LE CONTRÔLE HUMAIN DES SYSTÈMES ALGORITHMIQUES - UN REGARD CRITIQUE SUR L'EXIGENCE D'UN "HUMAIN DANS LA BOUCLE"

Winston Maxwell

LE CONTRÔLE HUMAIN DES SYSTÈMES ALGORITHMIQUES - UN REGARD CRITIQUE SUR L'EXIGENCE D'UN "HUMAIN DANS LA BOUCLE"

Mémoire original pour présenter l'habilitation à diriger des recherches de l'Université Panthéon-Sorbonne, soutenue le 25 novembre 2022 par Winston Maxwell*

Jury:
Célia Zolynski, Professeure, Université Panthéon Sorbonne, Présidente du Jury
Brunessen Bertrand, Professeure, Université de Rennes, Rapporteure
Nicolas Curien, Professeur, Conservatoire National des Arts et Métiers, Rapporteur
Judith Rochfeld, Professeure, l'Université Panthéon-Sorbonne, Garante
Jean-Yves Ollier, Conseiller d'Etat
Céline Castets-Renard, Professeure University of Ottawa

*Directeur d'études, droit et numérique, Télécom Paris - Institut Polytechnique de Paris
Laboratoire i3 (UMR 9217),winston.maxwell@telecom-paris.fr

Résumé
Un consensus dans l'univers du droit a émergé sur la nécessité d'un "humain dans la boucle" pour les systèmes IA en particulier quand ils ont un impact important pour les droits humains (déclenchement d'un contrôle policier, refus d'un prêt…). Simple en apparence, le principe de l'humain dans la boucle pose cependant de nombreuses questions, à commencer par ; "à quoi sert-il ?" et "comment l'organiser pour qu'il soit efficace?"

L'intervention humaine est supposée être un remède à plusieurs maux algorithmiques : elle permettrait de détecter des erreurs algorithmiques, de rendre le processus algorithmique "plus juste", et de contribuer à une meilleure responsabilisation des décisions algorithmiques. Mais comment l'humain peut-il faire tout cela en même temps, surtout lorsqu'il n'a que quelques secondes pour agir face à une quantité vertigineuse de données à examiner ?

Cet ouvrage examine la question de "l'humain dans la boucle" sous l'angle des droits fondamentaux, du futur règlement européen IA Act, et des sciences informatiques et cognitives, une approche qui permet de passer du principe éthique du contrôle humain à ses réalités opérationnelles. Mettre un "humain dans la boucle" peut être une fausse bonne idée, l'humain devenant lui-même un automate qui valide les résultats algorithmiques. Comment éviter un nivellement par le bas, et s'assurer que l'humain et la machine gardent chacun leurs spécificités et leur valeur ajoutée ? L'auteur propose une approche analytique qui allie respect des droits humains et efficacité opérationnelle des systèmes d'IA.

TABLE DES MATIÈRES Page

INTRODUCTION - POURQUOI ÉTUDIER LE CONTRÔLE HUMAIN DES DÉCISIONS ALGORITHMIQUES ?

La présente étude explore le "pourquoi" et le "comment" du contrôle humain des systèmes de décision algorithmiques. Elle explorera les objectifs du contrôle humain, les modalités et limitations de celui-ci, et les textes qui l'encadrent, dans l'objectif de pouvoir mieux spécifier, sur le plan réglementaire, le type de contrôle humain adapté à chaque situation. L'idée d'étudier le contrôle humain de décisions algorithmiques peut se justifier à la lecture de l'arrêt de la Cour de Justice de l'Union Européenne (CJUE) du 6 octobre 2020 dans l'affaire *La Quadrature du Net*[1]. Pour qu'un système algorithmique de signalement de risques terroristes soit compatible avec la Charte des droits fondamentaux de l'Union Européenne, chaque signalement algorithmique doit être réexaminé par des moyens non-automatisés. Le rôle central du contrôle humain dans la protection des droits a été réaffirmé par la CJUE le 21 juin 2022[2] dans une affaire concernant l'analyse de données PNR de passagers aériens : un système de détection qui génère plus de 80% de faux signalements peut néanmoins être compatible avec la Charte des droits fondamentaux de l'Union européen, à condition de prévoir un contrôle humain efficace. Mais l'obligation de réexaminer chaque signalement algorithmique peut laisser perplexe. À partir de quelles informations les contrôleurs humains sont-ils supposés vérifier ces signalements ? Comment s'assurer que ce réexamen sera effectif, sachant que le contrôle humain se heurte à des biais cognitifs humains, notamment les biais d'automatisation ?

---

[1] CJUE, 6 octobre 2020, *La Quadrature du Net*, aff. jointes C-511/18, C-512/18 et C-520/18, qui reprend les conditions déjà imposées par la Cour dans son avis 1/15 du 26 juillet 2017, *Accord PNR UE-Canada*. Répertoire IP/IT et Communication Données à caractère personnel – Décision automatisée et justice – Liane HUTTNER – Novembre 2020; B. Bertrand, Chronique Droit européen du numérique - Les enjeux de la surveillance numérique RTD eur. 2021, 175; Chronique UE et droits fondamentaux - Droit au respect de la vie privée (art. 7 Charte), droit à la protection des données personnelles (art. 8 Charte) – Florence Benoît-Rohmer – RTD eur. 2021. 973; N. Mallet-Poujol, Droit des communications électroniques (1re partie), Légipresse 2021, 240; J. Larrieu, Ch. Le Stanc, P. Tréfigny, Droit du numérique D. 2020, 2262; W. Maxwell, La CJUE dessine le noyau dur d'une future régulation des algorithmes, Légipresse 2020, 671; M. Lassalle, Protection des données, renseignement, procédure pénale et enquêtes administratives : l'approche française remise en cause par la CJUE, D. 2021,406; E. Daoud, I. Bello, O. Pecriaux, Données de connexion et sauvegarde de la sécurité nationale : l'exception confirme la règle, Dalloz IP/IT 2021, 46; D. Simon, Droits fondamentaux - Protection des données Europe n° 12, Décembre 2020, comm. 374; Union européenne - Un an de Droit pénal de l'Union européenne (Février 2020 – Février 2021) - Chronique par Olivier CAHN Droit pénal n° 4, Avril 2021, chron. 4; Données de connexion et lutte contre la criminalité - Dispositif de surveillance des personnes versus droit à la protection des données à caractère personnel - Commentaire par Anne Danis-Fatôme Communication Commerce électronique n° 7-8, Juillet 2022, comm. 52 ; Protection des données - La Cour de justice revient sur l'interdiction absolue des mesures générales de conservation et de traitement des données à caractère personnel, pour finalement en dresser le régime dérogatoire - Note sous arrêt par Dominique Berlin La Semaine Juridique Edition Générale n° 48, 23 Novembre 2020, 1323 ; Numérique - Droit de la donnée - Chronique Sous la direction de Matthieu Bourgeois et Louis Thibierge Avec la collaboration de Julie Dehavay La Semaine Juridique Entreprise et Affaires n° 25, 23 Juin 2022, 1225 ,; Droit de la communication - Droit de la communication - Chronique par Pascale Idoux et Laurence Calandri La Semaine Juridique Edition Générale n° 17, 26 Avril 2021, doctr. 478

[2] CJUE, 21 juin 2022, Ligue des droits humains c. Conseil des ministres, aff. C-817/19; Vie privée - Conditions de transfert, de conservation et de traitement des données PNR - Veille par Dominique Berlin La Semaine Juridique Edition Générale n° 27, 11 Juillet 2022, 857; Répertoire de droit européen / La protection des données personnelles dans les relations internes à l'Union européenne Eur. – Céline CASTETS-RENARD – Mise à jour de juin 2022

A la même époque que la décision de la CJUE dans l'affaire *La Quadrature du Net*, plusieurs articles sont parus dans les *law reviews* américaines questionnant le rôle de l'intervention humaine par rapport au principe constitutionnel américain de *due process*. Dans le cadre de décisions administratives, un décisionnaire humain est-il toujours nécessaire ? Or, le débat outre-atlantique sur le rôle de l'humain dans les décisions administratives a de nombreux points communs avec le débat en France sur les décisions administratives entièrement automatisées[3]. Le principe de contrôle humain a également été mis en exergue par les recommandations du groupe de travail de haut niveau de la Commission européenne sur l'intelligence artificielle[4] et par la proposition de règlement européen sur l'IA[5].

L'ensemble de ces textes a mis en évidence plusieurs questions méritant une réflexion plus approfondie. Premièrement, qu'entend-on par contrôle humain ? Le terme a de nombreuses significations opérationnelles, allant du rôle humain dans la conception des objectifs et des paramètres de l'algorithme en amont, jusqu'à l'intervention humaine pour traiter des contestations individuelles en aval. Le terme "contrôle humain" n'est donc pas suffisamment précis et il s'avère nécessaire de développer un vocabulaire pour différencier les différents types de contrôle afin de mieux analyser leur utilité. Ce sera l'objet de la première partie.

Deuxièmement, quels sont les textes juridiques et décisions de justice aux États-Unis, en France et au sein de l'UE qui imposent un contrôle humain, et comment ces textes et décisions évoquent-ils les différents types de contrôle humain que nous avons répertoriés ? Est-ce que ces textes nous donnent des indications sur les modalités et les objectifs du contrôle humain ? Nous y réfléchirons dans une deuxième partie.

En troisième lieu, comment le contrôle humain contribue-t-il à la détection d'erreurs ? Dans les décisions de la CJUE le premier objectif annoncé est de réduire le nombre de faux positifs ; le contrôle humain servirait donc principalement à détecter les erreurs. Mais quels types d'erreurs, et comment le contrôle humain peut-il les détecter ? Dans certains domaines, les résultats algorithmiques sont moins erronés en moyenne que les décisions humaines. De plus, l'esprit humain se prête mal à la détection d'erreurs algorithmiques, surtout lorsque le temps est limité. La détection de discriminations, un autre objectif mis en avant par la CJUE, soulève d'autres questions sur le "comment" du contrôle humain. La finalité de la détection des erreurs et des discriminations peut donc paraître difficile à atteindre, ce que nous verrons dans une troisième partie.

En outre, la détection d'erreurs et de discriminations n'est pas la seule finalité du contrôle humain. Les articles traitant de la question de *due process* américain soulignent les valeurs procédurales associées à l'intervention humaine, quelle que soit son utilité dans la détection d'erreurs. Une procédure impliquant un décisionnaire humain serait plus respectueuse des valeurs humaines, indépendamment de son rôle dans la qualité non-erronée de la décision. Quelles sont ces valeurs procédurales, et comment le contrôle humain peut-il contribuer à les préserver ? La question se pose surtout dans le secteur privé où les garanties de *due process*, et les garanties du procès équitable, ne

[3] Cons. const. Décision n° 2018-765 DC du 12 juin 2018.
[4] Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle, *Lignes directrices en matière d'éthique pour une IA digne de confiance*, 8 avril 2019.
[5] Commission européenne, *Proposition de règlement établissant des règles harmonisées concernant l'intelligence artificielle*, COM(2021) 206 final, 24 avril 2021 (proposition de règlement européen sur l'IA).

s'appliquent pas directement. Il apparaît nécessaire d'identifier certaines valeurs procédurales clefs afin de les importer éventuellement dans les systèmes privés de prise de décisions algorithmiques. Nous nous interrogerons sur ce point dans une quatrième partie.

Le dernier objectif de la mise en place d'un contrôle humain est de permettre à l'exploitant du système de démontrer sa vigilance et sa conformité. Le contrôle humain serait une mesure de prudence, une mesure "appropriée" pour réduire les risques. La question de la responsabilité civile pour les dommages causés par les systèmes d'IA fait toujours débat. Néanmoins, que ce soit aux États-Unis ou en Europe, la responsabilité pour faute reste un point de référence solide. L'absence de contrôle humain constituerait-elle une faute ? Dit autrement, est-ce que le devoir de prudence et de vigilance impose un contrôle humain ? Le RGPD[6] et le futur règlement européen sur l'IA imposent un système de conformité, appelé *accountability* ou redevabilité, qui consiste à faire une analyse de risques, et mettre en place des mesures pour éliminer les risques ou les réduire à un niveau "acceptable". Quel est le rôle du contrôle humain dans ces mesures de réduction du risque, et quel est l'impact des coûts de mise en œuvre du contrôle humain dans la détermination du niveau "approprié" de ces mesures ? Nous aborderons ces questions dans une cinquième partie.

Quant à la démarche adoptée dans l'ensemble du travail, elle est comparatiste, s'appuyant à la fois sur des éléments de droit américain et français. En plus des travaux de chercheurs en droit, le mémoire se nourrit de la littérature sur interactions humain-machine (HCI ci-après), une discipline en sciences informatiques qui s'est penchée sur le rôle du contrôle humain. La littérature HCI met en évidence le lien entre le contrôle humain et l'explicabilité des algorithmes. Ainsi, même si le sujet de l'explicabilité des résultats algorithmiques n'est pas au centre de cette étude, il sera évoqué à plusieurs endroits.

L'étude est donc organisée en six parties. La Partie I présentera les modalités de contrôle humain, qui seront divisées en deux grandes familles : les contrôles humains "système" et les contrôles humains "individuels". L'étude se penchera particulièrement sur ce second cas, les contrôles humains "individuels" du type évoqué par la CJUE dans l'affaire *La Quadrature du Net*. La Partie II présentera ensuite les principaux textes qui encadrent et exigent du contrôle humain en France et aux Etats-Unis. La Partie III examinera le premier objectif du contrôle humain : la détection d'erreurs et de discriminations. De quelles erreurs s'agit-il ; et le contrôle humain est-il capable de les détecter ? Quel est l'impact des biais cognitifs humains ? La Partie IV examinera le deuxième objectif du contrôle humain : la préservation des valeurs procédurales. Le contrôle humain peut favoriser la contestabilité, la discussion, et le caractère "essentiellement humain"[7] de certaines décisions. La Partie V présentera le troisième objectif du contrôle humain : utiliser le contrôle humain pour exécuter son obligation de vigilance et montrer sa conformité. On examinera dans cette partie le rôle de la responsabilité civile et de la conformité dans la mise en place des mesures de contrôle humain, et notamment la question de ce qu'est un niveau "approprié" de contrôle humain. Enfin, la Partie VI présentera quelques suggestions pour améliorer le cadre réglementaire pour encourager l'adoption de mesures de contrôle humain efficaces.

---

[6] Règlement n° 2016/679/UE du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive n° 95/46/CE (RGPD).

[7] Rapport de la Commission des lois du Sénat sur le projet de loi informatique et libertés, Rapporteur Jacques Thyraud, 10 novembre 1977, p. 22.

I - LES MODALITÉS DU CONTRÔLE HUMAIN

Cette partie jettera les fondations opérationnelles nécessaires pour étudier le contrôle humain : quand et comment intervient-il ? Pour bien différencier les différents scénarios de contrôle humain, nous diviserons le cycle de vie d'un système IA en trois phases : la phase d'élaboration et de tests préalables de l'algorithme (phase 1), la phase opérationnelle (phase 2), et la phase de tests et d'audits après la mise en service de l'algorithme (phase 3)[8]. Nous distinguerons ensuite entre le contrôle humain "système" et le contrôle humain "individuel". Cette classification nous aidera à créer un vocabulaire commun pour désigner les différents types de contrôle humain.

A. Les multiples noms donnés au contrôle humain

Il existe de nombreux termes pour désigner le contrôle humain : intervention humaine[9], intervention humaine significative[10], contrôle de décision significatif[11], surveillance et vérification humaines[12], contrôles humains significatifs[13], contrôle humain effectif[14], véritable contrôle humain[15], supervision humaine complète à tout moment[16], évaluation humaine[17], examen humain[18], contrôle effectif par des personnes physiques[19], human-in-the-loop (HITL)[20], human-on-the-loop (HOTL)[21], human in

[8] D'autres classifications sont possibles. Ainsi, le NIST américain distingue la phase de pré-conception, la phase de conception et de développement, et la phase d'exploitation. v. NIST, Special Publication 1270, A Proposal for Identifying and Managing Bias in Artificial Intelligence, juin 2021, p. 6.

[9] Règlement nº 2016/679/UE du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive nº 95/46/CE (RGPD), art. 22; Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission concernant un cadre pour les aspects éthiques de l'intelligence artificielle, de la robotique et des technologies connexes, 2020/2012(INL), point 89; CNIL, « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle », Rapport de la CNIL, décembre 2017.

[10] Résolution du parlement européen du 20 octobre 2020, précité, point 69.

[11] Groupe de Travail Art 29, Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679, 3 oct. 2017 révisée le 6 févr. 2018 WP 251rev.01 p. 23

[12] Règlement 2021/784 sur du 29 avril 2021 relatif à la lutte contre la diffusion des contenus à caractère terroriste en ligne, art. 5(3).

[13] Résolution du parlement européen du 20 octobre 2020, point 68 et considérant 10.

[14] Commission nationale consultative des droits de l'homme (CNCDH), Avis relatif à l'impact de l'intelligence artificielle sur les droits fondamentaux, Avis A-2022-6 du 7 avril 2022.

[15] Résolution du Parlement européen du 20 janvier 2021 sur l'intelligence artificielle: questions relatives à l'interprétation et à l'application du droit international dans la mesure où l'Union est concernée dans les domaines des utilisations civiles et militaires ainsi qu'à l'autorité de l'État en dehors du champ d'application de la justice pénale, point 3.

[16] ibid art. 7(1).

[17] ibid point 110.

[18] ibid point 35.

[19] Proposition de règlement européen sur l'IA, art. 14(1).

[20] Wu, Xingjiao & Xiao, Luwei & Yixuan, Sun & Zhang, Junhang & Ma, Tianlong & He, Liang. (2021). A Survey of Human-in-the-loop for Machine Learning. 2 August 2021, arXiv 2108.00941.

[21] Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies, Lignes directrices en matière d'éthique pour une IA digne de confiance, Publications Office, 2019, https://data.europa.eu/doi/10.2759/54071 (Lignes directrices HLEG), point 65

command (HIC)[22], maîtrise par l'utilisateur[23], *Meaningful Human Control* (MHC)[24], réexamen individuel par des moyens non-automatisés[25], *human review*[26], *human monitoring*[27], *meaningful human review*[28], garantie humaine[29], ou encore "un examen, un jugement, une intervention et un contrôle humains significatifs"[30]. Ces termes s'accompagnent parfois de définitions[31] qui nous permettent d'extraire trois conditions cumulatives qui caractérisent un contrôle humain efficace : la personne effectuant le contrôle doit (i) avoir une connaissance du fonctionnement de l'algorithme et de ses limitations[32], (ii) s'engager dans un processus cognitif de réflexion qui, pour certains[33], doit tenir compte d'autres informations telles que le contexte de la décision, et (iii) avoir l'autorité et la capacité matérielle d'intervenir dans le système et changer la décision[34].

Le contrôle humain doit, par ailleurs et d'un point de vue temporel, s'effectuer tout au long du cycle de vie d'un système algorithmique[35]. Le contrôle humain se place sur un continuum dans le temps, de la conception de l'algorithme jusqu'à sa mise en oeuvre[36]. Le contrôle humain peut s'effectuer avant la prise d'effet d'une décision algorithmique, ou après. Pour simplifier, on peut diviser le contrôle humain en deux catégories : le contrôle humain relatif au fonctionnement du système dans sa globalité - le contrôle "système" - et le contrôle humain relatif à une décision algorithmique individuelle - le contrôle "individuel".

---

[22] Ibid.

[23] Conseil de l'Europe, Commission européenne pour l'efficacité de la justice (CEPEJ), "Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement", 2018, p. 12.

[24] F. Santoni de Sio, J. van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account", *Front. Robot. AI*, 28 Feb. 2018 https://doi.org/10.3389/frobt.2018.00015

[25] Directive (UE) 2016/681 du 27 avril 2016 relative à l'utilisation des données des dossiers passagers (PNR) pour la prévention et la détection des infractions terroristes et des formes graves de criminalité, ainsi que pour les enquêtes et les poursuites en la matière, art. 6(6) (directive PNR).

[26] Proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme 9 décembre 2021 COM(2021) 762 final, art. 8.

[27] Ibid., art. 7.

[28] Revised Code of Washington (RCW) §43.386.010, par. 10.

[29] Comité consultatif national d'éthique (CCNE) Avis 129: "Contribution à la révision de la loi de bioéthique 2018-2019", 18 septembre 2018.

[30] Résolution du parlement européen du 20 octobre 2020, considérant 10.

[31] Voy. notamment GROUPE D'EXPERTS INDÉPENDANTS DE HAUT NIVEAU SUR L'INTELLIGENCE ARTIFICIELLE, *Lignes directrices en matière d'éthique pour une IA digne de confiance*, 8 avril 2019, point 65 pour les termes "human-in-the-loop", "human-on-the-loop" et "human-in-command"; Revised Code of Washington (RCW) §43.386.010, par. 10., pour le terme "*meaningful human review*", F. Sauer, ICRAC INT'L COMM. FOR ROBOT ARMS CONTROL, "ICRC Statement on Technical Issues to the 2014 UN CCW Expert Meeting", (May 14, 2014) pour le terme "*meaningful human control*"; voy. également B. Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms" (sept. 2021), SSRN: 3921216, p. 12 pour une analyse de documents exigeant un contrôle humain "significatif".

[32] Proposition de règlement européen sur l'IA, art. 14(4); voy égal. la loi de l'Etat de Washington (RCW Chapter 43.386, op cit.) qui exige une formation préalable des personnes chargées du contrôle.

[33] B. Green, op. cit., p. 17.

[34] voy. notamment la Proposition de règlement européen sur l'IA, art. 14(4), et F. Sauer "ICRC Statement on Technical Issues", op cit.; Parlement européen résolution du 20 janvier 2021, op. cit., point 3.

[35] Sur le cycle de vie d'un système IA voir D. Leslie, C. Burr, M. Aitken, J. Cowls, M. Katell et M. Briggs, Intelligence artificielle, droits de l'homme, démocratie et État de droit. Guide introductif, Conseil de l'Europe, 2021; voy. également CNCDH avis du 7 avril 2022, point 71.

[36] Pour une illustration de tous les stades au cours desquels le contrôle humain peut s'effectuer, voir le diagramme de l'iceberg utilisé dans M. Ekelhof, G. Persi Paoli, L'élément humain dans les décisions relatives à l'utilisation de la force, United Nations Institute for Disarmament Research (UNIDIR), 2020.

Mais avant de rentrer dans ces distinctions, précisons de quel "humain" il s'agit lorsque l'on évoque le contrôle humain. On se réfère là à un par les personnes responsables du système, non par les personnes affectées par celui-ci. Ces dernières peuvent exercer un certain contrôle à travers les choix qu'elles font lorsqu'elles interagissent avec le système[37]. Les dispositions du RGPD visent en partie à garantir la réalité de ces choix par l'utilisateur final. Cet aspect du contrôle humain est important, particulièrement lorsque l'on évoquera, dans la troisième partie, les valeurs procédurales. La personne affectée par une décision algorithmique bénéficie dans certains cas d'un droit de participer à la décision et d'interagir avec le décisionnaire[38]. Mais à part cet aspect de participation, le contrôle humain qui nous intéresse sera celui exercé par l'utilisateur de l'algorithme[39], et dans un moindre mesure, celui exercé par le fournisseur de l'algorithme[40].

B. Les contrôles humains pendant les trois phases de la vie de l'algorithme

Dans les sections qui suivent, nous allons dresser une liste des types de contrôle humain susceptible d'intervenir durant les trois principales phases de la vie de l'algorithme : la phase de conception, de test et de validation du système (phase 1), la phase d'exploitation pendant laquelle les résultats individuels sont générés et les décisions individuelles prises (phase 2), et la phase de tests et d'audits du système après sa mise en production (phase 3). Les phases 2 et 3 peuvent coïncider sur le plan temporel, mais la phase 2 s'insère dans le processus d'exploitation alors que la phase 3 se passe en dehors du processus d'exploitation.

1. La phase 1 : contrôles humains pendant la conception, les tests, et la validation de l'algorithme

La phase 1 nécessite uniquement un contrôle humain "système" la phase d'exploitation et la génération de résultats n'ont pas démarré. Pendant cette phase, contrôle humain "système" concernera la définition du problème à résoudre et sa traduction en objectifs mathématiques, phase pendant laquelle certains biais peuvent s'introduire, puisque la chose que l'on souhaite prédire - par exemple quel candidat saura être un "bon" salarié - doit être traduit en objectifs quantitatifs s'appuyant sur d'autres critères de substitution[41]. Pour un modèle à base de règles, l'humain définira lui-même les critères de décision[42]. Ce point est important car, comme nous le verrons ci-après[43], les décisions de la CJUE et du Conseil

[37] CNCDH, avis du 7 avril 2022, op cit., point 70.
[38] Mantelero, A. (2022). Regulating AI. In: Beyond Data. Information Technology and Law Series, vol 36. T.M.C. Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-531-7_4, §4.2.1.2.
[39] La proposition de règlement européen sur l'IA utilise le terme "utilisateur" du système pour désigner l'entreprise qui exploite le système. La résolution du parlement européen du 20 octobre 2020 sur la responsabilité civile utilise le terme "opérateur aval". v. C. Mangematin, "Les propositions européennes visant à encadrer la responsabilité civile découlant de dommages causés par l'intelligence artificielle. Bien mais peut mieux faire!", Responsabilité Civile et Assurances, LexisNexis, n° 5, mai 2022, p. 7; J. Eynard, L'identification des acteurs dans le cycle de vie du système d'intelligence artificielle, Dalloz IP/IT 2022 71.
[40] Le contrôle humain exercé par le fournisseur de l'algorithme se limitera généralement au contrôle "système" dans l'élaboration de l'algorithme.
[41] H. Suresh, J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle", *Proceedings of EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO '21). ACM, New York 2021; David Lehr et Paul Ohm, Playing with the Data: What Legal Scholars Should Learn About Machine Learning, 51 U. CAL. DAVIS L. REV. 653 (2017).
[42] CNIL, « Comment permettre à l'homme de garder la main ?" op cit, p. 20.
[43] Infra, par. II-A-1.

constitutionnel français font référence à des “modèles et critères préétablis spécifiques", ce qui implique un contrôle humain dans la définition de ces critères.

Lorsqu’il s’agit en revanche d’un modèle d’apprentissage automatique (*machine learning*), l’algorithme d’apprentissage crée les règles de décision, mais l’humain devra quand même définir l’objectif à optimiser par l’algorithme d’apprentissage, sélectionner les données d’apprentissage, et veiller à leur nettoyage et étiquetage[44]. Lors des phases de test et de paramétrage, l’humain choisira les seuils de sensibilité, et la priorisation de certains types d’erreurs par rapport à d’autres, par exemple le taux de faux positifs *versus* faux négatifs[45]. L’humain devra également décider du caractère approprié de la technologie par rapport au problème à résoudre, et décider de son déploiement ou non. Le contrôle humain “système” pendant la phase 1 peut s’accompagner d’une étude d’impact et d’une consultation publique, comme nous le verrons ci-après[46]. Même si les contrôles humains système sont extrêmement importants pendant la phase de conception de l’algorithme, cette étude se penchera principalement sur le problème du contrôle humain des décisions individuelles pendant la phase 2.

2. La phase 2 : les contrôles des résultats algorithmiques pendant l’exploitation

Pendant la phase 2, les contrôles peuvent concerner une décision individuelle, un contrôle humain “individuel”, ou le fonctionnement de l’ensemble, un contrôle humain système. On se penchera en premier sur le contrôle humain individuel.

Le contrôle humain des décisions individuelles concerne l’examen des résultats algorithmiques dans un cas individuel : un score de crédit, un score de risque de récidive, un score de risque de terrorisme ou de fraude, la probabilité d’une concordance entre deux images, le choix d’une cible pour une attaque militaire, etc. sont des décisions auxquelles peuvent participer un algorithme, dont le processus doit connaître certains contrôles. Comment le contrôle humain s’opère-t-il lorsqu’il s’agit de contrôler ces résultats individuels ?

On peut diviser le contrôle individuel pendant la phase 2 en deux sous-groupes : le contrôle individuel avant la prise d’effet de la décision individuelle (contrôle individuel *ex ante*), et le contrôle individuel après la prise d’effet de la décision individuelle (contrôle individuel *ex post*)[47]. La distinction binaire entre contrôle individuel *ex ante* et contrôle individuel *ex post* est pertinente pour les systèmes qui génèrent un résultat algorithmique conduisant à une décision individuelle. Elle sera moins pertinente pour les systèmes dynamiques, qui génèrent

[44] L’étiquetage ne sera pertinent que pour l’apprentissage supervisé. Pour une présentation de l’ensemble des phases pour lesquelles un contrôle humain “système” est nécessaire, voy. D. Lehr, P. Ohm, “Playing with the Data: What Legal Scholars Should Learn About Machine Learning”, 51 *U. of California Davis law review*, 653, 2017; V. P. Besse, C. Castets-Renard, A. Garivier et J.-M. Loubes, L'IA du Quotidien peut-elle être Éthique ? Loyauté des Algorithmes d'Apprentissage Automatique, 2018, https://hal.archives-ouvertes.fr/hal-01886699v2.

[45] FRA (EU Agency for Fundamental Rights), Technologie de reconnaissance faciale : considérations relatives aux droits fondamentaux dans le contexte de l’application de la loi, 21 novembre 2019, p. 9-10.

[46] Infra, para.IV-C-3-d.

[47] CNCDH rapport du 7 avril 2022, op cit., point 66.

des résultats algorithmiques en continu. Pour ces systèmes dynamiques - une voiture autonome par exemple -, l'idée même d'un contrôle humain de chaque résultat algorithmique n'a pas de sens, car il n'existerait pas un seul résultat algorithmique et une seule décision à contrôler, mais un flux de résultats algorithmiques qui conduisent à une série d'actions. Seul un contrôle "système" serait approprié dans ce cas. La discussion qui suit partira donc de l'hypothèse d'un système plus statique, qui génère un seul résultat algorithmique, par exemple une alerte de blanchiment de capitaux, conduisant à une seule décision, le contrôle humain s'insérant soit avant (*ex ante*), soit après (*ex post*), la prise d'effet de la décision.

a. Le contrôle individuel *ex ante* en phase 2

Le contrôle individuel *ex ante* signifie que l'humain, non l'algorithme, prend la décision, en s'appuyant en tout ou en partie sur le résultat algorithmique. Grâce à cette intervention humaine, la décision ne sera pas considérée comme entièrement automatisée au sens de l'article 22 du RGPD[48]. Il s'agira d'un système d'aide à la décision, non un système de décision autonome[49]. Pour rappel, l'article 22 du RGPD concerne les décisions fondées "exclusivement" sur un traitement automatisé, donc aux décisions où il n'y a pas d'intervention humaine significative[50]. Dans notre classification, le contrôle individuel ex ante serait considéré comme une intervention humaine significative au sens du RGPD, même si l'efficacité de cette intervention sera, comme nous le verrons, très variable en fonction des circonstances. Néanmoins, l'idée de départ pour le contrôle humain individuel *ex ante* est que c'est l'humain, non la machine, qui "prend" la décision.

Au sein de la catégorie de contrôle individuel *ex ante*, on peut distinguer trois situations[51]. Dans la première, l'algorithme génère un élément qui contribue à éclairer le décisionnaire humain, mais le résultat algorithmique n'est pas central, et le décideur s'appuie sur d'autres éléments pour prendre sa décision. La "sortie" de l'algorithme ne se confondra pas avec la question posée au décisionnaire humain, mais sera l'un des éléments pris en compte par celui-ci[52]. Par exemple, la législation de certains états des Etats-Unis prévoit l'utilisation d'un système algorithmique pour aider le juge d'application des peines à évaluer le risque pour la société d'une mise en liberté conditionnelle, ou d'une solution alternative telle que le port d'un bracelet

[48] J. Rochfeld, "Données à caractère personnel - Droit de ne pas subir une décision fondée sur un traitement automatisé", D. Rép. IP/IT et Communication, mai 2020, point 12.

[49] La frontière entre une aide à la décision et une décision entièrement automatisée est souvent floue, voy. CNIL, « Comment permettre à l'homme de garder la main ?" op cit, p. 27.

[50] Groupe de Travail Art 29, Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679, 3 oct. 2017 révisée le 6 févr. 2018 WP 251rev.01 p. 23; J. Rochfeld, Droit de ne pas subir une décision fondée sur un traitement automatisé, op. cit.

[51] Cette classification est similaire à celle faite par l'ENA dans son rapport de juin 2019, voy. L'Ecole nationale de l'administration (l'ENA), Ethique et responsabilité des algorithmes publics, Rapport établi à la demande de la mission Etalab, Juin 2019, p. 29.

[52] En 1975, le Rapport Tricot a attiré l'attention des responsables de décisions "sur la nécessité de ne compter sur l'analyse de système que comme un instrument de travail parmi d'autres et de ne s'en remettre jamais à ses seules conclusions", Rapport Tricot, op cit., p. 83.

électronique[53]. L'objectif de ces mesures est de désengorger les prisons, et à également rendre ces décisions moins imprévisibles et subjectives. Un score de probabilité de récidive va figurer dans un dossier, ainsi que d'autres éléments nécessaires pour éclairer le décisionnaire. En théorie[54], le décisionnaire n'accordera donc pas un poids excessif au score algorithmique. Celui-ci sera considéré comme un élément de preuve parmi d'autres. On appellera cette première situation l'approche "tribunal", car le décisionnaire aura en théorie une variété de sources d'information, y compris celles éventuellement fournies par la personne affectée, et il aura le temps de les consulter. Les résultats algorithmiques étant un élément accessoire, le décisionnaire ne se laissera pas aveugler par leur présence. Un autre exemple de cette approche "tribunal" serait une proposition algorithmique de diagnostic médical à partir d'une image. Cette proposition serait confrontée à d'autres données du patient, ainsi que de l'avis des médecins. Elle ne serait pas l'élément central de la décision.

Dans la deuxième situation *ex ante*, l'algorithme génère un résultat qui constitue l'élément central de l'action à entreprendre, et le décisionnaire humain devra valider le résultat avant d'agir[55]. Pour effectuer cette validation - et c'est un élément clé de notre système de classification - le décisionnaire humain consultera d'autres informations contextuelles pour savoir si le résultat algorithmique est fondé. Dans le cas d'une alerte anti-blanchiment, par exemple, le décisionnaire va consulter les informations dans le dossier client pour contextualiser l'alerte algorithmique. La banque appellera peut-être le client pour demander une explication. Sur le fondement de ces informations supplémentaires, l'expert de la banque va décider de valider, ou de classer sans suite, l'alerte anti-blanchiment. On appellera cette deuxième situation l'approche "validation avec informations supplémentaires". Deux éléments clés distinguent cette situation de la première : d'abord, le temps de validation sera très limité, le contrôle humain s'insérant dans un système de décision algorithmique générant des dizaines voire des centaines d'alertes par jour. Ensuite, contrairement à la première situation, le résultat algorithmique sera l'élément déclencheur, et central, du processus de décision, non un élément accessoire.

Dans la troisième situation *ex ante*, l'algorithme génère un résultat qui constitue l'élément déterminant de l'action à adopter (comme dans la deuxième situation), et l'humain doit la valider sans consulter d'autres informations que celles utilisées par l'algorithme lui-même. Ce cas est rare, car généralement on ne va pas demander à un humain de refaire le travail de l'algorithme à partir des mêmes données. Cela serait contre-productif, sauf dans les situations rares où l'humain est aussi performant que l'ordinateur dans la tâche de calcul[56]. C'est justement le cas de la reconnaissance d'images issues de la vie courante. Un humain est généralement

[53] Le recours à des outils algorithmiques est interdit en France en application de l'article xx de la loi 78-17 du 6 janvier 1978. L. Huttner, op. cit. point 13.
[54] Ces situations sont simplifiées pour faciliter l'analyse. Les situations réelles seront plus complexes.
[55] C'est l'approche envisagée par la CJUE dans l'affaire *La Quadrature du Net*.
[56] Il peut exister d'autres situations où une intervention humaine sera souhaitable, par exemple pendant une période de rodage d'un nouveau système.

aussi performant qu'un ordinateur pour reconnaître l'image d'un chat, l'image d'une infraction routière, ou le visage d'une personne qu'il connaît. Il suffit de regarder le (ou les) image(s) et le cerveau humain tirera sa propre conclusion presque aussi rapidement que l'ordinateur. Cette validation humaine est loin d'être parfaite, comme nous le verrons dans la deuxième partie lorsque l'on évoquera les biais humains. Mais il s'agit de l'une des rares situations où le cerveau humain rivalise avec la performance de l'ordinateur dans l'analyse des mêmes données d'entrée. Dans d'autres situations, l'analyse d'opérations bancaires ou de données de connexion, l'humain aura du mal à tirer des conclusions des données d'entrée alimentant l'algorithme. Sans autres informations, le contrôle de l'humain sera inutile. On appellera cette troisième situation "validation sans informations supplémentaires".

Parmi ces trois situations de contrôle individuel *ex ante*, les deux premières, approche tribunal et validation avec informations supplémentaires, seront dépendantes de la qualité des autres informations disponibles pour le décideur humain. L'approche tribunal nécessitera l'organisation de moyens pour permettre à la personne affectée, et/ou à d'autres experts[57], d'apporter des informations et arguments supplémentaires. L'approche validation avec informations supplémentaires nécessitera la préparation d'informations consultables rapidement par le décisionnaire humain afin d'évaluer la cohérence du résultat algorithmique par rapport à ces autres informations. La deuxième et la troisième situations présenteront des défis importants en termes de biais d'automatisation, car le temps disponible au décideur humain sera très limité. Les biais d'automation peuvent se produire dans le contexte "tribunal" aussi, mais leurs effets seront contrebalancés par le temps de réflexion disponible au décideur, et l'apport d'informations ayant vocation à mettre en doute les résultats algorithmiques[58].

b. Le contrôle individuel *ex post* en phase 2

Le contrôle individuel *ex post* intervient toujours pendant la phase d'exploitation (phase 2), mais après la prise d'effet de la décision algorithmique, le plus souvent dans le cadre d'une contestation individuelle. Cette intervention *ex post* est prévue par l'article 22 du RGPD dans le cas des décisions entièrement automatisées, à savoir des décisions sans contrôle humain individuel *ex ante*[59]. Mais une contestation peut également survenir après une décision bénéficiant d'un contrôle humain individuel ex ante. Dans le cadre d'une contestation, le décisionnaire humain prendra en considération la conclusion algorithmique mais également d'autres informations fournies par la personne qui conteste la décision. L'intervention humaine dans le cadre d'une contestation pourrait ressembler à l'approche "tribunal" du contrôle individuel *ex ante*, sauf que le contrôle humain intervient après la prise d'effet de la

[57] Dans le domaine médical, une réunion staff permet de confronter le point de vue de différents spécialistes autour d'un problème. Le point de vue du patient ne sera pas nécessairement sollicité.
[58] Sur le biais cognitifs humains, voy. infra. par. III-B et suiv..
[59] Sur les conditions de l'article 22 RGPD, voy. § xx, supra, et J. Rochfeld, Droit de ne pas subir une décision fondée sur un traitement automatisé, op. cit.

décision, et seulement en cas de contestation. On appellera cette situation l'approche "contestation".

Certaines situations feront appel à plusieurs modalités de contrôle individuel. Une approche "validation avec informations supplémentaires" peut se transformer en approche "tribunal", si le contrôleur humain a un doute sur la situation et doit faire des recherches approfondies en consultant la personne affectée. De plus, la présence d'un contrôle individuel *ex ante* ne préjuge en rien la présence d'un contrôle individuel *ex post* en cas de contestation.

Les quatre approches de contrôle humain individuel se résument ainsi :

Trois types de contrôle individuel *ex ante*
L'approche "tribunal" : un décideur humain prend en considération une sortie algorithmique parmi d'autres éléments pour fonder une appréciation globale différente de ce que vise l'algorithme
L'approche "validation avec informations supplémentaires" : l'algorithme propose une décision, le décideur humain la valide ou pas en fonction d'autres informations à sa disposition
L'approche "validation sans informations supplémentaires" : l'algorithme propose une décision, le décideur humain la valide ou pas **sans** consulter d'autres informations
Un seul type de contrôle individuel *ex post*
L'approche "contestation": l'algorithme prend une décision avec effet automatique, l'humain réexamine la décision en cas de contestation

Tableau 1 : Les types de contrôle humain individuel en phase 2

c. Le contrôle "système" en phase 2

Le contrôle du type "*human-on-the-loop*" est un exemple de contrôle système qui a lieu pendant la phase d'exploitation : un humain surveille le fonctionnement de l'algorithme, et il a la capacité et l'autorité d'intervenir pour l'arrêter en cas d'anomalie[60]. L'approche *human-on-the-loop* est destinée à garantir qu'un humain pourra à tout moment intervenir, y compris pour arrêter le système, en cas de dysfonctionnement. L'approche *human-on-the-loop* a échoué tragiquement dans l'accident de la voiture autonome Uber en Arizona en 2018. Un conducteur humain était censé surveiller le fonctionnement du système de conduite autonome et

[60] Methnani et al., op cit.

intervenir en cas de problème. Mais l'humain est intervenue trop tard pour éviter la collision avec un piéton sur la voie[61].

3. Phase 3 : le contrôle humain lors des tests et audits ultérieurs

Après la mise en production, les tests et audits seront conduits pour détecter des biais, et ces tests et audits seront conduits par des humains[62]. Ces interventions système sont extrêmement importantes, contribuant à la contestabilité par conception (*contestability by design*) des systèmes algorithmiques[63]. Pour le système de détection de risques terroristes, la CJUE existe des vérifications à "intervalles réguliers" pour s'assurer que les critères du modèles restent pertinents et non-discriminatoires[64]. Des audits seront conduits, notamment pour vérifier l'impact sur les droits fondamentaux[65]. Nous appellerons ces tests et audits des mesures de contrôle "système" phase 3. Comme nous l'avons signalé[66], la présente étude porte principalement sur les contrôles humains individuels en phase 2, ce qui ne diminue en rien l'importance des contrôles "système", y compris en phase 3.

Les différentes phases du contrôle humain dans le cycle de vie d'un algorithme se présentent ainsi:

[61] NTSB (National Traffic Safety Bureau) Accident Report NTSB/HAR-1903, Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, 19 Nov. 2019, https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf

[62] CNIL, « Comment permettre à l'homme de garder la main ?" op cit, p. 20.

[63] Marco Almada. 2019. Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19). 2–11; J. Rochfeld, "Données à caractère personnel - Droit de ne pas subir une décision fondée sur un traitement automatisé", op. cit. point 27.

[64] CJUE, La ligue des droits humains, aff. C-817/19, point 201.

[65] Mökander, J., Morley, J., Taddeo, M. *et al.* Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Sci Eng Ethics* 27, 44 (2021). https://doi.org/10.1007/s11948-021-00319-4

[66] V. Introduction, supra..

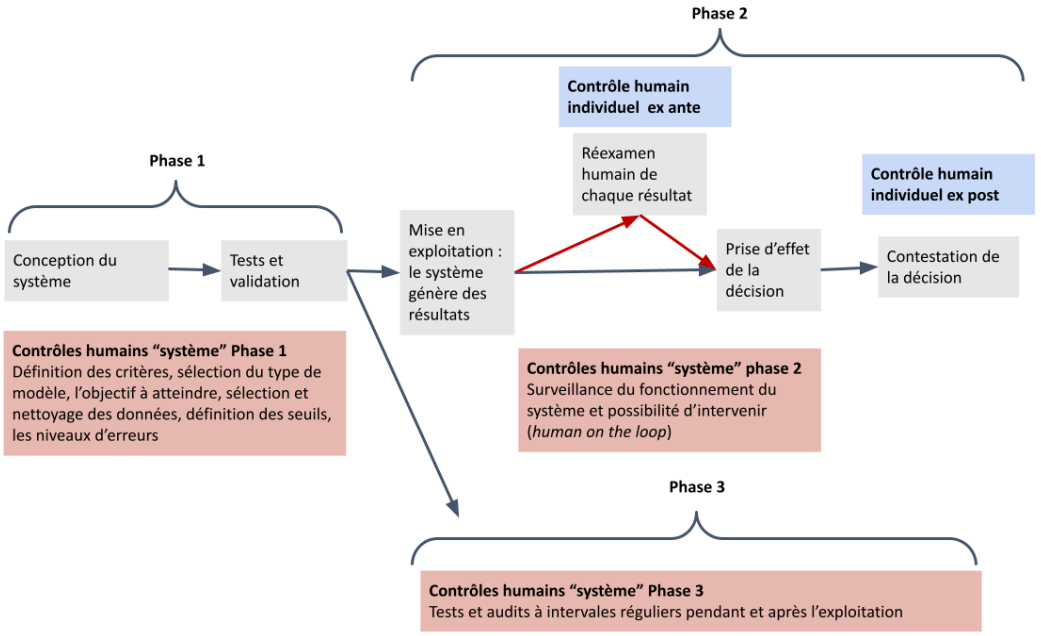

Schéma 1 : illustration des contrôles humains dans les différentes phases de vie d'un algorithme

II - LES EXIGENCES LÉGALES D'UN CONTRÔLE HUMAIN

Ayant établi une typologie des contrôles humains - contrôles système ou individuels, *ex ante* ou *ex post* - nous nous tournons maintenant vers les textes légaux qui imposent et encadrent un contrôle humain, que ce soit en phase 1, 2 ou 3. En général, ces textes ne précisent ni les modalités du contrôle humain, ni les finalités de celui-ci. On présume que le contrôle humain est bon, sans savoir pourquoi il est bon, ni comment le mettre en œuvre pour qu'il atteigne ses objectifs. L'ambition de cette étude est à l'inverse de ne plus présumer, et de se poser la question de la pertinence de ce contrôle et d'apporter une réponse partielle à ces questions.

Cette partie examinera les exigences légales qui imposent et encadrent le contrôle humain.
Cette partie présentera les différentes exigences légales en France et aux États-Unis d'un contrôle humain, en se concentrant d'abord sur les exigences d'un contrôle pendant la phase 1, celle de l'élaboration du système avant sa mise en exploitation (II-A). Ensuite, nous traiterons le cas des contrôles pendant la phase 2, celle de l'exploitation (II-B), et le cas des contrôles pendant la phase 3, celle des tests et audits en aval (II-C). La proposition de règlement IA sera traitée à part (II-D), car hormis le cas de l'identification biométrique, elle n'impose pas de mesure de contrôle humain spécifique, mais plutôt un cadre permettant de définir un contrôle humain effectif, qu'il soit "système" ou individuel. Enfin, la section (II-E) se penchera sur les exigences légales de l'explicabilité - la présentation d'informations sur le fonctionnement de l'algorithme - car cet aspect est important dans le contexte des valeurs procédurales que nous étudierons par la suite[67].

Cette description de l'environnement juridique autour du contrôle humain nous permettra à examiner plus en détail le "pourquoi" du contrôle humain dans les parties suivantes[68].

A. Les exigences d'un contrôle humain pendant la phase 1

La phase 1 concerne, on s'en souvient, l'élaboration de l'algorithme, la définition de son objectif, le choix du modèle, le choix et le nettoyage des données, et les tests de performance et de biais.

1. Le contrôle humain dans la définition des règles algorithmiques

En France, le législateur a précisé dans la loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles que dans le cadre de systèmes de décision entièrement automatisés, l'administration devait garder la maîtrise de l'algorithme et de ses évolutions[69]. Pour le Conseil constitutionnel, qui s'est prononcé sur la question dans sa décision n° 2018-765 DC du 12 juin 2018, cela signifie que les outils d'apprentissage automatique ne peuvent être utilisés pour les décisions entièrement automatisées[70] : les règles et critères

[67] V. infra partie IV.
[68] V. infra parties III, IV et V.
[69] L'article 47 de la loi n° 78-17 du 6 janvier 1978, modifiée par l'article 21 de la loi n° 2018-493 du 20 juin 2018..
[70] Décision n° 2018-765 DC du 12 juin 2018, point 71: "Il en résulte que ne peuvent être utilisés, comme fondement exclusif d'une décision administrative individuelle, des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement". V. A.

doivent être “définis à l’avance” par l’administration[71]. A défaut, il existait un risque de délégation illégale des pouvoirs de l’administration à une machine en contradiction avec l’article 21 de la Constitution[72]. Cela revient à exiger un contrôle humain système en amont, au stade de la définition des critères et paramètres de l’algorithme[73].

On retrouve cette même exigence d’un contrôle humain dans la définition des critères dans la directive PNR[74] et dans les décisions de la CJUE concernant l’utilisation d’algorithmes pour détecter des menaces terroristes[75]. Dans les deux cas, il ne s’agit pas de décisions entièrement automatisées, puisqu’un humain doit contrôler chaque résultat positif à sa sortie, avant la prise de décision. Malgré ce contrôle individuel *ex ante*, la directive PNR précise que le traitement automatique de données passagers s’appuie sur des “critères préétablis”[76]. Le texte de transposition de la directive 2016/681 en droit français précise que ces “critères sont définis en coopération avec les autorités mentionnées à l'article R. 232-15 [du code de la sécurité intérieure]. Ils doivent être ciblés, proportionnés, spécifiques aux infractions et non discriminatoires[77]”.
Dans son avis du 26 juillet 2017 sur l’accord PNR entre l’Union européenne et le Canada, la CJUE utilise les termes “modèles et critères préétablis spécifiques”[78]. Elle utilise les mêmes termes dans sa décision sur les algorithmes de détection de menaces terroristes dans son arrêt du 6 octobre 2021 *La Quadrature du Net*[79] et celui du 21 juin 2022 *La ligue des droits*

Debet, Une validation presque complète de la loi Informatique et Libertés par le Conseil constitutionnel (Partie I), CCE n° 9, sept. 2018, comm. 65.

[71] Ibid, point 69

[72] L’article 21 de la Constitution du 4 octobre 1958 stipule que Le Premier ministre dirige l'action du Gouvernement, qu’il exerce le pouvoir réglementaire et qu’il peut déléguer certains de ses pouvoirs aux ministres. Déléguer la prise de décision à une machine auto-apprenante, dont l’administration n’aurait pas la maîtrise, risquerait de constituer un abandon par l’administration de son pouvoir réglementaire, ainsi qu’une violation du principe de la publicité des règlements. V. commentaire du Conseil constitutionnel de la décision n° 2018-765 DC, p. 23-24
https://www.conseil-constitutionnel.fr/sites/default/files/as/root/bank_mm/decisions/2018765dc/2018765dc_ccc.pdf .

[73] Restrepo Amariles, op. cit., p. 286.

[74] Directive 2016/681 du 27 avril 2016 précitée.

[75] CJUE Accord PNR Canada avis 1/15 du 26 juillet 2017, c. V. Correia, À propos de l'avis 1/15 relatif au projet d'accord entre le Canada et l'Union européenne sur le transfert des données des passagers aériens, CCE n° 6, Juin 2018, étude 10, N. Le Bonniec, L'avis 1/15 de la CJUE relatif à l'accord PNR entre le Canada et l'Union européenne : une délicate conciliation entre sécurité nationale et sécurité numérique RTD eur. 2018. 617; CJUE La Quadrature du Net, affaires jointes C 511/18, C 512/18 et C 520/18, 6 octobre 2020, D. Berlin, La Cour de justice revient sur l'interdiction absolue des mesures générales de conservation et de traitement des données à caractère personnel, pour finalement en dresser le régime dérogatoire, La Semaine Juridique Edition Générale n° 48, 23 Novembre 2020, 1323, B. Bertrand, Chronique Droit européen du numérique - Les enjeux de la surveillance numérique RTD eur. 2021. 175, E. Daoud, I. Bello, O. Pecriaux, Données de connexion et sauvegarde de la sécurité nationale : l'exception confirme la règle, Dalloz IP/IT 2021. 46; CJUE, Ligue des droits de l’homme c. Conseil des ministres, aff. C-817/19, 21 juin 2022, D. Berlin, Conditions de transfert, de conservation et de traitement des données PNR, La Semaine Juridique Edition Générale n° 27, 11 Juillet 2022, 857.

[76] Directive 2016/681 du 27 avril 2016 précitée, art. 6(3)(b).

[77] Code de la sécurité intérieure, article R232-13; L. Huttner, “Données à caractère personnel - Décision automatisée et justice”, D. Rép. IP/IT et Communication, nov 2020, para. 32.

[78] CJUE avis 1/15 point 172; C. Castets-Renard, « L'accord PNR UE-Canada : validation conditionnelle par la CJUE », Dalloz IP/IT, N° 9, 2017, p. 420.

[79] CJUE La Quadrature du Net, affaires jointes C 511/18, C 512/18 et C 520/18, point 180.

humains[80]. Dans l'arrêt du 21 juin 2022, la CJUE apporte des précisions sur la signification des mots "modèles et critères préétablis spécifiques", en précisant notamment que le mot "préétabli" exclut l'utilisation d'un algorithme de *machine learning*[81]. Cette approche est cohérente avec celle du Conseil constitutionnel, qui a utilisé, dans sa décision du 12 juin 2018 les termes "critères définis à l'avance par le responsable du traitement" et a expressément exclu le recours à l'apprentissage automatique[82]. Cette exclusion de *machine learning* est renforcée, selon la Cour, par le fait qu'un algorithme de *machine learning* risquerait de priver d'effet utile le réexamen individuel des concordances positives ainsi que le contrôle de licéité requis[83]. A cause de l'opacité de l'algorithme d'apprentissage automatique, il serait impossible de comprendre la raison du signalement[84]. La CJUE relève également que les critères ne doivent en aucun cas s'appuyer sur "l'origine raciale ou ethnique d'une personne, ses opinions politiques, sa religion ou ses convictions philosophiques, son appartenance à un syndicat, son état de santé, sa vie sexuelle ou son orientation sexuelle[85]". L'exploitant des algorithmes doit pouvoir démontrer que les modèles et critères préétablis sont "ciblés, proportionnés et spécifiques"[86]. Les critères préétablis doivent tenir compte tant des éléments "à charge" que des éléments "à décharge"[87]. Ainsi, les critères préétablis doivent incorporer des règles qualitatives définies par les humains responsables du système, ce qui exclurait un recours au *machine learning*.[88] S'il existait un doute sur la compatibilité entre le *machine learning* et les décisions nécessitant un contrôle de légalité, ce doute semble levé par l'arrêt *La ligue des droits humains* du 21 juin 2022.

2. Les exigences de contrôle humain par rapport à d'autres aspects d'élaboration de l'algorithme

En matière de contrôle humain dans l'élaboration de l'algorithme, on peut mentionner la loi de l'Etat de Washington du 31 mars 2020 sur la reconnaissance faciale, qui prévoit la conduite de tests et l'élaboration d'un rapport de responsabilisation (accountability report) soumis à consultation publique[89]. Cette approche est également prévue par la proposition de loi *Algorithmic Accountability Act of 2022*, qui exigerait une étude d'impact, des tests de performance et de biais, et une consultation des parties intéressées[90]. Cette proposition de loi obligerait le responsable du système à justifier pourquoi le recours au système est

[80] CJUE, 21 juin 2022, Ligue des droits de l'homme c. Conseil des ministres, aff. C-817/19.
[81] Ibid., point 194.
[82] Décision n° 2018-765 DC du 12 juin 2018, point 69.
[83] CJUE, 21 juin 2022, Ligue des droits de l'homme c. Conseil des ministres, aff. C-817/19, point 195.
[84] Ibid.
[85] Ibid, point 196.
[86] Ibid, point 198.
[87] Ibid, point 200.
[88] V. Beaudouin, I. Bloch, F. d'Alché-Buc, D. Bounie, J. Eagan, W. Maxwell, S. Clémençon, P. Mozharovskyi, J. Parekh, "Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach" SSRN 3559477, 2020.
[89] Revised Code of Washington (RCW) Chapter 43.386; DLAPiper, In Washington State's landmark facial recognition law, public sector practices come under scrutiny and regulation, April 22, 2020 https://www.dlapiper.com/en/us/insights/publications/2020/04/in-washington-states-landmark-facial-recognition-law-public-sector-practices-come-under-scrutiny/
[90] Proposition de loi S. 3572 Algorithmic Accountability Act of 2022 des Sénateurs Wyden, Booker et Clarke, 3 février 2022.

nécessaire, et notamment en quoi il apporte des bénéfices supplémentaires par rapport aux processus existants.

L'obligation de tests et d'une étude d'impact en phase 1 est également présente dans la proposition de règlement européen sur l'IA, que nous examinerons à part[91].

B. Les exigences de contrôle humain pendant la phase 2

Le contrôle humain en phase d'exploitation, la phase 2, et plus particulièrement le contrôle humain individuel, est au cœur des questions abordées dans cette étude. Nous examinerons successivement, pour la phase 2, les exigences en matière de contrôle individuel ex ante, à savoir avant la prise d'effet de la décision (i), ensuite les exigences en matière de contrôle individuel ex post, à savoir après la prise d'effet de la décision (ii), et enfin les exigences en matière de contrôle "système" pendant la phase d'exploitation (iii).

1. Les exigences d'un contrôle humain individuel ex ante

a. Décisions concernant la reconnaissance faciale

Aux États-Unis, la loi du 31 mars 2020 de l'État de Washington sur la reconnaissance faciale[92] prévoit un examen humain significatif (*meaningful human review*) à l'égard de décisions algorithmiques fondées sur la reconnaissance faciale à but d'identification ayant des effets juridiques ou des effets significatifs similaires à l'égard d'un individu[93]. Un examen humain significatif existe aux yeux de cette loi si (i) une ou plusieurs personnes physiques examinent ou supervisent les décisions algorithmiques, (ii) ces personnes ont reçu une formation spécifique sur le contrôle de décisions automatiques de reconnaissance faciale et notamment sur les limitations et risques de ces systèmes et comment leurs résultats doivent être vérifiés, et (iii) les personnes disposent d'un pouvoir de modifier les recommandations algorithmiques[94]. La loi ne précise pas que l'examen humain doit nécessairement intervenir *avant* la prise d'effet d'une décision fondée sur la reconnaissance faciale, mais le caractère ex ante du contrôle semble évident compte tenu du contexte. D'une part, la loi indique que les décisions ayant un impact sur les droits des individus "seront soumises" à un examen humain, ce qui suggère que l'examen devra intervenir pour l'ensemble des décisions, et pas seulement celles qui font l'objet d'une contestation. D'autre part, une intervention humaine sera nécessaire de toute façon pour entreprendre une mesure découlant de la reconnaissance faciale, un mandat d'amener, par exemple[95]. La loi de l'État de Washington définit le contrôle humain comme "un examen ou une supervision" (*review or oversight*) par une ou plusieurs personnes physiques. Le mot "examen"

[91] V. infra, par. II-E.
[92] Revised Code of Washington (RCW) Chapter 43.386.
[93] La loi utilise les mêmes termes que l'article 22 du RGPD en ce qui concerne l'impact de la décision.
[94] Revised Code of Washington (RCW) § 43.386.010.
[95] La loi de l'Etat de Washington précise que les résultats de reconnaissance faciale ne peuvent jamais suffire à eux seuls à justifier une perquisition ou autre mesure d'enquête nécessitant l'autorisation d'un juge sur le fondement d'un faisceau de preuves concordantes (*probable cause*). RCW §43.386.090(5).

suggère un rôle actif : la validation individuelle de chaque résultat algorithmique par l'humain. Le mot "supervision" a une connotation plus passive, l'humain pouvant se contenter de regarder passer les résultats algorithmiques et intervenir en cas de désaccord. Il n'existe pas à ma connaissance de lignes directrices sur l'application de la loi de l'État de Washington qui nous éclairerait sur ce point. Cependant, la formation obligatoire des personnes chargées du contrôle doit couvrir notamment "les procédures pour interpréter et agir sur les résultats produits par le service de reconnaissance faciale[96]". Les mots "interpréter et agir" suggèrent un rôle actif de la personne responsable du contrôle, non un rôle de simple spectateur.

Le projet de règlement européen sur l'intelligence artificielle, que nous examinerons plus loin[97], contient néanmoins une disposition précise sur le contrôle humain individuel ex ante qu'il convient de mentionner ici. La proposition de règlement prévoit qu'aucune mesure ou décision ne pourra être prises par l'utilisateur sur la base de l'identification résultant d'un système d'identification biométrique sans vérification et confirmation par au moins deux personnes physiques[98]. Les mots "vérifier et confirmer" utilisés dans la proposition de règlement suggèrent un rôle actif, comparable au rôle actif suggéré par les mots "interpréter et agir" utilisateur dans la loi de l'Etat de Washington. Le projet de règlement européen va plus loin que la loi de l'État de Washington en imposant la vérification par au moins deux personnes. Le projet de règlement ne prévoit pas, en revanche, une formation spécifique des personnes effectuant ces vérifications, même si une formation peut être prévue par le fournisseur du système dans le cadre de son système de gestion des risques[99].

b. Algorithmes de détection de menaces terroristes

Nous avons vu dans la section précédente que la directive PNR 2016/681 ainsi que les trois décisions de la CJUE[100] concernant les algorithmes de détection de risques terroristes imposent un contrôle humain système en amont, pour définir les modèles et critères "préétablis" de l'algorithme. La directive et les trois décisions de la CJUE imposent également un contrôle humain individuel de chaque résultat positif en aval, à la sortie de l'algorithme, avant que l'autorité ne prenne d'autres mesures par rapport aux risques identifiés. La directive PNR 2016/681 prévoit que "toute concordance positive obtenue à la suite du traitement automatisé des données PNR … est réexaminée individuellement par des moyens non automatisés[101]". Cette disposition est transposée en droit français à l'article R232-13 du Code de la sécurité intérieure qui précise que "toute concordance positive obtenue à la suite de

[96] RCW §43.386.060 "Procedures to interpret and act on the output of the facial recognition service"
[97] infra, par. II-D.
[98] Proposition de règlement européen sur l'intelligence artificielle, art. 14(5).
[99] L'article 9(4)(c) de la proposition de règlement prévoit que la formation des utilisateurs sera un facteur pris en compte dans le système de gestion des risques mis en place par le fournisseur.
[100] CJUE accord PNR Canada Avis 1/15 du 26 juillet 2017; La Quadrature du Net affaires jointes C 511/18, C 512/18 et C 520/18 du 6 octobre; et Ligue des droits de l'homme, affaire C-817/19 du 21 juin 2022.
[101] Directive 2016/681, art. 6(6).

l'évaluation réalisée au titre du présent article est réexaminée individuellement par des moyens non automatisés avant transmission[102]".

Alors que la CJUE disait peu sur les conditions du contrôle humain individuel e*x ante* dans ses arrêts du 26 juillet 2017 (accord PNR Canada) et du 6 octobre 2020 (La Quadrature du Net), elle a livré plus de précisions dans son arrêt du 21 juin 2022 (Ligue des droits de l'homme) sur le "pourquoi" et le "comment" du contrôle humain individuel *ex ante*. Sur le "pourquoi", la CJUE indique que le contrôle individuel *ex ante* est destiné à "déceler, dans toute la mesure du possible, l'existence éventuelle de 'faux positifs'", ainsi qu'à "exclure d'éventuels résultats discriminatoires".[103] Ce deuxième objectif, détecter les discriminations, est nouveau par rapport aux objectifs cités par la CJUE dans ses deux arrêts précédents du 26 juillet 2017 et du 6 octobre 2020. Dans ces arrêts précédents, la CJUE a mentionné uniquement l'objectif de contrôle des faux positifs, alors que dans son arrêt du 21 juin 2022, la Cour cite également la détection de résultats discriminatoires. Nous examinerons ces objectifs dans la troisième partie[104].

Sur le "comment" du contrôle individuel ex ante, l'arrêt du 21 juin 2022 donne quelques indications sur le niveau de diligences attendu des contrôleurs humains. Ceux-ci doivent pouvoir confirmer l'existence "d'éléments de nature à fonder, à suffisance de droit, un soupçon raisonnable de participation à des infractions terroristes ou à des formes graves de criminalité"[105]. Si ces éléments ne sont pas confirmés par les contrôleurs, ou s'il existe des éléments indiquant que les traitements conduisent à des résultats discriminatoires, l'autorité ne doit pas agir sur le signalement[106]. La CJUE semble exiger que les humains en charge du contrôle humain puissent vérifier l'existence d'éléments de soupçon indépendamment de l'algorithme, comme s'il s'agissait d'une enquête classique. Pour permettre cette prise de responsabilité, le contrôleur humain doit pouvoir "comprendre la raison" du signalement, ce qui exclut, selon la Cour, le recours à du *machine learning*[107].

Quant à l'encadrement de ce contrôle, la CJUE indique que "les Etats membres doivent prévoir des règles claires et précises de nature à guider et à encadrer l'analyse effectuée par les agents en charge du réexamen individuel, aux fins d'assurer le plein respect des droits fondamentaux consacrés aux articles 7, 8 et 21 de la Charte et, notamment de garantir une pratique administrative cohérente au sein de l'UIP respectant le principe de non-discrimination"[108]. Compte tenu du nombre élevé de faux positifs, la CJUE impose aux autorités en charge du contrôle de "définir des critères de réexamen objectifs permettant à ses agents de vérifier, d'une part, si et dans quelle mesure une concordance positive (*hit*) concerne effectivement

[102] L. Huttner, "Données à caractère personnel - Décision automatisée et justice", D. Rép. IP/IT et Communication, nov 2020, para. 32.
[103] CJUE, Ligue des droits humains, affaire C-817/19, point 203.
[104] infra, partie III-A.
[105] CJUE Ligue des droits humains, affaire C-817/19, point 204.
[106] Ibid.
[107] Ibid., point 195.
[108] Ibid. point 205.

un individu qui est susceptible d'être impliqué dans les infractions terroristes ou les formes graves de criminalité… ainsi que, d'autre part, le caractère non-discriminatoire des traitements, et notamment des critères préétablis, et des bases de données utilisées[109]." En clair, les contrôleurs humains doivent suivre une procédure bien précise pour valider une alerte. Ces procédures opérationnelles doivent prévoir que le résultat du réexamen individuel prévaudra sur le signalement automatique[110]. Pour permettre la traçabilité et la responsabilisation, chaque contrôle humain doit faire l'objet d'un compte rendu, pour permettre un contrôle éventuel du contrôle humain[111].

En résumé, la CJUE considère que les deux objectifs du contrôle humain individuel *ex ante* sont, d'une part, de réduire les faux positifs et, d'autre part, de vérifier l'absence de discriminations. Comme nous le verrons ci-après[112], la capacité pour un humain d'effectuer ces deux tâches de contrôle n'est pas évidente. La CJUE semble avoir pris conscience de ces difficultés pratiques puisqu'elle impose aux autorités en charge de ce contrôle l'obligation d'élaborer des règles claires et précises pour guider et encadrer les agents humains, ainsi qu'un système pour assurer la traçabilité de ces contrôles.

c. Algorithmes de détection et de retrait de contenus terroristes en ligne

Le règlement européen 2021/784 du 21 avril 2021 relatif à la lutte contre la diffusion des contenus à caractère terroriste en ligne impose aux hébergeurs une obligation de mettre en place des "mesures spécifiques appropriées, efficaces et proportionnées" pour lutter contre la diffusion de ces contenus[113]. Bien que le règlement n'impose pas le recours à des algorithmes de détection automatique de ces contenus, le règlement encourage néanmoins l'utilisation de mesures techniques. Mais pour réduire les faux positifs liés à l'utilisation de ces mesures, le règlement impose une surveillance et une vérification humaines pour s'assurer de l'exactitude des signalements et éviter le retrait de matériel qui ne constitue pas un contenu à caractère terroriste[114]. Les modalités de cette surveillance et vérification humaines ne sont pas précisées dans le règlement, mais il s'agit bien d'un contrôle *ex ante* dont l'objectif est d'éviter les retraits injustifiés (faux positifs)[115].

---

[109] Ibid., point 206.
[110] Ibid., point 208.
[111] Ibid. point 207.
[112] infra, partie III-B.
[113] Règlement 2021/784 du 29 avril 2021 relatif à la lutte contre la diffusion des contenus à caractère terroriste en ligne, art. 5(2); W. Maxwell, "The GDPR and private sector measures to detect criminal activity" *Revue des Affaires européennes,* Bruylant / Larcier, 2021.
[114] Règlement 2021/784, art. 5(3) et considérant 23. L'article 5(3) précise que "lorsque les mesures spécifiques impliquent le recours à des mesures techniques, des garanties appropriées et efficaces, notamment au moyen d'une surveillance et d'une vérification humaines, sont prévues pour s'assurer de l'exactitude et éviter le retrait de matériel qui ne constitue pas un contenu à caractère terroriste."
[115] E. Dreyer, "Présentation rapide du Règlement (UE) 2021/784 du 29 avril 2021 relatif à la lutte contre la diffusion des contenus à caractère terroriste en ligne", Dalloz IP/IT 2021 p.527.

d. Décisions judiciaires ou administratives s'appuyant sur des résultats algorithmiques

Une décision judiciaire qui s'appuie sur un résultat algorithmique peut être considérée comme une forme de contrôle humain individuel ex ante des résultats algorithmiques. Il en est de même pour une décision administrative, qui s'appuie en tout ou en partie sur un résultat algorithmique. Dans les deux cas, l'algorithme reste une aide à la décision, et c'est l'humain qui prend la décision, selon une approche "tribunal"[116].

En France, l'article 47 de la loi du 6 janvier 1978 interdit la prise en considération par un juge d'une évaluation algorithmique des aspects de la personnalité de la personne si l'affaire implique une appréciation sur le comportement d'une personne [117]. Le recours à un outil de prédiction de récidive du type COMPAS semble donc exclu. Pour les décisions administratives, la loi française permet aux agents de l'administration de recourir aux algorithmes, mais impose des conditions strictes de transparence qui nous examinerons par la suite[118].

Aux États-Unis la prise en considération de résultats algorithmiques par un juge ou par un agent de l'administration dans le cadre d'une décision individuelle a été examinée sous l'angle du principe constitutionnel de *due process*. Le principe de *due process* s'applique à toute décision prise par une autorité publique, qu'elle soit administrative ou judiciaire, qui conduit à la perte d'un droit. Il impose une notification préalable à la personne affectée des détails de la décision envisagée, et une possibilité pour celle-ci de bénéficier d'une "audience" (*hearing*) pour présenter ses arguments[119]. La compatibilité d'outils de décision algorithmiques avec les exigences de *due process* a été longuement débattue par la doctrine[120]. L'une des questions principales est le niveau d'intervention humaine nécessaire pour satisfaire à l'exigence constitutionnelle d'une "audience" (*hearing*)[121]. Selon le professeur Aziz Huq, la décision de la Cour Suprême des États-Unis dans l'affaire *Mathews c. Eldridge*

[116] Etalab, Guide des Algorithmes Publics, §1 https://etalab.github.io/algorithmes-publics/guide.html.
[117] L. Huttner, "Données à caractère personnel - Décision automatisée et justice", D. Rép. IP/IT et Communication, nov 2020, points 12 et 20.
[118] infra, par. II-E.
[119] É. Zoller, « Procès équitable et due process », Rec. Dalloz, 2007; Pascal MBONGO, Procès équitable et Due Process of Law, Nouveaux cahiers du Conseil Constitutionnel N° 44 (Le Conseil constitutionnel et le procès équitable) - Juin 2014
[120] Voy. notamment D. K. Citron, « Technological due process », *Washington university law review*, vol. 85, nº 6, 2008, p. 1249 ; K. Crawford, J. Schultz, « Big data and due process: toward a framework to redress predictive privacy harms », *Boston college law review*, vol. 55, 2014, p. 93 ; C. Coglianese, D. Lehr, « Regulating by robot: administrative decision making in the machine-learning era », *The Georgetown law journal*, vol. 105, nº 5, 2017, p. 1147 ; E. Berman, « A government of laws and not of machines », *Boston university law review*, vol. 98, 2018, p. 1277 ; A. Z. Huq, « Constitutional rights in the machine learning state », *Cornell law review*, vol. 105, 2020, p. 1875 ; A. Z. Huq, « A right to a human decision », *Virginia law review*, vol. 106, 2020, p. 611.
[121] H. J. Friendly, « Some kind of hearing », *University of Pennsylvania law review*, vol. 123, 1975, p. 1267

[122] admet la possibilité de se dispenser entièrement d'une intervention humaine dans certains cas mineurs[123].

Deux affaires aux États-Unis illustrent comment un résultat algorithmique s'intègre dans une décision judiciaire ou administrative – une approche du type "tribunal" dans notre système de classification – et comment le principe constitutionnel de *due process* s'applique à ces cas. Nous nous situons bien dans des cas de contrôle individuel ex ante, avec des décisions humaines intervenant avant la prise d'effet de la décision algorithmique. La tâche du décisionnaire humain est donc d'accorder un poids approprié au résultat algorithmique et d'en tirer les conséquences en fonction des autres éléments de preuve à sa disposition[124].

L'affaire *State c. Loomis*[125] jugée par la cour suprême de l'état du Wisconsin le 13 juillet 2016 est surtout connue pour son traitement de la question des biais algorithmiques dont souffre l'outil statistique COMPAS[126] qui prédit le risque de récidive. En ce qui concerne la question du contrôle humain individuel, l'affaire *Loomis* nous éclaire sur les modalités d'utilisation par un juge des résultats d'un outil algorithmique imparfait et opaque, et la compatibilité de cette utilisation avec le principe de *due process*. L'affaire Loomis concerne la décision d'un juge de condamner M. Loomis à une peine de prison après un accord conclu entre M. Loomis et le procureur. Dans cet accord M. Loomis a admis sa culpabilité pour des crimes relativement mineurs comparés aux crimes violents (tentative de meurtre en bande) pour lesquels il a été arrêté. Il avait été prévenu que compte tenu des circonstances, la peine maximale pour ces crimes mineurs serait appliquée. Le juge en charge de la fixation des peines a consulté de nombreux éléments dans le dossier de Loomis, y compris un score algorithmique concernant son risque de commettre un autre crime s'il était relâché. Mais la motivation du juge dans la fixation des peines s'est appuyée sur les éléments non-algorithmiques, et a simplement mentionné que les conclusions découlant de ces autres éléments étaient cohérentes avec les scores algorithmiques. Loomis a fait appel, estimant que le score algorithmique dans le dossier a violé ses droits à *due process*.

---

[122] *Mathews v. Eldridge*, 424 U.S. 319, 1976

[123] A. Z. Huq, « A right to a human decision », art. cit.

[124] Selon Meg Leta Jones, la décision de la cour suprême de l'Etat de Wisconsin dans l'affaire *Loomis* impose le principe d'un contrôle humain individuel ex ante ("*human in the loop*"): voy. M. Leta Jones, "The Right to a Human in the Loop: Political Constructions of Computer Automation & Personhood", 47 *Soc. Stud. Sci.* 216 (2017).

[125] 881 N.W. 2d 749 (Wis. 2016), cert. denied, 137 S.Ct. 2290 (2017); pour une description complète de l'affaire, voy. W. Maxwell, "La régulation des algorithmes aux États-Unis : quelles leçons pour l'Europe ?", in Bertrand B. (dir.), *La politique européenne du numérique*, Bruxelles, Bruylant, 2022, à paraître; E. Marique, A. Strowel, Gouverner par la loi ou les algorithmes : de la norme générale de comportement au guidage rapproché des conduites, Dalloz IP/IT 2017. 517; Y. Meneceur, Les systèmes judiciaires européens à l'épreuve du développement de l'intelligence artificielle,Revue pratique de la prospective et de l'innovation n° 2, Octobre 2018, dossier 7.

[126] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), voy. C. Castets-Renard, « L'IA en pratique : la police prédictive aux États-Unis », Dalloz IP/IT, N° 5, 2019, p. 314; R; Pons, L. Risser, Biais et discriminations dans les systèmes d'intelligence artificielle, Dalloz IP/IT 2022. 75; L. Huttner, op. cit., point 9.

La décision de la cour suprême de l'État de Wisconsin prend acte des faiblesses de l'algorithme, du fait de l'existence de certains biais statistiques et du caractère opaque des modèles. La cour exige que l'utilisation de tels outils soit accompagnée d'une mise en garde sur leurs faiblesses. La cour précise également qu'un juge ne pourra jamais, sans violer les droits de *due process* de la personne condamnée, fonder une décision de fixation de peine sur un score algorithmique en tant qu'élément déterminant. En revanche, elle admet que des résultats algorithmiques peuvent éclairer la décision du juge dès lors que celui-ci s'appuie sur d'autres éléments du dossier pour motiver sa décision. En d'autres termes, la cour laisse au juge une marge d'appréciation sur le poids à accorder à un résultat algorithmique dans la totalité des éléments de preuve à sa disposition.

Dans l'affaire *Loomis*, le résultat algorithmique avait un caractère accessoire par rapport à l'ensemble des autres éléments de preuve dans le dossier. Ce ne sera pas le cas dans la seconde affaire importante, l'affaire *Houston Federation of Teachers*[127] décidée le 4 mai 2017 par le tribunal fédéral de district du Texas. Là, le résultat algorithmique avait un poids plus important. L'algorithme calculait des scores de performance pour les enseignants de la ville de Houston, et ces scores étaient pris en considération pour augmentation de salaire, et pouvaient même conduire à des licenciements. Le syndicat des enseignants a contesté l'utilisation des scores devant le tribunal fédéral du Texas, et celui-ci a décidé que l'utilisation des scores violait les droits de *due process* des enseignants parce que ceux-ci n'avaient pas la possibilité vérifier l'absence d'erreurs. Dans sa décision du 4 mai 2017, le tribunal fédéral s'est appuyé sur la jurisprudence en matière de tests anti-dopage, une jurisprudence qui exige que la personne bénéficie de la possibilité de demander un deuxième test à partir du même échantillon d'urine pour vérifier l'absence d'erreur dans les résultats [128].

Ces deux affaires, *Loomis* et *Houston Federation of Teachers,* montrent qu'en matière de *due process*, l'utilisation d'un résultat algorithmique par un juge ou par un agent de l'administration dans une décision individuelle devra s'entourer des mêmes précautions que d'autres éléments de preuve scientifique, des tests anti-dopage, par exemple. Ces précautions comprennent notamment la possibilité de répliquer les résultats algorithmiques pour vérifier l'absence d'erreurs. De plus, en matière judiciaire, le juge bénéficie d'une marge d'appréciation sur le poids à accorder à ces éléments de preuve, dès lors que le juge est informé des limitations de l'algorithme. Dans le cas *Loomis*, la cour suprême de Wisconsin était convaincue que la présence du score COMPAS n'avait rien changé à la décision du juge. C'est pour cette raison que la cour a considéré que les droits de *due process* de Monsieur Loomis n'avaient

[127] *Houston Federation of Teachers v. Houston Ind't School Dist.,* 251 F. Supp. 3d 1168 (S.D. Tex. 2017) ; M. A. Paige, A. Amrein-Beardsley, « "Houston, we have a lawsuit": a cautionary tale for the implementation of value-added models for high-stakes employment decisions », *Educational researcher*, vol. 49, n° 5, 2020, p. 350-359; K. Crawford, J. Schultz, AI Systems as State Actors, 119 Columbia L. Rev. 1941 (2019); W. Maxwell, "La régulation des algorithmes aux États-Unis : quelles leçons pour l'Europe ?", op cit.

[128] Dans l'affaire *K.W. v. Armstrong*, 180 F. Supp. 3d 703 (D. Idaho 2016), le tribunal a également estimé que la possibilité de vérifier l'absence d'erreurs est un élément essentiel de *due process.* Restrepo Amariles, op. cit., p. 290.

pas été violés, ce qui n'aurait pas été le cas si la recommandation avait été un facteur déterminant dans la décision.

La décision de la cour suprême de Wisconsin dans l'affaire *Loomis* soulève alors la question du poids d'une prédiction algorithmique dans la prise de décision judiciaire. Les juges sont considérés comme des experts de la preuve, capables d'évaluer la valeur probante d'un élément de preuve scientifique et d'accorder à cet élément le poids qu'il mérite. Si le juge est pleinement informé des limites d'un outil statistique et ne fonde pas sa décision uniquement sur les conclusions de l'outil mais se contente de l'utiliser pour éclairer d'autres éléments du dossier, pourquoi interdire son utilisation ? C'est le point de vue de la cour suprême de l'État de Wisconsin.

Du point de vue de la protection des droits individuels, l'approche "tribunal" illustrée par l'affaire *Loomis*, et recommandée en 1975 par le Rapport Tricot "informatique et les libertés"[129], est préférable aux autres modes de contrôle humain, car l'algorithme apporte simplement un éclairage statistique à la question posée à l'humain. Selon ce modèle, le décideur humain prend en considération d'autres éléments non-statistiques, et est parfaitement informé des limitations de l'outil statistique. Le décideur n'accorde pas, en théorie, un poids excessif à ses prédictions. Hélas, ce modèle de prise de décision apparaît réservé aux situations où le temps de décision est long : un procès, ou un diagnostic médical complexe, par exemple, qui permet un échange collégial. Dans beaucoup de situations où le recours à un algorithme est nécessaire, le temps de décision sera court, et le volume de décisions élevé. L'approche "tribunal" pour un contrôle individuel ex ante ne sera donc pas possible.

2. Les exigence d'un contrôle individuel *ex post*

Pendant la phase d'exploitation (phase 2), le contrôle humain individuel peut intervenir après la prise d'effet d'une décision individuelle. Ce contrôle individuel ex post se déclenche généralement à la suite d'une contestation. Cette forme de contrôle humain est exigée par plusieurs textes.

a. Droit de contestation prévu par les textes sur la protection des données à caractère personnel

Le contrôle individuel ex post est imposé à travers le droit de contestation prévu par l'article 22 du RGPD et l'article 9 de la Convention 108+ pour les décisions entièrement automatisées. L'article 22 du RGPD impose une interdiction générale de

[129] Rapport Tricot, op. cit., p. 15, qui recommande que l'humain utilise l'informatique "pour éclairer et préparer sa décision, non pour prendre celle-ci".

décisions entièrement automatisées[130] et prévoit ensuite trois dérogations[131]. Chacune de ces dérogations doit s'accompagner de mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée. Pour la deuxième dérogation, à savoir les décisions entièrement automatisées prévues par la loi, la dérogation dispose que la loi doit prévoir des mesures appropriées, mais cet alinéa du RGPD s'abstient d'indiquer le contenu de ces mesures. En revanche, pour la première et la troisième dérogation, à savoir une décision entièrement automatisée nécessaire à la conclusion où à l'exécution d'un contrat, ou une décision entièrement automatisée fondée sur le consentement, le RGPD nous dit que les mesures appropriées doivent inclure au moins le droit pour la personne concernée d'obtenir une intervention humaine de la part du responsable de traitement, d'exprimer son point de vue et de contester la décision. Les mesures appropriées contiennent donc trois éléments : une intervention humaine de la part du responsable du traitement, la possibilité pour la personne concernée d'exprimer son point de vue, et enfin la possibilité pour celle-ci de contester la décision[132]. En outre, le RGPD prévoit la fourniture d'informations utiles sur la logique sous-jacente, un point qui sera examiné plus loin[133].

Le droit de contestation *ex post* est également prévu par la nouvelle version de la Convention 108 du Conseil de l'Europe pour la protection des personnes à l'égard du traitement des données à caractère personnel du 18 mai 2018[134]. L'article 9(a) de la convention dispose que toute personne a le droit de ne pas être soumise à une décision l'affectant de manière significative, qui serait prise uniquement sur le fondement d'un traitement automatisé de données, sans que son point de vue soit pris en compte.[135] Les commentaires officiels qui accompagnent la Convention 108+ indiquent que l'article 9(a) est destiné à fournir aux individus un droit de contester la décision algorithmique en faisant valoir de manière effective leurs points de vue et leurs arguments. Les individus doivent avoir la possibilité de prouver l'inexactitude éventuelle des données à caractère personnel avant l'utilisation, l'inadéquation du profil qu'il est prévu d'appliquer à leur situation particulière ou d'autres facteurs qui auront un impact sur le résultat de la décision automatisée.[136] Ce droit de contestation permet à la personne notamment de montrer que le profil algorithmique généré par l'algorithme n'est pas adapté à sa situation personnelle.

---

[130] Ce point est débattu dans la doctrine en raison du langage ambigu de l'article 22, mais la position du groupe de travail article 29 est qu'il s'agit d'une interdiction de principe. Groupe de travail Art. 29 Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679, WP251, Adoptées le 3 octobre 2017 Version révisée et adoptée le 6 février 2018; J. Rochfeld, "Données à caractère personnel - Droit de ne pas subir une décision fondée sur un traitement automatisé", D. Rép. IP/IT et Communication, mai 2020, para. 12

[131] J. Rochfeld, "Données à caractère personnel - Droit de ne pas subir une décision fondée sur un traitement automatisé", op. cit.

[132] Ibid., point 25.

[133] infra., par. II-E.

[134] Conseil de l'Europe, Protocole d'amendement à la Convention pour la protection des personnes à l'égard du traitement des données à caractère personnel, adopté par le Comité des Ministres lors de sa 128e session à Elseneur, le 18 mai 2018 (la "Convention 108+").

[135] Convention 108+, article 9(a).

[136] Rapport explicatif de la Convention 108+, para. 75.

Cela suppose que la personne puisse avoir connaissance du profil et des données d'entrée sur laquelle il repose. Les commentaires officiels concernant la Convention 108+ donnent plus d'indications sur les modalités du contrôle individuel ex post que celles que l'on peut glaner du RGPD et des lignes directrices du Groupe de Travail Article 29[137]. Ces commentaires contiennent cependant une incohérence, car ils indiquent que l'individu doit pouvoir prouver l'inexactitude éventuelle des données avant leur utilisation. Cela semble impossible car la contestation intervient forcément après la prise de décision, donc après l'utilisation des données. La référence à "l'inadéquation du profil" dans les commentaires officiels est également remarquable, car cela suggère qu'un individu aurait le droit de ne pas être évalué uniquement sur le fondement d'un profil statistique, ou d'exiger un profil plus adapté à sa situation individuelle. Il est fréquent, notamment dans le domaine de l'assurance, d'appliquer un traitement à un individu en raison de son profil statistique[138]. Il semble peu probable que la Convention 108+ confère aux individus un nouveau droit de contester une prime d'assurance en raison de la supposée inadéquation du profil avec sa situation personnelle.

b. Un droit de contestation en matière de retrait automatisé de contenus en ligne

Une grande partie des retraits de contenus en ligne s'effectue sans intervention humaine, par l'application d'algorithmes[139]. La loi française n° 2021-1109 du 24 août 2021 confortant le respect des principes de la République confère aux utilisateurs de réseaux sociaux le droit de contester le retrait de contenus ou la suspension d'un compte[140]. Lorsqu'une plateforme bloque ou rend inaccessible un contenu, la plateforme doit informer l'utilisateur à l'origine de la publication, en indiquant les raisons qui ont motivé cette décision, en précisant si la décision a été prise au moyen d'un outil automatisé, et en l'informant des voies de recours internes et judiciaires. Les plateformes doivent mettre en place des dispositifs de recours internes permettant aux utilisateurs de contester la décision. Le traitement des recours ne doit pas s'appuyer uniquement sur l'utilisation de moyens automatisés.

La proposition de règlement européen sur les services numériques[141], dit "*Digital Services Act*" ou "DSA", prévoit également un système de notification aux personnes ayant posté le contenu et un droit de contestation impliquant nécessairement l'intervention d'un humain. Le projet de règlement prévoit à son article 17 que les fournisseurs de plateforme doivent mettre en place un système interne de traitement des réclamations permettant à tout utilisateur de contester une décision de blocage ou de retrait. L'article 17(5) du projet européen dispose que ce système

[137] Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage, WP251 précité.
[138] L. Barry, A. Charpentier, "L'équité de l'apprentissage machine en assurance" op. cit.; F. Schauer, "Profiles, Probabilities and Stereotypes", op cit.
[139] Cambridge Consultants, Use of AI in Online Content Moderation, Report produced on behalf of OFCOM, 2019; W. Maxwell, Applying Net neutrality rules to social media content moderation systems, Enjeux numériques, Annales des Mines, 2022.
[140] Article 42 de la loi n° 2021-1109 insérant un nouvel article 62 dans la loi n° 86-1067 du 30 septembre 1986.
[141] Proposition de règlement du Parlement européen et du Conseil relatif à un marché intérieur des services numériques (Législation sur les services numériques) et modifiant la directive 2000/31/CE, SEC(2020) 432 final.

interne de réclamation ne peut pas s'appuyer uniquement sur des moyens automatisés. Le contrôle individuel ex post s'impose, donc, même si la détection et les décisions de retrait initiales peuvent être effectuées par un algorithme.

Le droit de contestation devant un décideur humain est également prévu par les "Santa Clara Principles", adoptés en 2018 et révisés en 2021 par quatre ONG américaines dédiées à la défense des droits et libertés individuels en ligne.[142] Bien qu'ils n'aient aucun caractère contraignant et ne soient pas encore appliqués par les grandes plateformes, les Santa Clara Principles méritent notre attention parce qu'ils exigent un moyen de contestation 'ayant du sens' (*meaningful*), comprenant une analyse par une ou plusieurs personnes (humaines) non-impliquées dans la décision initiale, l'idée étant de fournir une certaine indépendance par rapport à la décision initiale. Cette indépendance du décisionnaire rejoint les recommandations du Conseil de l'Europe du 8 avril 2020[143] et la résolution du Parlement européen du 20 octobre 2020, qui prévoient toutes les deux un examen humain impartial[144].

c. Le droit pour les travailleurs de plateforme, après une décision, d'avoir une discussion avec un représentant humain

La proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme prévoit la possibilité pour un travailleur, après une décision nourrie par un algorithme, d'avoir une discussion avec un représentant humain de la plateforme afin de "clarifier les faits, les circonstances, et les raisons qui ont conduit à la décision"[145]. Le représentant humain de la plateforme devra avoir les compétences, la formation, et l'autorité pour exercer ses fonctions[146].

C. Phase 3 : Les exigences légales d'un contrôle humain "système" lors de tests et audits

Le lecteur se rappellera que la phase 3 concerne les tests et audits du système qui ont lieu en parallèle de l'exploitation du système, ou après l'exploitation.

Les décisions de la CJUE concernant les algorithmes de prédiction de risques terroristes exigent un contrôle "système" après la mise en exploitation et la génération de résultats, pour vérifier la fiabilité et l'actualité des modèles et des critères préétablis, ainsi que des bases de données utilisées. Selon la

[142] Electronic Frontier Foundation, American Civil Liberties Union Foundation for Northern California, Center for Democracy and Technology, Open Technology Institute, The Santa Clara Principles: On Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org; voy. également Frederick Mostert, 'Digital due process': a need for online justice, Journal of Intellectual Property Law & Practice, Volume 15, Issue 5, May 2020, Pages 378–389, https://doi.org/10.1093/jiplp/jpaa024

[143] Conseil de l'Europe, Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme, 8 avril 2020, paragraphe C-4.4 qui prévoit que les réclamations par rapport à un traitement algorithmique devraient faire l'objet d'un examen impartial et indépendant

[144] Résolution du Parlement européen du 20 octobre 2020, op cit, point 35: garantir l'examen humain impartial de toutes les réclamations.

[145] Proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme 9 décembre 2021 COM(2021) 762 final, art. 8(1).

[146] Idem.

Cour, ces modèles, critères et bases de données doivent faire l'objet de "réexamen régulier"[147], ce qui se traduit par une obligation de contrôle humain système tout au long de la vie de l'algorithme. Dans sa décision du 21 juin 2022, la CJUE confirme que le réexamen régulier des critères préétablis doit servir à les actualiser en fonction de l'évolution dans la lutte contre le terrorisme et les formes graves de criminalité, et de l'expérience acquise dans leur application, l'objectif étant de vérifier le caractère strictement nécessaire de l'application de ces critères[148].

La proposition de règlement européen sur l'IA, que nous examinerons plus en détail dans la section suivante, prévoit la mise en place de contrôles tout au long du cycle de vie du système, donc des contrôles systèmes pendant la phase d'exploitation. Il en est de même pour la proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme, qui imposerait un contrôle humain continu des décisions algorithmiques pour s'assurer que le système algorithmique ne crée pas de risques pour la santé physique ou mentale de travailleurs[149]. Cette proposition précise que les personnes en charge du contrôle devront disposer des compétences, de la formation et de l'autorité nécessaires à ses fonctions. Elles devront être protégées contre des licenciements, des sanctions disciplinaires, ou d'autres traitements défavorables en raison de leur rejet de décisions automatiques[150].

D. La proposition de règlement européen sur l'IA (AI Act)

La proposition de règlement européen sur l'IA impose aux fournisseurs de systèmes d'IA à haut risque une série d'obligations liées notamment à la gouvernance des données et à la mise en place d'un système de gestion des risques[151]. Ces mesures seront prises avant la mise sur le marché du système, et tout au long de son cycle de vie. Par ce biais, la proposition de règlement prévoit une série de contrôles humains "système", à la fois en phase 1 (élaboration et tests avant exploitation), 2 (exploitation) et 3 (tests et audits en opération). Notamment en ce qui concerne le contrôle "système" phase 2, la proposition de règlement précise que la personne en charge du contrôle humain doit disposer d'un bouton d'arrêt[152].

En ce qui concerne le contrôle humain individuel, la proposition de règlement n'impose aucune obligation spécifique hormis le cas de l'identification biométrique en temps réel, qui nécessite une vérification et une confirmation par au moins deux personnes avant la mise en œuvre de la décision [153]. La proposition est étonnamment silencieuse par rapport à d'autres situations où un contrôle humain individuel s'imposerait. Sur le "pourquoi" du contrôle humain, la proposition confirme bien qu'il s'agit d'une mesure pour prévenir ou à réduire au minimum les risques pour la santé, la sécurité

[147] CJUE avis 1/15 point 173; affaires jointes C 511/18, C 512/18 et C 520/18 point 182; affaire C-817/19, point 201..
[148] CJUE affaire C-817/19, point 201.
[149] Proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme 9 décembre 2021 COM(2021) 762 final, art. 7(3).
[150] Ibid.
[151] Proposition de règlement européen sur l'IA, art. 9; sur la proposition de règlement, v. C. Castets-Renard, Quel droit de l'intelligence artificielle dans l'Union européenne ? Ou les multiples ambitions normatives de l'AI Act, Dalloz IP/IT 2022. 6; C. Castets-Renard, Quelle politique européenne de l'intelligence artificielle ? RTD eur. 2021. 297, G. Marti, L. Cluzel-Métayer et S. Merabet, Chronique Intelligence artificielle - Droit et Intelligence artificielle - La Semaine Juridique Edition Générale n° 26, 28 Juin 2021, doctr. 720..
[152] Proposition de règlement européen sur l'IA, art. 14(4).
[153] Ibid, art. 14(5).

ou les droits fondamentaux, notamment lorsque d'autres mesures n'élimineront pas entièrement les risques[154]. Le contrôle humain serait une forme de voiture balai qui attraperait les risques résiduels qui passeraient à travers les autres mesures de protection.

Sur le "comment" du contrôle humain, la proposition indique seulement que les outils et interfaces du système doivent permettre aux personnes en charge du contrôle humain d'être en capacité de comprendre et agir sur le système. En particulier, les outils prévus par le fournisseur doivent leur permettre de "d'appréhender totalement les capacités et les limites" du système, de surveiller correctement son fonctionnement, afin de pouvoir détecter et traiter dès que possible les signes d'anomalies, de dysfonctionnements et de performances inattendues. Les outils et interfaces doivent permettre aux personnes d'avoir conscience des risques liés aux biais d'automatisation, d'être en mesure d'interpréter correctement les résultats compte tenu des outils des méthodes d'interprétation disponibles, et de décider, dans une situation particulière, de ne pas utiliser le système d'IA à haut risque ou de négliger, passer outre ou inverser le résultat fourni par ce système. Enfin, le système doit permettre aux personnes en charge du contrôle humain de décider de passer outre ou inverser les résultats, et d'intervenir sur le système y compris au moyen d'un bouton d'arrêt [155]. La proposition de règlement prévoit, en résumé, l'installation de manettes et de jauges permettant à un humain d'effectuer un contrôle effectif. Pour certains auteurs, la proposition de règlement imposerait des exigences irréalisables en matière d'interfaces. Comment comment permettre à un contrôleur humain "d'appréhender totalement" les capacités et les limites du système[156]? La proposition de règlement semble tomber dans le piège de surestimer les capacités de l'humain, et partant, de surestimer le pouvoir du contrôle humain.

Le manque de détails sur le "comment" du contrôle humain est compréhensible, puisque le contrôle dépendra du contexte de déploiement de chaque système. Il incombera au fournisseur du système de prévoir les modalités du contrôle humain en fonction des risques, et de les mettre dans la notice d'utilisation. Néanmoins, la proposition de règlement aurait pu donner une grille d'analyse en précisant, par exemple, qu'un contrôle humain individuel *ex ante* serait nécessaire pour toute situation comportant des risques similaires à ceux de l'identification biométrique à distance pour laquelle un tel contrôle est déjà exigé par le règlement. La jurisprudence de la CJUE[157] nous donnent des indications sur le type de contrôle humain nécessaire pour les systèmes de détection de risques de criminalité grave et de terrorisme. Dans son dernier arrêt du 21 juin 2022, la CJUE va jusqu'à nous donner des exigences sur les mécanismes de contrôle du contrôle humain : procédures de contrôle claires, documentation de chaque instance de contrôle humain permettant une traçabilité et un contrôle de la licéité du contrôle humain.

L'utilisateur, pourtant en première ligne pour effectuer le contrôle humain, n'a actuellement aucune obligation directe au titre de la proposition de règlement. Ses seules obligations en matière de contrôle humain sont de suivre les indications du fournisseur[158]. Et même là, l'utilisateur dispose

[154] Ibid., art. 14(2).
[155] Proposition de règlement européen sur l'IA, art. 14(4).
[156] Ebers, M.; Hoch, V.R.S.;Rosenkranz, F.; Ruschemeier, H.; Steinrötter, B. The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). J 2021, 4, 589–603, p. 597.
[157] Infra, par. II-B-1-b.
[158] Ibid, art. 29(1).

d'une marge de manoeuvre pour "organiser ses propres ressources et activités aux fins de la mise en œuvre des mesures de contrôle humain indiquées par le fournisseur"[159].

Nous effectuerons quelques recommandations à la fin de cette étude[160] pour améliorer la proposition de règlement.

E. Les exigences d'explicabilité des algorithmes

Lorsque nous aborderons les biais humains en partie II, et les valeurs de procédure ainsi que les problèmes liés à la participation de l'individu en partie III, la question de l'explicabilité des algorithmes se posera. En prévision de ces discussions, un rapide résumé des exigences légales d'explicabilité s'impose.

Le RGPD prévoit, pour les décisions entièrement automatisées, la mise à disposition d'informations "utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée"[161]. Le champ d'application de ce droit est limité aux décisions entièrement automatisées, et le contenu des informations devant être transmises n'est pas clair[162]. Pour certains, il s'agit uniquement d'informations sur le fonctionnement de l'algorithme en général[163]. Pour d'autres, il s'agit d'informations sur la décision individuelle[164]. Seule cette dernière interprétation permettrait à la personne de prendre connaissance des facteurs individuels contribuant à la décision algorithmique et de contester utilement celui-ci[165]. L'article 9(1)(c) de la Convention 108+ du Conseil de l'Europe prévoit la communication du raisonnement qui sous-tend le traitement[166]. Le rapport explicatif de la convention précise que les personnes ont le droit non seulement de prendre connaissance du raisonnement qui sous-tend le traitement, mais également des conséquences de ce raisonnement et des conclusions qui peuvent en avoir été tirées, en particulier lors de l'utilisation d'algorithmes pour une prise de décision automatisée, notamment dans le cadre du profilage.[167] Selon ce rapport, la connaissance du raisonnement utilisé par l'algorithme contribue à l'exercice effectif d'autres garanties essentielles comme le droit d'opposition et le droit de recours auprès de l'autorité compétente.[168] Le champ d'application du droit à

---

[159] Ibid, art. 29(2).
[160] Infra, partie VI.
[161] Règlement nº 2016/679/UE du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive nº 95/46/CE (RGPD), arts. 13, 14 et 15.
[162] J. Rochfeld, "Données à caractère personnel - Droit de ne pas subir une décision fondée sur un traitement automatisé", op. cit. point 23.
[163] Wachter, S., B. Mittelstadt, and L. Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." International Data Privacy Law 7 (2): 76–99.
[164] Groupe de travail Art. 29 Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679, WP251, Adoptées le 3 octobre 2017 Version révisée et adoptée le 6 février 2018, p. 29 qui donne l'exemple d'informations devant être fournies dans le cadre d'un octroi de prêt, et p. 36.
[165] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. International Data Privacy Law, 7, 4, 233–242.
[166] Conseil de l'Europe, Protocole d'amendement à la Convention pour la protection des personnes à l'égard du traitement des données à caractère personnel, adopté par le Comité des Ministres lors de sa 128e session à Elseneur, le 18 mai 2018 (la "Convention 108+"), art. 9(1)(c).
[167] Rapport explicatif de la Convention 108+, para. 77.
[168] Ibid.

l'information prévu par l'article 9(c) de la Convention 108+ est plus large que celui prévu par le RGPD, puisqu'il s'applique à tout traitement algorithmique, même si ce traitement n'est pas l'unique fondement de la décision.

Dans sa décision du 21 juin 2022 sur les données PNR, la CJUE exige, d'une part, que les personnes chargées du contrôle humain individuel ex ante puisse comprendre les raisons d'un signalement de risque terroriste[169]. D'autre part, les individus visés par ces signalements doivent pouvoir comprendre le fonctionnement des critères et programmes de manière à décider, en pleine connaissance de cause, d'introduire ou non un recours[170]. En cas de recours juridictionnel, le juge et l'intéressé[171] doivent pouvoir prendre connaissance de l'ensemble des motifs et éléments de preuve sur la base desquels la décision était prise, y compris des critères d'évaluation préétablis et du fonctionnement des programmes appliquant ces critères[172].

Pour les décisions administratives prises sur le fondement d'un résultat algorithmique, les articles L.311-1-3 et R311-3-1-2 du Code des relations entre le public et l'administration prévoient une communication, sous une forme intelligible, d'informations sur la décision individuelle, y compris les données traitées et leurs sources, le degré et le mode de contribution du traitement algorithmique à la prise de décision, les paramètres du traitement et leur pondération appliqués à la situation de l'intéressé[173]. Les articles L.300-2 et L.300-3 prévoient aussi une communication des codes sources, considérés comme un document administratif[174].

L'accès général à l'information a par ailleurs été examiné par la CrEDH dans l'affaire *Sigurdur Einarsson c. Islande*[175] et par le Tribunal de District de la Haye dans l'affaire *SyRI*[176]. Dans la première affaire, la CrEDH a estimé que les défendeurs devaient avoir accès aux mêmes outils informatiques pour le tri de documents que ceux dont disposait l'office du procureur, afin de respecter l'égalité des armes. Dans la seconde affaire, le tribunal de district de la Haye a estimé que l'absence d'information faite aux personnes visées par un score algorithmique contribuait à rendre le dispositif incompatible avec l'article 8 de la CEDH, et notamment l'exigence de proportionnalité. L'absence d'information rendait inefficace le droit de contester le rapport algorithmique.

Aux Etats-Unis plusieurs décisions ont imposé le principe d'accès à l'information au titre du droit à *procedural due process*. Dans l'affaire *Houston Federation of Teachers*, le tribunal a ainsi exigé que les personnes ciblées par un score algorithmique puissent tester l'algorithme afin de vérifier l'absence

[169] CJUE, Ligue des droits humains, affaire 817/19, point 195.
[170] Ibid, point. 210.
[171] L'accès par l'intéressé aux informations peut être limité en cas de menaces pour la sûreté de l'Etat, Ibid.
[172] Ibid., point 211.
[173] Code des relations entre le public et l'administration, arts. L.311-1-3 et R311-3-1-2; J. Rochfeld, Droit de ne pas subir une décision fondée sur un traitement automatisé, op. cit., point 5.
[174] Ibid. arts. L.300-2 et L.300-3; L'Ecole nationale de l'administration (l'ENA), Ethique et responsabilité des algorithmes publics, Rapport établi à la demande de la mission Etalab, Juin 2019.
[175] CEDH, Sigurdur Einarsson a. o. v. Iceland, n. º 39757/15, 4 June 2019, p 4.
[176] Nederlands Juristen Comité voor de Mensenrechten et autres c. les Pays Bas, Tribunal de district de la Haye, affaire n° C/09/550982 / HA ZA 18-388, 5 février 2020. Une version anglaise de la décision est disponible sur le site rechtspraak.nl: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878.

d'erreurs, et la réplicabilité des scores individuelles[177]. Les exigences de *due process* ne vont pas jusqu'à imposer, néanmoins, une communication des codes sources de l'algorithme, protégés par le secret des affaires[178]. Il s'agit plutôt d'un droit de connaître toutes les données d'entrée afin de vérifier leur exactitude, et et de s'assurer de la réplicabilité des résultats[179]. Or, le droit de connaître les données d'entrée ne permet pas de remettre en cause l'approche méthodologique de l'algorithme lui-même, dès lors que les limitations, y compris les biais, de l'algorithme sont portés à l'attention du décideur[180].

Ces lois et décisions de justice en Europe et aux Etats-Unis prévoient en définitive un accès à différents types d'information : les données d'entrée, les paramètres et leur pondération dans une décision individuelle, les informations sur le fonctionnement général de l'algorithme, les codes sources, et l'accès à l'algorithme pour vérifier la réplicabilité des résultats. Une transparence totale peut se heurter naturellement à d'autres principes, tels que le secret des délibérations du jury[181], la nécessité de protéger une enquête en cours[182], la sûreté de l'Etat[183], ou d'autres secrets protégés par la loi[184]. Le principe de l'accès à l'information souffrira d'exceptions, mais ces exceptions devront être justifiées, limitées à ce qui est strictement nécessaire, et entourées d'autres garanties afin de respecter le principe de proportionnalité[185].

S'il s'agit d'une décision individuelle qui s'appuie même en partie sur un procédé algorithmique, l'administration a l'obligation de communiquer à l'intéressé, à sa demande, les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé[186]. Pour un modèle issu de l'apprentissage machine, la pondération exacte des paramètres restera un mystère. Grâce à des méthodes d'explicabilité post hoc telles que SHAP, il sera néanmoins possible de visualiser les poids approximatifs des différents paramètres. Mais cette estimation restera une évaluation approximative. On peut dès lors se poser la question de la compatibilité de modèles issus de l'apprentissage automatique avec les obligations de transparence imposées par le Code des

---

[177] 251 F. Supp. 3d 1168 (S.D. Tex. 2017); Paige MA, Amrein-Beardsley A. "Houston, We Have a Lawsuit": A Cautionary Tale for the Implementation of Value-Added Models for High-Stakes Employment Decisions. Educational Researcher. 2020;49(5):350-359.

[178] State v. Loomis, 881 N.W. 2d 749 (Wis. 2016), cert. denied, 137 S.Ct. 2290 (2017).

[179] *K.W. v. Armstrong*, 180 F. Supp. 3d 703 (D. Idaho 2016), la possibilité de vérifier l'absence d'erreurs est un élément essentiel de *due process.* Restrepo-Amariles, David, Algorithmic Decision Systems: Automation and Machine Learning in the Public Administration (30 nov 2020). The Cambridge Handbook of the Law of Algorithms, SSRN: 3974564, p. 290.

[180] State v. Loomis, précité..

[181] Conseil constitutionnel, décision n° 2020-834 QPC du 3 avril 2020, point 13, sur l'accès aux algorithmes locaux dans l'affaire Parcoursup.

[182] Affaires jointes C-511/18, C-512/18 et C-520/18, La Quadrature du Net, 6 octobre 2020, point 191. Voy. également Règlement (UE) 2021/784 du Parlement européen et du Conseil du 29 avril 2021 relatif à la lutte contre la diffusion des contenus à caractère terroriste en ligne, art. 11(3) qui limite la fourniture d'informations lorsque la limitation est nécessaire pour protéger une enquête en cours.

[183] CJUE Ligue des droits humains, affaire 817/19, point 211.

[184] Code des relations entre le public et les administrations, art. L 311-3-1, L311-5, RGPD, considérant 63..

[185] Conseil constitutionnel, Commentaire de la décision n° 2020-834 QPC du 3 avril 2020, p. 23-26; voy. également Commission nationale consultative des droits de l'homme (CNCDH), Avis relatif à l'impact de l'intelligence artificielle sur les droits fondamentaux, Avis A-2022-6 du 7 avril 2022, para. 76, sur la nécessité de limiter strictement les exceptions à la transparence,

[186] Code des relations entre le public et l'administration, art. R311-3-1-2; J. Rochfeld, Droit de ne pas subir une décision fondée sur un traitement automatisé, op. cit., point 9.

relations entre le public et l'administration. Ni le guide sur les algorithmes publics d'Etalab[187], ni le rapport de l'ENA sur les algorithmes publics[188], n'aborde le sujet.

[187] Etalab, Guide des Algorithmes Publics, https://etalab.github.io/algorithmes-publics/guide.html
[188] L'Ecole nationale de l'administration (l'ENA), Ethique et responsabilité des algorithmes publics, Rapport établi à la demande de la mission Etalab, Juin 2019.

III - LE PREMIER OBJECTIF DU CONTRÔLE HUMAN : DÉTECTER LES ERREURS ET LES DISCRIMINATIONS

Le premier objectif du contrôle humain est de détecter des erreurs et des discriminations. Réduire les faux positifs est l'objectif mentionné expressément par la CJUE dans les affaires *La Quadrature du Net* et *l'accord PNR Canada* pour justifier un contrôle individuel *ex ante* de chaque alerte algorithmique concernant un risque de terrorisme. Dans sa décision du 21 juin 2022 dans l'affaire *La ligue des droits humains*, la CJUE a réitéré l'objectif de réduire les faux positifs, mais a également mentionné l'objectif de détecter des discriminations[189]. Le règlement européen sur la lutte contre les contenus terroristes en ligne vise la réduction de faux positifs comme objectif du contrôle humain. Si le futur règlement européen AI Act prévoit un contrôle individuel *ex ante* de chaque résultat positif d'un système de reconnaissance faciale, c'est bien dans le but de réduire le risque de faux positifs.

Quelles sont ces erreurs et ces discriminations algorithmiques contre lesquelles le contrôle humain doit nous protéger ? L'objectif de cette partie est d'examiner le concept d'erreur algorithmique, et de développer une typologie de ces erreurs afin de mieux comprendre l'utilité d'un contrôle humain pour chaque type d'erreur (III-A). La discrimination sera considérée comme un type d'erreur. Ensuite seront présentés les biais cognitifs humains qui peuvent constituer un obstacle à la détection d'erreurs (III-B).

A. Une classification des erreurs algorithmiques

1. Le compromis entre faux positifs et faux négatifs

Cette section présentera une liste des différents types d'erreurs algorithmiques générés par un système d'IA. L'objectif est de comprendre les différents types d'erreurs pour mieux analyser la pertinence du contrôle humain dans leur détection.

Les systèmes d'IA fournissent généralement une prédiction, soit en forme de notation (par exemple un score de 88/100[190]), soit en forme de classification (éligible pour le prêt/non-éligible). Dans un but de simplification, on peut se concentrer sur les algorithmes de classification binaire, et les différents types d'erreurs qui peuvent en découler. Pour ces algorithmes, les erreurs algorithmiques sont soit des faux négatifs, à savoir une prédiction négative (par exemple, "cette opération bancaire n'est pas suspicieuse") qui en réalité devrait être positive, soit un faux positif, à savoir une prédiction positive (par exemple, "cette opération bancaire est suspicieuse") qui en réalité ne devrait pas l'être.

L'objectif du contrôle humain sera souvent de contrôler les décisions positives ("cette opération bancaire est suspicieuse") pour détecter, si possible, des faux positifs[191]. Il sera plus difficile de contrôler les négatifs car cela reviendrait à analyser toutes les observations analysées par l'algorithme, y compris les observations qui n'ont pas conduit à une

[189] CJUE aff. 817/19, point 203.
[190] En langage d'IA, il s'agit d'un modèle de régression.
[191] La détection de faux positifs a été citée expressément par la CJUE dans sa décision du 6 octobre 2020 La Quadrature du Net, aff. jointes C-511/18, C-512/18 et C-520/18, point 182.

classification positive. Dans certains cas individuels, par exemple dans une décision d'accorder un prêt, un humain pourrait analyser la décision négative ("vous n'êtes pas éligible"). Mais dans la plupart des cas, les négatifs, et *a fortiori* les faux négatifs, seront trop nombreux pour contrôler de manière systématique. Le contrôle humain individuel, s'il a lieu, se concentrera donc sur les décisions positives dans l'objectif d'identifier les faux positifs.

Pour un algorithme de classification, la performance prédictive est représentée graphiquement par une courbe ROC - *receiver operating characteristics*[192]. En général, plus l'espace en dessous de la courbe[193] est important, plus l'algorithme sera performant dans ses prédictions. La sensibilité de l'algorithme par rapport aux deux types d'erreurs, les faux positifs et les faux négatifs, pourra être ajustée par le concepteur, voire l'utilisateur, de l'algorithme. On peut demander au modèle de réduire le nombre de faux positifs, mais cela conduira à augmenter le nombre de faux négatifs, et *vice versa*. Le bon compromis entre le taux de faux positifs et le taux de faux négatifs dépendra du contexte. Par exemple, dans un système de reconnaissance faciale pour retrouver un enfant disparu, on pourrait tolérer un taux de faux positifs plus élevé que dans un système destiné à localiser un fugitif. Chaque cas d'usage nécessitera un arbitrage qui tiendra compte des préjudices découlant des deux types d'erreurs[194]. Il est hélas impossible de réduire les deux types d'erreurs en même temps.

Le diagramme qui suit illustre le caractère de vases communicantes entre le taux de faux positifs et faux négatifs. Le choix du seuil Beta dans le diagramme conduira soit à augmenter le taux de faux positifs (FP), soit à le diminuer, mais entraînera un effet, en sens inverse, sur le taux de faux négatifs (FN).

[192] Alaa Tharwat, "Classification assessment methods", *Applied Computing and Informatics*, Vol. 17 No. 1, 2021 pp. 168-192

[193] AUC - *Area Under Curve.*

[194] FRA (EU Agency for Fundamental Rights), Technologie de reconnaissance faciale, op cit., p 9-10; S. Clémençon, W. Maxwell, "Why facial recognition algorithms can't be perfectly fair", *The Conversation*, 20 juillet 2020; A. D. Selbst, D. Boyd, S. Friedler, S. Venkatasubramanian, J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems", *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, 2019.

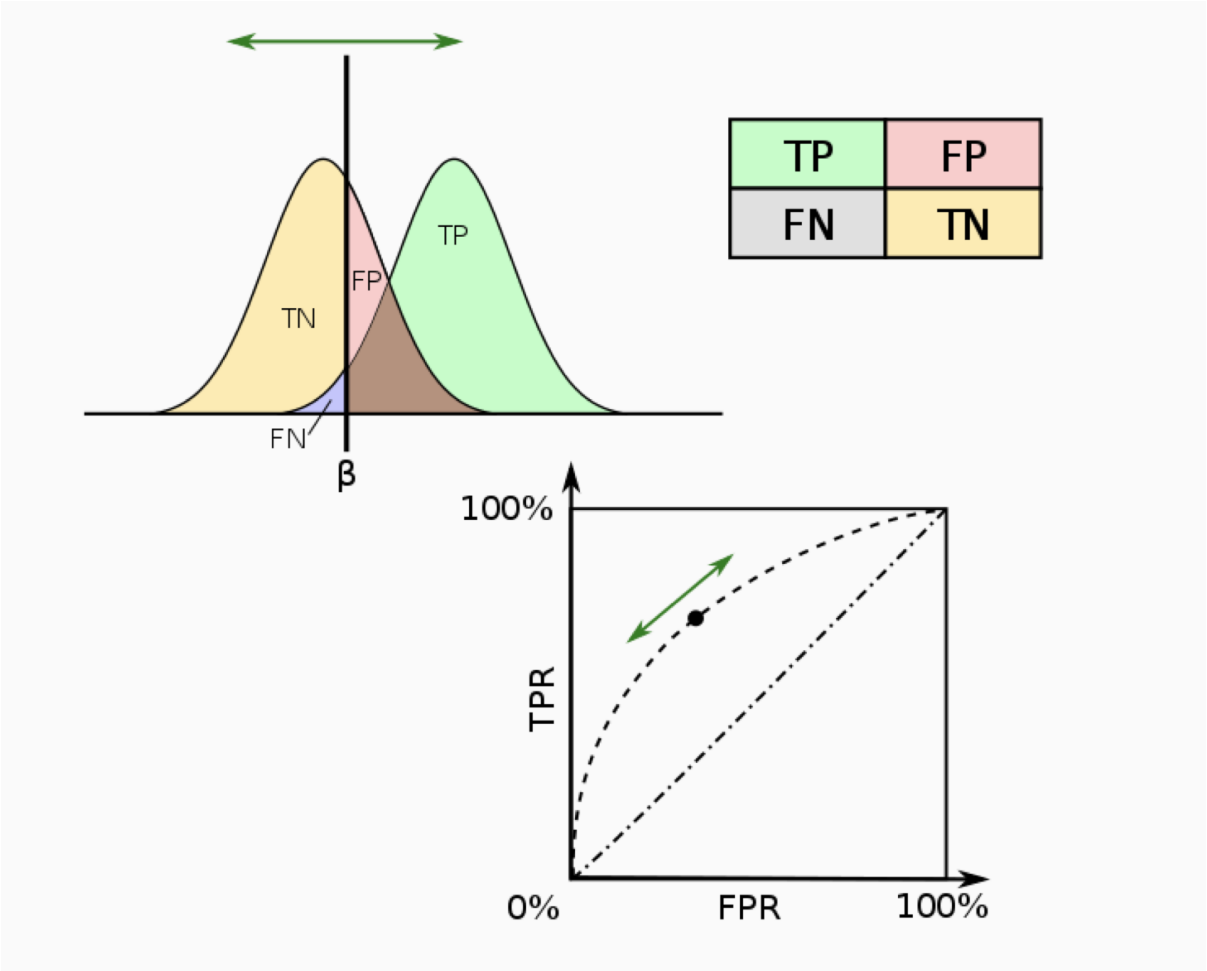

Schéma 2 : illustration du compromis entre les faux positifs et les faux négatifs. Source: M. Sozio, T. Viard, Introduction to Big Data Mining https://tiphaineviard.com/files/SD201_Intro.pdf, accessed July 21, 2022.

2. La différence entre biais et erreur aléatoire

Sur le plan statistique, un biais signifie "toute différence entre l'espérance mathématique d'un estimateur et la grandeur à estimer"[195]. Le terme biais signifie une erreur systématique par rapport à la vérité[196], par opposition à une erreur aléatoire[197]. La différence entre un biais et une erreur aléatoire est liée au problème du compromis entre biais et variance, et au phénomène de surapprentissage versus sous-apprentissage[198]. En effet, un modèle trop fortement adapté aux particularités des données d'apprentissage ne pourra pas se généraliser[199]. Par exemple, un algorithme qui apprend uniquement à partir d'images de

[195] Dictionnaire Larousse en ligne de la langue française, consulté le 18 janvier 2022.
[196] Dans le langage de l'intelligence artificielle, on parle de la "vérité du terrain" (ground truth)
[197] Vocabulaire international de métrologie – Concepts fondamentaux et généraux et termes associés (VIM), 3e édition, Projet final 2006-08-01, p. 51
[198] Wikipédia, Exactitude et précision, consulté le 18 janvier 2022,Wikipédia, Dilemme biais-variance, consulté le 18 janvier 2022.
[199] Wikipédia, Surapprentissage, consulté le 18 janvier 2022; G. James, D. Witter, T. Hastie, R. Tibshirani, "An Introduction to Statistical Learning with Applications in R", Springer 2013, §2.2.

voitures de la marque Renault créera une fonction mathématique compliquée qui permettra de ne jamais se tromper lorsqu'il s'agit d'une Renault. Mais une telle fonction mathématique sera, parce qu'elle a "sur-appris", complètement inadaptée au problème général de reconnaître une voiture. Pour éviter ce problème, il faut lisser la fonction[200] pour qu'elle soit plus souple et adaptable, mais au prix de créer des biais - des erreurs systématiques - pour certains types de cas individuels peu fréquents dans les données.

Le sur-apprentissage peut conduire notamment à des fausses corrélations apprises à tort à partir d'informations sans pertinence (du "bruit") dans les données d'entraînement, et qui provoquent des conséquences catastrophiques lorsque le modèle est exposé à de nouvelles données. Par exemple, le modèle peut apprendre à tort que l'existence de la neige dans une photo est un critère important pour prédire qu'il s'agit d'un loup plutôt que d'un husky[201]. Ou pour reprendre l'exemple des voitures, l'algorithme va associer le losange de Renault avec le fait que l'objet est une voiture. Une autre source importante de biais et d'erreurs aléatoires provient des données utilisées pour l'entraînement du modèle, et/ou des données d'exploitation, comme nous le verrons ci-après[202].

Le but des *data scientists* est d'assurer un taux élevé d'exactitude en moyenne, ce qui signifie qu'il existera nécessairement des compromis et des erreurs. Même si ces erreurs restent statistiquement rares, leur gravité peut être élevée. Les critères de performance de l'algorithme ne tiendront pas nécessairement compte du niveau de gravité de ces erreurs à la marge, ce qui peut conduire à des situations inacceptables dans des contextes critiques pour lesquels le coût d'une erreur individuelle est très élevé[203].

3. Le problème des classes déséquilibrées

La recherche d'incidents rares, des activités terroristes par exemple, peut conduire à la production d'un nombre très élevé de faux positifs en raison du déséquilibre entre les classes de données d'observation. Par exemple, sur un million d'opérations bancaires, seulement 100 opérations seront véritablement suspicieuses. La classe des cas "non-suspicieux" sera 10 000 fois plus grande que la classe des cas véritablement "suspicieux". Dans ces conditions, l'algorithme va générer un nombre très élevé de faux positifs quel que soit la performance prédictive de l'algorithme[204]. Ce phénomène concerne l'ensemble des algorithmes utilisés dans le domaine de la détection de risques terroristes, de fraudes ou de blanchiment de capitaux.

Il ne sera pas rare que ces algorithmes, malgré leur performance prédictive élevée, aient un taux de faux positif de plus de 90%[205]. Ce phénomène peut surprendre. Comment est-il

[200] En langage machine learning, il s'agit de "régulariser" la fonction.
[201] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", Aug. 2016, arXiv: 1602.04938.
[202] Infra, par. III-A-4.
[203] Jean-Louis Dessalles, Des intelligences très artificielles, Ed. Odile Jacob 2019.
[204] La proportion de faux positifs par rapport à des vrais positifs peut facilement atteindre 99%, même pour un algorithme qui au départ à un taux de performance prédictive de 99%.
[205] Pour une illustration en matière de lutte contre le terrorisme, voy. Wikipédia, Base rate fallacy, consulté le 2 février 2022; Astrid Bertrand, Winston Maxwell, Xavier Vamparys, Do AI-based anti-money laundering (AML)

possible qu'un algorithme ayant une performance prédictive de 99%, donc générant moins de 1% de faux positifs, génère plus de 90% de faux positifs en conditions de déploiement ? L'explication réside dans la rareté des cas vraiment positifs. Dans un système de détection de fraudes, par exemple, sur 100 000 opérations, il n'y aura que 10 opérations vraiment frauduleuses. Un algorithme ayant un taux de performance de 99% va quand même générer 1000 alertes positives pour les 100 000 opérations analysées. Résultat : pour les 10 opérations vraiement frauduleuses détectées, le système va générer 990 faux positifs, soit 99% de faux positifs par rapport aux vrais cas de fraudes détectés[206]. Pour un système de détection de risque de criminalité, comme ceux analysés par la CJUE dans les affaires Accord PNR Canada, La Quadrature du Net, et Ligue des droits humains, cela signifie que l'algorithme va "accuser" entre 90 et 99 personnes innocentes pour chaque personne vraiment coupable. Du point de vue du respect des droits fondamentaux, c'est consternant. Et pourtant, la CJUE estime que ce système est compatible avec la Charte, grâce en partie au réexamen individuel de chaque résultat par un expert humain.

4. Les causes de biais

Dans les systèmes de classification, les biais se traduisent souvent par un niveau de performance différent entre deux groupes. Par exemple un système va générer plus de faux positifs pour des personnes ayant une peau noire que pour les personnes ayant une peau blanche[207]. Les biais ont différents origines[208]: les données utilisées pour développer l'algorithme peuvent être non-représentatives de la population qui sera analysée par la suite [209]; les données peuvent contenir des paramètres d'évaluation qui ne captent pas exactement le phénomène que l'on souhaite évaluer, ou bien des paramètres qui créent des effets de rétroaction[210]; le choix du modèle, et son niveau de complexité, peut favoriser certains biais[211]; et bien sûr les données étudiées pour développer le modèle peuvent

systems violate European fundamental rights?, International Data Privacy Law, Volume 11, Issue 3, August 2021, Pages 276–293, https://doi.org/10.1093/idpl/ipab010

[206] Il existe des solutions techniques pour atténuer ce problème, notamment en rééquilibrant les données par voie d'échantillonnage. voy. Mai, Jianning & Chuah, Chen-Nee & Sridharan, Ashwin & Ye, Tao & Zang, Hui. (2006). Is sampled data sufficient for anomaly detection?. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. 165-176. 10.1145/1177080.1177102.

[207] J. Buolamwini, T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of Machine Learning Research* 81:1–15, 2018

[208] P. Bertail, D. Bounie, S. Clémençon, P. Waelbroeck. "Algorithmes : Biais, Discrimination et Équité", hal-02077745 2019; H. Suresh, J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle", arXiv:1901.10002, 2021.

[209] A. Chouldechova « Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments », *Big Data*, vol. 5, n°2, pp. 153-163, 2017. Ce type de biais existe également dans les outils de notation de solvabilité. Les données historiques détenues par les banques ne concernent que les personnes ayant contracté un prêt avec la banque. Au moment de la demande d'un prêt, la population qui nous intéresse est celle de tous les demandeurs de prêt, non seulement les personnes ayant effectivement obtenu un prêt. Il y a une différence entre la population représentée dans l'échantillon et la population que l'on souhaite étudier. Évidemment ce biais est impossible à corriger car on ne peut pas savoir si une personne n'ayant pas contracté un prêt l'aurait remboursé!

[210] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, S. Venkatasubramanian, "Runaway Feedback Loops in Predictive Policing", *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research, Vol. 81 2018), S. A. Friedler, C. Wilson (Eds.). PMLR, New York.

[211] P. Bertail et al. op. cit.

contenir des phénomènes sociaux – racisme, sexisme – qui existent dans la société mais que l'on ne souhaite pas reproduire à l'avenir[212].

Si l'erreur ne provient pas d'un biais, il s'agit d'une erreur aléatoire. Les erreurs aléatoires peuvent être liées à des cas statistiquement rares, que l'algorithme va ignorer pour éviter les pièges liés au sur-apprentissage. Comme nous l'avons vu, un algorithme de prédiction doit s'accommoder d'un certain taux de variance pour pouvoir garder son pouvoir prédictif face à de nouvelles données. Le modèle se focalise sur les cas les plus fréquents, et les cas rares seront ignorés. Si les cas rares appartiennent à un groupe de personnes, par exemple à des personnes aux cheveux roux, il s'agit d'un biais, car le taux d'erreur sera systématiquement plus élevé pour ce groupe de personnes. Si les erreurs apparaissent de manière aléatoire dans la population - les erreurs n'épargnent aucune couleur de cheveux, par exemple - il s'agira d'une erreur aléatoire. D'autres erreurs aléatoires peuvent se produire en raison d'une erreur de saisie ponctuelle dans les données d'entrée, une erreur dans une date de naissance, ou dans le nom du patient par exemple.

5. Les fragilités particulières des réseaux de neurones

Les modèles à base de réseaux de neurones vont générer une notation inférieure à 100% pour chaque observation, même s'il s'agit d'un cas évident, par exemple une prédiction que le soleil va se lever demain. Des procédures de seuillage (*softmax*) vont ensuite choisir le résultat avec la plus forte probabilité. Un cas qui devrait conduire à une décision sans équivoque à 100% pourrait tomber dans une zone grise de probabilité et passer en deuxième rang si le réseau identifie une autre option ayant une probabilité plus forte[213]. Ce phénomène est exploité par l'apprentissage dit adversarial pour piéger l'algorithme par des attaques. Une modification de quelques pixels dans une image peut changer la classification de manière radicale[214]. Un panneau de signalisation "stop" devient un ballon[215]; un avion devient une plage[216].

6. Les prédictions statistiquement fondées dans l'ensemble mais fausses dans un cas individuel

---

[212] Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining Knowl Discov. 2020;10:e1356. https://doi.org/10.1002/ widm.1356;
N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes", *Proceedings of the National Academy of Sciences* 115, 16 (2018); Restrepo-Amariles, David, Algorithmic Decision Systems: Automation and Machine Learning in the Public Administration (30 nov 2020). The Cambridge Handbook of the Law of Algorithms, SSRN: 3974564, p. 281-282.

[213] Jean-Louis Dessalles, Des intelligences très artificielles, Ed. Odile Jacob 2019, pages 113-114.

[214] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2014). Intriguing properties of neural networks. International Conference on Learning Representations (ICLR 2014); Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I. & Sohl-Dickstein, J. (2018). Adversarial examples that fool both human and computer vision. ArXiv, , 1802.08195.

[215] S.T. Chen, C. Cornelius, J. Martin, D. H. Chau, "Robust Physical Adversarial Attack on Faster R-CNN Object Detector", arXiv: 1804.05810, 2018.

[216] L. Chen, G. Zhu, Q. Li, H. Li, "Adversarial Example in Remote Sensing Image Recognition", arXiv: 1910.13222, 2019.

Enfin, un algorithme peut avoir raison sur le plan statistique, mais se tromper sur un cas individuel. Un algorithme peut classer une personne comme non-éligible pour un prêt parce que son espérance de vie est plus courte que la durée du prêt, par exemple. En effet, de nombreuses décisions concernant un traitement individuel reposent sur une classification statistique moyenne. L'industrie des assurances repose ainsi sur ces classifications. Les risques actuariels sont calculés par rapport à des groupes de personnes, des groupes définis en fonction de différents paramètres tels que l'âge des personnes, leur genre, le quartier où elles habitent, leur profession, leur statut de fumeur, ou le modèle de voiture qu'elles conduisent[217]. Ces classifications sont bien entendues des simplifications de la réalité, et ne tiennent pas compte des caractéristiques de chacun. Elles seront fausses appliquées à un cas individuel, mais utiles dans leur ensemble[218]. Peut-on parler d'erreur algorithmique ?

7. La différence entre biais et discrimination

Dans l'affaire La ligue des droits humains, la CJUE impose le contrôle humain en partie pour détecter des discriminations dans les résultats. Quelle est la différence entre biais et discriminations[219] ? Un biais est, comme nous l'avons vu, une erreur systématique qui affecte un certain groupe. Par exemple, un système de reconnaissance faciale va avoir un taux de faux positifs plus élevé pour les barbus que pour les non-barbus. Il y a erreur, et un groupe frappé plus que d'autres. Une discrimination se réfère généralement à une différence de traitement défavorable à l'égard d'un groupe de personnes partageant un attribut protégé par la loi, par exemple une même religion, origine ethnique ou genre. Contrairement à un biais, une discrimination n'implique pas nécessairement une erreur. Par exemple, il peut s'avérer statistiquement exact que les femmes ont moins d'accidents de voiture que les hommes, mais la loi interdira une différence de prime d'assurance entre les hommes et les femmes[220]. Ensuite, contrairement aux biais, les discriminations ne concernent que certains attributs protégés par la loi. Dans notre exemple du système de reconnaissance faciale, le taux de faux positifs plus élevés pour les barbus n'est pas une discrimination parce que le fait d'être barbu n'est pas protégé par la loi. En revanche, un taux de faux positifs plus élevé pour les peaux foncées serait à la fois un biais, et une discrimination, car la couleur de peau est un attribut sensible, protégé par la loi.

Comme l'a rappelé la CJUE dans l'affaire La ligue des droits humains, une discrimination peut être indirecte, à savoir ne pas résulter d'une classification effectuée directement en fonction de l'attribut protégé, par exemple la religion, mais résulter d'autres facteurs, parfois cachés et insidieux - le nom, l'adresse postale, certains achats sur Internet, etc. - qui conduisent indirectement à une différence de traitement selon cet attribut protégé. La détection de ces

[217] L. Barry, A. Charpentier, "L'équité de l'apprentissage machine en assurance" hal- 03561709, 2022.
[218] Le statisticien George Box a créé l'aphorisme "tous les modèles sont faux, mais certains sont utiles". Box, George E. P. (1976), "Science and statistics" (PDF), Journal of the American Statistical Association, 71 (356): 791–799, doi:10.1080/01621459.1976.10480949.
[219] v. R. Pons et L. Risser, Biais et discriminations dans les systèmes d'intelligence artificielle, Dalloz IP/IT 2022 75.
[220] CJUE, 1 mars 2011, *Assoc. belge des Consommateurs Test-Achats*, aff. C-236/09; Barry et Charpentier, op cit., p. 4..

discriminations indirectes est l'un des objectifs du contrôle individuel ex ante, selon la CJUE [221].

8. Tous les modèles sont faux…

En conclusion, un bon modèle correspondra au phénomène observé la plupart du temps, mais il existera toujours des cas marginaux (appelés *outliers*) où le modèle se trompera par rapport au phénomène observé. Un certain taux d'erreurs algorithmiques ne serait donc pas nécessairement un défaut, mais un état inévitable. Même un algorithme extrêmement performant conduira à une proportion impressionnante de fausses accusations lorsque l'algorithme a pour tâche d'identifier de potentielles activités criminelles. Ces faux positifs sont inévitables dès lors que les cas de criminalité recherchés dans les données sont rares par rapport aux cas non-criminels. Ce taux élevé de faux positifs est la principale raison de l'imposition de contrôle humain individuel par la CJUE dans les affaires Accord PNR Canada, La Quadrature du Net, et La ligue des droits humains.

La CJUE compte également sur le contrôle humain pour détecter des discriminations, à savoir des alertes algorithmiques qui visaient de manière disproportionnée les personnes d'une certaine religion, couleur de peau ou origine ethnique. Comme nous l'avons vu, la question des classifications statistiques est mélangée avec la question des erreurs. Chaque classification statistique repose sur une moyenne, ce qui ne garantit nullement qu'un individu se conformera à la moyenne. S'agit-il pour autant d'une erreur pour laquelle le contrôle humain doit apporter une réponse? Les commentaires officiels de la Convention 108+ indiquent que la personne affectée par une décision automatisée pourra prouver "l'inadéquation du profil qu'il est prévu d'appliquer à leur situation particulière[222]". Ce commentaire laisse penser que la personne pourra contester qu'elle appartienne à tel ou tel profil statistique. Dans son avis du 7 avril 2022, la CNCDH mentionne le danger de discriminations entre personnes selon des propriétés diverses, être propriétaire d'un chien, par exemple[223]. Cela va dans le même sens. Mais interdire toute classification statistique des personnes nierait l'ensemble des avantages liés à cette classification, qui permet de simplifier la vie - réduire les coûts de transaction en langage économique - en traitant des cas semblables à la louche. Cette approche a permis à l'industrie des assurances d'exister[224].

On voit que la question des erreurs et des discriminations est riche de nuances. En définissant plus précisément la typologie des erreurs et des discriminations, il devient plus facile d'imaginer le type de contrôle humain qu'il convient d'appliquer. Entre notre typologie

[221] CJUE, Ligue des droits humains, aff. 817/19, point 197. L'élimination de toute discrimination indirecte semble
difficile à appliquer dans le domaine de la lutte contre le terrorisme, car certains critères de classification s'appuient explicitement sur des listes de pays et organisations considérés à risque élevé pour le terrorismes. Ces pays et organisations peuvent représenter, en majorité, des personnes appartenant à la même religion. Le pays ou l'organisation deviendrait donc un indicateur indirect (un "proxy") de religion.

[222] Conseil de l'Europe, Protocole d'amendement à la Convention pour la protection des personnes à l'égard du traitement des données à caractère personnel, adopté par le Comité des Ministres lors de sa 128e session à Elseneur, le 18 mai 2018 (la "Convention 108+"), rapport explicatif para. 75.

[223] CNCDH avis du 7 avril 2022, op cit., note de bas de page 7.

[224] L. Barry, A. Charpentier, op. cit.; F. Schauer, "Profiles, Probabilities and Stereotypes", Belknap Press of Harvard University Press, 2003.

de contrôles - individuels, système, ex ante, ex post - et notre typologie d'erreurs et de discriminations, un tableau de concordance commence à se dessiner. De nombreux événements de la vie courante, l'éligibilité pour un prêt par exemple, dépendent de prédictions statistiques qui s'appuient sur des moyennes. Ces prédictions pourraient s'avérer fausses lorsqu'elles sont appliquées à un cas individuel. Le but du contrôle humain des décisions algorithmiques serait-il de nier toute discrimination entre les personnes fondées sur des profils statistiques ?

Types d'erreurs	Leurs caractéristiques	Pertinence pour le contrôle humain
Faux positifs et faux négatifs	La performance prédictive d'un algorithme de classification sera mesurée par rapport au taux de faux positifs et de faux négatifs. En privilégiant un niveau faible de faux positifs on va augmenter les cas de faux négatifs, et vice versa. Chaque cas d'usage nécessitera une décision sur le taux acceptable de faux positifs par rapport aux faux négatifs. Il s'agit d'une décision politique et juridique, non mathématique. Dans la détection de menaces terroristes ou de cas de blanchiment d'argent, le taux de faux positifs sera nécessairement très élevé en raison du déséquilibre des classes.	Le contrôle humain se concentrera généralement sur les cas positifs pour détecter d'éventuels faux positifs. Les cas négatifs seront trop nombreux en général pour se prêter à un contrôle systématique.
Équilibre entre biais et variance	Chaque modèle comportera un certain taux d'erreur - biais et/ou variance - assumé et souhaitable.	Le contrôleur humain ne saura pas si l'erreur provient d'erreurs "assumées" ou d'erreurs excédant le seuil prévu.
Déséquilibre des classes	Un algorithme très performant générera quand même un nombre très important de faux positifs pour chaque cas de vrai positif.	Le contrôleur humain peut être noyé sous un déluge de 99 faux positifs pour chaque vrai cas positif.

Les biais	Les biais génèrent des erreurs systématiques, se traduisant souvent par des différences de performance, entre différents groupes de la population.	Le contrôle humain d'une décision individuelle ne permettra pas de savoir s'il s'agit d'une erreur aléatoire ou d'un biais. Les conséquences de biais seront généralement plus importantes car les erreurs ont vocation à se répéter pour le même groupe.
Erreurs dans les données d'entrée	Ces erreurs seront considérées comme aléatoires si elles se produisent de manière ponctuelle. Ce seront des biais si les erreurs sont systématiques, par exemple provenant d'un défaut dans un instrument de mesure.	Le contrôle humain d'une décision individuelle ne permettra pas généralement de déceler ces erreurs hors le cas d'un audit ou contestation.
Les attaques (*adversarial examples*)	Les réseaux de neurones donnent des probabilités de moins de 100%, même pour les cas évidents. Cela ouvre une brèche pour des manipulations.	Les erreurs grossières en reconnaissance d'images seront détectées par l'humain, mais d'autres peuvent s'avérer plus difficiles à détecter.
Prédictions statistiquement fondées mais inexactes dans un cas individuel	L'algorithme peut avoir raison sur les caractéristiques moyennes d'une personne appartenant à un certain groupe, mais se tromper dans un cas individuel.	S'il s'agit de la prédiction d'un événement futur, par exemple le remboursement d'un prêt, le contrôle humain ne permettra pas de vérifier l'exactitude de la prédiction sans attendre l'événement futur.
Discriminations	Les résultats algorithmiques affectent de manière disproportionnée certains groupes partageant des attributs protégés par la loi : religion, origine ethnique, couleur de peau, genre, orientation sexuelle. Contrairement aux biais, une discrimination peut exister même en l'absence d'une erreur systématique.	Comme pour les biais, la détection de discriminations nécessitera généralement un contrôle humain de type "système"

Tableau 2 : Les types d'erreurs et leur pertinence pour le contrôle humain

B. Les biais cognitifs humains qui peuvent nuire à la détection d'erreurs

Pour évaluer l'utilité du contrôle humain dans la détection d'erreurs, il faut tenir compte des obstacles qui peuvent nuire à ce contrôle, et notamment les biais cognitifs des personnes chargées du contrôle.

1. Confiance ou défiance excessive

Le terme biais d'automatisation désigne généralement la tendance à se fier automatiquement ou excessivement aux résultats produits par un système d'IA[225], même si ce terme peut également désigner l'effet contraire, à savoir l'aversion aux résultats algorithmique[226]. Les recherches en sciences cognitives et interfaces hommes-machines (*HCI-human-computer interaction*) confirment qu'il est souvent illusoire de demander à un humain de contrôler les résultats d'une machine[227]. L'effet des biais de l'automatisation a d'ailleurs tout particulièrement été démontré dans la reconnaissance faciale. Dans plusieurs études[228], les humains en charge du contrôle des concordances positives proposées par l'algorithme ont été induits en erreur par les résultats algorithmiques, le niveau de performance des contrôleurs humains diminuant par rapport aux contrôleurs ne s'appuyant pas sur les recommandations algorithmiques. Les meilleurs résultats en reconnaissance faciale consistent à demander aux humains et aux algorithmes d'évaluer les images séparément, et ensuite de combiner les résultats[229]. Les algorithmes les plus performants en reconnaissance faciale ne commettent pas davantage d'erreurs que les humains, la performance humaine étant particulièrement mauvaise lorsqu'il s'agit d'images de personnes inconnues[230], où de personnes appartenant à une autre ethnie[231].

2. Les biais de l'inattention

Les biais d'automatisation sont très présents dans les interactions entre pilotes (ou conducteurs) et outils de pilotage (ou de conduite) automatique. Demander à un être humain de superviser les décisions d'ordinateur et d'intervenir en cas de besoin va se heurter à une réalité cognitive. L'attention de l'être humain a tendance à décrocher au bout

[225] Proposition de règlement européen sur l'IA, art. 14(4)(b).
[226] D. Leslie, C. Burr, M. Aitken, J. Cowls, M. Katell et M. Briggs, Intelligence artificielle, droits de l'homme, démocratie et État de droit. Guide introductif, Conseil de l'Europe, 2021
[227] M. De-Arteaga, R. Fogliato, A. Chouldechova, "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores", Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, 2020.
[228] M. C. Fysh, M. Bindemann, "Human-Computer Interaction in Face Matching", *Cogn. Sci*. 2018 Jun 28;42(5):1714–32; J.J. Howard, L.R. Rabbitt, Y.B. Sirotin, "Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making". PLoS ONE 15(8), 2020.
[229] A.J. O'Toole, H. Abdi, F. Jiang, P.J. Phillips, "Fusing face-verification algorithms and humans" IEEE Trans Syst Man Cybern B Cybern. 2007 Oct;37(5):1149-55.P; J. Phillips, A. Yates, Y. Hu, C. Hahn, E. Noyes, K. Jackson, J. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J.-C. Chen et al."Face recognition accuracy of forensic examiners, super recognizers, and face recognition algorithms", *Proceedings of the National Academy of Sciences* Jun 2018, 115 (24) 6171-6176.
[230] D. White, R.I. Kemp, R. Jenkins, M. Matheson, A. M. Burton "Passport Officers' Errors in Face Matching". PLoS ONE 9(8), 2014.
[231] Wong, Hoo Keat et al. "The Own-Race Bias for Face Recognition in a Multiracial Society." *Frontiers in psychology* vol. 11 208. 6 Mar. 2020.

de 30 minutes environ[232]. Une fois son attention perdue, l'humain mettra beaucoup de temps à reprendre le contrôle de la situation en cas d'incident[233]. Le rôle de l'humain surveillant un ordinateur est comparable à celui d'une personne surveillant de la peinture qui sèche[234]. Les humains peuvent également être surpris par le comportement de la machine et réagir de manière inappropriée[235]. Selon Banks et al., un système autonome est "le plus dangereuse lorsqu'il se comporte de manière cohérente et fiable la plupart du temps"[236]. Même pour les systèmes qui se limitent à générer des alertes pour informer un décisionnaire humain d'éventuels risques, la création d'une quantité excessive d'alertes conduit à une lassitude (*alert fatigue*), voire à la désactivation complète du dispositif par le décisionnaire humain[237].

3. La charge cognitive

En plus des problèmes d'inattention et de complaisance, l'être humain sera dans l'incapacité de tenir compte d'un nombre élevé de variables, ce qui rend le contrôle de certains résultats algorithmiques difficile. Le cerveau humain fonctionne en simplifiant les variables au maximum, pour placer les variables-clés dans un narratif structuré mais simple, un narratif déjà bien compris dans l'esprit de l'individu[238]. Par exemple, lorsque l'œil d'un conducteur aperçoit un mouvement au bord de la route, cette information sera immédiatement insérée dans le narratif : "un animal risque de traverser la route devant moi". Certains de ces narratifs sont instinctifs - le bruit d'un serpent dans l'herbe entre ainsi dans un narratif de danger préprogrammé dans notre cerveau. Certains narratifs sont appris par l'expérience. En revanche, lorsque le nombre de variables d'entrée est très élevé et ne relèvent pas de signaux visuels ou sonores usuels, l'ordinateur sera plus efficace que l'humain à en tirer des conclusions. L'analyse d'images est une exception : pour analyser une image de la vie courante, le cerveau humain sera aussi performant, voire plus, qu'un ordinateur. Mais dans les autres cas, comme l'analyse d'opérations bancaires par exemple, demander à un humain de contrôler une décision algorithmique s'appuyant sur un nombre important de variables d'entrée sera futile, l'humain étant dans l'incapacité de vérifier les conclusions de l'ordinateur dans un court laps de temps[239]. Pour ces situations, le contrôle humain s'effectuera plutôt en amont, dans la création de l'algorithme et dans les processus de test et de validation avant sa mise en exploitation[240]. Ce contrôle s'effectuera ensuite par des tests

---

[232] Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 776; J. Zerilli, A. Knott, J. Maclaurin, et al. "Algorithmic Decision-Making and the Control Problem", Minds & Mac*hines* 29, 555–578 (2019), p. 560

[233] Pour le conducteur d'une voiture, le temps est entre 10 et 40 secondes. Wagner, B., "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems", *Policy & Internet*, 11 (2019) p. 109.

[234] N.A. Stanton, "Responses to autonomous vehicles", *Ingenia*, 62, 9 (2015) p. 9.

[235] A. Rankin, R. Woltjer, J. Field, "Sensemaking following surprise in the cockpit—a re-framing problem", *Cogn Tech Work* 18, 623–642 (2016).

[236] Banks, V. A., Plant, K. L., & Stanton, N. A. (2018b). Driver error or designer error: Using the per- ceptual cycle model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety Science, 108, 283*.

[237] M. Susan Ridgley, M. D. Greenberg, *"Too Many Alerts, Too Much Liability: Sorting Through the Malpractice Implications of Drug-Drug Interaction Clinical Decision Support",* 5 *S.L.U. J. HEALTH L. & POL'Y* 257, 259 (2012)

[238] J.-L. Dessalles, op cit., p. 146.

[239] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics Inf Technol* 6**,** 175–183 (2004); Responsabilité et IA, Étude du Conseil de l'Europe, DGI(2019)05, Rapporteur: Karen Yeung.

[240] A. Z. HUQ, « A right to a human decision », *Virginia law review*, vol. 106, 2020, p. 611.

périodiques au cours de la vie de l'algorithme. Un contrôle individuel en aval, décision par décision, sera impossible, sauf si l'humain s'appuie sur d'autres informations non prises en compte par l'ordinateur pour vérifier la cohérence du résultat algorithmique.

4. Les biais émotionnels

Les émotions et les préjugés diminuent également la qualité du contrôle humain. Sans le savoir, les humains peuvent être plus racistes ou injustes dans leurs décisions que les machines. Certains humains sont hostiles par principe aux algorithmes et vont adopter des décisions contraires, même lorsque la recommandation algorithmique est fondée[241]. Le poids des préjugés et des émotions est ainsi très débattu dans le contexte d'outils d'aide à la décision en matière de justice. Des études ont montré que dans les décisions de justice relative à la fixation des peines, les juges fixent les peines les plus lourdes avant le déjeuner [242]. Une autre étude montre que les juges ont tendance à réduire la sévérité d'une recommandation algorithmique lorsque celle-ci est sévère, mais à respecter la recommandation algorithmique lorsque celle-ci est clémente, peu importe le caractère fondé de la recommandation[243]. D'ailleurs, les algorithmes de prédiction de récidive ont été introduits par le législateur aux États-Unis justement pour limiter les aléas liés à la personnalité du juge[244]. Selon le professeur Aziz Huq, les décisions algorithmiques pourraient s'avérer plus juste dans certains cas que les décisions humaines[245], une position également soutenue par le professeur Cass Sunstein[246]. Pour ce dernier, les décisions humaines sont particulièrement biaisées, affectées par des signaux non-pertinents que Sunstein compare à du bruit dans les jeux de données. En réduisant le niveau bruit, et en éliminant les biais cognitifs humains classiques[247] les algorithmes peuvent jouer un rôle important dans la lutte contre les discriminations humaines[248].

5. Biais induits par le régime de responsabilité

Un autre facteur réduisant l'efficacité du contrôle humain concerne le régime de responsabilité civile, qui peut inciter la personne en charge du contrôle humain à suivre excessivement les recommandations algorithmiques. Si l'humain se trompe en suivant la recommandation algorithmique, il sera généralement considéré comme moins fautif que s'il

---

[241] M. De-Arteaga et al., op cit. p. 2.

[242] S. Danziger, J. Levav, L. Avnaim-Pesso "Extraneous factors in judicial decisions", *Proceedings of the National Academy of Sciences*, 108 (17), 2011; pour d'autres exemples, voy. D. Kahneman, O. Sibony, C. Sunstein, "Noise - Pourquoi nous faisons des erreurs de jugement et comment les éviter", Odile Jacob, 2021.

[243] S. D. Bushway, E. G. Owens, A. Morrison Piehl, "Sentencing guidelines and judicial discretion: Quasi-experimental evidence from human calculation errors", *Journal of Empirical Legal Studies* 9, 2 (2012), 291–319.

[244] *State v. Loomis*, 881 N.W. 2d 749 (Wis. 2016), para. 40-42.

[245] A. Z. Huq, « A right to a human decision », *Virginia law review*, vol. 106, 2020, p. 611.

[246] C. R. Sunstein, "Governing by Algorithm? No Noise and (Potentially) Less Bias" Harvard Public Law Working Paper No. 21-35, 2021, SSRN 3925240; D. Kahneman, O. Sibony, C. Sunstein, "Noise - Pourquoi nous faisons des erreurs de jugement et comment les éviter", Odile Jacob, 2021.

[247] Ces biais comprennent généralement les biais de disponibilité, les biais de confirmation, les biais d'ancrage, voy. A. Tversky, D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases". *Science*. 185 (4157) 1974.

[248] Sunstein, "Governing by Algorithm?", art. cit.

se trompe en contredisant l'algorithme[249]. Cette situation poussera l'humain à suivre la recommandation, même lorsqu'elle est erronée. Ce biais se retrouvera dans toute situation où l'humain devra justifier, à son supérieur hiérarchique par exemple, le choix de ne pas suivre une recommandation algorithmique alors que le choix inverse ne nécessiterait aucune justification de la part du décideur humain.

6. Connaître les faiblesses du cerveau humain pour concevoir des contrôles efficaces

Le problème est donc de concilier le principe d'un contrôle humain avec les biais de l'automatisation, les limitations cognitives et les incitations qui peuvent conduire un humain à suivre les résultats algorithmiques au lieu de les examiner avec un œil critique. S'il s'agit de vérifier qu'une image contient bien l'image d'un vélo, le vérificateur humain pourra contrôler, rapidement et sans trop d'effort, la classification effectuée par l'ordinateur pour identifier une éventuelle erreur de classification. En revanche, s'il s'agit de vérifier le niveau de risque de terrorisme fondé sur l'analyse de milliers de données de connexion et de localisation, le contrôle humain sera difficile voire impossible. Cette conclusion signifie que la mesure imposée par la CJUE dans l'affaire *La Quadrature du Net*[250], à savoir un réexamen par un humain de chaque résultat algorithmique avant d'entreprendre d'autres actions d'enquête, serait potentiellement inefficace. L'humain sera dans l'incapacité d'interpréter de manière indépendante l'ensemble des données de connexion analysées par l'ordinateur pour générer son alerte. Le contrôle s'effectuera de manière plus globale, à travers des tests périodiques d'exactitude et de non-discrimination. Il en sera de même pour une décision algorithmique devant se traduire immédiatement par une action, le contournement d'un obstacle sur la route par exemple. L'humain n'aura pas le temps d'intervenir dans la prise de décision individuelle[251]. Enfin, même si les données d'entrée sont parfaitement compréhensibles pour un contrôleur humain, celui-ci aura du mal à suivre la cadence si le volume de résultats et de décisions est élevé. Les actions de retrait de contenus sur les réseaux sociaux font partie de cette dernière catégorie, car les données d'entrée (images, textes) sont généralement compréhensibles pour les humains, mais le nombre d'images et de textes à examiner chaque minute dépasse les capacités des équipes humaines[252].

7. Le rôle des explications dans la diminution des biais humains

Les explications algorithmiques peuvent avoir un impact sur la qualité du contrôle humain, mais pas forcément dans le bon sens : les explications peuvent augmenter la probabilité pour les humains de suivre les recommandations algorithmiques, même lorsque celles-ci sont erronées[253]. Dans le domaine de la lutte contre le blanchiment des capitaux et le

[249] W. N. Price 2nd, S. Gerke, I.G. Cohen, "Potential liability for physicians using artificial intelligence", JAMA 322, 1765-1766 (2019).
[250] CJUE La Quadrature du Net, affaires jointes C-511/18, C-512-18 et C-520-18.
[251] Ce phénomène a été mis en exergue par les conclusions du NTSB sur l'accident mortel du véhicule Uber en Arizona en 2018: National Transportation Safety Board, Accident Report NTSB/HAR-19/03 PB2019-101402, Adopted November 19, 2019.
[252] Cambridge Consultants, Use of AI in Online Content Moderation, Report produced on behalf of OFCOM, 2019.
[253] G. Bansal et al. "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance", Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).

financement du terrorisme, certaines banques évitent de fournir des explications aux analystes humains afin de ne pas altérer leur jugement sur la suite à donner à une alerte[254]. Dans le domaine de la reconnaissance d'image, un outil d'explication mettant en exergue la partie de l'image que l'algorithme a considéré comme importante dans la classification peut néanmoins permettre au contrôleur humain d'identifier des fausses corrélations et d'augmenter la confiance dans l'utilisation de ces outils[255]. Une explication peut aider l'utilisateur à prendre conscience des limitations de l'algorithme[256]. Mais l'explication d'un résultat algorithmique issu d'un modèle d'apprentissage automatique sera d'une faible utilité si l'objectif est de justifier une décision par rapport à des normes morales ou juridiques[257]. Ces explications se limitent à décrire les facteurs qui ont probablement contribué au résultat, mais ne pourront établir un lien de causalité ou une justification par rapport à une règle de droit[258]. Hénin et Le Métayer proposent un système d'explication algorithmique qui puiserait ses explications à partir de sources normatives externes, ce qui pourrait aider à combler le vide qui sépare l'explication algorithmique et la justification d'une décision algorithmique[259].

Association for Computing Machinery, New York, 2021; D. Eiband, A. Buschek, A. Kremer, H. Hussmann, "The Impact of Placebic Explanations on Trust in Intelligent Systems," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, 2019.

[254] Cette information a été fournie à l'auteur par une banque dans le cadre de des recherches de Télécom Paris sur l'IA explicable dans la lutte contre le blanchiment de capitaux (XAIforAML).

[255] C. Meske, Christian & Bunde, Enrico, "Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support" *Proceedings of the International Conference of AI in HCI* 2020, pp. 54-69.

[256] A. Springer, S. Whittaker, "Progressive disclosure: empirically motivated approaches to designing effective transparency,"Proceedings of the 24th International Conference on Intelligent User Interfaces, New York, 2019, pp. 107–120.

[257] S. Robbins, "A Misdirected Principle with a Catch: Explicability for AI", *Minds and Machines* 29:495–514, 2019.

[258] Ibid.

[259] Clement Henin, Daniel Le Métayer. A Framework to Contest and Justify Algorithmic Decisions. AI and Ethics, Springer, 2021, 1, pp.463-476.

IV - LE DEUXIÈME OBJECTIF DU CONTRÔLE HUMAIN : PRÉSERVER LES VALEURS PROCÉDURALES

A. Les valeurs procédurales en tant qu'objectif distinct du contrôle humain

La correction d'erreurs n'est pas le seul objectif du contrôle humain. Un autre objectif, et non le moindre, est d'assurer un processus de décision respectueux de valeurs humaines. Il est possible d'imaginer un système algorithmique qui se tromperait moins qu'un décideur humain[260]. Si l'humain ne sert à rien dans la détection d'erreurs, faut-il pour autant renoncer à l'humain dans la boucle ? On sent que le contrôle humain sert à autre chose que la simple détection d'erreurs. Mais à quoi ? Le contrôle humain préserve des "valeurs procédurales" (*process values*)[261], indépendamment de la "valeur instrumentale"[262] du contrôle humain dans la réduction des erreurs. Le processus de décision, et l'implication de l'humain dans celui-ci, ont une valeur intrinsèque qu'il convient de préserver, quel que soit leur impact sur le résultat de la décision.

Souvent le contrôle humain servira à la fois à la détection d'erreurs et au respect des valeurs procédurales, faisant d'une pierre deux coups. Mais ce n'est pas toujours le cas. Une garantie de procédure peut s'avérer inutile dans la recherche de la vérité, mais restera indispensable pour préserver la légitimité et le caractère équitable de la procédure[263]. Inversement, un contrôle humain peut s'avérer indispensable dans la détection d'erreurs, mais inutile dans le respect des valeurs procédurales[264]. D'où l'importance de séparer les deux objectifs, car ils feront parfois appel à des formes d'intervention humaine différentes.

Les valeurs procédurales seront particulièrement importantes dans des situations où la vérité objective fait défaut. Dans ces situations, le caractère non-erroné d'une décision sera impossible à vérifier, et un principal critère de qualité de la décision sera sa procédure. Dans les situations complexes impliquant les relations humaines, la découverte d'une seule vérité objective est parfois impossible. Comme le dit Antoine Garapon, "juger une personne, ce n'est pas seulement apprécier un acte, mais surgir dans un enchaînement d'événements inextricables et en imputer un à une histoire particulière."[265] Dans ces situations, une procédure respectueuse des droits, et un décisionnaire humain impartial, deviennent un indice important de qualité, créant une présomption de décision non-erronée sans que l'on sache si c'est vraiment le cas. Selon Antoine Garapon, le rituel social autour du processus de décision compte autant que le résultat[266].

Cela est d'ailleurs compris de longue date. La question du respect de valeurs procédurales, et de l'intervention humaine dans les systèmes algorithmiques, n'est pas nouvelle. Dans son examen du

---

[260] D. Kahneman, O. Sibony, C. Sunstein, "Noise - Pourquoi nous faisons des erreurs de jugement et comment les éviter", Odile Jacob, 2021; Cass R. Sunstein, *Governing by Algorithm? No Noise and (Potentially) Less Bias*, 71 Duke L.J. 1175-1205 (2022); A. Z. Huq, « A right to a human decision », *Virginia law review*, vol. 106, 2020, p. 611.

[261] RS Summers, Evaluating and Improving Legal Processes - A Plea for "Process Values", 60 Cornell L. Rev. 1, 12 (1974).

[262] Martin H. Redish and Lawrence C. Marshall, Adjudicatory Independence in the Values of Procedural Due Process, 95 Yale L. J. 455 (1986).

[263] Joint Anti-Fascist Refugee Committee v. McGrath, 341 U.S. 123, 171-72 (1951) (Frankfurter J., concurring).

[264] Des exemples sont présentés infra, para. IV-D-2.

[265] Garapon, Antoine. Bien juger. Essai sur le rituel judiciaire. Odile Jacob, 2001, p. 310

[266] Ibid.

projet de loi informatique et libertés en 1977, la Commission des lois du Sénat soulignait ainsi l'importance de l'intervention humaine pour garantir la qualité humaine du processus décisionnel :

> "L'intention de votre commission est de faire qu'en aucune manière ce mode de jugement ne supplante les moyens traditionnels et introduise l'automatique là où la nuance, pour ne pas dire la délicatesse, sont souvent de mise. La deuxième idée est de maintenir délibérément à la décision de justice son caractère, certes faillible mais essentiellement humain."[267]

Préserver l'aspect "essentiellement humain" d'une décision revient à mettre un humain dans la boucle, même si le terme "dans la boucle" se prête à confusion[268]. Définir ce que l'on entend par "humain dans la boucle" dépend des valeurs procédurales que l'on souhaite défendre, et ce sont ces aspects que nous examinons dans cette partie.

B. Définition des trois valeurs procédurales clé

Mais en quoi consiste une procédure régulière dans le contexte de systèmes algorithmiques ? Il serait tentant de se référer simplement au principe américain de *due process* et au principe européen du procès équitable. Même si cette réponse est exacte, elle n'est pas satisfaisante car elle ne permet ni d'identifier les aspects précis de la procédure qui sont importants dans un système de décision algorithmique, ni de choisir les modalités de contrôle humain qui vont contribuer à les préserver. De plus, les protections du *due process* et du procès équitable ne s'appliquent qu'à des décisions prises par l'Etat. Les décisions du secteur privé y échappent. Il faut donc identifier les aspects de la procédure qui peuvent s'exporter vers le secteur privé. Plusieurs auteurs soutiennent alors que les garanties de procédure issues du droit administratif devraient s'appliquer aux décisions algorithmiques, y compris celles prises dans un contexte privé[269]. Une transposition au secteur privé de l'ensemble des ces garanties procédurales de droit public serait pourtant inappropriée. Le retrait d'un message sur Twitter n'est pas l'équivalent d'une décision administrative, même si les deux types de décision peuvent être alimentés par un score algorithmique. Même pour les décisions de l'administration, les garanties de procédure varieront en fonction des conséquences de la décision pour l'individu. L'approche élaborée par la Cour Suprême dans l'affaire *Mathews c. Eldridge* en témoigne. La Cour a admis une souplesse dans l'application des garanties procédurales, le niveau de protection dépendant, entre autres, de la gravité de l'affaire[270]. Même dans le cadre de décisions administratives, il faut se résoudre ainsi à créer des garanties de procédure sur mesure, adaptées au contexte de la décision. Pour Rebecca Williams, le droit administratif fournit une boîte à outils intéressante pour trouver des solutions pour le secteur privé[271]. Mais le choix des outils doit

---

[267] Rapport de la Commission des lois du Sénat sur le projet de loi informatique et libertés, Rapporteur Jacques Thyraud, 10 novembre 1977, p. 22.

[268] Crootof, Rebecca and Kaminski, Margot E. and Price II, William Nicholson, Humans in the Loop (March 25, 2022). Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011, Available at SSRN: https://ssrn.com/abstract=4066781 or http://dx.doi.org/10.2139/ssrn.4066781.

[269] R Williams, 'Rethinking Administrative Law for Algorithmic Decision Making' (2021) Oxford Journal of Legal Studies;; Frederick Mostert, 'Digital due process': a need for online justice, Journal of Intellectual Property Law & Practice, Volume 15, Issue 5, May 2020, Pages 378–389.

[270] *Mathews v. Eldridge*, 424 U.S. 319, 1976; E. Zoller, « Procès équitable et *due process of law* », *Recueil Dalloz*, 2007, p. 517.

[271] R Williams, 'Accountable Algorithms: Adopting the Public Law Toolkit Outside the Realm of Public Law' (2022) Current Legal Problems

s'effectuer en fonction de chaque situation. Une simple transposition des garanties du public vers le privé serait excessive. La décision prise par une plateforme de retirer une vidéo ne nécessitera pas les mêmes garanties de procédure qu'une décision de lancer une enquête administrative pour suspicion de fraude. De même, la participation de la personne concernée par la décision ne sera pas toujours possible, au moins dans un premier temps. Un signalement algorithmique concernant un risque de terrorisme ne pourra pas être suivi d'une discussion avec la personne ciblée pour recueillir son point de vue. Ces différences dans l'application des principes d'une procédure régulière ne changent rien quant à la nécessité de rechercher les valeurs procédurales qui doivent être défendues dans chaque situation, et de choisir les modalités de contrôle humain adaptées.

Quelles sont ces valeurs ? Le principe du respect d'une procédure régulière puise ses sources dans plusieurs droits fondamentaux : dignité humaine[272], autonomie[273], procès équitable, droit à un recours effectif[274], égalité, et état de droit[275]. Mais peut-on, sur ce fondement, faire une liste précise des valeurs dites de procédure ? Dans un article de 1974 le professeur Robert Summers a identifié une dizaine d'éléments qui caractérisent une procédure régulière : (1) la participation, (2) la légitimité, (3) le caractère non-violent de la procédure (4) le respect de la dignité, (5) le respect de la vie privée, (6) l'aspect volontaire de la procédure, (7) l'équité, (8) la légalité, (9) la rationalité, et (10) le cadre temporel[276]. Summers n'est pas le seul à avoir établi une liste de valeurs de procédure. En 1975 le juge Henry Friendly a énuméré onze éléments présents dans la jurisprudence américaine sur le *procedural due process*[277]. Dans son guide sur l'application de l'article 6 de la CEDH, la Cour européenne des droits de l'homme (CrEDH) a également établi une liste d'une dizaine d'éléments présents dans la jurisprudence européenne sur le procès équitable, dont les plus importants sont (1) un tribunal indépendant et impartial, (2) une association effective à la procédure (3) l'égalité des armes et une procédure contradictoire, (4) la motivation des décisions, (5) un délai raisonnable, (6) une information sur l'accusation et la possibilité de préparer sa défense, et (7) l'équité dans la

[272] Commission Européenne, Groupe Européen d'Éthique des Sciences et des Nouvelles Technologies Déclaration sur l'intelligence artificielle, la robotique et les systèmes « autonomes » 9 mars 2019, p. 11, "La dignité humaine, en tant que fondement des droits de l'homme, implique qu'une intervention et une participation humaines significatives doivent être possibles pour ce qui concerne les hommes et leur environnement. Par conséquent, contrairement à l'automatisation de la production, il n'est pas approprié de gérer le sort des hommes et d'en décider de la même manière que nous gérons et décisions de ce qu'il advient des objets ou des données, même si c'est techniquement concevable."

[273] Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M., and Briggs, M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer. The Council of Europe, p. 18.

[274] Ibid., p. 20 et 22; voir également Commission européenne pour l'efficacité de la justice (CEPEJ) – Charte éthique européenne d'utilisation de l'intelligence artificielle dans les systèmes judiciaires et leur environnement, Conseil de l'Europe, 4 décembre 2018, p. 50-51.

[275] Ibid. p. 22.

[276] RS Summers, Evaluating and Improving Legal Processes - A Plea for "Process Values", 60 Cornell L. Rev. 1, 12 (1974)

[277] Henry J. Friendly, "Some Kind of Hearing", 123 U. Pa. L. Rev. 1267 (1975). Les onze éléments sont: i. L'existence d'un tribunal indépendant et impartial (unbiased); ii. Une information préalable de la personne quant à l'action envisagée et les raisons de l'action. Cet aspect renvoie au principe de "notice", l'un des deux piliers avec le principe de "hearing", du concept de due process; iii. Une opportunité pour la personne de présenter les raisons pour lesquelles l'action en question ne devrait pas être menée ; iv. Le droit de faire entendre des témoins ; v. Le droit de connaître les preuves à charge ; vi. Le droit de s'assurer que la décision est prise uniquement sur la base des preuves présentées dans le cadre de la procédure ; vii. Le droit d'être assisté par un avocat; viii. L'existence de procès-verbaux et autres formes de documents dans lesquels les preuves et arguments sont consignés ; ix. Une décision motivée expliquant les raisons de la décision ; x. L'ouverture du procès au public ; xi. La possibilité d'un recours effectif devant une juridiction.

collecte de preuves[278]. Les Nations unies ont de leur côté identifié huit éléments nécessaires pour qu'un système de réclamation non-judiciaire soit efficace et équitable.[279] Margot Kaminski et Jennifer Urban ont étudié les modalités de contestation de décisions algorithmiques en identifiant plusieurs valeurs de procédure essentielles[280]. Frederick Mostert a identifié huit éléments de ce qu'il appelle le "*digital due process*"[281].

Dans un objectif de simplification, on a tenté de consolider ces différentes valeurs autour de trois conditions : (i) la participation effective et égalitaire de la personne dans le processus de décision, ce qui nécessite un accès aux informations pertinentes et une opportunité pour la personne de présenter ses arguments, (ii) l'existence d'un décisionnaire humain, pour asseoir la légitimité de la procédure et respecter la dignité de la personne affectée par la décision, et (iii) la rationalité, celle-ci supposant une délibération sereine, impartiale, et une motivation logique de la décision. La première condition, "participation égalitaire", regroupe le premier élément (participation) et le septième élément (équité) dans la classification de Summers. La deuxième condition, "décisionnaire humain", regroupe le deuxième élément (légitimité) et quatrième élément (dignité) dans la classification de Summers. La troisième condition (rationalité) reprend tel quel le neuvième élément de Summers. Les trois conditions identifiées ici sont également cohérentes avec les autres listes, et notamment celle de Henry Friendly de 1975, le guide de la CrEDH, les listes dressées par Kaminski et Urban, et par Mostert. On peut donc examiner tour à tour ces trois conditions, ou "valeurs", en les plaçant dans le contexte de décisions algorithmiques. Pour chaque valeur procédurale, nous décrirons ses caractéristiques avant d'examiner les obstacles à sa réalisation et d'éventuelles solutions, y compris techniques, dans un contexte algorithmique.

C. Première valeur procédurale : la participation effective et égalitaire de la personne dans le processus de décision

1. Présentation de la valeur de participation

La personne qui fait l'objet d'une décision doit pouvoir participer au processus de décision, et non seulement la subir. Cet aspect est parfois appelé "contestabilité"[282]. Le principe de

[278] CrEDH, Guide sur l'article 6 de la Convention européenne des droits de l'homme - Droit à un procès équitable (volet pénal), mis à jour au 30 avril 2022.

[279] Nations unies, Principes directeurs relatifs aux entreprises et aux droits de l'homme, 2011, principe n° 31. Les principes sont: (i) la légitimité, (ii) l'accessibilité, (iii) la prévisibilité, (iv) l'équité, (v) la transparence, (vi) la compatibilité avec les droits, (vii) l'évolutivité en fonction des retours d'expérience, et (viii) élaboration en collaboration avec les acteurs affectés.

[280] Kaminski, Margot E. and Urban, Jennifer M., The Right to Contest AI (November 16, 2021). Columbia Law Review, Vol. 121, No. 7, 2021, Available at SSRN: https://ssrn.com/abstract=3965041. Pour ces autrices, les trois éléments clés d'une procédure équitable sont (i) une notification de la personne, (ii) l'opportunité de présenter des arguments, et (iii) un décisionnaire légitime, pages 2032-2040.

[281] Frederick Mostert, 'Digital due process': a need for online justice, Journal of Intellectual Property Law & Practice, Volume 15, Issue 5, May 2020, Pages 378–389, https://doi.org/10.1093/jiplp/jpaa024. Pour Mostert, les huit éléments d'une procédure équitable sont (i) un décisionnaire indépendant et impartial, (ii) une notification préalable, (iii) l'opportunité de présenter des arguments et éléments de preuves, (iv) le droit d'être assisté par un avocat, (v) le droit de faire appel de la décision, (vi) le droit d'avoir accès à un tribunal judiciaire à tout moment, (vii) une décision motivée, (viii) un remède effectif.

[282] Lyons, Henrietta, Eduardo Velloso, and Tim Miller. "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions." *ArXiv:2103.01774 [Cs]*, February 23, 2021; Kaminski, Margot E. and Urban, Jennifer M., The Right to Contest AI (November 16, 2021). Columbia Law Review, Vol. 121, No. 7, 2021,

participation se retrouve dans la jurisprudence américaine sur le *due process*[283], et dans la jurisprudence européenne sur le procès équitable[284]. Le rapport Tricot en 1975 avait déjà souligné, également, l'importance pour les personnes affectées par une décision algorithmique de "discuter les données et les processus dont les résultats leur seront opposés"[285]. Dans sa note de présentation du projet de Directive 95/46, la Commission Européenne a souligné l'importance pour l'individu de pouvoir participer dans la prise de décision, au lieu de la subir[286].

Le RGPD et la Convention 108+ permettent à l'individu, en cas de décision entièrement automatisée, de mettre en avant son point de vue, ses arguments, et de prouver l'inexactitude des données ou l'inadéquation du profil établi par l'algorithme[287]. La recommandation du Conseil de l'Europe sur l'intelligence artificielle souligne le droit pour chacun d'être entendu[288]. La proposition de directive sur les travailleurs de plateforme donnerait le droit aux travailleurs affectés par une décision algorithmique de discuter avec un humain afin de clarifier les raisons de la décision[289]. La personne affectée par la décision doit pouvoir réagir, exercer son autonomie et être entendue par un autre humain[290]. Une procédure qui ignorerait la participation active de la personne dans sa propre défense ne serait pas considérée comme une procédure équitable aux yeux de la société[291]. Pour le CEPD, une absence de participation de la personne concernée pourrait constituer une traitement "déloyal"[292]. La participation est un signe de respect de la personne[293].

L'aspect égalitaire de la participation implique alors comme corollaire, en premier lieu, que la personne concernée par la décision puisse bénéficier d'une symétrie dans l'organisation de la procédure, dans le temps de parole, et dans les modalités d'accès aux informations[294]. En deuxième lieu, la personne devra pouvoir connaître l'ensemble des éléments de la décision

---

Available at SSRN: https://ssrn.com/abstract=3965041; Hirsch T, Merced K, Narayanan S, Imel ZE, Atkins DC. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. DIS (Des Interact Syst Conf). 2017 Jun;2017:95-99. doi: 10.1145/3064663.3064703. PMID: 28890949; PMCID: PMC5590649; Marco Almada. 2019. Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19). 2–11.

[283] Friendly, op cit.; Redish et Marshall, op cit.

[284] CEDH, Guide sur l'article 6 de la Convention européenne des droits de l'homme - Droit à un procès équitable (volet pénal), mis à jour au 31 décembre 2021, §152.

[285] Rapport de la Commission informatique et libertés du 27 juin 1975, Président Bernard Chenot, Rapporteur Général Bernard Tricot, La Documentation Française (Rapport Tricot), p. 81.

[286] COM (92) 422 final - SYN 287, p. 26.

[287] RGPD, article 22(3) et considérant 71; Convention 108+, article 6(1)(a) et commentaires point 75.

[288] Recommandation du Conseil de l'Europe du 8 avril 2020, point B 4.3.

[289] Proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme 9 décembre 2021 COM(2021) 762 final, art. 8(1).

[290] Résolution du parlement européen du 20 octobre 2020, considérant 18 et 35 110

[291] CEDH op cit, § 152. Lu dans son ensemble, l'article 6 de la Convention garantit le droit pour tout accusé de participer de manière effective à son procès (*Murtazaliyeva c. Russie* [GC], § 91), ce qui inclut, entre autres, le droit non seulement d'y assister, mais aussi d'entendre et suivre les débats".

[292] CEPD, Lignes directrices 4/2019 relatives à l'article 25, Protection des données dès la conception et protection des données par défaut, Version 2.0 Adoptées le 20 octobre 2020, point 70.

[293] Schauer, Frederick. "Giving Reasons." *Stanford Law Review*, vol. 47, no. 4, Stanford Law Review, 1995, p. 658, https://doi.org/10.2307/1229080.

[294] CrEDH, Guide sur l'article 6 de la Convention européenne des droits de l'homme - Droit à un procès équitable (volet civil), mis à jour au 31 décembre 2021, p. 90.

pour pouvoir utilement les contester, et notamment les demandes ou reproches de son adversaire[295]. En troisième lieu, la personne doit disposer des délais et moyens intellectuels pour préparer sa défense[296]. L'égalité des armes peut nécessiter l'accès à l'algorithme utilisé, ainsi que les données d'entrée, pour être en mesure de contester les résultats[297].

Une égalité absolue dans la procédure ne conduit pas nécessairement à améliorer le caractère non-erroné de la décision, mais elle contribue au respect de la dignité des personnes concernées, et à la légitimité de la procédure. En 1951, la Cour Suprême a confirmé que même si une procédure ne semblait pas idéalement conçue dans l'objectif de conduire à une décision non-erronée, il était néanmoins important de disposer de telles procédures à l'égard du public au nom duquel ces procédures sont créées[298]. La justice doit non seulement être juste, elle doit être perçue comme étant juste par les citoyens[299].

Même si le principe de "contestabilité" des décisions algorithmiques couvre ces mêmes éléments essentiels de participation, les auteurs mettent également en avant l'importance de la participation des parties-prenantes dans l'élaboration du système algorithmique[300]. Il s'agit d'une participation collective en amont, plutôt qu'une participation individuelle en aval. Nous y reviendrons lorsque l'on évoquera la question des consultations publiques dans le cadre d'analyses d'impact.

2. Les obstacles à la participation

La participation effective et égalitaire de la personne pose plusieurs difficultés. Comment rendre disponibles et intelligibles les informations nécessaires pour cette participation ? Comment faciliter la compréhension de la personne et faciliter sa participation effective ? Comment savoir si la personne participe vraiment ? Deux obstacles vont réduire la possibilité d'une participation effective : le temps de décision, et les limitations liées à l'interface informatique.

Le premier obstacle sera le temps. Un système algorithmique où la décision suit rapidement la génération du résultat algorithmique ne laissera aucune place à l'échange et à la discussion avant la prise de décision. Comme évoqué précédemment[301], les différents scénarios de contrôle humain - "système", "individuel", ex ante, ex post - s'inscrivent dans des fenêtres temporelles différentes. Une participation de la personne peut se manifester au stade du contrôle individuel *ex ante*, ou au stade du contrôle individuel *ex post*, mais cette

[295] CrEDH, Guide sur l'article 6 de la Convention européenne des droits de l'homme - Droit à un procès équitable (volet pénal), mis à jour au 30 avril 2022, point 408; Friendly, Some Kind of Hearing, op. cit., p. 1272.
[296] Ministère de la Justice, Le droit à un procès équitable, 1er juillet 2002 https://www.justice.gouv.fr/organisation-de-la-justice-10031/les-fondements-et-principes-10032/le-droit-a-un-proces-equitable-10027.html
[297] CrEDH, Sigurdur Einarsson a. o. v. Iceland, n. º 39757/15, 4 juin 2019; Houston Federation of Teachers v. Houston Independent School District, 251 F. Supp. 3d 1168 (S.D. Tex. 2017) ; M. A. Paige, A. Amrein-Beardsley, « "Houston, we have a lawsuit": a cautionary tale for the implementation of value-added models for high-stakes employment decisions », Educational researcher, vol. 49, no 5, 2020, p. 350-359
[298] Joint Anti-Fascist Refugee Committee v. McGrath, 341 U.S. 123, 171-72 (1951) (Frankfurter J., concurring).
[299] Redish et Marshall, op cit. p. 483.
[300] Almada, op cit.
[301] supra, n° xx

participation dépendra du temps disponible. Pour les décisions les plus importantes, l'approche "tribunal" du contrôle individuel *ex ante* permettra à la personne concernée de prendre le temps de comprendre les informations et fournir ses arguments au décisionnaire humain. Pour le contrôle individuel *ex ante* selon le modèle "validation avec informations supplémentaires", la participation sera partielle voire inexistante. Par exemple, dans le cas de signalements anti-blanchiment, la banque pourra appeler son client pour demander des précisions, mais cette participation ne sera pas systématique et, lorsqu'elle sera présente, elle sera limitée par le temps, sans réel dialogue entre le décisionnaire humain et la personne concernée. Dans l'approche "validation sans informations supplémentaires" du contrôle individuel ex ante, la participation de l'individu est exclue par définition, puisque le décisionnaire humain se limite à revoir les mêmes informations que celles analysées par l'algorithme. Par exemple, la personne chargée de la validation d'un résultat de reconnaissance faciale ne va pas solliciter l'avis de l'individu figurant dans l'image.

La contrainte temporelle sera moins présente pour un contrôle individuel *ex post*, car il s'agit généralement d'une contestation. La procédure de contestation peut incorporer des délais nécessaires pour permettre à la personne d'accéder aux informations pertinentes, les comprendre, et exprimer ses arguments, en adoptant les garanties procédurales présentes dans l'approche "tribunal". La contrainte temporelle ne sera pas non plus un facteur dans le cadre de participations collectives au stade de l'élaboration de l'algorithme[302].

Le deuxième obstacle à la participation effective de l'individu tiendra aux imperfections dans la présentation des informations et plus généralement dans les interactions humain-machine évoquées dans la Partie II. En effet, la plupart du temps la participation s'effectuera en ligne. L'ergonomie de l'interface utilisateur, la présentation des explications, et les possibilités de dialogue seront donc déterminantes. Or, les méthodes actuelles pour expliquer les décisions algorithmiques se heurtent à de nombreux biais cognitifs qui limitent la qualité de la participation[303]. Dans la deuxième partie, nous avons évoqué ces biais dans le contexte du contrôle humain des résultats algorithmiques : ce sont les biais de l'automatisation. Les mêmes biais affectent les personnes ciblées par une décision algorithmique, lorsque les échanges entre ces personnes et les responsables des décisions, dans le cadre d'une contestation par exemple, s'effectuent à travers une interface informatique. Un défi majeur est donc de concevoir des systèmes capables de provoquer la curiosité de l'individu, d'instaurer un dialogue avec lui, voire de stimuler son esprit critique[304]. Mais, malgré ces approches, la principale barrière à une participation effective de l'individu sera l'interface informatique, qui aura tendance à réduire le niveau d'engagement de l'individu comparé à une discussion face à face.

3. Les solutions pour faciliter une participation effective

Dans la discussion qui suit, on évoquera principalement des solutions qui favorisent la participation de l'individu par rapport à une décision individuelle qui l'affecte. La

[302] Infra, par. I-B-1

[303] Astrid Bertrand, Rafik Belloum, James R. Eagan et Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA. https://doi.org/ 10.1145/3514094.3534164

[304] Ibid.

participation collective des parties-prenantes dans la conception de l'algorithme, et dans les phases ultérieures de contrôle - ce que l'on appelle dans cette étude le contrôle humain "système" - est importante, et sera évoquée à la fin de cette section, mais de manière moins détaillée.

La valeur de participation comporte deux volets : un volet informationnel - comment s'assurer que la personne a accès aux informations utiles pour la contestation ? - et un volet sur la présentation d'arguments - comment faciliter l'échange des points de vue ? Comme nous le verrons, les deux volets sont intimement liés.

a. Les solutions pour présenter des information utiles à l'individu

Comme indiqué dans la Partie I, plusieurs textes prévoient une obligation de fournir des informations "utiles" à l'individu[305]. Mais utiles pour quelle finalité ? Dans un contexte algorithmique, les informations seront "utiles" dans la mesure où elles facilitent la contestation d'une décision algorithmique[306]. La présentation d'information contribue ainsi au respect d'une valeur procédurale, à savoir la contestabilité de l'algorithme[307].

Avant d'évoquer le rôle du contrôle humain, examinons les solutions techniques d'explicabilité qui permettent à la personne affectée de se doter d'une compréhension générale du fonctionnement de l'algorithme, ce qui contribuerait à la contestabilité des décisions algorithmiques. L'explicabilité doit être suffisante pour permettre à l'individu de contester les résultats algorithmiques[308], ce qui conduit à une conclusion contrariante pour les entreprises responsables de l'exploitation de l'algorithme : pour fournir des informations utiles, elles doivent privilégier des approches qui facilitent la contestation des décisions issues de leurs propres

[305] RGPD, arts. 13(2)(f) et 14(2)(g). L'obligation de fournir des informations utiles ressort également de l'article 5 de la Régulation (UE) 2019/1150 du 20 juin 2019 prouvant l'équité et la transparence pour les entreprises utilisatrices de services d'intermédiation en ligne (Règlement « P2B »), même si ce règlement vise l'information fournie à des entreprises utilisatrices de la plateforme.

[306] Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission concernant un cadre pour les aspects éthiques de l'intelligence artificielle, de la robotique et des technologies connexes, 2020/2012(INL), considérant 19, "Le niveau d'explicabilité de ces procédés doit dépendre du contexte de ces procédés techniques et de la gravité des conséquences liées aux résultats erronés ou inexacts, et il doit être suffisant pour pouvoir contester ces résultats…"; voy. également Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies, Lignes directrices en matière d'éthique pour une IA digne de confiance, Publications Office, 2019, https://data.europa.eu/doi/10.2759/54071 (Lignes directrices HLEG), point 53.

[307] Hirsch T, Merced K, Narayanan S, Imel ZE, Atkins DC. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. DIS (Des Interact Syst Conf). 2017 Jun;2017:95-99. doi: 10.1145/3064663.3064703. PMID: 28890949; PMCID: PMC5590649.

[308] Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission concernant un cadre pour les aspects éthiques de l'intelligence artificielle, de la robotique et des technologies connexes, 2020/2012(INL), considérant 19, "Le niveau d'explicabilité de ces procédés doit dépendre du contexte de ces procédés techniques et de la gravité des conséquences liées aux résultats erronés ou inexacts, et il doit être suffisant pour pouvoir contester ces résultats…"; voy. également Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies, Lignes directrices en matière d'éthique pour une IA digne de confiance, Publications Office, 2019, https://data.europa.eu/doi/10.2759/54071 (Lignes directrices HLEG), point 53.

systèmes. Elles doivent sélectionner les informations pertinentes dans le cadre d'une contestation, à l'instar du travail d'un avocat qui chercherait les points faibles d'une décision. Cette obligation créerait un conflit d'intérêts au sein de l'entreprise responsable de l'exploitation, ce qui soulève la question de l'utilisation d'outils neutres, éventuellement fournis par un tiers, pour effectuer le travail de sélection et de présentation d'informations pertinentes[309].

Une stratégie possible pour l'entreprise serait de fournir à l'individu toutes les informations liées à la décision algorithmique. Mais est-ce que cette information serait utile ? Noyé sous un déluge d'informations, l'individu serait alors dans l'incapacité de participer utilement à la décision. Une symétrie dans l'information est, certes, mise en avant par certaines décision en matière de procès équitable[310]. Mais, au stade initial de la contestation, la symétrie serait à double tranchant, le volume de l'information réduisant l'utilité de celle-ci.

Dans la présentation d'informations utiles, l'intervention humaine serait potentiellement moins performante que des solutions techniques, sauf dans le cas où la personne en charge de la sélection et de la présentation des informations serait parfaitement objective, capable de tenir compte des objectifs de l'individu cherchant à contester une décision défavorable. Si la personne était salariée de l'entreprise responsable de l'exploitation de l'algorithme, une telle objectivité ferait défaut. Si elle n'était pas salariée de l'entreprise, la personne n'aurait pas accès aux informations nécessaires. De ce point de vue, un algorithme du type "*argument mining*" pourrait être plus efficace qu'un humain dans la sélection et la présentation d'informations pertinentes[311]. Pour développer un esprit critique des résultats algorithmiques, Bansal et alii proposent un outil d'explication qui présente des contre-arguments, mettant en avant les limitations de la prédiction algorithmique[312]. D'autres proposent des stratégies de jeux pour motiver les personnes à comprendre les informations qui leur sont présentées et à mieux chercher les informations pertinentes[313]. Une autre stratégie mise en avant par la doctrine sur la question consiste à permettre à l'individu de tester l'algorithme en changeant les données d'entrée, par exemple en changeant l'âge ou le salaire mensuel de la personne pour observer l'effet sur le résultat de sortie[314]. Cette approche est de nature à faciliter la compréhension de l'algorithme et sa contestabilité.

---

[309] Imposer l'utilisation d'un outil tiers soulèverait de nombreuses questions par rapport à la sécurité des informations que je n'évoquerai pas ici.

[310] CrEDH, Sigurdur Einarsson a. o. v. Iceland, n. º 39757/15, 4 June 2019.

[311] Ionim, N., Bilu, Y., Alzate, C. et al. An autonomous debating system. Nature 591, 379–384 (2021). https://doi.org/10.1038/s41586-021-03215

[312] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1-81. https://doi.org/10.1145/3411764.3445717

[313] A. Simkute, E. Luger, M. Evans, and R. Jones, "Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context," in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, New York, NY, USA, juillet 2020, pp. 263–268. doi: 10.1145/3393914.3395862.

[314] Pour une description des différentes techniques permettant à une personne de tester les résultats algorithmiques en changeant les données d'entrée, voir Andrew D. Selbst et Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085 (2018), p. 1122.

b. Les solutions pour favoriser la présentation d'arguments par l'individu

En ce qui concerne la présentation des arguments, le meilleur cadre reste celui d'un tribunal respectant les principes d'un procès équitable. Dans le cadre de systèmes de décision algorithmiques, le contrôle humain individuel *ex ante* selon l'approche "tribunal" reste l'étalon d'or, car il permet une participation active de la personne. Mais cette approche sera rarement compatible avec les contraintes de temps d'un système algorithmique. En l'absence d'une procédure assortie des mêmes garanties qu'une procédure judiciaire, comment faciliter la présentation d'arguments ?

Le sujet est lié, comme nous l'avons vu, à celui des explications utiles. Certaines solutions ont été étudiées dans le domaine de l'interaction humain-machine pour favoriser la contestabilité de résultats algorithmiques. La mise en avant d'arguments nécessite la construction d'un contre-narratif, à savoir une histoire qui contraste avec celle privilégiée par l'algorithme. Ce contre-narratif doit s'appuyer sur des éléments de preuves supplémentaires[315]. Selon Lyons, Velloso et Miller, il n'existerait aucun consensus sur les caractéristiques nécessaires d'un système permettant à l'individu de mettre en avant ses arguments[316]. Selon ces chercheurs, le seul point de convergence concerne le fait que les systèmes de présentation des arguments devraient imiter, dans la mesure possible, les mécanismes utilisés dans la contestation de décisions humaines. Les difficultés relatives à la conception d'un système de contestation sont également liées à l'ambiguïté des objectifs poursuivis par ces systèmes : s'agit-il d'un mécanisme pour aider à corriger des erreurs - l'objectif examiné dans la deuxième partie - ou d'un mécanisme servant à favoriser la participation de la personne en tant que valeur distincte[317] ? Un mécanisme destiné uniquement à corriger des erreurs pourrait se dispenser de mesures visant à entendre l'ensemble des arguments de la personne, choisissant uniquement les éléments ayant une incidence sur le caractère éventuellement erroné de la décision. Cependant une telle approche ignorerait la valeur intrinsèque de la participation dont il est question dans cette Partie.

On a évoqué l'idée de présenter des informations pour une éventuelle contestation. Pour aller plus loin, est-ce qu'un outil automatique pourrait aider la personne à formuler ces arguments ? Les chatbots sont utilisés dans la gestion de réclamations de consommateurs[318] mais, à notre connaissance, il n'existe pas de travaux sur la création d'un chatbot dont l'objectif serait d'aider l'individu dans une procédure de

[315] Hirsch T, Merced K, Narayanan S, Imel ZE, Atkins DC. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. DIS (Des Interact Syst Conf). 2017 Jun;2017:95-99. doi: 10.1145/3064663.3064703. PMID: 28890949; PMCID: PMC5590649.
[316] Lyons, Henrietta, Eduardo Velloso, and Tim Miller. "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions." *ArXiv:2103.01774 [Cs]*, February 23, 2021;
[317] Ibid.
[318] Voy. notamment Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, et Rama Akkiraju. 2018. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (*CHI '18*). Association for Computing Machinery, New York, NY, USA, Paper 415, 1–12. DOI:https://doi.org/10.1145/3173574.3173989

contestation. Les chatbots actuels visent à augmenter la satisfaction du consommateur lorsque celui-ci a un problème, de manière à préserver l'image de l'entreprise. La capacité du chatbot à comprendre l'individu et à le guider dans ses démarches contribue à maximiser la satisfaction client. Ainsi, un chatbot qui vise à satisfaire le client va probablement utiliser des mécanismes pour permettre à l'individu de présenter ses arguments efficacement, puisqu'il existera une corrélation entre la présentation des arguments de l'individu et le niveau de satisfaction du client. Mais on peut imaginer que les deux objectifs divergent. Par exemple, dans la gestion d'une réclamation, un chatbot pourrait conclure qu'il serait plus efficace pour l'entreprise d'offrir des bons d'achat plutôt que d'écouter les arguments du client, coupant court ainsi à la présentation d'arguments.

Si l'on se place uniquement dans l'objectif de fournir à l'individu un moyen de présenter efficacement ses arguments, le rôle de l'outil automatique serait comparable à celui d'un avocat qui écoute son client afin de déterminer s'il y matière à contestation, ou un responsable hiérarchique qui écoute un salarié qui ne comprend pas pourquoi il n'a pas reçu de bonus. Dans les deux cas, la première démarche est de comprendre, à travers des questionnements, le point de vue de la personne souhaitant contester la décision. Ce point de vue sera généralement construit par la personne à partir d'une connaissance incomplète des faits. La deuxième démarche sera donc d'amener la personne à prendre connaissance d'autres éléments pertinents, pour voir si ces autres éléments modifient l'appréciation de la personne sur sa situation et la décision prise. La dernière démarche sera d'aider la personne à formuler ses arguments compte tenu de cette vision plus complète de la situation. Présenté ainsi, le processus de construction et de présentation des arguments se confond avec le processus d'explication, les deux fonctionnant en même temps, de manière dynamique. Pour être utile, l'explication devra être écoutée et comprise par la personne, ce qui nécessite une curiosité de sa part. Certains travaux en explications algorithmiques visent justement à éveiller la curiosité de la personne, en lui posant des questions pour que celle-ci se rende compte qu'elle n'a pas connaissance de tous les éléments pertinents à la décision. Ces questions donneront envie à la personne de connaître ses autres éléments[319]. Comme il a été mentionné dans la section précédente, certains chercheurs ont imaginé un outil d'explication qui fournirait des contre-arguments[320]. Pour imiter le plus possible les conditions d'un tribunal, on pourrait imaginer un chatbot qui aiderait la personne à définir précisément ses points de désaccord, formuler ses arguments, et produire des éléments de preuve pour étayer ses arguments. Un

[319] Astrid Bertrand, Towards Informed Decision-making: Triggering Curiosity in Explanations to Non-expert Users, Draft Working Paper, HAL hal-03651368, v1; voy. également, Shin, Dajung Diane, and Sung-il Kim. "Homo Curious: Curious or Interested?" *Educational Psychology Review* 31, no. 4 (December 1, 2019): 853–74. https://doi.org/10.1007/s10648-019-09497-x. Kidd, Celeste, and Benjamin Y. Hayden. "The Psychology and Neuroscience of Curiosity." *Neuron* 88, no. 3 (November 4, 2015): 449–60. https://doi.org/10.1016/j.neuron.2015.09.010.
[320] Bansal et al., op cit.

chatbot de ce type pourrait s'appuyer sur la technologie permettant l'extraction d'arguments (*argument mining*)[321].

c. Le rôle de l'humain pour faciliter la présentation des arguments

Dans l'optique de faciliter la participation, le rôle de l'humain serait de suppléer aux solutions techniques en privilégiant une discussion directe avec la personne, une discussion qui faciliterait la présentation des arguments, mais également la réversibilité des rôles discutée dans la section suivante. La proposition de directive sur les travailleurs de plateforme vise précisément ce rôle pour le représentant humain de la plateforme. Une personne physique devra se rendre disponible pour "discuter et clarifier les faits, circonstances et raisons qui ont conduit à une décision" alimentée par un système algorithmique[322]. L'article 8 de cette proposition privilégie expressément la valeur de discussion humaine en tant que telle, non en tant que moyen pour détecter des erreurs. On retrouve le même souci pour les valeurs de participation dans les recommandations du Conseil de l'Europe sur les impacts des systèmes algorithmiques sur les droits de l'homme, qui préconisent la mise en place de points de contact et de permanences téléphoniques faciles d'accès, et la possibilité pour les individus et les groupes à exprimer leurs doléances[323].

d. Les solutions pour favoriser une participation collective

La participation individuelle est nécessaire, mais non suffisante[324]. La doctrine sur la contestabilité des décisions algorithmiques souligne l'importance de la participation collective dans la phase de conception du système[325]. Au stade de l'élaboration de l'algorithme, des responsables humains feront des choix sur les caractéristiques de l'algorithme, sur les critères de performance à privilégier, sur l'acceptabilité de certains biais. Pour les algorithmes du secteur public, la participation collective s'organisera à travers les mécanismes démocratiques classiques, notamment l'adoption d'une loi, et l'organisation d'un cadre réglementaire qui prévoit une consultation publique, ou l'avis d'une commission représentant les personnes affectées, avant l'adoption de choix techniques. Pour un algorithme du secteur privé, une consultation peut être prescrite par la loi. C'est déjà le cas pour les algorithmes utilisés pour le contrôle de l'activité des salariés[326], même si, selon certains, le champ d'application de cette disposition est trop étroit et devrait s'étendre à toute technologie ayant un impact sur les conditions de travail[327]. Un rapport du Parlement

[321] Ionim, N., Bilu, Y., Alzate, C. et al. An autonomous debating system. Nature 591, 379–384 (2021). https://doi.org/10.1038/s41586-021-03215-

[322] Proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme 9 décembre 2021 COM(2021) 762 final, art. 8(1).

[323] Conseil de l'Europe, Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme, 8 avril 2020, paragraphe C-4.2.

[324] Mantelero, Regulating AI, op cit., § 4.2.1.3.

[325] V. notamment Kaminski et Urban, The Right to Contest AI, op cit.

[326] Art. L. 2312-38 code du travail.

[327] M.-C. Amauger-Lattes, Le dialogue social : outil de régulation de l'intelligence artificielle dans l'entreprise,Droit social 2021 p.146.

européen recommande la consultation systématique des instances représentatives du personnel quant aux choix de conception et de déploiement d'algorithmes ayant un impact potentiel sur les conditions du travail[328]. La proposition de loi américaine *Algorithmic Accountability Act 2022* imposerait une consultation des parties intéressées (*stakeholders*) à l'intérieur et à l'extérieur de l'entreprise par rapport à l'analyse d'impact préparée par le responsable de l'algorithme[329]. Une consultation publique est également prévue par la loi sur la reconnaissance faciale de l'Etat de Washington[330]. La proposition de règlement AI Act, quant à elle, reste silencieuse sur le sujet.

D. L'existence d'un décisionnaire humain

1. Présentation de la valeur d'un décisionnaire humain

Selon Brennan-Marquez et Henderson, la légitimité d'un processus de décision dépend de la possibilité pour les deux personnes concernée, le décideur et la personne ciblée par la décision, de se mettre à la place de l'autre, et d'évaluer l'équité de la décision du point de vue de l'autre[331]. Cette approche fait écho au principe du voile d'ignorance de Rawls, selon lequel les personnes qui élaborent des règles justes doivent toujours garder à l'esprit que les règles pourraient demain s'appliquer à elles[332]. Aux États-Unis, le droit d'être jugé par un jury de pairs[333] s'inspire de cette même idée de réversibilité des rôles. Aujourd'hui une personne peut être dans le rôle d'un membre du jury, mais demain la même personne pourrait être à la place de l'accusé. Selon Antoine Garapon, le fait de juger un autre humain "exige de prendre conscience que celui qui juge partage la condition de celui qui est jugé".[334] La même logique s'applique dans le contexte des conflits armées. Un soldat humain qui prend la décision de tuer un combattant ennemi doit pouvoir imaginer le cas où les rôles seraient inversés. Sans cette réversibilité des rôles, il n'existerait plus de barrière morale et éthique dans la conduite du combat.[335] Brennan-Marquez et Henderson soutiennent que la réversibilité des rôles, et par conséquent la responsabilité morale qui doit accompagner l'acte de juger[336], nécessitent l'existence d'un décisionnaire humain[337], capable de considérer

[328] European Parliament Research Service, AI and digital tools in workplace management and evaluation - An assessment of the EU's legal framework, May 2022,
[329] Proposition de loi S. 3572 Algorithmic Accountablity Act of 2022, Section 3(b)(1)(G).
[330] La loi de l'Etat de Washington sur la reconnaissance faciale prévoit la publication de "l'accountability report" et la prise en compte de commentaires publics avant l'exploitation du système. RCW XXXXX
[331] K. Brennan-Marquez et S. Henderson, "Artificial Intelligence and Role-Reversible Judgment, 109 J. of Crim. L. & Criminology 137 (2019).
[332] J. Rawls,Théorie de la Justice, Points, 2009.
[333] Le droit d'être jugé par ses pairs a ses origines dans le Magna Carta de 1215, et a été repris dans le Sixième Amendement de la Constitution des Etats-Unis. https://www.law.cornell.edu/wex/magna_carta consulté le 25 mai 2022.
[334] Garapon, op cit. p. 310
[335] Yeung 2019, p. 29; Anderson et Waxman 2017..
[336] Commission Européenne, Groupe Européen d'Éthique des Sciences et des Nouvelles Technologies Déclaration sur l'intelligence artificielle, la robotique et les systèmes « autonomes » 9 mars 2019, p. 11, "La responsabilité morale, quel que soit son sens, ne peut pas être attribuée ou transférée à une technologie 'autonome'."
[337] K. Brennan-Marquez et S. Henderson, "Artificial Intelligence and Role-Reversible Judgment, 109 J. of Crim. L. & Criminology 137 (2019).

la personne qui subit une décision comme son égal[338]. Cette égalité ferait défaut en l'absence d'un décisionnaire humain. Dès lors, pour asseoir la légitimité de la procédure et respecter la dignité de la personne affectée par la décision, il est nécessaire de mettre un "humain dans la boucle".

2. Les obstacles à l'existence d'un décisionnaire humain

L'existence d'un décisionnaire humain n'est pas en soi une difficulté. Comme l'a remarqué Crootof et alii[339], certains considèrent, à tort, qu'il suffit d'insérer un humain dans la boucle pour résoudre la plupart des problèmes liés aux décisions algorithmiques. L'humain serait un module supplémentaire qu'il suffirait de brancher au système. La difficulté – et tout l'objet de cette étude – est de définir ce que ce décisionnaire humain va décider, dans quel délai, à partir de quelles informations et dans quel contexte procédural. Le contrôle humain individuel aura du mal à suivre le volume et la vitesse des résultats algorithmiques, ce qui conduira, au moins dans les domaines où la vie humaine n'est pas en jeu, à éviter un contrôle individuel *ex ante* pour privilégier un contrôle système, complété par un contrôle individuel *ex post*. Pour ces situations, le ou les décisionnaires humains seront ceux responsables des contrôle systèmes, et ceux qui statuent sur les contestations individuelles. Même si un contrôle individuel *ex ante* est prévu, la présence d'un décisionnaire humain ne contribuera pas forcément à la préservation des valeurs procédurales. Des dizaines voire des centaines de personnes peuvent être affectées au contrôle individuel *ex ante* d'alertes algorithmiques, que ce soit dans le cadre de signalements anti-blanchiment ou d'alertes sur des contenus illicites sur les réseaux sociaux, mais ces personnes seront généralement inconnues de l'individu affecté par la décision, et n'auront aucune interaction avec lui. Si l'objectif est de promouvoir le respect de valeurs procédurales, ce type de contrôle humain individuel ne sert rien, car le décisionnaire, bien qu'étant humain, reste anonyme, noyé dans une masse de contrôleurs humains qui travaillent dans l'ombre. Dans ces conditions, la réversibilité des rôles qui fonde la nécessité d'une intervention humaine, est absente. En conclusion, un décisionnaire humain sera peu utile, *en tant que garant de valeurs procédurales,* en dehors des situations "tribunal" ou "contestation" car seules ces situations là donnent le temps à un échange humain permettant au décideur et à la personne affectée de se mettre à la place de l'autre. Cela ne veut pas dire qu'un contrôle humain en dehors de ces situations ne sert à rien. Il peut contribuer à la réduction des erreurs (l'objectif examiné dans la Partie II), et à se conformer à ses devoirs de vigilance (l'objectif examiné dans la Partie IV). Mais dans l'optique de promouvoir le respect de valeurs procédurales, la présence d'un décisionnaire humain n'a d'utilité que dans les situations où un temps de dialogue est possible.

3. Les solutions pour garantir un échange avec un décisionnaire humain

Pour promouvoir le respect de valeurs procédurales, l'existence d'un décisionnaire humain doit permettre la réversibilité des rôles, un processus par lequel le décisionnaire se met à la place de l'individu affecté par la décision, et *vice versa*. La réversibilité des rôles instaure un sentiment de symétrie et d'égalité dans les relations. Mais, pour favoriser la réversibilité, un

[338] M. Pritchard, Human Dignity and Justice, Ethics, vol. 82, n° 4 (1972) p. 299; H. Spiegelberg, "Human Dignity: A Challenge to Contemporary Philosophy," Philosophy Forum 9, nos. 1/2 (March 1971): 39-64
[339] Crootof et al., Humans in the Loop, op cit.

décisionnaire humain doit disposer du temps nécessaire pour écouter et comprendre le point de vue de la personne affectée. Ces deux aspects ne sont pas évidents dans un contexte algorithmique.

Pour permettre la réversibilité des rôles, le décideur humain doit avoir un sentiment de responsabilité morale envers la personne, un sentiment qui sera favorisé par un échange verbal et nominatif (à savoir non-anonyme) entre le décideur et la personne affectée, par exemple par voie téléphonique. Le sentiment de responsabilité découlera également de l'obligation pour le décideur de motiver et signer sa décision

Le décisionnaire doit également disposer du temps nécessaire pour permettre un échange verbal et l'expression de l'empathie, ce qui est loin d'être évident dans un contexte de systèmes de décisions algorithmiques. Si l'échange verbal est bien intégré dans l'approche "tribunal" du contrôle individuel *ex ante*, cet échange est difficile, voire impossible, dans le cadre des autres modèles de contrôle individuel *ex ante* (validation avec ou sans informations supplémentaires). Même en cas de contrôle humain *ex post*, le volume des contestations peut rendre l'échange verbal et l'empathie difficiles à organiser.

Le problème du respect des valeurs procédurales et l'existence d'un décisionnaire humain se pose en dehors des systèmes algorithmiques. Dans les litiges de consommation, une procédure de réclamation écrite peut s'avérer frustrante pour le consommateur lorsque l'entreprise répond sans donner le nom et le numéro de téléphone d'une personne qui pourrait échanger avec le consommateur. L'enjeu du litige est souvent trop faible pour justifier la saisine d'un tribunal. Dès lors, le recours à une procédure de règlement extrajudiciaire des litiges de consommation est dorénavant obligatoire pour les entreprises en Europe[340]. La directive 2013/11 impose le recours à une personne physique impartiale pour le règlement extrajudiciaire des litiges, ainsi que la possibilité pour chaque partie d'exprimer son point de vue[341]. La directive précise que la présence physique de la personne n'est pas obligatoire, mais ne dit rien sur l'existence ou non d'un échange verbal[342]. Comme nous l'avons vu, la proposition de directive sur les travailleurs de plateforme imposerait également une "discussion" avec une personne physique pour clarifier la situation[343]. Une telle discussion favorisera en théorie le respect de la personne affectée, l'interlocuteur humain pouvant se mettre à la place de celle-ci et exprimer de l'empathie.

Si les systèmes de contrôle individuel *ex ante* et *ex post* ne permettent pas au décisionnaire humain d'avoir un échange avec la personne suffisante pour permettre l'expression de l'empathie, la valeur procédurale attachée à cette intervention humaine sera amoindrie voire nulle. Il pourrait être utile, dans ce cas là, de prévoir le recours à un organisme de médiation comme en matière de litiges de consommation.

---

[340] Directive 2013/11/UE du Parlement européen et du Conseil du 21 mai 2013 relative au règlement extrajudiciaire des litiges de consommation et modifiant le règlement (CE) n o 2006/2004 et la directive 2009/22/CE.
[341] Ibid., arts. 6(1) et 9(1).
[342] Ibid., considérant 42.
[343] Proposition de directive sur l'amélioration des conditions de travail dans le cadre du travail via une plateforme 9 décembre 2021 COM(2021) 762 final, art. 8(1).

E. La rationalité

1. Présentation de la valeur de la rationalité

L'exigence de rationalité implique celle d'un lien logique entre la décision prise et les éléments sur lesquels la décision est censée s'appuyer.[344] La rationalité est particulièrement développée par Summers dans son article de 1974. La rationalité suppose que le décisionnaire : (i) prenne en considération l'ensemble des arguments et des éléments de preuve, (ii) pondère ces éléments, (iii) délibère sereinement, (iv) produise une décision impartiale, à savoir une décision qui s'appuie uniquement sur des arguments et des éléments de preuve produits dans la procédure, et (v) motive la décision en expliquant les raisons de celle-ci. La rationalité nécessite un lien logique et explicable entre les entrées du processus décisionnel (les arguments et éléments de preuve) et la sortie (la décision).

Or, ce type de lien et d'explication peut s'avérer difficile voire impossible si la décision s'appuie sur un modèle d'apprentissage automatique. Un modèle à base de règles s'y prêtera mieux, permettant de relier le résultat à une série de règles logiques prédéterminées par les programmeurs humains. La rationalité nécessite une prévisibilité et une logique à la fois dans la procédure et dans les décisions qui en découlent. Une procédure régulière suppose que les faits et arguments présentés contribuent à un raisonnement menant logiquement à la décision prise et que l'on puisse, à l'image d'un arbre de décision, comprendre à quel moment, et pourquoi, tel ou tel fait ou argument a été écarté, ou au contraire pris en considération, par le décideur[345]. Comme l'a souligné le tribunal fédéral américain dans l'affaire *Houston Federation of Teachers*, les règles de *substantive due process* imposent un lien logique entre le critère pris en compte par l'algorithme et la finalité recherchée. Karen Yeung (2019) donne l'exemple d'une enseignante licenciée parce qu'elle était rousse[346]. Cette décision de licenciement a été annulée en raison de l'absence de lien logique entre la performance de l'enseignante et la couleur de ses cheveux[347]. Fournir des raisons pour une décision peut souvent contribuer à améliorer la qualité de celle-ci, mais au-delà de cette contribution à la qualité de la décision, l'obligation de motiver une décision contribue aux valeurs de procédure, rendant le processus de décision plus acceptable[348]. La motivation permet aux individus d'intégrer la décision dans un système plus large et cohérent, et

[344] Houston Federation of Teachers, précité.
[345] CrEDH, Guide sur l'article 6 de la Convention européenne des droits de l'homme - Droit à un procès équitable (volet pénal), mis à jour au 31 août 2021, §183 "La motivation a pour finalité de démontrer aux parties qu'elles ont été entendues et, ainsi, de contribuer à une meilleure acceptation par elles de la décision. En outre, elle oblige le juge à fonder son raisonnement sur des arguments objectifs et préserve les droits de la défense
[346] Karen Yeung, Why Worry about Decision-Making by Machine? in Algorithmic Regulation Karen Yeung and Martin Lodge Ed. , Oxford University Press 2019.
[347] Short v. Poole Corporation [1926] Ch 66, 90-91, cité dans Yeung (2019), p. 27.
[348] Chad M. Oldfather, "Writing, Cognition, and the Nature of the Judicial Function" (2008) 96 Geo LJ 1283; CEDH op cit. §183; CrEDH Guide sur l'article 6, op cit., point 185.

adapter leur comportement en conséquence[349]. La rationalité se retrouve naturellement dans les exigences de l'Etat de droit[350] et du droit à un procès équitable[351].

2. Les obstacles à la rationalité

Le principe de rationalité exige un corps de règles stable et compréhensible, et une application logique et impartiale de ces règles aux cas individuels. L'exigence de rationalité peut se heurter aux algorithmes d'apprentissage automatique dont les règles internes sont dépourvues de liens logiques apparents. Les approches d'explicabilité *post hoc* donneront quelques indices sur le fonctionnement interne de l'algorithme, mais ces approches ne permettront pas de dégager des liens de causalité. Un score algorithmique fondé seulement sur une approche connexionniste serait ainsi antinomique par rapport au principe de rationalité[352]. Cela signifie-t-il qu'une prédiction statistique ne pourra jamais fonder une décision rationnelle ? Certainement pas. De nombreuses décisions - dans les domaines des assurances, de la sécurité, et du traitement médical par exemple - peuvent s'alimenter en tout ou en partie de prévisions statistiques tout en étant considérées comme "rationnelles". Dans ce cas, l'outil statistique s'intègre dans un raisonnement logique plus large sur la sécurité d'un médicament, sur la probabilité d'un sinistre, ou sur la météo. Lorsqu'une prévision statistique s'intègre dans un système de règles logiques, les méthodes d'analyse statistique sont utilisées pour assurer la qualité de la prédiction. Ces mêmes méthodes font souvent défaut dans l'apprentissage automatique. Ainsi, un procédé de *machine learning* peut difficilement servir à lui seul à justifier une décision. Tout au plus, il peut contribuer à une décision logique prise par rapport à plusieurs critères, dont le résultat algorithmique.

Autre élément de rationalité, l'impartialité de la décision se heurtera à deux types d'obstacles : premièrement, les biais humains ; deuxièmement, le lien de subordination entre le décideur humain et l'entité responsable de l'exploitation du système algorithmique. Nous ne reviendrons pas sur les biais humains examinés en détail dans la deuxième partie. En ce qui concerne le lien de subordination, ce problème existera, à des degrés différents, pour tout mécanisme de contrôle humain organisé par le fournisseur ou l'exploitant du système de décision algorithmique. Le décideur humain sera généralement salarié ou sous-traitant de l'entreprise qui exploite l'algorithme, ce qui soulève des risques de conflits d'intérêts. Les éventuels conflits d'intérêts nuiront à la rationalité.

3. Les solutions pour favoriser la rationalité

La rationalité exigera une décision sereine, impartiale et logique, fondée sur l'application de règles connues. En tant que fondements de l'Etat de droit aux Etats-Unis, la rationalité sera

[349] Schauer, Frederick. "Giving Reasons." Stanford Law Review, vol. 47, no. 4, Stanford Law Review, 1995, pp. 633–59, https://doi.org/10.2307/1229080.
[350] Conseil de l'Europe, Commission européenne pour la démocratie par le droit (Commission de Venise), Liste des critères de l'Etat de droit, CDL-AD(2016)007, adoptée le 11-12 mars 2016, p. 15.
[351] CrEDH Guide de l'article 6, op cit. §183
[352] L'absence de rationalité dans les modèles de machine learning a conduit le Conseil constitutionnel et au législateur français à exclure leur utilisation par l'administration dans le cadre de décisions entièrement automatisées. Conseil constitutionnel, Décision n° 2018-765 DC du 12 juin 2018, point 71; voy. également CJUE Ligue des droits humains, affaire C-817/19 du 20 juin 2022, point 194.

également un élément important pour une procédure régulière dans le contexte de décisions algorithmiques privées. La rationalité est déjà exigée dans le cadre de décisions algorithmiques de retrait de contenus terroristes, puisque la plateforme d'hébergement doit publier ses règles sur la lutte contre des contenus terroristes, s'assurer que les mesures prises pour lutter contre de tels contenus sont ciblées, proportionnées, et appliquées de manière diligente et non-discriminatoire[353]. L'obligation d'objectivité et de non-discrimination dans l'application des règles traduit une forme de rationalité[354]. L'article 42 de la loi française du 24 août 2021 impose aussi l'envoi d'une notification expliquant les raison du retrait d'un contenu[355], et une obligation similaire figure dans le projet de règlement DSA[356]. La rationalité requiert la possibilité pour le décideur humain de se référer à des règles auxquelles tous les utilisateurs sont soumis de manière non-discriminatoire[357]. Cela se traduit par la nécessité de se référer non à un résultat algorithmique comme seul fondement d'une décision, mais au résultat algorithmique comme élément qui contribue à l'application d'un système de règles externes à l'algorithme, les règles de l'entreprise ou les textes de loi. Par exemple, le refus d'un crédit pourrait se justifier par rapport au règlement intérieur de la banque sur les critères d'octroi de crédit, ce règlement intérieur précisant que les scores algorithmiques seront un facteur, parmi d'autres, pris en considération. Dans le secteur médical, le recours à une recommandation algorithmique pourrait se justifier par rapport au protocole de traitement qui prévoit expressément cet élément dans le cadre du parcours de soins. Intégré dans un cadre rationnel plus large, le résultat algorithmique, même fondé sur l'apprentissage automatique, ne serait pas contraire au principe de rationalité.

La rationalité englobe aussi l'idée d'une délibération impartiale, sereine et réfléchie de la part du décideur humain. Mais, comme nous l'avons vu dans la deuxième partie, les biais de l'automatisation rendent difficile un processus de décision impartial, serein et réfléchi. Le décideur humain se retrouvera généralement dans une situation où la meilleure stratégie pour lui sera d'approuver les recommandations algorithmiques. Certains outils d'explication peuvent aider à ralentir le processus de décision, voire créer un doute chez le décideur pour que celui-ci hésite avant d'approuver le résultat algorithmique. Mais les contraintes de temps, et la structuration de la responsabilité pour les acteurs, auront souvent le dernier mot. Le doute et la réflexion pourront retrouver une place dans la procédure de contestation, à condition que celle-ci soit structurée selon l'approche "tribunal"[358].

L'impartialité, quant à elle, serait favorisée par une séparation fonctionnelle entre les personnes qui gèrent les contrôles individuels *ex post* (les contestations) et celles qui gèrent les contrôles individuels et système *ex ante*[359]. Une impartialité totale nécessiterait le recours

[353] Règlement 2021/784 du 29 avril 2021 relatif à la lutte contre la diffusion des contenus à caractère terroriste en ligne; W. Maxwell, Applying Net neutrality rules to social media content moderation systems, Enjeux numériques, Annales des Mines, N° 18, juin 2022..
[354] Summers, Evaluating and Improving Legal Processes, op. cit. p. 26.
[355] Loi n° n° 2021-1109 du 24 août 2021 confortant le respect des principes de la République, art. 24.
[356] Proposition de règlement du Parlement européen et du Conseil relatif à un marché intérieur des services numériques (Législation sur les services numériques) et modifiant la directive 2000/31/CE, SEC(2020) 432 final.
[357] Clément Henin, Daniel Le Métayer, Beyond explainability: justifiability and contestability of algorithmic decision systems, AI & SOCIETY, July 2021 https://doi.org/10.1007/s00146-021-01251-8.
[358] Pour une description de l'approche tribunal, voy. la section xxx, supra.
[359] Cette séparation fonctionnelle est préconisée dans les Principes de Santa Clara sur la modération de contenus en ligne: Electronic Frontier Foundation, American Civil Liberties Union Foundation for Northern

à un organisme indépendant, à l'image des systèmes de médiation pour les consommateurs [360].

F. Conclusion sur le rôle de l'humain dans le respect des valeurs procédurales

Les valeurs procédurales examinées dans cette troisième partie sont indépendantes de l'objectif de détection d'erreurs examinées dans la deuxième partie de l'étude. Un contrôle humain de résultats algorithmiques peut être utile dans la lutte contre les faux positifs, mais complètement inutile dans le respect des valeurs procédurales. Et inversement : une procédure de dialogue humaine peut contribuer au respect des valeurs procédurales sans apporter une valeur ajoutée en matière de détection d'erreurs. Ce constat pourrait conduire à une approche cynique par laquelle une procédure de dialogue est instaurée dans l'unique but de faire participer la personne, sans que cela ait un impact quelconque sur la décision. Heureusement, la séparation entre la détection d'erreurs et la défense des valeurs procédurales n'est pas totale. Dans la plupart des situations, les actions qui contribuent au respect des valeurs de procédure ont aussi un rôle dans la détection d'erreurs. En plus, un système qui ferait participer la personne uniquement pour lui donner l'impression de participer, sans en tenir compte dans la décision finale, bafouerait les principes mêmes que la participation est censée promouvoir, dont le respect de la dignité humaine. Ainsi, les deux rôles du contrôle humain, détection d'erreurs et respect des valeurs procédurales, font partie d'un ensemble. Il n'empêche que, dans le choix des modalités pratiques du contrôle humain, les deux objectifs doivent être clairement séparés.

California, Center for Democracy and Technology, Open Technology Institute, The Santa Clara Principles: On Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org.

[360] Directive 2013/11/UE du Parlement européen et du Conseil du 21 mai 2013 relative au règlement extrajudiciaire des litiges de consommation et modifiant le règlement (CE) n o 2006/2004 et la directive 2009/22/CE

V - LE TROISIÈME OBJECTIF DU CONTRÔLE HUMAIN : EXÉCUTER SES OBLIGATIONS DE VIGILANCE ET DÉMONTRER SA CONFORMITÉ

A. Présentation du troisième objectif lié à la conformité

La troisième finalité du contrôle humain est de démontrer sa conformité, afin d'atténuer les risques de responsabilité et de sanctions. Cet objectif découle des deux premiers. Si un certain niveau de contrôle humain s'avère utile et nécessaire dans la détection d'erreurs et dans la protection des valeurs procédurales, alors ce contrôle fera partie des obligations de vigilance qui pèsent sur l'exploitant de l'algorithme. Une absence de contrôle humain sera synonyme d'absence de vigilance, et donc, faute. Même en l'absence d'un accident causant un préjudice, un défaut de contrôle humain peut constituer un défaut de conformité, les systèmes de conformité étant organisés en grande partie autour de structures humaines de gouvernance. Une absence de contrôle humain sera synonyme de défaut de conformité, entraînant un risque de sanctions au titre du RGPD et du futur règlement européen sur l'IA.

Pour l'exploitant d'un système d'IA, un contrôle humain contribue à montrer sa prudence et ses diligences dans la supervision du système, ainsi que sa conformité au RGPD et au futur règlement européen AI Act. Une entreprise exploitant un système IA s'efforcera de mettre en place un contrôle humain "approprié" dans l'optique de réduire sa responsabilité civile et ses risques de sanctions administratives. Dans les sections qui suivent, nous essayerons de définir ce qu'est un contrôle humain approprié au niveau du RGPD, du futur règlement AI Act et de la responsabilité délictuelle. Cette recherche nous conduira naturellement aux deux premiers objectifs du contrôle humain : un contrôle humain approprié sera celui qui réduit les risques d'erreurs (objectif n° 1), et qui contribue à préserver les valeurs de procédure (objectif n° 2). Le troisième n'est finalement que la suite des deux premiers, sans existence autonome. Néanmoins, examiner le contrôle humain à la lumière des règles de responsabilité civile, du RGPD et du futur règlement AI Act nous aidera à apprécier l'impact de l'état de l'art et des coûts. Une mesure de prévention "appropriée" résultera généralement d'une pondération entre le risque - son niveau de gravité et sa probabilité -, l'état de l'art, et le coût de mise en œuvre des mesures de protection.

Mais avant d'examiner ces questions juridiques, abordons brièvement un autre aspect de la responsabilité qui a son importance, l'aspect moral. Selon certains philosophes, le contrôle humain serait nécessaire pour établir une responsabilité morale pour une action. Certaines décisions importantes, la décision d'utiliser la force létale sur le champ de bataille par exemple, nécessitent une décision humaine et une justification par rapport à un système normatif, par exemple les règles de la guerre. Sans cette décision humaine et cette justification, il existerait un vide de responsabilité (responsibility *gap*) inacceptable. Le contrôle humain devient ainsi un impératif moral, le fondement de la responsabilité morale qui doit s'attacher à chaque décision ayant des implications morales pour autrui. Cet aspect du contrôle humain est certes important, mais appartient à mon sens au deuxième objectif du contrôle humain, la préservation des valeurs de procédure. L'existence d'un décisionnaire humain, et l'exigence de rationalité, visent la même idée : certaines décisions nécessitent par nature une décision humaine et une justification par rapport à un système de normes partagé. La question du contrôle humain et la responsabilité morale ne sera donc pas traitée dans cette partie.

Le terme “vide de responsabilité” (*responsibility gap*) peut également prêter à confusion. Dans la littérature philosophique sur les armes autonomes, le vide de responsabilité désigne l’absence de responsabilité morale[361]. Pour le juriste, le vide de responsabilité désigne la situation où il s’avère difficile d’identifier un responsable d’un dommage causé par l’IA en raison de l’évolutivité du système, celui-ci échappant au contrôle de ses concepteurs. Le vide de responsabilité au sens juridique sera probablement comblé par la jurisprudence et/ou par un règlement qui désignera un ou plusieurs responsables par défaut. La résolution du parlement européen propose à cet égard de désigner les opérateurs d’IA en aval et en amont comme responsables par défaut[362]. Le RGPD désigne, quant à lui, un responsable du traitement sur qui l’ensemble des obligations pèsent en matière de protection des données à caractère personnel.

Dans les sections qui suivent, on examinera le contrôle humain dans un régime de responsabilité délictuelle pour faute (A), le contrôle humain en tant que mesure de protection appropriée imposée par le RGPD (B), et le contrôle humain en tant que système de gestion de risque imposé par le futur règlement AI Act (C).

B. La responsabilité civile délictuelle

La résolution du Parlement européen du 20 octobre 2020 sur la responsabilité civile pour l’intelligence artificielle[363] propose une responsabilité sans faute pour l’exploitation d’un système IA à haut risque, et une responsabilité pour faute en cas de dommages causés par un système d’IA qui ne serait pas considéré comme à haut risque. Cependant, dans le régime de responsabilité pour faute proposé par le Parlement européen, la faute serait présumée et pour s’exonérer l’entreprise devra apporter la preuve qu’elle s’est conformée au devoir de diligence[364].

Dans un régime de responsabilité sans faute, la présence ou l’absence d’une intervention humaine par l’exploitant ne changera rien à l’analyse de la responsabilité en cas de dommage causé par la décision algorithmique. Dans un régime sans faute, l’objectif de l’intervention humaine sera surtout d’éviter des erreurs créatrices de dommages, car chaque erreur créatrice de dommage générera une responsabilité[365]. Dans un régime de responsabilité fondé sur la faute, la présence ou l’absence d’une

---

[361] Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI. https://doi.org/10.3389/frobt.2018.00015, p. 4..Roff, H.M. and R. Moyes. 2016. “Meaningful Human Control, Artificial Intelligence and Autonomous Weapons.” Briefing Paper Prepared for the Informal Meeting of Experts on Lethal AutonomousWeapons Systems, UN Convention on Certain Conventional Weapons, Geneva, Switzerland.

[362] Cette approche est également celle du RGPD, qui désigne un “responsable du traitement”.

[363] Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission sur un régime de responsabilité civile pour l’intelligence artificielle https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_FR.html (2020/2014(INL))

[364] Ibid., point 20 “la personne lésée devrait néanmoins bénéficier d’une présomption de faute de l’opérateur, qui devrait avoir la possibilité de se disculper en apportant la preuve qu’il s’est conformé au devoir de diligence”

[365] Sur les régimes de responsabilité potentiellement applicable à l’IA, voir Conseil de l’Europe, Comité d’experts sur les dimensions des droits de l’homme dans le traitement automatisé des données et les différentes formes d’intelligence artificielle (MSI-AUT), Étude sur les incidences des technologies numériques avancées (dont l’intelligence artificielle) sur la notion de responsabilité, sous l’angle des droits humains, rapporteuse Karen Yeung, DGI(2019)05, 2019; S. Dormont, « Quel régime de responsabilité pour l’intelligence artificielle? », *CCE* 2018, Étude 19, p 58-62; C. Mangematin, “Les propositions européennes visant à encadrer la responsabilité civile découlant de dommages causés par l’intelligence artificielle. Bien mais peut mieux faire!”, Responsabilité Civile et Assurances, LexisNexis, n° 5, mai 2022, p. 7; C. Lachieze, Vers un régime de

intervention humaine sera calibrée en fonction du devoir de diligence. Dans cette optique, l'absence de contrôle humain pourrait constituer une faute. Lorsque la charge de la preuve pèse sur l'exploitant de l'IA pour démontrer l'absence de faute, celui-ci sera d'autant plus incité à mettre en œuvre un système de contrôle humain pour montrer que l'entreprise a mis en œuvre toutes les mesures raisonnablement nécessaires pour éviter des préjudices. Prouver l'absence de faute revient à démontrer que l'exploitant s'est conformé à son devoir de diligence. Quel est ce devoir de diligence, et quelle est la place de l'intervention humaine dans ce devoir?

La résolution du Parlement européen du 20 octobre 2020 sur la responsabilité civile pour l'intelligence artificielle[366] propose un devoir général de diligence pour l'exploitation de systèmes d'IA qui ne serait pas considéré comme étant à haut risque.[367] Ce devoir consisterait à appliquer "toute la diligence requise… *en exécutant toutes les actions suivantes: en sélectionnant un système d'IA adapté au regard des tâches à accomplir et des capacités requises, en mettant correctement en service le système d'IA, en contrôlant ses activités et en maintenant la fiabilité opérationnelle par l'installation régulière de toutes les mises à jour disponibles."*[368]

Comme l'a constaté la CNIL, le choix des modalités de l'intervention humaine sera un élément de ce devoir de diligence ou de vigilance[369]. La CNIL, quant à elle, propose un devoir général de *vigilance* en matière d'IA:

> *"Il s'agit d'organiser, par des procédures et mesures concrètes, une forme de questionnement régulier, méthodique, délibératif et fécond à l'égard de ces objets techniques de la part de tous les acteurs de la chaine algorithmique, depuis le concepteur, jusqu'à l'utilisateur final, en passant par ceux qui entraînent les algorithmes.*[370]"

Le devoir de *vigilance* envisagé par la CNIL est-il différent du devoir de *diligence* envisagé par le Parlement européen dans sa résolution du 20 octobre 2020 ? En droit bancaire, le devoir de vigilance va au-delà du simple devoir de diligences, impliquant l'existence d'un devoir fiduciaire[371]. En matière de protection de l'environnement et des droits humains, la loi du 17 mars 2017[372] définit le devoir de vigilance comme les mesures propres à identifier et à prévenir les atteintes graves envers les droits humains et les libertés fondamentales, la santé et la sécurité des personnes ainsi que l'environnement. La loi du 17 mars 2017, comme le rapport de 2017 de la CNIL, associe le terme vigilance à une double obligation : une obligation de curiosité et d'exploration dans l'identification des risques, et une obligation de mettre en place des mesures pour prévenir ces risques. Le devoir de curiosité et d'exploration se retrouve dans les obligations qui pèsent sur les membres d'un conseil

---

responsabilité propre à l'intelligence artificielle? JCP G 2021, 457; D. Galbois-Lehalle, Responsabilité civile pour l'intelligence artificielle selon Bruxelles: un initiative à saluer, des dispositions à améliorer, D, 2021, p. 87.

[366] Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission sur un régime de responsabilité civile pour l'intelligence artificielle (2020/2014(INL))

[367] Cette résolution propose une responsabilité sans faute pour les systèmes à haut risque.

[368] Ibid. art. 8(2)(b).

[369] Ce lien est également mis en avant par Wagner, B. (2019), Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. Policy & Internet, 11: 104-122. https://doi.org/10.1002/poi3.198

[370] CNIL 2017, p. 6.

[371] N. Mathey, Les devoirs fiduciaires du banquier, Revue de Droit bancaire et financier n° 5, Septembre 2021, étude 14, p. 10.

[372] L. n° 2017-399, 27 mars 2017 relative au devoir de vigilance des sociétés mères et des entreprises donneuses d'ordre, JORF n° 0074, 28 mars 2017.

d'administration, qui ont le devoir de questionner les dirigeants sur leurs actions et sur les informations présentées[373]. A ce stade on peut considérer que le devoir de diligence mentionné par le Parlement européen et le devoir de vigilance mentionné par la CNIL sont similaires, sauf dans la mesure où le devoir de vigilance ajouterait une obligation de curiosité et de diagnostic en amont pour identifier les risques. Quoi qu'il en soit, le contrôle humain sera un élément parmi d'autres dans les obligations de prévention découlant du devoir de diligence (ou de vigilance)[374].

Le contrôle humain n'est cependant pas mentionné explicitement, ni dans la définition du devoir de diligence proposée par le Parlement européen, ni dans celle du devoir de vigilance proposée par la CNIL. La CNIL précise toutefois que le choix des modalités du contrôle humain s'effectuera en fonction de l'obligation de vigilance. Plus l'enjeu pour les droits et libertés individuels est important, plus l'obligation de vigilance pourrait exiger un contrôle humain fort. La résolution du Parlement européen ne vise pas expressément le contrôle humain, mais il est nécessairement présent. Pour le Parlement européen le devoir de diligence comprend l'acte de sélectionner un système adapté, de le mettre en service, et surtout de le contrôler. Il serait difficile d'imaginer un contrôle des activités d'un système IA sans un élément humain. La CNIL évoque un questionnement "régulier, méthodique, délibératif et fécond" par rapport aux risques, un devoir de questionnement essentiellement humain.

La doctrine américaine sur la responsabilité civile délictuelle s'est penché sur le cas des dispositifs médicaux utilisant de l'IA. Dans le cadre d'hôpitaux, l'introduction de nouveaux dispositifs d'aide à la décision pourrait s'avérer fautive si l'hôpital n'a pas effectué les vérifications nécessaires de l'outil et ne s'est assuré de la formation des médecins[375]. La nécessité ou non d'une intervention humaine du médecin dépendra des conditions d'utilisation définies par le fabricant dans le cadre de l'homologation du produit, ainsi que l'état de l'art[376]. Si l'algorithme est une aide à la décision, ce sera le décisionnaire humain et éventuellement son employeur qui seront responsables en cas de faute dans la prise de décision[377].

---

[373] Cass. com. « Crédit Martiniquais », 30 mars 2010, commentaires A. Couret, L'administrateur, dirigeant de droit responsable de l'insuffisance d'actif : sévérité circonstancielle de la Cour de cassation, La Semaine Juridique Entreprise et Affaires n° 37, 15 Septembre 2011, 1655.

[374] CNIL 2017, p. xx. Ce lien est également mis en avant par Wagner Wagner, B. (2019), Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. Policy & Internet, 11: 104-122. https://doi.org/10.1002/poi3.198

[375] Price II, William Nicholson, Medical Malpractice and Black-Box Medicine (February 2, 2017). I. Glenn Cohen et al., eds., Big Data, Health Law, and Bioethics (Cambridge University Press, 2018),

[376] Daniel Schönberger, Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications, *International Journal of Law and Information Technology*, Volume 27, Issue 2, Summer 2019, Pages 171–203

[377] Diamantis, Algorithms Acting Badly: A Solution from Corporate Law, 89 Geo. Wash. L. Rev. 801 (xxxx); Selbst, "Negligence and AI's Human Users, 100 Boston U. L. Rev. 1315 (2020); Price II, William Nicholson, Medical Malpractice and Black-Box Medicine (February 2, 2017). I. Glenn Cohen et al., eds., Big Data, Health Law, and Bioethics (Cambridge University Press, 2018), U of Michigan Public Law Research Paper No. 536, Available at SSRN: https://ssrn.com/abstract=2910417; Greenberg M, Ridgely MS. Clinical Decision Support and Malpractice Risk. *JAMA.* 2011;306(1):90–91. doi:10.1001/jama.2011.929; Price II, William Nicholson, Medical Malpractice and Black-Box Medicine (February 2, 2017). I. Glenn Cohen et al., eds., Big Data, Health Law, and Bioethics (Cambridge University Press, 2018), U of Michigan Public Law Research Paper No. 536, Available at SSRN: https://ssrn.com/abstract=2910417

Une responsabilité pour les erreurs algorithmiques poussera les entreprises à augmenter le niveau de contrôle humain pour les décisions les plus critiques[378].

Parmi les différentes options de responsabilité proposées, la résolution du Parlement européen propose une responsabilité pour faute avec renversement de la charge de la preuve. Il incomberait à l'opérateur du système de prouver l'absence de faute en démontrant qu'il a été diligent. Ce renversement de la charge de la preuve est similaire au système imposé par le RGPD, comme nous le verrons ci-après. L'obligation de mettre en place des contrôles, notamment humains, du système d'IA s'inscrit dans une tendance plus générale de définir la faute comme l'absence de mesures préventives suffisantes[379]. La responsabilité civile aurait un rôle de prévention autant que de réparation[380].

C. Le contrôle humain et conformité RGPD

Hormis le cas des décisions entièrement automatisées, le RGPD reste muet sur le contrôle humain. Néanmoins, même s'il n'est pas mentionné expressément, le contrôle humain fera partie des "mesures techniques et organisationnelles appropriées" que le responsable du traitement doit mettre en oeuvre en application des articles 24 et 25 du RGPD[381]. Le responsable du traitement doit concevoir et mettre en place une série de mesures, dont des contrôles humains, pour assurer le respect du règlement. Les mesures doivent être "appropriées", à savoir effectives pour atteindre le but de protection, compte tenu du contexte[382]. Selon le CEPD, le contrôle humain constitue un élément de loyauté dans le traitement des données[383]. L'absence de contrôle humain pourrait constituer une violation du principe de loyauté, même hormis le cas spécifique des décisions entièrement automatisées visées par l'article 22 du règlement. Dans la logique du RGPD, les modalités du contrôle humain seraient définies en fonction de l'analyse d'impact conduite par le responsable du traitement[384]. Cette analyse évaluera les risques et identifiera les mesures envisagées pour faire face aux risques[385]. Parmi les risques figuereront les erreurs algorithmiques examinées

[378] Wagner, B. (2019), Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. Policy & Internet, 11: 104-122. https://doi.org/10.1002/poi3.198 , p. 117.

[379] J. Lefebvre, La responsabilité délictuelle face aux mesures préventives, LPA 09 sep. 2020,n° 153n3, p.5.

[380] Ibid.

[381] RGPD, art. 24(1); W. Maxwell et C. Gateau, Les sources d'inspiration du Règlement général sur la Protection des Données: la conformité, la réglementation de l'environnement, la responsabilité des produits défectueux, Enjeux Numérique, Annales des Mines, n° 2, juin 2018, p. 34.; Thibault Douville, Droit des données à caractère personnel, Gualino, Lextenso, 2021, p 204; Metallinos N., "Le principe d'accountability: des formalités préalables aux études d'impact sur la vie privée", Comm. com. électr. 201, étude 11, n° 4; G29, avis n°3/2010, 13 juill. 2010, WP 173 sur le principe de responsabilité.

[382] Pour le CEPD, "Approprié signifie que les mesures et les garanties nécessaires doivent être adaptées pour atteindre le but visé, c'est-à-dire qu'elles doivent mettre en œuvre les principes de protection des données de façon effective . Ainsi, l'exigence du caractère approprié est étroitement liée à l'exigence d'effectivité." CEPD, Lignes directrices 4/2019 relatives à l'article 25, Protection des données dès la conception et protection des données par défaut, Version 2.0 Adoptées le 20 octobre 2020, point 8.

[383] Ibid., point 70.

[384] Un système IA qui analyse des données à caractère personnel sera généralement un traitement à haut risque au regard de l'article 35 du RGPD, soit parce qu'il évalue de manière systématique et approfondie les aspects personnels concernant des personnes physiques sur la base de laquelle sont prises des décisions affectant des individus de manière significative, soit parce qu'il s'agit de catégories spéciales de données ou de surveillance systématique à grande échelle d'une zone accessible au public.

[385] RGPD, art. 35(7)(d)

dans la Partie III et les risques pour le respect des valeurs procédurales examinés dans la Partie IV. Les mesures appropriées seront construites en considérant l'utilité du contrôle humain pour atténuer chacun des risques, "l'état des connaissances, des coûts de mise en œuvre et de la nature, de la portée, du contexte et des finalités du traitement ainsi que des risques, dont le degré de probabilité et de gravité varie, que présente le traitement pour les droits et libertés des personnes physiques"[386]. L'existence de "bonnes pratiques" sur la nécessité d'une intervention humaine sera un critère important pour juger du caractère approprié ou non des mesures[387].

De manière générale, le RGPD impose des mécanismes de gouvernance qui incluent une formation et une sensibilisation des personnes, une allocation claire des responsabilités opérationnelles, et une structure de gouvernance permettant la supervision au plus haut niveau de l'exécutif[388]. L'intervention humaine s'inscrira dans ce programme de gouvernance[389].

Le système de gouvernance et de conformité envisagé par le RGPD exige non seulement de se mettre en conformité, mais de pouvoir le démontrer[390]. S'agit-il d'un renversement de la charge de la preuve comme le préconise le Parlement européen dans sa résolution du 20 octobre 2020 ? Même si la démonstration de la conformité n'est pas explicitement présentée comme une condition d'exonération de responsabilité au titre de l'article 82 du RGPD, l'idée est là, puisque l'article 82(3) du RGPD permet une exonération de responsabilité si le responsable du traitement apporte la preuve que "le fait qui a provoqué le dommage ne lui est nullement imputable". Le fait générateur de responsabilité au titre de l'article 82(1) est la violation du règlement[391]. L'article 82(3) prévoit quant à lui une exonération de responsabilité si le responsable démontre sa conformité au règlement[392]. Démontrer la conformité revient aussi à pouvoir justifier les décisions issues du système. L'acte de justification nécessitera généralement une intervention humaine, soit en amont, soit en aval, soit aux deux niveaux.[393] Insérée dans un schéma de contrôles internes et de gouvernance, l'intervention humaine nécessitera une allocation des responsabilités au sein des organes de contrôle ainsi que la mise en place d'un programme de formations[394].

D. Le projet de règlement européen AI Act

---

[386] RGPD, art. 25(1); CEPD Lignes directrices 4/2019 précité.
[387] CEPD, WP 253 op. cit., p. 14.
[388] Metallinos, Le principe d'accountability, op.cit. p. 2; CNIL, Délibération n° 2017-219 du 13 juillet 2017 portant modification du référentiel pour la délivrance de labels en matière de procédures de gouvernance tendant à assurer la protection des données
[389] La réglementation bancaire impose également des systèmes de gouvernance reposant en partie sur le contrôle humain : voy. Directive (UE) 2019/878 du Parlement européen et du Conseil du 20 mai 2019 modifiant la directive 2013/36/UE en ce qui concerne les entités exemptées, les compagnies financières holding, les compagnies financières holding mixtes, la rémunération, les mesures et pouvoirs de surveillance et les mesures de conservation des fonds propres; sur la gouvernance en tant qu' outil de prévention des risques, voy. M. Storck, Le risque, 10 ans après l'affaire Enron - . - Rapport de synthèse La Semaine Juridique Entreprise et Affaires n° 24, 14 Juin 2012, 1393.
[390] RGPD, art. 5(2)
[391] RGPD art. 82(1).
[392] W. Maxwell et C. Gateau, Les sources d'inspiration du Règlement général sur la Protection des Données: la conformité, la réglementation de l'environnement, la responsabilité des produits défectueux, Enjeux Numériques, Annales des Mines, n° 2, juin 2018, p. 34.
[393] Cummings, p. 25; Skitka et al. 2000; Yeung 2019
[394] G29, avis n°3/2010, 13 juill. 2010, WP 173 sur le principe de responsabilité, points 31 et 41.

La proposition de règlement européen AI Act oblige le fournisseur d'un système à haut risque à définir les mesures de contrôle humain nécessaires pour assurer une utilisation conforme du dispositif[395]. Ces mesures de contrôle humain, y compris la formation des personnes manipulant le dispositif, seront documentées dans la notice d'utilisation. Selon le projet de règlement, l'exploitant aura l'obligation de suivre les consignes contenues dans la notice d'utilisation. A défaut, l'exploitant s'exposera à des sanctions administratives. Même si ce n'est pas spécifié dans le projet de règlement, un exploitant qui n'appliquerait pas les consignes contenus dans la notice d'utilisation commettrait également une faute engageant sa responsabilité si la faute cause un préjudice[396].

Ainsi le contenu spécifique du contrôle humain pour les systèmes IA à haut risque sera précisé en partie dans la notice d'utilisation après une analyse de risques conduite par le fournisseur. L'analyse de risque suit la même logique que l'analyse d'impact prévue par le RGPD : identifier les risques, évaluer leur niveau de gravité et de probabilité, et définir des "mesures appropriées de gestion des risques" pour éliminer ou réduire les risques[397]. Ces mesures "prennent en considération l'état de la technique généralement reconnu, notamment tel qu'il ressort des normes harmonisées ou des spécifications communes pertinentes"[398]. La réduction des risques tiendra également compte "des connaissances techniques, de l'expérience, de l'éducation, de la formation pouvant être attendues de l'utilisateur et de l'environnement dans lequel le système est destiné à être utilisé"[399].

E. La responsabilité peut être un frein à un contrôle humain adapté

L'objectif de responsabilité et de conformité peut être détourné. Dans le rapport du Conseil de l'Europe sur la responsabilité et l'IA, Karen Yeung mentionne la possibilité qu'une intervention humaine soit prévue pour s'assurer de l'existence d'un bouc émissaire :

> "...continuer à insister pour qu'un être humain se trouve « dans la boucle » pour assurer une supervision risque de transformer les intéressés en «amortisseurs moraux», en totems dont le rôle central deviendra de prendre la faute sur eux, même s'ils ne maîtrisent que partiellement le système, et susceptibles de servir de boucs émissaires aux entreprises et organisations cherchant à se dégager de leurs responsabilités.[400]"

Ce danger est également souligné par plusieurs auteurs américains, pour qui le contrôle humain individuel (*human-in-the-loop*) peut servir de fusible pour réduire le risque de responsabilité pesant

---

[395] Proposition de règlement AI Act, arts. 9 et 14.

[396] Le Bureau européen des unions de consommateurs regrette que le projet d'AI Act ne contienne pas une disposition explicite conférant un droit d'obtenir réparation pour toute violation du règlement : BEUC, Regulating AI to Protect the Consumer - Position Paper on the AI Act, 7 oct. 2021, p. 24.

[397] Proposition de règlement AI Act, art. 9(2).

[398] Ibid., art. 9(3).

[399] Ibid., art. 9(4).

[400] Conseil de l'Europe, Comité d'experts sur les dimensions des droits de l'homme dans le traitement automatisé des données et les différentes formes d'intelligence artificielle (MSI-AUT), Étude sur les incidences des technologies numériques avancées (dont l'intelligence artificielle) sur la notion de responsabilité, sous l'angle des droits humains, rapporteuse Karen Yeung, DGI(2019)05, 2019, p. 68; Elish, M.C. (2016) : 'Letting Autopilots Off the Hook: Why do we blame humans when automation fails?' 16 juin. Disponible à l'adresse suivante :http://www.slate.com/articles/technology/future_tense/2016/06/why_do_blame_humans_when_automation_fails.html

sur l'entreprise[401]. Crootof et alii mentionnent le cas où les règles de responsabilité civile pourraient avoir l'effet inverse, à savoir réduire la motivation d'insérer un humain dans la boucle lorsque le risque d'erreur humaine est supérieur au risque d'erreur algorithmique[402].

Un autre obstacle au déploiement d'un contrôle humain est son coût. Dans le cadre de la lutte contre la détection de blanchiment de capitaux et de financement de terrorisme (LCB-FT) certains établissements financiers emploient des centaines de personnes pour revoir les signalements algorithmiques. Il en est de même pour les grands réseaux sociaux qui emploient des milliers de personnes, souvent en sous-traitance, pour revoir les signalements de contenus prohibés par les conditions d'utilisation du site. Si le nombre de personnes employées pour le contrôle humain devient un critère de conformité en soi[403], cela peut inciter les entreprises à sur-investir dans le contrôle humain, quelle que soit son efficacité.

F. Définir un niveau de contrôle humain "approprié" dans le cadre d'une analyse de risque

Lorsque les règles de responsabilité et de conformité ont un objectif de prévention, les règles doivent inciter à l'adoption de mesures "appropriées", à savoir suffisamment strictes pour prévenir la plupart des préjudices, sans forcément atteindre un taux zéro de préjudice. Dans l'analyse économique du droit, le niveau optimal de l'investissement dans le contrôle humain se situerait au point où le coût marginal d'un contrôleur humain supplémentaire est égal au coût de l'erreur évité, ou des valeurs de procédure préservées, par cet investissement supplémentaire[404]. Un certain niveau résiduel d'erreur, et un certain niveau de non-respect de valeurs de procédure, seraient tolérés. Cette analyse économique des règles de responsabilité atteint ces limites lorsque les préjudices concernent les droits fondamentaux, pour lesquels toute évaluation économique est contestable. Ainsi la formule de Hand sur le coût des accidents ne pourra s'appliquer telle quelle au contrôle humain, faute d'un moyen objectif de quantifier le respect des valeurs de procédure. Néanmoins, le RGPD adopte implicitement cette logique économique lorsqu'il impose la mise en place de mesures "appropriées" [405]. A travers le mécanisme d'analyse d'impact, le RGPD oblige le responsable du traitement à identifier les risques et mettre en place des mesures pour les réduire à un niveau acceptable, sans

[401] Crootof, Rebecca and Kaminski, Margot E. and Price II, William Nicholson, Humans in the Loop (March 25, 2022). Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011, Available at SSRN: https://ssrn.com/abstract=4066781 or http://dx.doi.org/10.2139/ssrn.4066781, p. 20; Madeleine Claire Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI., TECH., & SOC'Y 40 (2019);

[402] Crootof et al., op cit., p. 26.

[403] Cette critique est parfois exprimée dans le cadre de la conformité LCB-FT. Voy. notamment M. Sciurba, 'The Incompatibility of Global Anti-Money Laundering Regimes with Human and Civil Rights – Reform Needed?' (Nomos, Baden-Baden, 2019) 99.

[404] Deffains, Bruno. « Analyse économique de la responsabilité civile », *Archives de philosophie du droit*, vol. 63, no. 1, 2021, pp. 117-140; W. Maxwell, Smart(er) Internet Regulation Through Cost-Benefit Analysis - Measuring harms to privacy, freedom of expression, and the internet ecosystem, Presses des Mines, 2017.

[405] W. Maxwell et D. Ouandji, Les mesures appropriées et analyses de risques dans le RGPD, La Revue du DPO, Retrospective 2018, Université Paris 1, Mars 2019, pp 17-30; CEPD, Lignes directrices 4/2019 relatives à l'article 25, Protection des données dès la conception et protection des données par défaut, Version 2.0 Adoptées le 20 octobre 2020.

définir ce qu'est un niveau acceptable[406]. Le RGPD admet une prise en compte du coût des mesures, ainsi que l'état des connaissances[407].

La proposition de règlement AI Act adopte une méthodologie similaire, avec un objectif de déployer des mesures de gestion de risque permettant d'atteindre un niveau de risque "acceptable"[408]. Les mesures tiennent compte des "connaissances techniques, de l'expérience, de l'éducation, de la formation pouvant être attendues de l'utilisateur et de l'environnement dans lequel le système est destiné à être utilisé"[409]. Selon la proposition de règlement, le contrôle humain "vise à prévenir ou à réduire au minimum les risques", sans préciser ce qu'est le "minimum"[410]. Comme la proposition de règlement AI Act, la proposition de loi américaine Algorithmic Accountability Act 2022 imposerait à l'exploitant de l'algorithme l'obligation d'identifier les risques et mettre en oeuvre des mesures pour les éliminer ou les "réduire de manière raisonnable"[411]. L'existence de certains risques résiduels est donc possible, mais l'exploitant devra justifier sa démarche[412].

Définir ce qu'est un risque "acceptable", "réduit au minimum", ou "réduit de manière raisonnable", fera appel nécessairement à une appréciation subjective, dépendant en partie de l'utilité sociale du système. Pour un système qui permet de sauver de nombreuses vies humaines, le niveau de risque "acceptable" sera généralement plus élevé que pour un système ayant peu d'utilité sociale. La prise en compte des bénéfices sociaux découlant de l'activité est expressément admise par la FTC dans son évaluation de pratiques déloyales[413]. Le niveau de risque "acceptable" sera plus élevé pour une activité apportant un bénéfice important aux consommateurs que pour une activité dont l'utilité sociale est faible voire nulle[414]. Même la Cour Suprême, dans sa jurisprudence sur les garanties de procédure, met en équilibre le coût des mesures de procédure proposées, et leur utilité dans la manifestation de la vérité[415].

La solution pour un contrôle humain efficace et approprié réside dans l'identification des risques liés à l'absence de contrôle humain – risques liés aux erreurs algorithmiques, risques liés aux valeurs de procédure – et l'évaluation des mesures de contrôle humain (contrôles individuels, systèmes) les plus adaptées pour réduire ces risques à un niveau acceptable. Ce qu'est un niveau acceptable du risque

---

[406] Groupe de travail Article 29 sur la protection des données, Lignes directrices concernant l'analyse d'impact relative à la protection des données (AIPD) et la manière de déterminer si le traitement est «susceptible d'engendrer un risque élevé» aux fins du règlement (UE) 2016/679, WP 248 rev. 01, 4 oct., p. 22, "dès lors, il appartient au responsable du traitement d'évaluer les risques pour les droits et libertés des personnes concernées et d'identifier les mesures envisagées pour réduire ces risques à un niveau acceptable…."

[407] RGPD, art. 25; CEPD, Lignes directrices 4/2019 relatives à l'article 25, Protection des données dès la conception et protection des données par défaut, Version 2.0 Adoptées le 20 octobre 2020.

[408] Proposition de règlement AI Act, art. 9(4).

[409] Ibid.

[410] Proposition de règlement AI Act, art. 14(2).

[411] Proposition de loi Algorithmic Accountability Act of 2022, Sec. 4(a)(9)(B).

[412] Ibid., Sec. 4(a)(9)(C).

[413] Federal Trade Commission (FTC) (1980), "FTC Policy Statement on Unfairness," 17 December, available at http://www.ftc.gov/public-statements/1980/12/ftc-policy-statement-unfairness

[414] Maxwell, Winston, The Notion of 'Fair Processing' in Data Privacy Law (January 2, 2015), in "Quelle protection des données personnelles en Europe?", Céline Castets-Renard (ed.), University of Toulouse, 2015, Available at SSRN: https://ssrn.com/abstract=2544623; Winston J. Maxwell, Principles-based regulation of personal data: the case of 'fair processing', *International Data Privacy Law*, Volume 5, Issue 3, August 2015, Pages 205–216, https://doi.org/10.1093/idpl/ipv013

[415] *Mathews v. Eldridge*, 424 U.S. 319, 1976; E. Zoller, « Procès équitable et *due process of law* », *Recueil Dalloz*, 2007, p. 517.

restera une appréciation subjective. En l'absence de lignes directrices claires, une consultation avec des parties-prenantes et/ou le régulateur sur le niveau de risque résiduel pourrait augmenter la légitimité de la décision[416].

[416] La loi de l'Etat de Washington sur le reconnaissance faciale impose la mise en consultation publique de l'analyse de risques (*accountability report*) préparée par l'exploitant de l'algorithme; une obligation de consultation est également prévue dans la proposition de loi "Algorithmic Accountability Act of 2022" Sec. 3(b)(1)(G) (proposition des Sénateurs Wyden, Booker et Clarke, 3 février 2022).

VI - LES PISTES DE RÉFLEXION POUR FAVORISER UN CONTRÔLE HUMAIN EFFICACE

A. Une confiance excessive en l'humain dans la boucle

La décision de la CJUE du 21 juin 2022 confirme la tendance actuelle à surestimer le pouvoir salvateur de "l'humain dans la boucle". Mais comme cette étude a démontré, le contrôle humain individuel ne peut pas tout faire, surtout si l'on ne définit pas avec précision ses objectifs et les moyens pour les atteindre. Pour illustrer le problème, revenons sur le cas examiné par la CJUE, le contrôle humain des signalements de risques terroristes. L'algorithme de détection s'appuie sur des critères préétablis par des experts humains, non sur un modèle de *machine learning*. L'algorithme va générer automatiquement une alerte si certains critères sont remplis, par exemple le cas d'un voyageur qui a récemment voyagé dans un pays à risque, et a payé son billet en liquide. Ces deux critères vont générer un certain nombre de fausses alertes, à savoir des personnes innocentes qui ont voyagé récemment dans un pays à risque et ont payé leur billet en liquide. Pour la CJUE, le contrôleur humain individuel doit examiner l'alerte afin de décider s'il s'agit d'un faux positif. Or, dans le scénario que nous venons d'évoquer, les deux facteurs à risque sont objectivement remplis, ce qui constitue une base légitime de suspicion. On voit mal comment le contrôle humain pourrait faire autre chose que de confirmer que ce scénario correspond bien à un risque accru d'activité terroriste, car cette conclusion a déjà été le fruit d'une réflexion humaine en amont, au moment de la définition des critères. Le contrôleur humain pourrait évidemment consulter d'autres informations par rapport au passager qui pourrait expliquer pourquoi il a voyagé dans tel ou tel pays, et a payé en liquide. Mais cette démarche nécessite l'accès à d'informations supplémentaires et éventuellement des mesures d'enquête supplémentaires, tel qu'un entretien avec le passager, pour confronter le signalement de risque à d'éventuelles explications. Or, la CJUE ne mentionne pas cette étape, donnant ainsi l'impression que l'humain pourra déceler seul des faux positifs simplement en examinant attentivement l'alerte générée ainsi que les données de passagers y afférentes. Cela nous paraît difficile.

Le deuxième questionnement concerne la détection des discriminations. Selon la CJUE, le rôle du contrôle humain individuel est également de détecter des signalements discriminatoires. Si l'on reprend l'exemple d'un scénario qui génère une alerte à cause d'un voyage récent en Syrie et le paiement en liquide d'un billet, le résultat positif risque de cibler de manière disproportionnée les personnes ayant des liens au Moyen-Orient. Le scénario est-il de ce fait discriminatoire ? Et comment le contrôleur humain qui examine une alerte individuelle pourra-t-il conclure que le niveau de discrimination à l'égard de personnes ayant des liens au Moyen-Orient est excessif ? Là aussi, la CJUE donne l'impression que la détection des discriminations est à la portée des humains en charge du réexamen de chaque signalement, alors cette tâche appartient plutôt aux contrôles "système" en phase 1 ou phase 3.

Ce n'est pas le rôle de la CJUE de résoudre ces questions opérationnelles. La Cour renvoie le problème aux Etats membres qui doivent développer des solutions opérationnelles pour garantir un contrôle humain efficace sur les deux fronts : détection de faux positifs, et détection de discriminations. Si le contrôle humain individuel n'est pas en mesure de détecter des discriminations, il incombe aux Etats membres de prévoir d'autres mécanismes, et notamment des contrôles humains système, pour assurer cette tâche.

B. Le contrôle humain nécessite un encadrement réglementaire plus spécifique

Dans la première partie, on a distingué le contrôle humain "système" et le contrôle humain "individuel". Dans la section sur la typologie des erreurs, on a vu que pour certains types d'erreurs, seul un contrôle "système" sera efficace, notamment pour détecter des biais, ou définir des seuils de faux positifs par rapport aux faux négatifs. Les exigences de contrôle système sont relativement développées dans la proposition de règlement européen sur l'IA. En revanche, la question d'un contrôle individuel n'est pas traitée, sauf pour les systèmes d'identification biométrique en temps réel, pour lesquels une vérification humaine est expressément requise.

Compte tenu de la typologie des erreurs algorithmiques, les types de contrôle humain, les exigences légales de contrôle humain, et les obstacles à un contrôle humain efficace, faudrait-il prévoir un encadrement réglementaire plus spécifique pour assurer un contrôle humain individuel plus efficace ? Pour rappel, le contrôle humain individuel s'exerce pendant la phase 2, la période d'exploitation. Il peut s'effectuer en amont de la prise d'effet de la décision (contrôle individuel *ex ante*), ou en aval de la prise d'effet de la décision (contrôle individuel *ex post*). La première partie a identifié trois types de contrôle humain individuel *ex ante* - les approches "tribunal", "validation avec informations supplémentaires", et "validation sans informations supplémentaires" - et un seul type de contrôle humain individuel *ex post* - l'approche "contestation".

L'efficacité du contrôle individuel, qu'il soit *ex ante* ou *ex post*, dépend du temps et des informations disponibles au contrôleur humain. Pour les approches "tribunal" et "contestation", le temps sera moins limité que pour les approches "validation avec informations supplémentaires" et "validation sans informations supplémentaires". Pour les approches "tribunal" et "contestation", le principal défi sera d'organiser l'apport d'informations et d'arguments par la personne affectée par la décision dans un cadre qui respecte les principes d'un procès équitable, et le *due process* aux États-Unis. Pour être cohérent avec ces principes, le décideur humain devrait être indépendant et impartial, comme le recommande le Parlement européen dans sa résolution du 20 octobre 2020 et le Conseil de l'Europe dans sa recommandation du 8 avril 2020. Le RGPD et la Convention 108+ prévoient la possibilité de mettre en avant des arguments et de produire des informations dans le cas d'une contestation, mais ces textes ne vont pas jusqu'à imposer que le décideur humain soit impartial. Le futur règlement européen sur l'IA pourrait combler ce vide en prévoyant que pour les situations de contrôle humain du type "contestation", l'indépendance et l'impartialité du décideur humain devraient être prévues[417] .

La proposition de règlement européen sur l'IA prévoit un contrôle humain individuel *ex ante* pour vérifier les résultats d'identification biométrique à distance. La proposition de règlement ne fait pas de distinction entre la reconnaissance faciale et d'autres formes d'identification biométrique à distance. Or, les modalités de contrôle humain ne seront pas les mêmes. Pour la reconnaissance faciale, une vérification humaine ne nécessitera pas l'accès à d'autres informations. Les images parleront pour elles-mêmes. Pour d'autres formes d'identification biométrique à distance, la vérification humaine devra s'appuyer sur d'autres informations. Un système qui identifie des personnes à partir des mensurations de leur squelette ou de leur démarche ne pourra être contrôlé par un humain sans consulter d'autres informations. Il en serait de même pour un système qui

[417] Pour l'approche "tribunal", les règles de procédure civile, pénale et administrative prévoiront déjà l'indépendance et l'impartialité du tribunal.

identifie à distance une personne à partir du son de sa voix. La vérification humaine nécessitera la confrontation du résultat algorithmique à d'autres informations. On sera dans le même cas que pour la vérification d'une alerte de risques de terrorisme ou de blanchiment de capitaux à partir de données de connexion ou de transaction.

Ainsi, en dehors des cas où les images parlent pour elles-mêmes, le contrôle humain individuel est hautement dépendant des autres informations à la disposition du contrôleur. Imposer un contrôle humain individuel sans prévoir l'accès à ces informations supplémentaires mettrait l'humain dans une situation impossible où il doit contrôler une tâche que l'ordinateur exécute beaucoup mieux que lui.

C. Définir des tâches spécifiques pour l'humain et pour l'ordinateur

Pour éviter cette situation, il faudrait découper une tâche en plusieurs sous-tâches, dont certaines seraient confiées à l'ordinateur, et d'autres à l'humain. Pour Zerilli et ses co-auteurs[418], le contrôle humain efficace nécessite de diviser une tâche en une série de sous-tâches, et de déléguer seulement certaines sous-tâches à la machine. Dans un système d'alertes de risques terroristes, cela reviendrait à assigner à l'ordinateur la tâche de prédire le risque d'activités terroriste à partir d'un jeu de données prédéfini $x_1, x_2, x_3 \ldots x_n$, et à assigner à l'humain la tâche de confronter cette prédiction algorithmique à d'autres informations qualitatives prédéfinies, informations non prises en considération par l'algorithme. En cas d'incohérence entre la prédiction algorithmique et les informations qualitatives, le contrôleur humain aurait pour mission soit de rejeter la prédiction algorithmique, soit se référer à un niveau de décision supérieur. Dans sa démarche de vérification individuelle, le contrôleur humain n'aurait pas pour rôle en conséquence de vérifier la pertinence de la prédiction algorithmique par rapport aux données de départ ($x_1, x_2, x_3 \ldots x_n$,), car cela reviendrait à contrôler le travail de l'ordinateur et les choix des concepteurs de l'algorithme. Il n'aurait pas non plus pour tâche de vérifier l'absence de biais, ou l'absence d'erreurs dans les données d'entrée, tâches confiées au contrôle humain "système". La seule mission du contrôleur de décisions individuelles serait de vérifier la cohérence de la prédiction algorithmique par rapport à d'autres informations prédéfinies, et avoir un chemin de décision clair lorsqu'il constate une incohérence. Cette division des tâches entre ordinateur et humain, et entre le contrôle système et contrôle individuel, permettrait de compenser en partie les biais de l'automatisation, et irait dans le sens des recommandations de l'étude du Conseil de l'Europe sur la responsabilité[419]. Dans une approche similaire, Methnani et alii proposent un contrôle humain variable et adaptatif selon l'état du système [420]. Cette approche nécessite également de diviser la tâche de contrôle en sous-tâches bien définies, afin de permettre au système de solliciter, de manière dynamique, le niveau de contrôle approprié. Crootof et alii proposent également une définition claires des rôles de l'humain et de la machine, et surtout la nécessité de préciser la finalité de l'intervention humaine[421].

L'obligation de définir à l'avance ces tâches, ainsi que les informations sur lesquelles le contrôleur de décisions individuelles doit s'appuyer pour effectuer sa vérification de cohérence, aurait plusieurs avantages : premièrement, elle obligerait le concepteur et l'utilisateur de l'algorithme à tenir compte

[418] Zerilli et al, op cit., p. 560.
[419] Conseil de l'Europe, Responsabilité et IA, DGI(2019)05, Rapporteur: Karen Yeung, p. 85; Zerilli et al., op cit.
[420] Methnani et al., op cit.
[421] Crootof et al. Humans in the Loop, op cit. p. 62.

des contraintes cognitives et temporelles de la vérification individuelle *ex ante*, et organiser en amont l'accès aux informations nécessaires pour cette vérification.

Deuxièmement, elle obligerait le concepteur et l'utilisateur à assumer la responsabilité, par la mise en œuvre de contrôles humains systèmes adéquats, de la qualité des prédictions algorithmiques par rapport aux données d'entrée prédéfinies. La responsabilité du contrôleur humain individuel serait limitée à sa tâche de vérification de cohérence entre le résultat algorithmique et d'autres informations qualitatives à sa disposition. La délimitation claire de la mission du contrôleur individuel éviterait de déplacer sur lui la responsabilité d'autres erreurs dans l'algorithme, erreurs qui auraient dû être détectées dans le cadre de contrôles systèmes[422].

Troisièmement, la définition de la sous-tâche confiée à l'algorithme, de la sous-tâche confiée au contrôleur humain individuel, des sous-tâches confiées contrôleurs humains système, ainsi que des données utilisées pour chaque sous-tâche, faciliterait la mise en place de systèmes de gouvernance et de responsabilisation, la préparation de l'analyse d'impact exigée par l'article 35 du RGPD, l'analyse de risque exigée par l'article 9 du futur règlement européen sur l'IA, et le plan de gouvernance des données exigé par l'article 10 de ce futur règlement[423]. Lorsqu'un traitement est effectué par ou à la demande de l'État, la législation qui définit ce traitement devra de toute façon prévoir en détail les données utilisées ainsi que les mesures visant à garantir un traitement licite et loyal[424]. Les modalités du contrôle humain, qu'il soit "système" ou individuel, feront partie de ces mesures de protection et devront par conséquent être explicitées.

Quatrièmement, la définition précise de la mission du contrôleur humain individuel et des informations auxquelles il devra se référer pourrait libérer le contrôleur humain d'une partie de ses biais d'automatisation. Responsable d'une tâche différente de celle de l'ordinateur, mais tout aussi importante que celle-ci, l'humain restera focalisé sur les tâches pour lesquelles il garde l'entière responsabilité. Pour ces tâches-là, l'expertise métier du décideur humain sera sollicitée, permettant ainsi à l'humain de garder cette expertise par la pratique[425].

L'obligation de découper les tâches entre l'ordinateur et l'humain, et ensuite de définir les modalités du contrôle humain, devrait s'inscrire dans la démarche d'analyse de risque et de tests proposée par le futur règlement européen sur l'IA. Actuellement la proposition de règlement imposerait au fournisseur du système l'obligation de conduire une analyse de risques. Le contrôle humain constituerait l'une des mesures de protection prévues pour réduire les risques. Pour favoriser une meilleure prise en compte des spécificités du contrôle humain, la proposition de règlement pourrait être améliorée sur trois points.

Premièrement, l'analyse de risques conduite par le fournisseur du système pourrait être complétée par une analyse de risques faite par l'utilisateur, qui sera mieux placé pour identifier les risques

---

[422] Conseil de l'Europe, Responsabilité et IA, op cit., p. 68; B. Green, op. cit., p. 21.
[423] Sur l'approche de gouvernance et de responsabilisation envisagée par la proposition de règlement, voy. C. Castets-Renard, "Quelle politique européenne de l'intelligence artificielle ?" RTD Eur. 2021 p.297.
[424] RGPD, art. 6(3), et art. 23; W. Maxwell, "The GDPR and private sector measures to detect criminal activity" *Revue des Affaires européennes,* Bruylant / Larcier, 2021.
[425] Zerilli et al., op. cit. soulèvent le danger de perte d'expertise humaine.

d'exploitation[426]. Deuxièmement, ces analyses pourraient dresser une liste des types d'erreurs que le système serait susceptible de générer, et évaluer l'utilité d'un contrôle humain système et/ou individuel pour la détection de chaque type d'erreur, ainsi que les modalités de ces contrôles pour les rendre efficaces.

D. Rendre obligatoire des tests d'efficacité du contrôle humain

Enfin, l'efficacité du contrôle humain devrait faire l'objet de tests dans des conditions d'opération, ce qui rejoint les recommandations de Green[427]. Celui-ci propose des tests réguliers de systèmes humains-machines pour vérifier que les conditions de collaboration et de contrôle sont véritablement efficaces. Selon Crootof et alii la proposition de règlement européen AI Act n'impose pas suffisamment d'obligations directement sur l'utilisateur du système pour assurer le bon fonctionnement du système technico-social, à savoir l'équipe humain-machine, dans son ensemble[428]. Un programme de tests réguliers sur l'efficacité du contrôle humain obligerait l'utilisateur et le fournisseur du système à définir les critères d'efficacité de ce contrôle humain, critères qui font actuellement défaut. De tels critères existent en abondance pour mesurer le taux de faux positifs ou de biais. Mes tester le taux d'efficacité d'un contrôle humain est plus rare hormis le cas de l'aviation [429], et soulève de nombreuses questions. Or, mesurer l'effectivité du contrôle humain est déjà une obligation indirecte découlant de l'article 25 du RGPD. Pour le CEPD, le caractère "approprié" d'une mesure est liée à son effectivité, et l'effectivité doit faire l'objet de critères (KPI) et de tests[430].

Une méthodologie de test pourrait être développée pour évaluer l'efficacité d'un contrôle humain dans l'atteinte des différents buts identifiés dans cette étude : détection d'erreurs, détection de discriminations, protection de valeurs procédurales[431]. Même si les critères exactes des tests pourraient varier en fonction du contexte opérationnel, une méthodologie commune obligerait chaque fournisseur et utilisateur à définir les objectifs du contrôle humain dans son cas d'usage, et prévoir des critères de mesure, et des tests, par rapport à chaque objectif.

E. Les exigences de la CJUE comme étalon d'or du contrôle humain

En plus d'une méthodologie de tests, la réglementation, et en particulier le futur règlement européen sur l'IA, pourrait prévoir des lignes directrices sur le niveau de contrôle humain exigé en

[426] Dans son avis du 7 avril 2022, la CNCDH souligne également la nécessité d'assurer un contrôle humain au niveau de l'utilisateur. Avis du 7 avril 2022, point 62; Ebers et al., op cit, préconise la préparation d'une analyse d'impact par l'exploitant du système.
[427] B. Green, op. cit., p. 28; voy. également Ebers, op cit., p. 597.
[428] Crootof et al., Humans in the Loop, op cit., p. 67; voy. également Michael Veale & Frederik Zuiderveen Borgesius, Demystifying the Draft EU Artificial Intelligence Act, 22 COMPUT. L. REV. INT'L. 97.
[429] Le secteur de l'aviation fait exception. Les interactions entre pilotes et systèmes automatiques font l'objet de tests réguliers. V. The Role of Humans in Intelligent and Automated Systems (Le rôle de l'homme dans les systèmes automatisés intelligents) Papers presented at the RTO Human Factors and Medicine Panel (HFM) Symposium held in Warsaw, Poland, 7-9 October 2002, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.215.783&rep=rep1&type=pdf
[430] CEPD, Lignes directrices 4/2019 relatives à l'article 25, Protection des données dès la conception et protection des données par défaut, Version 2.0 Adoptées le 20 octobre 2020, points 13 et 16.
[431] Dans son article de 1974 sur les valeurs procédurales, Robert Summers a proposé des critères pour mesurer la qualité d'une procédure par rapport à ces valeurs procédurales. R. Summers, Evaluating and Improving Legal Processes, op. cit.

fonction des risques pour les droits fondamentaux. Cela aiderait dans l'élaboration des analyses de risques et dans la définition des mesures nécessaires pour réduire les risques. Pour les systèmes d'aide à la décision présentant des risques importants en matière de droits fondamentaux, les lignes directrices pourraient rappeler tout simplement les exigences de la CJUE dans sa décision du 21 juin 2022, à savoir (i) un contrôle humain phase 1 dans la définition des critères, (ii) un contrôle humain individuel *ex ante* durant la phase 2 pour détecter des faux positifs, avec un encadrement et une traçabilité de ce contrôle humain individuel, et enfin (iii) un contrôle humain "système" en phase 3 pour évaluer la pertinence des critères et détecter d'éventuelles discriminations. Pour des systèmes présentant moins de risques, le niveau des contrôles humains serait éventuellement ajusté à la baisse, mais les exigences de la CJUE seraient l'étalon d'or par rapport duquel les autres approches seraient évaluées.

En conclusion, un contrôle humain efficace nécessite de ne plus considérer ce contrôle comme une seule "chose" indivisible, mais une panoplie d'actions de contrôles plus spécifiques, chacun ayant des objectifs biens spécifiques, et chacun nécessitant des moyens spécifiques pour atteindre ces objectifs. Comme pour toute mesure de protection, l'efficacité du contrôle humain devra faire l'objet de tests, ce qui nécessitera de développer des critères de performance.