MED-TAID : Un outil pour évaluer la confiance des systèmes médicaux intelligents

Clotilde Brayé[1,2,3], Aurélien Bricout[1], Arnaud Gotlieb[2], Nadjib Lazaar[3], Quentin Vallet[2]

[1] Enovacom, France
{prenom.nom}@nehs-digital.com,
https://www.enovacom.com/
[2] Simula Research Laboratory, Norway
arnaud@simula.no
https://www.simula.no
[3] LISN, Université Paris-Saclay
{nom}@lisn.fr
https://www.lisn.upsaclay.fr/

Résumé. L'intelligence artificielle (IA) a transformé tous les aspects de la Santé en permettant de faire évoluer les systèmes médicaux intelligents (SMI). Toutefois, son utilisation comporte des risques pour les patients et les professionnels de santé. Dans cet article, nous présentons MED-TAID , un outil permettant d'assister le développement de SMI dignes de confiance. L'outil permet d'effectuer une gestion des risques éthiques en évaluant et mesurant les exigences d'une IA de confiance tout au long du cycle de vie. Nous avons effectué une première validation de MED-TAID en évaluant un SMI basé sur un modèle détectant automatiquement la COVID-19 sur des images de scanner thoracique. Son utilisation a démontré que l'évaluation des risques sur la base des critères éthiques permet de développer un SMI de confiance, en intégrant notamment des méthodes d'explicabilité pour réduire l'impact du manque de transparence du système pour le professionnel de santé.

1 Introduction

Les systèmes médicaux intelligents (SMI) intègrent des méthodes basées sur l'IA au service de la santé. L'utilisation de ces systèmes est vaste, allant de l'aide au diagnostic (Pacilè et al., 2020), au suivi des patients (Pivovarov et Elhadad, 2015), à l'aide aux soins (Saxena et al., 2022), etc. Utilisés en routine clinique, ces systèmes n'en restent pas moins vecteurs de risque, que ce soit pour les patients ou pour les professionnels de santé (Seyyed-Kalantari et al., 2021). C'est une des raisons pour lesquelles un cadre réglementaire autour du concept d'IA digne de confiance a été défini. Ce cadre repose sur trois composantes : (i) licite : respect de la loi, (ii) éthique : intégration et respect de critères éthiques et (iii) robuste : sur le plan technique et social. Ces trois composantes doivent être présentes tout au long du cycle de vie du SMI (High-Level Expert Group on Artificial Intelligence, 2019). Sur cette base, l'*AI Act* (European Commission, 2024) est le texte fondateur qui encadre la conception et l'usage des IA,

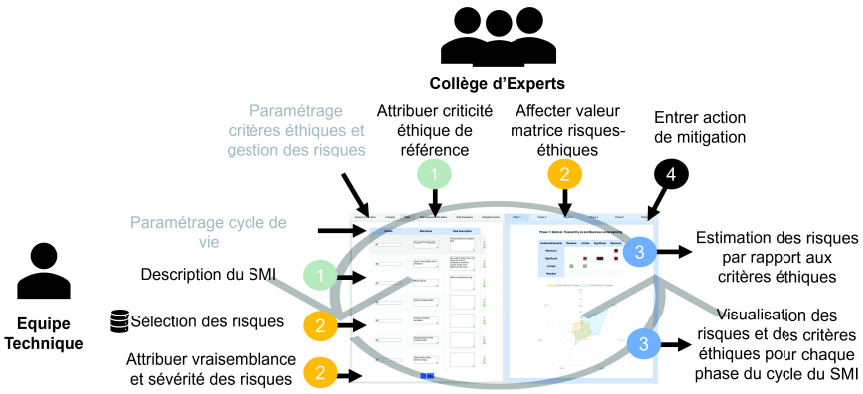


FIG. 1 – Présentation général du logiciel MED-TAID . Les numéros correspondent aux quatre fonctionnalités du logiciel décrit dans la section 1.

quel que soit le secteur d'activité. Concernant les SMI, l'*AI Act* induit la nécessité de mettre en œuvre une méthode de gestion des risques, comprenant (i) l'identification et l'évaluation des risques liés aux systèmes et aux utilisateurs de ces systèmes, (ii) la mise en place d'actions de mitigation permettant de réduire l'impact du risque. L'évaluation repose sur l'attribution de niveaux de vraisemblance et de sévérité pour chaque risque, représentés dans une matrice des risques. Cette démarche est pilotée par un collège d'experts constitué spécifiquement pour chaque SMI considéré. De nombreux travaux ont été menés pour guider l'application de la gestion des risques dans des domaines variés. Les travaux de Haakman et al. (2021) indiquent la nécessité de réviser les modèles de cycle de vie d'IA. Leur étude a menée à la révision des modèles de cycle de vie actuels (Shearer, 2000; Amershi et al., 2019; Ericson et al., 2000) en y ajoutant l'analyse des risques dans le cycle de vie une fois le modèle développé. Concernant la gestion des risques des dispositifs médicaux (DM), Khinvasara et al. (2023) explorent les défis et les meilleures pratiques pour proposer une méthodologie de gestion des risques en amont du développement. Leur processus inclut les étapes de la norme ISO 14971 (International Organization for Standardization, 2019). Vyhmeister et Castane (2024) proposent TAI-PRM, pour «*Trustworthy AI Project Risk Management*», une méthodologie pour intégrer des considérations éthiques d'une IA digne de confiance dans un processus de gestion des risques dédié à l'Industrie 5.0. Ces travaux permettent d'utiliser les métriques du processus pour évaluer les risques éthiques, mais ceci dans le contexte particulier des systèmes industriels du futur. Dans Brayé et al. (2023), nous avons proposé une méthodologie, nommée TAID , de gestion éthique des risques spécifiquement adaptée au domaine de la santé. TAID intègre l'évaluation de critères éthiques en lien avec un ensemble de risques spécifiques à un DM. Toutefois, cette méthodologie n'est pas accompagnée d'outils pratiques, ce qui limite son déploiement sur des cas d'études concrets.

Contributions. Dans cet article, nous présentons un outil nommé MED-TAID («*patients and medical staff-centered Trustworthy-AI-by-Design*») qui permet d'assister le développement d'une IA digne de confiance en santé dès sa conception et tout au long de son développement. Le logiciel permet (i) d'identifier le SMI ; (ii) d'effectuer une appréciation des risques

et de les projeter sur les critères éthiques ; (iii) d'évaluer les risques et l'impact sur les critères éthiques du SMI , et de (iv) satisfaire l'intégration d'une IA de confiance dans la gestion des risques. Ces étapes sont intégrées dans chaque phase du cycle de vie du SMI comme indiqué dans la figure 1. MED-TAID a été validé à travers un premier cas d'étude portant sur l'aide au diagnostic de la COVID-19 à partir d'images de scanner thoracique (Boussel et al., 2022).En déployant MED-TAID , nous montrons que la gestion éthique des risques a permis de développer un SMI satisfaisant les exigences de confiance.

2 Méthodologie TAID

La méthodologie TAID («*Trustworthy-AI-by-Design*») repose sur trois étapes visant à minimiser et contrôler les risques en les associant aux critères éthiques dès la conception du système d'IA (Brayé et al., 2023).

Étape 1 : Attribution de la criticité de référence. Soit $\mathcal{T} = \{T_1, \ldots, T_m\}$ un ensemble de critères éthiques, C_{ref} est le vecteur associé à la criticité de référence de $\mathcal{T}$, défini comme suit : $\forall i, C_{ref}[i] > 0$ est le degré de menace évalué pour T_i par rapport au fonctionnement du SMI. Plus le degré est proche de 0, plus l'exigence éthique est forte.

Étape 2 : Gestion éthique des risques. Soit $R = \{r_1, \ldots, r_n\}$ l'ensemble des risques identifiés pour un SMI donné. Ces risques peuvent être identifiés à partir d'une ontologie qui accompagne l'utilisateur via des moteurs de suggestion ou d'une évaluation manuelle par les experts. En notant qu'il existe un lien entre les risques $\mathcal{R}$ et les critères éthiques $\mathcal{T}$, il est possible de projeter les risques sur les critères éthiques sous forme d'une matrice Risques-Éthique. La matrice $\mathcal{M} \in \mathbb{R}^{m \times n}$ est définie comme suit :

$$\begin{bmatrix} m(T_1, r_1) & \cdots & m(T_1, r_n) \\ \vdots & \ddots & \vdots \\ m(T_m, r_1) & \cdots & m(T_m, r_n) \end{bmatrix} \text{ tel que : } \forall i \sum_{j=1}^{m} m(T_i, r_j) = 1$$

où chaque valeur $m(T_i, r_j) \in [0, 1]$ correspond à un coefficient défini soit par un collège d'experts, soit par l'expérience obtenue sur des cas d'usage.

Par ailleurs, chaque risque est caractérisé par un niveau de vraisemblance noté $V(r_i)$ et de sévérité noté $S(r_i)$. L'attribution de ces valeurs constitue la dernière étape de la gestion éthique des risques. Sur la base de la vraisemblance et de la sévérité, la quantification de chaque risque $q(r_i)$ est déterminée en utilisant une fonction standard de combinaison utilisée en gestion des risques, e.g., $q(r_i) = V(r_i) \times S(r_i)^2$. Ainsi, nous définissons $Q = [q(r_1), \ldots, q(r_n)]$ comme le vecteur de la quantification des risques.

Étape 3 : Contrôle des risques pour une IA digne de confiance. TAID repose sur un processus itératif de réduction des risques. À chaque itération k, la quantification des risques est représentée par Q_k. La criticité à l'itération k, C_k, est calculée via le produit matriciel : $C_k = \mathcal{M}.Q_k$. Les valeurs de C_k sont ensuite comparées à C_{ref}, la criticité de référence définie à l'étape 1. Si $\exists i, C_k[i] > C_{ref}[i]$, une action de mitigation doit être mise en œuvre pour réduire l'impact du risque concerné. Cette action modifie Q_k en Q_{k+1}, permettant ainsi de poursuivre le processus itératif jusqu'à ce que tous les critères soient satisfaits. Bien que la terminaison ne soit pas garantie dans le cas général, elle peut être assurée sous l'hypothèse qu'il existe toujours des actions de mitigation permettant une réduction significative des risques.

Autrement dit, si $\exists\epsilon$ suffisamment grand tel que $C_k[i] - C_{k+1}[i] > \epsilon$, alors la convergence vers un état acceptable est garantie.

3 Description de l'outil MED-TAID

Les premiers utilisateurs visés par cet outil sont les personnes en charge du développement et du déploiement du SMI, à savoir, le chef de projet, l'équipe de développement logiciel, de validation et d'intégration. Comme illustré sur la figure 1, ces utilisateurs peuvent alors identifier et évaluer les risques liés au SMI en prenant en compte les critères éthiques et proposer des actions de mitigation, jusqu'à ce que les valeurs soient satisfaisantes. Les autres utilisateurs visés sont ceux appartenant au « Collège d'Experts (CoE) ». Ce groupe d'utilisateurs est à même d'évaluer le SMI, définir les valeurs de la criticité de référence et traduire les actions de mitigation en exigences techniques pour le SMI.

3.1 Module d'appréciation des risques

Nous distinguons les données de paramétrage et les données utilisateur. Les *données de paramétrage* concernent les valeurs de $\mathcal{T}$ ainsi que les étapes du cycle de vie. Tandis que les *données utilisateur* sont les éléments de l'étape 2 de la méthodologie TAID . Notez aussi que la fonction de quantification des risques est paramétrable [1]. Dans le paramétrage de MED-TAID , nous utilisons les 7 critères Européen d'une IA de confiance : T_1 : Action humaine et contrôle humain ; T_2 : Robustesse technique et sécurité ; T_3 : Respect de la vie privée et gouvernance des données ; T_4 : Transparence ; T_5 : Diversité, non-discrimination et équité ; T_6 : Bien-être sociétal et environnemental ; T_7 : Responsabilité. Nous établissons 6 étapes successives pour le cycle de vie du SMI : 1) *Medical, Trustworthy AI and business understanding* ; 2) *Data Pipeline* ; 3) *AI Modeling and Implementation* ; 4) *Trustworthy Evaluation and Certification* ; 5) *Deployment* et 6) *Monitoring*. Nous proposons quatre niveaux pour $V(r_i)$ et $S(r_i)$: résiduel, limité, important, et maximum.

Initialement, l'utilisateur définit le périmètre de fonctionnement du SMI en décrivant la finalité du système, le cas d'usage, les données et les techniques d'IA à mettre en œuvre. Ces informations peuvent être enrichies tout au long du cycle de vie. La criticité de référence C_{ref} est définie par le CoE et est affichée sous la forme d'un *diagramme de Kiviat*. Ensuite, l'utilisateur peut sélectionner les risques spécifiques liés au SMI depuis une base de données (Slattery et al., 2024; Bilaney et al., 2024) enrichit avec une ontologies de risque en IA basé sur l'*AI Act* (Golpayegani et al., 2022), remplir la matrice $\mathcal{M}$ et qualifier les risques en attribuant un niveau de vraisemblance et de sévérité.

3.2 Module de contrôle des risques et de la criticité

A partir des données, MED-TAID calcule automatiquement la quantification des risques et la criticité dans une itération k. À partir de la quantification, MED-TAID calcule la criticité du système C_k et l'affiche sur le diagramme de Kiviat. Ainsi, sur un même diagramme, sont affichées la criticité de référence C_{ref} et C_k, celle calculée à l'itération k. L'utilisateur peut

1. Cette fonction est propre à chaque entreprise en charge du développement de dispositif médicaux.

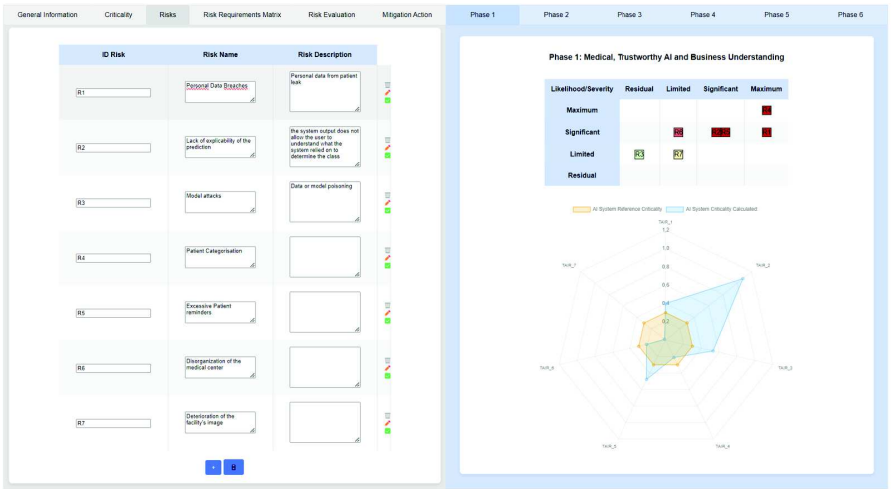


FIG. 2 – Capture d'écran de MED-TAID .

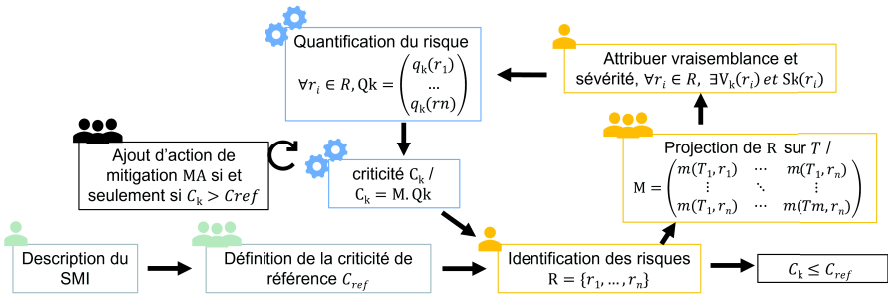


FIG. 3 – Processus de fonctionnement de MED-TAID .

ainsi détecter les axes où C_k est supérieur à C_{ref} et décider d'itérer le processus ou non. Pour chaque sommet du polygone de la criticité calculée, la ligne de la matrice $\mathcal{M}$ associée s'affiche indiquant les risques impactant le critère éthique.

3.3 Principe de fonctionnement de MED-TAID

Comme illustrée en figure 3, MED-TAID fonctionne selon un cycle dirigé par l'utilisateur. Tant que $C_k > C_{ref}$, MED-TAID invite l'utilisateur à appliquer des actions de mitigation sur les risques identifiés et itérer le calcul de la criticité. En effet, les actions de mitigation sont caractérisées par leur association à un risque et à une phase du cycle de vie et elles permettent de réduire la vraisemblance et la sévérité du risque, et par conséquent, réduire la quantification du risque.

Ce processus est itéré jusqu'à ce que $\exists k$ tel que $C_k \leq C_{ref}$ auquel cas MED-TAID assure que les risques vis-à-vis de la criticité des critères éthiques ont été suffisamment réduits par des actions de mitigation.

4 Cas d'usage : COVID-19 détection

Cette section illustre la mise en œuvre de MED-TAID dans le développement d'un modèle d'apprentissage profond visant à classifier la présence ou non de la COVID-19 sur des images de scanner thoracique. Le SMI est défini de la façon suivante : **Finalité du système :** Aide au diagnostic pour le radiologue pour la détection de la COVID-19 ; **Description du cas d'usage :** un patient avec des symptômes du COVID-19 effectue un scanner thoracique. Les images servent de données d'entrée au modèle dont la sortie sera «COVID-19» ou

NON COVID-19 ; **Caractérisation des données :** Images du scanner et métadonnées comprenant le genre des patients, leur tranche d'âge, etc. ; **Modèle d'IA :** Modèle d'apprentissage profond basé sur une architecture ResNet-50 ré-entraîné (Koonce, 2021).

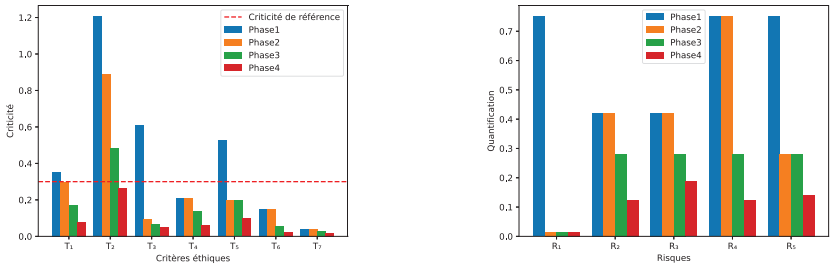
A partir de ces informations, le CoE établit la criticité de référence pour tous les critères éthiques. Puis, il identifie 5 risques auxquels sont attribués des valeurs dans la matrice risques-éthique $\mathcal{M}$: Fuite de données personnelles (r_1) Manque d'explicabilité de la prédiction (r_2) Attaque par inversion du modèle (r_3) Mauvaise prise en charge des patients (r_4) Sortie du modèle biaisée (r_5)

L'itération à la phase 1 du cycle de vie commence et les utilisateurs réalisent les étapes comme illustré dans la figure 3. A cette première itération, les critères éthiques T_1, T_2, T_3 et T_5 présentent des criticités supérieures à la criticité de référence fixée par le CoE, illustré en figure 4. Aussi, les actions de mitigation ci-dessous sont appliquées : Utilisation d'une base de données anonymisée, associée au risque r_1, évalué en Phase 2 ; Répartition égale pour les attributs protégés des patients, associée au risque r_5, évalué en Phase 2 ; Implémentation du Grad-CAM algorithme, associée au risque r_2, évalué en Phase 4 ; Utilisation d'une sigmoïde en fonction d'activation, associée au risque r_4, évalué en Phase 3. Nous montrons en figure 4, les résultats de l'application de quatre actions de mitigation pour réduire la criticité des critères éthiques.

5 Conclusion et évolutions futures

L'outil MED-TAID permet de concevoir et d'évaluer un SMI répondant aux trois exigences d'une IA digne de confiance : (i) licéité grâce à la gestion des risques, (ii) éthique via l'intégration des critères éthiques, et (iii) robustesse par l'imposition d'une criticité de référence. Il se distingue par l'introduction des concepts de criticité et de matrice risques-éthique, ainsi que par son approche matricielle de l'évaluation des risques. MED-TAID prévient les risques en intégrant les critères éthiques tout au long du cycle de vie des SMI de manière itérative.

La validation sur le cas de détection de la COVID-19 montre que l'outil aide à arbitrer les risques pour une IA digne de confiance. Cependant, l'attribution des valeurs dans la matrice risques-éthique dépend du CoE, une limite connue. Une amélioration prévue vise à quantifier les risques à priori pour atteindre la criticité éthique cible. En appliquant une résolution de



(a) Criticité (b) Quantification

FIG. 4 – Résultats de l'évolution de la criticité calculée des critères éthiques (a) et de la quantification des risques (b) en fonction des quatres premières phases.

contraintes linéaires sur $C_k = M.Q_k$, il sera possible de déterminer des valeurs de Q satisfaisant C_{ref}, guidant ainsi l'utilisateur dès l'étape d'attribution des valeurs.

Références

Amershi, S., A. Begel, C. Bird, R. Deline, H. Gall, E. Kamar, N. Nagappan, B. Nushi, et T. Zimmermann (2019). Software engineering for machine learning : A case study. In *IEEE/ACM 41st Int. Conf. on Soft. Eng. : Soft. Eng. in Practice (ICSE-SEIP)*, pp. 291–300. IEEE.

Bilaney, G., D. Edwards, et A. Kirchhof (2024). Citedrive AI incident database. https ://incidentdatabase.ai/.

Boussel, L., J. Bartoli, S. Adnane, J. Meder, P. Malléa, J. Clech, M. Zins, et J. Bérégi (2022). French imaging database against coronavirus (FIDAC) : A large COVID-19 multi-center chest CT database. *Diagnostic and Interventional Imaging 103*(10), 460–463.

Brayé, C., J. Clech, A. Gotlieb, N. Lazaar, et P. Malléa (2023). Towards trustworthy-ai-by-design methodology for intelligent radiology systems. *Platforme Française en Intelligence Artificielle (PFIA)*.

Ericson, G., W. A. Rohm, J. Martens, K. Sharkey, C. Casey, B. Harvey, et N. Schonning (2000). *Team Data Science Process Documentation*.

European Commission (2024). *Regulation of the European Parliament and of the Council. Artificial Intelligence Act*. Official Journal of the European Union. https ://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Golpayegani, D., H. Pandit, et D. Lewis (2022). *AIRO : An ontology for representing ai risks based on the proposed EU AI act and ISO risk management standards*, pp. 51–65. IOS Press.

Haakman, M., L. Cruz, H. Huijgens, et A. Van Deursen (2021). AI lifecycle models need to be revised : An exploratory study in fintech. *Empirical Soft. Eng. 26*(5), 95.

High-Level Expert Group on Artificial Intelligence (2019). *Ethics guidelines for trustworthy AI*. LU : Publications Office of the European Union.

International Organization for Standardization (2019). *Medical devices — Application of risk management to medical devices* (ISO Standard No. 14971 :2019 ed.). International Organization for Standardization.

Khinvasara, T., S. Ness, et N. Tzenios (2023). Risk management in medical device industry. *J. Eng. Res. Rep 25*(8), 130–140.

Koonce, B. (2021). Resnet 50. *Convolutional neural networks with swift for tensorFlow : image recognition and dataset categorization*, 63–72.

Pacilè, S., J. Lopez, P. Chone, T. Bertinotti, J. Grouin, et P. Fillard (2020). Improving breast cancer detection accuracy of mammography with the concurrent use of an AI tool. *Radiology : Artificial Intelligence 2*(6).

Pivovarov, R. et N. Elhadad (2015). Automated methods for the summarization of electronic health records. *Jour. of the American Medical Informatics Ass. 22*, 938–947.

Saxena, S., B. Jena, N. Gupta, S. Das, D. Sarmah, P. Bhattacharya, T. Nath, S. Paul, M. Fouda, et M. e. a. Kalra (2022). Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine. *Cancers 14*(12), 2860.

Seyyed-Kalantari, L., H. Zhang, M. B. McDermott, I. Chen, et M. Ghassemi (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature medicine 27*(12), 2176–2182.

Shearer, C. (2000). The CRISP-DM model : the new blueprint for data mining. *Journal of data warehousing 5*(4), 13–22.

Slattery, P., A. Saeri, E. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, et N. Thompson (2024). The ai risk repository : A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv :2408.12622*.

Vyhmeister, E. et G. Castane (2024). TAI-PRM : trustworthy AI—project risk management framework towards industry 5.0. *AI and Ethics*, 1–21.

Summary

Artificial Intelligence (AI) has advanced healthcare by improving intelligent medical systems (IMS). However, its use involves risks for patients and healthcare professionals. This paper introduces MED-TAID , a tool designed to support the development of trustworthy IMS. The tool enables ethical risk managment by assessing and quantifying compliance with trustworthy requirements throughout the IMS lifecycle. We evaluated MED-TAID on a ML model for automatically detecting COVID-19 in chest CT scans. Its use has demonstrated that risk assessment on the basis of ethical requirements enables the development of a trustworthy MIS, in particular by integrating explicability methods to reduce the impact of the system's lack of transparency for healthcare professionals.