Le contrôle humain n'est pas une présence : C'est une capacité sous contrainte

Deux doctrines, un seul mot

La doctrine française de la garantie humaine et le contrôle humain de l'AI Act ne désignent pas le même objet, et le débat public les confond depuis cinq ans. Cette confusion a une conséquence pratique : on continue d'exiger un contrôle humain sans jamais mesurer s'il en est un.

La première doctrine s'est cristallisée autour de l'article L. 4001-3 du code de la santé publique, créé par l'article 17 de la loi n° 2021-1017 du 2 août 2021. Ce texte prévoit l'information de la personne concernée, l'information des professionnels de santé recourant au traitement, l'accès aux données utilisées et aux résultats qui en sont issus, et l'explicabilité du fonctionnement à la charge des concepteurs. Il n'emploie pas lui-même l'expression de garantie humaine. Celle-ci appartient à la doctrine qui a entouré le texte, de l'avis n° 130 du CCNE de mai 2019 à l'avis conjoint CCNE 141 et CNPEN 4 de novembre 2022, qui reconnaît les collèges de garantie humaine associant professionnels et représentants des patients. Cette doctrine est *pluridisciplinaire et différée* : elle organise un regard collectif sur un dispositif, en amont et en aval de son usage.

Le contrôle humain de l'article 14 du règlement (UE) 2024/1689 relève d'une autre logique. Il impose que les systèmes à haut risque soient conçus, interfaces homme-machine comprises, de telle sorte que des personnes physiques puissent les superviser effectivement pendant la période de leur utilisation. Le règlement n'exige pas que cette supervision s'exerce décision par décision, ni en temps réel. Mais dans la classe de systèmes qui nous occupe ici, l'aide à la décision en série à cadence soutenue, l'exigence devient principalement *individuelle et synchrone* : elle doit pouvoir s'exercer au poste, dans le temps utile de la décision.

On a promis la première doctrine en finançant le second dispositif. Ce texte porte sur le second, et sur cette classe de systèmes seulement. Il ne vaut pas pour les décisions rares à fort enjeu unitaire, où le temps de délibération existe par construction.

La saturation, pas le volume

L'article 14, paragraphe 4, énumère ce que l'opérateur doit pouvoir faire. On peut ranger ces capacités en trois familles. Agir, d'abord : écarter la sortie du système, refuser de l'utiliser, interrompre son fonctionnement. Comprendre, ensuite : connaître les capacités

et les limites du système, interpréter correctement ce qu'il produit, surveiller son fonctionnement. Et une troisième, d'une nature différente des deux premières : rester conscient de sa propre tendance à se fier excessivement à la sortie produite. La première famille relève de l'ergonomie, la deuxième de la formation, la troisième de la métacognition.

Rien de tout cela n'est une propriété du système. Ce sont des capacités exercées par une personne, dans un poste, sous une charge.

Le titre de cette note dit débit. C'est une métaphore de travail, assumée comme telle : elle installe d'emblée une intuition d'ingénierie là où le vocabulaire réglementaire n'offre que des qualités morales. Ce n'est pas la variable fondamentale, et il faut le dire tout de suite. Le volume n'est pas le débit de contrôle, et le débit de contrôle n'est pas la saturation. Ce qui sature un poste n'est pas le flux de décisions mais le produit de ce flux par la quantité de jugement que chaque décision réclame, rapporté à la capacité cognitive réellement disponible. Deux cents dossiers par jour dont un pour cent présente une ambiguïté réelle peuvent être parfaitement contrôlables. Vingt dossiers par jour exigeant chacun un quart d'heure d'analyse indépendante peuvent ne pas l'être. Le même poste, la même personne, deux régimes opposés.

$$S = \frac{\lambda \times J}{C}$$

où λ est la fréquence des décisions soumises au poste, J la demande moyenne de jugement par décision, et C la capacité effective de jugement disponible. Trois régimes s'en déduisent.

1. Lorsque $S \ll 1$, le poste dispose d'une réserve cognitive et la supervision peut être ce qu'elle prétend être.
2. Lorsque $S \to 1$S→1, elle se fragilise sans que rien ne le signale.
3. Lorsque $S \geq 1$, la revue indépendante systématique n'est pas dégradée : elle est structurellement impossible, et ce qui subsiste sous ce nom est une procédure de ratification.

Ce rapport est une heuristique de conception proposée ici, pas un instrument calibré emprunté à l'ergonomie cognitive. Ni J ni C ne se mesurent directement, et l'écrire ainsi ne les rend pas mesurables. Le rapport est de surcroît écrit sur des valeurs moyennes. En exploitation, sa distribution importe davantage encore : un poste soutenable en moyenne peut saturer sous l'effet de grappes de cas à forte demande de jugement, et c'est précisément à ces moments que le contrôle est le plus nécessaire. Ce que la notation apporte est un déplacement de la question : cesser de demander combien de dossiers l'opérateur traite, et commencer à demander combien de jugements le poste réclame et combien il peut en produire.

Elle apporte aussi une conséquence que la formulation littéraire cachait. L'introduction d'un système d'IA n'agit pas sur un seul terme, elle agit sur les trois, et pas dans le même sens. Elle augmente λ, puisque c'est généralement l'objet du déploiement. Elle réduit J sur les cas simples, qu'elle traite bien. Elle augmente J sur les cas résiduels, qui sont précisément ceux qu'elle traite mal et qu'elle laisse à l'humain. Et elle réduit C à long terme, par désentraînement. Un dispositif peut donc dégrader sa propre supervision sans qu'aucun de ses indicateurs ne bouge.

Lorsque ce rapport approche l'unité, la supervision ne disparaît pas : elle change de nature. Elle devient une procédure de ratification ou de rejet par défaut, dont la sortie ne dépend plus du contenu du cas mais de la configuration du poste.

Sous charge, le contrôle converge vers l'action que le poste rend gratuite. Une alerte est une interruption, l'action gratuite est de passer outre, la charge produit du rejet massif. Un poste de revue est une station de validation, l'action gratuite est d'approuver, la charge produit de la ratification. Même mécanisme, deux réglages du défaut. La variable de conception n'est donc pas la vigilance, c'est le réglage de l'action par défaut et la distance qui sépare l'opérateur de la saturation.

Quatre conditions, et elles se multiplient

Le contrôle effectif suppose quatre conditions : une capacité de jugement, une indépendance épistémique, une autorité effective, et la soutenabilité du désaccord. Elles ne s'additionnent pas. Dans l'heuristique proposée ici, elles se comportent de façon multiplicative,

$$H = f(S) \times I \times A \times D, \qquad f'(S) < 0$$

où H désigne la capacité effective de supervision et ff une fonction décroissante de la saturation définie plus haut. La forme de f n'est pas spécifiée, et ne peut pas l'être : rien ne permet de dire aujourd'hui si la dégradation est graduelle, à seuil ou brutale. La hiérarchie des deux grandeurs importe davantage que leur forme. S est une propriété du poste sous charge. H est une propriété fonctionnelle du dispositif de supervision installé dans ce poste. Une valeur proche de zéro sur un seul terme suffit à annuler la contribution des trois autres. C'est une propriété du modèle, non une régularité mesurée.

La capacité de jugement, première condition, est ce que la saturation consomme : c'est le terme $f(S)$, et il ne se lit pas dans le nombre de dossiers traités.

L'indépendance épistémique. C'est probablement la condition la plus mal comprise, parce qu'on la réduit à l'information. Il serait excessif de soutenir que juger une sortie exige toujours une information que le système n'a pas consommée : un opérateur peut détecter une erreur à partir des mêmes données, par une représentation, une méthode ou un modèle mental différents. L'indépendance requise peut être informationnelle, méthodologique, temporelle, organisationnelle ou causale. **Le vrai problème n'est pas**

le partage des données, c'est la corrélation des erreurs. Si l'humain et le système exploitent les mêmes données, s'appuient sur les mêmes variables intermédiaires et ont été formés par les mêmes pratiques institutionnelles, l'ajout de l'humain produit une redondance apparente sans gain de sûreté proportionné. L'ingénierie de sûreté connaît cela depuis longtemps sous le nom de défaillance de mode commun, et le vocabulaire des barrières successives employé dans cette note est celui de James Reason (*Human Error*, Cambridge University Press, 1990 ; *Managing the Risks of Organizational Accidents*, Ashgate, 1997).

Ce point excède largement l'article 14, et c'est ce qui le rend utile. Redondance n'est pas indépendance. Deux lecteurs formés à la même école ne constituent pas nécessairement deux barrières indépendantes. Deux modèles partageant leurs données d'entraînement et leur architecture non plus. Et le cas limite mérite d'être nommé, parce qu'il est en train de devenir la norme d'implémentation : un opérateur qui vérifie une sortie à partir de l'explication fournie par le système qui l'a produite n'est pas une seconde barrière, c'est une pseudo-redondance. L'explication améliore la compréhension et cadre simultanément le raisonnement de celui qui la reçoit ; sa provenance fait donc partie de l'architecture de la barrière, au même titre que son contenu. **La valeur d'une seconde barrière ne dépend pas de sa performance propre, mais de la corrélation de ses erreurs avec celles de la première. Deux décideurs ne font pas deux barrières lorsque leurs modes de défaillance sont corrélés.**

L'autorité. Le règlement la nomme, mais seulement là où il conçoit vraiment quelque chose : le considérant 48 et l'article 14, paragraphe 5, exigent que les personnes chargées de la supervision disposent de la compétence, de la formation et de l'autorité nécessaires. L'autorité n'est pas le bouton. Un opérateur peut disposer du temps, détecter l'erreur, avoir sous la main une commande d'infirmation parfaitement ergonomique, et n'exercer aucun contrôle effectif si son désaccord déclenche une justification hiérarchique, dégrade ses indicateurs ou peut lui être reproché ultérieurement. J'ai proposé ailleurs de nommer cet écart entre l'autorité formelle d'un acteur et son emprise réelle sur la décision. Il constitue ici la troisième condition, et c'est la seule que le règlement mentionne explicitement, dans une disposition qu'il n'applique qu'à un cas.

La soutenabilité du désaccord. Distincte de l'autorité, avec laquelle on la confond volontiers. L'autorité est la permission de s'écarter. La soutenabilité est ce que s'écarter consomme, et combien de fois par jour on peut le consommer. Dans le régime ordinaire, ratifier consomme peu et s'écarter engage, documente, ralentit et se voit. L'asymétrie n'est pas une loi de nature : elle s'inverse dès qu'une organisation demande des comptes sur les ratifications autant que sur les infirmations, ce qui reste rare. On verra plus loin que lorsque cette asymétrie se retourne, c'est en général trop tard et devant une commission d'enquête.

Ici doit être nommée la contre-thèse la plus sérieuse, celle qu'un juriste opposera en premier. L'article 14 n'impose pas un résultat cognitif. Il impose au fournisseur de rendre la supervision possible par conception, et au déployeur de la confier à des personnes disposant de la compétence, de la formation et de l'autorité nécessaires. Reprocher au règlement de ne pas garantir la vigilance revient à lui reprocher de ne pas être ce qu'il n'a jamais prétendu être. L'objection est juste au plan juridique, et c'est précisément là qu'est le problème institutionnel : une obligation de moyens conçus, non instrumentée, produit une conformité documentaire dont personne, y compris a posteriori, ne peut établir l'efficacité. Le texte est défendable. L'usage qui en est attendu ne l'est pas.

On objectera aussi que la saturation des dispositifs d'alerte tient moins au volume qu'à la faible spécificité, donc aux faux positifs. L'objection est juste, et elle précise la thèse plutôt qu'elle ne la défait : améliorer la spécificité réduit la demande de jugement par décision, c'est-à-dire agit sur le numérateur du rapport de saturation. C'est une confirmation du modèle, pas une réfutation.

Reste une remarque sur l'état du droit. Le règlement (UE) 2026/1744, entré en vigueur le 27 juillet 2026, a réécrit l'article 4 : fournisseurs et déployeurs doivent prendre des mesures pour soutenir le développement de la maîtrise de l'IA de leur personnel, sans que cette obligation leur impose de garantir un niveau déterminé pour quiconque. Il serait facile, et faux, d'y voir une contradiction avec l'article 14. Celui-ci demeure une obligation autonome, et l'assouplissement de l'article 4 ne dispense personne d'organiser une supervision effective là où elle est requise. La tension est ailleurs, et elle est plus inquiétante que la contradiction : l'article 14 définit des capacités fonctionnelles que le dispositif de supervision doit permettre d'exercer, au premier rang desquelles la compréhension des limites du système et la résistance au biais d'automation, au moment précis où l'article 4 affaiblit l'un des moyens horizontaux susceptibles d'y contribuer. L'exigence de conception reste due. Le moyen est devenu facultatif dans son intensité.

L'aviation, lue correctement

Les deux sections qui suivent changent de régime argumentatif. Elles ne procèdent pas par mesure mais par autorité et par cas : ce que la littérature des facteurs humains a établi, et ce qu'une affaire documentée montre. Les données reviennent en section 6.

L'analogie aérienne circule dans les deux sens, et généralement mal.

Elle ne prouve pas que l'humain est le maillon faible. Les statistiques attribuant la majorité des accidents à une cause humaine doivent être maniées avec prudence : la catégorie même d'erreur humaine reflète en partie la manière dont les systèmes sociotechniques attribuent causalité et responsabilité. Madeleine Clare Elish a décrit ce mécanisme sous le nom de zone de déformation morale, où l'opérateur le plus proche

absorbe une responsabilité alors même que son contrôle effectif était limité, ce qui protège l'intégrité du système technique (*Engaging Science, Technology and Society*, 2019, vol. 5, p. 40-60, DOI 10.17351/ests2019.260). C'est une objection méthodologique sérieuse, pas une réfutation générale de la statistique.

Ce que l'aviation établit est plus utile à notre propos. Lisanne Bainbridge l'a formulé en 1983 dans *Ironies of Automation* (*Automatica*, vol. 19) : l'automatisation laisse à l'humain les tâches qu'il accomplit le moins bien, la surveillance prolongée et la reprise en situation anormale, tout en dégradant par le désusage les compétences dont il aura besoin le jour où l'automatisation échoue. Le rapprochement avec le rapport de saturation est direct, et il en constitue la lecture dynamique. Bainbridge décrit exactement ce que produit l'introduction d'un système sur les trois termes : la demande de jugement J se concentre sur les cas rares et difficiles, ceux que l'automatisation ne traite pas, pendant que la capacité C s'érode faute d'exercice sur les cas ordinaires. Le numérateur augmente là où il est le plus coûteux, le dénominateur diminue en silence. Ce n'est pas une référence historique, c'est le régime transitoire du modèle.

Le secteur qui investit le plus dans la vigilance humaine documente sa dégradation depuis quarante ans. C'est un argument a fortiori, non une transposition : trafic contraint, opérateurs sélectionnés, retour d'expérience institutionnalisé, rien de comparable avec un poste qui traite des dossiers scorés à la chaîne.

Et l'aviation n'a pas obtenu sa sécurité en exigeant de la vigilance. Elle a cessé d'en dépendre : enveloppe de vol, redondance, listes de vérification opposables, coordination d'équipage, double signature, enregistreurs, retour d'expérience non punitif. Cette lecture engage l'auteur et n'est pas ici étayée par une histoire de la sécurité aérienne. Elle se vérifie néanmoins sur pièces : dans ce secteur, le contrôle humain est un processus outillé et enregistré, pas une qualité morale attendue de l'opérateur.

Dans tout le règlement, un seul endroit procède ainsi. L'article 14, paragraphe 5, impose une vérification par au moins deux personnes physiques pour l'identification biométrique à distance. Un cas, pas une méthode.

Ce que devient un contrôle non mesuré

L'affaire néerlandaise des allocations familiales est le cas le mieux documenté, et il faut se garder de lui faire porter un modèle unique. Amnesty International, dans *Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal* (2021), y décrit un système de scoring, l'usage de la nationalité comme facteur de risque, une revue manuelle requise après signalement, l'absence d'information permettant aux agents de comprendre le score, et l'absence de supervision humaine significative. L'architecture décisionnelle est celle de l'article 14 : la machine signale, l'humain examine. Mais la causalité y est plus riche que le seul effet de charge :

discrimination incorporée dans les variables, opacité de la sortie, rigidité des règles administratives, asymétrie institutionnelle, incitations de recouvrement. Le contrôle humain nominal coexistait avec un score opaque et un environnement qui empêchait la revue manuelle de constituer une barrière indépendante. Notre modèle explique une partie du mécanisme. Il n'explique pas l'affaire.

Il éclaire en revanche ce qu'il advient de l'asymétrie du coût. Pendant des années, ratifier un signalement n'a rien coûté aux agents. Puis la démission du gouvernement en janvier 2021 a rendu ces ratifications extrêmement coûteuses, rétrospectivement, et pour les personnes les moins bien placées pour les contester. C'est la zone de déformation morale à l'échelle d'une administration.

Ben Green a passé en revue quarante et une politiques imposant une supervision humaine et en tire deux conclusions (*The Flaws of Policies Requiring Human Oversight of Government Algorithms*, *Computer Law & Security Review*, 2022, vol. 45, DOI 10.1016/j.clsr.2022.105681). La première est l'incapacité empirique des personnes à remplir les fonctions attendues. La seconde, plus dérangeante, est que ces politiques légitiment l'adoption de systèmes défectueux en produisant un faux sentiment de sécurité. Le contrôle nominal ne protège pas la personne exposée à la décision. Il protège le déploiement, et il expose l'opérateur.

Le contrôle humain est un classificateur, donc il s'évalue

Si le contrôle nominal est le problème, la mesure est la réponse. Encore faut-il mesurer la bonne chose, et la mesure évidente ne mesure rien.

Deux revues systématiques ont établi la variabilité des taux d'infirmation des alertes médicamenteuses en prescription informatisée. Van der Sijs et ses coauteurs, sur dix-sept publications, rapportent des taux d'infirmation de 49 % à 96 % (*Journal of the American Medical Informatics Association*, 2006, PMID 16357358). Une revue plus récente portant sur vingt-trois études publiées entre 2000 et 2019 rapporte une fourchette de 46,2 % à 96,2 %, avec des écarts marqués selon le type d'alerte, de 46 % à 95 % pour les alertes d'allergie médicamenteuse et de 82 % à 96,8 % pour les alertes de posologie (*JMIR Medical Informatics*, 2020, PMID 32706721).

À l'autre extrémité du spectre, Rosbach et ses coauteurs ont mesuré le mouvement inverse en pathologie computationnelle : dans une expérience en ligne réunissant vingt-huit pathologistes formés, chargés d'estimer un pourcentage de cellules tumorales, l'exposition à une recommandation erronée conduit à renverser une évaluation initialement correcte dans environ 7 % des cas (arXiv:2411.00998, 2024). La pression temporelle n'augmentait pas la fréquence du phénomène mais en aggravait la sévérité. Ce que cette étude prouve : le biais d'automation existe chez des experts, sur une tâche où ils sont compétents, et il est quantifiable. Ce qu'elle ne prouve pas : une prévalence

en routine clinique, l'effectif étant de vingt-huit et le protocole expérimental. À noter, et le texte serait malhonnête de l'omettre : dans la même étude, l'intégration de l'IA améliorait significativement la performance globale.

Ces deux séries de chiffres ne se cumulent pas. La première relève de la sous-confiance, la seconde de la sur-confiance. Ce qu'elles établissent ensemble est qu'un taux d'infirmation élevé ne prouve pas un contrôle efficace, et qu'un taux faible ne prouve pas un système fiable.

La revue JMIR va d'ailleurs plus loin, et c'est son résultat le plus utile ici : les auteurs ont classé les infirmations selon leur appropriation, et obtiennent une fourchette allant de 29,4 % à 100 % selon le type d'alerte. Autrement dit, deux services affichant le même taux d'infirmation peuvent l'un exercer un contrôle pertinent, l'autre produire du bruit. Le taux ne les sépare pas. La qualification de chaque infirmation, oui.

La bonne formulation est donc celle-ci, et elle doit être bornée pour tenir. L'article 14 attribue à la supervision plusieurs fonctions : interpréter, contextualiser, refuser d'utiliser le système, l'interrompre, détecter des anomalies. Une seule d'entre elles se prête au traitement qui suit. *Lorsqu'on attribue au contrôle humain la fonction de détecter et corriger les sorties erronées, il devient lui-même un classificateur de ces erreurs. Et un classificateur s'évalue.* Il prend en entrée une sortie machine et produit un verdict binaire, ratification ou infirmation. Un classificateur s'évalue par croisement avec l'état réel.

	IA incorrecte	**IA correcte**
Humain infirme	infirmation utile	infirmation inutile
Humain ratifie	erreur laissée passer	ratification correcte

De cette table se déduisent les quatre grandeurs qui décrivent réellement un dispositif de supervision : la sensibilité du contrôle, sa spécificité, la valeur prédictive du désaccord et la valeur prédictive de la ratification. Le cadre d'évaluation est celui, banal, de tout test diagnostique. Son application au dispositif de supervision humaine est en revanche une proposition de cette note, et elle en est le cœur : elle fait passer le contrôle humain du statut de disposition à documenter à celui de fonction dont la performance se mesure.

Cette exigence prolonge une position que je défends par ailleurs sur les exigences de preuve : une organisation enregistre ses preuves ou les reconstitue, et ce qu'elle n'a pas enregistré, elle ne le démontrera pas. Le contrôle humain n'échappe pas à la règle. Il en est même le cas le plus révélateur, puisque c'est le seul dispositif de sûreté dont on exige la présence sans jamais demander le journal.

Une précaution s'impose alors, et elle vaut avertissement contre l'usage naïf de la table. Une erreur de sortie n'est pas une décision incorrecte, et une décision incorrecte n'est

pas un dommage. Une sortie fausse peut rester sans conséquence, une sortie techniquement exacte peut contribuer à une décision injuste, une infirmation juste peut arriver trop tard, une ratification fautive peut être rattrapée par une barrière située en aval. Trois couches de performance doivent donc être distinguées : celle du système, celle de la supervision, celle du dispositif sociotechnique dans son ensemble. La sensibilité du contrôle mesure la deuxième. Elle ne mesure pas la réduction finale du risque, et personne ne devrait la présenter comme telle. C'est aussi la raison pour laquelle la supervision humaine ne peut être évaluée seule : dans une architecture de barrières, la performance d'une barrière ne dit rien de la sûreté de l'ensemble tant que l'on ignore ce qui subsiste derrière elle.

Un contrôle dont on ignore la sensibilité n'est pas faible. Il n'est pas évalué.

Reste la difficulté sérieuse, et elle est réelle : dans la plupart des systèmes à haut risque, la colonne de l'état réel n'est pas disponible au moment de la décision. En crédit, en recrutement, en protection sociale, en justice, l'erreur est observable avec retard, contrefactuelle, ou contestable. Exiger la mesure sans traiter ce point serait une injonction creuse là où elle est le plus nécessaire.

Il existe des réponses, et elles sont connues des dispositifs de qualité : vérité terrain différée lorsque le résultat devient observable, double lecture indépendante sur échantillon, adjudication experte, cas étalons injectés périodiquement dans le flux, revue systématique des désaccords, analyse contrefactuelle lorsque le protocole le permet. Aucune n'est gratuite. Toutes sont moins coûteuses que la démonstration a posteriori qu'un contrôle affiché n'en était pas un.

Cette exigence n'est pas seulement une bonne pratique. Dans l'arrêt *SCHUFA Holding* du 7 décembre 2023 (C-634/21), la Cour de justice ne s'est pas demandé qui signait, mais ce qui déterminait : un score constitue en lui-même une décision individuelle automatisée dès lors que le tiers qui le reçoit s'en prévaut de manière déterminante. La Cour n'a rien jugé sur la revue humaine nominale, la question ne lui était pas posée, et il serait imprudent de lui prêter cette solution. Mais elle rappelle au point 67 que le responsable du traitement doit être en mesure de démontrer sa conformité. Une organisation qui ne connaît pas la sensibilité de son contrôle ne dispose pas, par cette seule organisation formelle de la supervision, d'une mesure permettant d'établir dans quelle proportion la sortie machine est effectivement corrigée lorsqu'elle devrait l'être.

Le projet de norme européenne qui doit préciser ce que le contrôle humain signifie techniquement, prEN 18229-3:2026, préparé par le comité CEN/CLC/JTC 21, est soumis à enquête depuis cet été. Ce qui se joue dans ce document est de savoir si l'article 14

sera opérationnalisé par des exigences mesurables ou par une liste de dispositions à documenter.

Instrumenter, inverser, déclarer, ou remonter d'un étage

Il existe un dispositif documenté qui tient, et il agit sur le rapport de saturation plutôt que sur la vigilance. L'essai MASAI, mené sur 105 934 participantes en Suède, a comparé une lecture mammographique assistée par IA à la double lecture standard. Le système oriente les examens dont le score est de 1 à 9 vers une lecture unique et ceux dont le score est de 10 vers une double lecture. Les publications rapportent une réduction de 44,2 % de la charge de lecture (Lång et al., *Lancet Oncology*, 2023, 24(8), p. 936-944, DOI 10.1016/S1470-2045(23)00298-X), une augmentation de 29 % de la détection sans hausse significative des faux positifs (Hernström et al., *Lancet Digital Health*, 2025, 7(3), e175-e183, DOI 10.1016/S2589-7500(24)00267-X), et sur le critère principal un taux de cancers d'intervalle non inférieur, 1,55 contre 1,76 pour mille, ratio 0,88 avec un intervalle de confiance à 95 % de 0,65 à 1,18 (Gommers et al., *The Lancet*, 2026, 407(10527), p. 505-514, DOI 10.1016/S0140-6736(25)02464-X).

Ce que ces chiffres prouvent : à charge de lecture réduite de 44 %, le dispositif ne dégrade pas le résultat et détecte davantage. Ce qu'ils ne prouvent pas : une supériorité sur les cancers d'intervalle, l'intervalle de confiance recouvrant l'unité. Le protocole tient par ailleurs à un dépistage organisé, une tâche perceptive et des lecteurs experts. L'humain y reste pleinement dans la boucle interprétative, et il serait faux de présenter MASAI comme une absence de supervision. L'innovation est ailleurs : le dispositif ne demande pas à une quantité constante d'attention humaine de suivre un flux augmenté, il utilise la machine pour *allouer différentiellement* cette ressource rare. En termes de saturation, il agit d'abord sur le numérateur, en réduisant la demande de jugement là où elle n'est pas nécessaire. Il préserve peut-être aussi le dénominateur, en concentrant l'exercice des lecteurs sur les cas qui le méritent, mais l'effet documenté est la baisse de charge.

D'où deux étages de réponse, qu'il faut se garder de mettre sur le même plan.

Au niveau transactionnel, trois patrons. Mesurer, c'est-à-dire connaître la sensibilité du contrôle et en faire un actif auditable. Réduire la demande de jugement, en utilisant la machine pour ramener *JJ* au niveau où le jugement existe. Ou concevoir le système de telle sorte que l'humain n'ait pas à compenser une faiblesse structurelle, ce qui revient à déclarer le dispositif de supervision pour ce qu'il est et à reporter la charge de sûreté en amont.

Au niveau institutionnel, autre chose. Green ne conclut pas à un meilleur contrôle humain, il conclut au déplacement du centre de gravité vers le contrôle institutionnel. Cela signifie cesser de demander à l'opérateur de rattraper chaque erreur, et contrôler à la place ce qui relève réellement de l'organisation : l'admission du système, les

populations concernées, les seuils, les conditions d'usage, la surveillance après déploiement, les conditions de suspension et de retrait.

Gouverner → Admettre → Exploiter → Superviser → Surveiller → Suspendre

La supervision transactionnelle occupe une place dans cette séquence, et une seule. Elle est une barrière de dernier recours au niveau de la décision, elle n'est pas l'architecture de sûreté elle-même. La traiter comme telle est précisément l'erreur que le règlement encourage sans la prescrire.

Limites

Quatre, au moins.

1. La classe de systèmes visée est étroite. Le raisonnement ne s'applique pas aux décisions rares à fort enjeu unitaire, où la ressource de jugement n'est pas la contrainte active.
2. Le corpus empirique mobilisé sur les taux d'infirmation provient très majoritairement de la santé, souvent de systèmes d'alerte à base de règles, et non de systèmes de l'annexe III. Il vaut comme argument de mécanisme, pas comme mesure transposable. Les trois auteurs qui structurent les sections 4 et 5, Bainbridge, Elish et Green, appartiennent par ailleurs à la même tradition critique des facteurs humains, ce qui oriente le cadrage.
3. Le rapport de saturation est une heuristique, pas un instrument calibré. Ses deux termes s'approchent par des indicateurs indirects dont la validité reste à établir.
4. Enfin, la mesure proposée a un effet en retour. Rendre le taux d'infirmation visible, c'est en faire un objet de pilotage, donc d'optimisation. Un dispositif évalué sur son taux de désaccord produira du désaccord. C'est un risque connu de toute métrique de contrôle, et il ne se traite que par le croisement avec l'état réel, c'est-à-dire par la table entière et non par sa première ligne.

Il faut ajouter une cinquième limite, qui est la plus sérieuse. La vérité terrain n'est pas toujours ontologique. En imagerie, un cancer d'intervalle finit par exister ou non. En recrutement, en crédit, en protection sociale, la bonne décision est normative, multidimensionnelle et souvent contrefactuelle : il n'existe pas d'état réel à opposer au verdict, seulement un jugement second qu'il faudra bien légitimer. La table diagnostique reste alors un cadre d'évaluation utile, elle cesse d'être une matrice directement mesurable. Le seuil entre les deux situations n'est pas technique, il est politique, et cette note ne le tranche pas.

Conclusion

Le contrôle humain n'existe pas parce qu'un humain est dans la boucle. Il existe si cette boucle dispose d'une capacité de jugement, d'une indépendance, d'une autorité effective et d'une soutenabilité du désaccord suffisantes pour modifier la trajectoire du système, et si l'organisation peut démontrer qu'elle le fait lorsque le système se trompe.

Formulé dans le vocabulaire de la sûreté, cela donne une exigence plus dure que celle du règlement. Soit MM une erreur du système, HH l'échec de la supervision à la corriger, et BB l'échec des barrières situées en aval. La probabilité qu'une défaillance traverse les trois s'écrit

$$\boldsymbol{P(M \cap H \cap B) = P(M) \times P(H \mid M) \times P(B, \mid MH)}$$

Cette écriture ne suppose aucune indépendance : elle en est l'exact contraire. Elle rend les dépendances visibles et impose de les estimer. La performance de la supervision doit être mesurée conditionnellement aux erreurs que le système produit effectivement, et non sur l'ensemble des cas. Celle des barrières aval doit l'être conditionnellement aux erreurs qui ont déjà franchi les barrières précédentes. La corrélation des modes de défaillance n'invalide donc pas le calcul. Elle interdit une seule chose, et c'est la pratique courante : remplacer ces probabilités conditionnelles par des taux marginaux mesurés séparément, puis les multiplier. Le contrôle humain n'est pas une propriété de conformité. C'est une barrière dont il faut connaître la probabilité conditionnelle de défaillance.

La question réglementaire pertinente n'est donc pas de savoir s'il y a un humain, mais quelle fonction de réduction du risque lui est attribuée, quelle est sa probabilité de défaillance dans les conditions réelles d'exploitation, et quelles barrières indépendantes subsistent lorsqu'il échoue. Poser la question ainsi évite de faire du contrôle humain la nouvelle divinité réglementaire après avoir démonté l'ancienne.

La présence humaine n'est pas une barrière de sûreté. Elle ne le devient que si sa fonction de réduction du risque est définie, si ses modes de défaillance sont mesurables, et si ses défaillances ne sont pas corrélées à celles du système qu'elle est censée contrôler.

Un contrôle qui ne produit jamais de désaccord n'est pas nécessairement un contrôle. Un contrôle dont on ne connaît pas la sensibilité n'est même pas évalué.