IAG et irresponsabilité sociale des entreprises

Christine Marsal, Gautier Barlette

29e Conférence de l'Association Information et Management

27-29 mai 2024 à Montpellier - La Grande-Motte

IAG et irresponsabilité sociale des entreprises

Christine Marsal

Gautier Barlette

Université de Montpellier

MRM

Résumé :

Le développement rapide des IAG et les questions éthiques, morales, sociétale qui en découlent entrainent praticiens, gouvernements et chercheurs à poser les bases d'une IAG responsable. Notre posture d'étude est différente et nous abordons la question de la responsabilité des IAG sous l'angle de pratiques potentiellement irresponsables. L'irresponsabilité sociale des entreprises (ISE) est considérée comme un phénomène à part entière qu'il convient d'étudier en dehors de toutes les démarches RSE existantes. La première partie de notre article s'attache à clarifier les bases conceptuelles de l'ISE. Nous montrons ensuite en quoi, le développement des IAG est un terrain d'étude fécond pour illustrer les dimensions de l'ISE. La partie empirique (non réalisée à ce jour) permettra de tester un nouvel outil de *scoring* ISE et de montrer quelles sont les dimensions critiques de la diffusion massive des IAG.

Mots clés :

IAG ; irresponsabilité ; IA explicable ; Intentions

IAG and corporate social irresponsibility

Abstract :

The rapid development of IAGs and the ethical, moral and societal issues they raise are leading practitioners, governments and researchers to lay the foundations for responsible IAGs. Our approach is different, however, and we approach the question of corporate

responsibility from the angle of potentially irresponsible practices. Corporate social irresponsibility (CSR) is considered to be a phenomenon in its own right that should be studied separately from all existing CSR initiatives. The first part of our article sets out to clarify the conceptual foundations of SRI. We then show how the development of EMIs is a fertile field of study for illustrating the dimensions of SRI. The empirical part (which has not yet been carried out) will allow us to test a new SRI scoring tool and to show which dimensions are critical to the massive spread of SRI.

Keywords :

AIG, irresponsibility, explainable AI, intent

IAG et irresponsabilité sociale des entreprises

Introduction

Dans l'enquête consacrée aux IA , réalisée en janvier 2024[1],auprès de plus 1200 managers, les consultants du BCG rapportent que 89% des managers considèrent que les IAG sont dans les trois investissements technologiques les plus importants pour l'année 2024.51% la mettent en tête de liste après la cybersécurité et le *cloud computing.*

L'Intelligence Artificielle Générative (désormais IAG) est définie comme une branche de l'intelligence artificielle capable de générer un nouveau contenu, par opposition à l'analyse ou traitement de données existantes, comme le font les systèmes experts (Vaswani et al., 2017). Les IAG font référence à des modèles et techniques qui permettent de générer du contenu original et nouveau : du texte élaboré (non distinguable d'un texte écrit par un humain), des images, de la musique. Leur diffusion dans le grand public a été fulgurant. OpenAI a annoncé la mise à disposition d'une version gratuite de ChatGPT basé sur l'intelligence artificielle en novembre 2022. En l'espace d'une nuit, ChatGPT s'est fait connaître du grand public et, en quelques mois, il comptait plus de 100 millions d'utilisateurs dans le monde entier (Milmo, 2023).Cependant, l'utilisation d'un outil à la fois aussi démocratisé et aussi puissant que l'IAG posent des questions éthiques pour les entreprises, tant au niveau de l'irresponsabilité sociale relative à l'utilisation des IA (impact écologique, problème de vérification des sources, mauvaise utilisation de l'IA par ses utilisateur), qu'au niveau des décisions irresponsables engendré par l'IA comme outil stratégique.

Un enjeu bien compris en France, par le Comité National de Pilotage de l'Éthique Numérique (désormais, CNPEN), qui souhaite rappeler certains arguments exprimés dans son avis sur les enjeux éthiques des agents conversationnels en les étendant à l'utilisation des LLM. « *...La question de la responsabilité est posée dans toutes ses formes : responsabilité légale et morale, individuelle et collective, celle du concepteur, du fabricant, de l'utilisateur et du décideur politique, celle relative aux éventuels dysfonctionnements et celle liée aux conséquences de ces technologies à long terme.* »[2]. Prenant acte des risques engendrés, la Commission Européenne s'empare des IAG définies comme des systèmes : " *destinés à générer, avec différents niveaux d'autonomie, des contenus tels que des textes complexes, des images, des sons ou des vidéos*"[3] dont les fournisseurs doivent satisfaire des obligations complémentaires (à celles existantes pour les IA).

Alors que les débats nourrissent les craintes d'un remplacement de l'homme par la machine, plusieurs questions restent en suspens. Les IAG induisent-elles des pratiques et

[1] https://www.bcg.com/publications/2024/from-potential-to-profit-with-genai

[2] CNPEN, Avis n° 3, 15 septembre 2021, Agents conversationnels : enjeux d'éthique, p. 6

[3] Draft Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament
and of the Council on harmonised rules on Artificial Intelligence Act and amending certain Union Legislative
Acts. 16/5/2023. (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))

comportements intrinsèquement irresponsables ? Comment qualifier le degré d'irresponsabilité des IAG déployées dans les organisations ?

Pour répondre à ces questions, nous revenons dans une première partie sur le concept d'Irresponsabilité Sociale de l'Entreprise (désormais ISE) et nous en présentons les principales caractéristiques. La deuxième partie est consacrée à la diffusion des IAG et aux responsabilités engagées dans sa diffusion, elle se conclut par une proposition de grille de *scoring* d'irresponsabilité.

L'objectif de ce travail est de proposer une grille d'évaluation qui contribuera à la recherche académique de plusieurs façons : opérationnalisation du cadre conceptuel de Clark et al. (2022), identification des dimensions critiques de l'irresponsabilité de l'IAG, clarification des enjeux des IAG dites « responsables ». D'un point de vue managérial, la grille de score doit permettre d'alerter les dirigeants (et salariés) sur les risques d'irresponsabilité de l'utilisation des IAG et permettre de mettre en place des moyens de prévention ou de remédiation.

1. Qu'est-ce que le concept d'Irresponsabilité Sociale des Entreprises (ISE)?

1.1. Définition et enjeux de l'ISE.

L'irresponsabilité sociale des entreprises est souvent définie comme étant l'autre face de la RSE (Godfrey *et al.* (2009), Hull et Rothenberg (2008), Mac Mahon (1999) et Zyglidopoulos *et al. (*2012)), partent du principe que l'ISE n'a qu'un effet miroir de la RSE, et que l'ISE n'aurait pas d'effet en tant que telle.

Cependant au fil des années, les chercheurs spécialisés du domaine ont démontré que l'ISE à des effets qui lui sont propres (Riera et Iborra, 2017), et que de ce fait l'ISE doit être étudiée en tant que sujet propre, avec ses propres logiques internes et ses propres outils méthodologiques.

La première définition de l'ISE remonte à 1977 par Armstrong qui évoque « *un acte jugé socialement irresponsable est une décision de choisir une alternative inférieure à une autre alternative si toutes les parties sont considérées. Généralement, cela implique un gain pour une partie au détriment du système global* ». Clarks et al. (2022) suggèrent que cet effet de balancier entre gains et préjudices n'est pas toujours pertinent.

Il convient cependant de discerner les mauvaises conduites ou « *misconducts* » (Greve *et al.*, 2010, p. 54), qui sont des comportements isolés effectués par des individus, des violations répétées des normes sociales par les entreprises (Greve *et al.*, 2010, p. 54), qui ont des effets plus locaux et bien moindres. Cependant, si des « *misconducts* » répétitives (exemple du racisme envers un salarié) ne sont jamais sanctionnées par l'entreprise, les actes isolés deviennent un problème récurrent et rentrent dans le domaine de la responsabilité de l'entreprise. La seconde définition communément admise est celle de Campbell (2007). L'auteur considère que si une entreprise commet volontairement quelque chose qui nuit à une ou plusieurs de ses différentes parties prenantes, ou si l'entreprise ne corrige pas un préjudice qu'elle a causé dès sa découverte par le management/direction de l'entreprise, alors la relation

de confiance entre l'entreprise et les différentes parties prenantes est rompue, et l'entreprise se retrouve dans une situation d'irresponsabilité d'entreprise.

Cette définition va ensuite évoluer afin mieux mettre en valeur les différentes parties prenantes et surtout leur manière de percevoir une ISE : « *l'irresponsabilité d'entreprise est définie par ce que les parties prenantes considèrent comme socialement irresponsable* » (Lange et Washburn, 2012 ; Tench *et al.*, 2017 ; Wagner *et al.*, 2008 ; Williams et Zinkin, 2008).

Cette définition se bases sur les travaux de Lange et Washburn (2012), qui ancrent l'irresponsabilité dans « *l'attribution theory* », théorie de psychologie se basant sur la perception et l'interprétation d'une cause. Appliqué à l'ISE, la théorie aide à mieux prendre en compte la perception sur l'origine d'une ISE, la volonté de faire ladite ISE et l'impact de l'irresponsabilité sociale par les différentes parties prenantes.

Cela implique trois concepts importants :

- L'ISE doit être perçue par les différentes parties prenantes, tant que l'ISE n'est pas révélée, elle n'a pas d'impact.
- L'ISE en fonction de la façon elle est perçue, peut changer la notion de volonté de commettre et/ou de réparer la fautes (Clark et al 2022)
- l'ISE va avoir des impacts différents en fonction du cadre sociétal, une ISE aux États-Unis n'aura pas le même effet qu'en Chine, et ce qui n'était pas perçu comme une ISE dans le passé peut l'être aujourd'hui.

1.2 Caractérisation des pratiques socialement irresponsables

Clark et al. (2022) approfondissent le concept d'ISE et identifient ce qu'ils nomment des concepts "zones grises". Ces concepts font débat dans le monde académique et nuisent aux progrès des recherches dans ce domaine.

Les trois concepts qui font débat sont le rôle des préjudices et des bénéfices, le rôle de l'acteur et de l'intentionnalité et le rôle de la rectification.

Les auteurs récapitulent une revue de littérature réalisée à partir des mots clés que l'on retrouve dans les recherches dédiées à L'ISE. Ils effectuent leur recherche dans les journaux académiques consacrés à la gestion d'entreprise et à la comptabilité entre 1956 et septembre 2020. Les éléments de synthèse qu'ils proposent sont les suivants :

Le concept de préjudice reste problématique car il est difficile de déterminer le degré par lequel une entreprise est irresponsable, si on base l'analyse uniquement sur le caractère préjudiciable de ses actes. Les auteurs notent ainsi, qu'une entreprise peut porter préjudice sans que cela soit intentionnel. Le préjudice en lui seul ne suffit pas à qualifier une entreprise d'irresponsable.

Il est donc indispensable de lier le préjudice à l'intentionnalité.

La discussion du concept d'intentionnalité fait aussi débat et les auteurs proposent de l'envisager sous un angle large : connaissance, négligence, insouciance. Nous pouvons illustrer ces concepts par des exemples .

Les cadres dirigeants d'une entreprise ou ses salariés peuvent causer un préjudice en connaissance de cause. Si l'on prend l'exemple du scandale Volkswagen, les ingénieurs ont bien mis en place des dispositifs de contournement des tests antipollution, de façon avérée et non accidentelle. De même, les pratiques commerciales douteuses qui ont conduit à surendetter des ménages modestes, avant que n'éclate la crise des *subprimes,* ont mis en avant la négligence de la supervision interne de certaines banques et l'insouciance des chargés de clientèle. La négligence signifie que des modalités de contrôle n'ont pas été mises en œuvre de façon pertinente. Dans ces deux exemples, les préjudices causés sont de le fait d'une intentionnalité qui prend plusieurs formes possibles.

La question de la rectification et des moyens de remédiation engagés par l'entreprise fait aussi débat. Les auteurs notent que toutes les formes de rectification ne sont pas équivalentes : une rectification forcée (pression sociale, poursuites juridiques etc..) n'a pas le même "poids" qu'une rectification volontaire. Il est donc nécessaire de les distinguer afin de caractériser les pratiques observées.

A la suite de cette synthèse, les auteurs proposent un cadre d'analyse qui s'appuie sur l'idée qu'il existe des degrés d'intensité dans les trois dimensions. L'absence de préjudice, d'intentionnalité et l'existence de mécanismes de remédiation volontaires signifient que les entreprises ne sont pas irresponsables (il y a une absence d'irresponsabilité). À l'opposé du schéma, les entreprises dont les actes peuvent porter atteinte à la vie humaine, de façon intentionnelle et sans modalités de remédiation sont considérées comme étant fortement irresponsables. Entre ces deux situations extrêmes, il existe un continuum dans lequel évoluent les organisations. Le caractère synthétique de ce cadre d'analyse permet une base de travail solide pour l'étude de changements organisationnels comme peuvent l'être les changements induits par les IAG.

-

2. A la recherche d'une IAG responsable

La question de la responsabilité liée à la conception et à l'utilisation des IAG est au cœur de débats et discussions tant en Europe, au Royaume Uni, qu'aux États-Unis ou en Chine (Carter, 2020). Les principales parties prenantes perçoivent qu'il existe un ou des problèmes potentiels liés à l'usage des IAG. C'est ainsi que la Commission Européenne propose une réglementation en fonction de niveaux de risques : inacceptable, élevé, limité. Cela induit que, en fonction des risques identifiés, les fournisseurs de système d'IA (et par extension des IAG) doivent se conformer à des obligations de transparence. Notre revue de littérature consiste à relever les trois dimensions retenues par Clark et al. (2022) : dommages et intentionnalité d'une part, remédiation d'autre part. Les articles sont sélectionnés partir des mots clés suivants : IAG responsable, irresponsabilité, responsabilité, IAG et éthique, IAG et risques.

2.1 Les préjudices et l'intentionnalité en question

La question des préjudices que peuvent causer les IAG fait l'objet de débats et de tentatives de clarification. Les cadres légaux sont récentes et certains aspects sont encore en discussion (Carter, 2020).

2.1.1. Les risques identifiés et /ou les préjudices éventuels

Les risques et les préjudices associés à la conception et l'utilisation de l'IAG sont multiples. Ils font l'objet de discussions au sein des instances européennes, dans des comités d'éthique, au sein d'organisations internationales (Carter, 2020). Nous pouvons reprendre la liste proposée par Doyal et al. (2023) qui en offre une bonne synthèse :

- Les risques liés à la collecte et à l'utilisation des données personnelles :
- Les risques liés à la désinformation
- Le manque de transparence et la difficulté d'identifier la qualité des *outputs*, de corriger les erreurs
- Les risques de plagiat
- Les risques de remplacement d'emplois
- Risque de dépendance technologique accrue

De leur côté, Wach et al. (2023) évoquent moins des risques que des menaces : (i) absence de réglementation du marché de l'IA et besoin urgent de réglementation, (ii) mauvaise qualité, absence de contrôle de la qualité, désinformation, contenu "*deepfake*", biais algorithmique, (iii) pertes d'emplois dues à l'automatisation, (iv) violation des données personnelles, surveillance sociale et violation de la vie privée, (v) manipulation sociale, affaiblissement de l'éthique et de la bonne volonté, (vi) creusement des inégalités socio-économiques, et (vii) technostress lié à l'IA. Les auteurs poursuivent avec l'exemple suivant : si ChatGPT fournit des conseils financiers inexacts aux utilisateurs, entraînant des pertes financières, il peut être difficile de déterminer qui est responsable de ces pertes. Il peut en résulter des litiges juridiques et des difficultés à établir les responsabilités, ce qui peut entraîner des dommages financiers et des atteintes à la réputation pour les organisations et les particuliers.

Face à ces risques et préjudices potentiels, la question se pose de connaître la responsabilité des parties prenantes impliquées dans la diffusion des IAG. Deux parties prenantes attirent l'attention des autorités réglementaires : le concepteur et l'utilisateur.

2.1.2. Intentionnalité des acteurs

Baird et Maruping (2021) soulignent la part croissante que les IAG prennent dans les processus décisionnels. Cette délégation plus ou moins supervisée peut causer des préjudices tant à l'entreprise utilisatrice qu'à ses parties prenantes. Baek et Kim, (2023) notent que les IAG sont indifférentes à la vérité, elles peuvent produire un nombre considérable de fausses informations à des coûts très limités. La question de l'intentionnalité est particulièrement ardue à délimiter, deux acteurs sont considérés dans les travaux académiques mais aussi dans les discussions institutionnelles : les concepteurs et les utilisateurs.

L'intentionnalité des concepteurs :

Baek et Kim (2024) mettent en avant le rôle des concepteurs et leur responsabilité quant aux impacts positifs et négatifs que peuvent avoir les IAG. Ils ne posent pas la question de l'intentionnalité en tant que telle, mais cette posture semble indiquer que pour ces deux auteurs les concepteurs savent ce qu'ils font (ou doivent savoir ce qu'ils font). Cette posture n'est pas neutre car, les IAG peuvent amplifier la mauvaise conception des algorithmes (Manyika et al., 2019 ; Harfouche et al., 2023 ; Doyal et al., 2023). Cette amplification n'est

pas « voulue » par les concepteurs, mais cette responsabilité induit qu'ils auraient pu être négligents (au sens de Clark et al ; 2022).

Grinbaum et al. (2023) évoquant la manipulation possible des utilisateurs par une IAG (des utilisateurs qui se laisseraient abuser par le caractère « humain » des réponses données par l'IAG) soulignent que le concepteur ne peut pas prévoir toutes les "sorties" du système et peut être de bonne foi. Mais dans le même temps, ils soulignent aussi que si l'IAG permet la production de contenus potentiellement faux, cela engage la responsabilité des concepteurs. Selon les cas, le concepteur peut être responsable ou non responsable. Les auteurs cités n'utilisent pas le terme irresponsable.

L'intentionnalité des utilisateurs :

Kanitz et al. (2022) insistent sur les conséquences (potentiellement négatives) non intentionnelles de la mise en place des IAG. Ainsi, ils soulignent le manque de retour d'expérience sur les bénéfices et les risques d'utilisation des IAG (impacts sur la performance, l'utilisation des ressources, l'expérience des collaborateurs), sur la qualité de la supervision automatisée et de leurs effets sur les collaborateurs. Dans leur esprit la non-intentionnalité provient de la difficulté de prévoir ou d'imaginer les conséquences négatives de l'utilisation des IAG. Baird et Maruping (2021) remarquent que les IAG introduites dans les systèmes de décisions ont tendance à collecter plus de données que nécessaires ce qui peut conduire à des usages non prévus : surveillance accrue des employés. Monod et al. (2024) cherchent à comprendre comment la mise en place des IAG, destinée à améliorer les performances de l'entreprise peut conduire à des conséquences négatives. Ils ont conduit une étude de cas qualitative approfondie, dans une entreprise chinoise qui a introduit un outil d'IA dans le domaine de la vente. Ils ont constaté que si la conception initiale de l'outil d'IA visait apparemment à responsabiliser les vendeurs et a d'abord permis d'obtenir les résultats escomptés, au fil du temps, l'outil a été détourné pour être utilisé à des fins de contrôle managérial. Cette évolution a émergé organiquement d'une prise de conscience croissante par les managers des possibilités que l'outil d'IA leur offrait pour mieux accomplir leur travail et surveiller leurs subordonnés. Bernardo et Seva (2023) proposent une conception des IAG explicable tournée vers les besoins des utilisateurs afin de favoriser leur confiance et leur acceptation. Leur étude montre toutefois de possibles mauvaises utilisations de IAX (surveillance des salariés).

2.2 Les remédiations envisagées

Clark et al. (2022) évoquent une troisième dimension qui est là pour atténuer les préjudices causés par des comportements ou pratiques irresponsables, il s'agit des moyens de remédiation mis en place par les entreprises mises en cause.

2.2.1. Le débat des remédiations

Dans un souci d'anticiper les préjudices et les questions de responsabilité, Carter (2020) note que quelques géants de la tech (Google, Microsoft, IBM) ont développé leurs propres standards éthiques relatifs à l'AI alors qu'Amazon s'y refusait.

Jovanovic et Campbell (2022) suggèrent que les problématiques d'éthique, d'équité et la reddition de compte (*accountability*) puissent se résoudre via des boucles d'actions et de rétroactions. Il serait donc possible de corriger les "mauvais" résultats de l'IAG par une veille

régulière et des actions correctrices. Mais dans le même temps, Doyal et al. (2023) soulignent qu'il est très difficile pour un modèle déployé (comme celui de chat GPT) de corriger les biais qu'il véhicule après sa phase d'entraînement.

Entre chartes d'utilisation et solutions techniques, les moyens de remédiation semblent limités. Ils le sont d'autant plus que les IAG fonctionnent rapidement de manière indépendante. C'est ainsi que Wei et al. (2022) évoquent le concept de "capacité émergente" ou de "comportement émergent" des modèles LLM. Il s'agit de résultats inattendus ou surprenants pour les utilisateurs mais aussi pour les concepteurs. Ces "comportements" sont difficilement prédictibles. L'utilisation croissante des LLM devrait accroître ces comportements qui peuvent être néfastes dans des applications critiques ou sensibles.

2.2.1. Les IAG explicables : une réponse incomplète

L'IA explicable (IAX) est un concept qui tente de répondre au problème du manque de confiance des utilisateurs de l'IAG. Il s'agit de donner aux humains un niveau de compréhension de la façon dont l'IA (et par extension l'IAG) fonctionne (Bernardo et Seva, 2023). Arrieta et al. (2020) évoquent le concept d'intelligence artificielle responsable qui se traduit par une méthodologie visant à mettre en œuvre à grande échelle des méthodes d'IA avec au cœur l'équité, l'explicabilité des modèles et la responsabilité. Cette responsabilité consiste en l'adoption de principes pour la mise en œuvre pratique des modèles.

Outre l'explicabilité, les lignes directrices qui sous-tendent l'IA responsable établissent que l'équité, la responsabilité et la protection de la vie privée doivent également être prises en compte lors de la mise en œuvre de modèles d'IA dans des environnements réels.

Fjeld et al. (2019) proposent plusieurs principes : l'équité (lutte contre les discriminations), IA transparente et explicable, centrée sur le bien être des humains, la sécurité et la protection de la vie privée dès la phase de conception. Les auteurs soulignent que les entreprises publient déjà des principes d'IA intelligente, dans lesquelles elles affirment qu'elles prennent soin d'éviter les conséquences négatives de l'utilisation de l'IA. Cependant, il y a peu d'exemple de la façon dont ses principes sont mis en œuvre. De même, Ernst et al. (2019) soulignent le caractère persistant des biais, en donnant l'exemple des processus de recrutement. L'opacité des processus de décision n'est pas toujours dissipée.

Dans cette analyse rapide, la question de l'intentionnalité du concepteur n'est abordée que partiellement : la question de savoir si la mauvaise conception est une action en connaissance de cause, est le résultat d'une négligence ou d'une insouciance n'est pas complètement tranchée.

La question de l'intentionnalité de l'utilisateur (à l'exception de pratiques délibérées effectuées dans l'intention de nuire, comme la diffusion de fausses nouvelles) apparaît le plus souvent comme "accidentelle" (négligence/insouciance ?). L'exemple qui revient le plus souvent est l'utilisation des données abondantes collectées par les IAG pour surveiller étroitement les collaborateurs. Cette surveillance pouvant potentiellement supprimer des emplois (Bair et Maruping, 2021; Monod et al., 2024).

2.2.2. Les IAG sont-elles par nature irresponsables ?

Baek et Kim (2023) remarquent que les IAG sont conçues et entraînées par des grandes entreprises dont le but est de maximiser leurs profits, sans prendre en considération les fractures sociales, géographiques etc.

Grinbaum et al. (2023) rapportent les résultats d'une enquête réalisée par les autorités de contrôle italiennes et allemandes. Cette enquête montre que les principaux acteurs des IAG ne respectent pas la réglementation européenne de protection des données personnelles.

L'avis N°7 du CNPEN rend compte de l'absence de base légale pour la collecte extensive de données qui sont utilisées pour entraîner les IAG. Il souligne aussi, la non-convergence des principes européens avec les principes édictés dans l'État de Californie, où se situent les sièges sociaux des grands acteurs de l'IAG.

Diwivedi et al. (2023) soulignent que la technologie IAG actuelle repose sur des bases de données existantes, exploitées dans un cadre légal en construction, ce qui limite de facto les perspectives diverses et inclusives dans la prise de décision. Corchado et al. (2023), évoquent les biais perpétués par les IAG lorsqu'ils s'appuient sur des données elles-mêmes biaisées.

Grinbaum et al. (2023) soulignent le problème éthique majeur résultant de l'impossibilité de distinguer ce qui est généré par un être humain et ce qui est généré par une IAG. : *"Si un texte provoque une tension éthique, par exemple à cause du nudging ou de la fraude qu'il contient, il est impératif de pouvoir tracer son origine afin d'éviter la confusion entre un discours produit par un agent responsable, apte à répondre de ce qu'il dit, et la parole asémantique d'un système d'intelligence artificielle auquel on ne peut attribuer aucune responsabilité"*[4]

De Vries (2023) indique que les besoins en énergie des IAG ne se limitent pas à la phase d'apprentissage mais sont aussi conséquents pendant la phase d'exploitation des systèmes. L'auteur rapporte les propos d'un manager indiquant que le coût d'une interaction avec des systèmes LLM est près de 10 fois supérieur à une requête via un moteur de recherche. Il note enfin qu'une amélioration des langages employés peut aussi réduire les coûts de requête.

Acemoglu et Hazell (2022) notent que le remplacement des emplois humains par les IAG est en cours même s'il n'est pas encore perceptible au niveau macro-économique.

De tous ces éléments il ressort plusieurs constats : il existe des préjudices avérés observables (surveillance des salariés, collecte de données personnelles, coût environnemental, réduction massive d'effectif, décisions biaisées). L'intentionnalité des concepteurs reste difficile à observer pour des non experts, mais le développement des IAG explicables pourrait démontrer une non-internationalité de provoquer des dommages. L'intentionnalité des utilisateurs est aussi observable (pratiques de détournement des IAG, choix d'une utilisation éthique des IAG). Enfin, les moyens de remédiation sont aussi observables puisqu'ils sont au cœur des discussions entre parties prenantes (obligations réglementaires, pratiques d'entrainement et de correction des biais etc.).

[4] Systèmes d'intelligence artificielle générative : enjeux d'éthique. Avis 7 du CNPEN. 30 juin 2023, page 14

Cette observabilité sera testée par l'utilisation d'un outil que nous avons conçu dans le prolongement des travaux de Clark et al. (2022). Le tableau 1 donne les grandes lignes d'un *scoring* d'irresponsabilité que nous avons l'intention de tester à partir de plusieurs cas d'usage.

Préjudice	Pas de préjudice 0	Probabilité de préjudice 1	Actes ne portant pas préjudice à la vie humaine 2	Actes portant préjudice à la vie humaine 3
Intentionnalité	Pas d'intentionnalité 0			Intentionnalité 5
Rectification	Rectification volontaire 0	Rectification forcée 1		Pas de rectification 2
	Score = 0 **Pas d'irresponsabilité**	→		**Score = 10** **Forte irresponsabilité**

Réalisé à partir du cadre d'analyse proposé par Clark et al. (2022) , page 1498

Tableau 1 : Outil de score pour qualifier les activités et comportements irresponsables

Dans ce contexte, le statut de cet outil de *scoring* est ambigu et nous en sommes conscients. Au niveau des organisations utilisatrices, il s'agit de promouvoir un outil d'analyse et de prévention du risque de réputation lorsque des préjudices touchent les parties prenantes de l'entreprise (salariés, clients, fournisseurs, associations de protection de l'environnement). A ce titre, le résultat du *score* doit permettre d'identifier des actions préventives en termes d'organisation interne. Mais dans le même temps, il nous faut reconnaître le caractère imprécis de la grille. Par exemple, la case « probabilité de préjudice » indique qu'il existe un doute, que les connaissances que nous avons ne permettent pas d'inférer qu'un préjudice est exclu. Compte tenu des impacts négatifs potentiellement très importants (par le nombre de personnes affectées), le doute nécessite d'agir et de chercher à réduire l'ampleur du préjudice, voire à le supprimer. Cette posture est à rapprocher du principe de précaution développé par Vaissière (1999).Il est pertinent lorsque l'on s'intéresse à des cas d'usage, qui n'ont pas encore fait l'objet d'une exploitation à grande échelle.

Notre objectif est de proposer un outil qui puisse anticiper les risques d'irresponsabilité. En effet, en travaillant sur des cas d'usage, nous ne travaillons pas sur des cas réels mais des cas

en devenir, en discussion. Ce faisant, nous mettons en perspective les choix stratégiques de plusieurs secteurs (le secteur de l'immobilier, le secteur bancaire) au regard de l'intégrité future des hommes. Nous nous situons (ou tentons de le faire) dans une responsabilité identifiée comme : « une éthique orientée vers le futur »[5].

3. Méthodologie

3.1 Positionnement épistémologique

Notre positionnement épistémologique est motivé par les contraintes imposées par notre terrain de recherche. En effet, l'IAG est un domaine extrêmement nouveau, et encore assez peu utilisé par la plupart des entreprises, ce qui amène non seulement des difficultés d'accès au terrain, mais également un vrai manque de recul de la part de l'entreprise ainsi que les parties prenantes salariales.

Notre recherche est de type exploratoire, pour connaitre les différentes irresponsabilités d'entreprises possiblement générées par l'IAG. Tel que décrit par (Venkatesh et al, 2013) nous avons le choix entre trois positionnements épistémologiques : le pragmatisme *Critical Realism , et le transformative–emancipatory.* Le *critical realism* est un paradigme qui cherche à avoir une meilleure compréhension d'un phénomène (Archer et al. 1998; Bhaskar 1978; Danermark et al. 2002; Houston 2001)., partant du principe qu'il n'y a pas de vérité absolue dans laquelle un objet puisse être comparé (Maxwell, 1992). Il s'agit non seulement d'un paradigme particulièrement recommandé dans le cas d'une mix méthode, mais aussi d'un paradigme particulièrement adapté à nos questions de recherche, apportant de la nuance sur l'utilisation de l'IAG

3.3 Mix méthode

De ce fait, nous optons pour une méta analyse des données des grands cabinets de conseil tant qualitative que quantitative. Nous analysons les cas d'usage présentés par les rapports de cabinet de consultant.

Pendant longtemps les méthodes mixtes , mélangeant qualitatif et quantitatifs ont été mal considérées par la recherche (Venkatesh et al, 2013). Cependant (Venkatesh et al, 2013) et (Creswell et Creswell, 2018) montrent qu'une méthode mixte peut avoir son intérêt, pour une recherche exploratoire . Il y a de nombreuses variantes de mix méthode, et nous choisissons ici la méthode complémentaire telle qu'utilisée par (Soffer et Adder, 2007). Dans ce cadre, les données quantitatives et qualitatives sont combinées afin d'avoir une meilleure compréhension de ce phénomène.

3.4 Détails des méthodes et données collectées

L'accès au terrain est particulièrement ardu, non seulement car les entreprises utilisant l'IAG sont encore peu nombreuses, et parmi celle qui l'utilisent, peu la maitrisent ou souhaitent communiquer sur leur utilisation de l'IAG.

Les grands cabinets de conseil (Mckinsey, BCG…) restent à l'écoute des besoins, des questionnements et projets IAG de leurs principaux réseaux de clients. Dans ce but, ils ont

[5] Vaissière (1999), page 43

conduit de multiples enquêtes auprès des moyennes et de grandes entreprises. Les rapports sur lesquels nous avons travaillé , récapitulent des données quantitatives (enquêtes d'opinions auprès des managers, auprès de salariés, expérimentation à grande échelle d'une solution IAG) mais aussi des données qualitatives sous la forme de cas s'usage.

Nous travaillons donc sur des données secondaires, ces données sont à destination des décideurs, elles sont de nature à justifier le recours aux services des cabinets pour la mise en place des IAG. On peut donc logiquement supposer que les cas sont suffisamment crédibles et bien documentés pour les utiliser dans un article académique exploratoire. Notre apport consiste à analyser les cas d'usage sous l'angle de l'irresponsabilité possible de certains projets présentés. L'analyse de uses cases est une méthode notamment utilisé dans le cadre de l'analyse de système d'information ou dans le développement (Irwin et Turk 2005) notamment pour prévoir l'utilisation d'un nouvel outil en condition réelle. Cette méthode s'applique parfaitement dans le cadre de l'IAG, que ce soit dans notre cadre de recherche que dans un cadre plus général.

La prochaine étape est donc d'utiliser la matrice du tableau 1 sur les différents cas d'usage, afin de voir quels cas d'usages sont susceptible ou non d'engendrer des pratiques d'irresponsabilité sociale.

4. Premiers tests

4.1. Opérationnalisation des variables de la grille de *scoring*

Afin d'analyser ces cas, il nous faut définir les préjudices auxquels nous souhaitons nous intéresser : qualité de vie au travail (reconnaissance des compétences individuelles, degré de liberté accordé aux employés, valorisation de l'expérience humaine, etc..), impact environnemental, préservation de l'intégrité des autres parties prenantes (données personnelles, droits d'auteurs, discriminations etc..). Ces préjudices ont été présentés dans les dangers ou limites des IAG.

Le caractère intentionnel du préjudice consiste à considérer les démarches délibérées (licenciement de 70 % de l'effectif d'une entreprise par exemple). Cela n'exclut pas les stratégies de découvertes, un dirigeant saisit une opportunité de croissance qui pourra avoir des conséquences négatives, sans que cette conséquence soit « recherchée ». Le caractère non intentionnel du préjudice n'exclut pas la nécessité de corriger les pratiques concernées.

3.2. Les résultats attendus de l'application de cette grille.

Les cas d'usage envisagés par les cabinets de consultant sont des cas qui sont en réflexion dans les organisations, et lorsque ces cas entrent en application, nous n'en sommes qu'au début du processus. Le *scoring* proposé est un outil de prévention et d'anticipation des risques de réputation des entreprises concernées. Tous les secteurs, tous les pays, tous les citoyens sont concernés. Cet outil, par son coté simplificateur vise à alerter les dirigeants d'entreprise. Cette démarche est d'autant plus importante que l'étude réalisée par Mc Kinsey en 2020, indique qu'à l'époque seules 21 % des entreprises interrogées avaient mis en place des gardes

fous dans l'utilisation des IAG, les questions de cybersécurité ou de conformité à la réglementation.

L'originalité de notre démarche consiste à étudier l'implantation d'un nouvel investissement à l'aune de l'irresponsabilité. Notre démarche vise à identifier les préjudices récurrents entrainant des situations d'irresponsabilité sociale des entreprises. Certains cas, par exemple les IAG employées pour améliorer les diagnostics médicaux, n'entrainent pas de situation d'irresponsabilité. Nous pourrons les inclure dans nos résultats, non pas pour souligner le caractère ambivalent de cette technologie mais, simplement pour signaler que les IAG ne sont pas que des outils « nocifs ».

Enfin, à la suite d'une première catégorisation des cas d'usage (en fonction du score obtenu) nous cherchons à mieux comprendre les conditions qui favorisent les pratiques irresponsables (secteur fortement consommateur de technologies de l'information et de la communication, taylorisation actuelle du travail, entreprises cotées etc..).

3.3. Premiers tests de la grille

Le service après-vente [6]

Les agents conversationnels donnent des conseils personnalisés aux clients en ligne, les agents peuvent co participer à des entretiens en ligne entre salariés et clients. Les employés utilisent des argumentaires rédigés par les IAG et, pendant la conversation peuvent recevoir de nouvelles suggestions de réponses. Ils disposent en outre d'un accès rapide et complet au profil client. Après la conversation, l'employé reçoit un résumé succinct des points clef de l'entretien avec le client. Il utilise des informations automatisées et personnalisées générées par l'IAG, il reçoit aussi des suggestions de coaching personnalisées. Dans une entreprise ayant un service clientèle de 5000 agents, l'arrivée de l'IAG a augmenté de 14% par heure, le taux de résolution des plaintes clientèle, il a diminué le recours aux managers (moins 25%). Les employés les moins qualifiés ont augmenté leurs performances contrairement aux agents les plus qualifiés.

Dimensions de l'Irresponsabilité sociale	**Les domaines concernés**	**Le score que l'on peut attribuer**
Préjudice	Bien être du salarié (compétences/ degré de liberté/ collecte de données/ accélération du rythme de travail) Données personnelles client	2
Intentionnalité	Stratégie délibérée	5
Remédiation	Pas d'information à ce sujet	Neutre

[6] Page 14, The economic potential of generative AI: The next productivity frontier , Juin 2023

Le score serait de 7/10, les pratiques d'irresponsabilité concernent l'évolution des métiers commerciaux. Il s'explique par les bouleversements en cours : prise en compte des compétences et connaissances des meilleurs employés, statut des postes d'encadrement, rythme de travail et finalement sens du travail.

IAG et services marketing et vente (pages 16 et 17)

Les IAG vont collecter et analyser des données éparses (réseaux sociaux, actualités, recherche etc..). Les tendances en cours sont identifiées et donnent lieu à des proposition de communication plus efficaces (car personnalisés à l'extrême). Les agents conversationnels peuvent donner des recommandations comme des humains avec des signaux d'empathie et vont construire un rapport de confiance avec les consommateurs. Les clients sont fidélisés par les agents conversationnels qui personnalisent messages et récompenses, en minimisant les interactions négatives.

Dimensions de l'Irresponsabilité sociale	**Les domaines concernés**	**Le score que l'on peut attribuer**
Préjudice	Devenir de l'équipe de communication Données personnelles client, Manipulation du client , stratification de la société (voire discrimination)	2
Intentionnalité	Stratégie délibérée	5
Remédiation	Pas d'information à ce sujet	Neutre

Rapprocher la grille avec les éléments de réflexion dans les comités d'éthique et dans la réglementation en cours

Rapprocher avec la littérature illustrant l'irresponsabilité numérique

Ne pas laisser trop d'interprétation dans le choix des éléments constitutifs d'une IrSE numérique

Références

Acemoglu D, Autor D, Hazell J, et al. (2022) Artificial intelligence and jobs: Evidence from online vacancies. Journal of Labor Economics 40(S1): S293–S340

M., Bhaskar, R., Collier, A., Lawson, T., and Norrie, A.1998. Critical Realism: Essential Readings, London: Routledge.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115**.**

Baek, T. H., & Kim, M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics*, *83*, 102030.

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS quarterly*, *45*(1)

Bhaskar, R. 1978. A Realist Theory of Science, Hamel Hempstead,UK: Harvester.

Bryant, A., Griffin, J. J., & Perry, V. G. (2023). Irresponsible contagions: Propagating harmful behavior through imitation. *Business Ethics, the Environment & Responsibility*, *32*(1), 292-311.
Bernardo, E., & Seva, R. (2023, March). Affective Design Analysis of Explainable Artificial Intelligence (XAI): A User-Centric Perspective. In *Informatics* (Vol. 10, No. 1, p. 32). MDPI.
Carter, D. (2020). Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role?. *Business Information Review*, *37*(2), 60-68.
Clark C.E., Riera M., Iborra M. (2021). "Toward a theoretical framework of Corporate Social Irresponsibility: Clarifying the gray zones between responsibility and irresponsibility", *Business & Society*, June. doi:10.1177/00076503211015911
Corchado, J. M., López, S., Garcia, R., & Chamoso, P. (2023). Generative Artificial Intelligence: Fundamentals. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *12*(1), e31704-e31704.
DOI: https://doi.org/10.14201/adcaij.31704
Danermark, B., Ekstrom, M., Jakobsen, L., and Karlsson, J. 2002. Explaining Society: Critical Realism in the Social Sciences, London: Routledge
Doyal, A. S., Sender, D., Nanda, M., & Serrano, R. A. ChatGPT and Artificial Intelligence in Medical Writing: Concerns and Ethical Considerations. Cureus 2023; 15: e43292. DOI: 10.7759/cureus.43292
Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023a). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, Article 102642. https://doi. org/10.1016/j.ijinfomgt.2023.102642
Ernst E, Merola R and Samaan D (2019) Economics of artificial intelligence: Implications for the future of work. IZA Journal of Labor Policy 9(4). https://doi.org/10. 2478/izajolp-2019-0004.
Grinbaum A., Chatila R., Devillers L., Martin C. ,Kirchner C. Systèmes d'intelligence artificielle générative : enjeux d'éthique. Comité National Pilote d'éthique du numérique. 2023, pp. Avis 7 du CNPEN. cea-04153216
Irwin, Gretchen and Turk, Daniel (2005) "An Ontological Analysis of Use Case Modeling Grammar," *Journal of the Association for Information Systems*, 6(1), .
J. Fjeld, H. Hilligoss, N. Achten, M. L. Daniel, J. Feldman, S. Kagay, Principled artificial intelligence: A map of ethical and rights-based approaches (2019). URL https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf
Jovanovic, M., & Campbell, M. (2022). Generative artificial intelligence: Trends and prospects. *Computer*, *55*(10), 107-112.
Harfouche, A., Quinio, B., Saba, M., & Bou Saba, P. (2023). The recursive theory of knowledge augmentation (RTKA): Integrating domain knowledge with artificial intelligence to augment organizational knowledge. *Information Systems Frontiers*, *25*(3), 55–70. doi:10.1007/s10796-022-10352-8
Hull, C.E., Rothenberg S. (2008). "Firm performance: The interactions of corporate social performance with innovation and industry differentiation", Strategic Management Journal, 29, 781-789.
Houston, S. 2001. "Beyond Social Constructionism: Critical Realism and Social Work," British Journal of Social Work(31:6), pp. 845-861
Lange, D., & Washburn, N.T. (2012). Understanding attributions of corporate social irresponsibility. *Academy of Management Review*, 37(2), p.300-326. http://doi.org/10.5465/amr.2010.0522
Lin-Hi, N., & Müller, K. (2013). The CSR bottom line: Preventing corporate social irresponsibility. *Journal of Business Research*, 66(10), p.1928-1936. https://doi.org/10.1016/j.jbusres.2013.02.015
Monod, E., Mayer, A. S., Straub, D., Joyce, E., & Qi, J. (2024). From worker empowerment to managerial control: The devolution of AI tools' intended positive implementation to their negative consequences. *Information and Organization*, *34*(1), 100498.
Manyika J, Silberg J, Presten S (2019) What do we do about the biases in AI? Harvard Business Review, 25 October. Available at: https://hbr.org/2019/10/what-do-we-do-about-the-biases- in-ai

Maxwell, J. A. 1992. "Understanding and Validity in QualitativeResearch," Harvard Educational Review (62:3), pp. 279-300

Riera, M. and M. Iborra (2017). 'Corporate social irresponsibility:review and conceptual boundaries', European Journal of Management and Business Economics, 26, pp. 146–162.

Vaissière, T. (1999). L'éthique de responsabilité chez Hans Jonas à l'épreuve du droit international de l'environnement. *Revue interdisciplinaire d'études juridiques*, 43, 135-199. https://doi.org/10.3917/riej.043.0135

Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS quarterly*, 21-54.

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., ... & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(2), 7-30

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I., 2017.Attention is all you need. Advances in neural information processing systems, 30.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xiong, J.; Xu, L.; Li, Q.; Yuan, Z. & Chakraborty, S. (2023). Doing good and/or avoiding bad: the ambidextrous view of managing corporate social activities. *Management international*, 27(5), 25-36.