FIIDES : encourager la confiance pour le déploiement des systèmes d’intelligence artificielle

Rapport du groupe de recherche sur la confiance pour le déploiement des systèmes d’intelligence artificielle

Arnaud Latil
Michel Dubois
Servanne Monjour

Rapport déposé le 28 juillet 2023

Période de recherche : septembre 2022 – juin 2023

Résumé et recommandations

Le groupe de recherche a mené des entretiens semi-dirigés avec des *data scientists* collaborateurs de partenaires du programme confiance.ai entre janvier et juin 2023. Ces entretiens ont permis d'observer les méthodes de travail, les missions, l'organisation et, plus généralement, leurs perceptions des rapports entretenus avec les autres métiers et donc de l'intégration des outils d'intelligence artificielle dans l'organisation du travail.

Ces enquêtes ont permis de mettre à jour les enjeux et les croyances, réelles ou supposées, du déploiement de systèmes d'intelligence artificielle sur l'emploi et les savoir-faire. Elles ont aussi révélé les différentes stratégies de bricolage organisationnel mises en place par les *data scientists*, parfois en lien avec les métiers, afin de déployer des outils d'intelligence artificielle. Ces bricolages visent en particulier à faire accepter aux métiers les outils d'intelligence artificielle. Pour cela, les *data scientists* imaginent des techniques de communication afin d'expliquer le fonctionnement et les limites de l'intelligence artificielle. Ils développent aussi des méthodes destinées à assurer le contrôle des décisions par les personnes. Ces enquêtes ont aussi permis d'identifier des attentes de la part des *data scientists* en matière de formation à l'environnement éthique et réglementaire, d'explicabilité des systèmes d'intelligence artificielle et de collaboration avec les métiers.

À partir de ces enquêtes et des réflexions des membres du groupe de recherche, il est proposé aux pouvoirs publics et aux organisations de mettre en œuvre plusieurs initiatives en vue d'encourager le déploiement des systèmes d'intelligence artificielle. Ces propositions sont regroupées sous l'acronyme **« FIIDES »**, pour **Formation, Intégration, Information, Développement, Soutenabilité**, dont les principales actions sont les suivantes :

1. Formation
1.1. Former les collaborateurs à l'environnement éthique et réglementaire de l'IA
1.2. Recruter un ou une « Responsable des Systèmes d'IA » (RSIA) sur le modèle des DPO

2. Intégration
2.1. Recruter une personne « chargée de la transformation vers l'IA »
2.2. Mettre en place un comité éthique de l'IA et rédaction de chartes internes
2.3. Reconnaissance de droits de co-détermination

3. Information
3.1. Mettre en place une communication interne
3.2. Adapter les interfaces graphiques des systèmes d'IA
3.3. Anticiper les futures obligations d'information

4. Développement
4.1. Mettre en place des outils d'audit en IA et préparer l'entrée en vigueur de l'IA Act
4.2 Promouvoir les initiatives normatives privées
4.3. Mettre en place une gouvernance des données d'entraînement

5. Soutenabilité
5.1. Les considérations environnementales
5.2. La préservation des savoir-faire
5.3. La prise en compte du bien-être au travail
5.4. La protection de la souveraineté numérique

Sommaire

Introduction

1. Présentation du groupe de recherche

Le groupe de recherche s'est constitué en réponse à l'Appel à Manifestation d'Intérêt (AMI) publié en avril 2022 par le programme national Confiance.AI à destination de la communauté de recherche en Sciences Humaines et Sociales.

L'objectif de ce programme de recherche consistait à étudier les mécanismes de la confiance, tant sur un plan individuel que collectif, à travers les sciences humaines et sociales.

Les contributions scientifiques attendues dans le cadre de cet AMI étaient de deux ordres : dans un premier temps, il était attendu que les groupes de recherche sélectionnés apportent un regard critique sur le programme, qu'ils s'investissent personnellement en rencontrant les acteurs de l'IA et, enfin, qu'ils proposent des recommandations destinées à favoriser la confiance dans les systèmes d'IA.

Dans un second temps, le programme Confiance.AI envisage le financement de travaux de recherche supplémentaires (postdoctorat, temps d'ingénieur(e) de recherche, thèse, etc.).

Le présent rapport de recherche s'inscrit dans le cadre d'une réponse à l'AMI concernant cette première phase seulement.

Le présent groupe de recherche a répondu à cet AMI par une proposition en date du 23 mai 2022. Cette proposition visait à étudier les interfaces graphiques des systèmes d'IA.

Le groupe de recherche est composé des chercheuses et chercheurs suivants :

- **Arnaud Latil,** maître de conférences HDR en droit à Sorbonne Université, membre du Sorbonne Center for Artificial Intelligence (SCAI) et du Centre d'Etudes et de Recherches en droit de l'Immatériel (CERDI) de l'Université de Saclay
- **Michel Dubois**, sociologue, directeur de recherche CNRS, directeur du Groupe d'Etude des Méthodes de l'Analyse Sociologique de la Sorbonne (GEMASS — UMR8598), CNRS — Sorbonne Université
- **Servanne Monjour**, maîtresse de conférence en Humanités numériques à Sorbonne Université.

Les membres du collectif :

2. Évolution de la problématique de travail

La réponse à l'AMI prévoyait dans un premier temps l'étude des interfaces graphiques des systèmes d'IA (également dénommées « Interfaces Humain-Machine » ou IHM). L'hypothèse de recherche retenue consistait alors à supposer que les éléments de communication associés aux systèmes d'IA, comme les logos et plus largement les informations relatives au fonctionnement des SIA, participent, d'une manière ou d'une autre, à la confiance des utilisateurs.

Afin de mener à bien ce projet, il était attendu de pouvoir accéder auxdites interfaces graphiques.

Le cas d'usage retenu en concertation avec l'IRT Système X, porteur de cet AMI pour Confiance.AI, portait sur le « cas d'usage Renault », à savoir un système d'IA de type « Large Language Models » (LLM) dont la fonction consiste à analyser des avis clients disponibles sur le moteur de recherche « Google Search ».

Or, nous nous sommes aperçus après avoir mené une série d'entretiens avec les collaborateurs impliqués dans la conception de cet outil que nous ne pourrions pas y avoir accès. Les motifs de ce refus relevaient d'impératifs de secret des affaires. Nous avons donc tenté de mener des entretiens avec les usagers de cet outil (les « métiers »), à savoir les ingénieurs destinataires des avis clients. Or, il s'est avéré impossible de les rencontrer. Les motifs de ce refus sont restés incertains. La défiance envers ces outils, remettant en cause une part de l'activité d'analyse des avis clients par ces services, semble être à l'origine de l'impossibilité de poursuivre notre travail de recherche.

Dans ce contexte, nous avons décidé de poursuivre nos travaux en prenant acte de la difficulté d'étudier les IHM.

Plus encore, il nous est apparu que les crispations rencontrées lors de l'analyse de ce cas d'usage méritaient assurément une analyse. La question de la confiance dans les outils d'IA et dans leurs modalités de déploiement est en effet apparue comme le centre des difficultés. Pour cette raison, le groupe de recherche a décidé d'orienter son analyse vers celle des freins au déploiement des SIA dans les organisations.

Afin de mener à bien ce projet redimensionné, le groupe de recherche a mené des entretiens avec les services de data science (ie. les *data scientists*) des partenaires volontaires de Confiance.IA ayant répondu à nos sollicitations, à savoir Renault, Naval Group, Safran et Valéo. En outre, nous avons également mené un entretien avec des chercheurs de l'ICAm, membres du programme confiance.ai.

« *Data scientists* ». Le terme de « *data scientists* » semble s'être imposé en France pour désigner les professionnels notamment en charge de concevoir des systèmes d'intelligence artificielle (leurs compétences peuvent dépasser le cadre de l'intelligence artificielle et porter plus généralement sur le traitement de données).

Cette expression présente au moins une qualité, mais aussi deux faiblesses :

— la qualité de cette expression consiste à mettre l'accent sur la notion de *données* (data), distinguant en cela ces professionnels des informaticiens et des développeurs. Ces professions sont en effet différentes. Les *data scientists* sont issus du monde des mathématiques et de la statistique et sont parfois dépourvus de compétences en informatique.

— la première faiblesse de cette expression est qu'elle ne connaît pas de traduction française communément admise. On les désigne parfois comme des « spécialistes des données » ou des « consultants en informatique décisionnelle ».

— la seconde faiblesse porte sur le caractère trompeur du terme « *scientist* ». Ces professionnels ne sont généralement pas des « scientifiques », mais plutôt des ingénieurs. Bien qu'ils soient parfois titulaires d'un doctorat, leurs missions ne s'apparentent pas à de la recherche au sens strict. Les professionnels rencontrés n'exercent pas non plus leurs activités auprès de laboratoires de recherche privés (comme il peut en exister par ailleurs).

Compte tenu de ces faiblesses, nous proposons de nommer ces personnes des « **ingénieurs en intelligence artificielle** ». Par commodité, le présent rapport utilise indifféremment les deux expressions.

———

Enfin, le cadre juridique de l’intelligence artificielle se trouve en pleine construction. La présente étude a aussi pour objectif de mesurer le degré de préparation des organisations à l’entrée en vigueur d’un cadre juridique en préparation. Les recommandations formulées dans ce rapport de recherche intègrent certains éléments du futur cadre juridique de l’IA.

I. Méthode et champ de l'étude

La confiance provient du latin *fides* (la foi), *fidèle*, mais aussi de *fiducia* (la fiducie). « Faire confiance » signifie pour une personne qu'elle accorde du crédit à une personne ou à une chose (en réalité aux personnes qui ont conçu, produit ou qui exploitent la chose). La confiance s'analyse ainsi non comme un mécanisme de « croyance », mais plutôt comme la traduction d'une situation de « fiabilité » (McLeod C. [2011], « Trust », *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (dir.)[1]). Au-delà, le point le plus important est que la confiance se mesure toujours par rapport à des *attentes*, par nature variables suivants les personnes ou les contextes.

La confiance fait l'objet de très nombreuses études et réflexions en sciences humaines et sociales, aussi bien d'un point de vue sociologique, économique ou anthropologique, que psychologique, politique et juridique. Elle est fréquemment associée à deux dimensions : une dimension interpersonnelle et une dimension institutionnelle (*OCDE Guidelines on Measuring Trust*, 2017).

La nature des attentes relatives à chacune de ces dimensions peut différer. À un niveau interpersonnel, dire d'une personne qu'elle est « digne » de confiance c'est à la fois porter un jugement moral sur cette personne, mais c'est aussi définir plus ou moins explicitement un certain nombre d'attentes la concernant, par exemple pouvoir compter sur sa compétence ou son honnêteté, ou encore se sentir suffisamment en sécurité avec elle pour partager de façon ouverte ses pensées ou ses émotions. À un niveau institutionnel, dire d'une institution, d'une organisation voire d'une profession qu'elle génère de la confiance, c'est fréquemment décrire la capacité de cet acteur collectif à apparaître comme juste et équitable dans ses pratiques, ou encore responsable de ses actions comme de leurs conséquences en faisant preuve de transparence et de réactivité. Dans un cas comme dans l'autre la relation de confiance apparaît comme une situation de *vulnérabilité à autrui* dans des conditions d'*interdépendance* et d'*incertitude*. L'interdépendance signifie que l'intérêt d'une partie ne peut être satisfait sans dépendre d'une autre partie. L'incertitude renvoie à l'impossibilité de garantir une issue positive. Enfin, lorsque nous choisissons de faire confiance à autrui, nous nous exposons toujours à la possibilité d'être blessés, trompés ou déçus.

La confiance possède une valeur intrinsèquement liée à l'*incertitude* et au *risque* : lorsqu'une situation ne présente aucune incertitude ni aucun danger, la confiance n'est tout simplement pas mobilisée. En revanche, dès lors qu'une situation présente un risque, le choix de s'y engager dépend du degré de confiance accordé aux personnes ou aux choses. En reprenant l'expression proposée par Niklas Luhmann, la confiance constitue un « *mécanisme de réduction de la complexité* » (N. Luhmann, La confiance, un *mécanisme de réduction de la complexité sociale*, Economica, 2006 pour la trad. fr.). La confiance permet en effet aux personnes d'éviter d'impossibles, ou de difficiles, calculs de probabilité qu'un évènement

[1] Cf. en ligne [http://plato.stanford.edu/archives/spr2011/entries/trust/].

défavorable ne survienne. Il est cardinal de souligner ici que la confiance est un mécanisme destiné à éliminer le calcul des *incertitudes* et non les incertitudes elles-mêmes (N. Luhmann, *op. cit.*, p. 31).

L'étude de la confiance s'appuie généralement sur deux méthodes : les expériences et les enquêtes (E. Laurent, *Economie de la confiance*, La Découverte, 2019, pp. 43 à 68). À l'échelle internationale, de nombreuses enquêtes intègrent des mesures répétées des attitudes de confiance présentes dans les populations enquêtées. C'est le cas par exemple pour l'enquête *European Social Survey* (ESS) qui, à l'initiative de la Fondation européenne de la science, mesure tous les deux ans l'évolution des attitudes, des croyances et des comportements des Européens. À l'échelle mondiale, le World Values Survey (WVS) examine tous les cinq ans les valeurs et les croyances des individus dans différents pays, y compris la confiance interpersonnelle et institutionnelle. Ces enquêtes permettent souvent de souligner l'importance des facteurs nationaux et culturels associés à l'expression de confiance. En France par exemple, à l'occasion de la vague 2017-2022 du WVS, ils étaient 7 enquêtés sur 10 à considérer qu'« il n'est pas possible de faire confiance à la plupart des gens ». Pour la même enquête et sur la même période, en Norvège ou en Finlande, ils étaient à l'opposé 7 enquêtés sur 10 à considérer qu'il est tout à fait possible de faire confiance à la plupart des gens.

L'objectif d'encourager la confiance dans les systèmes d'IA fait l'objet d'un large consensus en termes de politiques publiques. En particulier, l'Union européenne vise à créer les conditions d'une « IA digne de confiance » (Rapport des Experts de haut niveau sur l'IA digne de confiance, 2020). L'OCDE organise également diverses actions destinées à parvenir à cet objectif (notamment par la création d'un espace ouvert de contribution sur les « outils et métriques »[2]).

Cette mobilisation institutionnelle doit toutefois tenir compte de la diversité des jugements et des attitudes à l'égard de l'intelligence artificielle à l'échelle internationale. En décembre 2020, le Pew Research Center[3] présentait les résultats d'une étude réalisée sur 20 pays. L'étude révèle de fortes disparités régionales. Les opinions sur l'IA sont généralement positives parmi les publics asiatiques interrogés. Environ deux tiers ou plus des populations enquêtées à Singapour (72 %), en Corée du Sud (69 %), en Inde (67 %), à Taïwan (66 %) et au Japon (65 %) affirment que l'IA est une bonne chose pour la société. Mais la plupart des autres pays interrogés ne parviennent pas à obtenir une majorité pour dire que l'IA est une bonne chose pour la société. En France, en particulier, les opinions sont particulièrement négatives : seulement 37 % des personnes interrogées estiment que l'IA peut être bénéfique pour la société, contre 47 % qui estiment qu'elle est néfaste pour la société. Ce résultat concernant la France a été confirmé par les résultats de l'enquête *Les français et la science 2021* coordonnée par M. Bauer, M. Dubois et P. Hervois[4] qui a montré que seuls 4

[2] Cf. en ligne [https://oecd.ai/en/catalogue/tools?terms=&page=1].

[3] Cf. en ligne [https://www.pewresearch.org/short-reads/2020/12/15/people-globally-offer-mixed-views-of-the-impact-of-artificial-intelligence-job-automation-on-society/].

[4] Cf. en ligne [https://www.science-and-you.com/sites/science-and-you.com/files/users/documents/les_francais_et_la_science_2021__rapport_de_recherche_web_v29112021_v2.pdf].

enquêtés sur 10 (39 %) considéraient que l'intelligence artificielle allait contribuer à améliorer leur qualité de vie. Il faut donc conserver à l'esprit que la question de la confiance accordée à l'intelligence artificielle singularise tout particulièrement la France par rapport à la plupart des pays développés étudiés.

L'analyse de la « confiance » dans les entreprises concerne classiquement ses relations entretenues avec au moins cinq catégories d'acteurs : les consommateurs, les citoyens, les actionnaires, les fournisseurs et ses collaborateurs (F. Caby, V. Louise et S. Rolland, *La qualité au XXIe siècle. Vers le management de la confiance*, Economica, 2002, p. 105). Construire la confiance avec ces différents partenaires implique alors la mobilisation de différents outils portant sur la qualité et la sécurité des services et des produits, les engagements éthiques, sociaux et environnementaux des entreprises ou encore plus largement le respect des normes plus ou moins obligatoires.

L'analyse des effets de l'IA sur le marché du travail appelle une attention particulière. Le thème de la « perte d'emploi » et du remplacement de certains métiers par l'IA se trouve en effet au centre du débat public. Ce thème s'est trouvé renforcé lors de l'hiver 2022/2023 sous l'effet de la forte notoriété acquise par *ChatGPT*, mais aussi à travers la diffusion massive d'images générées par des systèmes d'IA comme *Midjourney* ou *Stable diffusion*. De nombreux articles de presse et des reportages dans les médias ont ainsi discuté des effets possibles de l'IA sur les professions intellectuelles (graphistes, avocats, médecins, professeurs, journalistes, etc.).

Une étude publiée par l'OCDE en 2021 portant sur les suppressions d'emplois induits par l'IA ne met pourtant en valeur aucune corrélation nette entre l'usage de SIA et l'emploi. On peut y lire que « *globalement, il ne semble pas y avoir de relation claire entre l'exposition à l'IA et la croissance de l'emploi* ». Les auteurs du rapport avancent cependant avec prudence qu'il existerait toutefois une corrélation entre l'amélioration de la productivité grâce à l'IA et le degré d'informatisation d'une profession.

Enfin sur un plan plus descriptif, pour la situation française, il faut souligner l'intérêt des enquêtes INSEE sur *Les TICS et le commerce électronique dans les entreprises*[5] qui permettent de suivre la diffusion dans le milieu du travail du traitement automatique du langage naturel, l'analyse de données massives ou encore l'utilisation du Machine Learning.

II. Présentation des cas

Compte tenu du dimensionnement du programme confiance. AI et de l'identité des parties prenantes à ce programme, les entretiens réalisés comprennent un important biais de résultat, à savoir que nous n'avons recueilli que le point de vue des personnes en charge de développer des systèmes d'IA dans les organisations étudiées (les *data scientists*). Comme exposé plus

[5] Cf. en ligne [https://www.insee.fr/fr/statistiques/5349831]

haut, nous ne sommes pas parvenus à réaliser d'entretiens avec le management ni avec les métiers. L'objectif des entretiens a donc consisté à analyser les stratégies et les moyens déployés par les *data scientists* en vue de générer la confiance dans les SIA développés.

A. Icam (École d'ingénieurs à Toulouse, membre du programme confiance.ai)

Entretien avec Charly Pecost et Yann Ferguson (chercheurs à l'ICAM) le vendredi 6 janvier 2023

Ce premier entretien a été l'occasion d'échanger avec deux enseignants-chercheurs travaillant dans le cadre du LaborIA, un centre de ressources et de recherches créé en novembre 2021 et entièrement consacré aux transformations opérées par l'IA dans le monde du travail.

Charly Pecost et Yann Ferguson sont tous les deux sociologues de formation avec une spécialisation dans le domaine de l'ergonomie cognitive. En privilégiant l'étude des usages de l'IA en entreprise, leurs travaux visent à rendre compte de la manière dont les technologies modifient plus ou moins en profondeur les conditions de travail.

Les échanges ont permis de souligner un paradoxe relatif au déploiement de l'IA dans le cadre des entreprises étudiées avec d'un côté la nécessité affichée de réfléchir en profondeur à ses conditions d'usages et d'acceptabilité et de l'autre la difficulté, voire l'impossibilité, d'en parler ou simplement d'utiliser le terme même d'« intelligence artificielle », parfois jugée par le management trop problématique. Cette tension autour de l'IA en entreprise l'étudier sans en parler explique en partie la difficulté d'investir ce terrain pour les sciences humaines et sociales.

Charly Pecost et Yann Ferguson ont également décrit leurs propositions pour « outiller la confiance » chez l'industriel Renault avec un outil de diagnostic sous forme de bref questionnaire (passation en 15 min) construit autour de 4 familles de critères, au nombre desquels figurent l'engagement du professionnel dans son métier, son bien être et le sentiment reconnaissance liée à son activité, le sentiment de contrôle sur ses outils, etc.

Parmi les productions de LaborIA, il faut notamment souligner pour la thématique de notre groupe de travail le rapport rendu public en mars 2023, intitulé « Usages et impact de l'IA sur le travail au prisme des décideurs[6] ».

B. Renault

B1. Entretien avec Ayhan Uyanik, data scientist chez Renault et Vincent Feuillard, le 10 janvier 2023

[6] Cf. en ligne [https://travail-emploi.gouv.fr/IMG/pdf/enquete_laboria.pdf].

Lors d'un premier entretien avec l'équipe Renault [B1], nous avons échangé en présentiel avec Ayhan Uyanik, référent *data science* et satisfaction client chez Renault depuis 6 ans, accompagné de Vincent Feuillard (en visioconférence). Le service de Direction Qualité dans lequel travaille Ayhan Uyanik comprend une équipe d'une dizaine d'employés, parmi lesquels des profils de statisticiens (qui se concentrent sur l'évaluation de la fiabilité des véhicules) et des profils de *data scientist* (qui évaluent la satisfaction des clients).

Le cas d'usage analysé porte sur un outil d'analyse automatisé des avis clients. Il s'agit d'un outil de NLP destiné à capter, analyser et classifier les retours clients. Les données analysées, pour le moment, sont les verbatim des formulaires de satisfaction remplis par les clients suite à une plainte (un développement est prévu pour faire évoluer ce modèle d'analyse des plaintes vers un modèle d'analyse des opinions, prenant en compte des données plus larges, comme les évaluations sur Google, par exemple). L'objectif de l'outil est de limiter au maximum le temps de relecture humaine des verbatims et de proposer une catégorisation automatique des plaintes. Ce projet s'inscrit dans le cadre de travaux d'évaluation de la réputation des concessionnaires, destinés à aligner les expériences client et améliorer le service à la clientèle dans l'ensemble des enseignes Renault. Concrètement, l'outil se présente sous la forme d'une interface qui propose une catégorisation des avis clients. Nous avons pu brièvement voir l'outil lors de ce premier entretien, avant d'assister à une démonstration lors du second entretien [B2]. Nous n'avons en revanche pas pu tester nous-mêmes cet outil.

B2. Entretien avec Mathieu Sarazin, le 2 février 2023 (au technocentre de Guyancourt).

L'entretien est réalisé en présence de Ayhan Uyanik.

Mathieu Sarazin travaille dans le groupe Renault depuis près de 14 années. Après avoir été chef du service Statistiques & Data science, il est aujourd'hui Chef du service Qualité Client Numérique (Customer Quality Digital Leader). Son service a pour objectif de capter et d'analyser tous les signaux d'insatisfaction client (produit et service) à partir des nombreuses sources de données internes et externes au Groupe Renault, ainsi que de les prioriser et de les communiquer à l'ensemble des acteurs afin d'accélérer leur résolution.

Mathieu Sarazin rappelle l'origine des données traitées, qui découlent essentiellement des questionnaires associés aux ventes de véhicules ainsi que des formulaires remplis à l'occasion d'un passage chez un concessionnaire. L'ensemble consiste en un volume important de données, comprenant le verbatim en langage naturel de la description des problèmes rencontrés par les clients.

Par delà la demande initiale du management de réduire les coûts et les effectifs associés à la lecture des avis clients, l'enjeu technique du service est de transformer le « problème client » en un « problème d'ingénierie ».

L'entretien est l'occasion de réfléchir aux différents facteurs qui pèsent sur la diffusion des outils développés par ce service. Quatre facteurs sont ainsi fréquemment évoqués :

1) La perte d'emploi : l'IA est perçu comme une forme d'automatisation susceptible de se substituer à de l'emploi humain ;

2) La liberté de choix : les modèles développés permettent d'obtenir le plus souvent un résultat robuste dans l'affectation d'un avis client, mais ce faisant ils sont souvent perçus comme un retrait de liberté pour l'utilisateur. D'où la nécessité de maintenir une part décisionnelle à l'utilisateur qui peut passer notamment par l'utilisation d'un ranking avec des scores de vraisemblance.

3) La compréhension : la diffusion des outils IA suppose également une phase d'acculturation à l'IA qui n'est pas toujours aussi avancée que nécessaire. La présentation des modèles, de leurs comportements, permet d'informer, mais également le plus souvent de rassurer les métiers.

4) L'accessibilité des données : la difficulté des outils IA tient aussi parfois à l'invisibilisation des données brutes. Les utilisateurs ont besoin de conserver la possibilité de revenir vers le verbatim original.

Mathieu Sarazin souligne l'intérêt qu'il pourrait y avoir pour un industriel comme Renault à constituer et animer des communautés de consommateurs-experts susceptibles d'enrichir le matériau qualitatif traité par les outils IA en cours de développement.

C. Naval Group

Entretien avec Anthony Rossi, ingénieur data scientist, et Jérémy Trione, docteur en mathématique, le 17 avril 2023, par visioconférence.

Anthony Rossi dirige un service comptant quatre personnes, dont une alternante, situé à Toulouse. Plusieurs cas d'usage de systèmes d'intelligence artificielle ont été exposés. Les niveaux de développement des différents cas d'usage restent très inégaux. Le degré de maturité de ces projets reste à ce stade ici confidentiel et certains projets ont été abandonnés. Ces précisions étant apportées, les cas d'usage suivants ont été présentés :

- la maintenance prédictive et la détection des pannes susceptibles d'affecter des bâtiments ;
- L'aide à la gestion des plannings de maintenance des bâtiments (emploi du temps du personnel, occupation des sites de maintenance, etc.) ;
- L'affichage des données ;
- L'analyse de formulaires et de contrats à partir d'un outil de NLP, ainsi que l'automatisation du calcul des pénalités de retard et recherches documentaires.

En outre, un certain nombre de projets plus secondaires consistent en des assistances ponctuelles aux métiers.

D. Safran

Le groupe de recherche a réalisé deux entretiens avec l'entreprise Safran.

D1. Entretien réalisé avec Frédéric Jenson, responsable de programme IA, le 24 avril 2023, par visioconférence.

Monsieur Frédéric Jenson possède une formation de physicien avec une spécialité en matière d'imagerie médicale. Il est en charge de l'IA pour Safran. Il supervise une équipe de huit ingénieurs docteurs dans le cadre d'une division de recherche et développement.

Le cas d'usage présenté porte sur l'utilisation d'un système d'IA en matière de contrôle non destructif, c'est-à-dire d'inspection de pièces fabriquées à partir d'images de celles-ci. L'IA représente ici une aide aux opérateurs dans les opérations de vérification. Ces opérateurs sont des techniciens et des ingénieurs certifiés.

La certification des personnes occupe dans ce contexte une place très importante. L'usage de systèmes d'IA doit aussi se comprendre compte tenu des contraintes réglementaires fortes propres aux secteurs de l'industrie aéronautique.

D2. Entretien avec Bertrand Bouttier, responsable d'atelier, le 30 mai 2023, par visioconférence.

Le second entretien réalisé au sein du groupe Safran nous a donné l'occasion d'échanger avec Bertrand Bouttier, qui dirige un atelier de contrôle qualité chez Safran depuis 6 ans.

Bertrand Bouttier s'est spécialisé dans les contrôles non destructifs en lien avec la performance des installations industrielles et la sécurité des matériaux. Chez Safran, il œuvre dans le domaine du contrôle de la qualité afin d'améliorer les performances et les conditions de travail des opérateurs à l'aide de moyens numériques (intelligence artificielle, deep learning). L'outil qu'il nous a présenté consiste essentiellement en de l'aide à la sanction : l'outil offre une assistance aux opérateurs dans leur travail d'évaluation des pièces à partir de l'analyse de données visuelles (une série de photographies établies selon un protocole). Concrètement, l'outil intervient à la fin du processus de contrôle : les opérateurs doivent procéder à une première évaluation fondée sur l'examen et l'annotation des visuels de la pièce. L'algorithme opère ensuite un second contrôle à titre de comparaison. En cas d'écart entre les deux évaluations, un avertissement apparaît à l'écran du poste de travail de l'opérateur — à l'inverse, si aucun écart n'est relevé, alors l'IHM n'intervient pas. L'outil

fonctionne donc avant tout comme un filet de sécurité, à destination des opérateurs (profil de technicien ou technicien supérieur, niveau de recrutement Bac ou Bac+2).

E. Valeo

Entretien avec Monsieur Éric Feuilleaubois, ingénieur IA, le 16 mai 2023 (visioconférence)

Eric Feuilleaubois est responsable du développement d'un outil d'assistance à la conduite. Cet outil se fonde sur des briques d'IA capables de faire de la perception (reconnaissance des objets, des lignes de parking, etc.), en analysant des données issues de capteurs (caméra et ultrasons) installés sur les véhicules. L'outil est conçu selon un principe hybride, associant IA (en particulier *computer vision*) et logique classique sur les manœuvres de parking. L'IA a été développée à partir d'un modèle d'entraînement comprenant 130 scènes (au rythme de 10 à 15 images/seconde) captées dans des rues de Paris, Los Angeles, Francfort et Stuttgart.

Pour le moment, l'outil se cantonne à des tâches de parking autonome, mais l'entreprise évoque la possibilité d'expérimenter des tâches de planification de trajectoire. Par ailleurs, il s'agit d'un outil destiné à une clientèle limitée et aisée : les véhicules équipés du système d'assistance sont des véhicules haut de gamme (BMW, Mercedes).

III. Résultats de recherche

Les résultats que nous présentons reflètent essentiellement les discours recueillis lors des entretiens menés avec les *data scientists* que nous avons interrogés sur l'organisation de leur travail, sur leurs attentes et leurs réalisations. Les témoignages recueillis comprennent des informations précises sur les défis rencontrés par les équipes en charge du développement d'outils IA et sur les solutions mises en place pour les relever, mais ils livrent également un aperçu du ressenti des salariés au travail, en particulier sur les enjeux de confiance dans le cadre de leur mission.

Il faut souligner au préalable que le profil des *data scientists* interrogés est assez uniforme : il s'agit exclusivement d'hommes possédant des diplômes d'ingénieurs ou un doctorat, avec une forte appétence pour la recherche. La tranche d'âge de nos interlocuteurs s'est révélée en revanche très variable, comprenant des salariés diplômés depuis peu, d'autres bénéficiant de 15 ans d'expérience, mais également des seniors.

De manière générale, nous avons observé un fort esprit d'entreprise et un important enthousiasme concernant les virtualités de l'IA. En revanche, nous avons constaté une assez faible mobilité professionnelle des *data scientists* interrogés, qui restent fidèles à leur entreprise — au sein de laquelle ils peuvent cependant bénéficier d'une mobilité interne.

Trois thèmes, que nous présentons ci-dessous, ont dominé ces entretiens semi-dirigés : le premier sujet d'échange a porté sur l'organisation des *data scientists* au sein des organisations concernées (**A**). Les *data scientists* ont ensuite discuté des opportunités et des craintes liées aux conséquences des SIA sur l'emploi (**B**). Enfin, les entretiens ont permis de recueillir certaines attentes (**C**).

A. Les organisations face à l'IA

A.1. Stratégies d'entreprise

D'abord, d'un point de vue macro-organisationnel, la première observation saillante porte sur la place occupée par l'IA dans les stratégies d'entreprises. Certaines organisations se présentent en effet comme des entreprises de la « tech », bien que l'informatique ne soit pas le cœur de leur activité, tandis que d'autres ont conservé un discours plus traditionnel face au numérique. Concernant la première catégorie, on citera le cas de l'entreprise Renault qui se trouve engagée dans une stratégie résolument tournée vers la tech et l'IA. Le recrutement de Luc Julia, star de l'IA et concepteur de l'application « Siri » proposé par Apple, marque un virage notable vers l'IA. Notons cependant qu'un tel recrutement, sans aucun doute décisif pour consolider l'image « tech » de l'entreprise, n'a qu'un impact très limité sur les activités de la cellule IA avec laquelle nous avons échangé, qui n'a jamais été amenée à collaborer avec l'intéressé. Dans d'autres organisations, bien que l'informatique occupe pourtant une place importante, le discours d'entreprise relègue au second plan ces activités. Dans tous les

cas, les *data scientists* perçoivent le virage plus ou moins marqué de leur entreprise vers l'IA, et ils y sont sensibles.

Ainsi, nous avons nous-mêmes perçu une corrélation entre la stratégie d'entreprise et l'intégration des *data scientists* auprès des équipes métiers, sans toutefois que l'étendue de cette enquête soit suffisamment massive pour nous permettre de le démontrer.

Dans certains cas, les entretiens ont pourtant fait apparaître un sentiment de décorrélation entre le virage général de l'organisation vers l'IA et la tech et les difficultés rencontrées par certains métiers à l'acculturation au numérique. Dans ces cas précis, les impulsions du management en faveur de l'IA ne paraissent pas suivies d'effets tangibles quant aux modes d'organisation des métiers. Certains *data scientists* témoignent même d'un sentiment de méfiance qu'ils perçoivent à leur égard de la part de salariés inquiets de voir leurs habitudes de travail bouleversées.

A.2. Les conséquences sur l'emploi

Le second grand thème significatif porte sur les craintes ressenties par certains salariés spécialisés quant aux risques de remplacement et/ou de mutation de leur emploi. Les *data scientists* interrogés se sentent bien souvent en première ligne face aux craintes de leurs collègues auxquels ils ne savent pas toujours comment répondre — une tâche qui, d'ailleurs, ne leur incombe pas. S'il a été impossible durant cette enquête de mesurer avec précision la nature et l'étendue des craintes liées à la perte d'emplois (notre échantillon aurait en effet dû s'étendre au-delà des communautés de *data scientists*), il est en ressorti que ce sujet entraînait une perte de confiance notable dans les relations entre les *data scientists* et les métiers.

Concrètement, il est apparu à plusieurs reprises que des déploiements de projets IA ont été mis à l'épreuve en raison des enjeux liés aux mutations des conditions de travail. Il n'a cependant pas été possible de mesurer avec précision l'étendue précise de cette question, qui semble pourtant fondamentale en matière de confiance.

En outre, les enjeux liés à l'évaluation des performances des collaborateurs semblent aussi représenter une question débattue. Elle se trouve directement liée à l'amélioration supposée de la productivité grâce aux SIA. Il faut toutefois préciser que la présente étude n'a pas permis de mesurer ces gains de productivité, réels ou supposés. Il reste que l'enjeu de confiance est ici central : des attendus (réels ou non) de la part des utilisateurs ont semble-t-il contribué à les détourner des outils d'IA proposés par les *data scientist*.

Si les difficultés identifiées ont essentiellement porté sur la productivité, et donc sur le temps de travail, elles ne concernent en revanche pas le savoir-faire des métiers, qui demeure une valeur essentielle aux yeux des *data scientists*. Dès lors, l'enjeu central ne porte donc pas sur la *disparition* des métiers que nous avons identifiés, mais sur la *réduction des effectifs*, question étroitement liée à celle de l'*évaluation des performances.*

Enfin, on ne peut cependant pas non plus écarter le rôle certainement joué par le débat public relatif au problème de la désintermédiation — autrement dit au supposé « remplacement » des personnes par les machines animées par des SIA. Ce thème s'est en effet imposé dans le débat public, sans qu'il soit possible à ce stade de mesurer son effet. Notons cela dit que ce discours sur la désintermédiation a connu des précédents dans l'histoire de l'autonomisation du travail depuis le XIXe siècle au moins. Dans l'histoire récente, l'informatisation du travail a pu générer des craintes similaires. La quasi-totalité des *data scientists* que nous avons rencontrés a exprimé sa perplexité face à l'emballement médiatique qui se joue actuellement autour de l'IA, et à l'imaginaire que cette technologie cristallise à présent.

A.3. L'intelligence artificielle, une notion parfois nébuleuse

Dans un troisième temps, nous avons identifié une zone de tension autour de la complexité à définir ce que recouvre exactement la notion d'IA — certains de nos interlocuteurs estiment d'ailleurs que le terme est aujourd'hui employé de manière abusive, y compris au sein de leur entreprise, et préfèrent employer d'autres termes (par exemple, de la « perception avancée »). Le terme IA recouvre en effet plusieurs techniques différentes et, surtout, plusieurs usages (NLP, computer vision, IA générative, etc.). Dans certains cas l'IA est susceptible d'accélérer les activités humaines. Dans d'autres cas, l'IA ouvre de nouvelles applications auparavant impossibles à envisager. Or les conséquences organisationnelles de ces différentes applications sont très distinctes. Si bien qu'il apparaît difficile, au sein des organisations, d'envisager la « stratégie IA » de manière uniforme.

Lorsque l'utilisation de l'IA vise à améliorer la productivité en déchargeant les collaborateurs de tâches à faible valeur ajoutée, elle permet, en principe, de concentrer l'activité sur des opérations à plus haute valeur ajoutée. Toutefois, nous avons constaté que ce prétendu phénomène de transfert d'activité demeure loin d'être automatique. Plusieurs causes peuvent être avancées :

- la volonté des collaborateurs de conserver des tâches moins exigeantes intellectuellement. Ces activités à faible valeur ajoutée jouent en quelque sorte un rôle de sas de décompression essentiel à l'équilibre du salarié (par exemple, au sein d'un service de contrôle qualité, des ingénieurs peuvent consacrer 1h de leur journée à relire des formulaires de satisfaction des clients, avant de reprendre une activité plus exigeante) ;
- l'ancrage des habitudes de travail ;
- la volonté de conserver un volume de travail suffisant pour l'ensemble des collaborateurs concernés (cette cause peut d'ailleurs rejoindre la première : dans un cadre professionnel où le salarié évolue dans un environnement de travail particulièrement contraint — espace confiné, longues astreintes — la réalisation de tâches en apparence répétitives permet de rythmer et occuper le temps de travail).

Ajoutons à cela que la dénomination même d'« activités à faible valeur ajoutée » est sujette à caution, tant certaines de ces activités participent à la constitution d'une expertise essentielle à certains métiers (comme la connaissance précise de certaines pièces, machines, process…).

Dans d'autres cas, l'utilisation de l'IA permet de réaliser des tâches jusqu'alors inaccessibles aux collaborateurs, comme l'analyse de grands jeux de données. Mais parvenir à mettre en place ces activités de rupture nécessite une forte collaboration entre les *data scientists* et les métiers. Or les conditions de ces collaborations sont parfois imposées ou bien insuffisamment calibrées.

A.4. L'externalité des équipes de *data scientists*

À la suite de ce dernier point, l'équipe de recherche a pu constater des niveaux d'intégration très inégaux des groupes de *data scientists* avec les métiers. Le terme de « cellules IA » décrit assez bien l'observation qui a pu être faite des *data scientists* au sein des entreprises : ceux-ci travaillent bien souvent en équipe autonome « au service » d'une pluralité de métiers. Cette organisation présente l'avantage de pouvoir multiplier les opportunités d'innovation. Toutefois, une meilleure intégration pourrait être une source de confiance dans les SIA.

L'externalité des cellules IA conduit en outre à laisser à la sensibilité des responsables hiérarchiques le soin d'encourager la collaboration des *data scientists* avec les métiers. Une telle dépendance est de nature à rompre les liens de confiance au sein des organisations.

A.5. Des enjeux juridiques à prendre en compte

Enfin, les enjeux juridiques et réglementaires liés à l'IA, et notamment les débats relatifs à la mise en place au niveau européen d'un Règlement sur l'intelligence artificielle, sont apparus comme largement inconnus des *data scientists* qui ont d'ailleurs souvent exprimé un besoin de formation en ce sens (cf. partie C).

B. Environnement et conditions de travail

B.1. Des conditions de travail jugées satisfaisantes

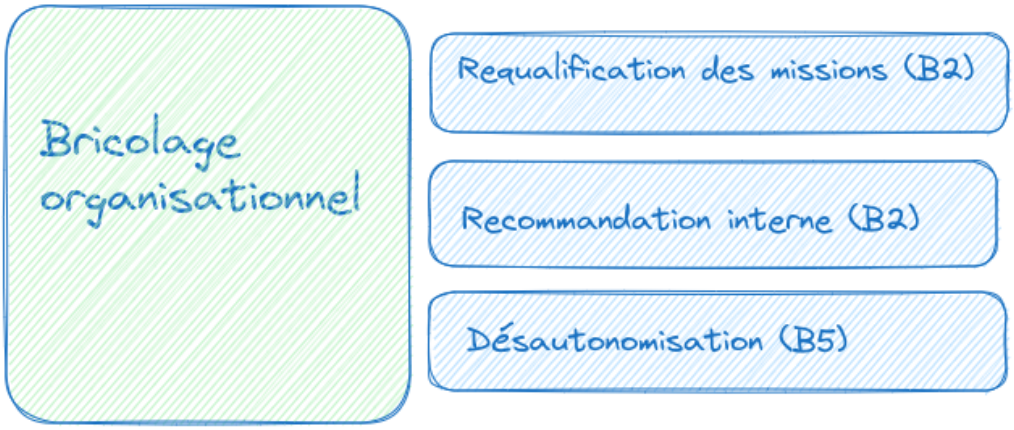
D'une manière générale, les *data scientists* ont exprimé une certaine satisfaction concernant leurs conditions de travail. Tous nos interlocuteurs ont en effet reconnu des conditions de travail plutôt favorables, aussi bien concernant les salaires, l'environnement de travail, mais également de rythme (la possibilité de télétravailler, notamment, a souvent été mentionnée comme un facteur positif). Les compétences en IA étant précieuses, les entreprises ont en effet réalisé des efforts notables pour attirer des profils spécialisés, et certaines n'hésitent pas à créer des primes pour encourager leurs salariés à débaucher des profils intéressants. Pour

autant, les entreprises françaises semblent rester « raisonnables » par rapport à des entreprises étrangères qui proposeraient, selon les *data scientists* interrogés, des salaires bien plus élevés.

Cet environnement de travail favorable se reflète dans l'expression des sentiments des *data scientists* envers l'organisation à laquelle ils appartiennent. Nous avons ainsi observé un attachement corporatif assez fort, accompagné d'un sentiment de fierté à travailler pour l'entreprise.

B.2. Un bricolage organisationnel

Le second point saillant porte sur l'articulation des missions des *data scientists* avec celles des métiers. Le sentiment général se dégageant est celui d'un bricolage organisationnel — loin d'être péjoratif, ce concept de « bricolage organisationnel » connaît un succès grandissant depuis une vingtaine d'années dans les sciences de la gestion (Baker & Nelson, 2005 ; Duymedjian & Rüling, 2010) où il désigne la capacité à innover et trouver des solutions en puisant dans les ressources déjà disponibles. Ce bricolage se manifeste généralement de trois manières :



- la première manifestation du bricolage organisationnel porte sur la requalification fréquente par les *data scientists* des missions qui leur sont confiées par le management.

Le périmètre des missions confiées aux *data scientists* fait ainsi généralement l'objet d'une redéfinition après leur attribution, soit que la mission ne puisse être menée à bon terme, soit que les instructions de départ soient insuffisamment précises. Le redécoupage *ex post* des missions s'explique en grande partie par la dimension novatrice et prospective des applications d'intelligence artificielle.

- la seconde manifestation du bricolage organisationnel se manifeste à travers le système de recommandation interne, et concerne l'opportunité des déploiements des systèmes d'IA dans les organisations.

Nous avons constaté que la nature des relations interpersonnelles, les recommandations internes et, parfois, le bouche-à-oreille entre les services sont souvent à l'impulsion des nouveaux projets impliquant les outils d'IA. La façon dont les *data scientists* parviennent à gérer leur notoriété au sein de l'entreprise et à assurer leur propre publicité auprès des autres services de l'entreprise apparaît donc essentielle. Inversement, le départ d'un collaborateur peut signifier l'abandon d'un projet IA.

- Enfin, le bricolage organisationnel se manifeste également à travers un travail de « désautomatisation » de l'IA, dont nous développons les caractéristiques ci-après (B.4).

B.3. Les processus décisionnels

Les enjeux liés à l'encadrement du processus décisionnel sont cardinaux dans le discours des *data scientists*. Pour eux, le maintien des personnes dans la boucle décisionnelle apparaît comme un outil de confiance essentiel au déploiement des outils d'IA. Ce constat n'a rien d'étonnant tant ce thème est central dans la littérature scientifique et dans les recommandations de politiques publiques.

Sur le plan des usages, nous avons pu constater un schéma récurrent, à savoir l'utilisation de l'IA à des fins de « vérification » d'une analyse humaine. L'objectif alors recherché consiste à écarter le risque de biais de confirmation en évitant que les personnes ne soient influencées ou même déstabilisées par les résultats de l'IA dans leurs prises de décisions (cf. B.5). Cette stratégie témoigne aussi d'une position de méfiance des utilisateurs de l'IA face à une technologie encore parfois peu mature.

B.4. La désautomatisation de la prise de décision

Plusieurs interlocuteurs nous ont expliqué opérer ce que nous proposons d'appeler une *désautomatisation* de l'IA, en déployant plusieurs stratégies pour invisibiliser celle-ci, ou plus exactement pour atténuer l'effet de désintermédiation qu'elle peut produire, en réintroduisant au niveau de l'IHM des éléments d'encapacitation (par exemple : un choix à réaliser par un opérateur entre plusieurs suggestions formulées par l'IA).

En effet, l'intégration d'un outil d'IA dans une chaîne de travail peut également se révéler un important facteur de déstabilisation au sein de certains métiers, comme nous avons pu le constater notamment dans des services assurant une mission de contrôle qualité. Le recours systématique à un outil IA (par exemple, un outil d'aide à la sanction qui va vérifier la conformité d'une pièce), lui-même placé en début de process, peut s'avérer contre-productif et chronophage dans la mesure où il finit par générer un manque de confiance de l'employé envers lui-même et ses propres compétences.

Conscients de cette difficulté, les *data scientists* travaillent à la conception d'outils les moins intrusifs possibles, et imaginent des IHM qui atténuent l'effet de désintermédiation pour

redonner des responsabilités à l'opérateur. Dans certains cas, l'IA restera muette et n'interviendra qu'en cas d'écart de diagnostic, à des fins de réajustement éventuel. Elle s'apparente ainsi à un filet de sécurité relativement peu intrusif, qui peut d'ailleurs rassurer les opérateurs dans leur travail.

B.5. L'objectif d'acceptabilité

Le panel de *data scientists* interrogé est apparu très soucieux de favoriser l'acceptabilité de leurs outils auprès des métiers. Certaines entreprises ont ainsi mis en place des processus de co-construction des outils avec les services concernés, en travaillant selon une méthode agile permettant aux métiers concernés de tester et d'émettre des recommandations.

Comme mentionné plus tôt, les *data scientists* ont également noté l'importance de ménager les métiers en utilisant l'IA à des fins d'encapacitation plutôt que de désintermédiation. Nous avons noté dans plusieurs entretiens le recours à des analogies avec des outils numériques grand public, en particulier des outils d'assistance désormais perçus comme non-intrusifs : les modèles du correcteur orthographique de Word ou encore des suggestions du moteur de recherche ont ainsi été cités (ce dernier exemple, en particulier, a même servi d'inspiration au développement de l'IHM d'un outil). À chaque fois, il s'agit de favoriser l'encapacitation des salariés, qui (re) deviennent maîtres de la décision finale en opérant un choix parmi les suggestions de l'IA.

C. Les attentes des *data scientists*

C.1 Clarifier la stratégie de l'organisation

La première des attentes perçues portait sur la clarification de la ou des stratégies IA déployées dans les organisations. Dans certains cas, ces stratégies étaient perçues comme insuffisamment claires. Cet état d'incertitude représente une source d'insécurité pour des collaborateurs soumis à un environnement très évolutif comme l'intelligence artificielle. Au contraire, la détermination d'une stratégie claire en matière d'IA est perçue comme un facteur structurant pour les *data scientists*.

Concernant les métiers, il conviendrait de vérifier empiriquement cette analyse. Mais relevons que d'une manière générale, l'utilisation d'un outil technique suppose le plus souvent sa bonne compréhension.

Proposer une réflexion stratégique sur l'IA semble dès lors nécessaire. Plusieurs grandes entreprises françaises (e.g. Renault, L'Oréal) ont d'ailleurs communiqué sur leur stratégie de se présenter comme des entreprises « Tech ». Au-delà de l'objectif de communication publique, placer la technique, et notamment l'IA au centre de la stratégie d'entreprise semble de nature à engendrer la confiance des partenaires. Les objectifs stratégiques du déploiement de l'IA, notamment, doivent être clarifiés : s'agit-il d'améliorer la qualité du travail,

d'améliorer la productivité, de mieux évaluer celle-ci ? Et si l'IA permet de gagner du temps, à quelles fins ce gain est-il utilisé au sein de l'entreprise : est-ce pour réduire les effectifs ou bien améliorer les conditions de travail ? Ces questions relèvent du management de l'entreprise, qui ont tout intérêt à définir clairement les objectifs de l'IA en termes de condition de travail, de manière à orienter les travaux des cellules IA, mais également de faciliter l'intégration de ces dernières auprès des métiers.

C.2. Collaborer avec les métiers

La seconde attente des *data scientists* porte sur le rapprochement avec les métiers. Un tel rapprochement permet une meilleure communication entre les acteurs en présence. Cette situation se révèle très avantageuse dans un contexte d'incertitude technologique. Déterminer les applications de l'IA relève en effet encore du tâtonnement : certaines applications en apparence évidentes se révèlent complexes à mettre en œuvre ou bien inefficaces ; d'autres applications n'apparaissent qu'à force d'essais répétés et de transformations.

Certaines organisations ont bien perçu ces difficultés en instaurant un dialogue institutionnel entre les *data scientists* et les métiers. Au contraire, en l'absence d'une telle organisation, les *data scientists* peuvent éprouver une forme d'isolement.

C.3. Se former aux enjeux éthiques et juridiques

Une autre attente forte porte sur la réflexion éthique et juridique autour de l'IA. Ces enjeux occupent en effet assez largement le débat public (surtout depuis la publication de ChatGPT par Open AI). En particulier, la transparence et la gouvernance des données figure au centre de l'attention.

C.4. Clarifier et formaliser les outils de la confiance

Les *data scientists* ont également exprimé la nécessité de formaliser leurs propres modèles et process à des fins de communication auprès des autres métiers. En d'autres termes, il s'agit pour eux d'obtenir ou de développer des outils capables de susciter la confiance auprès de leurs collègues : des éléments de type *model card*, par exemple, ou tout autre système de documentation permettant d'expliquer la conception, le développement et la fonction d'un outil, s'avèrent précieux pour donner plus de transparence aux outils, et générer plus de confiance. Pour être efficace, ces éléments de formalisation doivent s'adresser en priorité aux métiers : il s'agit d'outils destiner à fluidifier la communication entre ces derniers et les cellules IA.

Nous avons par ailleurs observé que plusieurs équipes avaient lancé de leur propre initiative des projets destinés à améliorer des services ou des process sans avoir toujours été sollicitées : dans ces cas précis, elles devaient ainsi, en plus de leur travail de conception de l'outil, convaincre leurs collègues de l'utilité de cet outil. Plusieurs équipes nous ont ainsi confié que

certains projets avaient finalement dû être abandonnés faute de pouvoir être adoptés par les services concernés. L'abandon de nombreux projet finit par créer des frustrations au sein des équipes.

C.5. Encourager l'innovation

Tous nos entretiens se sont déroulés au sein d'équipes motivées, qui ont démontré une forte volonté d'innover, notamment pour améliorer le travail de leurs collègues (via l'automatisation de certaines tâches, par exemple). Les cellules IA sont cependant bien souvent sous-dimensionnées et dans la plupart des entreprises, elles ne semblent pas répondre à un plan stratégique précis.

Leur formation de pointe explique très certainement leur engouement pour l'innovation dans un domaine qui a le vent en poupe dans les laboratoires de recherche. Mais cette appétence pour la recherche, et les pratiques qui en découlent — principalement de l'expérimentation, des preuves de concept dont la généricisation ou le développement à grande échelle n'est pas toujours assuré — peut sembler en décalage par rapport aux besoins de l'entreprise et de ses salariés.

Les entreprises, en ce sens, ont vocation à devenir à la fois des observatoires et de laboratoires du développement d'outils d'IA, et à renforcer leur politique de recherche et développement. Ces chefs de projet constituent un véritable atout pour l'entreprise, et pourraient être encouragés à conserver un pied dans le milieu académique (universités et grandes écoles), sous la forme de participation à des congrès et colloques, ou même sous la forme de vacations d'enseignement.

C.6. Généricizer et déployer les outils à plus grande échelle

Nous avons pu noter une forte demande en termes de générécisation, voire d'industrialisation, des outils développés par les cellules IA, qui semblent en pleine recherche de maturité. Dans la majeure partie des cas que nous avons pu observer, les projets réalisés par les cellules IA relèvent principalement de la preuve de concept ou restent à l'état de test. On note ainsi un décalage entre l'ambition d'un outil, les moyens humains et techniques mis en œuvre pour sa conception et son développement, ainsi que son déploiement puis usage effectif au sein de l'entreprise. La faible ampleur donnée à un projet auprès des métiers concernés, voire l'abandon de certains outils, a été évoquée dans plusieurs entretiens comme une source de frustration et de découragement au sein des équipes IA.

Cette volonté se trouve généralement renforcée lorsque les données nécessaires à la conception de systèmes d'intelligence artificielle sont disponibles (mais pas toujours utilisées), ce qui semble être assez souvent le cas. Ce constat mérite ici d'être souligné, car il s'oppose à l'idée répandue selon laquelle les *data scientists* manqueraient de données pour concevoir des systèmes d'intelligence artificielle. Bien que les situations soient nuancées, la

libération du potentiel de l’intelligence artificielle semble davantage relever d’enjeux de volonté que de disponibilité des données.

IV. Recommandations : la méthode FIIDES

FIIDES. Les recommandations du groupe de recherche visent à encourager la confiance des *data scientists*/ingénieurs en intelligence artificielle et, par capillarité, celle de l'ensemble des collaborateurs d'une organisation qui s'engage dans le développement d'outils d'intelligence artificielle. Ces recommandations consistent à favoriser la Formation, l'Intégration, l'Information, le Développement et, enfin, la Soutenabilité des usages de l'IA. **L'acronyme FIIDES synthétise ces propositions.**

A. Formation

A.1. Former les collaborateurs à l'environnement technique, éthique et réglementaire de l'IA

Les *data scientists* rencontrés ont largement exprimé leur volonté d'acquérir des connaissances techniques concernant l'IA (la plupart suivent avec attention l'actualité scientifique de l'IA), mais aussi éthique et réglementaire. Cette volonté s'inscrit assez logiquement dans un contexte général où les enjeux éthiques relatifs à l'intelligence artificielle font l'objet d'un intense débat public. Cette demande s'inscrit aussi dans la perspective d'un probable encadrement juridique contraignant.

Au-delà, une part importante des missions prises en charge par les *data scientists* consiste à expliquer aux collaborateurs (métiers, management, partenaires, etc.) le fonctionnement des systèmes d'IA proposés. Dès lors, le besoin de formation ne concerne pas seulement les *data scientists*, mais plus largement l'ensemble des collaborateurs en qualité d'utilisateurs des systèmes d'intelligence artificielle.

Les objectifs de connaissance et de formation représentent en effet des briques essentielles de la confiance. Ils correspondent aussi aux **Recommandations du Conseil sur l'intelligence artificielle** adoptées par l'OCDE le 22 mai 20191. Ces principes prévoient en effet de « **renforcer les capacités humaines et préparer la transformation du marché du travail** ». Pour ce faire, l'OCDE recommande aux pouvoirs publics de « travailler en étroite collaboration avec les parties prenantes en vue de préparer la transformation du monde du travail et de la société » en « dotant [les personnes] des compétences nécessaires » et en mettant en place des « **programmes de formation tout au long de la vie active, du soutien aux personnes affectées par les suppressions de postes et de l'accès aux nouvelles opportunités sur le marché du travail** ».

Le besoin de formation porte notamment sur l'environnement juridique en préparation au niveau de l'Union européenne. La proposition de Règlement relatif à l'intelligence artificielle, publiée par la Commission européenne le 21 avril 2021 (actuellement en cours de discussion

en trilogue) reste aujourd'hui plutôt méconnue. Or l'entrée en vigueur de ce texte à court terme risque de bouleverser le travail des data scientists et plus largement des utilisateurs de systèmes d'IA.

En particulier, les obligations en matière de gouvernance des données et d'explicabilité se trouvent au centre de la future réglementation de l'IA. L'anticipation de la mise en œuvre de ces obligations et, en tout état de cause, leur application selon le calendrier qui sera établi par le droit de l'UE, implique la mise en place de plans de formation.

Dans le même sens, la future réglementation repose sur une logique de certification et de marquage CE pour un certain nombre de systèmes d'IA (les systèmes d'IA dits à « haut risque »). Les formations devraient inclure l'exposé de ces travaux de normalisation, en particulier ceux en cours de discussion au CEN-CENELEC.

A.2. Recruter un ou une « Responsable des Systèmes d'IA » (RSIA) sur le modèle des DPO

L'environnement éthique et juridique en construction impose la mise en place d'une gouvernance interne des systèmes d'IA. Nommer des « responsables des systèmes d'IA », à la manière des Délégués à la Protection des Données personnelles, semble ainsi impératif.

Sans qu'une telle nomination ne soit rendue obligatoire, en l'état du droit positif ni de la proposition de Règlement européen relatif à l'IA, la complexité technique, éthique et juridique des cadres réglementaires à venir impose la création de ces nouveaux postes.

Une telle personne est susceptible d'être rattachée aux départements juridiques, compliance ou conformité (selon les modes d'organisation). Il reste que ce poste nécessite de solides compétences techniques en IA ainsi qu'une bonne culture juridique et réglementaire.

Cette personne serait notamment en charge de réaliser une cartographie de risques liés à la mise en œuvre des systèmes d'IA, assurerait la bonne gouvernance des données, prendrait en charge la procédure de marquage CE lorsqu'elle est requise, centraliserait les notifications d'incidents, assurerait le dialogue avec le comité d'éthique, se chargerait de l'application de la charte d'éthique, etc.

Du point de vue des pouvoirs publics, la structuration de cette nouvelle profession de responsables des systèmes d'IA pourrait emprunter la voie suivie par les délégués à la protection des données lors de l'entrée en vigueur du RGPD : encourager les formations universitaires et supérieures de niveau master, créer une entrée dans le Répertoire national des certifications professionnelles (RNCP) ou encore encourager le dialogue avec les autorités de régulation compétentes.

Une discussion concernant encadrement juridique de ces nouvelles professions, notamment en termes de responsabilité et de protection, mérite aussi d'être ouverte. Ces personnes ne devraient en effet pas être déclarées personnellement responsables des dysfonctionnements

des systèmes d'IA. Elles ne devraient pas non plus craindre les représailles de son employeur en cas de signalement des dysfonctionnements. Elles pourraient à cet égard bénéficier d'une protection équivalente à celle des lanceurs d'alertes.

B. Intégration

La seconde proposition consiste à favoriser l'intégration des *data scientists* auprès des métiers. L'objectif est d'encourager l'innovation en favoriser les échanges entre *data scientists* et personnes en charge de l'opérationnel, mais aussi des métiers supports. Plusieurs actions doivent permettre d'y parvenir. L'objectif ici visé consister à Intégrer l'IA dans la culture d'entreprise.

B.1. Recruter une personne « chargée de la transformation vers l'IA »

Que les missions confiées soient strictement déterminées ou non, il est apparu que les collaborateurs rencontrés dans le cadre de cette étude ne sont pas, ou faiblement, intégrés aux métiers. Or cette situation peut conduire à limiter l'innovation.

Sur le modèle des chargés de transformation numérique des organisations, nommer une personne en charge de la transformation vers l'IA semble indispensable. Compte tenu des enjeux économiques, techniques et stratégiques, l'acculturation à l'IA implique en effet une réflexion globale à l'échelle des organisations.

Ces personnes seraient chargées de faire le lien entre les *data scientists* au sein d'une même organisation, mais aussi d'assurer la diffusion des outils d'IA auprès des différents métiers. Ces personnes pourraient être recrutées en interne ou bien auprès d'organisations tierces.

Les missions confiées aux concepteurs de SIA sont plus ou moins strictement déterminées. Dans certains cas, un objectif déterminé est fixé, mais sans savoir si les outils techniques mobilisés seront de nature à l'atteindre. Cette stratégie repose sur le constat des évolutions rapides de l'IA et de l'incertitude inhérente à une technologique nouvelle à pouvoir remplir des objectifs innovants. Au contraire, dans d'autres cas, les concepteurs de SIA sont perçus comme des « chercheurs », très autonomes, dont la mission est de développer toutes sortes d'outils utiles pour les différentes fonctions opérationnelles du groupe.

B.2. Mettre en place un comité éthique de l'IA et rédaction de chartes internes

La mise en place d'un comité d'éthique de l'IA et la rédaction de charte sont susceptibles de structurer la stratégie d'entreprise en matière d'IA. Le recours aux comités (que l'on nomme parfois un peu péjorativement la « comitologie ») structure de nombreuses actions où les enjeux éthiques sont prégnants (les questions bioéthiques, la lutte contre les discriminations, la rémunération des dirigeants, etc.).

Les comités d'éthique de l'IA devraient être indépendants et composés de professionnels et de chercheuses et de chercheurs reconnus. Leurs missions consisteraient à évaluer les enjeux éthiques des systèmes d'IA, mais aussi à proposer un cadre stratégique à destination de tous les collaborateurs.

En plus d'une telle mission prospective, le comité pourrait prendre en charge les questions éthiques difficiles lorsqu'elles se posent (comme le recours à l'intelligence artificielle dans les processus de recrutement et de contrôle des salariés).

B.3. Reconnaissance de droits de co-détermination

L'intégration recommandée des *data scientists* vers les unités opérationnelles peut aussi prendre la forme de « droits de co-détermination » exercés par les salariés.

Une telle proposition est par exemple avancée par la Confédération allemande des syndicats (DGB) dans le rapport *Artificial Intelligence for Good Work* en 2020[7].

Une telle solution favoriserait l'exercice de la démocratie en entreprise et permettrait d'asseoir la légitimité des orientations stratégiques en matière d'intelligence artificielle.

C. Information

C.1. Mettre en place une communication interne

La mise en place d'une communication interne dédiée à l'intelligence artificielle semble impérative en raison des mutations importantes qu'engendre cette technologie sur l'emploi et les compétences. Des actions de communication (lettres internes, séminaires de réflexion, formations, etc.) permettraient de lever des craintes, parfois infondées, relatives à l'intelligence artificielle.

Ces actions paraissent d'autant plus nécessaires que l'intelligence artificielle n'est pas seulement une technologie nouvelle et prometteuse, mais aussi un véritable sujet de débat public. Or les passions entourant le débat public ne reflètent pas nécessairement l'état des techniques. Le domaine des véhicules autonomes semble particulièrement frappant à ce titre. Les craintes et les espoirs liés à la mise en circulation de véhicules autonomes ne correspondent en effet pas à la réalité des techniques actuelles.

Des actions de communication interne permettraient en conséquence d'encourager les relations entre les ingénieurs en intelligence artificielle et les métiers en permettant une égalité (relative) de connaissances technologiques. Ces actions apparaissent d'autant plus importantes que les technologiques d'intelligence artificielle sont généralement affectées par l'effet « boite noire ».

[7] Cf. en ligne [https://oecd.ai/en/catalogue/tools/artificial-intelligence-ai-for-good-work].

C.2. Adapter les interfaces graphiques des systèmes d'IA

L'analyse des interfaces graphiques consistait en la mission originelle de ce groupe de travail. Or, comme il a été précisé plus haut, un tel travail s'est révélé impossible. Il est cependant apparu au cours des entretiens que l'apparence des interfaces graphiques occupait une place importante dans le déploiement des systèmes d'intelligence artificielle.,

L'association de *designers graphiques,* en particulier d'*UX designers,* semble à ce titre particulièrement prometteuse. La lutte contre les *dark pattern* et l'encouragement aux bonnes pratiques graphiques sont essentiels au déploiement de systèmes d'intelligence artificielle de confiance.

C.3. Anticiper les futures obligations d'information

Les actions de communication interne et de création d'interfaces graphiques de confiance devraient enfin être déployées dans le respect des futures obligations d'information mises à la charge des producteurs de systèmes d'intelligence artificielle.

La proposition de Règlement européen relatif à l'intelligence artificielle contient en effet plusieurs obligations d'information de nature distincte :

— envers les utilisateurs des systèmes d'intelligence artificielle ;
— envers les auditeurs lorsque telles opérations sont prévues ;
— envers l'autorité régulatrice de l'intelligence artificielle.

Dans tous les cas il convient d'adapter les modalités de communication à ses destinataires. En outre, ces obligations d'information devraient être prises en compte dès la conception des systèmes d'intelligence artificielle, sans attendre le déploiement d'une législation contraignante.

D. Développement

D.1. Mettre en place des outils d'audit en IA et préparer l'entrée en vigueur de l'IA Act

Le développement des systèmes d'intelligence artificielle de confiance implique la mise en place de procédures de la part des producteurs et utilisateurs de systèmes d'intelligence artificielle.

Le recours à l'audit externe, notamment pour obtenir des labels de qualité, peut constituer un outil pertinent pour encourager la confiance des utilisateurs. À cette fin, les pouvoirs publics pourraient encourager le développement de telles initiatives. Dans tous les cas, les

producteurs de systèmes d'intelligence artificielle doivent réaliser une veille des labels de qualité proposés.

Il existe aujourd'hui de nombreuses entreprises qui proposent des outils d'audit de compliance pour l'IA. Il reste toutefois encore assez difficile d'accorder un juste crédit à ces solutions. Sans porter de jugement sur la qualité de leurs prestations, il reste que la législation demeure aujourd'hui en pleine construction, ce qui rend les opérations de compliance naturellement incertaines.

Plus globalement, les producteurs de systèmes d'intelligence artificielle devraient anticiper la mise en place de la Réglementation européenne par la formation des ingénieurs en intelligence artificielle et/ou des personnes concernées.

D.2. Promouvoir les initiatives normatives privées

L'écosystème normatif de l'intelligence artificielle se trouve aujourd'hui dans une situation d'intense productivité. Un certain nombre de référentiels et de labels ont en effet déjà fait l'objet de publications. L'objectif d'encourager la confiance se trouve en règle générale au centre de ces outils normatifs privés. A cet égard, il faut souligner que le rapport des experts de haut-niveau de l'Union européenne sur « l'IA de confiance » a profondément marqué les esprits.

L'association « Label IA » propose ainsi un référentiel, élaboré en 2019 de manière participative, visant une « data science responsable et de confiance[8] ». Six thèmes sont proposés à l'évaluation : les données personnelles, les biais, la performance, la reproductibilité et la chaîne de responsabilité, les modèles de confiance et les externalités négatives. L'évaluation proposée consiste à répondre à un série de questions concernant chaque thème afin d'obtenir un score et identifier ses zones de risques.

Une autre association, Positive AI, propose également un référentiel destiné à encourager l'IA de confiance. Ce label a été réalisé par des *data scientists* issus de plusieurs grandes entreprises (BCG Gamma, L'Oréal, Malakoff, Humanis et Orange France). Il vise en particulier à évaluer le degré de maturité des organisations concernant leur gouvernance, la gestion des systèmes d'intelligence artificielle et les algorithmes. Les points analysés portent sur la justice et l'équité, la transparence et l'explicabilité et, enfin, la supervision des décisions automatisées.

En partenariat avec Orange France, l'association Arborus propose, outre le label GEEIS IA[9] portant quant à lui sur le genre, l'équité et la diversité, la signature d'une Charte internationale pour une Intelligence Artificielle Inclusive. Elle porte sur la mixité et la diversité des équipes de *data scientists*, les biais de discrimination, la qualité des données, la formation des *data*

[8] Cf. en ligne [https://github.com/LabeliaLabs/referentiel-evaluation-dsrc/blob/master/referentiel_evaluation.md].
[9] Cf. en ligne [https://arborus.org/label/].

scientists et plus largement des utilisateurs et donneurs d'ordre et, enfin, la maîtrise en continue de la chaîne de valeur.

Leur degré de pénétration auprès des data scientists n'est pas pu être évalué dans le cadre de cette étude. Ces initiatives devraient en tout cas être encouragées et, le cas échéant, adaptés à la future réglementation européenne.

D.3. Mettre en place une gouvernance des données d'entraînement

La gouvernance des données d'entraînement figure au centre des objectifs visés par les encadrements éthiques et juridiques de l'intelligence artificielle.

La bonne gestion de ces données se trouve en effet à la base des politiques de risques et de prévention des biais dans les systèmes d'intelligence artificielle. Le lien entre les biais présents dans les données d'entraînement et ceux apparaissant dans les résultats proposés est en effet incontestable. Les exemples de discrimination (involontaire) à l'encontre de minorités, imputable aux biais présents dans ces jeux de données nourrissent ainsi l'actualité.

Outre la maîtrise de ces biais, la bonne gouvernance des données permet aussi de contrôler la légalité des données utilisées pour l'entraînement. Les données sont en effet couvertes par un certain nombre de dispositifs juridiques contraignants, dont le droit des données à caractère personnel, les droits de propriété intellectuelle et des secrets et les droits contractuels. Une mauvaise maîtrise du cadre juridique des données d'entraînement risque de compromettre la légalité de l'ensemble du système d'intelligence artificielle. Il est donc impératif que les données d'entraînement soient maîtrisées.

Les ingénieurs en intelligence artificielle recourent de plus, de manière semble-t-il très fréquente, à des jeux de données pré-entraînées, librement disponibles. Ils affinent ensuite l'entraînement afin de parvenir à leurs particuliers particuliers (on parle souvent de « fine-tuning » pour désigner cette opération). Or, si elle permet un gain de temps considérable, voire constitue dans certains cas la condition essentielle à la mise en place de systèmes d'intelligence artificielle, la pratique du fine-tuning soulève néanmoins d'importants enjeux de maîtrise des données d'entraînement.

E. Soutenabilité

E.1. Les considérations environnementales

La préoccupation quant aux conséquences environnementales du numérique occupe une place grandissante dans le débat public et dans les politiques publiques. Le gouvernement a ainsi présenté le 4 juillet 2023 une feuille de route pour la « décarbonation du numérique ». Elle s'inscrit à la suite d'une série d'initiatives législatives adoptées depuis 2020 visant à réduire l'empreinte environnementale des équipements numériques.

Or les systèmes d'intelligence artificielle engendrent des dépenses énergétiques, que ce soit pour l'entraînement des données ou pour l'utilisation. S'il reste aujourd'hui difficile de les calculer, il apparaît fondamental de prendre en compte cette dimension dans l'élaboration des futures politiques publiques et donc, dans la déploiement de l'intelligence artificielle de confiance.

L'expérience des cryptoactifs doit servir de leçon pour l'écosystème de l'intelligence artificielle. Perçus à leurs origines comme purement « immatériels », les cryptoactifs se sont révélés pour certains (ceux fondés sur une preuve de travail nécessitant des calculs complexes) comme de véritables gageurs énergétiques.

De telles prises de conscience d'un coût environnemental trop élevé risquent de ruiner les investissements et les espoirs placés dans de nouvelles technologies. Afin d'éviter de tels effets, il est nécessaire d'intégrer dans les cartographies des risques ainsi que dans les labels et recommandations la prise en compte de l'impact carbone des techniques d'intelligence artificielle.

E.2. La préservation des savoir-faire

Encourager la soutenabilité des systèmes d'intelligence artificielle implique encore de préserver les savoir-faire menacés par le déploiement de cette technologie. Ce risque concerne les compétences susceptibles d'automatisation totale ou partielle, comme en matière d'analyse médicale et de contrôle non destructif par exemple.

Le principal risque identifié porte sur l'appauvrissement des compétences liées à la diminution de l'attractivité d'un emploi jugé peu dynamique, car partiellement automatisé. Un exemple — qui n'est pas tiré de l'étude réalisée dans le cadre de confiance.AI) permet de prendre la mesure du phénomène : celui des médecins-radiologues. L'intelligence artificielle pourrait en effet améliorer la productivité de ces médecins concernant les tâches d'analyse de résultats. Dès lors, le nombre de médecins-radiologues à former serait plus faible, ce qui risque d'affaiblir cette spécialité, en réduisant la recherche et la formation. *In fine*, les compétences de radiologues risquent de s'appauvrir ou, à tout le moins, de croître moins rapidement.

E.3. Le bien-être au travail et l'organisation du temps de travail

Le temps gagné grâce à l'automatisation de certaines tâches, notamment des tâches répétitives, n'a pas forcément vocation à supprimer des emplois, mais peut aussi donner l'occasion de repenser l'organisation du travail et notamment du temps de travail. À l'inverse, comme nous l'avons déjà relevé dans les conclusions de ce rapport, certaines tâches jugées répétitives ou à faible valeur ajoutée sont utiles, car elles permettent d'équilibrer le travail d'un salarié en instituant des « pauses » dans la journée ou dans la semaine.

Ainsi, il semble nécessaire de proposer une réflexion collective sur la façon dont l'IA peut accompagner des stratégies de développement du bien-être au travail — sans menacer l'équilibre des salariés.

E.4. La protection de la souveraineté numérique

Enfin, le déploiement de systèmes d'intelligence artificielle pourrait entraîner une perte de souveraineté numérique de la France. Cette perte, liée à celle des savoir-faire, est d'abord susceptible de résulter de la délocalisation d'unités de production hors de France pour des motifs de coût du travail. L'automatisation de certaines tâches est en effet susceptible de produire un tel effet par le recours à de la main-d'œuvre située hors du territoire. Cette situation pourrait être accentuée par le recours au télétravail où la présence physique du collaborateur est jugée accessoire.

En outre, la localisation des données hors du territoire national est aussi susceptible de porter atteinte à la souveraineté nationale. Les enjeux en matière d'intelligence artificielle ne sont pas distincts de ceux rencontrés en matière de données en général (les défis rencontrés par la plateforme des données de santé constituent à cet égard des précédents importants).

V. Actions proposées à la suite du programme de recherche

A. La création d'une formation de « Responsable des systèmes d'IA » à Sorbonne Université

À l'initiative d'Arnaud Latil, et avec le soutien du Sorbonne Center for Artificiel Intelligence (SCAI), Sorbonne Université propose à partir de septembre 2024 une nouvelle formation, destinée aux professionnels et accessible en formation continue, visant à former des « responsables de systèmes d'IA ».

Ces « DPO de l'IA » seront en charge d'assurer la conformité réglementaire des SIA dans un contexte de forte évolution de la réglementation.

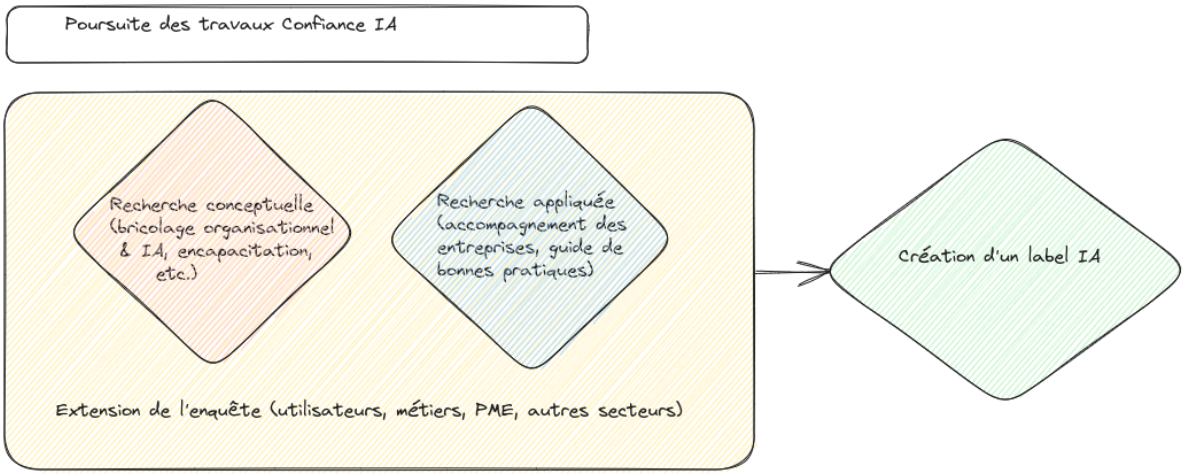
Cette nouvelle offre de formation découle directement des travaux menés dans le cadre de confiance.ai et implique des partenaires rencontrés dans le cadre du programme Confiance.AI, mais aussi des acteurs institutionnels de premier plan participent à cette formation, comme Monsieur Félicien Vallet, « chef de l'IA » à la CNIL et Monsieur Louis Morilhat pour le compte de l'AFNOR

B. Poursuite des recherches sur l'encouragement de la confiance dans les organisations

Dans l'objectif d'encourager le développement des SIA dans les organisations, les chercheurs et chercheuse du groupe de recherche proposent de poursuivre leurs travaux à travers les actions suivantes, dans le respect des orientations stratégiques du programme confiance. AI :

- L'extension du champ de l'enquête aux utilisateurs des systèmes d'intelligence artificielle. Une telle étude suppose une nouvelle définition du périmètre des personnes associées au programme confiance.ai
- Actions pratiques : La rédaction d'un guide de bonnes pratiques destinées à encourager le développement des systèmes d'intelligence artificielle
- Un accompagnement à la mise en place de la réglementation relative à l'intelligence artificielle, en particulier concernant le futur Règlement européen.
- La création d'un label « IA de confiance »

Ces travaux pourraient être menés de concert avec d'autres groupes de recherche du périmètre Confiance.AI.



C. Thèmes émergents pour les études sur l'intelligence artificielle par les SHS

Deux thèmes majeurs de recherche ont émergé grâce aux enquêtes menées dans le cadre de confiance.AI.

Le premier porte sur les modes d'organisation des *data scientists* et l'articulation de leurs actions avec les autres services de l'entreprise. Des études pluridisciplinaires, associant des sociologues, des spécialistes de sciences de gestion et des juristes, pourraient permettre de proposer des méthodes performantes visant à encourager le déploiement de l'intelligence artificielle de confiance dans les organisations.

Le second thème de recherche porte sur l'encapacitation des utilisateurs de systèmes d'intelligence artificielle. Parvenir à cet objectif est susceptible d'emprunter plusieurs voies : information, design, itération et tests, décompilation et retro-engineering, etc. De telles études impliquent aussi la collaboration de recherche issue de différentes disciplines, et notamment des designers, ergonomes, informaticiens et *data scientists*.

Éléments de bibliographie

- Boyer Bertrand, « Le soldat et les nouvelles technologies : la confiance a priori », *Inflexions* 2022/3, n° 51, pp. 79 à 84).

- Caby F., Louise V. et Rolland S., *La qualité au XXIe siècle. Vers le management de la confiance*, Economica, 2002

- CNIL, *Comment permettre à l'homme de garder la main ?*, 2017

- Doueihi Milad et Domenicucci Jacopo (dir)., *La confiance à l'ère numérique*, Editions rue d'Ulm, 2018

- Ellul Jacques, *Le système technicien*, préf. Jean-Luc Porquet, Cherche-midi, première ed., 1977, réed. 2012

- Fukuyama Francis, Tr*ust : The Social Virtues and the Creation of Prosperity*, Free Press, New York, 1995
- Laurent Eloi, *Economie de la confiance*, La Découverte, 2019

- Luhmann Niklas, *La confiance. Un mécanisme de réduction de la complexité sociale*, Economica, 2006

- OCDE , *Guidelines on Measuring Trust*, 2017

- Pasquale Franck, *Black box Society. Les algorithmes secrets qui contrôlent l'économie et l'information*, FYP Editions, 2015

- Pasquale Franck, *New laws of robotics. Defending human expertise in the age of IA*, The belknap press of Harvard University Press, 2020

- Senik Claudia (dir.), *Crises de confiance ?*, La Découverte, 2020

- Thaler Richard et Sunstein Cass, *Nudge. Improving Decisions About health, Wealth and Happiness*, Yale University Press, 2008

- Trouchaud Philippe, *La cybersécurité face au défi de la confiance*, préf. Pascal Andrei, O. Jacob, 2018

- Union européenne, *Lignes directrices en matière d'éthique pour une IA digne de confiance*, 2018 (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai)

- Verrier, Gilles, et Nicolas Bourgeois. F*aut-il libérer l'entreprise ?Confiance, responsabilité et autonomie au travail.* Dunod, 2016 (disponible sur cairn.info)

- Zuboff Shoshana, *L'âge du capitalisme de surveillance*, Zulma, 2020