JOURNAL DES ANALYSES STRATÉGIQUES

Pratiques de sécurité à l'ère de la transition numérique

novembre 2024

FONCSI
Fondation pour une culture de sécurité industrielle

OÙ EN EST-ON ?

Les premières réunions du groupe ont permis d'évoquer les attentes des participants en termes thématiques, et de définir le périmètre prévu pour les travaux, à l'intersection des nouvelles technologies (IA, big data), des FOH et des enjeux de sécurité.

LES MESSAGES CLÉS DE CE NUMÉRO

01 L'explicabilité et l'intelligibilité des composants logiciels intégrés sous forme de « boîte noire » est une problématique multidisciplinaire qui soulève d'importants enjeux en termes de confiance, de démonstration de sécurité, et d'acceptabilité sociale.

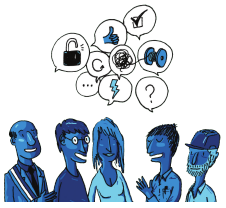
02 Mieux comprendre la manière dont les outils numériques viennent conforter, ou au contraire percuter, les pratiques et savoir-faire déjà bien établis dans différents contextes professionnels à fort enjeux de sécurité, permettra de mieux accompagner les différences d'approche souvent constatées selon les cultures métier.

03 Les jumeaux numériques pourraient permettre de reconcevoir la manière de penser les démonstrations de sécurité, pour passer d'une approche statique à la conception vers une approche plus dynamique centrée sur la capacité à mesurer la distance au danger et à piloter le système vers des zones moins dangereuses.

Vous avez dit « analyse stratégique » ?

Une analyse stratégique est un outil utilisé par la Foncsi pour étudier en profondeur une question sur une durée de 18 mois, en installant un continuum de l'innovation entre recherche et industrie. Elle est conduite par un groupe composé de représentants des mécènes et partenaires qui accompagnent l'analyse stratégique, d'experts académiques et de représentants de la Foncsi.

Cette analyse concerne les « *pratiques de sécurité à l'ère de la transition numérique* ». Elle vise à comprendre les enjeux et les opportunités que présentent le développement rapide des données massives et de l'intelligence artificielle basée sur l'apprentissage profond, pour les pratiques des professionnels (autant ceux pour lesquels la sécurité est le métier, que ceux pour lesquels elle constitue un enjeu parmi d'autres).



L'analyse stratégique sur les « *Pratiques de sécurité à l'ère de la transition numérique* » est l'une des trois analyses stratégiques de la première partie du programme scientifique « Foncsi 4 ». Les deux autres analyses lancées sur la même période sont « *Compétences et carrières à l'horizon 2040* » et « *Les nouveaux champions de la performance* ».

Une révolution en cours

Le développement rapide des technologies numériques liées à la collecte de quantités massives de données et leur traitement par des algorithmes de *machine learning*, produit une révolution dans les capacités de prévision, de pilotage et de surveillance de la sécurité industrielle, dans un monde devenu plus complexe, plus fragmenté, plus incertain et plus interdépendant.

Le thème de l'IA fait l'objet d'un engouement remarquable depuis cinq ans : les projets de recherche et développement nationaux et européens, les projets d'innovation dans les entreprises, et les initiatives liées aux enjeux de certification, foisonnent.

Confiance ai

DEEL
DEpendable & Explainable Learning

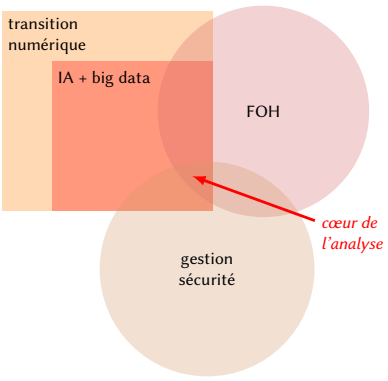
ISO/IEC TR 24028:2020
Information technology
Artificial intelligence
Overview of trustworthiness in artificial intelligence

INSTITUTE DATAIA
Data Science, Intelligence & Society

ANITI Université Fédérale Toulouse Midi-Pyrénées

NUMÉRIQUE LNE CERTIFICATION Intelligence artificielle

Ces travaux portent principalement sur les dimensions technologiques ou techniques du numérique et de l'intelligence artificielle. L'expertise et le réseau de la Foncsi étant davantage centrés sur les enjeux collectifs, organisationnels et inter-organisationnels de la sécurité industrielle que sur ses dimensions technologiques, nous avons choisi d'aborder l'analyse des pratiques et enjeux de sécurité liés au numérique à l'intersection des sciences humaines et sociales, comme le suggère le schéma à droite.



Les premiers échanges au sein du groupe font apparaître trois principaux **thèmes de réflexion** qui devraient être au centre de l'analyse à venir :

1. Les évolutions des pratiques d'audit, de **certifiabilité**, d'assurabilité à l'ère de la transition numérique ;
2. Les évolutions des **cultures professionnelles** liées à la transition ;
3. Compléter les analyses de risques *a priori* par des méthodes plus dynamiques.

Ces thèmes sont décrits sur les pages suivantes.

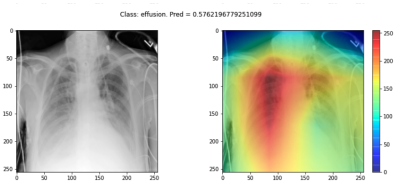
Thème 1 : Approche organisationnelle et systémique de l'explicabilité

Le développement du nombre de capteurs embarqués et la croissance des capacités de communication et de stockage conduisent à une augmentation massive de la quantité de données disponibles concernant le fonctionnement des systèmes. Ces données sont souvent traitées par des composants logiciels développés par des entreprises tierces qui sont intégrés par les développeurs de systèmes sous forme de « **boites noires** ». Les concepteurs, exploitants, et autorités de sécurité, ont aujourd'hui une capacité limitée à accéder au fonctionnement interne de ces composants logiciels et obtenir des informations fiables sur leur conception. Par exemple :

- Quelles hypothèses ont été adoptées concernant les **données d'entrée** qui ont servi à l'**apprentissage** du modèle numérique ?
- Quels processus ont permis de vérifier que les données d'apprentissage respectaient bien ces hypothèses ?
- Comment se comporte le composant sur des données qui ne correspondent pas à celles utilisées pour son apprentissage ?
- Comment encadrer la possibilité que le comportement du composant logiciel évolue au fil du temps grâce à sa capacité d'apprentissage ?

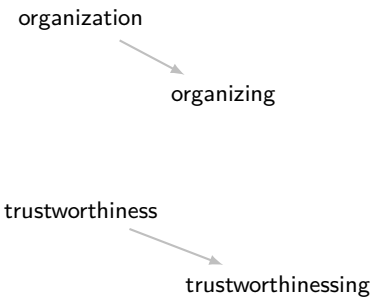
Cette évolution importante du paysage technologique interroge les pratiques d'audit, de certifiabilité et d'assurabilité de systèmes dans lesquels des composants logiciels assurent des fonctions importantes pour la performance et pour la sécurité.

D'importants travaux de recherche et développement sont actuellement conduits sur le sujet de l'**explicabilité** des composants logiciels complexes ("xAI" pour "explainable artificial intelligence"). S'agissant de logiciels de classification d'images, par exemple, les "saliency maps" permettent d'identifier les pixels d'une image qui sont les plus déterminants dans l'attribution d'une catégorie (*cf.* l'illustration à droite). Ces cartographies de la saillance permettent d'identifier des anomalies dans le fonctionnement de l'algorithme de classification. Ces approches restent toutefois très centrées sur le fonctionnement technique du composant, avec peu de prise en compte du contexte particulier dans lequel l'explication sera utilisée, ou des attentes des personnes et organisations recevant et interprétant l'explication.



Le groupe se focalisera davantage sur les **dimensions organisationnelles et systémiques de l'explicabilité** : quels types d'explications sont adaptés aux attentes et besoins de différents acteurs organisationnels (les concepteurs, les utilisateurs-opérateurs, les spécialistes sécurité, les tutelles) ? Comment prendre en compte la distinction entre *explication* (technique) d'un fonctionnement algorithmique et *intelligibilité* (contextuelle et sociale) de son comportement dans une situation particulière, où il se confronte à des pratiques et outils déjà établis ? Quels sont les **enjeux de responsabilité** associés au fait de recevoir une explication ? Quel impact sur la confiance de différentes catégories d'acteurs en le « bon » fonctionnement du composant logiciel pour l'utilisation qui en est faite ? Quelles caractéristiques attendre d'une explication utilisée dans une démonstration de sécurité liée à la certification d'un équipement ?

De la même façon que le célèbre spécialiste des organisations K. Weick a suggéré de déplacer le regard depuis l'organisation (vue comme une structure statique) vers les processus à l'œuvre pour maintenir et modifier l'organisation (le "organizing"), nous visons à adopter une **approche processuelle de l'explicabilité** : quels sont les activités, les enjeux de responsabilité et de pouvoir associés à la fabrique d'explications, de justifications et de preuves qui permettent d'obtenir une confiance justifiée dans le bon fonctionnement du système et sa capacité à répondre aux différentes attentes ("trustworthinessing", si l'on se permet des ajouts innovants à la langue anglaise) ?



Il existe une longue tradition de recherche et des connaissances pratiques établies sur l'**intégration d'automatismes** dans le **pilotage de systèmes critiques** : les travaux des années 1990 sur les "joint cognitive systems", sur les interfaces homme-machine, les retours d'expérience de l'introduction de dispositifs d'aide au pilotage dans l'aviation, par exemple. Il sera utile de réfléchir aux caractéristiques et enjeux spécifiques des données massives, de l'apprentissage profond, des IA génératives, et à quel point les connaissances établies peuvent guider la réflexion sur leur intégration dans des contextes à forts besoins de sécurité.

Thème 2 : Évolution des cultures professionnelles liée à la transition numérique

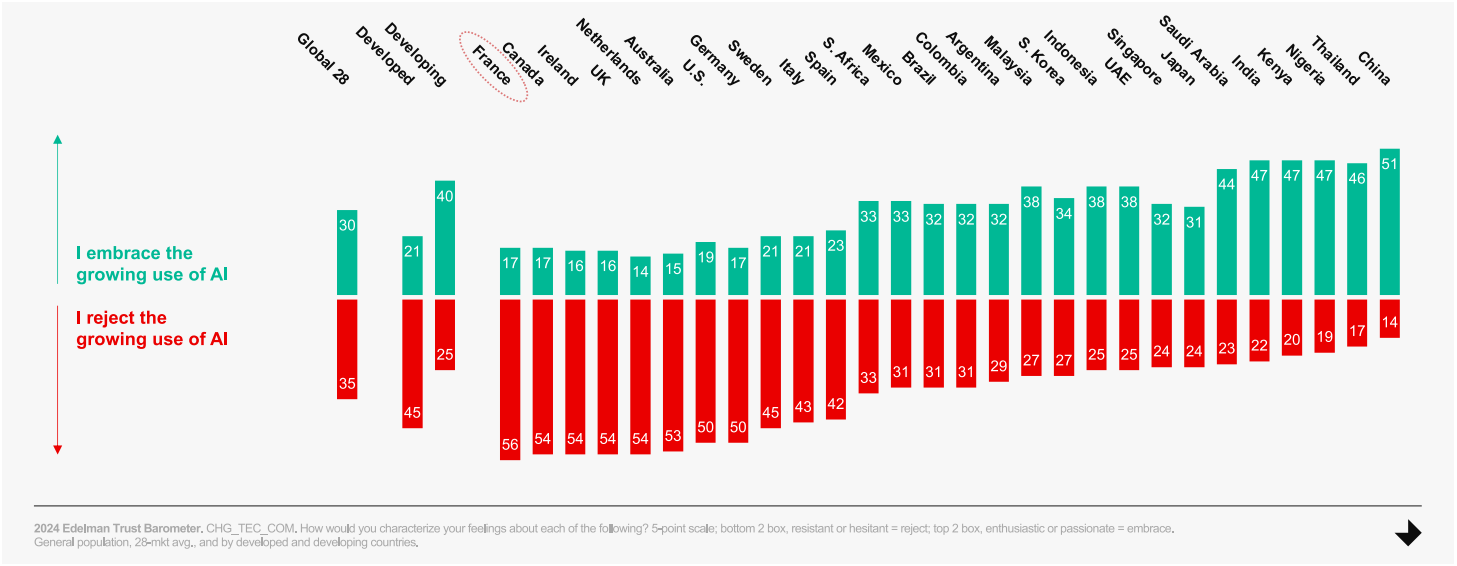
Lorsqu'ils sont intégrés dans des contextes organisationnels et professionnels, les outils numériques viennent compléter (ou interférer avec…) des outils, pratiques, **savoir-faire déjà bien établis**. Leur adoption, et la construction de la confiance en leur bon fonctionnement, varient considérablement selon les niveaux de compréhension et d'expertise des opérationnels concernant le numérique, ainsi que selon la nature conservatrice ou non de la **culture métier** de chaque spécialité disciplinaire, et de chaque individu.

> Move fast and break things. Unless you are breaking stuff, you're probably not moving fast enough.
>
> — Mark Zuckerberg (lettre aux investisseurs Facebook en 2012)

Ainsi, pour caricaturer, les experts de la tech adoptent le mantra "move fast and break things" de la Silicon Valley, valorisant l'innovation et la nouveauté et tolérant les défaillances comme sources d'apprentissage et de progrès, alors que la culture métier des professionnels de la sécurité tend à être plus conservatrice et considérer la défaillance comme un échec. Comment accompagner ces différences d'approche, chez des populations qui doivent travailler ensemble sur les grands projets ?

D'autre part, les cultures métier évoluent dans le temps avec l'introduction progressive d'automatismes et d'outils numériques d'aide à la décision. Historiquement, l'introduction de nouvelles technologies était généralement très progressive, et les professionnels d'un métier voyaient rarement leur activité bouleversée au cours d'une carrière. La **rapidité de l'innovation numérique** change la donne. Ainsi, les compétences et savoir-faire des pilotes de ligne ont très largement évolué au cours des trente dernières années avec l'introduction de nombreux automatismes. Les dispositifs numériques produisent des risques d'**habituation progressive** et de surconfiance chez les utilisateurs, qui ne sont pas toujours bien maîtrisés lorsque les systèmes en question sont jugés non critiques pour la sécurité (échappant ainsi aux analyses effectuées au nom de la certification ou des démonstrations de sécurité). De la même façon, ils peuvent modifier la perception de responsabilité de l'opérateur-décideur, produisant des effets sur la sécurité et des impacts juridiques qui ne sont pas toujours bien anticipés.

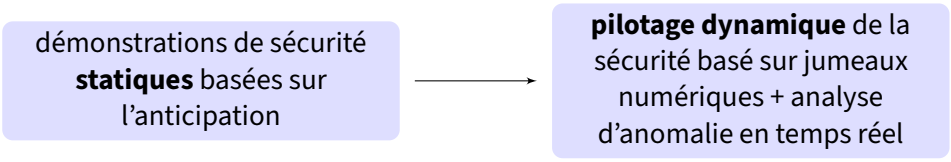
Notons en aparté que les **enjeux d'acceptabilité** sont particulièrement forts dans notre pays, puisque la France est le pays avec le plus faible niveau d'acceptation de l'adoption de technologies IA d'après le baromètre Endelman de 2024.



Thème 3 : Compléter les analyses de risques a priori par des méthodes plus dynamiques

La croissance de la quantité de données disponibles, et les capacités prédictives des « **jumeaux numériques** », ouvrent la porte à des changements du modèle conceptuel utilisé pour assurer la sécurité. La manière de piloter la sécurité pourrait devenir à la fois plus ambitieuse en matière d'identification et de surveillance (automatisée) des risques, et plus réactive et contextuelle. La démonstration de sécurité, qui se base aujourd'hui principalement sur la qualité de la conception et l'application des procédures d'exploitation et de maintenance, pourrait être complétée par la capacité à identifier des anomalies et des aléas en **temps réel**, à anticiper leurs conséquences possibles, et à ramener le système vers un espace de fonctionnement sûr.

Comme l'indique le schéma ci-après, les démonstrations de sécurité actuellement mises en œuvre sont statiques et basées sur une anticipation — se voulant — exhaustive des situations et aléas pouvant survenir. L'analyse de ces situations dangereuses permet d'identifier leur complément, l'espace de fonctionnement sûr, qui sera le **domaine de certification**. Les concepteurs imaginent des barrières techniques et organisationnelles qui permettront d'assurer que le système restera, dans son fonctionnement futur, dans le domaine de certification. Ainsi, la démonstration de l'assurance sécurité est **statique**, évoluant seulement lorsque des modifications sont apportées au système, ou lorsque l'expérience opérationnelle met en évidence des lacunes dans les analyses initiales.



Les jumeaux numériques performants et l'analyse d'anomalie en temps réel permettent d'envisager un **pilotage plus dynamique de la sécurité**. Ce mode de fonctionnement s'appuierait sur les composants suivants, sortes d'hypothèses à valider :

- C1 : des jumeaux numériques **fidèles** même dans les **situations non nominales** (ce qui a rarement été le cas des modèles de simulation historiques, même les plus performants) ;
- C2 : la capacité à assurer une mesure dynamique de la **distance à l'accident**, en s'appuyant sur des mesures en temps réel plutôt que sur un calcul effectué à la conception ;
- C3 : la capacité à mesurer les **marges adaptatives** encore disponibles au sein du système (sa capacité à maintenir le contrôle) ;
- C4 : l'existence de **moyens de pilotage et de contrôle** permettant d'éviter les zones de risque inacceptable, même si ces zones n'ont pas été identifiées à la conception, et de reconstruire les marges adaptatives lorsqu'elles se dégradent. Ces moyens de contrôle permettent un mode de pilotage qui ressemble, dans son esprit, à l'« approche par états » (APE) dans le secteur nucléaire.

Dans cette nouvelle approche du pilotage de la sécurité, la démonstration de sécurité consisterait à assurer la validité de ces quatre composants-hypothèses, plutôt que de viser à la conception une identification exhaustive des aléas et dangers qui pourraient survenir. Si ce mode de fonctionnement permettrait d'agrandir le domaine de certification et d'obtenir un niveau d'assurance sécurité supérieur face aux aléas, on anticipe qu'il peut être difficile de modéliser les réactions des humains et les dynamiques collectives dans les jumeaux numériques.

Le travail est lancé

L'analyse stratégique rassemble des experts académiques et des représentants des mécènes et partenaires de la Foncsi (exploitants, autorités, réseaux de réflexion déjà constitués sur ces sujets). Le groupe va régulièrement inviter des chercheurs, experts et opérationnels à présenter leurs travaux ou témoigner de leur expérience lors des réunions. Les prochains mois seront consacrés à la préparation de la "big picture", un état des problèmes et des acteurs clés, ainsi qu'à l'identification de chercheurs ayant travaillé sur des thèmes en lien avec les questions évoquées dans ce document.

Une question ou réaction ? N'hésitez pas à contacter Eric Marsden, qui anime cette analyse stratégique de la Foncsi avec Véronique Steyer, professeure à Polytechnique.

Courriel : eric.marsden@foncsi.org

Ce Journal des analyses stratégiques est publié par la Foncsi et diffusé à l'ensemble des mécènes et partenaires du programme scientifique «Foncsi 4».