

La gestion intégrée des risques de l'intelligence artificielle : enjeux techniques, réglementaires et sociotechniques pour la gouvernance des systèmes d'information

Integrated risk management of artificial intelligence: technical, regulatory and socio-technical challenges for information systems governance.

Auteur 1 : Yassin SELOUANI.

Auteur 2 : Hassan CHIDOUD.

Auteur 3 : Abdelhak TALI.

Yassin SELOUANI, <https://orcid.org/0009-0002-0096-9877>, Docteur
Université Abdelmalek Essaâdi, Ecole Nationale de Commerce et de Gestion, Tanger, Maroc

Hassan CHIDOUD, <https://orcid.org/0009-0000-3637-335X>, Docteur
Université Abdelmalek Essaâdi, Faculté des Sciences et Techniques, Tanger, Maroc

Abdelhak TALI, Enseignant chercheur
Université Abdelmalek Essaâdi, Faculté des Sciences, Juridiques, Economiques et Sociales, Tanger, Maroc

Déclaration de divulgation : L'auteur n'a pas connaissance de quelconque financement qui pourrait affecter l'objectivité de cette étude.

Conflit d'intérêts : Les auteurs ne signalent aucun conflit d'intérêts.

Pour citer cet article : SELOUANI .Y, CHIDOUD & TALI .A (2026) « La gestion intégrée des risques de l'intelligence artificielle : enjeux techniques, réglementaires et sociotechniques pour la gouvernance des systèmes d'information », African Scientific Journal « Volume 03, Num 38 » pp: 001 – 024.



DOI : 10.5281/zenodo.23664451.
Copyright © 2026 – ASJ



Résumé

L'intégration croissante de l'intelligence artificielle dans les systèmes d'information transforme le traitement et l'exploitation des données, tout en exposant les organisations à de nouveaux risques techniques, réglementaires et sociotechniques. La gestion de ces risques constitue ainsi un enjeu important pour une utilisation sûre et responsable de ces technologies.

Cette étude vise à examiner comment les organisations peuvent identifier, évaluer et maîtriser les risques liés à l'intelligence artificielle en articulant les dimensions techniques, réglementaires et sociotechniques, souvent abordées séparément.

L'étude repose sur une synthèse narrative de la littérature récente. L'analyse est organisée autour de trois ensembles de travaux portant sur les vulnérabilités techniques et les mécanismes de défense, la gouvernance de l'intelligence artificielle et la conformité réglementaire, ainsi que les risques sociotechniques et éthiques.

Les résultats montrent que les réponses techniques réduisent certaines vulnérabilités sans garantir une protection durable face à l'évolution des menaces et peuvent entraîner des arbitrages, notamment en matière de confidentialité. Les cadres de gouvernance offrent des repères importants, mais leur mise en œuvre organisationnelle demeure complexe. Les risques sociotechniques restent, quant à eux, insuffisamment intégrés aux processus d'évaluation et de décision.

L'étude met en évidence une fragmentation persistante entre ces différentes approches. Elle souligne l'intérêt d'une approche intégrée de la gestion des risques, associant l'évaluation continue des vulnérabilités, une gouvernance adaptée au niveau de risque, la conformité réglementaire et la prise en compte des conséquences sociotechniques dans les systèmes d'information.

Mots clés : intelligence artificielle ; gestion des risques ; gouvernance de l'IA ; systèmes d'information ; conformité réglementaire ; risques sociotechniques.

Abstract

The growing integration of artificial intelligence into information systems is transforming the processing and use of data while exposing organizations to new technical, regulatory, and sociotechnical risks. Managing these risks has therefore become an important challenge for the safe and responsible use of these technologies.

This study examines how organizations can identify, assess, and manage the risks associated with artificial intelligence by bringing together technical, regulatory, and sociotechnical dimensions that are often addressed separately.

The study is based on a narrative review of recent literature. The analysis is structured around three main areas: technical vulnerabilities and defense mechanisms, artificial intelligence governance and regulatory compliance, and sociotechnical and ethical risks.

The findings show that technical responses can reduce certain vulnerabilities without ensuring lasting protection against evolving threats and may involve trade-offs, particularly regarding privacy. Governance frameworks provide important guidance, but their implementation within organizations remains challenging. Meanwhile, sociotechnical risks are still insufficiently integrated into risk assessment and decision-making processes.

The study highlights persistent fragmentation among these approaches. It underlines the value of an integrated risk management approach combining continuous vulnerability assessment, risk-proportionate governance, regulatory compliance, and consideration of sociotechnical consequences within information systems.

Keywords : artificial intelligence; risk management; artificial intelligence governance; information systems; regulatory compliance; sociotechnical risks.

Introduction

L'intelligence artificielle occupe désormais une place croissante dans les systèmes d'information des organisations. Elle est mobilisée pour automatiser certaines décisions, analyser de grands volumes de données, détecter des anomalies, assister les utilisateurs ou encore produire et organiser des contenus. Qu'il s'agisse de l'évaluation d'une demande de crédit, de la détection d'une fraude, de l'aide au diagnostic ou de la génération de texte, ces usages reposent de plus en plus sur des systèmes capables d'influencer directement les processus organisationnels et les décisions qui en découlent. Cette diffusion rapide s'accompagne toutefois d'un constat devenu difficile à ignorer : l'intelligence artificielle ne permet pas seulement de réduire certains risques, elle peut également en créer de nouveaux.

Ces risques présentent des caractéristiques particulières. Ils ne résultent pas toujours d'une défaillance technique clairement identifiable, mais peuvent émerger de l'interaction entre les modèles, les données, les utilisateurs et l'environnement organisationnel dans lequel le système est déployé (Selbst et al., 2019). Un système peut ainsi obtenir des résultats satisfaisants dans un environnement contrôlé tout en révélant, en situation réelle, des vulnérabilités de sécurité, des biais ou des comportements inattendus. Les travaux consacrés aux attaques adverses ont notamment montré que des modifications parfois difficilement perceptibles peuvent conduire un modèle à produire des résultats erronés (Kurakin et al., 2018). L'essor des grands modèles de langage et des systèmes plus autonomes a, par ailleurs, élargi la nature des menaces auxquelles les organisations doivent faire face.

La gestion de ces risques ne relève donc pas exclusivement des spécialistes de la sécurité informatique. Elle soulève également des questions de gouvernance de l'information, de qualité et de protection des données, de conformité réglementaire, de responsabilité et de supervision humaine. Les systèmes d'intelligence artificielle constituent désormais une composante à part entière de nombreux systèmes d'information, qu'il s'agisse de plateformes de gestion des connaissances, de systèmes documentaires ou d'applications destinées à soutenir les activités décisionnelles. Leur déploiement modifie ainsi les conditions dans lesquelles l'information est produite, traitée, partagée et utilisée, tout en introduisant de nouvelles sources d'incertitude pour les organisations (Ibrahim, 2025).

Face à ces enjeux, plusieurs réponses se sont développées. Les recherches techniques se sont principalement intéressées à l'identification des vulnérabilités et au développement de mécanismes de défense, notamment l'entraînement adverse, qui vise à améliorer la robustesse des modèles face à des entrées volontairement manipulées (Goodfellow et al., 2015). Parallèlement, les organismes de normalisation, les institutions publiques et les régulateurs ont

élaboré différents cadres destinés à orienter la gouvernance et la gestion des risques liés à l'IA. Le cadre de gestion des risques du NIST et le règlement européen sur l'intelligence artificielle constituent, à cet égard, des références importantes. D'autres travaux se concentrent enfin sur les conséquences éthiques et sociotechniques de l'IA, en cherchant à identifier les préjudices que ces systèmes peuvent produire pour les individus et la société.

Ces différentes approches répondent à des préoccupations complémentaires, mais elles restent souvent développées dans des espaces disciplinaires distincts. Les vulnérabilités techniques sont généralement étudiées selon des critères de robustesse et de sécurité, tandis que les cadres réglementaires s'intéressent davantage aux obligations, aux responsabilités et aux mécanismes de contrôle. Les approches sociotechniques, quant à elles, permettent d'élargir l'analyse aux conséquences qui ne peuvent être réduites au fonctionnement d'un système. Cette séparation pose un problème pour les organisations, qui doivent gérer simultanément l'ensemble de ces risques dans des environnements informationnels de plus en plus complexes.

Dès lors, la question n'est plus seulement de savoir comment rendre un système d'intelligence artificielle techniquement plus performant ou plus robuste. Elle consiste également à déterminer comment identifier, évaluer et maîtriser les risques associés à son utilisation dans une organisation. Cette question est particulièrement importante dans le domaine des systèmes d'information, où les enjeux techniques se combinent avec la gouvernance des données, les exigences de conformité et les effets produits sur les utilisateurs et les processus organisationnels.

Le sujet de cette recherche porte sur la gestion intégrée des risques liés à l'intelligence artificielle dans les systèmes d'information des organisations, à l'articulation de leurs dimensions techniques, réglementaires et sociotechniques. Cet article propose d'examiner cette problématique à travers une synthèse de la littérature consacrée à trois dimensions complémentaires : les risques techniques et les limites des mécanismes de défense, les cadres de gouvernance et de conformité réglementaire, ainsi que les risques sociotechniques et éthiques. L'objectif de cette recherche est de mettre en relation ces travaux afin de répondre à la question suivante : **comment les organisations peuvent-elles identifier, évaluer et maîtriser les risques propres à leurs systèmes d'intelligence artificielle en articulant les dimensions techniques, réglementaires et sociotechniques de la gestion des risques ?**

L'intérêt de cette approche réside dans le constat que ces trois dimensions restent encore largement cloisonnées, alors qu'elles se rencontrent concrètement dans les systèmes d'information déployés par les organisations. Une mesure technique peut avoir des conséquences sur la confidentialité des données, tandis qu'un dispositif de conformité peut ne

pas permettre d'anticiper certains préjudices sociaux ou organisationnels. La gestion des risques liés à l'intelligence artificielle suppose ainsi une vision plus intégrée, capable de dépasser la juxtaposition des réponses techniques, réglementaires et éthiques.

La question présente également un intérêt particulier dans des contextes où les cadres réglementaires consacrés à l'intelligence artificielle restent en cours de structuration. Au Maroc, la protection des données personnelles repose notamment sur la loi 09-08, dont l'application aux traitements utilisant l'intelligence artificielle a été rappelée par la Commission nationale de contrôle de la protection des données à caractère personnel en 2025. Cette évolution illustre la nécessité, pour les organisations, d'adapter leurs pratiques de gouvernance à une technologie dont les usages et les risques évoluent rapidement.

La structure de la recherche est la suivante. La première partie présente la revue de littérature, organisée autour des trois dimensions retenues : les risques techniques et les limites des mécanismes de défense, la gouvernance et la conformité réglementaire, et les risques sociotechniques et éthiques. La deuxième partie expose la démarche méthodologique de synthèse narrative adoptée. La troisième partie présente les résultats de la synthèse, puis les discute afin de dégager les principaux éléments d'une approche intégrée de la gestion des risques de l'intelligence artificielle dans les systèmes d'information, ainsi que les implications, les limites et les perspectives de recherche. Une conclusion clôt l'article.

1. Revue de littérature

1.1. Les risques techniques et les limites des mécanismes de défense

La littérature consacrée à la sécurité de l'intelligence artificielle met en évidence une diversité de vulnérabilités susceptibles d'affecter les systèmes déployés dans les organisations. Les attaques par évasion cherchent à modifier les données au moment où elles sont analysées afin de provoquer une erreur de classification. Les attaques par empoisonnement ciblent, quant à elles, les données utilisées lors de l'apprentissage et peuvent altérer durablement le comportement du modèle. D'autres attaques visent la confidentialité des informations en tentant notamment de déterminer si certaines données ont été utilisées lors de l'entraînement d'un système (Song et al., 2019). Ces vulnérabilités ont été étudiées dans plusieurs domaines sensibles, notamment les véhicules autonomes, l'imagerie médicale et les services financiers (Sampathkumar & Veeras, 2025; Walter et al., 2025).

Face à ces risques, différents mécanismes de défense ont été proposés. Parmi les approches les plus étudiées figure l'entraînement adverse, qui consiste à intégrer, durant la phase d'apprentissage, des exemples conçus pour tromper le modèle afin d'améliorer sa résistance aux attaques. Plusieurs travaux montrent que cette approche peut renforcer la robustesse des

systèmes face à certaines menaces connues. Toutefois, les résultats disponibles soulignent également ses limites. (El Azzouzy et al., 2026; Olutimehin et al., 2025) mettent en évidence des améliorations de la robustesse dans certaines conditions, sans que celles-ci garantissent une protection durable face à des attaques plus élaborées. (Uesato et al., 2018) avaient déjà montré qu'une défense pouvait apparaître efficace lorsque son évaluation reposait sur des attaques insuffisamment exigeantes, puis révéler ses limites face à des méthodes d'optimisation plus robustes.

La littérature met également en évidence l'existence d'arbitrages entre différentes dimensions de la sécurité. Le renforcement d'un système contre une catégorie de vulnérabilités ne conduit pas nécessairement à une amélioration générale de sa sécurité. (Song et al., 2019) montrent notamment que certains mécanismes destinés à renforcer la robustesse peuvent avoir des conséquences sur la confidentialité des données. Cette interdépendance complique l'évaluation des solutions techniques et invite à dépasser une approche fondée sur l'optimisation isolée d'une seule vulnérabilité.

L'essor des grands modèles de langage a par ailleurs élargi la nature des risques techniques. Les attaques par injection d'instructions peuvent chercher à détourner le comportement d'un système ou à contourner les restrictions prévues par ses concepteurs. (Bassani & Sanchez, 2025; Zeng et al., 2024) soulignent ainsi la difficulté de garantir une protection complète contre différentes stratégies de contournement. Ces risques prennent une importance particulière lorsque les systèmes génératifs sont intégrés à des environnements organisationnels et disposent d'un accès à des données ou à d'autres ressources informationnelles.

Les systèmes dotés d'une plus grande autonomie soulèvent des interrogations supplémentaires. Lorsqu'un système ne se limite plus à produire une réponse, mais peut interagir avec d'autres applications ou exécuter certaines actions, les conséquences d'une erreur ou d'un comportement inattendu peuvent s'étendre au-delà du fonctionnement du modèle lui-même. Les travaux de (Vijayvargiya et al., 2026) illustrent ainsi les difficultés liées à l'évaluation des comportements des systèmes d'intelligence artificielle plus autonomes.

Dans l'ensemble, cette littérature montre que les vulnérabilités techniques constituent une dimension essentielle des risques liés à l'intelligence artificielle. Elle souligne cependant qu'aucun mécanisme de défense ne constitue une solution définitive et que l'efficacité des réponses dépend du type de menace, du contexte d'utilisation et de l'évolution des techniques d'attaque. Ces constats conduisent à considérer la sécurité technique comme une composante nécessaire, mais insuffisante, d'une gestion plus globale des risques liés à l'intelligence artificielle.

1.2. Gouvernance de l'intelligence artificielle et conformité réglementaire

Au-delà des vulnérabilités techniques, une partie croissante de la littérature s'intéresse aux mécanismes de gouvernance destinés à encadrer le développement et l'utilisation de l'intelligence artificielle. Ces travaux portent sur les principes, les dispositifs de contrôle, les responsabilités et les cadres permettant d'identifier et de gérer les risques associés aux systèmes d'intelligence artificielle.

La multiplication des cadres disponibles constitue toutefois un défi en soi. (Saeri et al., 2025), à partir d'une analyse de plusieurs cadres de gestion des risques publiés récemment, mettent en évidence une grande diversité de mesures et de recommandations. Cette diversité témoigne de la richesse des initiatives développées, mais elle contribue également à une certaine fragmentation des pratiques. Les organisations peuvent ainsi être confrontées à des définitions différentes des risques, à des taxonomies peu compatibles et à des exigences dont l'articulation reste complexe. (Bagehorn et al., 2025) soulignent également les difficultés d'interopérabilité entre les différentes ressources utilisées pour identifier et classifier les risques liés à l'intelligence artificielle.

La gouvernance influence également la manière dont les organisations définissent et hiérarchisent les risques. (Saari & Mügge, 2026) montrent que les cadres de gestion des risques ne constituent pas uniquement des instruments techniques ou administratifs, mais participent à déterminer les problèmes considérés comme prioritaires et les réponses jugées appropriées. La question de la gouvernance concerne ainsi directement la répartition des responsabilités et les processus par lesquels les organisations prennent des décisions relatives à l'utilisation de l'intelligence artificielle.

La conformité réglementaire constitue une autre dimension importante de cette littérature. L'existence d'un cadre juridique ou réglementaire ne garantit pas nécessairement sa traduction effective dans les pratiques organisationnelles. Les travaux de (Wagner et al., 2025) mettent notamment en évidence les difficultés rencontrées par les organisations dans la mise en œuvre de certaines exigences liées à la gouvernance des données, à la robustesse, à la précision et à la gestion des risques. Ces difficultés peuvent être particulièrement importantes pour les organisations disposant de ressources limitées ou d'une expérience réduite en matière de conformité.

Plusieurs recherches cherchent à rapprocher les exigences réglementaires et les pratiques de gestion des risques. (Weinzierl et al., 2024) proposent, par exemple, des approches permettant d'assister la classification des systèmes selon différentes catégories de risque. D'autres travaux, tels que ceux de (Eisenberg et al., 2025), s'intéressent à la possibilité de relier les taxonomies

de risques, les obligations réglementaires et les mécanismes de contrôle. Dans le secteur financier, (Giudici et al., 2024) mettent également en avant l'intérêt d'indicateurs permettant de traduire certains principes de gouvernance en critères plus directement mobilisables.

La diversité des cadres peut également devenir plus complexe lorsque les organisations opèrent dans plusieurs environnements réglementaires. (Rathi, 2025) souligne les difficultés qui peuvent résulter de l'existence d'exigences différentes, voire difficilement conciliables, selon les juridictions concernées. La gestion de la conformité devient alors étroitement liée aux capacités organisationnelles de coordination et d'adaptation.

Ainsi, la littérature montre que les cadres de gouvernance et les dispositifs réglementaires fournissent des repères essentiels pour structurer la gestion des risques liés à l'intelligence artificielle. Leur efficacité dépend néanmoins de leur capacité à être traduits en pratiques organisationnelles et à être articulés avec les dimensions techniques des systèmes ainsi qu'avec leurs conséquences sur les utilisateurs et les organisations.

1.3. Les risques sociotechniques et éthiques de l'intelligence artificielle

Une troisième catégorie de travaux adopte une perspective plus large en s'intéressant aux conséquences susceptibles de résulter de l'interaction entre les systèmes d'intelligence artificielle, les individus et leur environnement organisationnel ou social. Cette approche repose sur l'idée que les effets d'un système ne peuvent être compris uniquement à partir de ses caractéristiques techniques. Ils dépendent également des données utilisées, des conditions de déploiement, des pratiques des utilisateurs et des contextes dans lesquels les systèmes sont intégrés (Selbst et al., 2019).

Plusieurs travaux ont proposé des taxonomies destinées à identifier et à structurer ces risques. (Abercrombie et al., 2024) développent notamment une classification des préjudices liés aux systèmes algorithmiques et à l'automatisation. Dans le cas des grands modèles de langage, (Eisenberg et al., 2025) identifient différentes catégories de risques couvrant notamment la discrimination, la désinformation et les préjudices pouvant résulter des interactions entre humains et systèmes d'intelligence artificielle.

La question prend une dimension plus large avec le développement des modèles à usage général. (Uuk et al., 2024), à partir d'une revue de la littérature, identifient plusieurs catégories de risques systémiques susceptibles de concerner l'environnement, les structures sociales et les conséquences d'une perte de contrôle sur certains systèmes. Ces travaux contribuent à élargir l'analyse au-delà des risques directement associés au fonctionnement d'un système particulier. D'autres recherches s'intéressent à des formes de préjudices plus difficiles à identifier à travers les mécanismes traditionnels de gestion des risques. (Zhang et al., 2025), à partir de l'analyse

d'interactions entre utilisateurs et systèmes conversationnels, mettent notamment en évidence différentes formes de préjudices relationnels. Ces résultats montrent que certains risques peuvent émerger progressivement au cours des interactions et ne correspondent pas nécessairement à une défaillance technique clairement identifiable.

Toutefois, l'existence de taxonomies détaillées ne garantit pas leur intégration dans les pratiques organisationnelles. (Berman et al., 2026) soulignent que les classifications des risques liés à l'intelligence artificielle restent encore faiblement mobilisées dans certains processus réels de gouvernance. Ces outils permettent d'identifier un ensemble de préjudices potentiels, mais ils ne précisent pas toujours les responsabilités associées ou les mécanismes organisationnels permettant d'y répondre.

Les travaux consacrés à la personnalisation et à l'alignement des systèmes d'intelligence artificielle soulignent également la difficulté de concilier les préférences individuelles avec des limites collectives destinées à prévenir certains comportements indésirables (Kirk et al., 2023, 2024). Avec l'augmentation de l'autonomie des systèmes, d'autres interrogations concernent les conséquences humaines et sociétales à plus long terme des transformations technologiques et organisationnelles associées à leur diffusion (Critch & Russell, 2023; Kasirzadeh, 2025).

Cette littérature montre ainsi que les risques sociotechniques et éthiques dépassent largement les questions de performance ou de sécurité technique. Ils peuvent émerger des interactions entre technologies, utilisateurs et organisations et nécessitent une attention particulière aux conséquences susceptibles d'apparaître au cours des usages réels.

2. Méthode

Cet article adopte une démarche de synthèse narrative de la littérature consacrée à la gestion des risques de l'intelligence artificielle. L'objectif n'est pas de réaliser une revue systématique visant à recenser de manière exhaustive l'ensemble des publications disponibles, mais de mettre en relation plusieurs courants de recherche qui abordent les risques liés à l'IA selon des perspectives différentes et encore largement cloisonnées.

Le choix de cette démarche s'inscrit dans un positionnement épistémologique interprétativiste. Les risques liés à l'intelligence artificielle ne se réduisent pas à des propriétés techniques mesurables : ils émergent de l'interaction entre les systèmes, les données, les utilisateurs et les contextes organisationnels, et leur signification varie selon les cadres de gouvernance et les usages. Une lecture interprétative et critique de la littérature est donc apparue plus adaptée qu'une approche visant l'agrégation quantitative de résultats. Le mode de raisonnement est principalement abductif : la synthèse procède par allers-retours entre les travaux examinés et la problématique, afin de proposer une explication plausible de la fragmentation observée et d'en

tirer les composantes d'une approche intégrée. Il comporte également une dimension inductive, dans la mesure où les résultats sont dégagés à partir de la comparaison des constats rapportés dans chaque ensemble de travaux. Ce choix répond enfin à l'hétérogénéité des disciplines mobilisées (sécurité informatique, droit et gouvernance, sciences sociales) et à l'évolution rapide du domaine, qui rendent une revue systématique peu adaptée à l'objectif conceptuel de l'étude.

La recherche documentaire a été orientée autour de trois ensembles de travaux. Le premier concerne les vulnérabilités techniques des systèmes d'intelligence artificielle et les mécanismes destinés à améliorer leur sécurité et leur robustesse, notamment face aux attaques adverses. Le deuxième regroupe les travaux consacrés à la gouvernance de l'IA, à la gestion des risques et aux exigences de conformité réglementaire. Le troisième s'intéresse aux risques sociotechniques et éthiques susceptibles de résulter de l'interaction entre les systèmes d'intelligence artificielle, les utilisateurs et leur environnement organisationnel ou social.

Le tableau 1 synthétise les trois ensembles de travaux sur lesquels repose la synthèse narrative ainsi que leur objet d'analyse.

Tableau 1 : Ensembles de travaux mobilisés dans la synthèse

Dimension	Objet d'analyse	Apport recherché pour la gestion des risques
Risques techniques	Vulnérabilités des systèmes d'IA (notamment attaques adverses) et mécanismes de défense visant la sécurité et la robustesse	Identifier les menaces et les limites des défenses techniques
Gouvernance et conformité réglementaire	Cadres de gouvernance, gestion des risques et exigences de conformité	Comprendre les repères, les obligations et les difficultés de mise en œuvre
Risques sociotechniques et éthiques	Préjudices résultant de l'interaction entre systèmes, utilisateurs et environnement organisationnel ou social	Élargir l'analyse aux conséquences qui dépassent le fonctionnement technique du système

Source : élaboré par les auteurs.

La sélection des travaux s'est fondée sur leur pertinence par rapport à ces trois dimensions et sur leur contribution à la compréhension des mécanismes d'identification, d'évaluation ou de maîtrise des risques. Une attention particulière a été accordée aux publications récentes, compte tenu de l'évolution rapide des technologies et des cadres de gouvernance de l'intelligence

artificielle. Des travaux plus anciens ont néanmoins été mobilisés lorsqu'ils constituaient des références importantes pour comprendre certaines vulnérabilités techniques ou les fondements des approches actuellement développées.

L'analyse a ensuite consisté à examiner les principaux résultats et limites rapportés dans chacune des trois catégories retenues. Au-delà de la présentation des travaux, la synthèse cherche à identifier les points de convergence, les tensions et les lacunes qui apparaissent lorsqu'on envisage conjointement les dimensions techniques, réglementaires et sociotechniques. Cette mise en relation permet notamment d'interroger la capacité des organisations à gérer ces différents risques dans un cadre cohérent de gouvernance des systèmes d'information.

Cette démarche comporte certaines limites. La littérature sur l'intelligence artificielle évolue rapidement et plusieurs travaux récents, en particulier dans le domaine des grands modèles de langage et des systèmes agentiques, peuvent être rapidement complétés ou nuancés par de nouvelles publications. Par ailleurs, le caractère narratif de la synthèse implique une sélection raisonnée des sources plutôt qu'un processus d'inclusion exhaustif. Les résultats doivent donc être interprétés comme une lecture structurée des principaux enjeux et tendances identifiés dans la littérature, et non comme une mesure quantitative de l'ensemble des travaux existants.

3. Résultats et Discussion

3.1. Résultats

La synthèse de la littérature fait ressortir quatre résultats principaux concernant la gestion des risques liés à l'intelligence artificielle dans les organisations.

La persistance et l'évolution des vulnérabilités techniques

Le premier résultat concerne le caractère persistant et évolutif des risques techniques. Les travaux examinés montrent que les mécanismes de défense permettent de réduire certaines vulnérabilités, mais qu'ils ne garantissent pas une protection durable contre l'évolution des attaques. Leur efficacité dépend notamment du type de menace considéré et des conditions dans lesquelles les systèmes sont évalués.

La littérature met également en évidence l'existence d'arbitrages entre différents objectifs. Le renforcement de la robustesse d'un système peut avoir des conséquences sur d'autres dimensions, notamment la confidentialité ou la performance. L'émergence des grands modèles de langage et des systèmes plus autonomes élargit par ailleurs les surfaces d'attaque et les conséquences potentielles d'un comportement inattendu. La sécurité technique apparaît ainsi comme une condition nécessaire de la maîtrise des risques, mais elle ne peut être envisagée comme une réponse suffisante à elle seule.

Des cadres de gouvernance nombreux, mais difficiles à traduire en pratiques

Le deuxième résultat concerne l'écart entre la multiplication des cadres de gouvernance et leur mise en œuvre effective au sein des organisations. La littérature met en évidence une diversité importante de principes, de taxonomies et de mécanismes destinés à encadrer les systèmes d'intelligence artificielle. Cette diversité fournit aux organisations de nombreuses références, mais elle peut également accroître la complexité de la gestion des risques.

Les difficultés concernent notamment l'interprétation des exigences, l'attribution des responsabilités, la gouvernance des données et l'intégration des mécanismes de contrôle dans les pratiques existantes. Ces défis peuvent être particulièrement importants pour les organisations disposant de ressources limitées ou d'une expérience réduite en matière de conformité. Les résultats suggèrent ainsi que l'existence d'un cadre réglementaire ou d'un référentiel de gouvernance ne garantit pas, à elle seule, une maîtrise effective des risques.

Une intégration encore limitée des risques sociotechniques

Le troisième résultat met en évidence une prise en compte insuffisante des risques sociotechniques dans les dispositifs ordinaires de gestion des risques. La littérature identifie pourtant un ensemble croissant de conséquences susceptibles d'affecter les individus, les organisations et, plus largement, la société.

Ces risques peuvent résulter des interactions entre les systèmes, les utilisateurs et leur environnement plutôt que d'une défaillance technique clairement identifiable. Ils peuvent également apparaître progressivement après le déploiement d'un système. Malgré le développement de nombreuses taxonomies permettant d'identifier ces préjudices potentiels, leur intégration dans les processus organisationnels d'évaluation, de décision et de gouvernance reste limitée.

La fragmentation comme résultat transversal

Le résultat transversal le plus important de la synthèse concerne la fragmentation persistante entre les approches techniques, réglementaires et sociotechniques. Chaque courant apporte des éléments essentiels à la compréhension des risques liés à l'intelligence artificielle, mais leurs interactions restent encore insuffisamment prises en compte.

Cette fragmentation peut conduire à une gestion partielle des risques. Une organisation peut renforcer la sécurité technique de ses systèmes sans examiner suffisamment leurs conséquences sur les utilisateurs. Elle peut également satisfaire à certaines exigences de conformité sans disposer d'une visibilité suffisante sur les vulnérabilités techniques ou les préjudices émergents. Les résultats montrent ainsi que les différents types de risques ne doivent pas être considérés comme des catégories indépendantes. Ils sont susceptibles d'interagir et nécessitent une

coordination entre les acteurs responsables de la sécurité, des systèmes d'information, de la gouvernance, de la conformité et des usages.

Dans l'ensemble, la synthèse met donc en évidence une tension entre la richesse des connaissances disponibles dans chaque domaine et la faiblesse relative de leur articulation dans une perspective organisationnelle commune. Ce constat constitue le point de départ de la discussion suivante, consacrée aux conditions d'une approche plus intégrée de la gestion des risques liés à l'intelligence artificielle.

Le tableau 2 récapitule les quatre résultats principaux de la synthèse.

Tableau 2 : Synthèse des résultats

Résultat	Constat principal	Enseignement pour les organisations
1. Vulnérabilités techniques persistantes et évolutives	Les défenses réduisent certaines vulnérabilités sans garantir une protection durable ; des arbitrages existent (robustesse, confidentialité, performance)	La sécurité technique est nécessaire mais non suffisante
2. Cadres de gouvernance nombreux mais difficiles à mettre en pratique	Écart entre la multiplication des cadres et leur mise en œuvre effective (interprétation des exigences, responsabilités, données)	L'existence d'un cadre ne garantit pas, à elle seule, la maîtrise des risques
3. Intégration limitée des risques sociotechniques	Prise en compte insuffisante de ces risques dans les processus d'évaluation, de décision et de gouvernance	Intégrer ces risques au suivi, y compris après le déploiement
4. Fragmentation transversale	Interactions insuffisamment prises en compte entre approches techniques, réglementaires et sociotechniques	Coordonner les acteurs de la sécurité, des systèmes d'information, de la conformité et des usages

Source : élaboré par les auteurs.

3.2. Discussion

Vers une approche intégrée de la gestion des risques de l'intelligence artificielle

Les résultats de cette synthèse montrent que la principale difficulté de la gestion des risques liés à l'intelligence artificielle ne réside pas uniquement dans l'identification de nouvelles

vulnérabilités ou dans l'absence de mécanismes de contrôle. La littérature offre déjà de nombreuses solutions techniques, cadres de gouvernance et classifications des risques. Le défi central concerne davantage leur articulation au sein d'une démarche organisationnelle cohérente.

Le premier résultat souligne le caractère évolutif des vulnérabilités techniques. L'efficacité limitée ou contextuelle de certaines défenses montre qu'une évaluation réalisée uniquement avant le déploiement ne peut suffire. La gestion des risques doit donc s'inscrire dans une logique continue, permettant de réévaluer les systèmes à mesure que les menaces, les usages et les conditions d'utilisation évoluent.

Le deuxième résultat met en évidence les limites d'une approche exclusivement fondée sur la conformité. Les organisations peuvent disposer de cadres et de principes destinés à encadrer l'intelligence artificielle sans être nécessairement capables de les traduire en responsabilités, procédures et mécanismes de contrôle opérationnels. La gouvernance joue ici un rôle essentiel en créant un lien entre les risques identifiés, les décisions à prendre et les acteurs chargés d'intervenir.

La prise en compte des risques sociotechniques constitue une troisième condition importante. Ces risques ne sont pas toujours immédiatement visibles et peuvent apparaître progressivement à travers les usages réels des systèmes. Leur intégration suppose donc d'élargir l'évaluation au-delà des indicateurs techniques ou des seules exigences formelles de conformité. Les retours des utilisateurs et l'observation des conditions réelles d'utilisation peuvent notamment contribuer à identifier des conséquences qui n'avaient pas été anticipées lors de la conception. Ces trois constats conduisent à considérer la fragmentation comme le principal résultat transversal de l'étude. Les dimensions techniques, réglementaires et sociotechniques ne correspondent pas à des problèmes totalement indépendants. Elles peuvent interagir et créer des arbitrages. Une mesure destinée à renforcer la sécurité peut, par exemple, avoir des conséquences sur la confidentialité, tandis qu'un système respectant certaines exigences réglementaires peut néanmoins produire des effets indésirables dans son environnement d'utilisation.

Une approche intégrée apparaît dès lors moins comme l'adoption d'un nouveau cadre unique que comme la mise en relation de mécanismes existants. Elle suppose au minimum quatre composantes : une identification continue des vulnérabilités et des nouveaux risques, une évaluation tenant compte de la criticité du système et de ses conséquences potentielles, la mise en œuvre de mesures techniques et organisationnelles adaptées, et une surveillance permettant d'ajuster ces mesures après le déploiement.

La proportionnalité constitue également un principe important. Tous les systèmes d'intelligence artificielle ne présentent pas le même niveau de risque. Les dispositifs de contrôle doivent donc être adaptés à la nature du système, aux données mobilisées, à son degré d'autonomie et aux conséquences potentielles de ses décisions ou de ses actions.

L'enjeu est ainsi de favoriser une meilleure circulation de l'information entre les différentes fonctions concernées. Les équipes techniques disposent d'informations essentielles sur les vulnérabilités et les performances des systèmes, tandis que les responsables de la gouvernance et de la conformité interviennent dans l'interprétation des exigences et l'organisation des responsabilités. Les utilisateurs et responsables opérationnels peuvent, quant à eux, identifier des effets ou des difficultés qui apparaissent uniquement dans les conditions réelles d'utilisation.

Dans cette perspective, la gestion des risques liés à l'intelligence artificielle peut être comprise comme un processus transversal de gouvernance des systèmes d'information. Elle associe la sécurité technique, la gouvernance des données, la conformité réglementaire et l'attention portée aux conséquences sociotechniques. La contribution de cette approche consiste donc à dépasser la juxtaposition de ces dimensions et à souligner la nécessité de les considérer dans leurs interactions.

Le tableau 3 résume les quatre composantes minimales de l'approche intégrée proposée.

Tableau 3 : Composantes de l'approche intégrée de gestion des risques

Composante	Contenu	Dimensions mobilisées
Identification continue	Repérage continu des vulnérabilités et des nouveaux risques	Technique ; sociotechnique
Évaluation proportionnée	Évaluation tenant compte de la criticité du système et de ses conséquences potentielles	Technique ; réglementaire ; sociotechnique
Mesures adaptées	Mise en œuvre de mesures techniques et organisationnelles adaptées	Technique ; réglementaire
Surveillance	Suivi permettant d'ajuster les mesures après le déploiement	Technique ; sociotechnique

Source : élaboré par les auteurs à partir de la discussion.

3.3.Implications organisationnelles et contextuelles

Les résultats présentent plusieurs implications pour les organisations qui intègrent l'intelligence artificielle dans leurs systèmes d'information. La première concerne la répartition des

responsabilités. Une gestion efficace des risques suppose d'éviter une séparation excessive entre les équipes chargées du développement, de la sécurité, de la gestion des données, de la conformité et des activités opérationnelles.

La deuxième implication concerne le suivi des systèmes dans le temps. Les risques ne restent pas nécessairement identiques après le déploiement. Les performances peuvent évoluer, de nouvelles vulnérabilités peuvent apparaître et les utilisateurs peuvent développer des usages initialement non anticipés. Les organisations doivent donc disposer de mécanismes permettant de détecter les incidents, de recueillir des informations sur les usages et de réexaminer les mesures de contrôle.

Cette question est particulièrement importante dans les contextes où la gouvernance spécifique de l'intelligence artificielle demeure en cours de structuration. Les organisations peuvent alors s'appuyer sur des dispositifs existants relatifs à la protection des données, à la sécurité de l'information et à la gouvernance des systèmes d'information, tout en les adaptant aux caractéristiques propres des systèmes d'intelligence artificielle.

Le contexte marocain peut notamment être envisagé sous cet angle. La diffusion croissante des usages de l'intelligence artificielle renforce la nécessité d'intégrer ces nouvelles technologies dans les dispositifs existants de gouvernance et de protection de l'information. L'enjeu ne consiste pas nécessairement à créer immédiatement des structures entièrement nouvelles, mais à développer progressivement les capacités nécessaires pour identifier les systèmes concernés, évaluer leurs risques et clarifier les responsabilités associées à leur utilisation.

Pour les organisations disposant de ressources limitées, une approche proportionnée paraît particulièrement pertinente. La multiplication de procédures complexes ne constitue pas nécessairement une réponse adaptée à tous les systèmes. L'intensité des contrôles devrait plutôt être liée à la criticité du système, à la sensibilité des données utilisées, à son degré d'autonomie et aux conséquences potentielles de son fonctionnement.

3.4.Limites et perspectives de recherche

Cette étude présente certaines limites. La première tient au caractère narratif de la synthèse, qui repose sur une sélection raisonnée de travaux plutôt que sur une démarche d'exhaustivité propre à une revue systématique. Les résultats doivent donc être interprétés comme une lecture structurée des principaux constats et tendances identifiés dans la littérature.

La deuxième limite est liée à l'évolution rapide du domaine. Les technologies, les vulnérabilités, les mécanismes de défense et les cadres de gouvernance évoluent continuellement. Certains résultats peuvent ainsi être complétés ou nuancés par de nouvelles recherches.

Enfin, l'approche intégrée proposée dans cette étude constitue une contribution conceptuelle issue de la synthèse et n'a pas fait l'objet d'une validation empirique. Son application peut varier selon les secteurs d'activité, la taille des organisations et la nature des systèmes concernés.

Ces limites ouvrent plusieurs perspectives. Des recherches empiriques pourraient examiner comment les organisations articulent concrètement les dimensions techniques, réglementaires et sociotechniques dans leurs pratiques de gestion des risques. Des études de cas permettraient également d'analyser les mécanismes de coordination entre les équipes techniques, les responsables des systèmes d'information et les fonctions de gouvernance.

Il serait enfin utile d'étudier cette problématique dans différents contextes organisationnels et sectoriels, notamment au Maroc, afin de mieux comprendre comment les organisations adaptent leurs dispositifs existants de gouvernance des données et des systèmes d'information aux risques spécifiques de l'intelligence artificielle.

Conclusion

Cette étude avait pour objectif d'examiner la manière dont les organisations peuvent appréhender les risques liés à l'intelligence artificielle en articulant trois dimensions souvent traitées séparément dans la littérature : les risques techniques, la gouvernance et la conformité réglementaire, ainsi que les risques sociotechniques. La synthèse réalisée met en évidence la richesse des travaux disponibles dans chacun de ces domaines, tout en soulignant la persistance d'une fragmentation entre les différentes approches.

Les résultats montrent d'abord que les mécanismes techniques de sécurité et de défense constituent une composante indispensable de la gestion des risques, sans toutefois garantir une protection durable face à l'évolution des vulnérabilités et des menaces. Ils révèlent ensuite que la multiplication des cadres de gouvernance et des exigences réglementaires ne se traduit pas automatiquement par une maîtrise effective des risques, notamment en raison des difficultés liées à leur mise en œuvre dans les organisations. Enfin, la synthèse souligne que les risques sociotechniques, bien que de plus en plus documentés, restent encore insuffisamment intégrés aux processus d'évaluation et de décision.

Le principal apport de cette étude réside ainsi dans l'identification de la fragmentation comme un résultat transversal. Les risques techniques, réglementaires et sociotechniques ne sont pas indépendants. Ils peuvent interagir, se renforcer ou conduire à des arbitrages entre différentes exigences. Une gestion fondée sur des dispositifs isolés risque donc de ne fournir qu'une réponse partielle aux défis posés par l'intégration de l'intelligence artificielle dans les systèmes d'information.

L'étude souligne, en conséquence, l'intérêt d'une approche intégrée reposant sur une évaluation continue des risques, une gouvernance proportionnée à la criticité des systèmes, une clarification des responsabilités et une attention portée aux conséquences susceptibles d'émerger dans les conditions réelles d'utilisation. Une telle approche ne suppose pas nécessairement la création d'un cadre entièrement nouveau, mais plutôt une meilleure articulation entre les mécanismes techniques, organisationnels et réglementaires déjà disponibles.

Cette perspective présente un intérêt particulier pour les organisations confrontées à une diffusion rapide de l'intelligence artificielle et à l'évolution simultanée des technologies, des usages et des exigences de gouvernance. Plus qu'un ensemble de contrôles isolés, la gestion des risques de l'intelligence artificielle apparaît comme une capacité organisationnelle à identifier les risques émergents, à comprendre leurs interactions et à adapter les mécanismes de contrôle au cours du cycle de vie des systèmes.

Les recherches futures pourront prolonger cette réflexion par des études empiriques permettant d'examiner la mise en œuvre concrète de cette articulation dans différents secteurs et contextes organisationnels. Elles pourraient notamment contribuer à mieux comprendre les conditions nécessaires au développement de dispositifs de gouvernance capables d'accompagner une intégration durable, responsable et maîtrisée de l'intelligence artificielle dans les systèmes d'information.

BIBLIOGRAPHIE

- Abercrombie, G., Benbouzid, D., Giudici, P., Golpayegani, D., Hernandez, J., Noro, P., Pandit, H., Paraschou, E., Pownall, C., Prajapati, J., Sayre, M. A., Sengupta, U., Suriyawongkul, A., Thelot, R., Vei, S., & Waltersdorfer, L. (2024). A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms (arXiv:2407.01294). arXiv. <https://doi.org/10.48550/arXiv.2407.01294>
- Bagehorn, F., BrimiJoin, K., Daly, E. M., He, J., Hind, M., Garces-Erice, L., Giblin, C., Giurghi, I., Martino, J., Nair, R., Piorkowski, D., Rawat, A., Richards, J., Rooney, S., Salwala, D., Tirupathi, S., Urbanetz, P., Varshney, K. R., Vejsbjerg, I., & Wolf-Bauwens, M. L. (2025). AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources (arXiv:2503.05780). arXiv. <https://doi.org/10.48550/arXiv.2503.05780>
- Bassani, E., & Sanchez, I. (2025). On Guardrail Models Robustness to Mutations and Adversarial Attacks. JRC Publications Repository. <https://doi.org/10.18653/v1/2025.findings-emnlp.922>
- Berman, G., Cooper, N., Hwang, A. H.-C., Deng, W. H., Shelby, R., & Hutchinson, B. (2026). “Death by a thousand taxonomies?”: AI Risk Classification In Practice (arXiv:2608.06831). arXiv. <https://doi.org/10.48550/arXiv.2608.06831>
- Critch, A., & Russell, S. (2023). TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI (arXiv:2306.06924). arXiv. <https://doi.org/10.48550/arXiv.2306.06924>
- Eisenberg, I. W., Gamboa, L., & Sherman, E. (2025). The Unified Control Framework: Establishing a Common Foundation for Enterprise AI Governance, Risk Management and Regulatory Compliance (arXiv:2503.05937). arXiv. <https://doi.org/10.48550/arXiv.2503.05937>
- El Azzouzy, O., Chanyour, T., & Jai Andaloussi, S. (2026). Towards Robust and Transparent AI: Addressing Adversarial Threats, Biases, and Governance Challenges. In B.-E. Boudriki Semlali, I. Ben Abdel Ouahab, & L. Bragança (Eds.), *Computer Sciences for Sustainable Development* (pp. 394-405). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-22567-2_35
- Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial Intelligence risk measurement. *Expert Systems with Applications*, 235, 121220. <https://doi.org/10.1016/j.eswa.2023.121220>

-
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples (arXiv:1412.6572). arXiv. <https://doi.org/10.48550/arXiv.1412.6572>
 - Ibrahim, M. (2025). An Ethical Framework for Mitigating AI Algorithmic Bias in Information Resource Development. *Journal of Information Sciences*, 24(2), 76–86. <https://doi.org/10.34874/IMIST.PRSM/jis-v24i2.60583>
 - Kasirzadeh, A. (2025). Two types of AI existential risk: Decisive and accumulative. *Philosophical Studies*, 182(7), 1975–2003. <https://doi.org/10.1007/s11098-025-02301-3>
 - Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2023). Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. arXiv.Org. <https://arxiv.org/abs/2303.05453v1>
 - Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4), 383–392. <https://doi.org/10.1038/s42256-024-00820-y>
 - Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial Examples in the Physical World. In *Artificial Intelligence Safety and Security* (p. 474). Chapman and Hall/CRC. <https://doi.org/https://doi.org/10.1201/9781351251389>
 - Olutimehin, A. T., Ajayi, A. J., Metibemu, O. C., Balogun, A. Y., Oladoyinbo, T. O., & Olaniyi, O. O. (2025). Adversarial Threats to AI-Driven Systems: Exploring the Attack Surface of Machine Learning Models and Countermeasures | *Journal of Engineering Research and Reports*. <https://journaljerr.com/index.php/JERR/article/view/1413>
 - Rathi, Y. (2025). AI Governance for Multi-Cloud Data Compliance: A Comparative Analysis of India and the USA. *The American Journal of Interdisciplinary Innovations and Research*, 07(08), 32–42. <https://doi.org/https://doi.org/10.37547/tajir/Volume07Issue08-03>
 - Saari, L., & Mügge, D. (2026). Forking paths of AI governance - how risk management frameworks shape the politics of AI. *Critical Policy Studies*, 0(0), 1–21. <https://doi.org/10.1080/19460171.2026.2669063>
 - Saeri, A. K., George, S. L., Graham, J., Lacarriere, C. D., Slattery, P., Noetel, M., & Thompson, N. (2025). Mapping AI Risk Mitigations: Evidence Scan and Preliminary AI Risk Mitigation Taxonomy (arXiv:2512.11931). arXiv. <https://doi.org/10.48550/arXiv.2512.11931>
-

- Sampathkumar, V., & Veeras, K. (2025). Adversarial Attacks on FinTech AI Models: Threats and Mitigation Techniques. 2025 IEEE International Carnahan Conference on Security Technology (ICCST), 1–7. 2025. <https://doi.org/10.1109/ICCST63435.2025.11293895>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, 59–68. <https://doi.org/10.1145/3287560.3287598>
- Song, L., Shokri, R., & Mittal, P. (2019). Privacy Risks of Securing Machine Learning Models against Adversarial Examples. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 241–257. <https://doi.org/10.1145/3319535.3354211>
- Uesato, J., O'Donoghue, B., Oord, A. van den, & Kohli, P. (2018). Adversarial Risk and the Dangers of Evaluating Against Weak Attacks (arXiv:1802.05666). arXiv. <https://doi.org/10.48550/arXiv.1802.05666>
- Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P., & Prunkl, C. (2024). A Taxonomy of Systemic Risks from General-Purpose AI (SSRN Scholarly Paper No. 5030173). Social Science Research Network. <https://doi.org/10.2139/ssrn.5030173>
- Vijayvargiya, S., Soni, A. B., Zhou, X., Wang, Z. Z., Dziri, N., Neubig, G., & Sap, M. (2026). OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety (arXiv:2507.06134). arXiv. <https://doi.org/10.48550/arXiv.2507.06134>
- Wagner, M., Song, Q., Borg, M., Engström, E., & Lysek, M. (2025). AI Act High-Risk AI Compliance Challenge and Industry Impact: A Multiple Case Study (SSRN Scholarly Paper No. 5221279). Social Science Research Network. <https://doi.org/10.2139/ssrn.5221279>
- Walter, M. J., Barrett, A., & Tam, K. (2025). Preventing Adversarial AI Attacks Against Autonomous Situational Awareness: A Maritime Case Study (arXiv:2505.21609). arXiv. <https://doi.org/10.48550/arXiv.2505.21609>
- Weinzierl, S., Zilker, S., Zschech, P., Kraus, M., Leibelt, T., & Matzner, M. (2024). How Risky is my AI System? A Method for Transparent Classification of AI System Descriptions by Regulated AI Risk Categories (SSRN Scholarly Paper No. 4984202). Social Science Research Network. <https://doi.org/10.2139/ssrn.4984202>

- Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., & Shi, W. (2024). How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs (arXiv:2401.06373). arXiv. <https://doi.org/10.48550/arXiv.2401.06373>
- Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y.-C. (2025). The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, 1-17. <https://doi.org/10.1145/3706598.3713429>