



ENCADREMENT ET SURVEILLANCE DE L'IA DANS LE SECTEUR FINANCIER

RÉUNION DE PLACE DU 1^{ER} JUILLET 2026

Cette présentation repose sur l'hypothèse que l'ACPR sera désignée autorité de surveillance de marché pour les cas d'usage spécifiques du secteur financier. Cette désignation a été proposée dans le cadre du projet de loi DDADUE, mais doit encore être validée par le Parlement.



INTRODUCTION

EMMANUELLE ASSOUAN

SECRÉTAIRE GÉNÉRALE DE L'ACPR





TOUR D'HORIZON DES ACTUALITÉS DU RÈGLEMENT IA

JADE AL YAHYA, CYRIL CHHUN, JULIEN URI

*SERVICE DE SURVEILLANCE DES RISQUES TECHNOLOGIQUES (SRT)
DIRECTION DE L'INNOVATION, DES DONNÉES ET DES RISQUES TECHNOLOGIQUES (DIDRIT)*

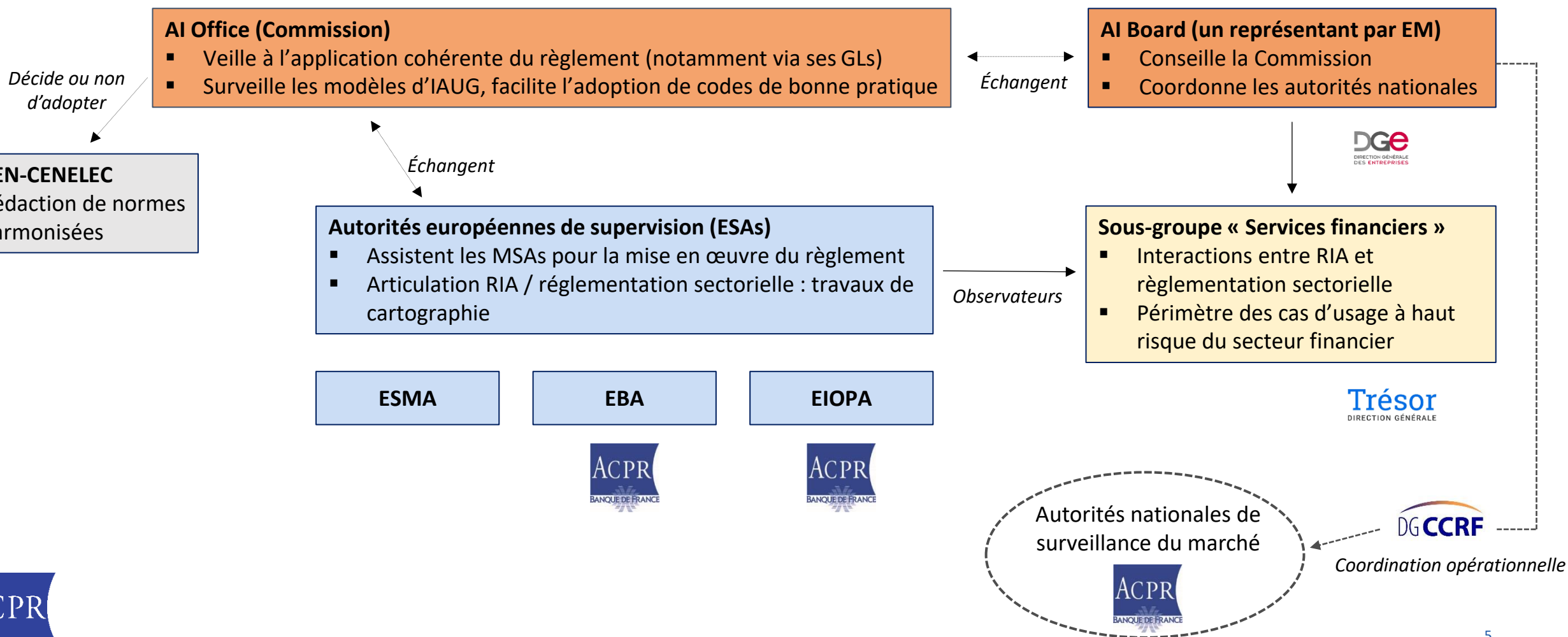


1. État des lieux de la réglementation

- i. Digital omnibus – volet IA : principaux points à retenir
- ii. Nouvelles *Guidelines* de la Commission européenne
- iii. La situation au niveau national
- iv. Questions-réponses



RAPPEL : L'ÉLABORATION DES NORMES DU SECTEUR FINANCIER AU NIVEAU EUROPÉEN





RÈGLEMENT IA : VUE D'ENSEMBLE

- Texte de niveau 1 : volet IA du Digital omnibus en voie de finalisation
- Publication de nouvelle Guidelines par la Commission européenne

Guidelines	Statut
Définition de l'IA	Publiée
Pratiques interdites	Publiée
IA à usage général – GLs et code de bonnes pratiques	Publiée
Notification des incidents sérieux	Publiée (<i>version draft</i>)
Classification des systèmes à haut risque	Publiée pour consultation publique
Exigences des systèmes à haut risque	×
Formulaire d'analyse d'impact sur les DF	×
Articulation avec les autres réglementations	×

- Normes harmonisées CEN-CENELEC : toujours attendues

DIGITAL OMNIBUS – VOLET « IA » - PRINCIPAUX POINTS À RETENIR (1/3)

8 novembre 2024

Déclaration de Budapest

Les dirigeants de l'UE ont appelé à "lancer une révolution en matière de simplification", visant à garantir un cadre réglementaire clair, simple et intelligent pour les entreprises

Depuis février 2025

La Commission européenne a présenté 10 trains de mesures « omnibus » visant à simplifier la législation existante dans de nombreux domaines, dont le numérique

19 novembre 2025

Adoption des propositions du **Digital Omnibus** par la Commission

Début 2026

Examen et négociations au Parlement européen et au Conseil

Avril-mai 2026

Trilogues et rapprochement des positions

7 mai 2026

Accord politique sur l'omnibus IA



DIGITAL OMNIBUS – VOLET « IA » - PRINCIPAUX POINTS À RETENIR (2/3)



Calendrier d'application des obligations relatives aux systèmes à haut risque

- **RIA initial** : entrée en application des obligations relatives aux SHR de l'annexe III au 2 août 2026.
- **Digital Omnibus** : proposition d'un report, conditionné à la disponibilité des normes.
- **Compromis final** : **report à dates fixes** > les règles concernant les SHR de l'annexe III **s'appliqueront à partir du 2 décembre 2027**.



Conditions d'utilisation des données sensibles pour détecter et corriger les biais algorithmiques

- **RIA initial** : possibilité pour les fournisseurs de systèmes d'IA à haut risque de traiter certaines données sensibles afin de détecter et corriger les biais.
- **Digital Omnibus** : extension de cette possibilité aux déployeurs de systèmes à haut risque, et aux fournisseurs et déployeurs d'autres systèmes et modèles d'IA.
- **Compromis final** : extension maintenue, mais associée à une exigence de « stricte nécessité ».

DIGITAL OMNIBUS – VOLET « IA » - PRINCIPAUX POINTS À RETENIR (3/3)



Obligation d'enregistrement des systèmes d'IA

- **RIA initial** : Article 6-4 > « Un fournisseur qui estime qu'un système d'IA visé à l'annexe III n'est pas à haut risque [...] est soumis à l'obligation d'enregistrement prévue à l'article 49, paragraphe 2 ».
- **Digital Omnibus** : proposait de supprimer cette obligation d'enregistrement.
- **Compromis final** : obligation maintenue sous une forme simplifiée, à des fins de surveillance de marché.



Coopération entre autorités publiques du Règlement IA

- **Bureau de l'IA (AI Office)** : compétence exclusive du Bureau de l'IA pour surveiller les SIA fondés sur un modèle GPAI, lorsque le système et le modèle sont développés par le même fournisseur, avec quelques exceptions où les autorités nationales restent compétentes, dont les SIA utilisés par les institutions financières.
- **Autorités de protection des droits fondamentaux** : obligation de bonne coopération avec les ASM.



GUIDELINES SUR LA CLASSIFICATION DES SYSTÈMES À HAUT RISQUE (1/2) : ÉLÉMENTS HORIZONTALS

La finalité est bien le critère déterminant

- Le seul critère permettant de déterminer si un système relève ou non de la catégorie des systèmes d'IA à haut risque est **sa finalité prévue** (*intended purpose*).
- **L'intervention d'un humain dans le processus** n'a aucune incidence sur la qualification du système en tant que système d'IA à haut risque.

Les mécanismes d'exemption de l'article 6(3) sont précisés

- Les *guidelines* précisent les contours des 4 hypothèses d'exemption pour certains systèmes à haut risque, ainsi que la notion de « profilage de personnes physiques ».
- La Commission confirme que les activités financières visées par les 5b et 5c de l'annexe III impliquent généralement un profilage (sauf tâches purement préparatoires).

Les architectures complexes (formées par plusieurs SIA) sont traitées comme un système unique

- Quand plusieurs systèmes d'IA forment ensemble un **système plus complexe**, de telle manière que leur combinaison ou leurs résultats peuvent **influencer significativement une décision individuelle**, la combinaison est traitée comme un système unique – et, le cas échéant, comme un système à haut risque.



GUIDELINES SUR LA CLASSIFICATION DES SYSTÈMES À HAUT RISQUE (2/2) : ÉLÉMENTS SPÉCIFIQUES À L'ANNEXE III

Précisions additionnelles sur le périmètre du cas d'usage 5b de l'annexe III

- **Sont considérés comme à haut risque** : tous les SIA qui effectuent une évaluation de solvabilité en vue d'accéder à des services publics ou privés essentiels.
- **Ne sont pas considérés comme à haut risque** :
 - Les SIA utilisés exclusivement pour la tarification (ex : taux du crédit)
 - Les SIA dont le but premier est la détection de fraude
 - Les modèles internes utilisés uniquement à des fins prudentielles

Précisions additionnelles sur le périmètre du cas d'usage 5c de l'annexe III

- **Évaluation des risques entendue largement** : vise à proposer, refuser, retirer ou fournir des services, et peut conduire à l'établissement ou à la modification des contrats (y.c. couverture, exclusions, franchises...)
- Inclusion des systèmes comprenant des fonctionnalités de **détection de la fraude** (même il s'agit de leur fonction principale), dès lors qu'ils font de l'évaluation des risques ou de la tarification.
- **« Assurance santé »** : ne peut recouvrir que des produits appartenant aux classes 1 et 2 d'assurance non-vie de la directive Solvabilité II (accidents, maladies).
- **« Assurance-vie »** : correspond aux classes de l'annexe II de la directive Solvabilité II ; ne couvre pas les produits d'investissement fondés sur l'assurance (IBIPs), mais intègre bien l'assurance emprunteur, les pensions personnelles, l'assurance dépendance.



FOCUS : RÉGRESSIONS LINÉAIRES ET LOGISTIQUES

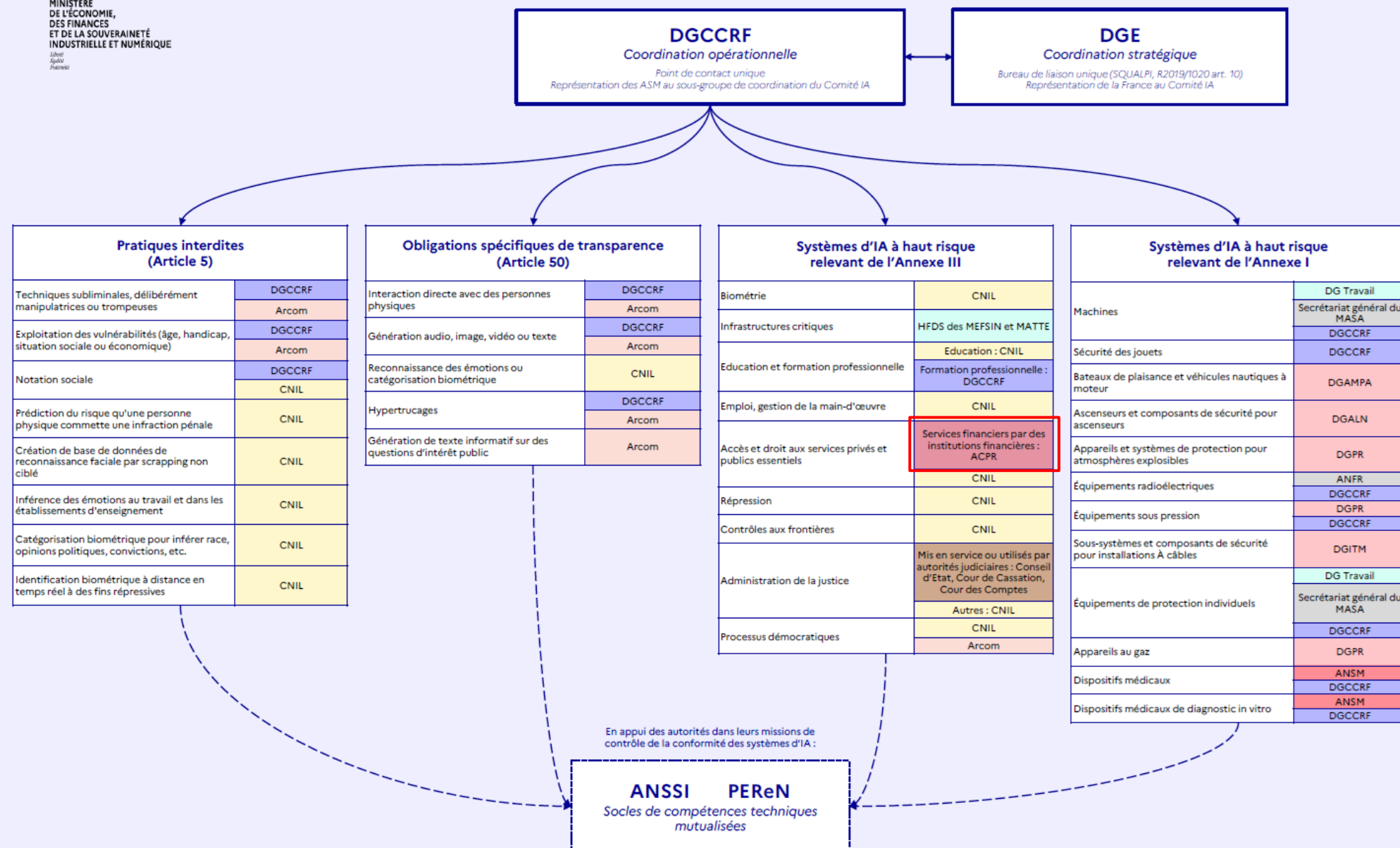
- Règlement IA : définition de l'IA en 7 critères, peu opérationnelle
- Guidelines sur la définition de l'IA en février 2025 (§ 42) : rédaction ambiguë
- Situation actuelle : incertitude pour l'ensemble des parties prenantes
- Démarches EBA / BCE / EIOPA pour obtenir des modifications :
 - Du Règlement via le Digital omnibus (échec)
 - De l'Annexe III que la Commission peut mettre à jour régulièrement (consultation publique sur les *Guidelines* classification des systèmes à haut risque)



FOCUS : PROJET DE *GUIDELINES* SUR LA TRANSPARENCE (ART 50 RIA)

- Texte de référence : Article 50 du Règlement IA sur les obligations de transparence pour certains systèmes d'IA (notamment IA générative)
- Mise en application : **2 août 2026** (pas de report).
- 4 dispositions principales :
 1. Fournisseurs : informer les personnes physiques qu'elles interagissent avec un système d'IA.
 2. Fournisseurs : rendre identifiable le contenu synthétique généré ou manipulé par IA.
 3. Déployeurs : informer les personnes physiques exposées à des systèmes d'IA effectuant de la reconnaissance d'émotion ou de la catégorisation biométrique.
 4. Déployeurs : rendre identifiables les hypertrucages et les textes d'intérêt public générés ou manipulés par IA.
- Publication d'un code de bonnes pratiques sur la transparence du contenu généré par IA (10 juin 2026), afin d'aider fournisseurs et déployeurs à se mettre en conformité

SITUATION EN FRANCE : LE SCHÉMA NATIONAL DE SURVEILLANCE DE L'IA (PROJET DE LOI DDADUE, EN COURS D'EXAMEN)





LES COMPÉTENCES DE L'ACPR EN MATIÈRE DE SURVEILLANCE DE L'IA PROVIENNENT DE DEUX SOURCES COMPLÉMENTAIRES

① Les nouvelles exigences issues du Règlement IA

- Cadre normatif **cohérent** (même si incomplet)
- Mais **périmètre** de systèmes d'IA **assez limité**, d'autant que ces domaines ne sont pas ceux où l'IA est la plus utilisée dans les établissements aujourd'hui



5b. Évaluation de la solvabilité des personnes physiques
5c. Évaluation des risques et tarification des personnes physiques en assurance santé et en assurance-vie



② Les exigences préexistantes issues des réglementations sectorielles

- Neutralité technologique de la réglementation : l'ACPR dispose déjà des compétences (même si en pratique, elle ne les utilise pas encore)



Tous les systèmes d'IA présentant des risques prudentiels ou de « conduite » (protection de la clientèle, LCB-FT)



Travaux européens de **cartographie** : **grande proximité** dans les exigences applicables, même si les objectifs affichés et la présentation diffèrent



Intérêt de concevoir un cadre commun de surveillance des systèmes d'IA



ASPECTS REGLEMENTAIRES

—

QUESTIONS - REPONSES



2. Exigences applicables et supervision

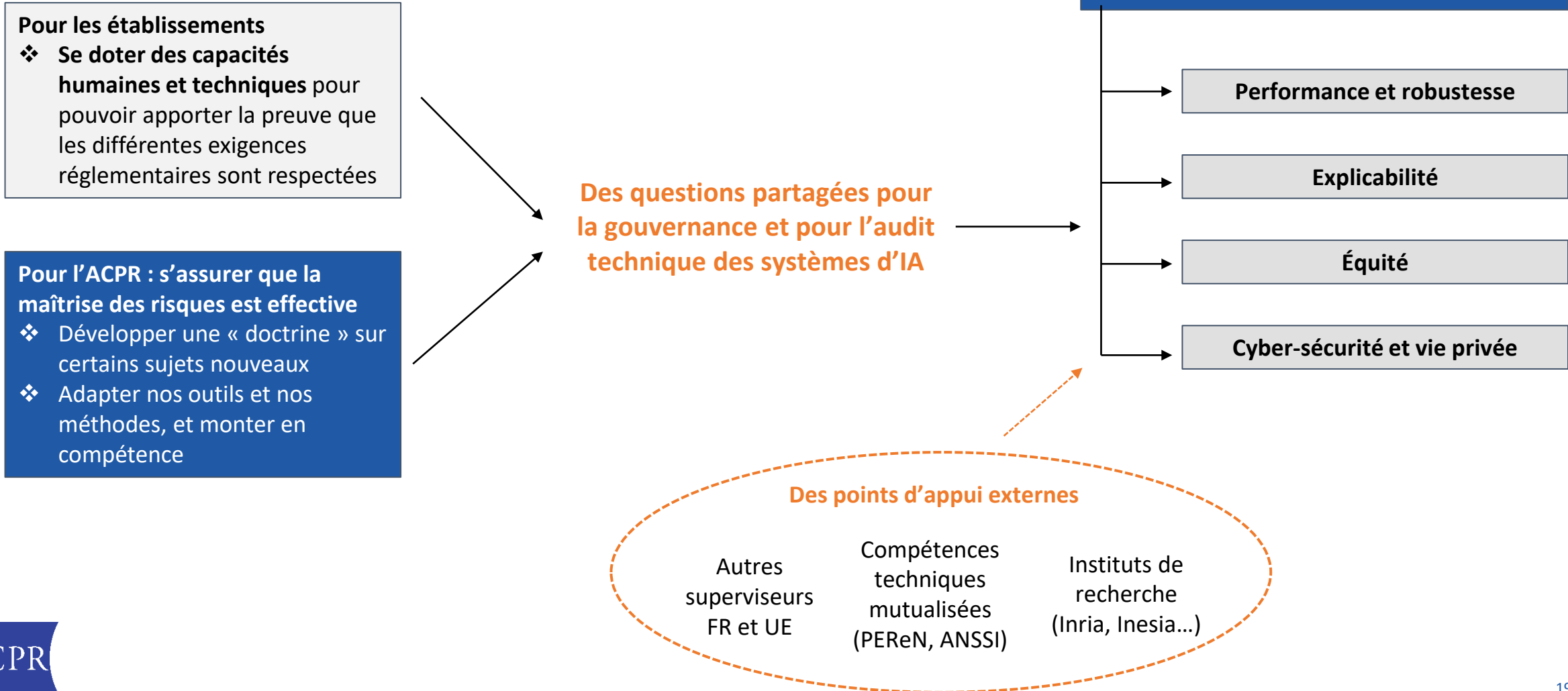
- i. La démarche générale de l'ACPR
- ii. Coopération entre autorités publiques
 - ❖ Intervention de F. Vallet, chef du service IA à la Cnil
 - ❖ Intervention de J.-M. Loubès, directeur de recherche à l'Inria
- iii. Les travaux méthodologiques de l'ACPR
 - ❖ Document de réflexion sur l'équité algorithmique
 - ❖ Travaux sur l'explicabilité
- iv. Questions-réponses



SURVEILLANCE DE L'IA : LA FEUILLE DE ROUTE DE L'ACPR

- **Objectif 1** : Concevoir une surveillance intelligente et proportionnée de l'IA dans le secteur financier
 - Définir une stratégie de surveillance (= *collecte d'informations, cotation des risques, priorisation et définition du programme de travail*)
 - Établir une méthodologie de l'audit de l'IA dans le secteur financier
 - Élaborer une organisation interne
- **Objectif 2** : Accompagner le secteur financier pour lui permettre de construire les « bons » outils de maîtrise des risques

ÉLABORATION D'UNE MÉTHODOLOGIE D'AUDIT : UNE DÉMARCHE DE CO-CONSTRUCTION



IA, RGPD et Règlement IA

Le rôle de la CNIL aujourd'hui et demain

Félicien Vallet
Chef du service IA

La CNIL en bref

NOS MISSIONS

› QU'EST-CE QUE LA CNIL?

- Créée par la loi Informatique et Libertés du 6 janvier 1978, la **Commission nationale de l'informatique et des libertés est une autorité administrative indépendante**. Elle veille à la protection des données personnelles contenues dans les fichiers et traitements informatiques ou papier, aussi bien publics que privés..
- Au quotidien, la CNIL s'assure que l'informatique est au service du citoyen et qu'elle ne porte atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.
- Depuis février 2019, **Marie-Laure Denis**, Conseiller d'État, est présidente de la CNIL.

Inform^{er} les personnes
et proté^{ger} leurs droits



Anticip^{er}
et innover



Accompagner la conformité
et conseiller



Contrôler
et sanctionner



Chiffres clés

Accompagner & conseiller

20

auditions parlementaires

133

délibérations

dont 90 avis sur projets de texte

1351

demandes de conseil traitées

Contrôler & sanctionner

323

contrôles

83

sanctions dont

486 839 500 €

d'amendes

31

rappels aux obligations
légales de la présidente

143

mis en demeure

Anticiper & innover

600

participants à l'événement *air2025*

900

participants au *Privacy Research Day*

26

articles et dossiers publiés sur
le site linc.cnil.fr

Informier & protéger

35 403

appels répondus

266

actions de sensibilisation
sur le terrain

20 150

plaintes reçues

18 123

plaintes traitées

6 999

demandes d'exercice des droits
indirect (EDI) reçues

32 262

demandes d'EDI traitées

dont 80 % demandes
FICOBA de 2024

10,5

millions de visites sur les sites
web de la CNIL

6 167

notifications de violation
de données personnelles

Le service IA de la CNIL

Qui nous sommes et ce que nous faisons

› 7 PROFILS MULTI-DISCIPLINAIRES

- › 2 docteurs en IA
- › 1 docteure en sciences cognitives
- › 2 ingénieurs IA
- › 2 juristes experts de la protection des données

› MISSIONS

- › **Appréhender** : Veille technologique et scientifique, Améliorer la culture de l'IA CNIL
- › **Guider** : Permettre et guider le développement d'une IA respectueuse de la vie privée
- › **Fédérer** : Établir des partenariats avec des institutions, des chercheurs et l'écosystème de l'IA
- › **Auditer** : Élaborer des méthodes d'audit et d'enquête des systèmes d'IA

L'IA à la CNIL

Des actions concrètes

APPRÉHENDER

- Assurer une veille technologique
- Acculturer l'institution sur les sujets IA
- Produire des contenus pédagogiques
 - Sensibiliser le grand public (rencontres des publics...)
 - Echanger avec les spécialistes (articles LINC, projets de recherche, conférences scientifiques...)
- Mener des expérimentation et aider le déploiement de technologies d'IA en interne



FÉDÉRER

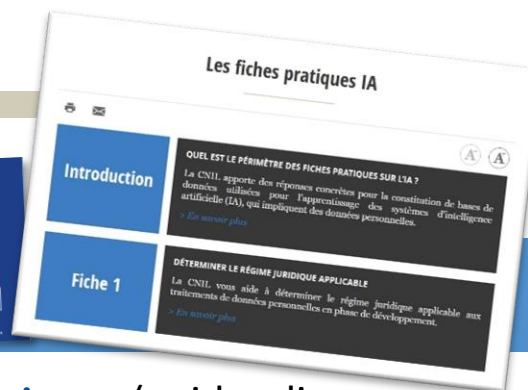
- Instaurer un dialogue de confiance avec les acteurs
- Animer des communautés à différents niveaux (institutions partenaires, au sein de l'écosystème IA, etc.)
- Proposer des dispositifs pour soutenir les acteurs (bac à sable, accompagnement renforcé, demandes de conseil...)



« Bac à sable » données personnelles : la CNIL lance un appel à projets sur l'intelligence artificielle dans les services publics

GUIDER

- Publier des contenus pratiques (guides, lignes directrices...)
- Former les professionnels (webinaires, journées de sensibilisation...)
- Mener des travaux sur le développement et le déploiement de systèmes d'IA (consultation publique, travaux sectoriels...)



AUDITER

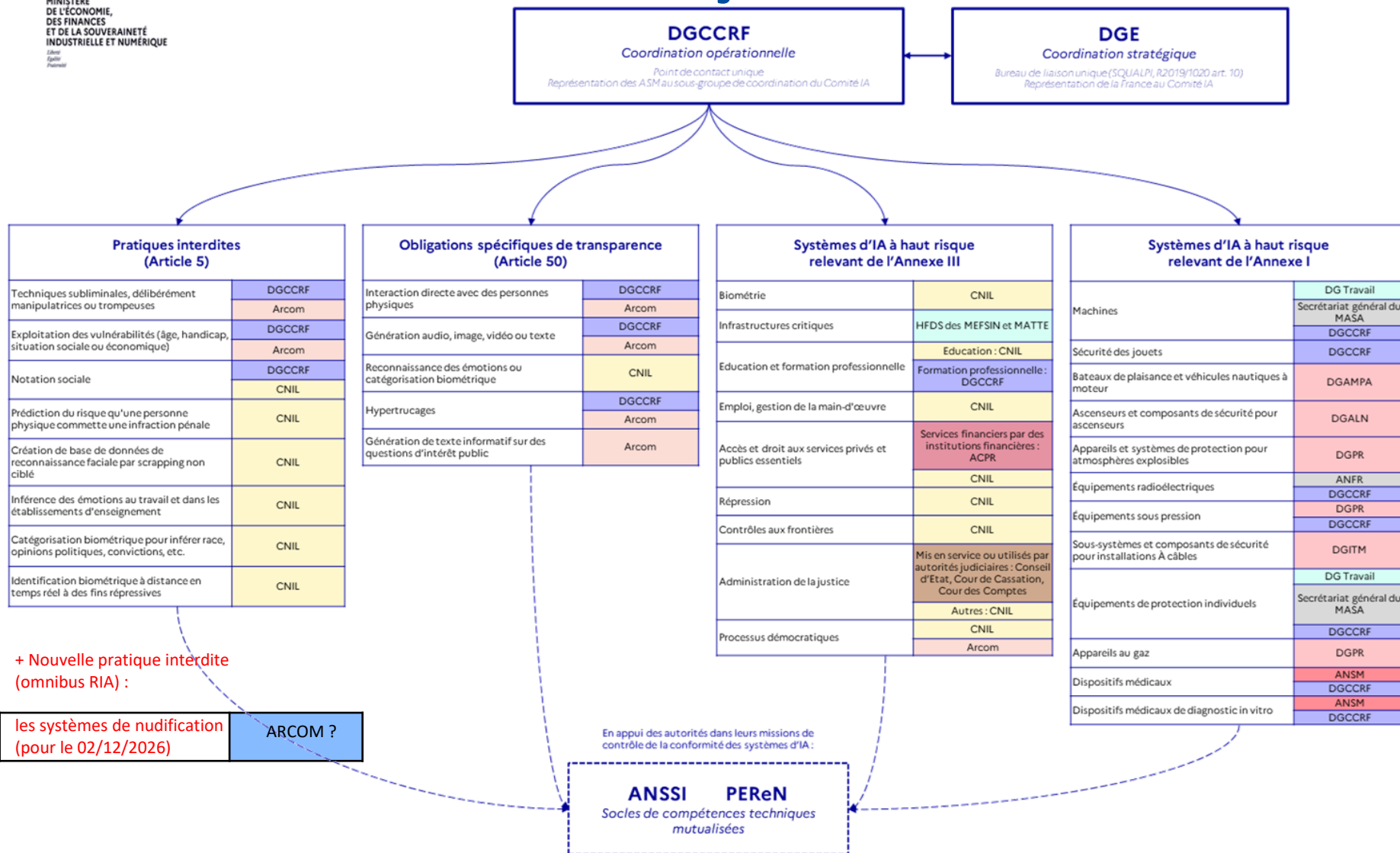
- Développer des méthodologies d'audit et d'enquête
- Outiller les contrôleurs en interne
- Anticiper les exigences du règlement IA et les systèmes de certification à venir



Les missions de la CNIL avec le RIA

De nouveaux rôles et pouvoirs

Gouvernance française



(cf zooms dans les planches suivantes)

Zoom

Pratiques interdites (Article 5)	
Techniques subliminales, délibérément manipulatrices ou trompeuses	DGCCRF
	Arcom
Exploitation des vulnérabilités (âge, handicap, situation sociale ou économique)	DGCCRF
	Arcom
Notation sociale	DGCCRF
	CNIL
Prédiction du risque qu'une personne physique commette une infraction pénale	CNIL
Création de base de données de reconnaissance faciale par scrapping non ciblé	CNIL
Inférence des émotions au travail et dans les établissements d'enseignement	CNIL
Catégorisation biométrique pour inférer race, opinions politiques, convictions, etc.	CNIL
Identification biométrique à distance en temps réel à des fins répressives	CNIL

+ Nouvelle pratique interdite
(omnibus RIA) :

les systèmes de <u>nudification</u> (pour le 02/12/2026)	ARCOM ?
---	---------

Obligations spécifiques de transparence (Article 50)	
Interaction directe avec des personnes physiques	DGCCRF
	Arcom
Génération audio, image, vidéo ou texte	DGCCRF
	Arcom
Reconnaissance des émotions ou catégorisation biométrique	CNIL
Hypertrucages	DGCCRF
	Arcom
Génération de texte informatif sur des questions d'intérêt public	Arcom

Systèmes d'IA à haut risque relevant de l'Annexe III	
Biométrie	CNIL
Infrastructures critiques	HFDS des MEFSIN et MATTE
Education et formation professionnelle	Education : CNIL
	Formation professionnelle : DGCCRF
Emploi, gestion de la main-d'œuvre	CNIL
Accès et droit aux services privés et publics essentiels	Services financiers par des institutions financières : ACPR
	CNIL
Répression	CNIL
Contrôles aux frontières	CNIL
Administration de la justice	Mis en service ou utilisés par autorités judiciaires : Conseil d'Etat, Cour de Cassation, Cour des Comptes
	Autres : CNIL
Processus démocratiques	CNIL
	Arcom

Systèmes d'IA à haut risque relevant de l'Annexe I	
Machines	DG Travail
	Secrétariat général du MASA
	DGCCRF
Sécurité des jouets	DGCCRF
Bateaux de plaisance et véhicules nautiques à moteur	DGAMPA
Ascenseurs et composants de sécurité pour ascenseurs	DGALN
Appareils et systèmes de protection pour atmosphères explosibles	DGPR
Équipements radioélectriques	ANFR
	DGCCRF
Équipements sous pression	DGPR
	DGCCRF
Sous-systèmes et composants de sécurité pour installations À câbles	DGITM
Équipements de protection individuels	DG Travail
	Secrétariat général du MASA
	DGCCRF
Appareils au gaz	DGPR
Dispositifs médicaux	ANSM
	DGCCRF
Dispositifs médicaux de diagnostic in vitro	ANSM
	DGCCRF

5. Accès et droit aux services privés essentiels et aux services publics et prestations sociales essentiels :

a) systèmes d'IA destinés à être utilisés par les autorités publiques ou en leur nom pour évaluer l'éligibilité des personnes physiques aux prestations et services d'aide sociale essentiels, y compris les services de soins de santé, ainsi que pour octroyer, réduire, révoquer ou récupérer ces prestations et services;

b) systèmes d'IA destinés à être utilisés pour évaluer la solvabilité des personnes physiques ou pour établir leur note de crédit, à l'exception des systèmes d'IA utilisés à des fins de détection de fraudes financières;

c) systèmes d'IA destinés à être utilisés pour l'évaluation des risques et la tarification en ce qui concerne les personnes physiques en matière d'assurance-vie et d'assurance maladie;

d) systèmes d'IA destinés à évaluer et hiérarchiser les appels d'urgence émanant de personnes physiques ou à être utilisés pour envoyer ou établir des priorités dans l'envoi des services d'intervention d'urgence, y compris par la police, les pompiers et l'assistance médicale, ainsi que pour les systèmes de tri des patients admis dans les services de santé d'urgence

Les différentes casquettes de la CNIL

- ▶ Notation sociale

- ▶ Prédiction du risque qu'une personne physique commette une infraction pénale

- ▶ Inférence des émotions au travail et dans les établissements d'enseignement

- ▶ Collecte non ciblée d'images faciales

- ▶ Catégorisation biométrique

Autorité nationale compétente

Pratiques interdite

Autorité de Surveillance de Marché

Système d'IA à haut risque

- ▶ Biométrie

- ▶ Éducation et formation professionnelle

- ▶ Emploi

- ▶ Accès/droits aux services privés/publics essentiels

- ▶ Répression

- ▶ Migration, asile et gestion des contrôles aux frontières

- ▶ Identification biométrique à distance en temps réel

- ▶ Reconnaissance des émotions et catégorisation biométrique

Autorité nationale compétente

Transparence

Organisme Notifié

Certification des systèmes d'IA pré mise sur le marché

- ▶ Systèmes biométriques point 1 annexe III destinés à être utilisé par des autorités répressives, les services de l'immigration ou les autorités compétentes en matière d'asile ou par les institutions, organes ou organismes de l'UE

- ▶ Dérogation à la pratique interdite biométrique à distance en temps réel à des fins répressives

Autorité administrative indépendante

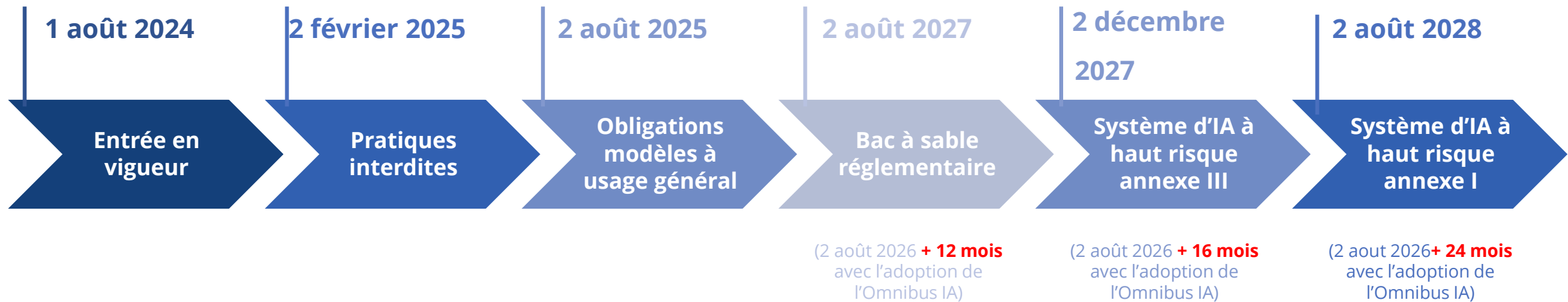
Autorité de protection des droits fondamentaux

- ▶ Uniquement sur le volet protection des données personnelles



Le calendrier

- **Un calendrier ajusté sur certaines obligations** en raison du retard de certains textes nécessaires à leur mise en œuvre (cf planche suivante)



*Sous réserve de l'adoption du compromis d'omnibus IA

Autres mesures importantes de l'omnibus IA

Données sensibles pour la détection et l'atténuation des biais

Actuellement dans le RIA

- La détection et l'atténuation des biais est une obligation pour les fournisseurs de systèmes d'IA à haut risque ;
- En cas de nécessité et sous réserves de garanties, l'utilisation de données sensibles est possible pour la détection et l'atténuation des biais ;

Dans l'omnibus IA

- Cette dérogation au traitement de données sensibles pour la détection et l'atténuation de biais est mobilisable pour l'ensemble des systèmes d'IA sous les mêmes conditions ;
- Cette dérogation n'est pas un prolongement de l'obligation de traiter les biais à l'ensembles des systèmes d'IA ;

Pouvoirs du bureau de l'IA

Pouvoirs du bureau de l'IA actuel sur les systèmes d'IA

- Pouvoirs équivalents à une autorité de surveillance de marché nationale pour le contrôle et la surveillance de certain systèmes d'IA ;
- Champ de compétence : systèmes d'IA basés sur des modèles à usage général si le fournisseur du modèle est le même que celui du système ;

Dans l'omnibus IA

- Le bureau de l'IA devient une ASM à part entière avec une **compétence exclusive** sur son champ de compétence ;

▸ **Systèmes d'IA basés sur des modèles à usage général si le fournisseur du modèle est le même que celui du système**

+

Systèmes d'IA constituant/ étant basés sur un VLOP ou VLOSE au sens du DSA

Exceptions

- SIA annexe I
- SIA infrastructure critique
- SIA dont le fournisseur est une autorité répressive, une autorité de contrôle des frontières, une institution financière
- SIAHR entrant dans le cadre de l'administration de la justice

Travaux de mise en place du RIA

- Gestion des risques IA
- Qualité et gouvernance des données
- Cadre de fiabilité de l'IA (x3) : enregistrements, exactitude et robustesse, transparence et contrôle humain
- Cybersécurité
- Système de gestion de la qualité

Normalisation

- Bac à sable réglementaire
- Plan de surveillance post-commercialisation
- Pouvoirs de contrôle de la Commission sur les modèles GPAI

Actes d'exécution

- ADCO : Travaux de coordination inter ASM européennes (coopération transfrontalières, gestion d'incidents graves et de pratiques interdites, ...)
- Comité IA : groupes de travail des Etats membres sur les thématiques du RIA

Comité IA et ADCO

Lignes directrices et autres

- LD pratiques interdites
- LD définition d'un système d'IA
- LD modèle à usage général (GPAI)
- **LD classification à haut risque annexe III**
- LD obligations de transparence
- LD incident à haut risque
- **LD interplay RIA/RGPD (avec la CNIL en lead)**
- LD obligations SIA à haut risque
- LD modifications substantielles
- LD répartition des responsabilités sur la chaîne de valeur
- LD sur les exigences en termes de contrôle humain
- Code de bonnes pratiques sur les modèles GPAI
- Code de bonnes pratiques sur les mesures de transparence
- Modèle d'analyse d'impact sur les droits fondamentaux (AIDF)
- Modèle de base de données pour les SIAHR

Adopté

En cours d'adoption

En attente

Gouvernance européenne du RIA

Niveau Européen



Forum consultatif

Soutient la contribution des parties prenantes



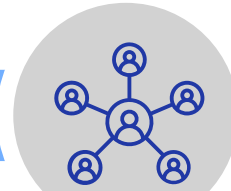
Panel scientifique

Apporte un soutien technique indépendant

Surveillance

Modèles à usage général

Régulation de l'IA



Comité IA

Avec les états membres pour une coordination au niveau de l'UE



Niveau des états membres



Organisme notifié

Responsable de la coordination nationale

Certification



Autorité de surveillance du marché

S'assure de la sécurité des produits

Surveillance

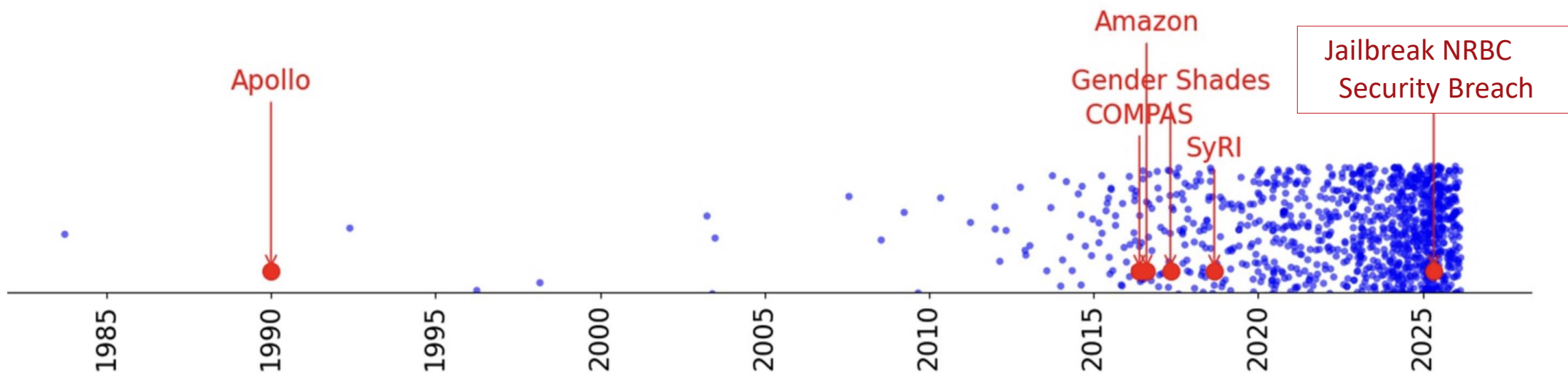
Systèmes d'IA
(toutes catégories de risques confondus)

Evaluer les Biais de l'IA

J-M. Loubes

INRIA - Head of REGALIA Lab - Research Chair TRIAL @ANITI

Directeur Scientifique Institut National Evaluation de l'IA



Accélération des incidents liés à l'IA (Source : AI Incident Database)

- Nécessité d'un programme national d'évaluation de l'IA :
 - Évaluer les modèles d'IA et GPAI avant déploiement et vérifier leur conformité
 - Anticiper, comprendre et objectiver les risques de l'IA
 - Réduire la dépendance aux outils d'évaluation de l'IA étrangers

Intelligence artificielle

L'Institut national pour l'évaluation et la sécurité de l'IA adopte sa feuille de route 2026-2027

 Date: 17 Fév. 2026

[Accueil](#) > [Actualités et événements](#) > L'Institut national pour l'évaluation et la sécurité de l'IA adopte sa feuille de route 2026-2027

Opéré par DGE et SGDSN

4 Partenaires :

- ANSSI
- **INRIA**
- LNE
- PReN

3 Poles de Recherche

- 1: Appui à la Régulation et Sécurité
- 2: Risques systémiques
- 3: Performance et fiabilité

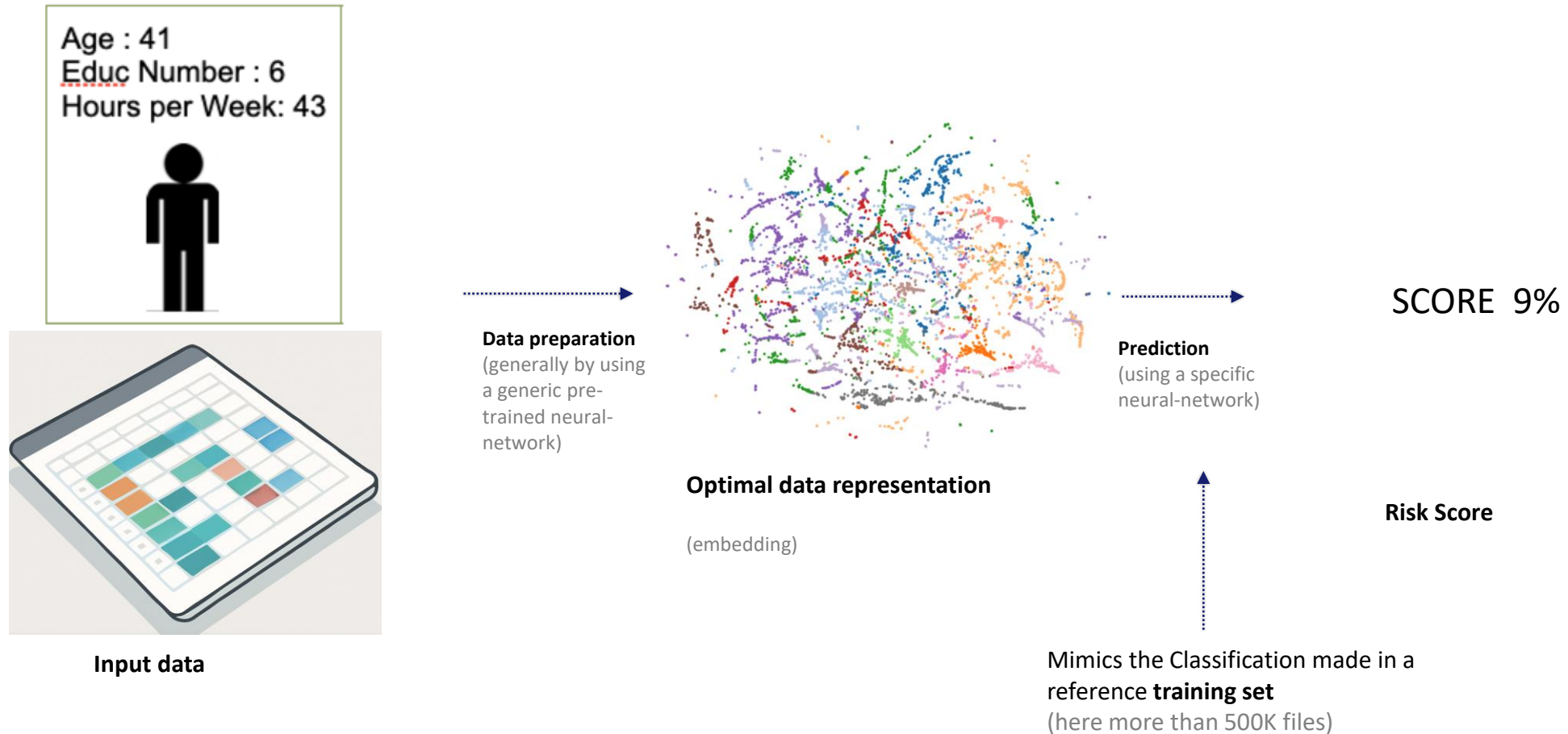
Doter la France d'une **capacité souveraine d'évaluation** fondée sur l'**excellence scientifique** nationale et traduite en outils, protocoles et plateforme directement mobilisables par les ASM et entreprises.

Prédiction automatique d'un score de risque pour application bancaire



Cas à haut risque (AI Act)

⌚ DIFFÉRENCIER LES INDIVIDUS SANS DISCRIMINER

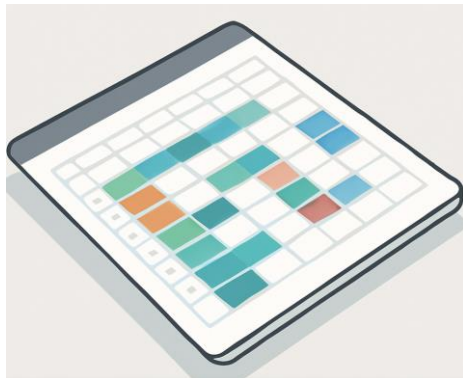
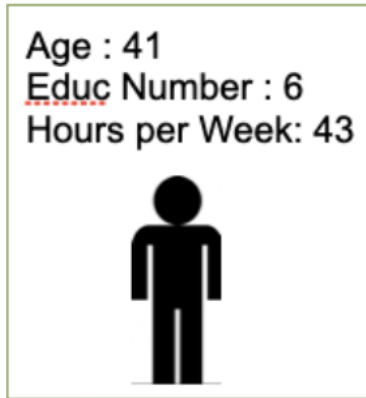


Prédiction automatique d'un score de risque pour application bancaire



Cas à haut risque (AI Act)

🕒 DIFFÉRENCIER LES INDIVIDUS SANS DISCRIMINER



Input data



Prediction
(using a specific
neural-network)

SCORE 9%

Evaluation du Système :

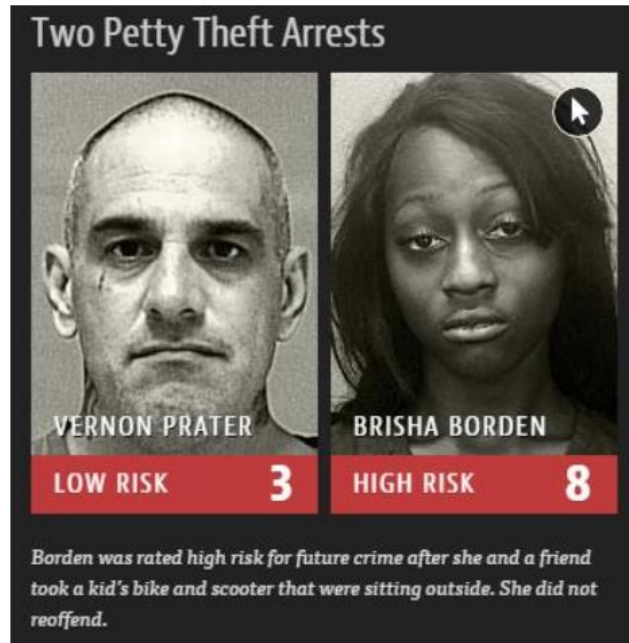
Intentionnalité ? Conformité performance / robustesse

Conformité : privacy des données des patients & **biais du système**

Cybersécurité : blocage / extraction

Risques systémiques : manipulation et risques pour le système bancaire ?

COMPAS



Source : *Propublica*

ALGORITHME DE RECRUTEMENT

Quand le logiciel de recrutement d'Amazon discrimine les femmes

En 2014, le géant du e-commerce a voulu confier ses candidatures à un algorithme, mais celui-ci a commencé à écarter les profils féminins.

Source : *Les Echos*

Amazon a dû désactiver une IA qui discriminait les candidatures de femmes à l'embauche

Source : *Numerama*

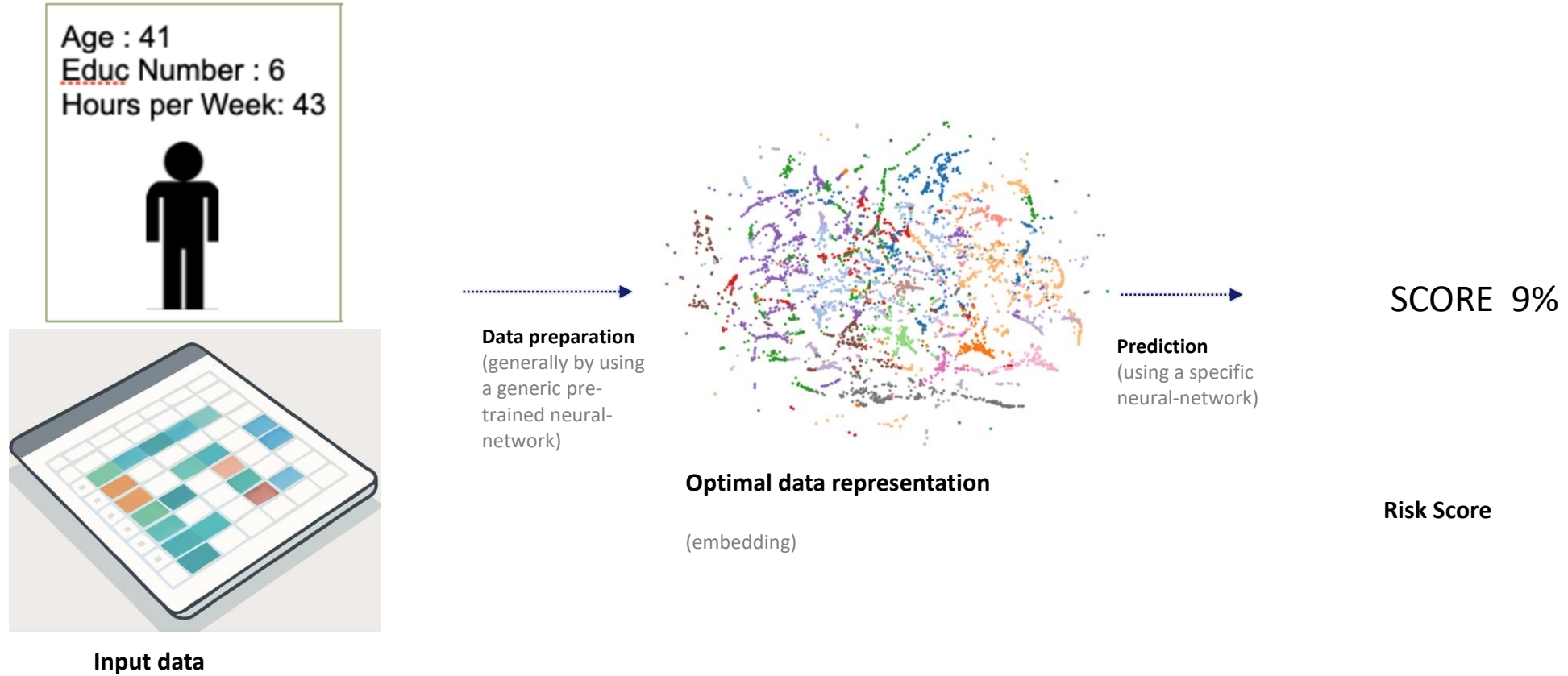


Prédiction automatique d'un score de risque pour application bancaire



Cas à haut risque (AI Act)

⌚ DIFFÉRENCIER LES INDIVIDUS SANS DISCRIMINER

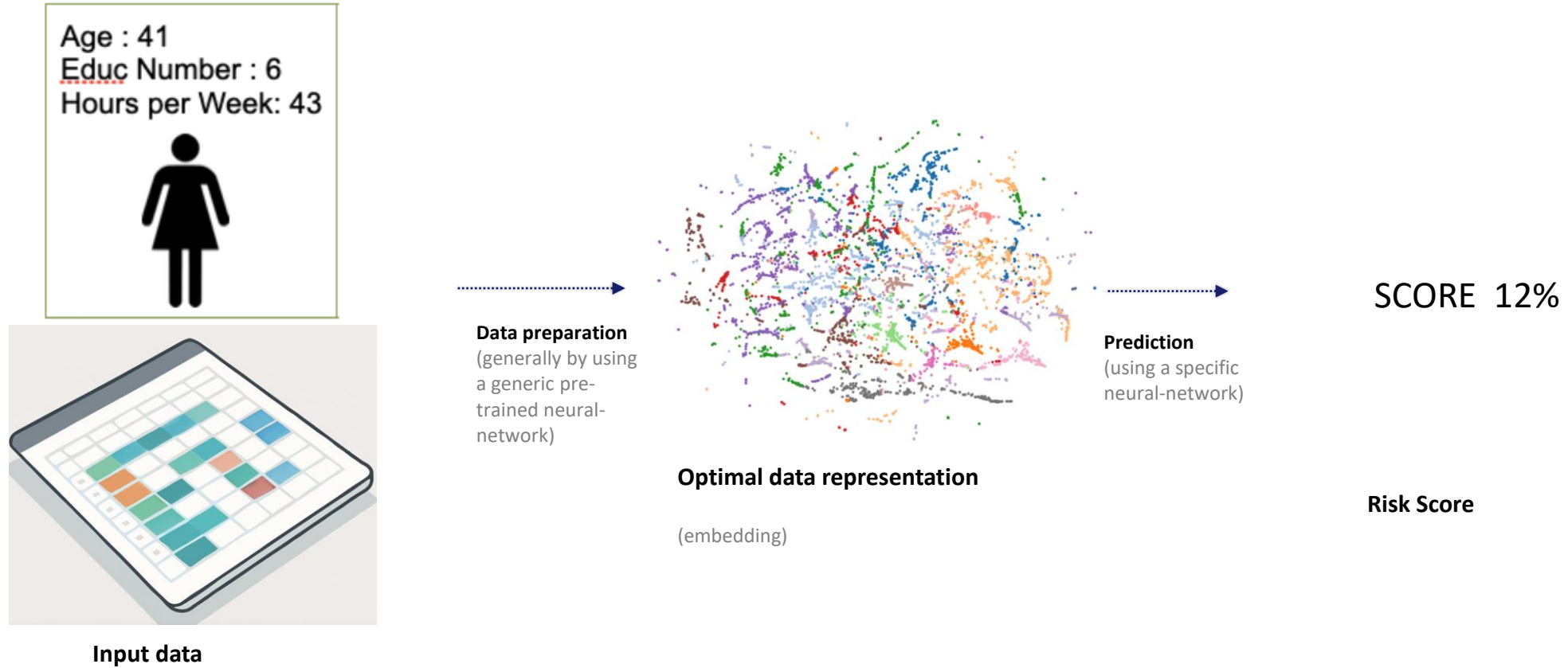


Prédiction automatique d'un score de risque pour application bancaire



Cas à haut risque (AI Act)

⌚ DIFFÉRENCIER LES INDIVIDUS SANS DISCRIMINER



Prédiction automatique d'un score de risque pour application bancaire



Cas à haut risque (AI Act)

⌚ DIFFÉRENCIER LES INDIVIDUS SANS DISCRIMINER

Age : 42
Educ Number : 8
Hours per Week: 41



Input data

Data preparation
(generally by using
a generic pre-
trained neural-
network)



Optimal data representation

(embedding)

Prediction
(using a specific
neural-network)

SCORE 14%

Risk Score

Prédiction automatique d'un score de risque pour application bancaire



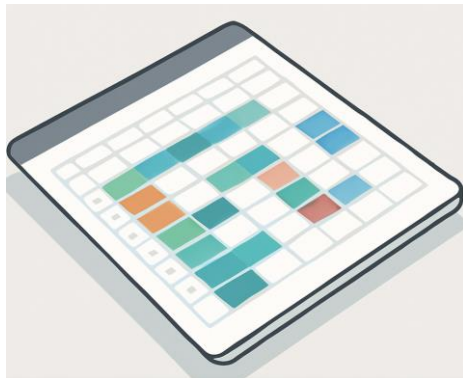
Cas à haut risque (AI Act)



⌚ DIFFÉRENCIER LES INDIVIDUS SANS DISCRIMINER

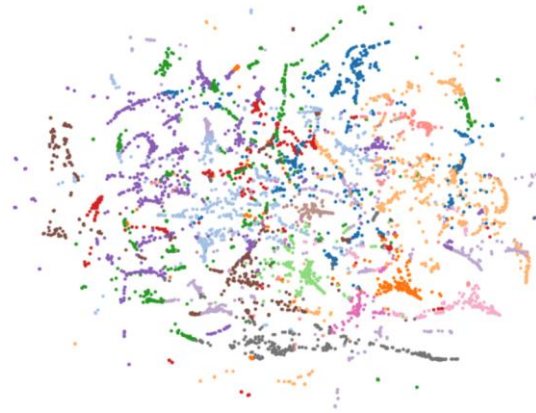
Lequité par l'ignorance n'est plus adaptée aux modèles d'IA.

Age : 42
Educ Number : 8
Hours per Week: 41

Input data

Data preparation
(generally by using
a generic pre-
trained neural-
network)



Optimal data representation

(embedding)

Prediction
(using a specific
neural-network)

SCORE 14%

Risk Score

Disparité ≠ Biais ≠ Discrimination

1

Disparité

Différence systématique entre groupes.

Un simple constat statistique, souvent légitime (un modèle utilisant le revenu produit des scores différents). Ne préjuge ni d'un biais ni d'une discrimination.

Disparity is everywhere

2

Biais

Écart par rapport à une norme pertinente.

Porte sur **toutes les propriétés du modèle** : exactitude, qualité de service, robustesse, explicabilité...

Caractériser un biais suppose d'**expliquer la norme par rapport à laquelle l'écart est jugé.**

3

Discrimination

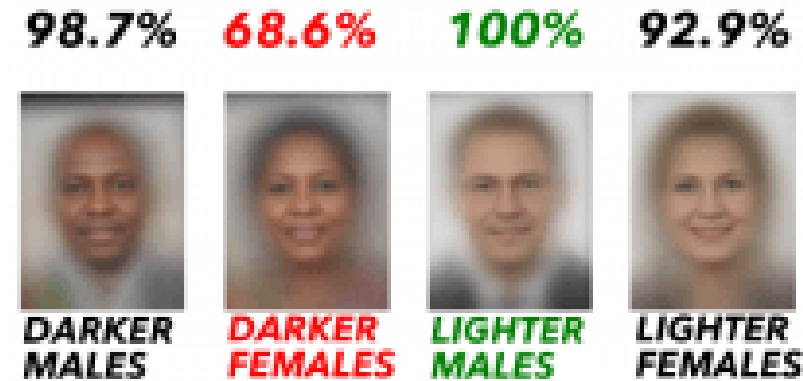
Un biais jugé inacceptable en contexte.

Impact : Au regard du préjudice, des droits affectés et du droit applicable. Implique un jugement normatif et contextuel.

Chaque niveau exige un type de preuve différent. « Disparity is everywhere »

CHOISIR LES METRIQUES DE BIAIS

Incompatibilité des propriétés : nécessité de choisir une définition de l'équité



Biais de décision (Indépendance) :

→ La même décision pour tous A

Notion probabiliste d'**indépendance**

$$\hat{Y} = f(X, A) \perp A$$

Biais de performance (séparation) :

→ Même performance pour tous A

Indépendance décision / attribut, à comportement réel égal

$$\hat{Y} = f(X, A) \perp A | Y$$

Biais de calibration (suffisance) :

→ La réalité est indépendante du groupe, à prédiction donnée.

Même **fiabilité des décisions**

$$Y \perp A | \hat{Y}$$

Implications pratiques différentes.

ESTIMER LES METRIQUES

L'algorithme comporte une partie aléatoire

La plupart des propriétés dépendent

- des **données d'entraînement**,
- du modèle d'IA et de l'**apprentissage**,
- des **données** sur lesquelles le système d'IA est déployé.

*-> Choix d'un niveau de contrôle des métriques
Moyenne, variabilité, une propriété particulière, ou
toute la distribution*

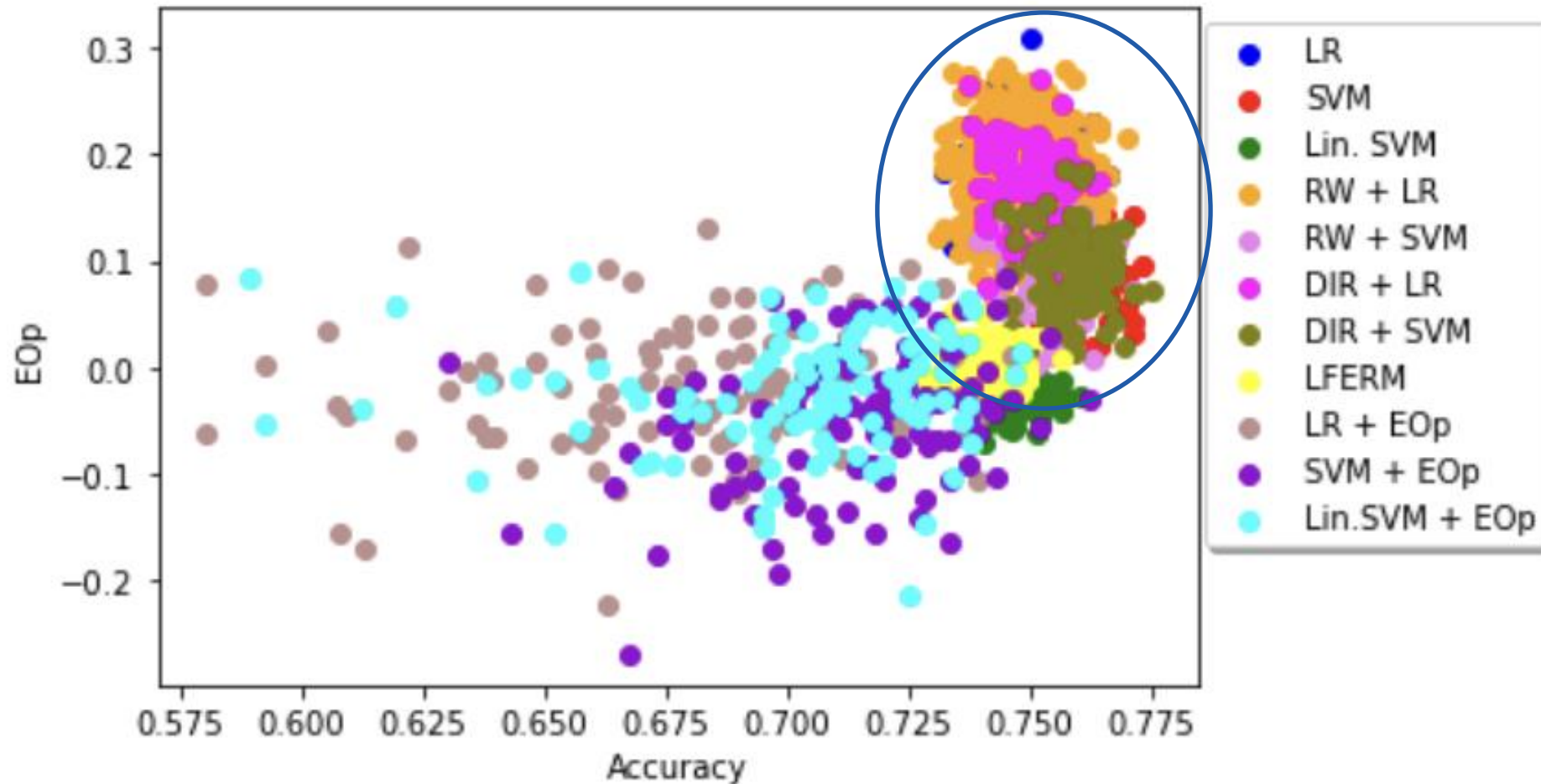
-> Contrôle des incertitudes liées à l'estimation



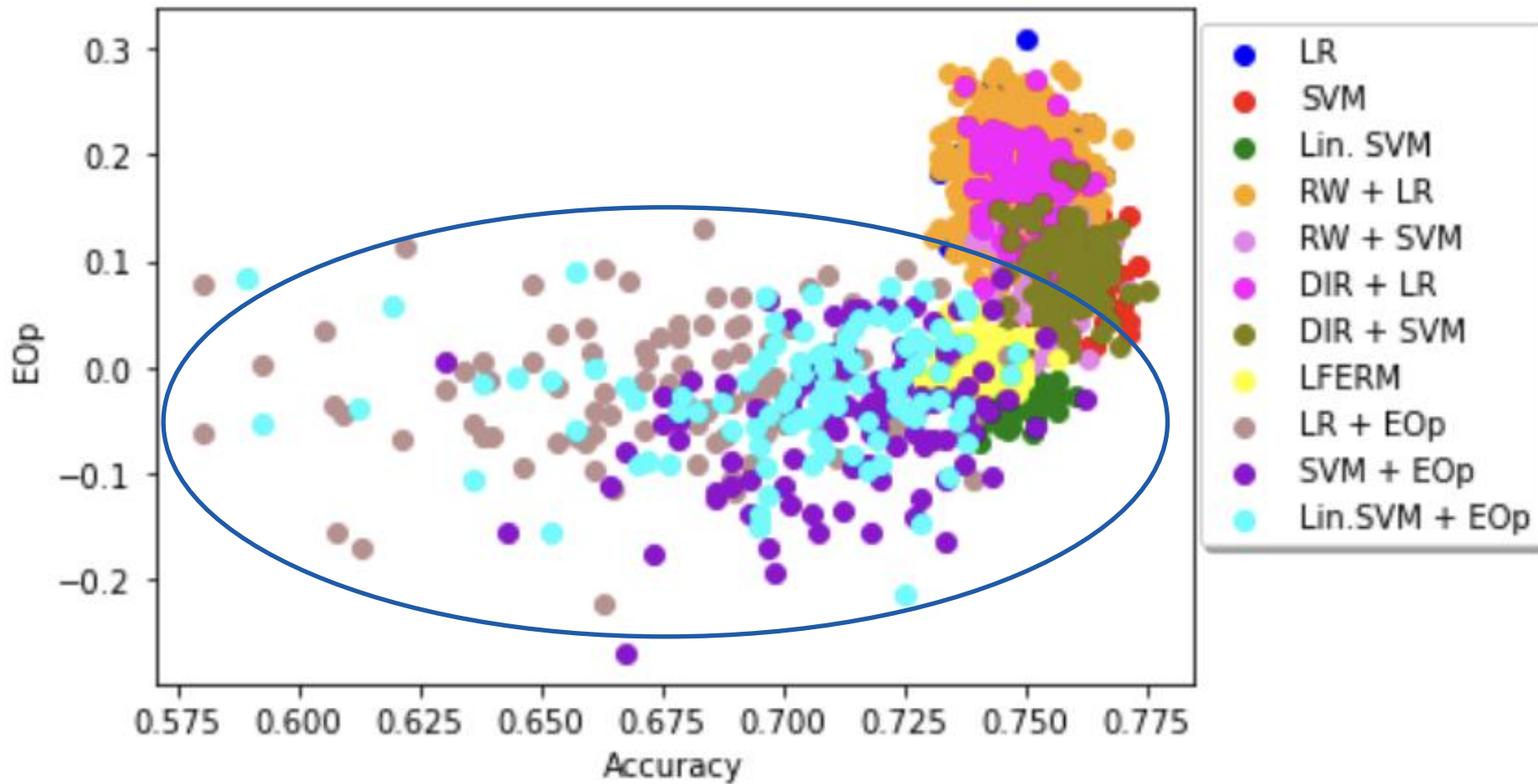
Métrique de séparation : « Egalité des Opportunités »

En moyenne autant d'erreurs d'attribuer à tort un crédit pour un groupe que pour l'autre

Variabilité des mesures de biais des modèles non contrôlés



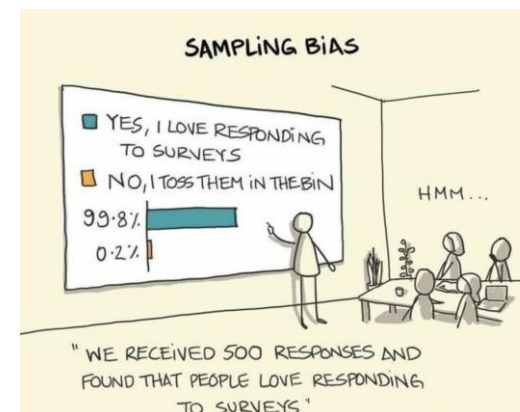
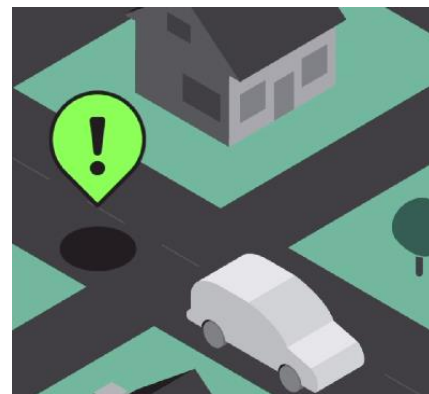
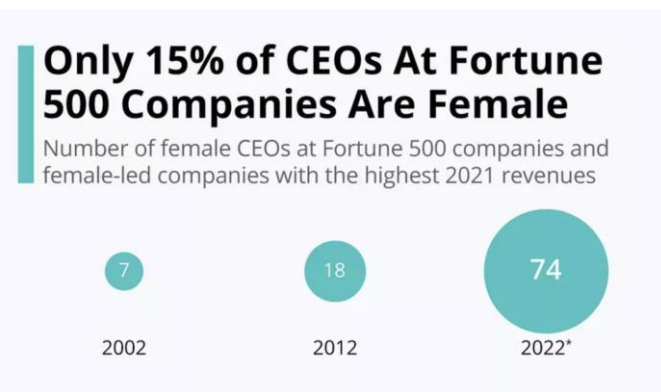
Variabilité des modèles corrigés par post-processing sur moyenne



Données, modèle et amplification

L'IA apprend à partir de données et les Biais sont présents dans les données

1. Biais historiques : les données sont le reflet des usages des sociétés et sont donc représentatifs des biais culturels déjà présents.
2. Biais d'acquisition : les conditions de collecte des données ne permettent pas d'avoir des échantillons représentatifs et avantagent une sous-population (biais géographiques, biais liés aux modalités de collecte). Les données peuvent aussi être pré-traitées par des algorithmes amplifiant les biais (ex : par agrégation).
3. Biais de sélection : seule une partie des données est observée car l'autre partie a été préalablement censurée (ex : on n'observe pas le risque de défaut de personnes à qui un prêt a été refusé).
4. Présence de variables confondantes : La sur-représentation de variables très corrélées entre elles crée des proxys. L'algorithme apprend une corrélation et la transforme en loi causale.



[A survey of bias in ML thought the prism of statistical parity Besse et al. 2022 The American Statistician]

Données, modèle et amplification



Biais dans la modélisation

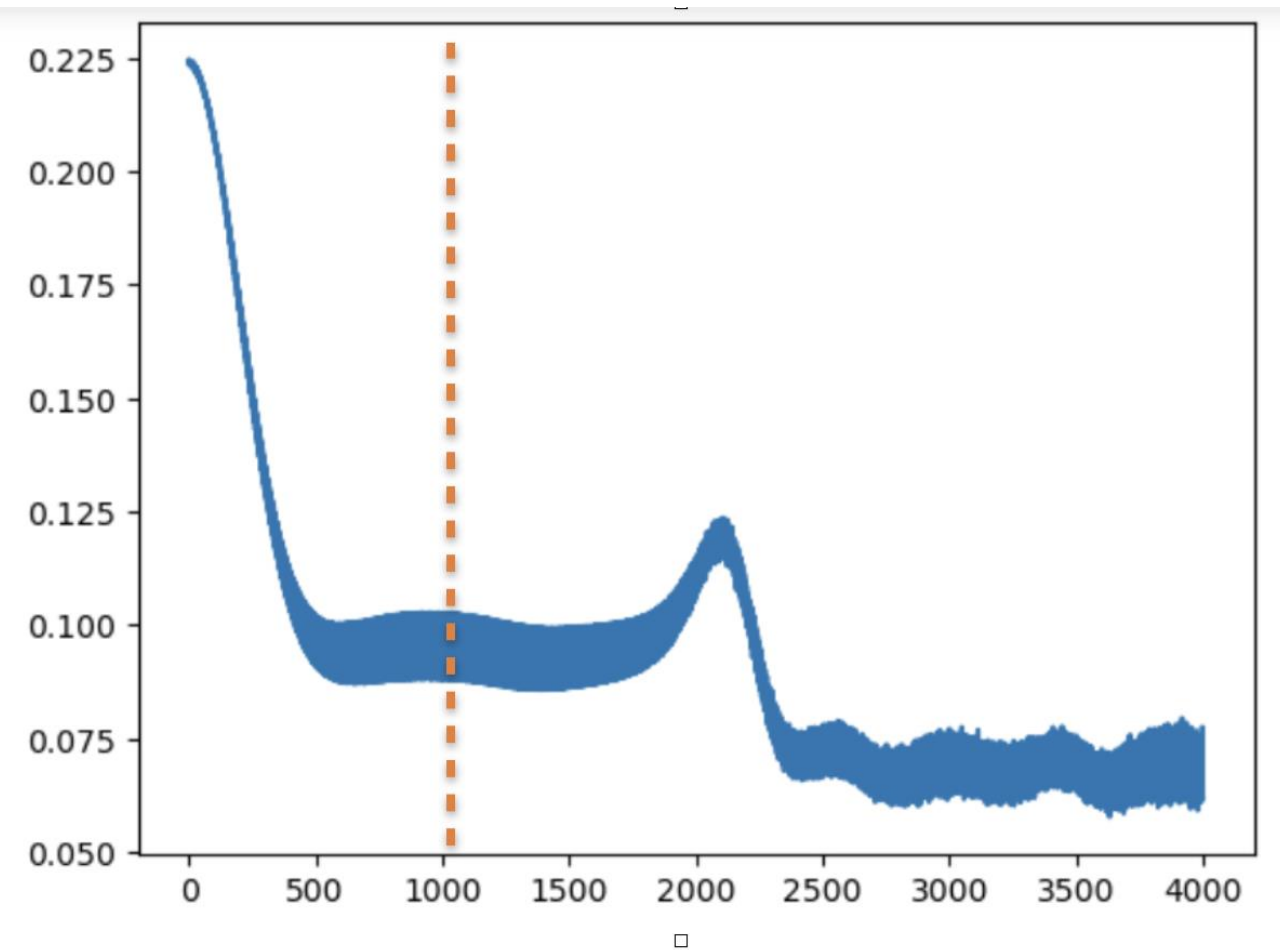
- Choix du proxy cible
- Choix de la Fonction d'objectif
- Choix seuils : amplification & rétroaction.



Amplification des biais par l'apprentissage

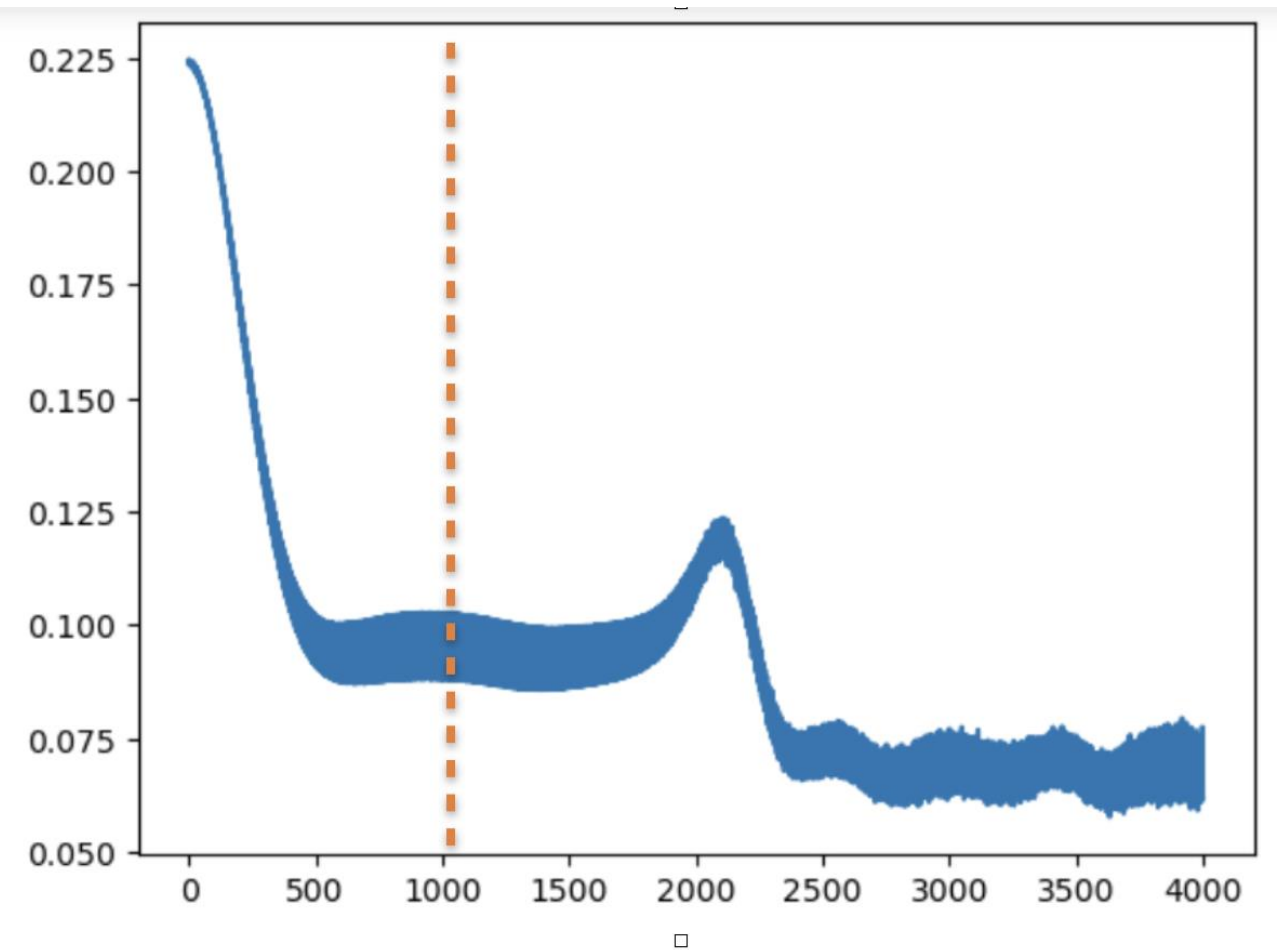
- La trajectoire d'entraînement n'est pas neutre.
- Les comportements minoritaires sont appris plus tard
- Choix des modèles, des optimiseurs et des méthodes d'entraînement (early stopping)

L'IA raisonne par stéréotypes lors de son apprentissage

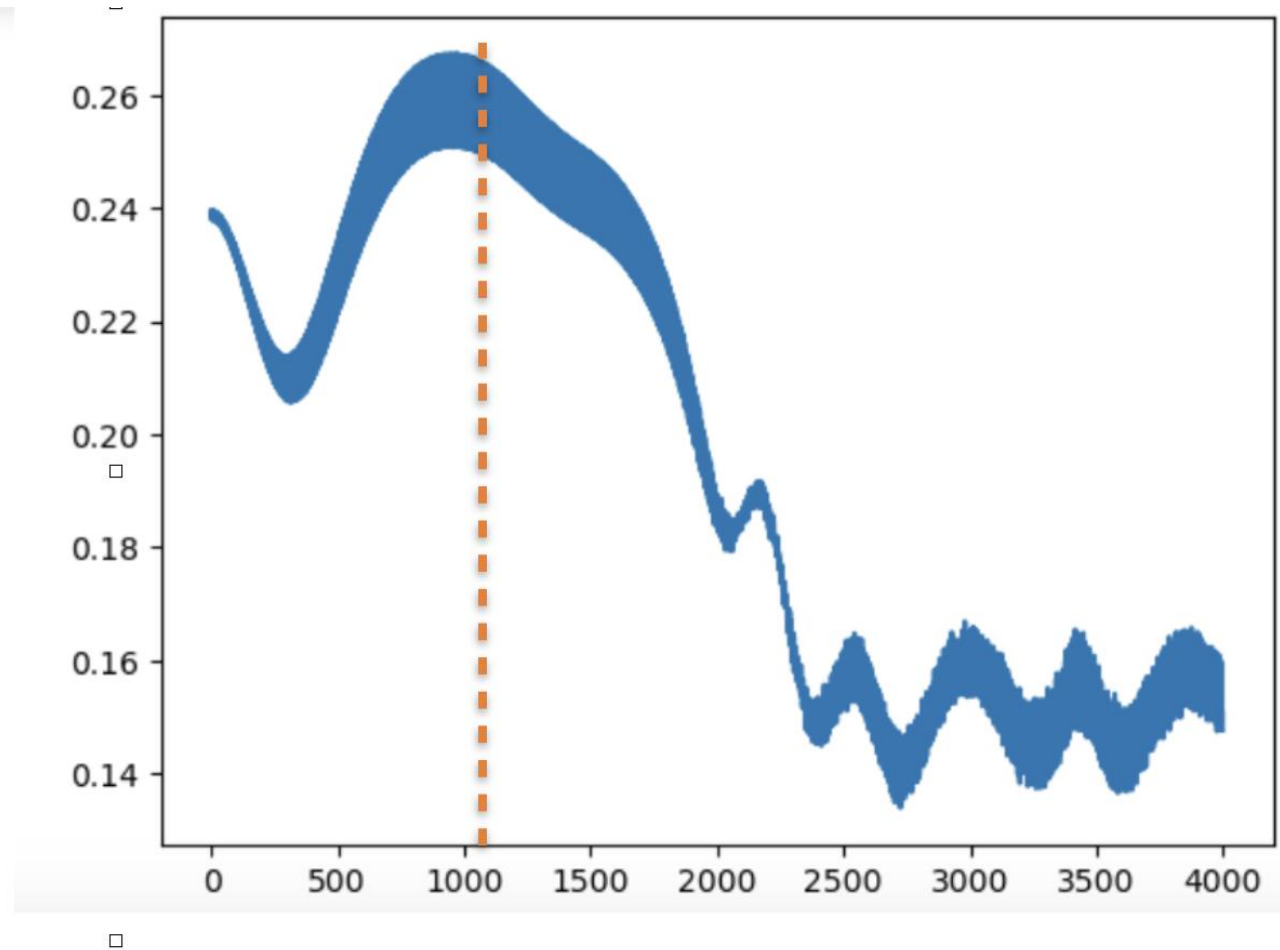


Perte générale

L'IA raisonne par stéréotypes

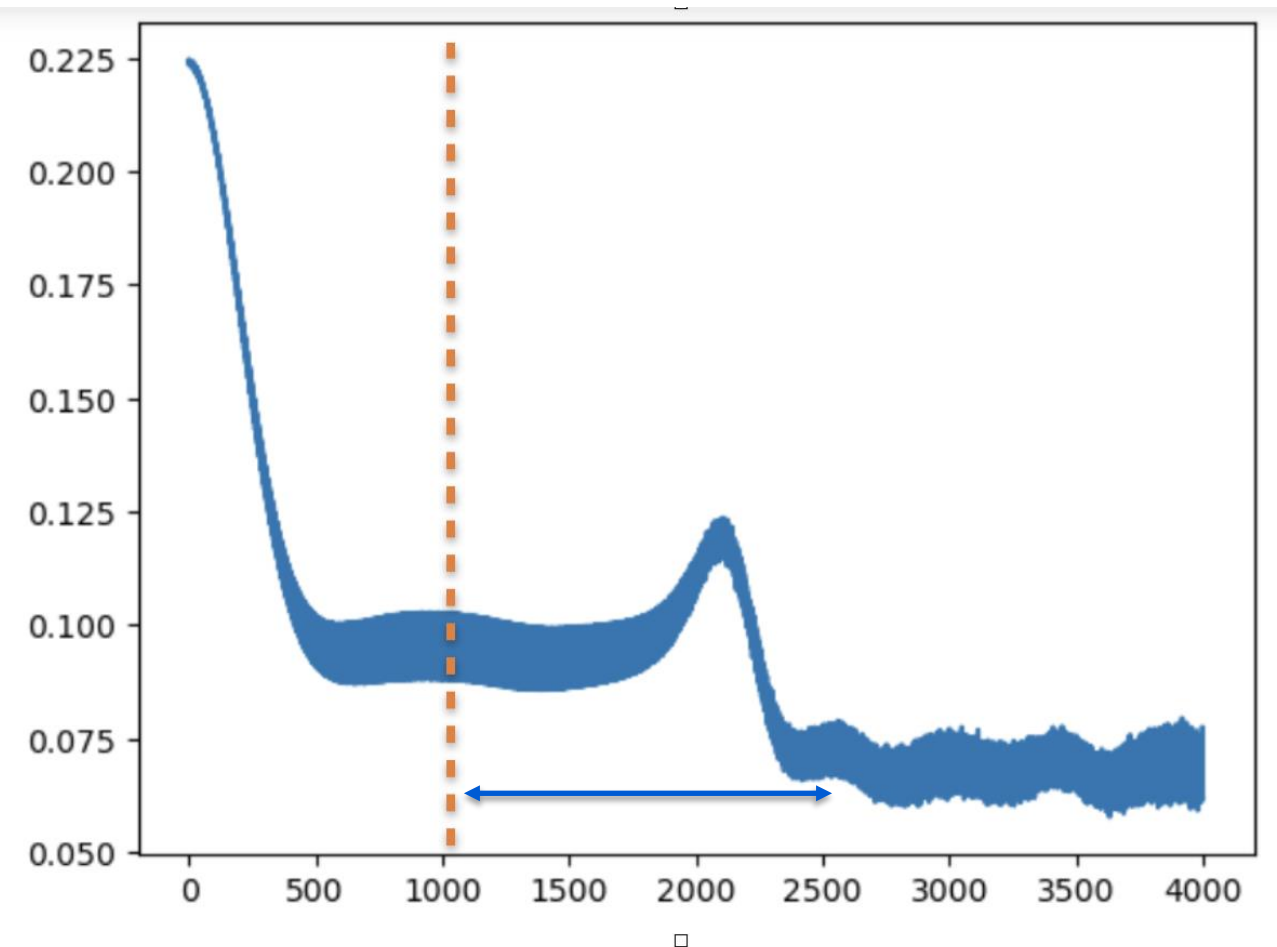


Perte générale

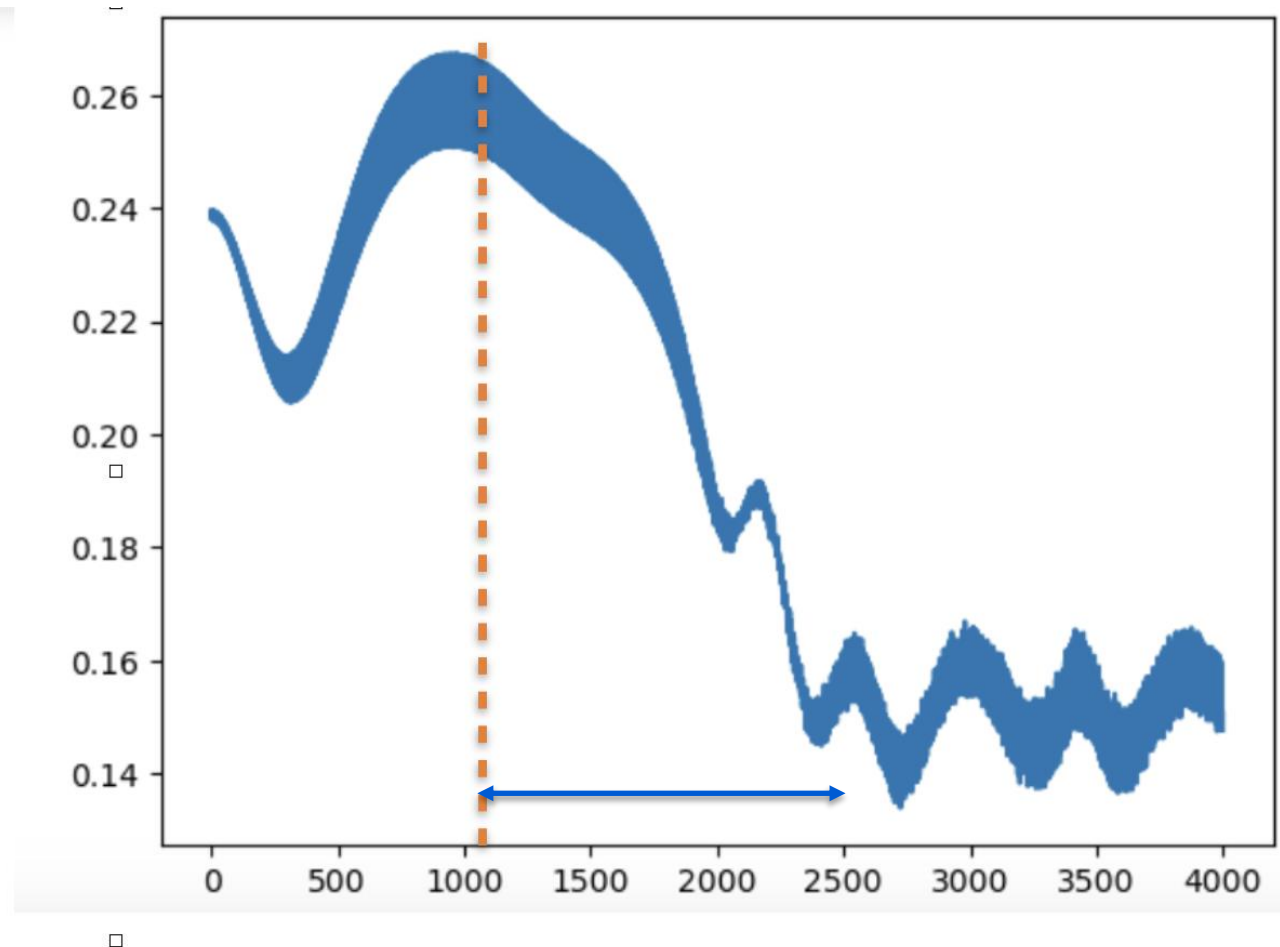


Perte d'un groupe minoritaire

Attention au cas du biais inconnu



Perte générale



Perte d'un groupe minoritaire

Coût de l'équité

IA GÉNÉRATIVE

Un cadre structurellement différent

Les méthodes conçues pour les modèles prédictifs ne se transposent pas directement : ni cible univoque, ni groupe déclaré, un préjudice dépendant de l'usage.

Quatre familles de préjudices

Représentationnels

rôles ou traits stéréotypés

Qualité de service

selon la langue ou les persona

Effacement de la diversité

convergence vers des profils dominants

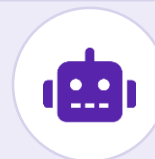
Allocatifs

en aval, via une décision automatisée



Le biais : une propriété interne du modèle d'IA

LLM as a judge : instrument suspect



Modèle du monde

Le biais vit dans la géométrie des représentations apprises



Intérêt caché & complaisance

Objectif non connu de l'utilisateur
sycophantie,
recommandations manipulatrices

TACHE 1 : CRÉATION D'UNE IMAGE D'UNE INSPECTION GÉNÉRALE D'UNE BANQUE



TACHE 2 : ANALYSE DES EXPRESSIONS FACIALES



Expert

Bossy

Le leurre de l'explication de surface

Un LLM sait formuler une justification crédible sans pour autant dévoiler ses mécanismes réels. **Loin de reconstruire le « modèle du monde », l'alignement n'en recouvre que la surface.**



L'alignement se réduit à un mince vernis

Centré sur les premiers tokens, le comportement « sûr » laisse le prior du monde presque inchangé



Raisonnement affiché ≠ explication

Issues du modèle lui-même, les traces de raisonnement tendent à rationaliser après coup au lieu de dévoiler les véritables critères

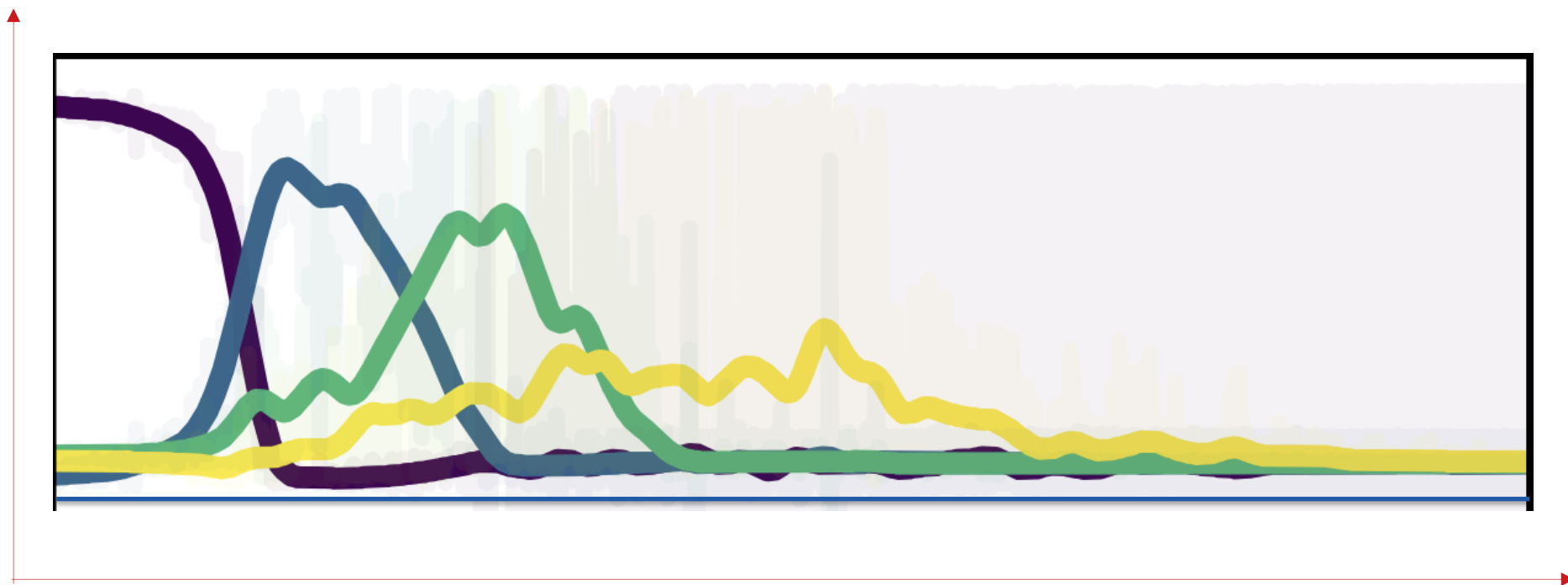


Niveau comportemental ≠ représentationnel

Effacer un stéréotype apparent ne modifie en rien la géométrie qui l'engendre : le biais subsiste, simplement dissimulé sur les seuls axes que l'on mesure.

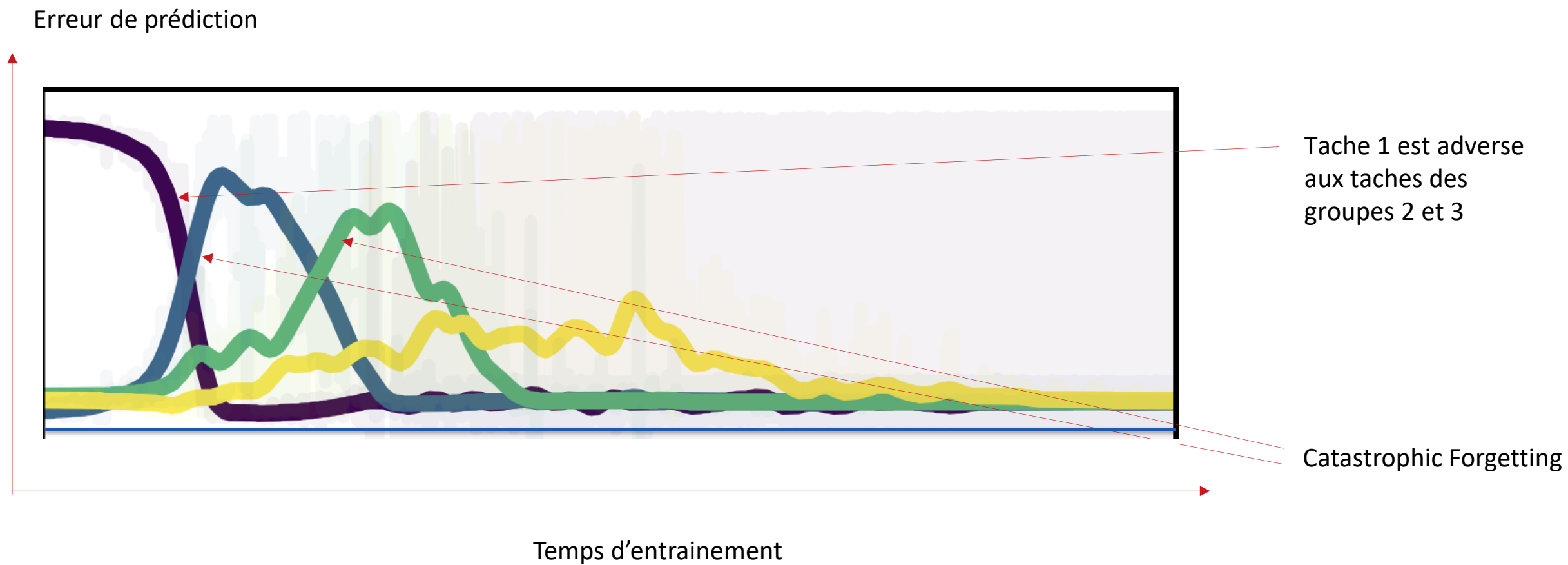
Le fine tuning désaligne les minorités

Erreur de prédiction



Temps d'entraînement

Le fine tuning désaligne les minorités

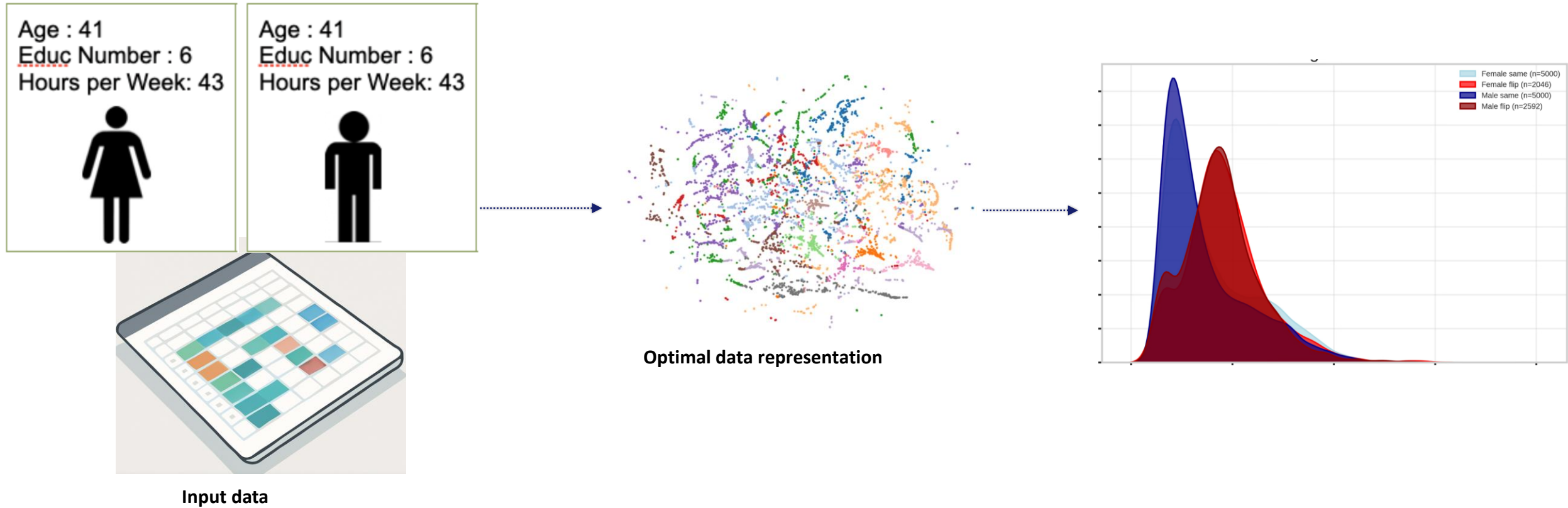


Probe et raisonnement counterfactual dans les embeddings du modèle



Cas à haut risque (AI Act)

⌚ DE NOUVEAUX OUTILS POUR COMPRENDRE LES LLM



Trois couches d'évaluation, à mobiliser ensemble



**DOCUMENT DE RÉFLEXION
« L'ÉQUITÉ ALGORITHMIQUE DANS LE
SECTEUR FINANCIER »**





DOCUMENT DE RÉFLEXION : POURQUOI S'INTÉRESSER À L'ÉQUITÉ ? (1/2)

■ Enjeu de conformité réglementaire :

- Le Règlement IA fait de l'**équité** une exigence
 - L'une des dimension techniques clés de l'évaluation de l'IA... probablement la plus nouvelle pour l'ACPR
- Enjeu plus large de **protection de la clientèle** et de confiance
- Dans un contexte de montée en puissance des enjeux de droits fondamentaux dans le secteur financier (ex : arrêt *Test-Achats* de la CJUE - 2011)

■ Enjeu de clarification réglementaire :

- Articuler réglementation trans-sectorielle et sectorielle, prendre en compte les enjeux spécifiques du secteur financier
- En l'occurrence : **comment segmenter sans discriminer ?**



DOCUMENT DE RÉFLEXION : POURQUOI S'INTÉRESSER À L'ÉQUITÉ ? (2/2)

- L'ACPR a conduit en 2025 une série d'**ateliers méthodologiques** avec des acteurs financiers volontaires :
 - Travaux assez récents et plutôt expérimentaux
 - **Fortes attentes** vis-à-vis des superviseurs – voire des législateurs – pour clarifier les règles.
- Le format du **document de réflexion** permet de « *développer des analyses sur des points qui n'ont pas encore fait l'objet de prises de positions officielles, en vue de les discuter avec les parties prenantes, notamment la profession, afin d'approfondir les analyses développées* »
 - **Consultation de la Place**
- Notre objectif : passer de la ***data science*** et du **droit à la pratique**

Structure du document de réflexion

1. Le contexte réglementaire
2. La notion d'équité dans le secteur financier
3. Les différentes conceptions de l'équité de groupe
4. Les méthodes d'estimation et de correction des biais
5. Mise en œuvre pratique
6. Anticiper l'essor de l'IA générative dans le secteur financier



1. CONTEXTE RÉGLEMENTAIRE : DE NOMBREUSES SOURCES À CONCILIER

Droits fondamentaux : droit anti-discrimination, droit des données

- Constitution française (principe), code pénal (26 critères protégés)
- Charte UE des droits fondamentaux, Convention européenne des droits de l'homme
- RGPD : interdiction de traiter les données « sensibles » (sauf exceptions)

Règlement IA

- Le droit à la non-discrimination doit être respecté par tous les systèmes d'IA
- Principe d'équité des systèmes à haut risque... à opérationnaliser

Règlementations sectorielles

- Prudentielle : l'équité n'est pas un sujet
- Protection de la clientèle : des exigences d'équité (CCD, MCD, DDA...), le plus souvent dans une logique d'abstention (de vente, d'octroi...), là où le Règlement IA porte plutôt une logique d'accès
- Exigences spécifiques assurance : interdiction des tarifs différenciés en fonction du genre (CJUE, Test-Achats, 2011)



2. CONCEPTS FONDAMENTAUX (1/2) : DISPARITÉ, BIAIS ET DISCRIMINATION

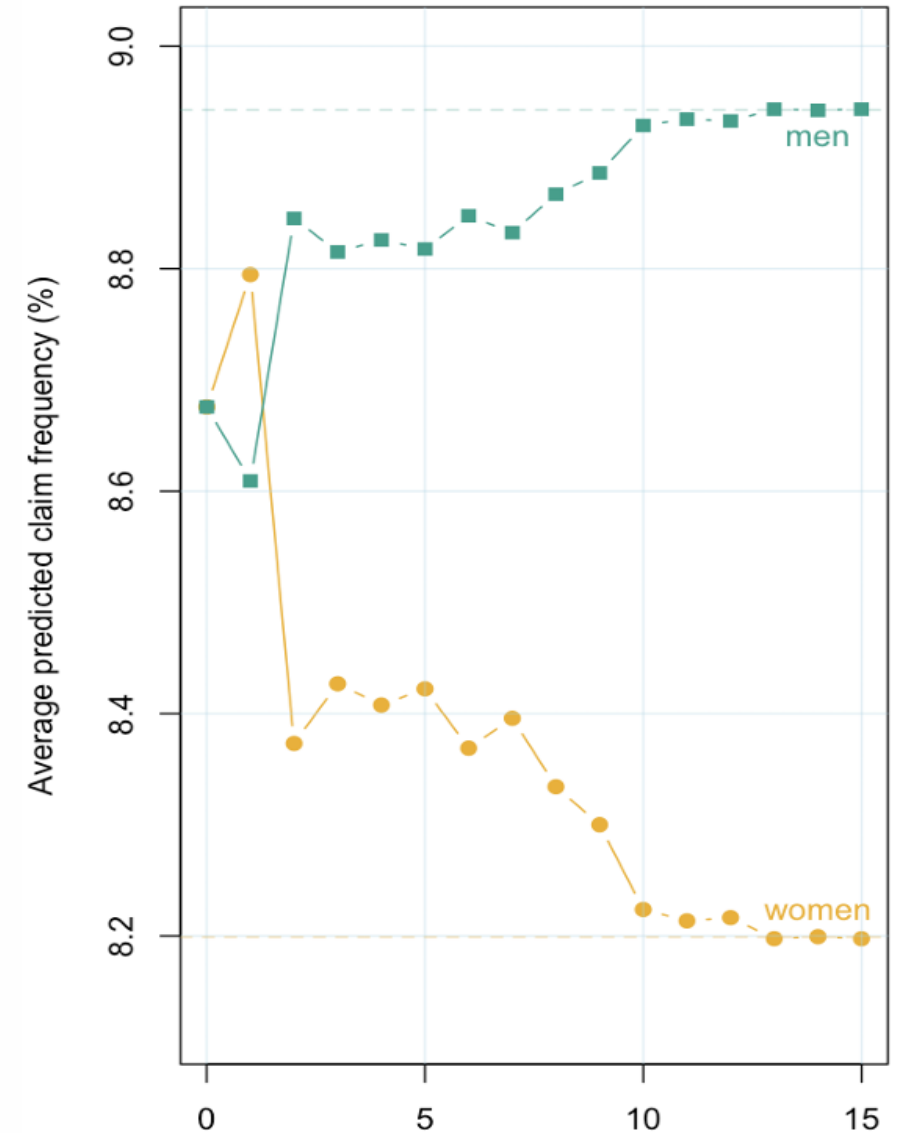
- **Disparité** : différence systématique dans le comportement, les sorties ou les performances d'un modèle entre des groupes.
- **Biais** : disparité traduisant un écart systématique par rapport à une norme d'évaluation pertinente (statistique, juridique, éthique)
- **Discrimination** : biais considéré comme inacceptable au regard du contexte d'usage, de la nature et de l'ampleur du préjudice (avéré ou potentiel), des droits affectés et du cadre juridique applicable.
 - Pour une autorité de contrôle, la norme sous-jacente est une **norme juridique** (critères protégés par le droit européen ou national)
 - Pour une institution financière, la norme peut également découler de sa politique interne (considérations éthiques, allant au-delà des exigences réglementaires)

2. CONCEPTS FONDAMENTAUX (2/2) : LES « PROXYs »

💡 Éliminer les variables sensibles ? (Équité par l'ignorance)

❌ **Insatisfaisant** : les modèles d'IA peuvent reconstruire les variables sensibles avec suffisamment de données grâce à des variables de substitution (*proxys*).

- Exemple : fréquence de déclaration de sinistre auto : 8,94 % (hommes) vs. 8,20 % (femmes).
 - Régression logistique pour estimer la fréquence de déclaration sur k variables, **en excluant le genre**
 - Dès $k = 2$, on voit une différence marquée.
 - À partir de $k = 10$, on a quasiment reconstruit la variable genre.



Source : Charpentier (2024).

3. L'ÉQUITÉ DE GROUPE : TROIS FAMILLES DE MÉTRIQUES

Critère d'équité	Indépendance (parité démographique)	Séparation (parité des taux d'erreur)	Suffisance (parité des valeurs prédictives)
Mesure	Même taux d'approbation entre groupes.	Mêmes taux de faux positifs et de faux négatifs entre groupes.	Mêmes valeurs prédictives positives et négatives entre groupes.
Principe	« La proportion d'emprunteurs approuvés est la même pour tous les groupes. »	« Parmi les emprunteurs viables, la proportion d'emprunteurs approuvés est la même pour tous les groupes. »	« Parmi les emprunteurs approuvés, la proportion d'emprunteurs viables est la même pour tous les groupes. »
Avantages	- Simplicité de mise en œuvre	- Ne pénalise pas l'accès au crédit des emprunteurs viables du groupe défavorisé	- Une prédiction a la même signification pour tous les groupes (même score = même niveau de risque)
Inconvénients	- Ne tient pas compte de la « viabilité » des emprunteurs	- Peut conduire à accepter un traitement différent d'individus similaires issus de groupes différents	- Risque d'exclusion accrue : critères plus stricts pour les groupes défavorisés



4. MISE EN ŒUVRE PRATIQUE (1/2) : CONSIDÉRATIONS GÉNÉRALES

- L'équité algorithmique n'est pas un sujet réservé aux seuls *data scientists* ; elle engage la **stratégie, la gestion des risques et la responsabilité** de l'institution financière :
 - Fixer les **grandes orientations** en matière d'équité
 - Puis les décliner au niveau technique en fonction des **différents cas d'usage métier**
 - Prendre en compte l'équité tout au long du **cycle de vie** (conception, validation, déploiement, suivi dans le temps)
- Exemple de **processus de réflexion structuré** :
 - **Quels biais** souhaite-t-on prévenir ou corriger ?
 - Préciser la **norme de référence** retenue pour caractériser une éventuelle discrimination
 - Expliciter son **niveau d'ambition** en matière d'équité (par exemple : ne pas introduire de biais supplémentaires, ou corriger plus activement les inégalités existantes ?)
 - **Quelles mesures techniques ou organisationnelles** mises en œuvre pour prévenir ou corriger ces biais?
 - Ces mesures introduisent-elles **de nouveaux risques** (performance, opacité, création d'autres biais...) ?
 - Comment **l'arbitrage entre ces différents effets** est-il réalisé ?



4. MISE EN ŒUVRE PRATIQUE (2/2) : QUESTIONS TECHNIQUES

Le document de réflexion discute plus en détail de certaines questions techniques pour fournir des **repères opérationnels** aux acteurs financiers :

Critères protégés et données sensibles	Analyse du statut de quelques variables usuelles dans le secteur financier (âge, genre, lieu de résidence, données de santé)
Détermination des groupes à comparer	Analyse univariée vs. analyse multivariée => l'analyse univariée est un minimum
Choix de la famille de métriques pertinente	Quelques éléments pour structurer le choix (règlementation, caractéristiques des données, erreurs les plus gênantes)
Seuils pertinents pour caractériser les biais	Pas de seuil universel : adopter une approche contextuelle
Prise en compte de l'incertitude	Intervalles de confiance ; estimations de l'écart minimal détectable entre les groupes
Correction des biais	3 familles de méthodes (pré-traitement, en cours de traitement, post-traitement) à choisir en fonction des données et de l'objectif visé



5. ANTICIPER L'ESSOR DE L'IA GÉNÉRATIVE DANS LE SECTEUR FINANCIER

- **Spécificité des systèmes génératifs :**

- Absence de cible univoque, sorties de nature symbolique ($\neq$ décision)
- Cela ne remet pas en cause la pertinence d'une analyse d'équité, mais en modifie les modalités

- **Origine des biais :**

- Apprentissage : un grand modèle de langage (LLM) construit un espace de représentation qui reflète la proximité entre les mots ; des associations sociales et des stéréotypes peuvent s'y ancrer
- Les techniques d' « alignement » permettent de réduire l'expression visible de certains biais mais agissent principalement en surface, sans transformer en profondeur la structure interne du modèle

- **L'évaluation des biais d'une IA générative** peut mobiliser 3 couches complémentaires (couche représentationnelle, couche comportementale, couche « allocative ») brièvement exposées dans le document de réflexion.



ÉQUITÉ : OUVERTURE D'UNE CONSULTATION PUBLIQUE

- Publication aujourd'hui du document de réflexion
- Ouverture d'une **consultation publique** jusqu'au 30 septembre
 - Questionnaire de consultation sur le site de l'ACPR
 - Pour contribuer : reglement-ia@acpr.banque-france.fr
- Synthèse de la consultation publique d'ici la fin 2026



TRAVAUX SUR L'EXPLICABILITÉ DE L'IA (1/2)

- L'explicabilité est un sujet qui présente deux facettes :

Production des
explications

- Approximation du comportement global ou local du modèle

Réception des
explications

- Communication de l'explication adaptée à l'audience visée

- Précédents travaux de l'ACPR sur l'explicabilité (dont le Tech Sprint de 2021) :
 - Le sujet est trop souvent considéré sous un angle uniquement technique : il est important de prendre en compte la réception, et donc l'audience visée
 - Il existe de multiples critères pour définir une « bonne explication » (précise, complète, actionnable, compréhensible, etc.).
 - Les explications peuvent également induire les humains en erreur.

TRAVAUX SUR L'EXPLICABILITÉ DE L'IA (2/2)



Revue de la littérature scientifique



Ateliers bilatéraux avec des acteurs volontaires du système financier

➤ Autour d'un ou plusieurs cas d'usage IA en production



Participation aux groupes de travail européens (MSU, EBA, EIOPA...)



Intéressés ? Ecrivez-nous :

Reglement-ia@acpr.banque-france.fr



**Document de réflexion
sur l'explicabilité
(Q4 2026 / Q1 2027)**



PROCHAINES ÉTAPES

- **A partir de l'été 2026 : ateliers bilatéraux sur l'explicabilité**
 - N'hésitez pas à candidater !
- **30 septembre 2026** : fin de la consultation publique sur l'équité
 - N'hésitez pas à contribuer !
- **Réunions de place** : la prochaine est prévue à l'automne 2026
- Page dédiée à la surveillance de l'IA sur le **site internet** de l'ACPR
 - Vidéos des réunions de place
 - Questions-réponses
- Des **suggestions** sur le contenu de cette page ou sur d'autres sujets ?
 - Écrivez-nous ! Reglement-ia@acpr.banque-france.fr



EXIGENCES APPLICABLES ET SUPERVISION

—

QUESTIONS - REPONSES





CONCLUSION

OLIVIER FLICHE

***DIRECTEUR DE L'INNOVATION, DES DONNÉES ET DES
RISQUES TECHNOLOGIQUES***





ANNEXES



LE CHAMP DE COMPÉTENCE PROBABLE DE L'ACPR

Fournisseur « amont »
d'IAUG

Modèle d'IA à usage
général

Fournisseur
« aval » d'IAUG

Système d'IA à **usage général**
pouvant être directement utilisé
pour au moins un cas d'usage
financier classé à haut risque

Déployeur
=
Institution financière

Système d'IA **reposant sur un
modèle d'IAUG** et utilisé pour un
cas d'usage financier classé à
haut risque

Déployeur
=
Institution financière

Fournisseur

Système d'IA « **classique** » pour
un cas d'usage financier classé à
haut risque

Déployeur
=
Institution financière