Lignes directrices de la FSA portant sur l'utilisation de l'intelligence artificielle (IA)

1. Introduction

L'IA est actuellement au cœur de toutes les discussions. Selon les éditeurs de logiciels, elle serait pour ainsi dire omniprésente. Les avocates et les avocats utilisent de plus en plus des systèmes d'IA, notamment pour des tâches telles que la traduction automatique ou l'analyse de vastes ensembles de données lors de *due diligences* ou d'enquêtes internes destinées aux clients. L'IA générative, qui produit elle-même des contenus comme des images ou des écrits, est également en plein essor, en particulier pour résumer, corriger ou améliorer des textes existants.

Les lignes directrices qui suivent se concentrent sur l'IA générative, mais les recommandations s'appliquent *mutatis mutandis* à tous les logiciels d'IA.

Comme c'est souvent le cas pour les nouvelles technologies, les systèmes d'IA présentent des avantages, notamment en termes d'efficacité accrue, mais aussi des risques. Ces lignes directrices visent à fournir un cadre de référence pour une utilisation responsable de l'IA dans la pratique des avocates et des avocats. Il est cependant fortement recommandé aux études d'établir des directives internes concernant le recours à l'IA et de définir des règles adaptées à leurs besoins spécifiques.

En raison de l'évolution fulgurante de l'IA, il est important de garder à l'esprit que les explications fournies ci-dessous pourront rapidement devenir obsolètes. C'est pour cette raison que nous avons fait le choix délibéré de ne pas mettre en avant des outils d'IA spécifiques.

2. Risques liés à l'IA et utilisation sécurisée

2.1 Secret professionnel de l'avocat, protection des données et autres obligations de confidentialité

Lors du choix et de l'utilisation de tout logiciel, y compris les applications d'IA, il convient de clarifier au préalable ce qu'il advient des données saisies, notamment en identifiant qui y a accès et où elles sont stockées, que ce soit de manière temporaire ou permanente. Cette démarche doit non seulement prendre en compte le secret professionnel de l'avocat, mais aussi se conformer à la loi sur la protection des données (LPD) ainsi qu'à d'autres obligations de confidentialité. De plus, les membres de la Fédération suisse des avocats sont soumis aux règles déontologiques en vigueur.

D'une manière générale, les options suivantes s'offrent aux avocates et avocats :

1. Le logiciel d'IA est installé et exploité sur le réseau interne de l'étude d'avocats (c.-à-d. *in situ*), avec la garantie qu'aucune donnée ne quitte ce réseau ou n'est stockée en dehors de l'infrastructure de l'étude ;

2. Le respect strict des règles relatives à l'externalisation, lorsque des applications sont obtenues *via* un fournisseur et éventuellement utilisées sur l'infrastructure de celui-ci. Pour plus de précisions, il est recommandé de consulter les directives de la FSA concernant la sous-traitance informatique et l'utilisation de services cloud (disponibles au téléchargement sur https://digital.sav-fsa.ch/fr/transition-numerique-de-l-etude-utilisation-du-cloud) ;

3. Il est également possible d'établir, avec le consentement éclairé de son client, une déclaration de renonciation concernant le secret professionnel et la LPD.

En l'absence de ces garanties, les informations confidentielles, les secrets d'entreprise (stratégies, données financières, etc.), les informations ou données personnelles des employés, des clients, des partenaires commerciaux ou d'autres personnes, sous quelque forme que ce soit (y compris les photos,

les vidéos, etc.), ainsi que les contenus protégés par des droits de propriété intellectuelle, notamment les droits d'auteur, ne doivent pas être introduits dans les systèmes d'IA.

2.2 Examen critique des résultats obtenus par l'IA

Les systèmes d'IA ne sont ni omniscients ni infaillibles. Les résultats qu'ils produisent peuvent être inexacts, incomplets ou insuffisants. Il est donc crucial de faire preuve de discernement et d'examiner de manière critique les réponses obtenues, en les corrigeant ou en les complétant si nécessaire.

Il est important de garder à l'esprit que les systèmes d'IA ne sont pas capables d'évaluer la justesse de leurs propres conclusions. Partant, il est impossible de demander à une IA de vérifier la précision de ses réponses, car cette capacité lui fait défaut.

Les erreurs ou les résultats insatisfaisants des applications d'IA peuvent provenir de plusieurs facteurs, notamment :

- Le phénomène d'hallucination, qui se produit lorsque l'IA génère des résultats qui ne correspondent à aucune réalité ;

- L'utilisation d'informations erronées ou incomplètes, due à l'absence de données de base pertinentes. Par exemple, un système d'IA peut être entraîné sur un ensemble de données obsolètes, ce qui lui fait omettre des événements ultérieurs. De plus, les systèmes d'IA ont tendance à générer davantage de fausses informations lorsque le nombre de sources consultées pour une question donnée est limité.

- La *sycophancy* est un phénomène dans lequel un modèle d'IA ajuste ses réponses de manière complaisante, les alignant sur le point de vue de l'utilisateur, de sorte que l'explication fournie par l'IA est objectivement biaisée (cf. *infra*).

Dans ce même contexte, il est important de souligner que les systèmes d'IA peuvent être affectés par des biais. Ceux-ci résultent du jeu de données employé pour l'entraînement du système, de la méthode appliquée, ainsi que des décisions prises par les développeurs lors de la modélisation. Bien que cette problématique soit souvent associée à l'utilisation de l'IA pour l'analyse de données, elle est également présente pour l'IA générative.

2.3 Responsabilité

En tant que mandataires, les avocates et les avocats endossent une responsabilité en cas de mauvaise exécution du mandat, et ils ne peuvent se soustraire à cette responsabilité en invoquant une erreur commise par l'IA.

Lors de l'utilisation de systèmes d'IA en accord avec son client (par exemple pour l'analyse de grandes quantités de données), il est recommandé de clarifier préalablement ces questions et de convenir d'une clause d'exclusion de responsabilité dans les limites autorisées par le droit.

2.4 Droit d'auteur

L'utilisation de matériel protégé par le droit d'auteur, en particulier pour l'entraînement de modèles de langage étendus (LLM pour *large language models*) ou de générateurs d'images et de vidéos, fait l'objet de vifs débats. La controverse porte sur le point de savoir si l'utilisation de tels ensembles de données d'entraînement sans le consentement des titulaires de droits viole leurs droits d'auteur ou si, au contraire, une utilisation limitée à des fins d'entraînement reste admissible. Cette question ne se pose généralement pas lors d'une utilisation standard d'applications d'IA. Elle peut toutefois devenir pertinente lorsque les études d'avocats entraînent leurs propres LLM ou enrichissent des LLM existants.

En l'absence de droits d'utilisation appropriés, les données d'entrée utilisées pour l'entraînement des systèmes d'IA ou pour la génération de résultats peuvent donc potentiellement violer les droits d'auteur de tiers.

Toutefois, dans la mesure où il s'agit d'une pure réalisation de l'IA, le résultat produit n'est en principe pas considéré comme une création intellectuelle au sens du droit d'auteur. Néanmoins, si un système d'IA est utilisé pour la collecte d'idées initiales ou pour une première ébauche, il est possible qu'une création intellectuelle individuelle soit ultérieurement produite, pouvant alors être considérée comme une œuvre protégée par le droit d'auteur malgré l'utilisation de l'IA. Il n'est pas non plus totalement exclu que le résultat produit par une IA générative porte atteinte aux droits d'auteur de tiers en ressemblant trop fortement à une œuvre protégée. Cette éventualité reste cependant peu probable.

2,5 Obligation d'informer les clients en cas d'utilisation de l'IA

Certains fournisseurs intègrent dans leurs conditions d'utilisation des clauses obligeant les utilisateurs à divulguer l'emploi de l'IA. Par conséquent, avant de recourir à une application d'IA, il est crucial de vérifier attentivement les conditions du prestataire concerné.

Par ailleurs, une obligation d'informer ses clients peut éventuellement se présenter lorsque ceux-ci requièrent ou s'attendent légitimement à une exécution strictement personnelle de l'avocate ou l'avocat.

3. Réglementation des systèmes d'IA

En novembre 2023, le Conseil fédéral a chargé le DETEC de présenter, d'ici fin 2024, des propositions pour réglementer l'IA en Suisse. Cette analyse doit se baser sur le droit en vigueur tout en identifiant les approches réglementaires envisageables et compatibles avec l'AI-Act de l'Union européenne et la Convention sur l'IA du Conseil de l'Europe (cf. *infra*). Une attention particulière sera portée au respect des droits fondamentaux, ainsi qu'aux normes techniques, implications financières et institutionnelles des différentes approches réglementaires. Le Conseil fédéral prévoit que son administration élaborera un projet de loi sur l'IA en 2025.

Le 12 juillet 2024, l'UE a publié au Journal officiel l'une des premières réglementations complètes sur l'IA (« loi sur l'IA » ou « AI-Act »). Cette loi produira les effets suivants :

- 1er août 2024 : entrée en vigueur de loi sur l'IA.
- 2 février 2025 : interdiction de certaines pratiques concernant les systèmes d'IA présentant un risque inacceptable.
- 2 août 2025 : applicabilité des règles sur les systèmes d'IA à usage multiple et du système de sanctions.
- 2 août 2026 : Entrée en vigueur des dispositions restantes de loi sur l'IA (à l'exception de ce qui suit).
- 2 août 2027 : entrée en vigueur du reste des dispositions relatives à la classification des systèmes d'IA autres que ceux visés à l'annexe 3 de l'AI-Act en tant que systèmes d'IA à risques élevés et des dispositions y afférentes.

La loi sur l'IA présente un large champ d'application, à la fois en termes de personnes concernées et de portée territoriale. Elle s'applique non seulement aux développeurs de systèmes d'IA et aux personnes qui les mettent sur le marché ou en service (appelés « fournisseurs » dans la loi), mais aussi à toutes les personnes qui utilisent un système d'IA dans le cadre de leur activité professionnelle (appelées « utilisateurs »). L'AI-Act englobe les fournisseurs et les utilisateurs au sein de l'UE, mais aussi, dans une certaine mesure, ceux qui se trouvent en dehors de l'UE et qui accèdent au marché de l'UE. Les personnes qualifiées de fournisseurs ou d'utilisateurs feraient bien de vérifier si l'AI Act s'applique à eux.

L'AI-Act se fonde sur une approche privilégiant une évaluation concrète des risques. À cet égard, les systèmes d'IA sont classés en quatre catégories :

1. Les logiciels d'IA présentant des risques inacceptables sont formellement interdits. Cela inclut, entre autres, les systèmes biométriques d'identification à distance et en temps réel (tels que la « reconnaissance faciale »), à moins qu'ils ne soient employés dans le cadre de poursuites pénales selon des protocoles stricts (par exemple, approbation préalable par une autorité judiciaire). Cette catégorie de risques inacceptables englobe le *social scoring*, c'est-à-dire un classement de la fiabilité des personnes physiques en fonction de leur comportement social, de caractéristiques personnelles ou de traits de personnalité connus ou prédits, si cela entraîne un désavantage ou une discrimination injustifiée ou disproportionnée à l'égard de ces personnes ;

2. Les systèmes d'IA à risques élevés sont autorisés sous réserve de conditions strictes, comprenant la gestion des risques, la gouvernance des données, la documentation technique, l'obligation de tenir un registre, la transparence des systèmes, la supervision humaine, les critères d'exactitude, de robustesse et de cybersécurité. Cette catégorie englobe notamment les systèmes d'IA utilisés pour les candidatures, promotions en cours d'emploi ou résiliations de contrats de travail, l'examen des demandes à l'aide sociale, l'évaluation des performances scolaires des élèves, les examens d'admission à l'université, ainsi que les prédictions de récidive des délinquants, etc. ;

3. Les systèmes d'IA présentant des risques limités sont soumis à des obligations de transparence restreintes, telles que l'information de l'utilisateur sur l'interaction avec un système d'IA ou l'identification des contenus générés par celui-ci ;

4. Les systèmes d'IA présentant un risque nul ou faible, comme ceux liés aux jeux vidéo, sont autorisés et restent largement non réglementés.

L'AI-Act prévoit la création d'un comité européen pour l'IA ainsi que l'établissement d'autorités de surveillance nationales habilitées à sanctionner les entreprises contrevenantes. Cependant, il n'est pas prévu que les individus puissent saisir directement une autorité en cas de litige.

Parallèlement aux initiatives de l'Union européenne, le Conseil de l'Europe a également travaillé à l'élaboration d'une convention sur l'IA. Le 17 mai 2024, le Comité des Ministres a adopté la Convention-cadre sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit (ci-après « Convention sur l'IA »). Ce traité international, initialement disponible en anglais et en français, vise à garantir que l'IA soit développée et utilisée dans le respect des droits fondamentaux, de la démocratie et des principes de l'État de droit. La Convention sur l'IA est ouverte à la signature et à la ratification des États membres du Conseil de l'Europe, ainsi que d'autres États ayant participé à son élaboration. Elle entrera en vigueur dès que cinq États, dont au moins trois États membres, l'auront ratifiée.

Conseil FSA, 14 juin 2024

Annexe : lexique

AI ou Artifical intelligence	IA ou intelligence artificielle
Syndrome de la boîte noire	Désigne l'impossibilité ou la difficulté de contrôler l'approche et la méthode de résolution des systèmes d'IA. Le fonctionnement interne n'est souvent pas vérifiable par les utilisateurs, voire par les programmeurs eux-mêmes.
Biais	Partialité d'un système d'IA due à l'ensemble des données d'entraînement, à la méthode d'entraînement utilisée ou aux décisions de modélisation prises par les programmeurs.
GPT	Transformateur pré-entraîné et génératif (« *Generative Pre-trained Transformer* »).
Hallucination	Informations prétendues ou erronées qui ne sont pas basées sur des données ou des événements réels, mais qui sont présentées comme telles.
Input	Donnée injectée dans un système d'IA.
IA ou intelligence artificielle	L'IA se réfère à la capacité des machines à effectuer des tâches qui nécessiteraient normalement l'intelligence humaine. Cela englobe la résolution de problèmes, l'apprentissage, la reconnaissance vocale, la prise de décision, etc. Les algorithmes et les systèmes d'IA peuvent analyser des données, reconnaître des modèles, et formuler des prédictions ou prendre des décisions sur cette base.
Loi sur l'IA ou AI-Act	Loi sur l'IA règlementant l'IA (https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=OJ:L_202401689).
LLM ou *large language model*	Modèle d'apprentissage automatique capable d'effectuer des tâches dans le domaine du traitement du langage naturel (NLP).
NLP ou *natural language processing*	Traitement du langage naturel par un système d'IA.
Output	Résultat produit à l'aide d'une IA générative
Prompt	Instruction donnée à une IA générative pour produire un résultat correspondant.