2019), « High-Level Expert Group on Artificial Intelligence, EU – Ethics guidelines for trustworthy AI », https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

HORVITZ E. (2014), *One-Hundred Year Study on Artificial Intelligence: Reflections and Framing*, Stanford University, [en ligne] disponible à l'adresse : https://ai100.stanford.edu/reflections-and-framing, consulté le 19 avril 2022.

HORVITZ E., SELMAN B. (2012), « Interim report from the panel chairs: AAAI Presidential Panel on Long-Term AI Futures », *in* EDEN H., MOOR J. H., SØRAKER J. H., STEINHART E., *Singularity Hypotheses*, Berlin, Springer, p. 301-308.

HORVITZ E., YOUNG J., ELLURU R. G., HOWELL C. (2021), « Key Considerations for the Responsible Development and Fielding of Artificial Intelligence », *National Security Commission on Artificial Intelligence*, [en ligne] disponible à l'adresse : https://www.nscai.gov/wp-content/uploads/2021/04/Key_Considerations_Extended_April_2021.pdf, consulté le 19 avril 2022.

HURLBUT J. B. (2015), « Remembering the future: Science, law, and the legacy of Asilomar », *in* S. JASANOFF, S.-H Kim (eds.), *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*, Chicago, University of Chicago Press, p. 126-51.

HUTTER M. (2004), *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*, Berlin, Springer.

JASANOFF S. (1990), *The fifth branch: science advisors as policy makers,* Cambridge, MA, Harvard University Press.

JORDAN M. (2018), « Artificial Intelligence: The Revolution hasn't happened yet », Medium, April 19, [en ligne] disponible à l'adresse : https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7, consulté le 19 avril 2022.

KATZ Y. (2020*), Artificial Whiteness: Politics and Ideology in Artificial Intelligence*, New York, Columbia University Press.

KEARNS M., ROTH A. (2019), *The ethical algorithm: The science of socially aware algorithm design*, New York, Oxford University Press.

KLINE R. (2011), Cybernetics, automata studies, and the Dartmouth conference on artificial intelligence, *IEEE Annals of the History of Computing*, vol. 33, n° 4, p. 5-16.

KOENE A. (2022), *A survey of artificial intelligence risk assessment methodologies. The global state of play and leading practices identified*, Trilateral Research Report, [en ligne] disponible à l'adresse : https://www.trilateralresearch.com/wp-content/uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf, consulté le 19 avril 2022.

LATOUR B. (2009), *Sur le culte des dieux faitiches*, Paris, Les Empêcheurs de penser en rond.

LAUMOND J.-P. (2012), *La robotique: une récidive d'Héphaïstos: Leçon inaugurale prononcée le jeudi 19 janvier 2012*, Paris, Fayard.

LOEB Z. (2021), « The lamp and the lighthouse: Joseph Weizenbaum, contextualizing the critic », *Interdisciplinary Science Reviews*, vol. 46, n° 1-2, p. 19-35.

MALLAT S. (2016), « Understanding deep convolutional networks », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, n° 2065.

MENECEUR Y. (2020), *L'intelligence artificielle en procès*, Bruxelles, Bruylant.

MENECEUR Y., HIBBARD L. (2021), « Les apports du Conseil de l'Europe à une réglementation globale de l'intelligence artificielle : Revue des instruments juridiques du Conseil de l'Europe relatifs à l'intelligence artificielle et des enjeux particuliers en matière de santé et de biomédecine », *Droit, Santé et Société*, n° 3, p. 55-63.

MORAVEC H. (1988) *Mind children: The future of robot and human intelligence*, Cambridge, Harvard University Press.

NEMITZ P. (2018), « Constitutional democracy and technology in the age of artificial intelligence », *Philosophical Transaction*, The Royal Society, vol. 376, n° 2133.

NEMITZ P. (2021), « Democracy through law. The Transatlantic Reflection Group and its manifesto in defence of democracy and the rule of law in the age of "artificial intelligence" », *European Law Journal*, p. 1-12.

ORSEAU L., ARMSTRONG S. (2016), « Safely interruptible agents », [en ligne] disponible à l'adresse : http://intelligence.org/files/Interruptibility.pdf, consulté le 19 avril 2022.

RAJARAMAN V. (2014), « JohnMcCarthy – Father of artificial intelligence », *Resonance*, n° 19, p. 198-207.

REED C. (2007), « Taking sides on technology neutrality », *SCRIPT-ed*, vol. 4, n° 3, [en ligne] disponible à l'adresse : https://script-ed.org/wp-content/uploads/2016/07/4-3-Reed.pdf, consulté le 19 avril 2022.

RUSSELL S. (2019), *Human compatible: Artificial intelligence and the problem of control*, London, Penguin.

RUSSELL S., DEWEY D., TEGMARK M. (2015), « Research priorities for robust and beneficial artificial intelligence », *AI Magazine*, vol. 36, n° 4, p. 105-114.

SHANNON C. E., McCARTHY J. (eds) (1956), *Automata Studies*, n° 34, Princeton, Princeton University Press.

SELBST A., BOYD D., FRIEDLER S. A., VENKATASUBRAMANIAN S., VERTESI J. (2019), « Fairness and abstraction in sociotechnical systems », *Proceedings of the conference on fairness, accountability, and transparency*, p. 59-68.

SMUHA N. A. (2019), « The EU approach to ethics guidelines for trustworthy artificial intelligence », *Computer Law Review International*, vol. 20, n° 4, p. 97-106.

SMUHA N. A. (2021), « Beyond the individual: governing AI's societal harm », *Internet Policy Review*, vol. 10, n° 3.

SMUHA N. A., AHMED-RENGERS E., HARKENS A., LI W., MACLAREN J., PISELLI R., YEUNG K. (2021), *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act,* Response to the European Commission's Proposal for an Artificial Intelligence Act from members of the Legal, Ethical & Accountable Digital Society (LEADS) Lab at the University of Birmingham.

SUPIOT A. (2005), *Homo juridicus. Essai sur la fonction anthropologique du droit*, Paris, Seuil.

TURING A. (1950), « Computing Machinery and Intelligence », *Mind*, vol. 59, n° 236, p. 433-460.

WEIZENBAUM J. (1976), *Computer Power and Hulan Reason: From Judgement to Calculation,* San Francisco, W.H. Freeman and Company.

WIENER N. (1950), *The Human Use of Human Beings. Cybernetics and Society*, London, Eyre & Spottiswoode.

YEUNG K. (2018), *A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility Within a Human Rights Framework*, MSI-AUT, Council of Europe.

ZUBOFF S. (2019), *The age of surveillance capitalism: The fight for a human future at the new frontier of power*, London, Profile Books.