Septembre 2024
Version 2.0

SOCIETE NUMERIQUE

POSITION

SUR LA RÉGLEMENTATION DES SYSTÈMES DÉCISIONNELS AUTOMATISÉS

Collaborateurs de la version 2.0 :

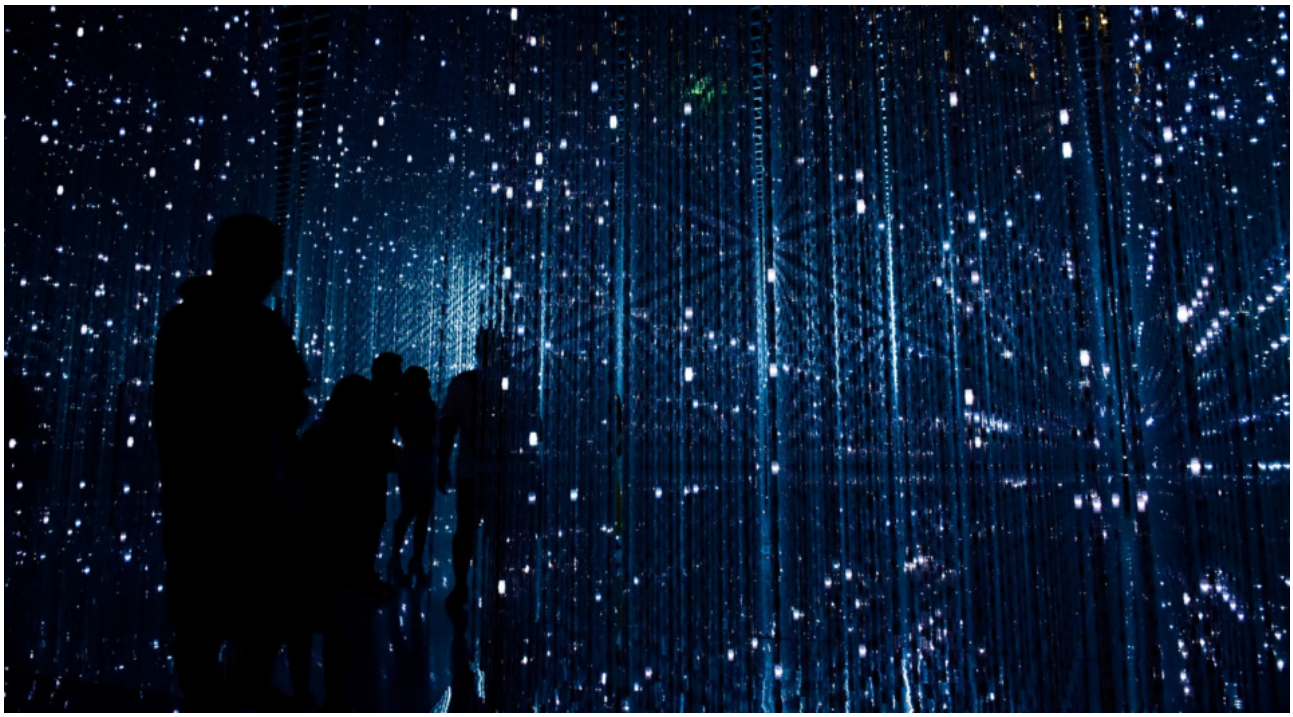
David Sommer, Lars Lünenburger, Erik Schönberger, Andreas Geppert, Luka Nenadic, Christian Sigg, Tanja Klankert, Peter Fröhlich, David Caspar, Thomas Mandelz und Jan Zwicky

Collaborateurs de la version 1.0 :

David Sommer, Andreas Geppert, Christian Sigg, David Caspar, Erik Schönenberger, Jan Zwicky, Lars Lünenburger, Luka Nenadic, Peter Fröhlich, Tanja Klankert und Thomas Mandelz

Traduction de la version 2.0 :

Charles Morel, Pierre Boizot et Philippe Gervais



SOCIÉTÉ NUMÉRIQUE / CH-4000 BÂLE
OFFICE@DIGITALE-GESELLSCHAFT.CH / SOCIETE-NUMERIQUE.CH

Résumé

Les systèmes automatisés de prise de décision (système ADM, ADMS) présentent un potentiel sociétal et économique important. Toutefois, ils comportent des risques considérables pour les individus et la société, contre lesquels la législation actuelle en Suisse n'offre pas une protection suffisante. C'est pourquoi Société Numérique propose un cadre réglementaire pour la réglementation des systèmes ADM, structurés autours de cinq piliers fondamentaux :

- Lors de l'utilisation de systèmes ADM, la protection des individus et de la société doit être garantie. Les **objectifs principaux**, en matière de protection individuelle, sont de respecter les droits fondamentaux et les droits humains. Lorsque ces systèmes affectent un grand nombre de personnes, des processus sociétaux ou des institutions démocratiques, ils relèvent d'objectifs de protection collective. Ces objectifs s'appuient sur le concept de protection des données développé par Société Numérique.
- Les systèmes ADM sont classés en **trois catégories de risque** : 1. *« Risque faible »* : systèmes qui ne devraient pas engendrer d'effets négatifs significatifs sur les individus ou la société. 2. *« Risque élevé »* : systèmes susceptibles de causer des dommages importants, notamment lorsqu'ils sont utilisés dans des secteurs sensibles comme les services sociaux. 3. *« Risque inacceptable »* : systèmes interdits par principe, tels que la surveillance biométrique de masse. Pour éviter une bureaucratie excessive, la classification repose sur des *auto-déclarations*. L'exactitude de ces déclarations est assurée par des sanctions *a posteriori* en cas de fausses informations.
- Afin de ne pas freiner l'innovation, les opérateurs privés ne doivent pas être soumis à des obligations légales « ex ante » excessives. Cependant, cette approche ne doit pas être perçue comme un blanc-seing permettant une utilisation irresponsable des systèmes ADM. Il est essentiel que les opérateurs privés et les autorités assument pleinement la responsabilité des dommages causés par une utilisation inappropriée des systèmes ADM. La proposition adopte ainsi une **approche mixte, combinant gestion des risques et des dommages**.
- En contrepartie de la liberté de conception accordée aux opérateurs, les systèmes classés « risque élevé » sont soumis à des **obligations de diligence et de transparence**, notamment concernant la qualité et l'origine des données utilisées. Les systèmes ADM déployés par les pouvoirs publics doivent respecter des standards encore plus stricts.
- **Des mécanismes efficaces de supervision et de sanction** sont indispensables. Société Numérique recommande une supervision étatique des systèmes ADM, l'imposition de sanctions dissuasives à l'encontre des opérateurs privés, ainsi que l'instauration d'un droit de recours collectif pour les personnes touchées par des pratiques non conformes à la loi.

Table des matières

1 Introduction

L'intelligence artificielle (IA) ou les systèmes automatisés de prise de décision (Automated Decision-Making Systems, ADM-System, ADMS, voir Glossaire) ne relèvent plus de la science-fiction. Ils sont déjà utilisés dans notre quotidien, nous aident à prendre des décisions, simplifient l'interaction avec les ordinateurs et peuvent générer des textes, des images ou de la musique. Les promesses de ces technologies sont très attrayantes. Par exemple, nous pouvons parler aux machines, automatiser davantage des tâches pénibles et innover dans des domaines tels que le diagnostic et le traitement personnalisés en médecine. La FDA, l'autorité américaine compétente en matière d'autorisation des produits médicaux, répertorie déjà plusieurs centaines de dispositifs médicaux utilisant l'IA et l'apprentissage automatique (Machine learning) (au 13 mai 2024, il y en avait 882). [1]

Cependant, comme pour la plupart des révolutions technologiques, des aspects négatifs émergent également, qui ne manquent pas en ce qui concerne l'IA et les systèmes ADM. Le cas le plus connu est probablement l'affaire des allocations familiales aux Pays-Bas, où des milliers de familles ont été injustement sommées de rembourser des allocations déjà versées, car la pluralité de nationalités avait été utilisée comme critère discriminant. [2] À l'heure actuelle, il est pratiquement impossible de dresser une vue d'ensemble de tous les domaines dans lesquels l'IA est utilisée ou ne l'est pas. Nous ne faisons qu'effleurer la surface d'un changement dont l'ampleur reste difficile à évaluer. Par conséquent, il est crucial d'orienter cette évolution afin de maximiser ses avantages tout en minimisant ses effets négatifs.

Des États et des entités, comme l'Union européenne, la Chine ou les États-Unis, ont reconnu le potentiel transformateur de l'IA et des systèmes ADM et s'efforcent activement d'introduire des cadres réglementaires dans ce domaine. Société Numérique s'engage [3] activement dans ces discussions et demande que la Suisse adapte son cadre réglementaire aux nouveaux équilibres et défis posés par les systèmes d'IA et ADM, afin de garantir un rapport satisfaisant entre les avantages et les risques.

Ce document propose un cadre réglementaire technologiquement neutre [4] pour la Suisse, reposant sur une approche centrée sur l'humain : l'utilisation de l'IA et des systèmes ADM doit viser à profiter aux personnes et à améliorer leur bien-être.

L'utilisation des systèmes ADM doit être soumise à des exigences strictes de *transparence* et de *traçabilité*. Un système ADM satisfait à la condition de *transparence* si les bases de ses décisions automatisées sont rendues accessibles. Par exemple, un système ADM utilisé pour évaluer la réintégration professionnelle des personnes accidentées est-il en mesure de démontrer qu'il a pris en compte et évalué tous les faits pertinents ? Le critère de *traçabilité* exige que les processus décisionnels d'un système ADM soient compréhensibles, afin que les personnes concernées puissent comprendre l'évaluation des faits pertinents et, si nécessaire, les contester.

Le cadre réglementaire proposé repose sur une évaluation des risques liés aux systèmes ADM. Les personnes concernées, les ONG accréditées et une autorité publique de surveillance des systèmes ADM doivent disposer d'un droit de regard sur les applications utilisant de tels systèmes. Ce cadre prend également en compte le fait que les risques associés à un système peuvent évoluer au fil du temps. Par ailleurs, il est compatible avec le nouveau concept de protection des données de Société Numérique.

En tant que société civile, nous n'avons plus la possibilité de décider si nous voulons ou non utiliser les systèmes ADM. Ils sont déjà une réalité. Cependant, nous pouvons encore choisir dans quels domaines nous souhaitons être assistés par de tels systèmes et dans quels domaines nous préférons rejeter leur utilisation.

2 Champ d'application

Notre proposition s'applique aux systèmes ADM qui prennent des décisions à l'aide de systèmes techniques, soit de manière entièrement automatisée,

[1] Voir Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA

[2] Voir l'article : Aufsicht und Transparenz : Wie die Niederlande aus KI-Skandalen lernen (netzpolitik.org) / Surveillance et transparence : comment les Pays-Bas tirent les leçons des scandales de l'IA (netzpolitik.org)

[3] C'est le cas par exemple de la Convention-cadre sur l'IA, les droits humains, la démocratie et l'État de droit du Conseil de l'Europe, voir le dossier DSM-Systèmes de la société numérique pour plus d'informations.

[4] Notre proposition de réglementation se concentre sur les effets et les risques des systèmes ADM et non sur les interdictions de technologies concrètes.

soit en tant qu'aide à la décision. Nous avons délibérément choisi d'éviter le terme controversé d'intelligence artificielle et d'utiliser à la place la définition suivante des systèmes ADM, basée sur une recommandation de l'AI Now Institute (Richardson et al., 2019, p. 20) pour la ville de New York :

An «automated decision system» is any software, system, or process that aims to automate, aid, or replace human decision-making. Automated decision systems can include both tools that analyze datasets to generate scores, predictions, classifications, or some recommended action(s) that are used by agencies to make decisions that impact human welfare, [5] and the set of processes involved in implementing those tools.

Nous définissons ainsi les systèmes DSM non pas du point de vue de la technologie utilisée, mais de celui de leur impact. Cela permet d'éviter les problèmes de définition auxquels l'OCDE a été confronté (OCDE 2024).

Les systèmes ADM utilisent des algorithmes et/ou des techniques d'intelligence artificielle pour prendre des décisions. Ils sont souvent (mais pas toujours) basés sur des données. Cela ne signifie toutefois pas que tous les algorithmes, systèmes d'intelligence artificielle ou outils de Big Data soient concernés par cette réglementation.

Bien qu'en principe, il soit techniquement correct d'affirmer que presque tous les programmes informatiques prennent des décisions en permanence, cette affirmation n'est pas pertinente dans le contexte de la réglementation. Les décisions concernées doivent être perceptibles en tant que décisions individuelles et distinctes, et avoir une certaine importance. En tant que telles, elles doivent avoir un impact concret sur les libertés, la vie, la santé ou la situation économique ou sociale des individus ou des groupes. Un système qui influence subtilement les comportements (par exemple, en modifiant les choix par un **nudging** algorithmique) relève également de la réglementation dans la mesure où ses décisions entraînent des répercussions mesurables sur la société.

Les systèmes techniques qui ne relèvent pas du champ d'application de cette réglementation n'ont pas besoin d'être catégorisés par niveau de risque (voir section 6), ce qui permet d'éviter une charge administrative inutile.

3 Résumé du cadre réglementaire proposé

Le cadre réglementaire suit une approche mixte, basée à la fois sur les dommages et sur les risques. Dans la première approche, les sanctions ne sont imposées qu'*a posteriori,* en cas de dommage, tandis que dans la deuxième approche, les applications à risque sont soumises à certaines exigences dès leur mise en œuvre. Toute personne utilisant un système ADM doit évaluer et classer le risque qu'il représente pour les individus et la société, dans le cadre des objectifs de protection. Le cadre réglementaire prévoit trois catégories : « risque faible », « risque élevé » et « risque inacceptable », définies par le pouvoir législatif. En principe, les catégories dépendent du risque que le système représente pour les individus et la société dans son ensemble. Par exemple, les « systèmes à risque faible ou nul » ne présentent aucun risque ou un risque négligeable pour les individus et aucun risque pour la société, tandis que les « systèmes à risque inacceptable » présentent un risque très élevé et inacceptable pour les individus ou pour la société. Entre ces deux extrêmes se situent les « systèmes à haut risque », soumis à des obligations étendues de transparence et de diligence, afin de permettre au public d'évaluer les risques et l'utilité des systèmes. Contrairement aux systèmes présentant un risque inacceptable, ceux à haut risque ne sont pas interdits.

Nous considérons en premier lieu l'impact des systèmes ADM sur les individus. Toutefois, l'utilisation à grande échelle de systèmes facilement déployables peut également engendrer des risques pour la société [6] qui ne sont pas suffisamment mesurables ou sanctionnables à l'échelle individuelle. Un exemple typique est la diffusion massive de publicités politiques individualisées sur les réseaux sociaux. Si le risque peut sembler négligeable au niveau individuel, ces systèmes peuvent, à grande échelle, avoir des conséquences notables, par exemple sur les résultats électoraux, portant ainsi atteinte à la démocratie. La ju-

[5] Impact on public welfare includes but is not limited to decisions that affect sensitive aspects of life such as educational opportunities, health outcomes, work performance, job opportunities, mobility, interests, behavior, and personal autonomy.

[6] https://booksummaryclub.com/weapons-of-math-destruction-book-summary/

risprudence suisse actuelle, notamment en matière de protection des données où elle se concentre essentiellement sur les individus, est insuffisante pour répondre à ces enjeux. Notre proposition vise à combler ce déficit en reconnaissant ces risques sociétaux et en proposant des recours collectifs, tels que les plaintes collectives et le droit de recours des organisations.

Le cadre réglementaire distingue les systèmes ADM utilisés dans le secteur privé de ceux employés pour l'accomplissement de missions de service public (voir sections 7.1, 7.2, 8.1 et 8.2). Dans les deux cas, afin de garantir la classification correcte des risques conformément à la section 6 et de veiller à l'application des obligations qui en découlent, nous demandons :

- un droit de recours ou d'action en justice pour les personnes concernées ;
- la possibilité pour l'autorité de surveillance gouvernementale et les ONG accréditées d'intervenir.

La nouvelle autorité de surveillance des ADMS (voir section 5.1) doit :

- recueillir les réclamations ;
- contrôler l'utilisation des systèmes ADM dans les entreprises et les services publics en cas de soupçon ;
- infliger des sanctions administratives en première instance, en fonction du chiffre d'affaires des entités concernées.

Ce comité d'experts, composé de spécialistes en sciences sociales ainsi que dans les domaines techniques et juridiques, doit pouvoir agir de manière indépendante sur les plans financier et organisationnel, sans recevoir d'instructions extérieures.

Pour ne pas entraver l'innovation, nous préconisons une approche fondée sur l'autodéclaration, plutôt que d'imposer des processus de contrôle bureaucratiques aux entreprises et à l'administration publique. Cette autonomie permet aux entités concernées d'organiser elles-mêmes la mise en œuvre concrète des règles dans les limites définies par le cadre réglementaire. Cependant, cette liberté doit être contrebalancée par des obligations supplémentaires, notamment :

- des exigences de transparence ;
- une obligation de diligence ;
- le renversement du fardeau de la preuve ;
- des sanctions efficaces en cas de non-respect.

Le fonctionnement exact d'un système ADM est souvent protégé par le secret industriel, ce qui complique pour les personnes externes l'obtention de preuves relatives à l'ampleur des risques associés. C'est pourquoi, en cas d'accusation justifiée (c'est-à-dire lorsqu'un tribunal accepte d'entrer en matière), le fardeau de la preuve doit être inversé. [7] L'entreprise incriminée doit alors, en tant que première instance dans la chaîne de recours, prouver la classification correcte de son système. Pour les systèmes utilisés dans le cadre de missions publiques, nous exigeons une large transparence, ainsi que la publication des systèmes et des données (voir glossaire), conformément à l'exigence « Public Money ? Public Code ! ». [8]

L'autorité de surveillance étatique des ADMS soutient les responsables du traitement [9] lors de l'évaluation des risques des systèmes ADM à l'aide de check-lists et de guides de bonnes pratiques dans le but d'assurer une sensibilisation et un traitement adéquats. Quiconque évalue incorrectement son système et ne respecte pas ses obligations, ou qui exploite un système ADM interdit parce que présentant un risque inacceptable, doit s'exposer à des sanctions sévères, proportionnelles à son chiffre d'affaires. Ces sanctions doivent rester administratives et ne pas viser pénalement les collaborateurs en tant qu'individus car il s'agit généralement de manquements collectifs. Les sanctions doivent néanmoins être utilisées en dernier recours.

Un nombre croissant d'institutions d'importance internationale, telles que l'Union européenne ou l'association professionnelle américaine des informaticien·ne·s (Association for Computing Machinery, ACM), s'intéresse désormais à la nécessité de réglementer les systèmes ADM (voir annexe). Nous sommes convaincus que le cadre proposé ici contri-

[7] L'UE a récemment introduit l'inversion du fardeau de la preuve pour les produits logiciels. Pour les produits techniquement complexes, les consommateurs peuvent exiger d'une entreprise qu'elle divulgue les preuves "nécessaires et proportionnées".

[8] Les logiciels payés par le secteur public et leur code source doivent être ouverts et accessibles à tous. https://publiccode.eu/

[9] Pour ce terme, nous utilisons exclusivement la forme masculine, car nous nous basons sur la définition de l'article 5, lettre j de la LPD.

buera à combler le vide juridique existant. Les aspects clés de ce cadre – en particulier les catégories de risques ainsi que les obligations de transparence et de diligence – sont détaillés ci-après.

4 La question sociétale

Une idée centrale est que **les système automatisé de prise de décision ne sont ni objectifs ni neutres**, car ils reflètent toujours les valeurs de leurs concepteurs ainsi que celles de la société. Ils peuvent donc être considérés comme le miroir socio-économique d'une société donnée. Cela pose problème car les valeurs culturelles et sociales varient d'une région du monde à une autre, tandis que les technologies, telles que les systèmes ADM, peuvent être utilisées de manière transfrontalière ou globale. De plus, ces valeurs peuvent évoluer au fil du temps, mais les systèmes, une fois paramétrés et utilisés sur le long terme, ne s'y adaptent pas nécessairement. En outre, les systèmes ADM sont contraints par leur fonction. Les décisions - conscientes ou inconscientes- prises par les développeurs dans le cadre de la conception de ces systèmes influencent leur fonctionnement et peuvent entraîner des conséquences négatives. Enfin, la fonction assignée à l'utilisation d'un système impose un cadre d'action restreint, qui est rarement remis en question.

Un exemple dans ce contexte serait l'utilisation de systèmes ADM pour réduire les frais de personnel dans le domaine des services sociaux. Ces systèmes prennent des décisions influençant massivement les ressources disponibles pour les personnes dépendant de l'aide sociale. D'une part, les principes décisionnels de ces systèmes sont discutables car les objectifs poursuivis par les développeurs ainsi que les valeurs sociétales sous-jacentes peuvent évoluer avec le temps. D'autre part, l'utilisation de tels systèmes est fondamentalement discutable : des études scientifiques récentes suggèrent qu'une augmentation des frais de personnel pourrait, au contraire, réduire les coûts globaux de ces programmes [10] (Eser Davolio 2020), ce qui pourrait plaider contre l'automatisation de ces tâches administratives.

Cependant, ces limitations sont souvent méconnues. On a trop tendance à considérer les résultats des systèmes ADM comme étant objectivement justes et corrects. Dans le contexte de la problématique liée à la prise de décisions, cette perception est toutefois trompeuse, car **dans de nombreuses situations, il n'existe pas de solution optimale**. En déléguant la prise quotidienne de décisions à un système ADM, ce dernier et tout ce qui le détermine ainsi que ses décisions deviennent partie intégrante de la réalité sociale. La sociologue Michele Willson décrit cela ainsi :

« An algorithm is delegated a task or process and the way it is instantiated and engaged with in turn impacts upon those things, people and processes that it interacts with - with varying consequences. » (Une tâche ou un processus est confié à un algorithme, et la manière dont ce dernier est conçu et manipulé influence à son tour les personnes et les processus avec lesquels il interagit – avec des conséquences variées.) (Willson 2017 : 139). Il en résulte des effets de rétroaction qui font que les systèmes automatisés de prise de décisions ainsi que leurs bases de données (parfois erronées ou imprécises) évoluent constamment et deviennent donc partie intégrante de la réalité sociale. Ces effets peuvent être intentionnels ou non.

Un effet particulièrement problématique est celui de la discrimination. Les systèmes automatisés de prise de décision, entraînés sur des ensembles de données, reproduisent les pratiques discriminatoires implicitement inscrites dans ces données, par exemple à l'encontre de groupes socialement défavorisés ou de personnes en situation de handicap. Ainsi, des systèmes de recrutement entraînés sur des données passées continueraient probablement à désavantager les femmes ou les personnes en situation de handicap, bien que de telles pratiques ne soient aujourd'hui plus tolérées.

Certes, les humains ne prennent pas nécessairement de meilleures décisions et ne sont pas exempts de préjugés. Cependant, ils sont capables de remettre ces décisions en question et en discuter, notamment grâce à l'aide de la technologie. **Cette capacité de réflexion fait défaut aux systèmes**, raison pour laquelle aucune décision ayant des impacts durables sur la société ne devrait leur être déléguée. Les effets discriminants des systèmes ADM risquent de s'aggraver, en particulier en raison de leur utilisation croissante et de leur mise en œuvre globalisée.

[10] Voir à ce sujet https://www.zhaw.ch/de/forschung/forschungsdatenbank/projektdetail/projektid/1668/

Un autre effet problématique réside dans le fait que les systèmes automatisés de prise de décision jouent un rôle croissant dans la sélection d'informations, par exemple sous forme de « systèmes de recommandation » ou d'outils qualifiés de « vérificateurs de faits » ou de « contrôleurs de droits d'auteur ». De cette manière, les opérateurs de ces systèmes se trouvent en position de renforcer sélectivement des messages et opinions via une orientation politique spécifique. Ces systèmes, par exemple lorsqu'ils fonctionnent comme vérificateurs de faits, n'ont pas nécessairement besoin de recourir à des données personnelles. Ce qui est commun à de nombreux effets des systèmes ADM, c'est qu'ils opèrent souvent de manière opaque et que leurs impacts ne deviennent visibles que tardivement ou par le biais d'effets secondaires indirects. La mise en œuvre et le fonctionnement de ces systèmes sont fréquemment inconnus, car leurs développeurs et opérateurs n'ont généralement aucun intérêt à les divulguer.

Enfin, l'interaction entre différents systèmes ADM engendre des risques supplémentaires, dont l'évaluation reste complexe. La survenue d'effets d'amplification par rétroaction (feedback loops, voir glossaire) est prévisible et constitue un risque sociétal. Cette **augmentation de la complexité** met en opposition des systèmes toujours plus sophistiqués et des humains qui, même dans la mesure où les systèmes individuels sont totalement transparents, peinent à en comprendre le fonctionnement.

Pour anticiper ces effets, **une obligation étendue de transparence et de diligence** doit être instaurée. Il faudrait, ne fusse que pour les systèmes automatisés de prise de décisions importants, créer un espace permettant un débat public pérenne sur les normes et principes sous-tendant la définition des métriques, des mesures ou des indicateurs utilisés dans l'évaluation de ces systèmes et l'interprétation de leurs résultats. **L'humain doit avoir la souveraineté sur les systèmes ADM**, et non l'inverse.

De plus, de nombreux effets ne peuvent pas être clairement identifiés d'un point de vue individuel, mais ne deviennent visibles qu'à travers l'accumulation des expériences de nombreuses personnes concernées. Malheureusement, les lois actuelles ne se focalisent que sur les conséquences individuelles. Par conséquent, il est nécessaire de développer des méthodes de **recours collectif**, qui sont encore rares dans le droit suisse.

En mettant l'accent sur les impacts et les risques, une **formulation technologiquement neutre** permet de s'adapter de manière flexible à de nouvelles méthodes ou à des changements dans l'utilisation des technologies existantes. L'objectif doit être que l'utilisation de ces systèmes améliore la condition humaine. Ces thèmes, et d'autres encore, sont intensément discutés dans les milieux scientifiques mais ne sont qu'esquissés ici dans un souci de concision.

5 Une proposition de réglementation pour les systèmes ADM

Nous sommes, en tant que société, de plus en plus confrontés aux conséquences spécifiques de l'utilisation des systèmes ADM. C'est pourquoi nous demandons **un élargissement substantiel des lois existantes** afin d'intégrer une réponse aux défis posés par ces systèmes, voire l'introduction d'une « loi ADMS », si cela devait favoriser l'atteinte du but visé. Nous exigeons **la transparence** dans l'utilisation des systèmes ADM, afin de permettre l'application des lois déjà en vigueur, ainsi que des **sanctions dissuasives** en cas de non-respect. Nous exigeons également que les systèmes ADM soient intelligibles, de manière à ce qu'ils puissent être compris rapidement et avec un effort intellectuel raisonnable par les êtres humains. Enfin, nous appelons à la mise en place d'une autorité de surveillance étatique ayant pour mission de surveiller les systèmes ADM critiques, avec la possibilité d'intervenir si nécessaire.

Nous nous focalisons en premier lieu sur les conséquences concrètes sur les individus et sur les moyens pour eux de faire valoir leurs droits. Cependant, certains risques touchent davantage la société dans son ensemble, comme par exemple l'influence politique exercée par le biais de campagnes publicitaires personnalisées ou bien les effets de rétroactions, par lesquels des systèmes interconnectés – avec ou sans intervention humaine – pourraient créer leurs propres référentiels de valeurs. **Si la transparence est essentielle, elle ne suffit cependant pas à elle seule sans mesures complémentaires**. Le droit à l'autodétermination individuelle en matière d'information nécessite, en plus de la connaissance des opérations, la possibilité d'exercer un certain contrôle sur ceux-ci.

Nous exigeons **des objectifs de protection clairs, à savoir le respect des droits fondamentaux et des droits humains, la préservation de la santé mentale et physique ainsi que de la sécurité de chaque individu, la préservation de l'égalité de traitement et des chances, ainsi que la protection des droits et processus démocratiques.** En outre, les personnes concernées et le public doivent avoir la

possibilité de contrôler effectivement le respect de ces objectifs de protection, de les questionner si nécessaire et de revendiquer leur application de manière simple et accessible. L'être humain doit conserver la prééminence : en d'autres termes, son interprétation doit, de manière générale, prévaloir sur celle de la machine, et les systèmes ADM doivent lui permettre de réaliser ses projets et d'atteindre ses objectifs de manière plus efficace, rapide et avec moins d'erreurs.

La réglementation des systèmes ADM ne doit ni freiner l'innovation ni imposer aux entreprises ou à l'autorité de surveillance des ADM, décrite ci-dessous, une charge administrative disproportionnée. Nous plaidons pour un cadre réglementaire large, introduisant la réglementation nécessaire de manière générique, tout en permettant aux différents secteurs économiques de déterminer eux-mêmes les méthodes les plus efficaces pour atteindre les objectifs de protection. Notre concept de classification, technologiquement neutre et basée sur les risques, **est compatible avec l'AI Act de l'Union européenne**. Il sera décrit en détail ci-dessous. Cependant, contrairement à la classification appliquée par l'UE basée sur les types d'utilisation , notre concept permet une classification tenant compte du contexte. Une vue d'ensemble des autres approches réglementaires, en Suisse et à l'étranger, se trouve en annexe.

Afin de soutenir le développement responsable des systèmes ADM, des mesures étatiques doivent être mises en place pour promouvoir les bibliothèques et socles d'application open-source destinés aux développeurs et développeuses d'ADMS et d'intelligence artificielle.

En principe, les lois existantes peuvent également s'appliquer aux systèmes ADM moyennant quelques modifications. Par exemple, la dispersion et l'utilisation des données personnelles peuvent être réglementées par la loi sur la protection des données. [11]

Les interdictions de discrimination représentent l'envers de la discussion émergente sur l'équité, [12] avec une large zone grise entre les deux [13], dans laquelle se trouvent la majorité des applications pratiques.

Le droit du travail ainsi que la législation sur la protection des données interdisent certaines pratiques de surveillance, qu'elles soient algorithmiques ou non, sur le lieu de travail. [14]

D'un point de vue purement juridique, l'adoption d'une loi spécifique aux ADMS pourrait avoir deux avantages principaux : premièrement, parce que les modifications nécessaires dans différents textes de loi nécessiteraient une définition commune des termes, des catégories et des risques et deuxièmement, parce que la définition de l'autorité de surveillance des ADMS, décrite plus loin, serait difficile à intégrer ailleurs.

Les personnes morales peuvent – tout comme les personnes physiques – tirer parti des systèmes ADM. Le spectre d'utilisation de l'intelligence artificielle est large, que ce soit dans la production, les prestations de services, etc., pour ne citer que ces exemples. À la date de cette publication, Société Numérique n'a pas connaissance d'exemples concrets de préjudice ou de discrimination envers des personnes morales. Néanmoins, Société Numérique est consciente que les personnes morales peuvent également subir des préjudices. À ce stade, il est cependant impossible d'évaluer si et dans quelle mesure cela pourrait se produire, ainsi que dans quelle mesure les lois existantes (comme le droit de la concurrence, par exemple) offriraient une protection suffisante aux personnes morales. Société Numérique se réserve la possibilité de prendre position sur ce sujet ultérieurement.

---

[11] Nous demandons une modification de l'article 21, alinéa 1 de la LPD (suppression du terme « exclusivement ») : « Le responsable du traitement informe la personne concernée de toute décision qui est prise sur la base d'un traitement de données personnelles automatisé et qui a des effets juridiques pour elle ou l'affecte de manière significative (décision individuelle automatisée).»

[12] L'équité, en lien avec les algorithmes décisionnels, implique l'évaluation et la correction des biais algorithmiques. Les résultats des algorithmes décisionnels sont considérés comme « équitables » lorsqu'ils sont indépendants de facteurs spécifiques comme le genre, l'âge, etc. Cependant, la formulation précise (mathématique) de l'équité reste un sujet de débat, certaines définitions étant même contradictoires.

[13] Alors que la discrimination, en tant qu'infraction, suppose une violation grave de l'équité, la notion d'équité parfaite ne peut généralement s'appliquer qu'à une métrique particulière, ce qui implique d'en négliger d'autres (également valides). Entre la discrimination et l'équité se trouve donc une importante zone grise.

[14] La surveillance des employés est autorisée dans une mesure limitée, comme l'enregistrement et le contrôle des heures de travail (art. 46 LTr), ou les données liées à l'aptitude pour le poste, etc. Les limites de la surveillance sont ancrées dans la protection de la personnalité (art. 328 ss CO), la protection des données dans la LPD et dans certains articles impératifs de la loi sur le travail. La surveillance systématique du comportement des employés est interdite (art. 26, al. 1 OLT 3), car elle peut avoir des répercussions sur la santé des travailleurs. Des exceptions peuvent être admises (art. 26, al. 2 OLT 3) pour certaines raisons, comme l'optimisation des performances ou l'assurance qualité, mais uniquement si la proportionnalité est respectée et si l'atteinte à la personnalité et à la santé est réduite au minimum (appréciation au cas par cas) (voir Bürgi et Nägeli, 2022).

5.1 L'autorité de surveillance des ADMS

L'**autorité de surveillance des ADMS** doit agir comme un **centre de compétence**. Elle conseille les entreprises, les administrations publiques et le grand public, tout en coordonnant d'éventuelles analyses à long terme. Elle recueille les plaintes des personnes concernées et procède, en cas de soupçon fondé, et indépendamment de toute directive, à des contrôles pour vérifier la conformité avec la réglementation et la classification des systèmes ADM publics et privés.

Cette autorité de surveillance des ADMS doit être compétente à tous les niveaux (fédéral, cantonal et communal). Elle peut, parallèlement aux voies de recours et d'action judiciaire accessibles aux individus et aux ONG accréditées, imposer des sanctions de première instance en cas de violation. Cette autorité centralise la compétence de veiller au respect de la réglementation des ADMS ainsi que de prévenir et sanctionner les risques pour les individus et la société. Elle se voit attribuer des tâches d'utilité publique uniformes pour l'ensemble du secteur public à tous les échelons (voir section 8).

Elle agit comme centre de compétence et de formation pour l'évaluation des risques associés aux systèmes ADM au travers de check-lists et de guides de bonnes pratiques. Elle a également comme but de sensibiliser à la problématique et de promouvoir une gestion adéquate de ces systèmes. Elle doit être constituée d'experts apportant des compétences complémentaires (dans le domaines des sciences sociales, de la technologie et du droit) et pouvoir agir indépendamment de toute directive et être dotée de son propre budget, à l'instar du PFPDT. Les modalités concrètes de mise en œuvre de sa mission de surveillance, conformément aux principes énoncés pour garantir la meilleure protection possible des individus et de la société, restent cependant à définir dans le détail.

5.2 La sécurité informatique

Les systèmes automatisés, tout comme d'autres systèmes informatiques, ne sont jamais sûrs à 100 % et peuvent donc être "piratés" ou utilisés à mauvais escient. Leurs fonctions peuvent donc être potentiellement détournées, par des initiés ou par des tiers. Nous estimons toutefois que les mesures visant à empêcher de telles attaques n'ont pas leur place dans une réglementation ADMS, mais dans une "**loi sur la sécurité informatique**" générique.

La loi ADMS doit plutôt s'intéresser aux effets spécifiques des systèmes automatisés de prise de décision. La classification des risques ne doit pas seulement tenir compte de l'utilisation prévue du système, mais également intégrer les cas prévisibles d'utilisations potentiellement erronées et abusives. [15] ("reasonably foreseeable misuse", EU AI Act Art 9.2(b)). Un exemple serait l'analyse de la communication de tous les collaborateurs dans le but d'améliorer la collaboration. Un abus prévisible de ce système est la surveillance et/ou l'évaluation des collaborateurs.

Un élément important pour la fiabilité des algorithmes réside dans l'application, encore manquante, de la **responsabilité du fait des produits aux logiciels**, applicable à tous les types d'algorithmes informatiques. Il ne devrait pas être possible pour les entreprises de logiciels de se soustraire à toute responsabilité grâce à une formulation habile de leurs conditions générales d'utilisation (CGU). En raison de sa portée générale, la responsabilité du fait des produits devrait toutefois être intégrée dans une loi sur la sécurité informatique et ne pas se limiter spécifiquement aux systèmes ADM.

6 Classification

Notre proposition suit une approche hybride combinant un modèle basé sur les préjudices et un modèle basé sur les risques. Dans le cadre d'une approche basée sur les préjudices, les sanctions ne sont imposées qu'à posteriori, en cas de préjudice avéré. En revanche, dans le cadre d'une réglementation basée sur les risques, les applications sont soumises dès leur conception à des obligations. Nous suivons ici les recommandations du «Gutachten der Datenethikkommission» (Rapport de la Commission sur l'éthique des données) du gouvernement fédéral allemand (page 43 et suivantes) [16] ainsi que l'analyse détaillée des implications en matière de droits fondamentaux des

[15] Il s'agit d'une part de l'utilisation illicite ou du "piratage" de ces systèmes, mais aussi d'effets spécifiques aux systèmes ADM, tels que l'extraction via les modèles eux-mêmes de données sensibles utilisées pour leur entraînement (voir glossaire), le manque de fiabilité des prédictions lors d'utilisation de données d'entrée n'ayant pas été testées (fragilité), de données d'entrée générées de manière spécifique, semblant être correctes pour les êtres humains mais qui aboutissent à des résultats erronés (Adversarial Examples), etc.

[16] Voir Datenethikkommission der Bundesregierung 2019

technologies de reconnaissance faciale réalisée par la FRA (cf. FRA 2019).

Grâce à cette approche hybride, les applications à risques et impacts importants sont soumises dès le départ à des obligations de diligence et de transparence, tandis que les applications présentant des risques et impacts moindres relèvent d'un système d'auto-déclaration. Les éventuelles violations des obligations ou erreurs de classification sont sanctionnées *a posteriori* dans le cadre de recours ou d'actions en justice. Cette approche permet aux exploitants de systèmes ADM de développer et de déployer leurs systèmes sous leur propre responsabilité, mais dans un cadre clairement défini. Cette responsabilité est renforcée et garantie par des mécanismes de sanctions.

Nous distinguons trois catégories de systèmes ADM : « risque faible », « risque élevé » et « risque inacceptable ».
La classification des systèmes automatisés de prise de décision repose sur une évaluation des risques associés à leur utilisation, tant au niveau individuel – perspective au cas par cas – qu'au niveau sociétal. Le risque pour la société est déterminé dans le contexte des objectifs de protection sur la base des préjudices potentiels pour la société dans son ensemble et de la probabilité d'occurrence. Pour les cas individuels, le risque est analysé en fonction des effets spécifiques sur les personnes concernées. L'arbre de décision permettant de définir ces catégories est établi par le législateur, et non par d'autres parties prenantes.

Nous détaillons ces risques dans la section suivante, avant de présenter les critères d'évaluation permettant de classer les systèmes ADM. Enfin, nous présentons les catégories spécifiques de manière concrète.

6.1 Objectifs de protection et risques pour les individus et la société dans son ensemble

Lors de l'utilisation de systèmes ADM, il est essentiel de garantir la protection des individus et de la société. Les objectifs fondamentaux de protection des individus sont le respect des droits fondamentaux et des droits humains. Nous nous référons ici aux objectifs de protection définis dans notre concept de protection des données (cf. Société Numérique, 2023), qui sont principalement axés sur la protection des individus, mais incluent également des risques sociétaux qui ne découlent pas uniquement de l'accumulation de risques individuels :

- protection contre les manipulations
- protection contre la discrimination
- protection contre la surveillance et droit à l'anonymat
- protection contre les atteintes à la santé ainsi qu'aux opportunités de vie et de développement
- droit à la transparence et devoir de diligence
- droit à l'oubli
- protection de la société ouverte et de la démocratie.

La manipulation se définit comme une influence intentionnelle sur les décisions d'une autre personne, influence ciblée et généralement dissimulée, dans le but de contourner son autodétermination et sa capacité de jugement. Les ADMS permettent une communication hautement automatisée et individualisée avec les personnes. La manipulation peut entraîner des préjudices pour les individus concernés. En exploitant les faiblesses humaines, elle vise à orienter le comportement des individus ou des groupes. Les personnes vulnérables sont particulièrement exposées à ce risque.

Il y a discrimination lorsque des décisions prises par des ADMS désavantagent des individus ou des groupes en raison de caractéristiques telles que l'appartenance ethnique, la couleur de peau, le genre, l'appartenance sociale, l'orientation sexuelle, etc. En règle générale, le biais sous-jacent est déjà présent dans les données d'entraînement et est perpétué, voire amplifié, par l'automatisation des décisions. Des exemples concrets de discrimination sont fournis dans le dossier du groupe spécialisé Tracking & Profiling de Société Numérique (DE).

Lorsque la protection contre la surveillance et le droit à l'anonymat sont violés, les individus peuvent être empêchés de développer librement leur identité. Par ailleurs, des « effets dissuasifs » (*chilling effects*) peuvent apparaître : la simple possibilité d'être surveillé peut dissuader des personnes de participer à des manifestations ou à des rassemblements, avec des conséquences extrêmement négatives sur les processus décisionnels démocratiques ; c'est la raison principale pour laquelle les ONG actives dans les droits numériques s'opposent fermement à l'identification biométrique et à la reconnaissance faciale dans les espaces publics.

Les ADMS qui interviennent directement ou indirectement dans des domaines comme les services sociaux, la justice, l'application de la loi, l'éducation ou l'économie quotidienne peuvent avoir une influence considérables sur les opportunités de développement des individus (par exemple, le refus d'un accès à des

études ou à un crédit). Pour des exemples et des analyses complémentaires, nous renvoyons à nouveau au dossier du groupe spécialisé Tracking & Profiling de Société Numérique.

Le droit à la transparence peut être violé à plusieurs niveaux. D'une part, les personnes peuvent ignorer que les décisions les concernant sont prises de manière automatisée. Par exemple, selon la nouvelle loi sur la protection des données, seules les décisions entièrement automatisées doivent être signalées. D'autre part, les individus concernés n'ont généralement pas la transparence nécessaire pour savoir quelles données d'entrée, modèles et coefficients (voir glossaire) sont utilisés pour la prise de décision. Là encore, nous renvoyons au dossier du groupe spécialisé Tracking & Profiling de Société Numérique.

Le droit à l'oubli est rarement abordé en lien avec les ADMS, mais il reste pertinent. Pendant combien de temps les données passées peuvent-elles être prises en compte pour des décisions automatisées, comme dans le calcul des cotes de crédit ? La question devient particulièrement complexe lorsqu'un système ADMS a déjà été entraîné sur ces données et nécessiterait une correction.

Dans le contexte des systèmes ADM, la société ouverte et le système démocratique sont en danger lorsque des personnes ou des groupes entiers sont discriminés en raison de certaines caractéristiques (cf. objectif de protection contre la discrimination) ou lorsque les processus démocratiques sont perturbés (cf. objectif de protection contre la surveillance et droit à l'anonymat). La démocratie est également mise en péril lorsque des individus ou des groupes entiers sont manipulés, par exemple via les réseaux sociaux. Cela peut inclure la diffusion ciblée et individualisée de messages à certaines catégories de population, sans aucune transparence quant aux informations diffusées. Cela fragmente l'espace du discours public, rendant plus difficile tout débat de fond.

6.2 Critères d'évaluation

Les risques sont compris comme la combinaison de l'ampleur du préjudice possible et de la probabilité que le préjudice se produise. En résumé, risque = ampleur du préjudice * probabilité d'occurrence du préjudice. Le préjudice est évalué en fonction des objectifs de protection (cf. 6.1). La probabilité de survenue du préjudice peut résulter de plusieurs facteurs. Une possibilité d'évaluation serait par exemple la probabilité qu'une situation problématique se produise (exposition) et la probabilité que celle-ci puisse être corrigée avant la survenue du préjudice. Dans certaines circonstances, il convient également de prendre en compte l'annulation du préjudice, par exemple un transfert ultérieur de fonds après leur blocage initial.

Pour la réglementation, les systèmes ADM doivent être classés en fonction de leurs risques. Pour la classification, nous reprenons et complétons certains concepts de l'AI Act de la Commission européenne (art. 7.2). Lors de l'attribution d'un système ADM à une catégorie de risque, les aspects suivants doivent être pris en compte :

- Quels sont l'objectif et le domaine d'application du système ADM ?
- À quelle échelle le système ADM sera-t-il (vraisemblablement) utilisé (ponctuellement ou à grande échelle) ?
- Dans quelle mesure l'utilisation du système ADM a-t-elle déjà porté atteinte à la santé, à la sécurité ou aux droits fondamentaux ? Y a-t-il des raisons de s'inquiéter sérieusement de la survenue de tels dommages, de tels préjudices ou de tels effets négatifs, sur la base de rapports ou d'affirmations documentées qui devraient être transmis aux autorités compétentes ?
- Quelle est l'ampleur potentielle de ces dommages, de ces préjudices ou de ces effets négatifs, notamment en termes d'intensité et de capacité à affecter un grand nombre de personnes ?
- Dans quelle mesure les personnes potentiellement lésées ou affectées dépendent-elles du résultat produit par un système ADM et dans quelle mesure dépendent-elles du système ADM, notamment parce qu'il est raisonnablement impossible de se soustraire à l'utilisation du système ADM pour des raisons pratiques ou légales ?
- Dans quelle mesure les personnes potentiellement lésées ou affectées sont-elles vulnérables face à l'entité qui met en place un système ADM, notamment en raison d'un déséquilibre en termes de pouvoir, de connaissances, de situation économique ou sociale, ou d'âge ?
- Dans quelle mesure et avec quelle facilité un résultat obtenu avec un système ADM peut-il être annulé ? Les résultats qui ont un effet sur la santé ou la sécurité des personnes ne peuvent pas être considérés comme facilement réversibles.

- Des boucles de rétroaction destructrices ou qui se renforcent d'elles-mêmes peuvent-elles se former et quelles sont les mesures prises pour les éviter ?

Si l'application d'un système ADM à un individu peut être évitée par ce dernier, en supposant des connaissances moyennes et des circonstances normales, et sans accepter d'inconvénients, ce système est classé dans une catégorie inférieure à celle où un individu est dépendant de ce système. Si l'effet d'une décision automatisée peut (facilement) être annulé (ou compensé), ce système est également classé dans une catégorie inférieure. Il faut pour cela que les individus aient connaissance non seulement de la décision automatisée en tant que telle, mais aussi de la possibilité de faire appel et de l'annuler, et que cette annulation puisse être demandée avec un niveau de connaissance moyen et dans les meilleurs délais, sans que cela n'entraîne de désavantages.

6.3 Les catégories de risque

Si un système technique concret entre dans le champ d'application de la présente loi (c'est-à-dire s'il s'agit d'un système ADM au sens de la section 2), il doit être classé dans l'une des trois catégories suivantes : **”risque faible”**, **”risque élevé”** et **”risque inacceptable”**. Les obligations de diligence et de transparence mentionnées ci-dessous ne s'appliquent qu'aux systèmes à ”risque élevé”.

Cette classification tient compte des risques au regard des objectifs de protection définis à la section 6.1, et utilise les critères d'évaluation définis à la section 6.2. L'entité qui utilise un système ADM détermine de manière autodéclarative s'il s'agit d'un système ADM et quel est le risque qui y est lié. L'objectif est de réduire autant que possible la charge administrative. En cas d'auto-déclaration erronée ou spécifiant une catégorie trop basse, des sanctions administratives élevées et proportionnelles au chiffre d'affaires peuvent être prises en fonction du degré de culpabilité. L'évaluation reviendra donc en fin de compte aux tribunaux.

Afin de ne pas freiner l'innovation, il est également important que la sécurité juridique soit garantie en matière d'autodéclaration. Les sanctions administratives ne doivent pas toucher les entreprises qui effectuent leur autodéclaration avec soin et en toute bonne foi. Ceci doit être pris en compte dans les actions de l'autorité de surveillance des ADMS, notamment par le biais de notices explicatives qui précisent les critères d'autodéclaration et donnent des exemples (cf. par exemple dans le droit de la protection des données : PFPDT 2023).

Cette classification basée sur les risques est compatible avec la loi sur l'IA de l'Union européenne (EU AI Act) dont la formulation est basée sur les applications. Toutefois, contrairement à cette loi, nous n'interdisons pas fondamentalement les applications, mais les examinons à la lumière des circonstances respectives. Ainsi, si la reconnaissance algorithmique des émotions peut être interdite lors des entretiens d'embauche, elle peut être utilisée dans le cadre d'une exposition artistique, car le risque pour la société et les individus est faible dans ce second cas. Le risque lié aux systèmes automatisés de prise de décision et, par conséquent, leur évaluation peuvent également évoluer au fil du temps, avec le développement de la technologie et de la société, ainsi qu'en cas d'interaction avec d'autres systèmes. Le schéma de classification proposé peut refléter ces évolutions.

La catégorie la plus basse (« risque faible ») comprend les systèmes

- qui présentent un risque faible pour la société
- dans le cas des individus, qui ne présentent aucun risque ou seulement un risque faible au regard des objectifs de protection

Les systèmes de cette catégorie se caractérisent donc par le fait qu'ils n'ont, selon toute vraisemblance, aucun effet négatif particulier sur les individus ou la société. Si des préjudices modérés ou des atteintes aux droits fondamentaux sont possibles, mais que le système peut être facilement évité sans connaissances particulières ou que les effets néfastes peuvent être facilement annulés, un système peut néanmoins être classé dans cette catégorie. Les systèmes dont les décisions peuvent avoir des effets néfastes sur la santé et/ou la sécurité des personnes ne peuvent en principe pas être classés dans cette catégorie.

Parmi les exemples de systèmes ADM appartenant à cette catégorie, on peut citer le contrôle automatique de l'intégrité des emballages alimentaires immédiatement après la production, ou les systèmes ADM de prévision de la concentration de pollen, qui sont d'une grande aide pour les personnes allergiques, mais dont un dysfonctionnement n’a que des conséquences négligeables.

La catégorie intermédiaire (« risque élevé ») comprend les systèmes

- qui présentent un risque élevé pour la société
- qui présentent pour les individus un risque élevé au regard des objectifs de protection.

Les systèmes de cette catégorie se caractérisent par le fait que leur utilité (positive) est contreba-

lancée par un potentiel négatif significatif. Le potentiel de dommages est toutefois encore acceptable (ou atténuable), sinon un tel système serait classé dans la catégorie supérieure. Les systèmes de cette catégorie sont généralement largement utilisés (et non pas de manière isolée) et ne laissent aux individus aucune possibilité d'échapper à la prise de décision automatisée. Dans cette catégorie, les dommages causés par les décisions ne peuvent pas non plus être annulés ou compensés en pratique.

Les systèmes ADM qui recommandent des contenus (qu'il s'agisse d'actualités comme dans le cas des algorithmes de fil d'actualité, de vidéos dans le cas des algorithmes de recommandation ou de contenus généraux comme dans le cas des moteurs de recherche) appartiennent à la catégorie « risque élevé » pour plusieurs raisons : ils concernent de grands groupes d'utilisateurs (potentiellement l'ensemble de la société), ils influencent la perception de leurs utilisateurs et il est prouvé qu'ils peuvent contribuer à la radicalisation (cf. Tufekci 2018, Frenkel et Kang 2021).

Les décisions individuelles dans les affaires sociales (par exemple, les évaluations d'éligibilité) sont à haut risque pour plusieurs raisons : elles concernent généralement des personnes vulnérables, elles ne peuvent être évitées/contournées et les personnes concernées ne peuvent généralement pas obtenir la correction d'une décision erronée sans recourir à des procédures judiciaires complexes et coûteuses. Les décisions qui conduisent à l'embauche ou à la sélection de personnes dans le cadre de procédures de candidature à un emploi présentent également un risque élevé, car elles ne peuvent être contournées par les personnes concernées, mais influencent leurs chances dans la vie et leur développement.

Il existe également des systèmes susceptibles d'avoir des conséquences graves et irréversibles sur les individus, par exemple dans le domaine du diagnostic médical. Ceux-ci seraient donc classés comme « inacceptables ». Tant que ces systèmes sont utilisés à des fins d'assistance et que la décision finale est prise par un spécialiste, il sensé de les reclasser dans la catégorie « risque élevé ». Cependant, la transition entre les systèmes de recommandation fréquemment utilisés comme systèmes d'aide et l'acceptation sans réserve de ces propositions est fluide, ce qui pourrait transformer de fait les systèmes initialement d'assistance en décideurs.

La catégorie la plus élevée (« risque inacceptable ») comprend les systèmes

- qui présentent un risque inacceptable pour la société dans son ensemble
- pour les individus, qui présentent un risque inacceptable au regard des objectifs de protection ou ont des conséquences graves et irréversibles.

Cette catégorie comprend donc les systèmes dont les conséquences négatives potentielles sont si importantes qu'on ne peut risquer de les utiliser. Pour de nombreux systèmes de cette catégorie, les dommages sont en outre connus et documentés, et ne sont donc plus potentiels, mais se produisent de manière fiable. Les décisions prises dans cette catégorie ne peuvent en outre être ni contournées (par exemple, la surveillance biométrique de masse) ni révisées, et elles sont souvent impossibles à vérifier (par exemple, les décisions automatisées/assistées en matière d'asile, de probation ou de justice). Les dommages craints ou avérés pour les individus et la société sont si importants dans cette catégorie que les risques ne peuvent être acceptés ni atténués. L'utilisation de tels systèmes est interdite.

La surveillance biométrique de masse déjà mentionnée ci-dessus (y compris la reconnaissance faciale [17]), qui non seulement constitue une atteinte massive aux droits fondamentaux tels que la dignité humaine, l'autonomie et la vie privée, mais a également des effets dissuasifs (cf. Assion 2014 et Penney 2016) sur les processus démocratiques et donc sur la société. Un autre exemple est l'évaluation automatisée des comportements (social scoring), qui concerne certes en premier lieu les individus, mais qui, par ses effets de contrôle et de modelage des comportements, peut avoir des répercussions considérables (et de surcroît non légitimées démocratiquement) sur la société.

Outre les décisions en matière d'asile, de probation ou de justice, nous constatons également des effets inacceptables pour les individus dans le cadre de la surveillance des employés, des élèves et des étu-

[17] par exemple https://gesichtserkennung-stoppen.ch, https://reclaimyourface.eu

[18] pour ces raisons, il a déjà été demandé que la surveillance et la gestion automatisée dans les contextes professionnels et éducatifs soient ajoutées à la liste des applications à interdire (cf. EDRi 2021) ; voir Crawford et al. pour une critique des techniques de reconnaissance automatisée des émotions (cf. Crawford 2021).

diants. Une évaluation automatisée à grande échelle sur le lieu de travail et les décisions de licenciement ou d'optimisation qui en découlent peuvent nuire de manière inacceptable à la santé physique des employés. [18]

7 Obligations de diligence et de transparence

En principe, l'entité qui, du point de vue des personnes concernées, utilise le système ADM (par exemple, l'entreprise exploitante) est responsable de son fonctionnement et de son classement correct dans les catégories de risque susmentionnées. S'il doit être possible de transférer certains risques commerciaux aux fabricants de composants ou de systèmes en vertu du droit civil, il convient toutefois d'empêcher (via la responsabilité du fait des produits mentionnée plus haut dans la loi distincte sur la sécurité informatique) que les fabricants puissent se dégager de toute responsabilité au moyen des conditions générales d'utilisation, comme c'est actuellement la pratique courante dans les contrats d'utilisation de logiciels.

Nous considérons qu'une obligation de certification n'est utile que dans des domaines d'application spécifiques et privés ayant un usage spécifique et facile à normaliser [19], comme c'est le cas pour les produits médicaux, par exemple un défibrillateur automatisé externe (DAE). Sinon, le risque est grand de voir la responsabilité transférée à grande échelle aux organismes de certification.

Une possibilité de faciliter le travail de diligence serait de réaliser des analyses d'impact qui mettent en lumière les conséquences possibles du développement et de l'utilisation d'un système ADM, jettent les bases de stratégies de réduction des risques mûrement réfléchies et servent d'indicateur d'une utilisation responsable de la technologie. Au moment de la rédaction du présent document, Société Numérique ne voit toutefois aucun avantage direct à rendre ces instruments obligatoires.

Le classement dans une catégorie de risque doit être documenté par l'exploitant. Pour les systèmes ADM présentant un « risque faible », une classification informelle mais justifiée de manière compréhensible suffit. Pour les systèmes ADM présentant un « risque élevé », une analyse systématique est nécessaire, qui peut par exemple être effectuée dans le cadre d'un processus de gestion des risques. Ce processus devant de toute façon exister pour certaines catégories de produits, tels que les dispositifs médicaux.

Les obligations de transparence suivantes s'appliquent uniquement aux systèmes ADM de la catégorie « risque élevé », les systèmes à risque « inacceptable » ne devant pas être utilisés de toute façon. Une déclaration erronée ou trop basse est passible de sanctions sévères. Le degré des obligations de transparence doit permettre d'évaluer les risques de chaque système, mais aussi fournir suffisamment d'informations pour évaluer l'impact de l'ensemble de l'écosystème ADMS. Nous recommandons donc des formats de rapports de transparence standardisés.

Nous faisons la distinction entre les obligations applicables aux systèmes utilisés dans le **secteur privé** et celles applicables aux systèmes utilisés **dans le cadre d'une mission de service public**. En règle générale, tous les systèmes (privés et publics) classés dans la catégorie « risque élevé » sont soumis à une obligation **d'étiquetage et d'information**, qui indique

1. qu'un système ADM est utilisé, [20]
2. un bref résumé de l'objectif du système et, concrètement, de ses sorties possibles,
3. des informations sur l'origine des données et des explications sur les features (voir glossaire) spécifiques utilisées par le système ADM et ce qu'elles représentent.

Les informations sur l'origine des données doivent également servir à garantir une qualité suffisante des données et à rendre visible l'**interconnexion** de plusieurs systèmes ADM (éventuellement de différents fabricants). En outre, nous

[19] Cela signifie un fonctionnement clairement vérifiable du système prenant des décisions de manière autonome.

[20] avec adaptation de l'art. 21, al. 1, nLPD (suppression de « exclusivement »)

[21] il s'agit de systèmes qui adaptent en permanence leur fonctionnement en fonction de nouvelles entrées. Ce mode de fonctionnement entraîne une multitude de problèmes, par exemple le fait que les systèmes désapprennent les garanties précédemment certifiées telles que « l'équité », ou qu'ils peuvent être alimentés de manière intentionnelle et difficilement détectable avec des données manipulées afin de modifier leur fonctionnement à l'avantage de l'attaquant.

[22] c'est-à-dire qu'elles sont statistiquement représentatives (par rapport à l'objectif et au domaine d'application du système), exactes, complètes et aussi cohérentes que possible, et qu'elles suivent une sémantique connue.

exigeons un contrôle périodique et continu du risque (c'est-à-dire du classement dans une catégorie particulière) ainsi que de la documentation relative aux obligations de transparence, en particulier pour les systèmes d'auto-apprentissage. [21]

La **qualité des données** mentionnée dans la section précédente signifie que les données utilisées doivent soit correspondre à la réalité, [22] assurant ainsi que les systèmes basés sur ces données fonctionnent de manière aussi irréprochable que possible, ou qu'elles n'ont été modifiées que dans des aspects intentionnels et généralement utiles, par exemple pour prévenir la discrimination.

En ce qui concerne l'**origine des données**, le droit à l'autodétermination informationnelle doit être respecté ; cela vaut également pour les données provenant de l'étranger. Les données doivent provenir de sources éthiquement acceptables. Par exemple, l'utilisation de données obtenues illégalement doit être systématiquement évitée. [23]

En outre, l'utilisation des sorties d'autres systèmes ADM doit être clairement indiquée. Cela vise à créer une transparence quant à l'**interconnexion** de ces systèmes, qui crée ses propres risques en raison de la complexité croissante prévisible des interactions, du flux d'informations (opaque) et des boucles de rétroaction qui en résultent.

Nous précisons ci-après les obligations de transparence applicables aux secteurs privés et publics.

7.1 Secteur privé

Pour les entités qui utilisent des systèmes dans le contexte du secteur privé, nous exigeons des informations sur l'origine de toutes les données utilisées ainsi que sur leur qualité et leur exhaustivité au regard de la finalité du système ADM. Cela inclut toutes les données utilisées pour la mise en place, l'entraînement, la validation, la prédiction, etc. du système. Cela comprend également la documentation relative à la finalité du système et des informations pertinentes sur les features utilisées comme données d'entrée afin de pouvoir évaluer l'impact sur les individus et la société ou le risque au regard des objectifs de protection, en particulier la santé, la sécurité ou les droits fondamentaux des individus ou de la société.

Une réglementation formelle peut prévoir, après avoir pesé les risques et les avantages, des exceptions aux obligations de transparence et à la responsabilité pour les produits standardisables, si elle crée en même temps un organisme de certification qui satisfait à des exigences de qualité au moins équivalentes à celles mentionnées ci-dessus (par exemple dans le domaine du diagnostic médical).

7.2 Dans l'accomplissement d'une mission de service public

Pour les acteurs publics, les réglementations existantes, telles que le droit public et le code de procédure pénale, imposent déjà des normes plus strictes en matière de traçabilité et de transparence. Ces normes s'appliquent également à titre minimal à l'utilisation des systèmes ADM. Outre les mêmes obligations de transparence que pour les systèmes ADM privés (informations sur l'origine et la qualité des données, les features et la finalité), nous demandons que les systèmes ADM des acteurs publics publient les coefficients des modèles (voir glossaire) dans un format standardisé [24] comme suit :

- Pour les systèmes ADM basés sur des données non personnelles, les données doivent être mises à disposition sous forme de données ouvertes (OpenData) dans la mesure du possible, avec les coefficients des modèles.
- Pour les systèmes ADM basés sur des données personnelles ou des données non personnelles qui ne peuvent être publiées, les règles suivantes s'appliquent : ils doivent soit
  - être entraînés sur des données synthétisées (ensemble de données synthétiques, voir glossaire), et celles-ci doivent être publiées avec les coefficients ;
  - ni les données ni les coefficients ne sont publiés si (dans des cas exceptionnels) la production de données synthétisées et l'utilisation de systèmes ADM correspondants entraînent des coûts disproportionnés ou si leurs coefficients permettent de déduire des données personnelles ou des données non personnelles qui ne peuvent être pu-

[23] Ou plutôt, leur utilisation n'est justifiée que dans certaines circonstances, après avoir pesé le pour et le contre d'un point de vue éthique (cf. Imhasly 2021).

[24] cela facilite l'évaluation automatique

bliées. Dans ce cas, toutefois, l'autorité de surveillance des ADMS et les ONG autorisées doivent avoir accès aux données afin de vérifier leur portée pour les individus et la société, ainsi que le risque pour la santé, la sécurité ou les droits fondamentaux des individus ou de la société.

L'obligation générale de divulgation exigée ici correspond à la revendication « Public Money ? Public Code ! » (Argent public ? Code public !), qui exige la publication du code source des logiciels financés par des fonds publics. Cela permet à d'autres autorités ou au public d'utiliser et de développer ces solutions.

Les lignes directrices de la Confédération en matière d'intelligence artificielle et leur suivi montrent l'importance de ce thème pour l'administration fédérale.

8 Contrôle, mesures et sanctions

Les violations des obligations de diligence et de transparence susmentionnées doivent être sanctionnées efficacement. Là encore, nous faisons la distinction entre le secteur privé et le secteur public. Dans les deux cas, le respect des obligations de diligence et de transparence est contrôlé d'une part par des individus et d'autre part par des associations habilitées (ONG) qui peuvent déposer plainte ou intenter une action en justice en cas de préjudice. Les associations doivent être habilitées à déposer plainte si elles sont actives dans toute la Suisse et si leurs statuts prévoient un objectif correspondant. Les possibilités de contrôle, les mesures et les sanctions ainsi que les voies de recours doivent être conçues de manière à garantir la meilleure protection possible aux personnes concernées ; si nécessaire, elles doivent également être complétées ou remaniées. Cela inclut également la révision et l'amélioration des moyens collectifs de faire respecter le droit. [25]

L'autorité de surveillance des ADMS doit pouvoir enquêter d'office sur les violations de la réglementation et rendre des décisions formelles. Elle peut exiger l'accès aux données et prononcer des sanctions. Afin de garantir la meilleure protection possible aux personnes concernées, les actes intentionnels ainsi que négligents doivent être punissables. Nous espérons que les systèmes douteux seront rapidement signalés afin d'attirer l'attention de la société civile et, par conséquent, des associations concernées ou de l'autorité de surveillance des ADMS.

Il convient de prévenir les fausses classifications des systèmes ADM, par exemple celles qui utilisent la catégorie de risque faible alors que le risque élevé est justifié, et les violations des obligations de diligence et de transparence qui en découlent, en prévoyant des sanctions suffisamment sévères. C'est sous cette condition que nous considérons que l'obligation d'autodéclaration proposée pour éviter les processus bureaucratiques et soulager les entreprises est suffisante. Nous attendons des entreprises qu'elles mettent en œuvre de manière autonome des règles analogues à celles relatives à la protection des données, par exemple en créant des points de contact internes en cas de violations présumées ou de déclarations incorrectes.

8.1 Secteur privé

La recherche d'une culpabilité individuelle ne semble pas pertinente, car les violations des obligations de diligence et de transparence susmentionnées relèvent généralement d'une faute au niveau de l'entreprise. L'autorité de surveillance des ADMS doit donc sanctionner les entreprises par des sanctions administratives et non les individus par des sanctions pénales. Cela évite également le risque de « rejeter » la faute sur des « boucs émissaires ». En outre, le barème des sanctions doit être fonction du chiffre d'affaires afin que les grandes entreprises ne puissent s'en tirer à bon compte. Les sanctions doivent être suffisamment élevées pour que les violations des obligations de diligence et de transparence ne soient pas considérées comme un risque commercial quotidien et donc « budgétisées ».

Une sous-estimation de la catégorie du système ADM par l'entité responsable est punissable.

L'autorité de surveillance des ADMS dispose notamment des instruments suivants : elle recueille les plaintes, peut exiger l'accès aux données et prononcer des sanctions et des décisions. L'évaluation finale incombe aux tribunaux.

[25] Au moment de la publication du présent document, l'introduction de moyens collectifs généraux de faire respecter le droit dans le code de procédure civile (CPC) est en cours de débat au Parlement. Une inscription dans le CPC serait certes bienvenue, mais nous demandons l'introduction de moyens collectifs de faire respecter le droit indépendamment de l'issue de ces délibérations.

En cas d'insuffisance présumée, nous envisageons les voies de recours suivantes :

1. Les personnes concernées peuvent intenter des actions en justice contre des entités du secteur privé et former des recours contre les décisions de l'autorité de surveillance des ADMS si celles-ci sont jugées insuffisantes. Les recours et plaintes collectives doivent également être expressément autorisés. Les voies de recours doivent être adaptées en ce sens.
2. Les associations habilitées (à l'échelle nationale et dont les statuts prévoient un objectif approprié) doivent pouvoir intenter une action contre des entités privées et former des recours contre les décisions de l'autorité de surveillance des ADMS sans être personnellement concernées (droit de recours ou d'action des associations). Compte tenu des frais de procédure élevés et à titre de mesure compensatoire, l'association concernée peut recevoir une partie du montant de la sanction à titre d'indemnisation.

En cas de condamnation possible, l'intérêt des accusés à dissimuler la vérité entraîne un fort déséquilibre des pouvoirs. En cas d'accusation grave, c'est-à-dire si un tribunal reconnaît la recevabilité de la plainte ou de l'action en justice, nous demandons donc le **renversement de la charge de la preuve**, de sorte que les entités accusées (exploitants des systèmes ADM) soit tenues de prouver de manière suffisante qu'elles n'ont pas enfreint les règles de classification, les obligations de diligence ou de transparence. Ce renversement de la charge de la preuve est l'une des contreparties à la confiance accordée via l'auto-déclaration ADM des entreprises.

Comme pour les produits technologiques et logiciels similaires, l'exploitant doit pouvoir demander des dommages-intérêts pour les dommages imputables à ses fournisseurs, tels que les développeurs ou les exploitants de systèmes. Toutefois, l'exploitant reste toujours responsable envers les personnes concernées.

8.2 Dans l'accomplissement d'une mission de service public

En principe, les mêmes contrôles, mesures et sanctions que ceux applicables aux entités privées doivent être possibles. Il reste à clarifier les modalités détaillées de la relation entre l'autorité de surveillance des ADMS et les entités chargées d'une mission des service public au niveau cantonal et communal.

Tant les particuliers que les associations doivent avoir la possibilité (comme le prévoit la section 8.1 consacré au secteur privé) d'agir contre les risques liés aux systèmes ADM et les sorties qu'ils génèrent. Afin d'éviter d'éventuels conflits de compétence, l'autorité de surveillance des ADMS pourrait, par exemple, toujours bénéficier de droits de partie dans la procédure cantonale lorsque des entités chargées d'une mission de service public au niveau communal ou cantonal sont concernées.

Ici aussi, comme c'est le cas pour les produits technologiques et logiciels similaires, l'autorité en tant qu'exploitante et mandante doit pouvoir demander des dommages-intérêts pour les préjudices imputables à ses fournisseurs, tels que les développeurs ou les exploitants de systèmes. L'autorité reste toutefois directement responsable envers les personnes concernées.

9 Quelques suggestions pour l'avenir

À l'avenir, les systèmes automatisés de prise de décision prendront en charge de plus en plus de tâches, de travaux et de fonctions, ce qui pourra entraîner de nouvelles conséquences, opportunités, défis et problèmes. Dans ce qui suit, nous souhaitons donc aborder, dans le cadre d'une évaluation des conséquences technologiques, différents points qui pourraient nécessiter une intervention réglementaire.

Le premier point concerne la **question du pouvoir** : qui crée, détermine et contrôle les systèmes, algorithmes et métriques utilisés ? Les personnes et organisations concernées ont une forte influence sur la perception et les possibilités de notre environnement naturel et social. Il convient donc de rester attentif aux conséquences futures de cette influence.

Le deuxième point concerne la **mise en réseau et l'interconnexion** de systèmes automatisés de grande envergure : dans un avenir proche, la sortie d'un système pourrait en partie constituer l'entrée d'un autre système, qui à son tour pourrait avoir une influence sur le premier système. Cela peut entraîner des effets de rétroaction complexes et multiformes, et donc des risques difficiles à évaluer, en particulier avec plus de deux systèmes. L'opacité partielle prévisible des systèmes interconnectés et l'imprévisibilité des effets qui en découle rendront nécessaire une réflexion à ce sujet. Une solution potentielle serait une modularisation claire des systèmes, de sorte que leur fonctionnement interne puisse être réduit à une abstraction simple, suffisante pour évaluer les conséquences de la rétroaction. On pourrait également envisager d'interdire le couplage des systèmes

à partir d'une certaine taille de cluster ou lorsque certains critères de sécurité ou d'objectifs ne sont plus remplis.

Dans divers cercles de discussion, on trouve la vision selon laquelle tous les problèmes sociaux et personnels peuvent être résolus avec plus de données et de meilleurs algorithmes, à condition de les autoriser. Cette vision du monde tente de réduire la réalité dans son ensemble à un ensemble de formules et de chiffres. Convaincus qu'il n'existe pas d'objectivité absolue et que, par conséquent, toutes les métriques, mesures et indicateurs ainsi que leur interprétation font l'objet de processus de négociation sociale, nous considérons cette voie comme trompeuse. Nous recommandons donc la **minimisation des données comme principe de base**, car, comme pour la protection des données, cela permet de réduire les problèmes à la source.

En outre, une **dépendance vis-à-vis des systèmes ADM** est prévisible. L'utilisation de l'automatisation pour faciliter et réduire la charge de travail permet d'accomplir des tâches plus nombreuses et plus complexes en moins de temps. Cependant, nous devons être conscients de ce que signifierait pour nous une panne de ces systèmes automatisés, de sa portée et des risques qui y sont associés, et préparer des stratégies d'urgence pouvant être mises en œuvre rapidement. En raison de l'interconnexion croissante et de la dépendance à l'égard de ressources individuelles, comme dans le cas d'Internet, le risque que de nombreuses fonctions tombent en panne simultanément augmente également. Il serait judicieux de discuter de mesures visant à créer des systèmes entièrement redondants.

Dans le même temps, on peut également entrevoir une perte potentielle **de compétences humaines** et une **perte de responsabilité**. Le transfert de tâches à des systèmes automatisés, le fait de se fier à leur exécution correcte et les effets d'accoutumance qui en découlent pourraient conduire à une perte des compétences qui seraient nécessaires sans ces systèmes et à une diminution de la responsabilité individuelle ou collective. Ces effets pourraient ne se manifester qu'au bout de plusieurs générations, par exemple lorsque certaines compétences ne seront plus transmises.

Les emplois simples ne nécessitant pas de longue formation vont progressivement disparaître. Cela peut avoir des conséquences sociales considérables, y compris en Suisse, car le statut social et le travail sont étroitement liés. Nous devons réfléchir à notre conception de la valeur sociale et, par conséquent, à la répartition des richesses, et éventuellement la redéfinir à long terme afin que **tout le monde puisse profiter des avantages de l'automatisation**.

Enfin, nous tenons à souligner l'importance des **évaluations continues des conséquences technologiques**, telles que celles réalisées par TA-Swiss pour le compte de la Confédération, entre autres organisations. L'objectif général de l'évaluation des conséquences technologiques est de fournir une analyse et une évaluation systématiques des effets et des conséquences des technologies dans tous les domaines concernés de l'environnement naturel et social.

A Propositions de réglementations pour les systèmes automatisés de prise de décision (ADM), l'intelligence artificielle et les algorithmes

Dans le cadre de la numérisation, les systèmes informatiques (basés sur les données) se sont répandus de manière pratiquement non réglementée (à l'exception des lois européennes sur la protection des données, principalement). Selon l'aspect mis en avant, ces systèmes sont désignés comme des « systèmes décisionnels automatisés », des « systèmes algorithmiques », des systèmes d'« intelligence artificielle » (« Artificial Intelligence »), ou encore des systèmes « Big Data ». Malgré leurs différences, ils ont en commun de traiter et d'analyser de très grandes quantités de données dans le but d'automatiser des décisions et/ou des processus.

Au cours des dernières années, un large consensus s'est dégagé : ces systèmes doivent être réglementés afin d'éviter leurs impacts et risques négatifs les plus graves. La demande d'une réglementation ne provient d'ailleurs pas seulement de la société civile, mais aussi des milieux politiques, économiques et scientifiques.

Par exemple, l'ACM (Association for Computing Machinery) a déjà élaboré en 2017 une prise de position sur la transparence et la responsabilité des algorithmes (cf. ACM 2017) et l'a mise à jour en 2022 (cf. ACM 2022). L'ACM est l'association professionnelle des informaticien·ne·s américain·e·s et donc l'organisation à laquelle appartiennent bon nombre des personnes qui sont à la pointe de la recherche, du développement et de l'utilisation des systèmes ADM. Bon nombre des déclarations et principes contenus dans cette position, par exemple en matière de transparence et de données, se retrouvent également dans notre proposition.

Le monde économique ne cesse également d'appeler les responsables politiques à réglementer ces systèmes. La déclaration faite en 2018 par Brad Smith, président de Microsoft, est particulièrement impressionnante. Il y souligne que la reconnaissance faciale doit être réglementée en raison de son potentiel dystopique et des dangers qu'elle représente pour la démocratie (cf. Smith 2018).

De nombreuses approches et propositions existantes (cf. DEK 2019, EU AI Act 2021) suivent une approche fondée sur les risques, qui tente, comme notre proposition, de classer les systèmes ADM en catégories en fonction de leur risque intrinsèque et de définir des règles plus strictes pour les catégories de risques les plus élevés. Le nombre de catégories varie selon les propositions. Cependant, il existe toujours une catégorie la plus basse (sans risque ou à faible risque) avec très peu de règles ou d'exigences, ainsi qu'une catégorie de systèmes ADM classés comme très risqués, dont l'utilisation est interdite. Dans la proposition de la Commission d'éthique des données, par exemple, la pyramide des risques a été développée.

La proposition de la Commission européenne « AI Act » (cf. EU AI Act 2024), adoptée en juin 2024, suit également cette approche fondée sur les risques. Contrairement à notre proposition, l'AI Act contient des listes concrètes d'applications interdites (telles que « l'identification biométrique à distance ») et à haut risque. Depuis sa publication, l'AI Act fait l'objet d'intenses discussions. Si la nécessité et la pertinence de cette proposition ainsi que son approche générale fondée sur les risques font l'objet d'un large consensus, la société civile formule également des critiques détaillées (ainsi, Société Numérique, en collaboration avec une large alliance du réseau EDRi, demande notamment un élargissement des catégories interdites et à haut risque ou la suppression des exceptions, en particulier dans le domaine de l'identification biométrique, cf. EDRi 2021). Les organisations de protection des consommateurs émettent également des critiques similaires (cf. VZBV 2021).

L'« AI Bill of Rights » (White House Office of Science and Technology Policy 2022) aux États-Unis n'est pas une tentative de réglementation à proprement parler, mais formule et renforce les droits fondamentaux dans le contexte de l'intelligence artificielle, tels que le droit à la protection contre la discrimination algorithmique, le droit à la transparence et à l'explication ou le droit à l'intervention humaine. Bien que le projet de loi porte le titre « AI », bon nombre de ses dispositions s'appliquent en particulier aux systèmes ADM. L'Accountability Act est une autre proposition de loi américaine. Le champ d'application de cette proposition concerne les décisions critiques concernant les consommateurs, par exemple dans les domaines de l'éducation, du travail ou de la santé. La proposition vise à minimiser les effets négatifs sur les consommateurs et propose divers droits pour ces derniers, tels

que l'obligation d'étiqueter les systèmes ADM, des possibilités d'opt-out, d'opposition et de rectification.

En République populaire de Chine, il existe également des propositions concernant l'utilisation de l'IA et des systèmes ADM (cf. National New Generation Artificial Intelligence Governance Specialist Committee). Cette proposition formule des directives éthiques dans le domaine de l'IA. Ces directives sont regroupées en normes fondamentales telles que la promotion du bien-être humain, la promotion de l'équité et de la justice, la protection de la vie privée ou le renforcement de la responsabilité.

L'AI Now Institute a compilé toute une série de « cas d'utilisation » publics pour la ville de New York (cf. AI Now Institute 2018), dont beaucoup sont également transposables à la Suisse. Le rapport contient également des références complémentaires et des exemples de cas.

B Bibliographie

B.1 Sources concernant les propositions de réglementation des systèmes ADM dans le contexte européen et intercontinental

- ACM (2017). Statement on Algorithmic Transparency and Accountability (Consulté le 13.0.7.2024). *ACM U.S. Public Policy Council.*
- ACM (2022). Statement on Principles for Responsible Algorithmic Systems (Consulté le 13.0.7.2024). *Europe/U.S. Public Policy Council.*
- OFCOM (2022). Monitoring der Leitlinien «Künstliche Intelligenz» für den Bund (Consulté le 13.0.7.2024). *Office fédéral de la communication OFCOM*.
- Conseil fédéral (2020). Die Leitlinien des Bundes für Künstliche Intelligenz (Consulté le 13.0.7.2024). *Le conseil fédéral*.
- CAHAI (2021). A legal framework for AI systems (Consulté le 13.0.7.2024). *Ad hoc Committee on Artificial Intelligence of the Council of Europe*.
- Datenethikkommission der Bundesregierung (2019). Gutachten der Datenethikkommission der Bundesregierung (Consulté le 16.12.2019). *Berlin : Bundesministerium des Innern, für Bau und Heimat.*
- EU AI Act (2024). REGULATION (EU) 2024/… OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). OJ 2024 L .... (Consulté le 28.06.2024). *Le parlement européen*.
- EU Digital Service Act (2020). Vorschlag für eine Verordnung des europäischen Parlaments und des Rates COM/2020/825 vom 15. Dezember 2020 über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie. *Le parlement européen*.
- European Commission adoption consultation (2021). Artificial Intelligence Act (Consulté le 13.0.7.2024). *EDRi*.
- European Union Agency for Fundamental Rights (2019). Facial recognition technology : fundamental rights considerations in the context of law enforcement (Consulté le 08.02.2022). *European Union Agency for Fundamental Rights*.
- National New Generation Artificial Intelligence Governance Specialist Committee (2021). Ethical Norms for New Generation Artificial Intelligence Released (Consulté le 13.0.7.2024). *Center for security and emergin technology*.
- Repräsentantenhaus (2022). H.R.6580 - Algorithmic Accountability Act of 2022 (Consulté le 13.0.7.2024). *117th Congress (2021-2022) of the United States of America*.
- Richardson, Rashida et al. (2019). Confronting Black Boxes : A Shadow Report of the New York City Automated Decision System Task Force (Consulté le 13.0.7.2024). *AI Now Institute*.
- Thouvenin, Florent et al (2021). Ein Rechtsrahmen für Künstliche Intelligenz. *Digital Society Initiative Universität Zürich.*
- Verbraucherzentrale Bundesverband (2021). Artificial Intelligence needs Real-World Regulation (Consulté le 13.0.7.2024). *Verbraucherzentrale Bundesverband*.
- White House Office of Science and Technology Policy (2022). Blueprint For An AI Bill Of

Rights Making Automated Systems Work For The American People. *The White House*.

B.2 Autre sources

- AI Now Insititute (2018). Automated Decision Systems - Examples of Government Use Cases (Consulté le 08.02.2022). *AI Now Insititute*.
- Assion, Simon (2014). Überwachung und Chilling Effect. *«Überwachung und Recht», Tagungsband zur Telemediucs Sommerkonferenz 2014, epubli GmbH, Berlin.*
- Bürgi, Urs (2022). Arbeitnehmerüberwachung (Consulté 25.01.2022). *Law Media*. Urheber : Bürgi Nägeli Rechtsanwälte.
- Crawford, Kate (2021). Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence. *Yale University Press, New Haven, MA*
- Société Numérique (2023). Datenschutz-Konzept der Digitalen Gesellschaft (Consulté le 24.06.2024). *Société Numérique*. Présenté au festival de la protection des données du 03.11.2023.
- Préposé fédéral à la protection des données et à la transparence (PFPDT) (2023) Aide-mémoire sur l'analyse d'impact relative à la protection des données personnelles visée aux art. 22 et 23 LPD (Consulté 2023) *Préposé fédéral à la protection des données et à la transparence (PFPDT).*
- Ensign, Danielle, et al. (2018). Runaway Feedback Loops in Predictive Policing. *FAT 2018,* 160-171.
- Eser Davolio, M. et al. (2020). Auswirkungen der Falllastreduktion in der Sozialhilfe auf die Ablösequote und Fallkosten : Entschleunigung zahlt sich aus (Consulté le 13.07.2024). *Schweizerische Zeitschrift für Soziale Arbeit*, 2019(25), pp. 31–51.
- Fanta, Alexander (2020). Datenschutzbehörde stoppt Jobcenter-Algorithmus (Consulté le 7.8.2022). Netzpolitik.org, 21.8.2020.
- Food and Drug Administration (2023). Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA (Consulté le 28.10.2023). fda.gov.
- FRA (2019). Technologie de reconnaissance faciale : considérations relatives aux droits fondamentaux dans le contexte de l'application de la loi. fra.europa.eu, 27.11.2019
- Frenkel, Sheera und Kang, Cecilia (2021). An Ugly Truth. *Harper Collins Publishers.*
- Gesichtserkennung-stoppen.ch (2021). https://www.gesichtserkennung-stoppen.ch/ (Consulté le 14.02.2022).
- Imhasly, Patrick (2021). Artikel Forschung an Raubgut. *NZZ am Sonntag*, 19.09.2021.
- NiederlandeNet (2020). RECHT : Gericht verbietet Betrugsbekämpfung mit Hilfe des Computerprogramms SyRi wegen Eingriffs in Privatsphäre (Consulté le 07.08.2022). *NiederlandeNet*, WWU Münster, 5.2.2020.
- OECD (2024). Explanatory memorandum on the updated OECD definition of an AI system. *OECD Artificial Intelligence Papers*, No. 8.
- Orwat, C. (2019). Diskriminierungsrisiken durch Verwendung von Algorithmen. I*nstitut für Technikfolgenabschätzung und Systemanalyse (ITAS)*, Karlsruher Institut für Technologie (KIT).
- O'Neil, Cathy (2016). Weapons of Math Destruction. *New York : Crown.*
- Penney, Jonathon (2016). Chilling Effects : Online Surveillance and Wikipedia Use (Consulté le 13.07.2024). *Berkeley Technology Law Journal*, Vol. 31, No. 1, p. 117, 2016.
- Public Code, Public Money (n.d.). https://publiccode.eu/ (Consulté le 14.02.2022).
- Reclaim your Face (2021). https://reclaimyourface.eu/ (Consulté le 14.02.2022).
- Smith, Brad (2018). Facial recognition technology : The need for public regulation and corporate responsibility (Consulté 08.02.2022), *blogs.microsoft.com.*
- Tufekci, Zeynep (2018). YouTube, the Great Radicalizer (Consulté 08.02.2022), *The New York Times*, 10.3.2018.
- Willson, Michele (2017). Algorithms (and the) everyday. *Information, Communication & Society*, 20:1, 137-150, DOI : 10.1080/1369118X.2016.1200645.

B.3 Crédit photo

- Photo de couverture : Foto - Robynne Hu, Unsplash- Lizenz

C Glossaire

- **Systèmes ADM** : Voir *systèmes automatisés de prise de décision*.
- **Algorithme** : Un algorithme est une suite finie et non ambiguë d'instructions ou d'opérations permettant de résoudre un problème ou une classe de problèmes. Il aboutit à une solution après un nombre fini d'étapes de calcul. Les humains peuvent également exécuter des algorithmes à l'aide d'un stylo et de papier.
- **Automated Decision-Making Systems (ADMS)** : Voir *systèmes automatisés de prise de décision*.
- **Systèmes automatisés de prise de décision** (anglais : *Automated Decision-Making Systems*) : Il s'agit de tout logiciel, système ou processus conçu pour automatiser, assister ou remplacer la prise de décision humaine. Les systèmes automatisés de prise de décision peuvent inclure :
  - **Des outils d'analyse de données** produisant des évaluations (numériques), des prévisions, des classifications ou des recommandations d'action.
  - **Des systèmes utilisés pour prendre des décisions** ayant un impact sur le bien-être des individus. Ce bien-être couvre, sans s'y limiter, des domaines sensibles tels que : les opportunités éducatives, les résultats de santé, la performance professionnelle, les perspectives d'emploi, la mobilité, les intérêts, les comportements, l'autonomie personnelle.

  Par ailleurs, les systèmes automatisés de prise de décision peuvent aussi désigner **les processus qui mettent en œuvre de tels outils** (d'après AI Now, Richardson *et al.*, 2019, p. 20, notre traduction).
- **Biais** : également appelé distorsion ou préjugé. Les biais algorithmiques apparaissent lorsqu'un système informatique reflète les valeurs implicites des personnes impliquées dans le codage, la collecte, la sélection ou l'utilisation des données pour entraîner l'algorithme.
- **Données** : Aussi appelé *jeu de données* (*dataset),* également ensemble de données. Ensemble de séries de données utilisées pour la mise en place, l'entraînement, la validation, la prédiction, etc. des systèmes ADM.
- **Série de données** : Collection de chiffres, de textes, d'images, de graphiques, etc. (pour les ordinateurs, tout cela correspond à des chiffres) se rapportant à une personne, un événement ou une situation mesurée.
- **Features** : les features sont des variables issues de séries de données et des valeurs dérivées qui sont utilisées comme séries de données d'entrée pour l'algorithme de décision (le modèle). Il peut s'agir, par exemple, de l'âge, du code postal, des préférences en matière d'eau minérale, mais aussi de méta-variables dérivées telles que la santé nutritionnelle.
- **Modèle de base (*Foundation Model*) :** il s'agit de systèmes d'IA et ADM qui, grâce à la généralité de leurs possibilités de traitement et de sortie, peuvent être utilisés pour une multitude de tâches connues et encore inconnues à ce jour.
- **Boucle de rétroaction** : dans le contexte des systèmes ADM, les boucles de rétroaction décrivent l'effet des résultats des systèmes ADM sur leurs entrées ou sur les entrées de systèmes fonctionnant de manière similaire. Pour des explications détaillées, voir annexe.
- **Coefficients** : les coefficients sont des chiffres concrets qui, selon la règle de calcul (l'architecture du modèle), sont utilisés avec les séries de données d'entrée afin d'obtenir une prévision par le modèle. Il s'agit par exemple des poids dans les réseaux neuronaux ou des limites de distinction dans les algorithmes d'arbres de décision (Decision-Trees).
- **Intelligence artificielle, IA :** il s'agit de systèmes ou d'algorithmes capables d'effectuer des tâches (complexes) à la place des humains. En raison du caractère controversé et de la portée du terme, nous renonçons ici à en donner une définition et faisons référence aux systèmes ADM dans le contexte du présent document.
- **Modèle (algorithmes de décision)** : algorithme capable d'appliquer certains types de modèles statistiques à des séries de données d'entrée. Les chiffres des séries de données

d'entrée sont utilisés avec d'autres chiffres (les coefficients du modèle) selon la règle de calcul (l'architecture du modèle).

- **Feedback-loop** : voir boucle de rétroaction.
- **Ensemble de données synthétiques** : séries de données générées artificiellement qui correspondent à des séries de données réelles dans toutes leurs caractéristiques essentielles. L'utilisation de données synthétiques permet d'éviter les problèmes liés à la protection des données lors de l'utilisation de données sensibles telles que les données personnelles. Les ensembles de données synthétiques sont certes générés artificiellement et leurs séries de données individuelles ne peuvent être attribuées à aucune personne ou objet réel. Mais ils peuvent refléter correctement les propriétés qu'un algorithme spécifique souhaite prédire, de sorte que les algorithmes entraînés sur ces données synthétiques peuvent également déduire correctement la propriété correspondante à partir de séries de données réelles. En termes simples, les ensembles de données synthétiques possèdent les mêmes propriétés pertinentes que les ensembles de données réels, de sorte que l'on peut former des algorithmes qui fonctionnent aussi bien sur des séries de données synthétiques que sur des séries de données réelles. Cependant, dès lors qu'il s'agit de déduire des propriétés à partir d'ensembles de données synthétiques qui n'ont pas été prises en compte lors de leur création à partir de l'ensemble de données réel, cela peut échouer.
- **Données d'entraînement** : ensemble de séries de données utilisées pour le développement ou l'entraînement d'un système ADM.
- **Données de validation** : ensemble de séries de données (différentes des données d'entraînement) utilisées pour évaluer la précision d'un système ADM entraîné.

D Liste des modifications

Tableau des modifications apportées entre la version 1.0 et la version 2.0 de ce document.

En général	Passage à la version 2, ajouts des noms des nouveaux contributeurs, et mention des précédents.
0. Résumé	Ajouté
1. Introduction	Ajouté
2. Champ d'application	Ajout de l'explication de la non-utilisation du terme ”IA”. Ajout d'une référence au ”nudging”
3. Résumé du cadre réglementaire proposé	Renversement de la charge de la preuve expliqué plus en détail. Explication de l'équilibre entre liberté d'auto déclaration et obligations étendues.
4. La question sociétale	Réécrit, exemple des systèmes ADM dans les services sociaux ajouté.
5. Une proposition de réglementation pour les systèmes ADM	Titre modifié. Ajout d'une explication concernant les personnes morales. Ajout d'une explication concernant le financement public des bibliothèques et outils open source.
6. Classification	Section 6.1 réécrite et alignée sur les objectifs de protection, versions précédentes supprimées. Définition du terme « risque » intégrée au chapitre 6.2. Au chapitre 6.2, « directives issues de la loi sur l'IA de la Commission européenne » remplacé par « concepts issus de la loi sur l'IA de la Commission européenne ». Dans la sous-section 6.3, un court paragraphe a été ajouté afin de garantir la sécurité juridique pour les entreprises (notamment grâce aux fiches d'information de l'autorité de surveillance ADMS).
7. Obligation de diligence et de transparence	Paragraphe d'introduction formulé plus clairement et responsable du traitement cité à titre d'exemple. Évaluations d'impact (Impact Assessment) discutées. Ajout d'une référence à la documentation relative au classement dans une catégorie de risque et à la gestion des risques. Ajout de références aux réglementations existantes pour les acteurs publics. Ajout d'une référence aux lignes directrices pour l'IA et l'évaluation.
8. Contrôle, mesures et sanctions	Mention de l'état actuel de la révision du CPC dans une note de bas de page. Suppression de la comparaison entre la surveillance de l'ADMS et celle de la FINMA. Renversement de la charge de la preuve expliqué plus en détail. Chaîne de recours mentionnée aux ch. 8.1 et 8.2. La responsabilité du responsable du traitement demeure.
9. Quelques suggestions pour l'avenir	Modifications mineures.
A. Boucles de rétroaction	Éliminé.
B. Propositions de réglementations pour les systèmes automatisés de prise de décision (ADM), l’intelligence artificielle et les algorithmes	Annexe B mise à jour, renommé annexe A. Titre légèrement adapté.