UCLouvain

UNIVERSITÉ DE NAMUR

Economics School of Louvain - ESL

Economics School of Namur - ESN

L'IMPACT DE L'INTELLIGENCE ARTIFICIELLE DANS LE SYSTEME BANCAIRE: UTILISATION DE L'IA DANS L'ÉVALUATION DU CREDIT-SCORING

Auteur: Mohamed Ali EDARAZI

Promoteur: Pr. Jean DESSAIN

Année académique 2025-2026

Master en Sciences économiques - 60 crédits

Université catholique de Louvain - Economics School of Louvain - Place Montesquieu 3, 1348 Louvain-La-Neuve
Université de Namur - Economics School of Namur - Rempart de la Vierge 8, 5000 Namur

Table des matières

INTRODUCTION GÉNÉRALE

Du jugement bancaire au scoring algorithmique : la transformation numérique de l'octroi de crédit

Depuis le début des années 2010, les grandes banques européennes ont largement automatisé l'évaluation du risque de crédit, mobilisant selon les établissements des moteurs réglementés, des modèles statistiques éprouvés et, plus récemment, des techniques de machine learning intégrées aux décisions d'octroi.

Considérons le cas de M. Dupont, qui dépose une demande de crédit à la consommation de 15 000 euros depuis son téléphone et se voit notifier d'eun refus instantané, sans entretien ni explication autre qu'une mention laconique de « solvabilité insuffisante ». Un algorithme de scoring a vraisemblablement agrégé des dizaines de variables, historique de crédit, comportements transactionnels, situation financière, profil de risque global (Baesens, Roesch et Scheule, 2016 ; Lessmann et al., 2015), sans que le demandeur en ait la moindre connaissance. Cette situation soulève plusieurs interrogations fondamentales : sur quelles données le modèle a-t-il été entraîné ? Les variables retenues respectent-elles le droit européen de la non-discrimination ? M. Dupont peut-il obtenir une explication intelligible de ce refus ? La banque sera-t-elle en mesure de justifier la robustesse du modèle auprès du superviseur ?

Le crédit n'a pas toujours fonctionné selon cette logique. Le relationship lending, tel que conceptualisé par Berger et Udell (2002), reposait sur la connaissance personnelle du client et le jugement individuel du banquier. L'introduction de la régression logistique dans les années 1960 a standardisé et statisticisé la décision, réduisant sa dépendance à l'appréciation subjective (Thomas, Edelman et Crook, 2002). Le machine learning, déployé à partir du milieu des années 2010, a ensuite apporté des gains de performance prédictive significatifs, confirmés par plusieurs études empiriques (Lessmann et al., 2015 ; Moscato et al., 2021). C'est cette trajectoire que la première partie retrace.

Contexte : l'essor de l'intelligence artificielle dans les banques européennes

Le crédit à la consommation constitue un marché hautement concurrentiel, où les grandes banques universelles font face tant aux spécialistes historiques (Cofidis, Cetelem, Sofinco) qu'à une vague de nouveaux entrants fintech (Klarna, N26, Revolut, Younited Credit) proposant des parcours d'octroi entièrement digitaux, conclus en quelques minutes. La crise COVID a accéléré la transformation numérique des établissements traditionnels et approfondi l'automatisation de leurs décisions (Vives, 2019 ; Boot, Hoffmann, Laeven et Ratnovski, 2021).

Les régulateurs ont accompagné ce mouvement. Sans remettre en cause le principe de l'automatisation, l'EBA, la BCE et le MSU convergent dans leurs travaux récents vers une

exigence commune : lorsqu'un algorithme intervient dans une décision de crédit, ses fondements doivent pouvoir être documentés et justifiés devant le superviseur.

La spécificité européenne : un cadre réglementaire dense

Les États-Unis ont inventé le credit scoring automatisé (Mester, 1997) ; l'Europe a choisi de l'encadrer strictement, produisant l'un des corpus réglementaires les plus denses au monde en la matière, articulé autour de quatre textes.

L'article 22 du RGPD restreint les décisions fondées exclusivement sur un traitement automatisé produisant des effets juridiques significatifs, couvrant directement le scoring de crédit (Wachter, Mittelstadt et Floridi, 2017 ; Kaminski, 2019). La directive (UE) 2023/2225 relative au crédit aux consommateurs, applicable à compter du 20 novembre 2026, impose en son article 18 un droit à l'intervention humaine et à une explication intelligible lorsque l'évaluation de la solvabilité est automatisée. L'AI Act (Règlement (UE) 2024/1689) classe les systèmes de scoring en tant que systèmes à haut risque, avec une première échéance au 2 août 2026. Le corpus prudentiel, CRR, CRD et supervision intégrée du MSU, complète le dispositif en conférant à la BCE un contrôle direct sur les modèles de risque des banques significatives.

Ce cadre laisse peu de latitude : les banques européennes peuvent automatiser, mais doivent être en mesure de rendre compte de chaque décision produite par leurs modèles.

Problématique, hypothèse et justification de la recherche

La question centrale de ce mémoire est la suivante : dans quelle mesure les banques de détail européennes peuvent-elles concilier performance prédictive et conformité réglementaire lors du déploiement de modèles de machine learning pour l'octroi de crédit à la consommation ?

La formulation « dans quelle mesure » signale d'emblée que la réponse ne sera pas binaire. La question se pose dans un contexte d'incertitude persistante quant aux conditions d'équité et d'absence de discrimination que requiert cette innovation, désormais intégrées pour la première fois dans l'environnement réglementaire européen, qui consacre le droit fondamental à la non-discrimination comme exigence non-negociable et non comme simple considération technique.

La thèse défendue est que conformité réglementaire et performance prédictive ne sont pas incompatibles, mais que leur articulation n'est pas automatique. Elle ne résulte pas du choix d'un algorithme particulier, mais de ce qui est construit autour : documentation rigoureuse, validation indépendante, explicabilité des décisions, intervention humaine effective et audit régulier des biais. Sans cette gouvernance du cycle de vie des modèles, tout algorithme, aussi performant soit-il, devient un vecteur de risque.

Sur le plan académique, ce travail cherche à faire dialoguer trois corpus qui se citent encore trop peu : la littérature technique sur la performance prédictive (Lessmann et al., 2015), la littérature FAccT sur les biais et l'équité algorithmique (Barocas et Selbst, 2016 ; Chouldechova, 2017), et une littérature juridique en plein essor sur le RGPD et l'AI Act, encore largement ignorée des chercheurs en machine learning (Wachter et al., 2017 ; Hacker, 2018 ; Veale et Borgesius, 2021).

Sur le plan professionnel, les banques font face à un arbitrage structurel : les modèles les plus performants sont souvent les moins interprétables, donc les plus vulnérables aux exigences

d'explicabilité du droit européen (Rudin, 2019). Cette tension entre précision prédictive et justifiabilité n'est pas théorique, elle est quotidienne dans les équipes risque.

Sur le plan sociétal, enfin, un scoring mal construit peut durablement exclure certaines populations de l'accès au crédit, en reproduisant ou amplifiant des iniquités structurelles, ce que prohibe explicitement l'article 21 de la Charte des droits fondamentaux de l'UE (Bartlett, Morse, Stanton et Wallace, 2022 ; Fuster et al., 2022).

La tension clé : un trilemme performance-conformité-éthique.

Le scoring algorithmique soumet les banques à trois exigences dont les logiques ne convergent pas toujours.

La performance prédictive désigne la capacité du modèle à discriminer les emprunteurs selon leur probabilité de remboursement. La conformité réglementaire impose le respect des exigences de transparence, d'explicabilité, de contrôle humain et de gouvernance prescrites par le RGPD, la Directive 2023/2225, l'AI Act et le cadre prudentiel. L'acceptabilité éthique, enfin, soulève la question des variables proxy : des indicateurs en apparence anodins, tels que le code postal ou la catégorie socio-professionnelle, peuvent en pratique recouvrir des attributs protégés par la loi, comme l'origine, le sexe ou le handicap. Le modèle n'y fait pas référence directement, mais en reproduit les effets discriminatoires, un mécanisme documenté dans le contexte du crédit par Bartlett et al. (2022) et Prince et Schwarcz (2020).

La littérature a par ailleurs établi qu'équité et précision peuvent entrer en conflit structurel : améliorer l'une revient parfois à dégrader l'autre. Chouldechova (2017) puis Corbett-Davies et Goel (2018) ont formalisé ces tensions de manière rigoureuse. Ce mémoire en prend acte, non pour les trancher, mais pour comprendre quels arbitrages les banques opèrent concrètement et avec quels outils elles cherchent à les atténuer.

Méthodologie et grille d'analyse

1. Méthodologie de la revue de littérature

La revue de littérature a été conduite selon une méthode systématique et reproductible. Les bases de données mobilisées sont Scopus, Web of Science, Google Scholar, SSRN, HAL, Cairn et JSTOR, interrogées à l'aide de mots-clés en français et en anglais combinés par opérateurs booléens : "credit scoring", "machine learning", "fairness", "AI Act", "RGPD", "explicabilité", "XAI", "biais algorithmique", "algorithmic bias", "conformité réglementaire", "regulatory compliance". La période couverte s'étend de 2015 à 2026, avec une actualisation ponctuelle pour les textes réglementaires les plus récents.

Les critères d'inclusion retiennent les articles issus de revues à comité de lecture, les rapports d'institutions de référence (EBA, BCE, BRI, Commission européenne) et les working papers de centres de recherche reconnus. Sont exclus les articles non académiques non vérifiés et toute source dépourvue de peer-review pour les affirmations techniques ou juridiques. La hiérarchisation adoptée est la suivante : les articles peer-reviewed constituent la source prioritaire pour les développements théoriques et empiriques ; les rapports institutionnels documentent les positions réglementaires et prudentielles ; les sources sectorielles et échanges avec praticiens n'interviennent qu'à titre illustratif, sans valeur démonstrative.

2. Méthode d’analyse documentaire des textes réglementaires

L'analyse des textes réglementaires repose sur une analyse thématique comparative appliquée à quatre corpus : le RGPD (2016/679), l'AI Act (2024/1689), la Directive 2023/2225 relative au crédit à la consommation, et le paquet prudentiel CRR3/CRD6. Une grille de lecture commune à cinq dimensions a été appliquée à chacun : (1) entrée en vigueur et date d'application, (2) champ d'application matériel et personnel, (3) obligations principales pesant sur les établissements de crédit, (4) autorités de contrôle compétentes, (5) régime de sanctions et voies de recours. La comparaison est restituée sous forme d'un tableau synthétique faisant apparaître les similitudes, lacunes et tensions entre les différents régimes applicables au credit scoring algorithmique.

3. Statut des échanges avec praticiens

Deux entretiens exploratoires ont été réalisés : l'un auprès d'un responsable risque crédit au sein d'une banque universelle européenne, l'autre auprès d'un consultant en conformité réglementaire. Conduits en personne entre Fevrier et Mai 2026 (durée moyenne : 45 minutes), ils s'articulaient autour de trois thèmes : (1) la mise en œuvre des algorithmes de machine learning dans les processus de décision de crédit, (2) les contraintes réglementaires perçues par les équipes opérationnelles, et (3) les arbitrages entre performance prédictive et exigences de conformité. L'exploitation a été réalisée par analyse thématique manuelle, avec anonymisation systématique des interlocuteurs et des établissements.

Il convient d'insister sur les limites de ce volet : deux entretiens ne constituent pas une base empirique suffisante pour produire des résultats généralisables, ni pour prétendre à une quelconque représentativité statistique. Ils interviennent en complément de l'analyse documentaire, à titre illustratif, pour confronter certaines hypothèses de la littérature aux contraintes opérationnelles réelles, sans vocation à valider ou infirmer les thèses du mémoire.

L'analyse s'organise autour de trois dimensions transversales, reconduites dans chacune des parties du mémoire : la performance prédictive (AUC-ROC, coefficient de Gini, calibration des probabilités de défaut, stabilité temporelle du modèle) ; la conformité réglementaire (explicabilité, minimisation des données, supervision humaine, documentation technique et auditabilité) ; l'équité et l'acceptabilité (détection des biais, traitement des variables proxy, mécanismes de recours, exigences de non-discrimination). Cette grille constitue le fil conducteur permettant de tester l'hypothèse centrale du mémoire.

Déclaration d’utilisation de l’intelligence artificielle générative

Ce mémoire a eu recours à l'outil d'intelligence artificielle générative Claude (Anthropic) de manière ponctuelle, pour trois usages distincts : (1) la reformulation de certains passages, (2) la relecture et la correction linguistique, (3) la structuration de certaines sections. L'ensemble des suggestions produites par l'outil a été systématiquement vérifié, validé et, le cas échéant, corrigé par l'auteur, qui assume l'entière responsabilité du contenu, de l'argumentation et des conclusions du mémoire. Les passages les plus sensibles, notamment l'analyse réglementaire, l'interprétation des sources et le positionnement académique, n'ont pas été produits par l'IA et ont fait l'objet d'un contrôle particulier.

Annonce du plan

La première partie retrace l'évolution technique du credit scoring, des modèles statistiques classiques au machine learning, et examine les conditions dans lesquelles les banques européennes déploient aujourd'hui ces outils. La deuxième partie analyse les risques que ces modèles génèrent : opacité structurelle, reproduction de biais, et conséquences juridiques et stratégiques pour les établissements. La troisième partie examine les obligations réglementaires découlant du RGPD, de la Directive 2023/2225, de l'AI Act et du cadre prudentiel, avant d'identifier les leviers dont disposent les banques pour y répondre : explicabilité algorithmique (XAI), gouvernance du risque modèle, contrôle humain et audits de biais.

Délimitation du champ

Géographique : l'Union européenne à 27, avec références comparatives ponctuelles aux États-Unis et au Royaume-Uni.

Thématique : exclusivement le credit scoring appliqué aux particuliers dans le cadre du crédit à la consommation, à l'exclusion du trading algorithmique, de la détection de fraude, des chatbots, de la lutte anti-blanchiment ou de la gestion d'actifs.

Méthodologique : approche théorique et analytique complétée par des échanges exploratoires, sans modélisation empirique originale sur données réelles.

Temporelle : 2018-2026, portée par l'entrée en vigueur du RGPD en mai 2018, la date d'application de la Directive 2023/2225 en novembre 2026 et des obligations de l'AI Act sur les systèmes à haut risque en août 2026.

Matrice de synthèse : articulation de la démonstration

Dimension de la recherche	**Question abordée**	**Littérature mobilisée**	**Cadre réglementaire**	**Résultat attendu**
Partie I : Transformation technique et opérationnelle	Comment le ML transforme-t-il le scoring ?	Lessmann et al. (2015), Baesens et al. (2016), Berg et al. (2020)	PSD2, Open Banking	Cartographie des gains et vulnérabilités
Partie II : Risques structurels	Quels biais et quelle opacité ?	Barocas et Selbst (2016), Rudin (2019), Chouldechova (2017)	RGPD art. 22, Charte des droits fondamentaux	Typologie des risques
Partie III : Conditions de conciliation	Quels leviers pour concilier les trois exigences ?	Wachter et al. (2017), Kozodoi et al. (2022)	AI Act, Directive 2023/2225, CRR3/CRD6	Architecture de gouvernance

Conclusion	Dans quelle mesure la conciliation est-elle possible ?	Ensemble du corpus	Ensemble du corpus	Gestion dynamique du trilemme

Cette matrice explicite le fil conducteur du mémoire. Chaque partie est construite pour répondre à un aspect de la problématique, en mobilisant un sous-ensemble spécifique de la littérature et du cadre réglementaire, et en produisant un résultat qui alimente la partie suivante. L'ensemble converge vers la thèse défendue : le trilemme ne se résout pas, il se gouverne.

PARTIE I : La transformation technique et opérationnelle du credit scoring

Cette première partie vise à montrer comment les techniques de machine learning et la numérisation des données ont réformé le credit scoring bancaire, pas seulement sur le plan de l'algorithme, mais aussi dans la gestion des infrastructures bancaires, des données et des logiques de décision. Les deux chapitres qui suivent posent les bases techniques et opérationnelles nécessaires à l'analyse des risques éthiques et réglementaires conduite dans les parties suivantes.

Chapitre 1 : Du scoring traditionnel au scoring algorithmique

Ce premier chapitre retrace la genèse du credit scoring, des approches statistiques anciennes aux modèles de machine learning d'aujourd'hui dans les banques. Passage technologique, mais non seulement technologique puisqu'ont changé dans les mêmes temps donnés mobilisées, designs décisionnels et métriques évaluatives. Ces éléments techniques sont la base du cadre au sein duquel les chapitres suivants traitent des tensions entre efficacité prédictive, contraintes légales et exigences éthiques.

1.1 Les fondements du scoring traditionnel : FICO, régression logistique et scorecards

1.1.1 Aux origines du credit scoring : de l'intuition du banquier à la formalisation statistique

Le credit scoring moderne est né aux États-Unis dans les années 1950, dans un contexte d'explosion du crédit à la consommation que les banques peinent à traiter manuellement. En 1956, Bill Fair et Earl Isaac fondent Fair, Isaac and Company, future FICO, avec l'idée de substituer des modèles mathématiques calibrés sur données historiques au jugement des analystes (Thomas, Edelman et Crook, 2002 ; Mester, 1997). L'intuition fondatrice est que le modèle statistique, en appliquant les mêmes critères à chaque dossier, corrige l'incohérence et la variabilité inhérentes au jugement humain (Hand et Henley, 1997 ; Siddiqi, 2006).

Le score FICO, dont l'adoption massive date des années 1980 et 1990, est une réalité avant tout américaine, portée par la croissance des demandes de crédit, les exigences d'objectivité issues de l'Equal Credit Opportunity Act (ECOA, 1974) et la montée en puissance des capacités informatiques (Thomas et al., 2002). Il se présente publiquement comme un nombre entre 300 et 850, construit à partir de cinq catégories de facteurs : historique de paiement (35 %), encours

total (30 %), ancienneté de l'historique (15 %), types de crédit détenus (10 %) et demandes récentes (10 %) (Mester, 1997 ; FICO, s.d.). Ces proportions restent toutefois approximatives : l'algorithme réel est propriétaire et n'a jamais été rendu public.

En Europe, le scoring statistique ne s'est généralisé qu'à partir des années 1990, de manière inégale. Le marché bancaire européen demeure découpé en espaces nationaux distincts, chacun doté de ses propres registres de crédit : la Schufa en Allemagne, le FICP géré par la Banque de France, le Central Credit Register de la Banque Nationale de Belgique, le CRIF en Italie (Thomas et al., 2002 ; Anderson, 2007). Cette architecture morcelée a durablement freiné l'émergence de modèles paneuropéens et explique pourquoi les banques européennes ont longtemps privilégié des approches relationnelles fondées sur la connaissance directe du client plutôt que sur des scores standardisés (Berger et Udell, 2002 ; Degryse et Ongena, 2005).

.

1.1.2 Périmètre de l'étude : application scoring vs autres formes de scoring

Un point de méthode s'impose avant d'aller plus loin. La littérature et la pratique bancaire distinguent plusieurs formes de scoring, qui ne répondent pas aux mêmes objectifs. L'application scoring intervient au moment de la demande de crédit, pour décider de l'octroi. Le behavioral scoring suit l'évolution du risque d'un client tout au long de la vie du crédit. Le collection scoring a pour rôle de hiérarchiser les actions de recouvrement en cas d'incident. Le fraud scoring, enfin, sert à la détection de la fraude. Ce mémoire s'intéresse à la problématique du scoring dans le crédit à la consommation, c'est-à-dire l'évaluation du risque à l'instant de la demande. Le behavioral scoring sera évoqué par endroits, où sa mention est éclairante. Le fraud scoring est exclu du périmètre : ses logiques techniques et réglementaires sont assez distinctes pour justifier un traitement séparé.

1.1.3 Le modèle de régression logistique : socle du scoring moderne

La régression logistique a dominé le credit scoring bancaire pendant plus de quarante ans et demeure une référence dans de nombreuses banques de détail européennes, notamment pour les modèles réglementaires et les scorecards où l'interprétabilité est aussi déterminante que la performance. Son principe repose sur la classification supervisée : une fonction logistique transforme la combinaison linéaire de variables explicatives en une probabilité de défaut (PD) comprise entre 0 et 1 (Hand et Henley, 1997 ; Thomas et al., 2002). Ce que le modèle cède en flexibilité par rapport aux approches plus récentes, il le compense en lisibilité : chaque coefficient est interprétable, chaque décision justifiable, ce qui en a fait un outil naturellement compatible avec les exigences réglementaires bien avant leur formalisation.

Formellement, le modèle s'écrit :

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

où : Y est la variable dépendante binaire (1 = défaillance, 0 = non-défaillance), X1,…,Xk les variables indépendantes, et β0,…,βk les coefficients estimés par maximum de vraisemblance. Chaque coefficient βi correspond au log-odds ratio associé à Xi, ce qui permet de lire directement le poids de chaque variable dans la décision finale. Cette propriété a largement contribué à la durabilité du modèle : un analyste peut expliquer pourquoi un dossier a été refusé, un auditeur

peut vérifier la cohérence des pondérations, un superviseur peut s'assurer que les critères retenus sont conformes aux exigences réglementaires (Hand et Henley, 1997 ; Lessmann et al., 2015).

Les variables classiquement utilisée dans les modèles de régression logistique appliqués au credit scoring sont, par ordre d'importance décroissante : revenus mensuels nets, taux d'endettement, ancienneté professionnelle, type de contrat de travail, historique de remboursement, ancienneté de la relation bancaire, situation familiale, statut résidentiel, montant et durée du crédit sollicité (Thomas et al., 2002 ; Baesens et al., 2016 ; Siddiqi, 2006). Ces informations sont structurées et bien documentées, mais leur apparente neutralité mérite d'être questionnée : la situation familiale, le statut résidentiel ou la nature du contrat de travail peuvent produire des effets discriminatoires indirects, une question qui sera développée en deuxième partie au regard des principes de proportionnalité, de minimisation des données et de non-discrimination.

La régression logistique présente plusieurs atouts qui expliquent sa résistance face aux approches plus récentes : transparence et interprétabilité directe des coefficients, stabilité des prévisions d'un échantillon à l'autre, et parcimonie limitant le risque de sur-apprentissage (Hand et Henley, 1997).

Ses limites sont néanmoins structurelles. L'hypothèse de linéarité, qui postule que chaque variable agit de façon linéaire sur le log-odds, ne tient pas dans tous les cas : la relation entre revenu et risque de défaut présente des effets de seuil que le modèle standard ne capte pas (Khandani, Kim et Lo, 2010). Les interactions entre variables ne sont pas détectées automatiquement et doivent être construites manuellement, ce qui implique un feature engineering long, coûteux et fortement dépendant des choix du modélisateur (Lessmann et al., 2015 ; Baesens et al., 2016).

.

1.1.4 Le scoring expert : le jugement humain structuré

Avant les modèles statistiques, et souvent en parallèle de leur introduction, les banques ont recouru à des systèmes de jugement expert structuré. Le principe consiste à codifier le savoir-faire des analystes crédit expérimentés sous forme de règles conditionnelles (« si…, alors… ») organisées en grilles ou arbres de décision, sans algorithme ni optimisation statistique. Par exemple : « Si taux d'endettement > 35% ET ancienneté professionnelle < 12 mois, ALORS refus » (Siddiqi, 2006 ; Thomas et al., 2002).

L'atout distinctif du scoring expert réside dans sa capacité à intégrer de la soft information : qualité de la relation bancaire, contexte économique local, événement exceptionnel dans la vie du demandeur, autant d'éléments qu'un analyste expérimenté peut pondérer et qu'aucun coefficient statistique ne capte véritablement (Berger et Udell, 2002 ; Petersen et Rajan, 2002). Pour les dossiers limites, cette capacité de nuance peut s'avérer déterminante.

Ses limites sont néanmoins bien documentées. Deux analystes confrontés au même dossier peuvent parvenir à des conclusions différentes, sans qu'aucun calibrage statistique ne permette de vérifier si les règles produisent de bonnes décisions en moyenne. Surtout, le modèle ne passe pas à l'échelle : à mesure que les volumes augmentent avec la digitalisation, l'instruction manuelle des dossiers devient un goulot d'étranglement structurel (Siddiqi, 2006 ; Hand et Henley, 1997).

1.1.5 Les scorecards : l'outil opérationnel dominant

De la tension entre rigueur statistique et lisibilité opérationnelle est née la scorecard : une grille standardisée qui traduit les coefficients d'un modèle statistique en points attribués à chaque modalité de variable, dont la somme constitue le score final. Elle demeure aujourd'hui l'outil de décision le plus répandu dans les banques de détail (Siddiqi, 2006).

Le principe repose sur un découpage de chaque variable en catégories, par exemple les revenus en tranches (moins de 1 500 €, 1 500-2 500 €, 2 500-4 000 €, plus de 4 000 €), auxquelles sont associés des points issus d'une transformation logarithmique standardisée des coefficients logistiques (Siddiqi, 2006 ; Anderson, 2007 ; Thomas et al., 2002). L'avantage sur le modèle statistique brut est opérationnel : un analyste peut lire et interpréter le résultat directement, sans recourir à la console.

La construction des scorecards mobilise deux indicateurs standards. Le Weight of Evidence (WoE) mesure, pour chaque modalité, sa capacité à discriminer bons et mauvais payeurs. L'Information Value (IV) agrège ces mesures en un score synthétique pour l'ensemble de la variable. Ces indicateurs présentent plusieurs avantages pratiques : ils imposent une relation monotone entre variable et risque, ce qui renforce la défendabilité lors de la validation ; ils améliorent la stabilité en production ; et ils produisent une documentation directement exploitable par les équipes de contrôle, y compris dans les environnements IRB où les exigences de traçabilité sont élevées (Siddiqi, 2006 ; Anderson, 2007).

Exemple simplifié de scorecard :

Variable	Modalité	Points
Revenus mensuels nets	< 1 500 €	15
	1 500 - 2 500 €	35
	2 500 - 4 000 €	55
	> 4 000 €	70
Ancienneté professionnelle	< 1 an	10
	1 - 3 ans	25
	3 - 10 ans	40
	> 10 ans	50
Incidents de remboursement	Aucun	60
	1 incident	25
	2 incidents ou plus	5

La classe de crédit résultant du score total est comparée à un cut-off : si le score est supérieur au cut-off, le crédit est autorisé, s'il est inférieur, il peut être refusé ou à l'inverse être soumis à un examen humain.

La pérennité des scorecards repose sur un ensemble d'atouts qui expliquent la pérennité des scorecards. Leur structure additive est lisible par l'analyste : d'un coup d'œil, on identifie les variables contributrices, à quel sens, pour quel poids. Cette lisibilité est aussi un outil de

conformité réglementaire : l'utilisation des modèles internes IRB a été explicitement subordonnée dans Bâle II d'abord, et dans Bâle III ensuite, à disposer de documentation, de validation et d'audibilités que les scorecards valident aussi aisément. (European Banking Authority, 2021 ; Siddiqi, 2006).

1.2 L'apport des algorithmes de machine learning

1.2.1 Les forêts aléatoires (Random Forests)

Formalisée par Leo Breiman en 2001, la forêt aléatoire repose sur l'entraînement d'un grand nombre d'arbres de décision indépendants plutôt que d'un seul. Deux mécanismes produisent la diversité nécessaire à la robustesse de l'ensemble : le bagging, par lequel chaque arbre est entraîné sur un sous-échantillon tiré avec remise, et la sélection aléatoire de variables à chaque nœud, qui empêche les variables dominantes de structurer l'ensemble des arbres. La prédiction finale agrège les résultats par vote majoritaire en classification ou par moyenne en régression. Breiman démontre que lorsque le nombre d'arbres croît, l'erreur de généralisation converge vers une limite dépendant de deux facteurs en équilibre : la qualité individuelle de chaque arbre et le degré de corrélation entre eux (Breiman, 2001).

Cette double injection de hasard produit des effets substantiels. Le modèle capte des relations non linéaires et des interactions entre variables que la régression logistique ne détecterait pas. L'agrégation d'arbres faiblement corrélés réduit la variance globale sans dégrader le biais, conférant au modèle une robustesse au sur-apprentissage que peu d'algorithmes de l'époque pouvaient revendiquer (Breiman, 2001 ; Lessmann et al., 2015).

Appliquées au credit scoring, les forêts aléatoires présentent trois atouts concrets : capture des non-linéarités et interactions sans spécification préalable, robustesse aux valeurs aberrantes, et production native de mesures d'importance des variables, utiles pour justifier une décision de crédit (Lessmann et al., 2015 ; Baesens et al., 2016).

La qualification de boîte noire mérite néanmoins d'être nuancée. Des outils tels que les mesures d'importance des variables ou les graphiques de dépendances partielles permettent d'interroger le comportement global du modèle, sans pour autant restituer le raisonnement à l'origine de chaque prédiction individuelle. La forêt aléatoire occupe ainsi une position intermédiaire : moins transparente que la régression logistique, mais nettement moins hermétique que les réseaux de neurones profonds.

1.2.2 Le Gradient Boosting et XGBoost : une référence en performance prédictive sur données tabulaires

Les machines à gradient boosté, et en particulier XGBoost (Chen et Guestrin, 2016), constituent une autre famille algorithmique ayant acquis une réputation solide sur données tabulaires. Leur principe diffère fondamentalement de celui des forêts aléatoires : plutôt que de construire des arbres en parallèle pour agréger leurs votes, le boosting les construit séquentiellement, chaque arbre corrigeant les erreurs du précédent en accentuant le poids des observations mal prédites. À chaque itération, un nouvel arbre est ajouté pour minimiser la perte en suivant le gradient négatif

des résidus, produisant une logique d'amélioration itérative qui explique la précision souvent supérieure du boosting sur données structurées (Friedman, 2001 ; Chen et Guestrin, 2016).

XGBoost se distingue par plusieurs innovations techniques : gestion native des données creuses, compression du calcul des quantiles, optimisation des accès cache, et régularisation L1 et L2 intégrée directement dans la fonction objectif. Cette dernière répond au risque historique de sur-apprentissage inhérent au boosting, qui, en s'acharnant sur les cas difficiles, tend à mémoriser le bruit plutôt que le signal. Ces caractéristiques permettent au modèle de passer à l'échelle sur des volumes de données inaccessibles aux implémentations classiques.

Dans le contexte du credit scoring, sa capacité à traiter des données tabulaires hétérogènes avec peu de prétraitement en fait un choix naturel. L'EBA retient les techniques de gradient boosting, XGBoost, LightGBM et CatBoost, parmi les candidats sérieux au scoring bancaire en Europe (EBA, 2021 ; 2023), pour des raisons pragmatiques : rapport favorable entre précision prédictive, capacité d'absorption de données hétérogènes et coût de calcul.

L'EBA pointe néanmoins un paradoxe structurel : ces algorithmes permettent d'exploiter des flux de données plus riches et plus granulaires, mais cette capacité a un prix. Plus le modèle est performant, moins il est lisible. La complexité croissante n'est pas un détail technique, c'est un problème de gouvernance.

1.2.3 Les réseaux de neurones et le deep learning : puissance et opacité

Au milieu des années 2010, les réseaux de neurones profonds ont introduit une rupture majeure. LeCun, Bengio et Hinton (2015) ont formalisé des architectures multicouches capables d'apprendre des représentations hiérarchiques des données sans spécification préalable des structures à rechercher. À chaque couche, les données sont traitées de façon non linéaire et transmises à la suivante, produisant des représentations de plus en plus abstraites. Le théorème d'approximation universelle établit qu'un réseau à couche cachée suffisamment large peut approcher n'importe quelle fonction continue, ce qui explique leurs performances sur des tâches complexes (LeCun et al., 2015).

Dans le credit scoring, cet enthousiasme se heurte à des réalités prosaïques. Gunnarsson, Vanden Broucke et Baesens (2021) ont montré que les gains de performance par rapport au gradient boosting restent faibles sur données tabulaires, malgré une complexité et un coût computationnel nettement supérieurs. Le contexte réglementaire européen ajoute une contrainte rédhibitoire : un modèle inexplicable est un modèle indéfendable devant un régulateur.

Le problème central est celui de l'opacité. Avec des milliers de paramètres, la décision devient impossible à reconstituer, sans équivalent aux coefficients d'une régression ou aux règles d'un arbre de décision. Des outils comme LIME ou SHAP tentent d'éclairer les mécanismes internes, mais ce sont des approximations dont la fiabilité reste débattue (Rudin, 2019).

Cette opacité freine l'adoption pour trois raisons distinctes : elle complique la validation réglementaire, point souligné explicitement par l'EBA (2021) ; elle entre en tension avec les exigences de transparence et de contrôle humain du RGPD, de la Directive 2023/2225 et de l'AI Act (Wachter et al., 2017) ; elle rend enfin plus difficile le diagnostic des erreurs et des biais, qui deviennent structurels dans le contexte du credit scoring.

1.2.4 Autres algorithmes dignes d'attention : SVM et modèles d'ensemble avancés

Deux autres familles méritent une mention. Les machines à vecteurs de support (SVM), formalisées par Vapnik dans les années 1990, reposent sur la recherche de l'hyperplan séparant les classes en maximisant la marge entre les observations les plus proches, les support vectors. Le kernel trick étend ce principe aux frontières non linéaires sans surcoût computationnel majeur. En credit scoring, les SVM atteignent parfois des performances comparables aux forêts aléatoires ou au gradient boosting, mais sans les surpasser systématiquement, pour un effort de calibration plus élevé et une interprétabilité moindre : le rapport effort-performance leur est globalement défavorable (Lessmann et al., 2015).

Les modèles d'ensemble de dernière génération, stacking et blending, mobilisent un méta-modèle qui synthétise les prédictions de plusieurs modèles hétérogènes. Si leur potentiel théorique est réel, leur valeur ajoutée en production bancaire reste limitée : la complexité et l'opacité qu'ils introduisent excèdent généralement les gains marginaux de performance qu'ils apportent (Lessmann et al., 2015).

1.2.5 Comparaison synthétique des arbitrages

Critère	**Régression logistique**	**Random Forest**	**XGBoost / GBM**	**Deep Learning**
Performance prédictive	Référence robuste	Supérieure dans de nombreux cas	Très élevée sur données tabulaires	Très élevée sur données tabulaires
Interprétabilité	Forte (coefficients)	Moyenne (importance des variables)	Faible à moyenne	Faible
Validation réglementaire	Aisée	Plus complexe	Complexe	Très complexe
Robustesse au sur-apprentissage	Bonne	Très bonne	Bonne (avec régularisation)	Variable (risque élevé)
Coût computationnel	Très faible	Modéré	Modéré	Élevé

Capture des non-linéarités	Faible	Bonne	Très bonne	Excellente

En situation de mise en œuvre réelle des modèles avancés, leur valeur ajoutée apparait souvent plus limitée que prévue. Il résulte de l'étude comparative de Lessmann et al. (2015), portant sur 41 jeux de données et 8 familles d'algorithmes, que le machine learning surpasse en moyenne la régression logistique, mais que l'écart de performance (mesuré en AUC) reste modeste sur des jeux de données structurés standards, et varie significativement selon la taille de l'échantillon et la nature des variables. La complexité apportée ne se traduisant pas nécessairement par des gains mesurables.

Sources : synthèse de l'auteur à partir de Lessmann et al. (2015) - 41 jeux de données, 8 familles d'algorithmes, métrique AUC ; Barboza et al. (2017) - données de défaut d'entreprises, modèles ML vs logit ; Gunnarsson et al. (2021) - deep learning vs gradient boosting sur données tabulaires ; Moscato et al. (2021) - benchmark de 15 classifieurs sur 5 jeux de données de crédit.

Dans un écosystème où transparence, interprétabilité et conformité réglementaire constituent des contraintes dures, toute progression de performance doit être confrontée au coût en lisibilité et en auditabilité qu'elle induit. La réponse varie selon les établissements, les volumes traités, le cadre réglementaire et l'appétit au risque. Un constat demeure néanmoins constant : les gains restent modestes, la complexité est réelle, et ce déséquilibre explique en partie la résistance de la régression logistique là où on aurait pu s'attendre à la voir supplantée.

1.3 Les nouvelles sources de données mobilisées

1.3.1 Les données comportementales et transactionnelles

La transformation du credit scoring par le machine learning ne se joue pas seulement au niveau des algorithmes : elle tient aussi à l'élargissement considérable des données mobilisées. Là où les modèles traditionnels travaillaient sur un ensemble restreint de variables structurées (revenus, historique de crédit, ratios d'endettement), les modèles ML ouvrent un périmètre bien plus large, regroupé sous le terme de données alternatives (Berg, Burg, Gombović et Puri, 2020 ; EBA, 2020b).

Parmi celles-ci, les données comportementales financières occupent une place centrale : fréquence des transactions, habitudes d'épargne, fluctuations de revenus, dépassements de découvert autorisé. Ces signaux révèlent des schémas que les méthodes traditionnelles ne captaient pas. Ce n'est pas l'intérêt pour ces données qui a manqué, mais la capacité technique à les exploiter à grande échelle, que le machine learning a rendue possible.

La véritable rupture en Europe est cependant d'ordre réglementaire. L'entrée en vigueur de la directive PSD2 en 2018 oblige les établissements de crédit à ouvrir l'accès aux données de comptes de paiement à la demande expresse des clients, via des interfaces standardisées. Les tiers agréés peuvent ainsi agréger des historiques de paiement multi-bancaires, permettant une évaluation de la solvabilité bien au-delà des seules données internes disponibles. C'est le principe de l'Open Banking, dont l'effet sur l'analyse du risque crédit est tangible (Boot et al., 2021 ;

Vives, 2019), et qui redéfinit fondamentalement ce qu'un prêteur peut connaître d'un emprunteur et dans quel cadre.

1.3.2 Les données des réseaux sociaux et des empreintes numériques

Berg, Burg, Gombović et Puri (2020), travaillant à partir de données d'une plateforme allemande de e-commerce, montrent que des variables comportementales simples, type d'appareil utilisé, heure de connexion, navigateur, domaine de l'adresse email, rivalisent avec les scores produits par les bureaux de crédit traditionnels et améliorent la qualité prédictive lorsqu'elles y sont intégrées. Ces variables, en apparence anodines, révèlent des formes d'homogénéité socio-économique fortement corrélées à la probabilité de remboursement, que les données déclaratives ne capturent pas.

Ce phénomène dépasse le contexte européen. En Asie, Sesame Credit exploite les données transactionnelles, sociales et comportementales d'Alipay pour produire un score de crédit. Au Kenya, M-Shwari mobilise les flux de paiement mobile de M-Pesa pour évaluer des emprunteurs sans dossier bancaire. Óskarsdóttir et al. (2019) montrent que des modèles construits à partir des seules métadonnées d'appels peuvent parfois surpasser les approches traditionnelles en performance ajustée au risque.

En Europe, ces approches se heurtent au RGPD. Le principe de minimisation des données exige que les variables mobilisées soient pertinentes, limitées à ce qui est nécessaire et collectées pour la finalité déclarée. L'inférence de catégories sensibles, croyances, orientation sexuelle, origine ethnique, est par ailleurs interdite sauf exceptions strictement encadrées. Le cadre juridique européen ne bloque pas l'innovation, mais en fixe les limites de façon claire.

L'EBA reconnaît elle-même que les applications de big data et d'analytique avancée dans la banque en sont encore à un stade précoce (EBA, 2020b) : le chemin entre ce que les données permettent techniquement et ce que le droit autorise concrètement reste long.

1.3.3 Les données alternatives au sens large : vers le scoring inclusif

Les données de transaction et les empreintes numériques ne constituent pas les seules sources alternatives mobilisées dans la littérature. Un intérêt croissant se manifeste pour des données permettant d'évaluer la solvabilité de populations traditionnellement exclues des modèles faute d'historique de crédit : jeunes adultes, récents immigrés, travailleurs indépendants à revenus irréguliers.

Parmi les sources les mieux documentées, les paiements de loyers apparaissent comme un signal fiable de rigueur budgétaire, bien qu'absents des bases de données des bureaux d'information européens. Les factures d'énergie et de téléphonie sont également recensées dans la littérature, avec des résultats variables selon les contextes (Bazarbash, 2019 ; Berg et al., 2021).

D'autres sources soulèvent des questions plus difficiles. La géolocalisation présente un risque documenté de substitution indirecte à l'origine ethnique ou au niveau de revenu, pris au sérieux par les régulateurs. Dans les pays en développement, certaines fintechs exploitent les

comportements numériques ou les caractéristiques techniques des appareils faute d'alternatives. L'exploitation de profils LinkedIn pour retracer un parcours professionnel commence à être explorée de façon encore marginale. Ce que toutes ces approches ont en commun est de repousser la frontière de ce qui est considéré comme une donnée pertinente pour l'évaluation du risque, avec les enjeux de proportionnalité et de non-discrimination que cela implique au regard du cadre juridique européen.

1.3.4 Opportunités et premières tensions éthiques

L'intégration des données alternatives ouvre des opportunités considérables tout en soulevant des tensions éthiques fondamentales. Sur le plan des opportunités, ces données permettent d'évaluer des emprunteurs sans historique de crédit, élargissant l'accès au financement à des populations a priori exclues (Commission européenne, 2021 ; Bazarbash, 2019), de réduire l'asymétrie d'information entre prêteur et emprunteur (Berg et al., 2020), et d'affiner la personnalisation des offres au profil réel de l'emprunteur.

Ces opportunités se heurtent néanmoins à quatre tensions structurelles. La première est celle de la vie privée : l'usage intensif des traces numériques menace des libertés fondamentales protégées par l'article 7 de la Charte européenne et le RGPD, posant la question de ce qu'il est légitime de savoir sur un individu pour lui accorder du crédit. La deuxième est la confusion entre corrélation et causalité : des variables en apparence neutres, comme le code postal ou le type d'appareil utilisé, peuvent produire des effets proches du redlining en reproduisant des inégalités existantes sous couvert de neutralité statistique, sans que les indicateurs de performance standard ne les détectent (Barocas et Selbst, 2016). La troisième concerne le consentement éclairé : dans la grande majorité des cas, l'emprunteur ignore quelles données sont mobilisées et quel poids elles exercent dans la décision finale. La quatrième est la fracture numérique : les populations laissant peu de traces numériques, âgées, rurales ou à faibles revenus, risquent d'être systématiquement défavorisées par des modèles interprétant l'absence de données comme un signal négatif, transformant un outil d'inclusion en vecteur d'exclusion (O'Neil, 2016 ; Hurley et Adebayo, 2016).

Plusieurs enseignements se dégagent de cet état des lieux. Le recours aux données structurées dans l'évaluation de la solvabilité précède le machine learning : la régression logistique et les scorecards constituaient déjà le standard bancaire bien avant l'essor de l'apprentissage automatique, et ce socle fonde les exigences actuelles de transparence et de reproductibilité. Les gains prédictifs des algorithmes plus récents restent modestes sur données structurées et s'accompagnent d'une complexité accrue qui pèse sur la lisibilité des décisions et les processus de validation. Enfin, l'Open Banking et la montée des données alternatives redessinent en profondeur le périmètre informationnel du prêteur, sans que ce mouvement soit encore stabilisé.

Ce double mouvement, technique et informationnel, impose aux banques un arbitrage que l'optimisation seule ne peut résoudre, entre puissance prédictive, conformité réglementaire et légitimité éthique vis-à-vis des emprunteurs. La suite de ce mémoire examine successivement les risques structurels induits par cette mutation, le cadre réglementaire européen qui s'y applique, et les voies opérationnelles dont disposent les banques pour rendre ces tensions plus soutenables.

Chapitre 2 : Le déploiement de l'IA dans le secteur bancaire européen

La présence croissante de l'intelligence artificielle dans le secteur bancaire et financier ne génère pas des risques accidentels, elle génère des risques structurels. Les grandes institutions concentrent les compétences technologiques et infrastructures de données, creusant ainsi pour les acteurs plus petits encore plus un écart de performance. Au cœur de cette logique, on trouve l'automatisation des décisions de crédit, dont la tension entre efficacité prédictive et droits individuels est loin d'être révolue.

L'analyse de ce travail s'effectue principalement sur l'Union européenne et ponctuellement sur des pays références en IA, ici l'Asie et l'Amérique du Nord, non pas pour comparer des systèmes juridiques radicalement incomparables, mais parce que ces pays semblent illustrer des usages algorithmiques ou des intégrations de données alternatives qui n'ont pas encore été stabilisés en Europe, ce qui permet d'anticiper des évolutions sans en importer.

2.1 État des lieux de l'adoption de l'IA par les banques opérant dans l'Union européenne.

2.1.1 Cartographie des initiatives : une mosaïque en mutation rapide

En 2025, le déploiement de l'intelligence artificielle dans le secteur bancaire européen demeure profondément inégal, entre grandes institutions et banques régionales, mais aussi entre États membres. Trois facteurs structurels expliquent cette dispersion : des rythmes de transformation numérique différenciés, des systèmes informatiques legacy résistant à la modernisation, et des cultures managériales accordant un poids variable à l'innovation algorithmique.

L'Evident AI Index constitue l'un des indicateurs les plus mobilisés pour situer les établissements sur ce terrain, bien que sa structure, articulée autour des ressources humaines, de l'innovation et de la communication publique, mesure autant la visibilité des actions que leur profondeur réelle. En 2023, JPMorgan Chase occupe la première position pour la troisième fois consécutive, confirmant la domination des banques américaines. Parmi les établissements opérant dans l'UE, trois profils de maturité se distinguent.

BNP Paribas figure au troisième rang mondial en 2024 et au premier rang européen pour ses investissements dans les start-ups d'IA. La banque mobilise près de 3 000 experts en données, dont plus de 700 data scientists, et a officialisé un partenariat avec Mistral AI, complété par le déploiement d'une plateforme d'IA générative à l'ensemble des collaborateurs de sa banque d'investissement. ING Group déploie le machine learning depuis 2020 sur un périmètre élargi, détection de fraude, tarification, octroi de crédit et expérience client, selon un cadre interne structuré en 140 contrôles sur 20 phases, avec un accent sur le suivi continu en production. BBVA progresse de la 26e à la 13e place mondiale en 2025, s'appuyant sur des équipes Data et IA dédiées et une Data University revendiquant la formation de 50 000 employés, chiffre à interpréter avec prudence au regard des pratiques de communication institutionnelle.

Santander a expérimenté des modèles de machine learning en Amérique latine avant de les adapter au cadre réglementaire européen. Société Générale travaille à concilier explicabilité et

décision de crédit sous double contrainte réglementaire et de transparence. Deutsche Bank accuse un retard plus marqué, hérité de plusieurs vagues de restructuration et de contraintes de marges persistantes.

Les néobanques, Klarna, N26 et Revolut, ont construit leurs modèles sans systèmes legacy, ce qui leur confère une réactivité structurellement supérieure. Leur positionnement mérite cependant d'être nuancé : N26 opère sous licence bancaire européenne, Klarna reste principalement un acteur du paiement fractionné, et Revolut poursuit l'obtention progressive de licences bancaires selon les marchés. Ces acteurs présentent un profil de risque de crédit sensiblement moindre que celui des banques universelles et font l'objet d'une surveillance prudentielle moins contraignante. Ils n'en exercent pas moins une pression concurrentielle réelle sur les établissements traditionnels, contraints de s'adapter sur le terrain de l'ouverture de compte, de la personnalisation et de l'expérience utilisateur.

Tableau 1 - Niveaux de maturité IA de certains acteurs bancaires opérant dans l'UE et fintechs comparatives

Institution	**Périmètre**	**Maturité IA**	**Usages principaux**	**Enjeux principaux**
BNP Paribas	UE	Élevée	IA générative, data science, conformité	Industrialisation à grande échelle
ING	UE	Élevée	Crédit, fraude, pricing, durabilité	Gouvernance et validation des modèles
BBVA	UE	Élevée	Plateforme data/IA, formation interne	Centralisation des données
Deutsche Bank	UE	Intermédiaire	Risque de crédit, conformité	Legacy IT, restructuration
Société Générale	UE	Intermédiaire	Scoring, risque de contrepartie	IA explicable, transparence
Santander	UE	Intermédiaire / élevée	Crédit, marchés internationaux	Adaptation réglementaire européenne
N26 / Revolut / Klarna	UE / variable	Variable	Scoring rapide, expérience client	Statut réglementaire, robustesse des modèles

Source : élaboration de l'auteur à partir des rapports annuels, communications institutionnelles et classements sectoriels cités.

2.1.2 De l'expérimentation à la production : des niveaux de maturité différenciés

À partir de 2023, certaines banques franchissent le seuil de l'expérimentation pour déployer l'IA à plus grande échelle, avec des effets concrets sur leur activité. Mais les moyens disponibles restent très inégaux. Les grandes banques internationales progressent sensiblement plus vite que les établissements régionaux ou coopératifs, qui manquent souvent des infrastructures et des compétences nécessaires. L'écart tend à se creuser, au point que certains observateurs évoquent un "AI readiness gap" susceptible de redistribuer les cartes du secteur.

L'étude Deloitte EMEA de début 2025 projette un bond de l'adoption de l'IA dans les petites banques, de 22% à 52% sur deux ans, porté notamment par l'accessibilité croissante des outils SaaS. Les grandes banques progressent plus lentement, leur maturité technologique existante limitant les marges de progression rapide. L'Evident AI Index confirme des performances encore très disparates parmi les grands établissements européens. Si l'écart entre banques qui avancent et banques qui stagnent continue de se creuser, une recomposition partielle du secteur n'est pas à exclure.

2.1.3 Facteurs d'adoption et freins structurels

Les banques traditionnelles subissent des pressions simultanées : concurrence des fintechs, maîtrise des coûts opérationnels, attentes clients en temps réel et ouverture des données via l'Open Banking et la PSD2, autant de dynamiques qui redéfinissent le périmètre de l'innovation financière.

Depuis 2021, plusieurs freins structurels entravent l'appropriation de l'IA dans le secteur (EBA, 2021). Sur le plan technique, la prévalence d'architectures legacy (COBOL, Fortran) rend la migration vers des environnements de machine learning longue, coûteuse et risquée. Sur le plan humain, la mutation des métiers génère des résistances, aggravées par la rareté des compétences en data science et la concurrence directe des fintechs et plateformes numériques pour les mêmes profils. Sur le plan réglementaire, la superposition des cadres (RGPD, CRR/CRD, normes EBA et BCE, AI Act) alourdit le coût de conformité au point de le rendre difficile à absorber pour les établissements de taille modeste ; moins d'un quart des institutions interrogées par Deloitte se déclaraient prêtes à répondre aux exigences de l'AI Act.

Un quatrième risque, moins débattu, est la dépendance croissante aux prestataires externes, qui s'articule à trois niveaux : logiciel et infrastructure (fournisseurs cloud, plateformes de déploiement), algorithmique (API d'IA générative, modèles pré-entraînés) et informationnel (agrégateurs de données). Cette imbrication complexifie la traçabilité des décisions, la gestion des risques et la répartition des responsabilités en cas de biais, tout en concentrant une part croissante de l'infrastructure financière entre les mains de quelques grands acteurs technologiques. La souveraineté numérique devient ainsi un enjeu concret ; DORA et la directive NIS2 y répondent en renforçant les exigences de résilience opérationnelle et de maîtrise des tiers.

Lorsque plusieurs établissements partagent les mêmes architectures et bases de données, leurs modèles peuvent remonter simultanément les mêmes signaux face à un choc identique. En période de crise, cette synchronisation algorithmique risque d'amplifier le resserrement du crédit plutôt que de l'atténuer, un phénomène de procyclicité dont la BRI (2022) confirme la pertinence et qui sera examiné ultérieurement en lien avec la question de la dérive des concepts.

Les entretiens professionnels font ressortir une hiérarchie claire des difficultés : au premier plan, la fiabilité des données d'entrée, l'insuffisance de la documentation et l'absence de validations tierces ; en second lieu, le clivage entre équipes métier et experts en données ; enfin, l'incertitude réglementaire entourant l'AI Act, qui rend difficile l'anticipation des exigences à venir.

2.2 Architectures techniques : modèles en production, pipelines de décision, systèmes hybrides

2.2.1 Le pipeline typique d'une décision de crédit assistée par l'IA

Le traitement d'une demande de crédit s'articule en six phases successives, où chaque défaillance amont se répercute en aval : la fiabilité de la décision finale dépend de la qualité de chaque maillon.

Collecte des données. Le profil du demandeur est constitué à partir de données déclaratives, de l'historique interne de la banque, des informations issues des bureaux de crédit, et, depuis la PSD2, des données externes accessibles via les API d'Open Banking (Thomas et al., 2002 ; Siddiqi, 2006).

Prétraitement et feature engineering. Les données brutes font l'objet d'un nettoyage (correction des anomalies, imputation des valeurs manquantes, encodage des variables catégorielles), suivi d'une phase de construction des variables. Bien qu'invisible dans les résultats finaux, cette étape conditionne la performance prédictive du modèle à chaque inférence (Baesens et al., 2016 ; Siddiqi, 2017).

Scoring. Le système de scoring, fondé sur des architectures telles que XGBoost, LightGBM ou les forêts aléatoires, produit trois sorties : la probabilité de défaut, la recommandation d'octroi, et les variables les plus contributives, ces dernières répondant aux exigences de justification imposées par le cadre réglementaire européen. Le tout est restitué en quelques secondes.

Application des règles métier. Le score est ensuite soumis à un ensemble de règles qui peuvent court-circuiter la recommandation du modèle : exclusions réglementaires (fichage FICP, procédure de surendettement), plafonds d'exposition par contrepartie, limites de concentration sectorielle ou géographique, et politiques internes de la banque. Le modèle n'est ainsi qu'un composant parmi d'autres du dispositif décisionnel.

Décision. Trois issues sont possibles : octroi automatique pour les profils conformes à score suffisant, rejet immédiat en cas de score insuffisant ou de blocage réglementaire, et revue manuelle pour les dossiers à signaux contradictoires ou en situation limite. C'est cette troisième voie qui concentre l'essentiel du jugement humain dans le processus.

Monitoring. Après déploiement, les performances du modèle font l'objet d'un monitoring actif : l'AUC-ROC vérifie le pouvoir discriminant, le taux de défaut prédit est confronté au taux observé, et des outils de surveillance détectent deux types de dérive, le data drift (évolution de la distribution des variables d'entrée) et le concept drift (rupture de la relation entre variables et cible). Le franchissement d'un seuil déclenche une alerte et peut engager une révision du modèle (Žliobaitė et al., 2016 ; EBA, 2023b).

Tableau 2 - Risques associés aux étapes du pipeline décisionnel

Collecte	Données excessives, consentement ambigu, qualité variable
Prétraitement	Introduction de biais par imputation ou feature engineering
Scoring	Opacité, instabilité, discrimination par proxy
Règles métier	Effets de seuil, rigidité, exclusion automatique
Décision	Automatisation excessive, validation formelle : rubber stamping
Monitoring	Détection tardive du drift, recalibrage procyclique

Source : élaboration de l'auteur.

2.2.2 Les systèmes hybrides homme-IA : réponse architecturale au défi réglementaire

L'article 22 du RGPD (2018) encadre strictement les décisions entièrement automatisées affectant les droits d'une personne. En matière de crédit à la consommation, un refus ou une offre fortement défavorable générés sans intervention humaine relèvent de ce régime. Trois bases légales permettent de le justifier : nécessité contractuelle, obligation légale ou consentement explicite. Quelle que soit la base retenue, le droit à contestation, à révision humaine et à être entendu demeure obligatoire. L'invocation de la nécessité contractuelle pour un processus entièrement automatisé reste toutefois contestable : l'évaluation de la solvabilité ne requiert pas structurellement l'absence de regard humain (Wachter et al., 2017 ; Kaminski, 2019).

Depuis 2024, le règlement (UE) 2024/1689 classe les systèmes de scoring crédit et d'évaluation de solvabilité parmi les systèmes à haut risque (annexe III, point 5b). Les obligations afférentes incluent une documentation technique renforcée, une surveillance continue, un contrôle de la qualité des données d'entraînement, la traçabilité des décisions et une supervision humaine effective. Seuls les systèmes dédiés exclusivement à la détection de fraudes financières sont exemptés. Ce cadre est substantiellement plus contraignant qu'une simple conformité RGPD, et les établissements n'ayant pas anticipé ces exigences lors du développement de leurs modèles devront probablement revoir leur architecture.

La littérature distingue trois configurations d'hybridation selon leur degré d'automatisation.

Dans la configuration **human-in-the-loop**, la décision humaine est guidée par la recommandation algorithmique. En pratique, un phénomène de rubber stamping tend à se substituer à l'examen critique : volume de dossiers excessif, formulation autoritaire des recommandations et absence de dispositifs permettant de les contester conduisent à une supervision formellement présente mais fonctionnellement neutralisée. C'est précisément ce glissement que le cadre réglementaire entend prévenir en exigeant que l'intervention humaine soit effective et non simplement tracée (EBA, 2021 ; Wachter et al., 2017).

Dans la configuration **human-on-the-loop**, le traitement est automatisé et un superviseur peut interrompre le processus en cas de problème détecté. Ce dispositif ne satisfait toutefois pas nécessairement aux exigences de l'article 22 du RGPD : si la décision reste automatique dans les faits, la présence d'un superviseur distant ou intervenant a posteriori ne modifie pas sa nature. Ce qui importe est la capacité réelle à modifier la décision avant qu'elle produise ses effets.

La configuration **fully automated** est admissible dans le cadre strict de l'article 22, paragraphe 2, pour des montants limités, des profils à risque maîtrisé et un cadre contractuel ou légal explicite. Elle doit s'accompagner de trois garanties non optionnelles : information du client sur le caractère exclusivement automatisé de la décision, existence d'un recours traçable, et accessibilité d'un examen humain à la demande. Une automatisation sans voie de recours effective ne satisfait en aucun cas aux exigences du règlement, indépendamment de la robustesse du modèle.

2.2.3 La stratégie des champion/challenger et le monitoring des modèles

Dans les systèmes de gouvernance des modèles, le modèle en production (champion) est en permanence comparé à des modèles challengers opérant en parallèle sans produire de décisions réelles, ce qui permet de les tester en conditions réelles sans prise de risque. Le remplacement du champion n'intervient qu'après des performances durables du challenger et une validation par plusieurs instances : équipes d'audit et de contrôle interne (robustesse, stabilité), équipes de conformité (implications réglementaires), et le cas échéant l'autorité de supervision pour les modèles à portée systémique. Le changement de modèle de crédit est une décision qui engage bien au-delà de la dimension technique (EBA, 2021 ; Siddiqi, 2017).

La pandémie de COVID-19 a constitué un cas d'école de concept drift : la rupture brutale des comportements de remboursement, conjuguée aux moratoires, a rendu les données historiques inopérantes pour décrire la réalité du moment, sans que les modèles le détectent automatiquement. Les équipes risque ont dû recalibrer en urgence, souvent avec un historique récent insuffisant pour le faire de manière fiable. Certains établissements ont temporairement suspendu leurs modèles automatiques au profit du jugement expert (Žliobaitė et al., 2016 ; EBA, 2023b).

Lorsque plusieurs établissements s'appuient sur des architectures voisines, leurs modèles tendent à réagir de manière similaire aux mêmes signaux. Des recalibrages convergents et non concertés peuvent ainsi conduire à un resserrement simultané des politiques d'octroi, amplifiant les tensions existantes : c'est le risque de convergence algorithmique et d'effet procyclique évoqué en 2.1.3. Les observations directes demeurent rares, les données nécessaires n'étant pas publiques et les établissements ne coordonnant pas explicitement leurs choix. Toutefois, l'uniformisation croissante des outils, des fournisseurs de données et des méthodologies rend ce scénario de plus

en plus vraisemblable. La contagion financière, longtemps pensée comme organisationnelle, revêt désormais également une dimension algorithmique (BRI, 2022).

2.3 Des promesses ambivalentes : performance, personnalisation et nouvelles vulnérabilités

2.3.1 Une meilleure segmentation du risque et la réduction des défauts de paiement

Les modèles de machine learning surpassent la régression logistique sur données conventionnelles, avec un gain moyen d'AUC-ROC de 2 à 5 points, pouvant atteindre 7 points sur des jeux de données enrichies (Lessmann et al., 2015 ; Barboza et al., 2017 ; Berg et al., 2020). Dans le cadre d'un portefeuille large, même un gain marginal de précision prédictive se traduit par une réduction mesurable des pertes attendues. Ces gains restent toutefois sensibles aux conditions de marché, de risque et de composition du portefeuille : ce qui performe sur un échantillon historique donné ne se généralise pas nécessairement en production, notamment en raison de la vitesse de changement des environnements.

Depuis les années 2010, plusieurs études montrent que les modèles de machine learning captent des signaux que la régression logistique ne détecte pas : flux de dépenses atypiques, irrégularités dans les paiements réguliers, légères inflexions comportementales pouvant précéder la défaillance de plusieurs mois. Ces signaux faibles ne se substituent pas nécessairement au jugement de l'analyste, mais le complètent. Des établissements enregistrent des gains de productivité, bien que ces résultats demeurent très contextuels, dépendants à la fois de la nature des données disponibles et des modalités d'intégration des alertes dans le processus décisionnel. Toute extrapolation brute de ces résultats est risquée. Ce qui s'avère le plus prédictif est parfois précisément ce qui échappe à la lecture directe.

2.3.2 La personnalisation du risque et le risk-based pricing

Une évaluation du risque plus précise rend possible une tarification plus fine. Les modèles de machine learning détectent des relations entre variables que les approches classiques ignorent ou lissent : comportements atypiques, corrélations inattendues, dynamiques non linéaires. Cette capacité permet d'affiner la segmentation des profils et d'ajuster les conditions proposées au plus près du risque réellement porté. Nous sommes donc face au principe du risk-based pricing, dont la granularité est décuplée par les algorithmes d’apprentissage (Thomas et al., 2002).

Le risque pourrait mieux être segmenté, ce qui améliorerait les conditions d’accès à l’emprunt de certains emprunteurs considérés comme dignes de confiance. Mais c’est ce que montre Fuster et al. 2022, cette précision accroît, en même temps, le fossé entre les groupes démographiques. En tant que variable prédictive, le quartier de résidence est susceptible de faire varier le taux proposé de manière très significative. Ce critère n’est pas pertinent pour mesurer la solvabilité du client ; il encode des ségrégations historiques, des dynamiques du marché immobilier, des inégalités d’accès, aux services. Deux emprunteurs au profil financier très proche peuvent se voir proposer des conditions d’emprunt très différentes selon leur adresse. Le modèle ne choisit pas de discriminer. Il amplifie ce que contiennent les données, et les données contiennent l’Histoire (Barocas et Selbst, 2016 ; Prince et Schwarcz, 2020).

2.3.3 L'inclusion financière : promesse et ambivalences

Les systèmes de scoring traditionnels excluent mécaniquement les profils sans historique bancaire : jeunes adultes, travailleurs aux revenus irréguliers, nouveaux arrivants. Des travaux engagés à partir de 2019 explorent l'usage de données alternatives (relevés de consommation, historiques de recharge mobile, comportements de paiement informel) pour inférer un risque de défaut là où les fichiers classiques sont vides. Des initiatives comme Tala au Kenya, M-Shwari ou Ant Group en Chine ont démontré la viabilité de ces approches à grande échelle. Leur transposition en Europe soulève toutefois des obstacles substantiels : le RGPD encadre strictement les données exploitables, les autorités prudentielles imposent une traçabilité que ces systèmes peinent à produire, et les populations cibles diffèrent sensiblement. Ce qui fonctionne dans un contexte de sous-bancarisation massive ne se transpose pas directement dans un marché déjà régulé.

Le cadre juridique européen est en effet peu compatible avec les fondements de ces modèles. Le RGPD impose la minimisation des données et interdit le traitement des catégories sensibles ; le consentement doit être explicite et renouvelable, ce qui contredit les modèles reposant sur des flux comportementaux continus. Les directives anti-discrimination (2000/43/CE et 2004/113/CE) prohibent les effets disparates, même indirects, fondés sur l'origine ou le sexe. L'EBA (2020) soulignait à cet égard que l'objectif d'élargissement de l'accès au crédit via les données alternatives ne saurait justifier a priori le choix de chaque variable ni ses effets sur les populations concernées.

Ces modèles prédictifs peuvent effectivement ouvrir l'accès au crédit à des profils qu'un scoring classique aurait écartés. Mais lorsqu'ils sont calibrés pour maximiser la rentabilité, leur logique change : ils identifient avec précision les emprunteurs les plus fragiles et les orientent vers les produits les plus coûteux. Plus le profil est vulnérable, plus le taux est élevé. La logique actuarielle n'est pas irrationnelle ; elle est simplement indifférente aux effets redistributifs qu'elle produit. Inclusion et extraction peuvent ainsi coexister dans le même système, selon la manière dont l'objectif du modèle est défini.

2.3.4 Rapidité de décision et réduction des coûts opérationnels

L'approbation d'un prêt, qui nécessitait autrefois plusieurs jours, s'effectue désormais en quelques minutes, standard imposé par les fintechs auquel l'ensemble du secteur bancaire s'est aligné. Parallèlement, le rôle des équipes humaines se redéfinit : à mesure que les dossiers courants sont absorbés par l'automatisation, les analystes se concentrent sur les cas résistant au traitement algorithmique, ceux qui requièrent jugement, contextualisation et interaction.

Cette évolution soulève une question sur la qualité de l'analyse des risques. L'expertise se maintient par la pratique : lorsque les cas complexes se raréfient dans le quotidien, la capacité à les détecter s'érode progressivement, sans rupture visible. C'est précisément ce glissement que le RGPD et l'AI Act entendent prévenir en exigeant une supervision humaine effective et non une simple validation formelle en bout de chaîne. En l'absence de cadre strict, ce phénomène reste difficilement perceptible jusqu'à ce qu'un dossier mal évalué révèle l'exposition à un risque de non-conformité.

Le machine learning améliore la précision du risque, accélère les traitements et réduit les coûts : ces apports ne sont pas contestés. Ce qui l'est, c'est ce qui les accompagne. La segmentation fine

véhicule des biais invisibles aux indicateurs standards. Le risk-based pricing est justifié jusqu'au moment où l'on examine qui supporte les hausses les plus sévères. L'élargissement de l'accès au crédit peut coexister avec des logiques d'extraction que personne ne défendrait explicitement. Les décisions se succèdent à un rythme qui ne laisse pas de place au regard humain entre elles. Ce n'est pas une dérive accidentelle : c'est le résultat structurel de la cohabitation entre données chargées d'histoire, exigences d'efficacité et attentes éthiques, sans arbitrage clair sur les priorités.

C'est sur ce terrain que les tensions propres à l'usage des algorithmes dans l'octroi du crédit se manifestent : risques de discrimination indirecte, opacité des modèles, fragilisation des droits des emprunteurs. L'Europe tente d'y répondre par un cadre réglementaire encore en construction. Performance, légalité et éthique constituent trois exigences qui ne convergent pas spontanément. Les sections suivantes analysent chacun de ces nœuds.

PARTIE II : Les risques structurels du scoring algorithmique

L'IA s'installe dans le scoring crédit, et deux problèmes concrets vont avec. Le premier : les algorithmes reproduisent ce que les données portent, et les données portent souvent des inégalités que personne n'a décidé d'y mettre. Le second : quand un modèle refuse un crédit, expliquer pourquoi devient difficile pour la banque, compliqué pour le régulateur, et à peu près impossible pour le client. Ce qui rend la situation inconfortable, c'est que ces deux problèmes se nourrissent l'un l'autre. Un modèle plus précis est souvent moins lisible. Un modèle plus transparent perd parfois en performance. Et la conformité réglementaire ne dit rien sur l'équité réelle des décisions.

Chapitre 3 : Les biais algorithmiques dans le credit scoring : mécanismes, effets et enjeux stratégiques

L'un des problèmes les plus concrets que pose l'IA dans le crédit, c'est la persistance des biais. Pas des anomalies isolées. Des erreurs qui se répètent, sur les mêmes profils, dans le même sens, sans que le modèle le signale. En matière de scoring, ça pèse : un algorithme ne se contente pas de prévoir un comportement, il décide qui obtient un prêt. Et un prêt, c'est souvent ce qui permet d'acheter un logement, de financer une formation, de démarrer une activité. Quand l'erreur est systématique, ses effets s'accumulent. Sur des individus, sur des ménages, parfois sur plusieurs générations. À ce stade, le problème ne se règle pas en ajustant un hyperparamètre.

3.1 Origines des biais : une taxonomie des mécanismes discriminatoires

3.1.1 Les biais dans les données d'entraînement

Un algorithme de crédit apprend sur des décisions passées. C'est son principe de base. Le problème, c'est que ces décisions n'étaient pas neutres. Si les refus des années 1980 ou 1990 reflétaient des critères discriminatoires, le modèle les absorbe et les reconduit, mécaniquement, sans que personne ne l'ait voulu. La machine ne discrimine pas par intention. Elle généralise ce qu'on lui a fourni, avec une régularité qu'aucun agent humain n'aurait pu atteindre. Un analyste biaisé faisait des erreurs de temps en temps. Un modèle biaisé les applique à des dizaines de milliers de dossiers, de façon identique, sans que rien dans les métriques ne le signale. La

discrimination ne disparaît pas. Elle devient simplement plus difficile à voir, et plus difficile à contester (Barocas et Selbst, 2016 ; O'Neil, 2016).

Historiquement, certains dossiers de crédit ont bien trop souvent été préjudiciables parce qu'ils concernent des populations dont les caractéristiques ne correspondent pas à celles de l'immense majorité, à savoir les femmes, les personnes issues des minorités ethniques et les habitants des territoires les plus défavorisés. Non pas parce que leur situation était intrinsèquement plus risquée, mais parce que les conditions mises en œuvre les rendaient plus dangereuses. Les difficultés ainsi accumulées produisaient bien les défauts qu'elles étaient censées anticiper. Ainsi, lorsque l'algorithme fait son apprentissage sur ces données, il ne comprend pas l'histoire qui se cache derrière les chiffres. Il constate que ces profils ont moins bien remboursé, et il reproduit les mêmes restrictions. La statistique finit par valider ce qui n'était au départ qu'un traitement inégal. Ce que le modèle apprend, c'est un résultat, pas une réalité. (Barocas et Selbst, 2016 ; Mehrabi et al., 2021).

3.1.2 Le biais de sélection

Les données de remboursement ont un angle mort évident : elles ne contiennent que les personnes à qui on a accordé un crédit. Les refusés n'y figurent pas. Le modèle apprend donc sur une population triée en amont, sans jamais observer ce qu'auraient fait ceux qu'on a écartés. C'est le biais de sélection, une forme de *survivorship bias* propre au crédit (Thomas et al., 2002). La reject inference tente de combler ce vide en reconstituant les comportements hypothétiques des refusés. Le problème, c'est que ces reconstitutions reposent sur des hypothèses qu'on ne peut pas vraiment tester. C'est le cas de la reject inference, dont certains travaux indiquent même qu'elle peut aggraver la distorsion au lieu de la réduire. Sur ce point, la littérature ne s'est pas accordée, et les résultats restent suffisamment hétérogènes pour que la méthode soit encore discutée (Lessmann et al., 2015).

3.1.3 Le biais historique et le biais de mesure

L'écart de revenus entre femmes et hommes dans l'UE s'établissait à 11,1 % en 2024 (Eurostat), reflet de biais structurels : inégalités d'accès à l'emploi, interruptions de carrière, sur-représentation dans certains secteurs. Lorsqu'un algorithme de crédit est entraîné sur des historiques salariaux, ces inégalités s'y trouvent encodées. Un modèle peut ainsi scorer différemment deux profils financièrement comparables selon leur niveau de revenu, et donc indirectement selon leur genre, sans qu'aucune variable ne discrimine explicitement sur ce critère. Comme le montrent Suresh et Guttag (2021), la performance algorithmique ne constitue pas une garantie d'équité : ce que le modèle apprend, ce sont des réalités contemporaines ancrées dans une histoire non neutre (Kozodoi et al., 2022).

Les indicateurs de solvabilité produisent par ailleurs des résultats structurellement défavorables pour certains profils. Pour un jeune adulte ou un emprunteur nouvellement bancaire, l'historique est trop court pour être interprété de manière fiable. Un score classique fondé sur l'ancienneté pénalise mécaniquement ces profils, indépendamment de leurs capacités réelles de remboursement. Le problème ne tient pas au profil lui-même, mais à ce que l'indicateur choisit de mesurer et à ce qu'il laisse de côté. Sur ces populations, l'écart entre score affiché et risque réel peut être substantiel (Mehrabi et al., 2021).

3.1.4 La discrimination par proxy

Retirer une variable sensible d'un modèle ne suffit pas à éliminer le biais qu'elle transporte. D'autres variables, en apparence anodines, peuvent la reconstituer. Le code postal en est l'exemple le plus documenté : il ne dit rien de l'origine ethnique ni de la situation sociale, mais il recoupe bien souvent la composition d'un quartier avec un degré de précision suffisamment bon pour réactiver des logiques proches du redlining. De manière analogue, le type de contrat fonctionne : un statut intérimaire ou précaire ne revendique explicitement aucun groupe, mais il punit beaucoup plus durement certains. Qui s'appelle redundant encoding : sans avoir accès à une information protégée, le modèle la recombine à partir de ce qui l'entoure. Un assemblage de détails apparemment neutres suffit. La suppression d'une variable n'efface pas ce qu'elle encode (Barocas et Selbst, 2016 ; Chouldechova et Roth, 2020).

3.1.5 Tableau de synthèse

Tableau 3 - Taxonomie des biais dans le credit scoring algorithmique

Type de biais	Mécanisme	Exemple	Risque associé
Biais historique	Données reflétant des inégalités passées	Écarts de revenus structurels	Différences d'accès ou de tarification
Biais de sélection	Refusés absents des données	Comportement non observable	Sous-estimation des profils atypiques
Biais de mesure	Variable moins fiable pour certains groupes	Score faible pour nouveaux arrivants	Mauvaise estimation de solvabilité
Discrimination par proxy	Variable neutre corrélée à un groupe protégé	Code postal, type de contrat	Discrimination indirecte
Biais de représentation	Sous-représentation dans les données	Peu de données sur indépendants	Performance dégradée sur sous-groupes

Source : élaboration de l'auteur à partir de Suresh et Guttag (2021), Barocas et Selbst (2016), Mehrabi et al. (2021).

3.2 Effets discriminatoires : cas documentés et jurisprudence émergente

3.2.1 Le cas Apple Card / Goldman Sachs (2019)

Au mois de novembre 2019, David Heinemeier Hansson fait une série de tweets qui se répandent rapidement dans le monde tech. Sa conjointe et lui ont des situations financières équivalentes. Mais la limite de crédit qui lui a été assignée sur l'Apple Card est plusieurs fois supérieure à la sienne. Ce constat est fait par Steve Wozniak dans son propre foyer. Les autorités chargées du contrôle des pratiques bancaires de l'État de New York (New York Department of Financial

Services) ouvrent une enquête sur Goldman Sachs. La banque affirme que l'algorithme sur lequel repose le traitement des crédits ne prend pas le genre comme variable. C'est peut-être exact. Ça ne suffit pas à expliquer les écarts. D'autres variables, en apparence neutres, peuvent reconstituer ce qu'on a voulu exclure, par corrélations successives, sans que personne ne l'ait programmé délibérément. L'enquête conclue en 2021 n'a pas établi d'intention discriminatoire. Mais l'intention n'est pas le seul problème. Ce que l'affaire met en évidence, c'est qu'à aucun moment Goldman Sachs n'a été en mesure d'expliquer ces écarts de façon satisfaisante, ni aux clients concernés, ni aux régulateurs.

3.2.2 L'étude Bartlett et al. (2022)

Une étude publiée en 2022 dans le Journal of Financial Economics examine les crédits immobiliers américains et relève des écarts de prix entre emprunteurs aux profils comparables, selon leur appartenance à différents groupes. Bartlett, Morse, Stanton et Wallace constatent que ces écarts se retrouvent aussi bien dans les fintechs que dans les établissements traditionnels. Ce résultat fragilise l'idée que l'automatisation suffirait à neutraliser les biais. Les algorithmes peuvent les intégrer de façon plus systématique qu'un jugement humain, et surtout sans que les métriques habituelles ne le détectent.

3.2.3 Le cas SyRI et l'arrêt SCHUFA

En 2020, le tribunal de La Haye invalide le système SyRI, outil automatisé de détection des fraudes aux prestations sociales. Le raisonnement du jugement dépasse le cas d'espèce : un algorithme opaque appliqué à des décisions touchant aux droits fondamentaux ne satisfait pas aux exigences du droit européen. La transparence y est posée non comme critère souhaitable, mais comme condition de légitimité. Significativement, le tribunal n'exige pas la preuve d'une discrimination effective : l'opacité du système, jointe à un risque plausible de traitement injuste, suffit. Ce renversement de la charge de la preuve produit des effets immédiats sur tout dispositif algorithmique déployé dans des situations analogues.

Le 7 décembre 2023, la CJUE rend son arrêt dans l'affaire Schufa (C-634/21). La Cour juge que le score produit par une agence d'information commerciale relève de l'article 22 §1 du RGPD dès lors qu'il est déterminant pour la décision d'octroi de crédit. Les banques ne peuvent dès lors plus traiter le scoring externe comme une donnée tierce dont elles ne seraient pas responsables. Par ailleurs, la directive (UE) 2020/1828 sur les actions de groupe élargit substantiellement l'exposition des établissements : un modèle contestable n'expose plus seulement à des recours individuels, mais à des actions collectives portées par des populations entières de demandeurs, ce qui constitue un point d'achoppement susceptible de fragiliser l'ensemble du modèle économique de l'établissement dans son environnement concurrentiel.

3.2.4 Discriminations structurelles persistantes en Europe

Jusqu'en 1965, en France, une femme ne pouvait pas ouvrir un compte bancaire sans l'autorisation de son mari. Les modèles entraînés sur des données historiques longues ne neutralisent pas ce genre de logique. Ils l'absorbent, sans variable explicite, par les corrélations que ces décennies ont laissées dans les données. Le code postal fonctionne de façon analogue dans les grandes villes européennes : il ne désigne aucun groupe directement, mais reflète souvent la composition sociale ou ethnique d'un quartier avec assez de précision pour produire

des effets proches du redlining. Des quartiers entiers se retrouvent pénalisés sans que le modèle les vise. Ce que la technologie fait ici, ce n'est pas créer de nouvelles discriminations. C'est reprendre les anciennes dans un format qui les rend moins visibles et plus difficiles à contester (Fuster et al., 2022).

3.3 Mesures d'équité algorithmique : apports et limites

3.3.1 Les métriques de fairness

Définir l'équité algorithmique est plus complexe qu'il n'y paraît : plusieurs critères coexistent, chacun cohérent en lui-même, mais mutuellement incompatibles.

La **parité démographique** exige des taux d'acceptation identiques entre groupes. Facile à mesurer, elle peut néanmoins contraindre le modèle à ignorer de réelles divergences de solvabilité, au risque d'exposer certains emprunteurs au surendettement.

L'égalité des chances (equal opportunity) impose un taux de vrais positifs identique entre groupes : à capacité de remboursement égale, la probabilité d'obtenir le crédit doit être la même.

L'égalité des odds (equalized odds) ajoute une contrainte symétrique : le taux de faux positifs doit également être équivalent entre groupes. Hardt, Price et Srebro (2016) ont démontré que ces deux derniers critères ne peuvent être satisfaits simultanément lorsque les taux de base diffèrent entre populations.

La **fairness contrefactuelle** pose une question distincte : si le demandeur appartenait à un autre groupe, la décision serait-elle identique ?

La **calibration** exige qu'à score égal, la probabilité de défaut soit la même pour tous les profils.

Chouldechova (2017) démontre formellement que ces critères sont incompatibles dès lors que les taux de défaut diffèrent entre groupes. Il faut donc choisir, et ce choix est autant politique que technique.

3.3.2 Le théorème d'impossibilité et les limites des métriques

Kleinberg et al. (2016), puis Chouldechova (2017), ont établi une limite formelle : sauf si les taux de défaut sont identiques entre groupes, aucun système prédictif ne peut satisfaire simultanément toutes les définitions usuelles de l'équité algorithmique. Ce n'est pas un problème de modèle mal calibré ou de données insuffisantes. Ce que nous avons là, c’est une impossibilité mathématique. Choisir un critère plutôt qu’un autre d’équité, c’est donc trancher sur quelle valeur : protège-t-on plus celui qui rembourse mais qui n’a pas pu obtenir de crédit, ou celui qui aura obtenu un crédit mais ne parviendra pas à rembourser ? Cette nécessité de choix ne peut pas se lire dans les données. Elle renvoie à des choix politiques et sociaux à faire explicitement, et non plus à dissimuler derrière la neutralité d’un algorithme qui aurait tout réglé.

Les indicateurs d'équité se heurtent à plusieurs difficultés concrètes. Ils tendent à ignorer les discriminations croisées, celles qui touchent des personnes appartenant simultanément à plusieurs groupes défavorisés (Chouldechova et Roth, 2020). Intégrer des contraintes d'équité dans un

modèle peut aussi dégrader ses performances économiques de façon mesurable (Kozodoi et al., 2022). La difficulté la plus structurelle est d'ordre juridique. Détecter un biais envers un groupe suppose de savoir à quel groupe appartient chaque demandeur. C'est exactement ce que l'article 9 du RGPD interdit en Europe, la collecte de données sur l'origine ethnique ou d'autres caractéristiques sensibles. Le résultat est un angle mort réglementaire : les outils pour mesurer le biais supposent des données que le droit européen ne permet pas de collecter. Ce problème, connu sous le nom de fairness without demographics, structure le débat depuis les travaux d'Hacker (2018).

3.4 Conséquences économiques et stratégiques pour les banques

3.4.1 Risques juridiques et réputationnels

L'affaire Apple Card l'a montré en 2019 : une polémique sur un algorithme peut, en quelques heures, forcer une enquête officielle et installer durablement un doute sur les pratiques d'un établissement. L'opacité des modèles aggrave la situation : quand une banque ne peut pas expliquer pourquoi un crédit a été refusé, elle n'est pas en mesure de se défendre non plus. Sur le plan réglementaire, les risques sont chiffrés. Le RGPD prévoit des amendes jusqu'à 20 millions d'euros ou 4 % du chiffre d'affaires mondial. L'arrêt Schufa de 2023 ajoute une obligation de gouvernance sur les algorithmes décisionnels qui pèse directement sur les établissements bancaires. Intervenir après une mise en cause revient systématiquement plus cher que concevoir le système correctement en amont.

3.4.2 Risques de gouvernance et de supervision

Dès lors que l'automatisation entre en jeu dans l'octroi du crédit, les contraintes s'amoncellent vite. En ces temps tendus, la possibilité d'un modèle interne de scoring commercial, prévue par l'approche IRB, pèse sur tout établissement. Réglementation sur les données, encadrement à venir de l'intelligence artificielle, exigences du droit des consommateurs : les dispositifs se télescopent. On observe la stabilité ou la lisibilité d'un tel modèle avec minutie. On découvre, dans le domaine prudentiel, que toute vulnérabilité perçue appelle réaction des autorités. Ça se traduit crûment. Des ajustements de capital peuvent être nécessaires. L'accès aux marges s'étiole. Parfois, le retour au standard est nécessaire. L'EBA l'a précisé dans ses orientations de 2021 : les exigences de robustesse et d'explicabilité sur les modèles IRB sont vérifiées, non seulement posées. Les établissements qui traitent ces sujets trop tardivement le paient doublement, en coûts de remédiation d'abord et en crédibilité ensuite auprès des superviseurs.

3.4.3 Algorithm aversion et asymétrie concurrentielle

Dietvorst, Simmons et Massey (2015) ont documenté le phénomène d'aversion algorithmique : dès qu'un algorithme commet une erreur, même mineure, la confiance des utilisateurs s'érode plus rapidement et plus profondément qu'elle ne le ferait envers un agent humain dans la même situation. Dans le crédit, cet écart est significatif. Un refus prononcé par un conseiller laisse une marge d'explication, de dialogue et de recours. Un refus automatisé sans justification accessible ne laisse rien de tel. La défiance ainsi générée ne se limite pas au dossier concerné : elle affecte le rapport à l'institution dans son ensemble.

La pression réglementaire n'a pas pesé de manière uniforme sur le secteur. Depuis 2014, le dispositif BCE/SSM a soumis les banques traditionnelles à un cadre strict, tandis que les néo-

banques et plateformes digitales relevaient souvent de régimes moins contraignants, créant des distorsions concurrentielles. L'AI Act modifie ce paradigme : la distinction ne s'opère plus selon le statut juridique de l'entité, mais selon son rôle dans la chaîne de valeur. Tout concepteur ou déployeur d'un système d'IA à haut risque dans le domaine du crédit, qu'il s'agisse d'une banque historique, d'une néo-banque ou d'une plateforme technologique, relève désormais du même régime d'obligations (Commission européenne, 2024).

Les algorithmes de scoring ne sont pas neutres. Les biais traversent l'ensemble du processus : choix des variables, construction du modèle, traitement des résultats. Une correction ponctuelle est insuffisante. Les affaires SyRI et Schufa l'ont illustré différemment mais convergement : l'opacité d'un système produisant des effets contestables transforme un problème technique en crise juridique et en perte de confiance durable. Chouldechova et Kleinberg ont établi une frontière formelle : satisfaire une définition de l'équité algorithmique se fait presque toujours au détriment d'une autre. Ce choix entre définitions concurrentes n'est pas d'ordre technique ; il relève de décideurs politiques et sociaux, non de concepteurs de modèles. La section suivante examine ce que le RGPD, l'AI Act et la jurisprudence récente apportent à cet égard, et quelle marge de manœuvre il reste aux banques pour construire une gouvernance des modèles techniquement défendable et éthiquement tenable.

Chapitre 4 : L'opacité des modèles et le défi de l'explicabilité

4.1 Le problème de la « boîte noire »

4.1.1 Qu'est-ce qu'un modèle « boîte noire » ?

Dans des domaines aux enjeux élevés comme la médecine, la justice ou le crédit bancaire, des décisions aux conséquences durables sont guidées par des modèles que leurs propres concepteurs ne comprennent pas toujours pleinement. Ces systèmes, qualifiés de boîtes noires, produisent un résultat sans exposer leur logique interne. Rudin (2019) soutient que les outils d'interprétation post-hoc ne résolvent pas ce problème de fond : en donnant une apparence de lisibilité à des modèles fondamentalement opaques, ils peuvent introduire de nouvelles distorsions. Dans des contextes où les erreurs produisent des effets durables sur des individus et des groupes, cette opacité n'est pas un détail technique.

Burrell (2016) identifie trois sources d'opacité algorithmique : le secret industriel, la complexité technique inaccessible au non-spécialiste, et l'opacité radicale des systèmes dont même les concepteurs ne peuvent reconstituer tous les chemins logiques. C'est cette dernière qui pose le problème le plus sérieux dans le scoring crédit : un modèle fondé sur un réseau de neurones profond ou un empilement massif d'arbres décisionnels ne peut être entièrement retracé, même par ses auteurs. Lipton complexifie davantage le tableau : la notion d'explicabilité elle-même n'est pas stabilisée. Ce qu'un client attend face à un refus, ce qu'un régulateur recherche lors d'un contrôle de conformité, et ce qu'un ingénieur analyse pour corriger un modèle sont trois attentes distinctes. Formuler une exigence de transparence sans préciser laquelle revient à poser une contrainte que personne ne peut véritablement vérifier.

4.1.2 La question de l'opacité dans le crédit.

Lorsqu'un algorithme accorde ou refuse un crédit, la logique de la décision reste le plus souvent inaccessible, au client, au conseiller, parfois à l'ingénieur lui-même. Wachter, Mittelstadt et Russell (2018) montrent qu'une explication n'est acceptable que si elle permet une contestation effective et indique ce qui doit changer pour obtenir une décision différente. Sans cette dimension actionnelle, la transparence reste purement déclarative.

La lisibilité d'un modèle ne garantit pas son équité. Un système parfaitement auditable peut discriminer si le modèle reconstruit les variables protégées via des proxys : le code postal se substitue à l'origine, la fréquence des retraits au niveau de revenus déclaré. La discrimination indirecte ne disparaît pas ; elle trouve un autre chemin. Tracer chaque prédiction et auditer chaque couche du modèle peut laisser intacte une distribution disproportionnée des refus sur certains groupes. La lisibilité est nécessaire, mais insuffisante : l'équité se construit, elle ne s'obtient pas par déduction de la transparence.

L'opacité produit des effets particulièrement sévères en période de crise. Depuis 2020, les moratoires ont masqué les défauts réels et maintenu artificiellement la stabilité apparente des portefeuilles. Les modèles entraînés sur des comportements antérieurs ont continué à scorer sur une réalité qui n'existait plus, sans que leurs mécanismes internes permettent de localiser le problème. Les alertes sont arrivées tard, non par absence de signaux, mais parce que l'outil ne permettait pas de les lire (Žliobaitė et al., 2016 ; EBA, 2023b).

Le problème juridique est de même nature. Transparence et imputabilité ne sont pas des exigences secondaires : elles conditionnent la responsabilité des décideurs et la contrôlabilité des systèmes. Un décideur qui ne justifie pas sa décision n'est pas imputable. Un système qu'on ne peut pas auditer n'est pas gouvernable. Citron et Pasquale (2014) ont nommé ce phénomène le "gouvernement par l'algorithme" : lorsque des acteurs publics ou privés délèguent leurs décisions à des modèles opaques, le pouvoir ne disparaît pas, il devient illisible.

4.1.3 La distinction fondamentale entre interprétabilité et explicabilité

L'interprétabilité intrinsèque désigne une propriété du modèle lui-même, et non une couche d'explication ajoutée après coup. Cette lisibilité par construction n'implique pas de sacrifier la complexité : les Generalized Additive Models, les Explainable Boosting Machines ou certains arbres de décision optimisés modélisent des relations non linéaires, intègrent des contraintes de monotonie et restent entièrement traçables. Ce ne sont pas des régressions logistiques déguisées.

Rudin et al. (2022) ont documenté que des modèles intrinsèquement interprétables atteignent, sur plusieurs benchmarks réels, des performances comparables aux boîtes noires. Pas systématiquement, ni dans tous les contextes, mais suffisamment pour que l'équation "interprétable égale simpliste" ne puisse plus être tenue pour acquise. La compétition lancée par le CFPB en 2019 l'a illustré concrètement : le groupe de recherche de Cynthia Rudin (Duke University) y a démontré qu'un modèle conçu pour être intrinsèquement interprétable peut atteindre des performances similaires à celles des boîtes noires. Complexité technique et lisibilité ne s'excluent pas.

En pratique, lorsque l'interprétabilité intrinsèque n'est pas retenue, des méthodologies d'explicabilité post-hoc permettent d'éclairer partiellement les décisions de modèles opaques. Ce

sont ces outils, examinés plus avant, qui dominent aujourd'hui le paysage, reflétant un choix de fait en faveur de l'explicabilité a posteriori plutôt que de la transparence dès la conception. Ce choix reste contestable, et le préjugé selon lequel interprétabilité et précision seraient structurellement incompatibles s'avère empiriquement fragile.

4.2 Les outils d'IA explicable (XAI)

4.2.1 SHAP (SHapley Additive exPlanations)

SHAP (Lundberg et Lee, 2017) constitue un cadre fondamental pour l'explicabilité des modèles de scoring crédit. Fondé sur la théorie des valeurs de Shapley, il attribue à chaque variable du profil client sa contribution marginale, calculée sur l'ensemble des configurations possibles de variables. Deux implémentations coexistent : KernelSHAP, fondé sur l'échantillonnage et applicable à tout type de modèle, et TreeSHAP, exact et plus rapide, mais restreint aux modèles à base d'arbres. Lundberg et Lee démontrent que les valeurs de Shapley sont les seules, parmi toutes les méthodes de décomposition additive, à satisfaire simultanément trois propriétés : précision locale, nullité des variables absentes et cohérence. Ces trois axiomes forment un système : abandonner l'un d'eux fait disparaître l'unicité de la solution.

SHAP présente toutefois des limites significatives en pratique. Étant une méthode à référentiel fixe, les explications produites évoluent avec la population sans signal explicite, ce qui pose problème dans des environnements soumis au concept drift. Par ailleurs, dans le crédit, les variables sont quasi systématiquement corrélées entre elles : revenus, patrimoine, ancienneté bancaire, catégorie professionnelle s'influencent mutuellement. La contribution individuelle de chaque variable devient alors difficile à interpréter, deux variables liées pouvant se redistribuer mutuellement leur importance calculée. Pour un régulateur souhaitant auditer un modèle de manière stable et reproductible, ces instabilités constituent un obstacle réel.

4.2.2 LIME et explications contrefactuelles

LIME (Ribeiro, Singh et Guestrin, 2016) adopte une logique d'approximation locale : des versions légèrement perturbées du profil à expliquer sont générées, pondérées par leur proximité à l'observation originale via un noyau gaussien, puis un modèle simple est entraîné sur cet échantillon local. Pour un profil donné, l'outil produit une explication du type : "autour de ce profil, le taux d'endettement et les retards récents sont les variables les plus discriminantes." Cette approche se heurte toutefois à deux limites rédhibitoires dans un contexte réglementaire. D'une part, l'instabilité : pour un même individu, deux exécutions successives peuvent produire des classements de variables différents, rendant l'explication difficile à défendre devant un contrôleur. D'autre part, la portée : LIME ne renseigne que sur le voisinage local d'une observation, sans rien dire du comportement global du modèle, ce qui est insuffisant pour un auditeur ou un superviseur prudentiel (Ribeiro et al., 2016 ; Molnar, 2019).

Les explications contrefactuelles (Wachter, Mittelstadt et Russell, 2018) répondent à une question différente : quelles modifications minimales du profil auraient conduit à une décision inverse ? Pour un refus de crédit, cela se formule ainsi : "votre demande aurait été acceptée si votre taux d'endettement était passé sous 33 % et en l'absence d'incident de paiement au cours des six derniers mois." L'emprunteur dispose d'une information actionnelle, sans que l'établissement n'ait

à exposer la mécanique interne du modèle, ce qui s'articule favorablement avec les exigences du RGPD en matière de décisions automatisées.

La méthode présente néanmoins plusieurs angles morts. Un contrefactuel peut être techniquement valide mais concrètement inutilisable, indiquer à un emprunteur qu'il aurait été accepté avec dix ans d'ancienneté supplémentaire est exact sans être actionnable. Plusieurs contrefactuels peuvent coexister pour une même décision, parfois dans des directions incompatibles. Enfin, la méthode ne renseigne pas sur l'équité du modèle sous-jacent : elle indique ce qui aurait modifié la réponse, non si cette réponse était légitime.

4.2.3 Limites générales des outils d'XAI et pluralité des destinataires

Les notions statistiques d'importance globale des variables sont d'une autre nature. Elles ne nous renseignent pas sur la raison pour laquelle M. Dupont se voit refuser l'accès, mais elles nous indiquent pour quelles variables l'impact sur le modèle de prévision des comportements des acteurs de la sélection est le plus fort, sur l'ensemble des individus. Cette information peut être un bon point de départ pour un auditeur, afin de vérifier que l'algorithme ne repose pas sur des variables proxy, de faire du monitoring de dérives entre deux périodes, ou de rendre compte des choix effectués dans la modélisation d'un modèle, pour un superviseur prudentiel. Le problème posé est de nature légale : il repose sur le RGPD, et en particulier son article 22, qui octroie un droit à l'explication sur base individuelle, mais nullement de statistiques agrégées. Être informé que dans le refus, le taux d'endettement est en « moyenne » la variable la plus influente ne permet pas à un emprunteur de comprendre la décision. Les deux niveaux d'analyse sont complémentaires, mais non substituables (EBA, 2021 ; Molnar, 2019).

Les défauts de chaque outil pris seul existent, mais certains défauts ne sont pas corrigibles par un paramétrage plus fin. Trois méritent d'être nommés.

La première, c'est l'infidélité. Une explication post hoc n'est pas le raisonnement du modèle : c'est une reconstruction après coup. Elle peut être cohérente, lisible, et quand même inexacte. Rudin (2019) le dit sans ménagement : ce type d'explication risque de donner une autorité injustifiée à des boîtes noires, en créant l'illusion qu'on a compris ce qu'on n'a pas ouvert.

La deuxième est plus préoccupante. Slack et al. (2020) mettent en avant une approche superficielle qui, au lieu d'aiguiser les discriminations, permet de manipuler les explications post hoc des modèles discriminatoires afin qu'elles paraissent acceptables. Les explications d'un classifieur biaisé peuvent être rendues acceptables. C'est ce que le recours à la littérature a nommé le fairwashing, où le modèle n'est pas changé, mais uniquement l'explication.

La troisième est en fonction des destinataires. Le client souhaite obtenir des informations relatives à ce qu'il peut changer afin d'obtenir un accord. Le data scientist formule un diagnostic portant sur le comportement du modèle. L'auditeur interne a besoin d'une documentation traçable. Le superviseur prudentiel a besoin d'une preuve de robustesse dans le temps. Ces quatre attentes ne recouvrent pas les mêmes champs. Aucun dispositif existant ne se prévalant de les assurer toutes pourrait être retenu. Bref, une explicabilité sérieuse devrait être spécifiquement mise en œuvre en fonction de la cible.

Les outils d'XAI améliorent la lisibilité des modèles. C'est réel, mais c'est insuffisant comme critère. Une explication instable ou déconnectée du raisonnement effectif du modèle ne protège

personne : elle donne l'impression d'un contrôle qui n'existe pas. Ce qui détermine leur utilité concrète, c'est moins la sophistication technique que leur stabilité dans le temps, leur fidélité au modèle sous-jacent, leur adéquation au public visé, et leur place dans un dispositif de gouvernance qui les traite avec autant de rigueur que le modèle lui-même.

Méthode	**Niveau**	**Avantages**	**Limites**	**Pertinence pour le crédit**
SHAP	Locale et globale	Fondement axiomatique, Visualisations, Comparaison entre individus	Sensibilité au jeu de référence, Corrélations entre variables	Forte pour audit interne et justification
LIME	Locale	Intuitif, model-agnostic	Instabilité, faible portée globale	Pédagogique, fragile réglementairement
Contrefact-uels	Locale	Actionnable, compréhensible pour le client	Plausibilité, cohérence causale, multiplicité	Très pertinent pour les refus de crédit
Feature importance	Globale	Vision d'ensemble, utile pour gouvernance	Pas d'explication individualisée	Utile pour audit et documentation

4.3 Tensions entre performance prédictive et explicabilité

4.3.1 Le trade-off : mythe ou réalité ?

L'idée selon laquelle performance et interprétabilité seraient structurellement incompatibles, les modèles opaques d'un côté, les modèles lisibles mais moins précis de l'autre, est une grille de lecture commode qui résiste mal à l'examen empirique (Lessmann et al., 2015 ; Gunnarsson et al., 2021).

Sur des données tabulaires structurées, qui constituent l'essentiel des données de credit scoring, des modèles intrinsèquement interprétables tels que les arbres de décision optimaux, les GAM ou les EBM peuvent atteindre des performances proches des boîtes noires. Rudin (2019) le démontre : lorsqu'un modèle transparent est moins performant, cela tient le plus souvent à la qualité de sa conception, non à une contrainte structurelle. Hlongwane, Ramabao et Mongwe (2024) montrent que le mouvement inverse est également possible : en appliquant les valeurs de Shapley à XGBoost, forêt aléatoire, LightGBM et CatBoost, il est possible de construire des scorecards plus lisibles qu'une régression logistique sans perte de qualité prédictive.

Le compromis existe, mais il n'est pas fixe : il dépend des choix de conception et de l'effort consenti. Il demeure réel sur des données massives ou non structurées, et il serait aussi inexact de le minimiser que de le surestimer. La question pertinente n'est pas de le trancher dans l'absolu, mais de l'évaluer dans son contexte : dans quelles conditions se manifeste-t-il, et quelle marge de performance le cadre européen laisse-t-il implicitement ou explicitement en faveur de l'explicabilité ? C'est cette contrainte externe qui rend le problème traitable, en fixant des seuils que les établissements ne choisissent pas eux-mêmes.

4.3.2 Stratégies d'arbitrage et position réglementaire

Trois stratégies de conciliation entre performance et explicabilité ont émergé, avec des niveaux de robustesse distincts.

La première, le modèle substitut, consiste à entraîner un modèle simple pour imiter un modèle complexe. Dès que l'imitation est imparfaite, les explications produites ne reflètent plus le raisonnement réel. Rudin (2019) y voit non un défaut corrigeable, mais une impasse structurelle de la démarche.

La deuxième consiste à construire dès la conception un modèle intrinsèquement transparent et compétitif sur les métriques : GAM, Explainable Boosting Models (Lou et al., 2013), arbres de décision optimaux (Bertsimas et Dunn, 2017), systèmes de scoring à restrictions entières (Ustun et Rudin, 2016). La transparence est alors une propriété de conception, non un ajout a posteriori.

La troisième, probablement la plus répandue en pratique, est une segmentation par usage : modèles interprétables pour les décisions de refus, soumises à une exigence d'explication individuelle directe, et modèles complexes pour le pricing ou le scoring comportemental, où les contraintes réglementaires sont plus diffuses. L'EBA (2021) a reconnu cette approche comme une réponse proportionnée à un cadre qui ne requiert pas le même niveau d'explicabilité sur l'ensemble de la chaîne décisionnelle.

Le cadre réglementaire applicable est lui-même hétérogène. Les modèles de scoring intervenant dans la décision de crédit et les modèles prudentiels internes utilisés pour le calcul des fonds propres en approche IRB ne relèvent pas des mêmes exigences, ne s'adressent pas aux mêmes interlocuteurs et ne reposent pas sur les mêmes bases juridiques.

Du côté du RGPD, une ambiguïté persistante structure le débat. L'article 22 encadre les décisions exclusivement automatisées produisant des effets juridiques ou significatifs ; les articles 13 à 15 imposent de communiquer des "informations utiles sur la logique du traitement", sans que le texte ne précise ce que cela signifie concrètement. Wachter, Mittelstadt et Floridi (2017) et Kaminski (2019) aboutissent à des lectures divergentes, laissant les équipes juridiques dans une incertitude réelle au moment de conseiller des choix de conception.

L'AI Act, applicable depuis août 2024, impose aux systèmes évaluant la solvabilité des personnes physiques un niveau de transparence suffisant pour que les déployeurs puissent interpréter les résultats, sans préciser les méthodes acceptables. L'EBA et l'ACPR reconnaissent LIME et SHAP comme outils d'interprétation de second niveau, sans se prononcer sur leur fiabilité, que Slack et al. (2020) ont fragilisée et que Rudin (2019) remet en cause plus fondamentalement.

Les banques investissent donc sur la base de règles non finalisées : les choix techniques sont faits aujourd'hui, les standards définitifs seront précisés ultérieurement.

Les chapitres précédents ont établi la densité du problème : biais enracinés dans les données historiques, opacité des modèles, outils d'XAI utiles mais manipulables, réponse réglementaire européenne substantielle mais incomplète. Reste à déterminer ce que ce cadre régule vraiment, s'il permet de résoudre le trilemme entre performance, conformité et acceptabilité éthique, et dans quelle mesure les banques peuvent en faire un levier stratégique plutôt qu'une contrainte subie. La suite examine ces implications.

PARTIE III : Les conditions de conciliation entre performance, conformité et éthique

Cette troisième partie interroge la façon dont les banques de détail européennes peuvent articuler trois exigences mal ajustées : performance prédictive, conformité réglementaire, acceptabilité éthique. Elle s'appuie sur le cadre normatif européen s'appliquant au credit scoring algorithmique et interroge les traductions concrètes : stratégies de déploiement, choix techniques, organisation de l'entreprise, gouvernance du risque de modèle. L'objectif n'est pas de proposer une solution unique au trilemme. C'est de montrer comment il peut être géré, c'est-à-dire contenu et surveillé, sans qu'on puisse prétendre l'avoir définitivement tranché.

Chapitre 5 : Le cadre réglementaire européen

5.1 Le RGPD, les décisions automatisées et les exigences d'explication

5.1.1 L'article 22 du RGPD : une garantie centrale du droit européen de la protection des données

Parmi les textes structurant le cadre juridique européen applicable au credit scoring algorithmique, l'article 22 du RGPD occupe une place singulière. Il pose une limite que les autres textes, AI Act, Directive 2023/2225, règles prudentielles, ne formulent pas aussi directement : toute personne a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé produisant des effets juridiques ou l'affectant de manière significative. En matière de crédit, ces deux conditions sont presque toujours réunies (art. 22, §1 RGPD).

Ce droit n'est pas absolu. L'intervention humaine permet d'en sortir du champ, mais pas n'importe laquelle : un conseiller qui consulte le score, valide et passe au dossier suivant ne constitue pas une intervention au sens du texte. Ce qu'exigent la jurisprudence et la doctrine, c'est une intervention informée, susceptible d'influer sur l'issue, et non une simple validation formelle.

L'article 22, §2 prévoit trois exceptions : nécessité contractuelle, obligation légale, consentement explicite. Aucune ne va de soi en matière de crédit.

L'exception contractuelle est la plus fréquemment invoquée, souvent de manière imprécise. Pour être valide, la banque doit démontrer que l'automatisation n'est pas simplement utile ou plus efficace, mais constitue le seul moyen d'exécuter l'obligation contractuelle. La réduction des

délais ou des coûts ne satisfait pas cette exigence de nécessité, entendue dans son acception stricte.

L'exception fondée sur le consentement explicite se heurte à une difficulté symétrique. Si refuser de consentir à la décision automatisée revient en pratique à se voir refuser le financement, le consentement n'est pas libre : il est captif de la relation. Dans les deux cas, contractuel ou consensuel, la banque doit en outre garantir une intervention humaine effective et une voie de contestation authentique, non simulée.

5.1.2 Les exigences d'explication : un débat académique toujours ouvert

L'obligation d'explication en matière de credit scoring algorithmique ne découle pas de l'article 22 seul : elle résulte d'une lecture conjuguée des articles 13 et 14 (information au moment de la collecte), de l'article 15 (droit d'accès à la logique du traitement), de l'article 22 (encadrement des décisions automatisées) et du considérant 71 (droit à une explication de la décision). C'est cette articulation, et non un texte isolé, qui fonde l'exigence d'explicabilité.

Cette lecture a suscité une controverse académique réelle. Goodman et Flaxman (2017) concluent à l'existence d'un droit à l'explication ex post individualisée. Wachter, Mittelstadt et Floridi (2017) contestent cette lecture : le texte contraignant n'accorderait qu'une information ex ante sur la logique mise en oeuvre. L'enjeu pratique est considérable : informer ex ante du recours à un traitement automatisé est gérable ; expliquer ex post les raisons d'un refus suppose que le système ait été conçu pour le permettre. Selbst et Powles (2017) proposent une position médiane, selon laquelle une "information signifiante" doit permettre à la personne de comprendre, discuter et agir. Kaminski (2019) prolonge cette position en distinguant trois finalités de l'explication, comprendre le résultat, le contester en justice, ou identifier les conditions d'une décision différente, qui n'appellent ni les mêmes dispositifs techniques ni le même niveau de détail.

L'arrêt Dun & Bradstreet (C-203/22, 27 février 2025) apporte un soutien jurisprudentiel significatif à la thèse ex post. Rendu sur le fondement de l'article 15, il établit que l'intéressé a le droit de comprendre quelles données ont été utilisées et comment elles ont produit la décision. La mise à disposition de l'algorithme ne satisfait pas à elle seule cette obligation : c'est une intelligibilité substantielle qui est requise, non un accès à la machinerie. La Cour précise que cette exigence peut être conciliée avec la protection des secrets d'affaires, ménageant un espace pour les opérateurs, mais sans exemption de fournir des explications substantielles.

À ces exigences s'ajoutent deux niveaux distincts d'analyse d'impact. L'article 35 du RGPD impose une DPIA pour tout traitement susceptible de présenter un risque élevé, catégorie dans laquelle le scoring automatisé entre explicitement. L'article 27 de l'AI Act impose quant à lui une DPIA spécifique aux droits fondamentaux pour les systèmes d'IA à haut risque. Ces obligations s'articulent par ailleurs avec les exigences de validation prudentielle de Bâle et de gouvernance du risque de modèle. Ces quatre cadres ne sont pas redondants : leurs périmètres se recoupent partiellement, mais leurs exigences sont distinctes, et aucun ne dispense des trois autres.

5.1.3 L'arrêt SCHUFA (CJUE, C-634/21, 2023) : une clarification décisive

L'arrêt Schufa du 7 décembre 2023 précise le champ d'application de l'article 22. Un score produit par une agence de référence constitue une décision automatisée au sens du texte dès lors que les banques lui attribuent un rôle déterminant dans l'octroi du crédit. Ce qui importe n'est pas l'existence d'un score, mais la place qui lui est accordée dans le processus décisionnel. Si le prêteur s'en remet au score sans exercer d'appréciation autonome, la décision demeure automatisée, quand bien même un conseiller l'aurait formellement contresignée. Maintenir la qualification d'acte préparatoire aurait ouvert une brèche évidente, permettant de sortir du champ de protection par une simple validation humaine d'apparat. La Cour n'a pas retenu ce raisonnement.

La protection ainsi reconnue n'est pas absolue. La CJUE admet qu'elle puisse être limitée, sous réserve d'un arbitrage conduit par l'autorité de contrôle compétente entre la protection des intérêts commerciaux du responsable de traitement et le droit de la personne concernée à l'information et à l'explication. Le responsable de traitement ne peut invoquer systématiquement le secret commercial pour se soustraire à toute obligation de transparence : la légitimité de cette invocation doit être justifiée et appréciée par l'autorité compétente, non décidée unilatéralement par la banque.

5.1.4 Portée pratique et régime de sanctions

Les décisions de crédit reposant exclusivement sur un traitement automatisé, ou dans lesquelles le score joue un rôle déterminant sans véritable intervention humaine, tombent sous le coup de l'article 22. La banque doit alors satisfaire trois exigences : fournir une explication compréhensible, permettre à l'emprunteur de contester la décision, et garantir qu'une intervention humaine réelle est possible. Pas une validation de façade : une révision effective.

La Directive 2023/2225 sur le crédit à la consommation prolonge le dispositif sur un périmètre ciblé. Pour les contrats couverts, elle impose, en cas d'évaluation automatisée de la solvabilité, une explication claire sur la logique du traitement, ses risques et ses effets, le droit pour l'emprunteur d'exprimer son point de vue, une intervention humaine effective et un droit de révision. Elle a un champ d’application restreint : elle n’intègre pas les crédits aux professionnels et aux personnes morales. Elle ne s’appliquera qu’à compter du 20 novembre 2026, ce qui laisse une période de mise en conformité resserrée au regard des dispositions qu’elle impose techniquement.

Le problème concret demeure posé. SHAP, LIME et les contrefactuels sont les outils disponibles mais leurs limites sont bien documentées : instabilité d’une exécution à l’autre, fidélité potentielle au raisonnement du modèle, risque de fairwashing avéré par Slack et al. (2020) La question de savoir s'ils suffisent à remplir des obligations juridiques dont le contenu précis reste lui-même débattu n'a pas de réponse arrêtée. C'est une incertitude que les banques gèrent en ordre dispersé, sans filet réglementaire clair.

Sur le terrain des sanctions, la visibilité est meilleure. L'article 83, §5, du RGPD prévoit des amendes pouvant atteindre 20 millions d'euros ou 4 % du chiffre d'affaires annuel mondial, selon le montant le plus élevé. Pour une grande banque universelle européenne, le chiffre d'affaires mondial se chiffre en dizaines de milliards : l'exposition est réelle, et elle porte précisément sur les manquements aux obligations découlant de l'article 22.

5.2 L'AI Act (2024) : un cadre horizontal ambitieux de régulation de l'IA

5.2.1 Le credit scoring classé « haut risque »

L'AI Act, Règlement (UE) 2024/1689, est entré en vigueur le 1er août 2024. C'est le premier cadre européen à réguler l'IA de façon horizontale, c'est-à-dire tous secteurs confondus. Le texte classe les usages par niveau de risque : certains sont purement interdits, d'autres soumis à des obligations lourdes, d'autres encore à peu près libres. Le credit scoring tombe dans la catégorie haute risque. L'Annexe III vise les systèmes d'évaluation de la solvabilité et de scoring des personnes physiques, à l'exception, notable, des outils de détection de fraude financière.

5.2.2 Les obligations pour les systèmes à haut risque

Les obligations ne sont pas les mêmes selon le rôle. La banque qui développe son propre système est fournisseur (provider), avec tout ce que ça implique en termes de contraintes. Est considéré comme déployeur (deployer) qui utilise un outil acheté à un tiers, au régime en principe plus léger. En pratique, la frontière est plus mouvante. La banque qui adapte de façon approfondie un système d'un prestataire, ou qui l'intègre dans sa propre chaîne de décision, risque d'être requalifiée, au moins partiellement, en fournisseur. Ce n'est pas un cas d'école : c'est précisément le cas de nombreuses banques qui s'appuient sur des solutions de scoring externes tout en les configurant à leur main.

Le Titre III, Chapitre 2 impose notamment :

Art. 9 - Gestion des risques : système continu d'identification et de traitement des risques (biais, discrimination, instabilité, concept drift).

Art. 10 - Gouvernance des données : données pertinentes, représentatives, aussi exemptes d'erreurs que possible ; attention aux biais potentiels - objectif difficilement atteignable dans des environnements de données historiques imparfaits, sans en garantir l'absence totale.

Art. 11 - Documentation technique : démonstration de la conformité auprès des autorités.

Art. 12 - Tenue de registres : enregistrement automatique des événements pertinents, essentiel pour l'audibilité ex post.

Art. 13 - Transparence : permettre aux déployeurs d'interpréter et d'utiliser correctement les résultats.

Art. 14 - Contrôle humain : human oversight by design, renforçant l'article 22 du RGPD.

Art. 15 - Exactitude, robustesse et cybersécurité : fonctionnement cohérent toute la durée d'exploitation.

Art. 26 - Obligations des déployeurs : utilisation conforme aux instructions, contrôle humain, contrôle des données d'entrée, conservation des journaux lorsqu'ils sont sous leur contrôle, information des personnes concernées dans certains cas.

Art. 27 - Évaluation d'impact sur les droits fondamentaux (FRIA) : les déployeurs de systèmes d'évaluation de la solvabilité des personnes physiques doivent réaliser, avant déploiement, une analyse de l'impact sur les droits fondamentaux. Cette exigence, voisine de la logique DPIA du RGPD, élargit le contrôle aux risques de discrimination, d'exclusion et d'atteinte à l'accès aux

services essentiels. Les deux évaluations doivent être articulées de manière cohérente dans la gouvernance interne.

5.2.3 Calendrier d'application, sanctions et double conformité

Les obligations applicables aux systèmes à haut risque de l'Annexe III entreront en vigueur le 2 août 2026. Le délai paraît raisonnable en théorie ; en pratique, les modèles de credit scoring sont souvent anciens, incomplètement documentés et imbriqués dans des processus décisionnels non conçus pour l'auditabilité. Construire la conformité à partir de cet état prend du temps. Les dispositions sur les pratiques interdites, quant à elles, sont applicables depuis février 2025, sans délai de grâce.

Le barème des sanctions suit une logique de plafonds alternatifs, le montant le plus élevé étant retenu : pratiques interdites, 35 M€ ou 7 % du chiffre d'affaires mondial ; manquements aux obligations des opérateurs, 15 M€ ou 3 % ; informations incorrectes transmises aux autorités, 7,5 M€ ou 1,5 %. Pour un grand groupe bancaire, le pourcentage constituera presque systématiquement le plafond effectif.

La difficulté centrale ne réside pas dans le respect de chaque texte pris isolément, mais dans leur coexistence. Le RGPD protège les données personnelles, l'AI Act régule le système d'IA lui-même, les exigences prudentielles s'y superposent. Chacun de ces cadres obéit à sa propre logique, impose ses propres obligations et relève d'autorités de contrôle distinctes. Une banque déployant un modèle de credit scoring doit coordonner son DPO, sa direction des risques et potentiellement plusieurs régulateurs aux priorités partiellement divergentes. Cette coordination n'est pas un détail organisationnel : c'est là que la conformité se joue ou se rate.

Les échéances mentionnées dans ce travail sont celles en vigueur à la date de rédaction. Le calendrier d'application de l'AI Act ayant fait l'objet de discussions récentes au niveau européen, il est conseillé de vérifier l'état du texte à la date du dépôt.

5.2.4 Comparaison internationale : le « Brussels Effect »

Anu Bradford (Columbia Law School) a théorisé dans *The Brussels Effect* (2020) que l'UE exporte ses normes sans jamais avoir à les imposer. L'idée est simple : quand un marché est assez grand, les entreprises qui veulent y accéder s'alignent, puis n'ont plus vraiment envie de maintenir deux standards en parallèle. Dans la finance, c'est encore plus direct. Les banques sont déjà sous Bâle, MiFID II et DORA. Ajouter un régime de gouvernance de l'IA propre au marché européen ne leur coûte pas grand-chose de plus. En garder un second pour les autres juridictions, si. Le standard de l'UE s'impose donc moins par conviction que par calcul.

Trois grandes zones réglementaires, trois logiques différentes. Aux États-Unis, il n'existe pas de cadre fédéral unifié pour l'IA : la FTC, le CFPB, l'OCC et la Fed interviennent chacun dans leur domaine, pendant que certains États comblent le vide à leur façon. Le Colorado AI Act de 2024, l’illustration la plus frappante, se réfère directement au modèle de régulation européen là où Washington se divise. Le Royaume-Uni opère un choix tout autre : au lieu d’une loi obligatoire, il laisse à la FCA et à la PRA la liberté de créer les conditions de mise en œuvre d’une régulation souple sur l’IA tout en affirmant jouer la carte de l’innovation. Quant à la Chine, elle a bien sûr

ses propres règles mais dans un cadre institutionnel si original que sa valeur de comparaison avec l'Europe ou les États-Unis est faible.

5.3 Les directives prudentielles : Bâle III et les modèles IRB

5.3.1 L'approche IRB et les différents types de scoring

Le cadre prudentiel issu des réformes finales de Bâle III, que certains appellent "Bâle IV", a été transposé en droit européen par le CRR et la CRD. L'approche IRB, introduite par Bâle II en 2004, est au cœur du sujet ici : elle permet aux banques qui y sont autorisées d'estimer elles-mêmes leur probabilité de défaut (PD), leur perte en cas de défaut (LGD) et leur exposition au moment du défaut (EAD). Ce n'est pas anodin. Ces trois paramètres déterminent les actifs pondérés par le risque (RWA), et les RWA déterminent les fonds propres exigés. Autrement dit, la qualité du modèle interne a une incidence directe sur le bilan.

Trois catégories d'outils de scoring doivent être distinguées selon leur régime réglementaire :

Type de scoring	**Finalité**	**Destinataire principal**	**Cadre réglementaire principal**	**Risque d'explicabilité**
Scoring d'octroi	Décision de crédit au moment de la demande	Client / emprunteur	RGPD art. 22, AI Act Annexe III, Dir. 2023/2225	Élevé : explication individuelle requise
Scoring comportemental	Suivi du risque après octroi	Institution financière	RGPD (selon effets), cadre prudentiel	Moyen à élevé selon les effets produits
Modèles IRB	Calcul prudentiel des fonds propres	Superviseur prudentiel (BCE, MSU)	CRR/CRD, Guidelines EBA	Élevé : capacité de validation par le superviseur

Les modèles IRB ne tombent pas automatiquement sous l'AI Act. Tout dépend de leur intégration : s'ils participent à un processus d'évaluation de solvabilité visé par l'Annexe III, la qualification de système d'IA à haut risque devient plausible. Sinon, ils restent en dehors du champ direct du règlement.

5.3.2 Validation des modèles, Guidelines EBA et machine learning

Les modèles IRB sont soumis à un examen prudentiel complet par la BCE dans le cadre du MSU : documentation technique, backtesting, stress testing. L'exigence la plus contraignante est d'ordre qualitatif : la banque doit démontrer qu'elle comprend son modèle, pas seulement qu'il produit des résultats dans les normes. Cette exigence devient particulièrement inconfortable face à des algorithmes peu interprétables.

Les guidelines EBA sur l'octroi et le suivi du crédit (EBA/GL/2020/06) posent une exigence structurante : l'explicabilité des décisions doit être démontrable tout au long du cycle de vie du modèle, de la conception à la désappropriation, en passant par l'usage, le recalibrage et la validation. À chaque étape, les choix de modélisation doivent pouvoir être justifiés.

L'EBA admet le recours au machine learning pour les modèles IRB, sous trois conditions : explicabilité démontrable, stabilité dans le temps, et validation par une fonction indépendante. La consultation de 2021 fournit quelques cas d'usage, quatorze au total, sans prétention à la représentativité statistique. Les résultats constituent un signal faible : les valeurs de Shapley arrivent en tête (40 %), devant l'amélioration de la documentation méthodologique (28 %) et les outils graphiques (20 %). Sur le terrain, le machine learning reste cantonné à des usages ciblés, principalement la modélisation de la PD, et les banques n'ont pas encore arrêté de position définitive sur la forme attendue du retour explicatif.

5.3.3 CRR3, l'output floor et ses implications stratégiques

CRR3, applicable depuis le 1er janvier 2025, introduit l'output floor : les RWA calculés par les modèles internes ne peuvent excéder 72,5 % des RWA qu'aurait produits l'approche standard. Ce plancher monte en charge jusqu'au 31 décembre 2029 et s'applique pleinement à partir du 1er janvier 2030.

L'effet secondaire est significatif. Jusqu'alors, un modèle plus sophistiqué pouvait justifier son coût par les économies de capital qu'il permettait. Ce raisonnement perd de sa force dès lors que le gain est borné à 72,5 % de la méthode standard. Les modèles simples et interprétables deviennent comparativement plus compétitifs : moins coûteux à construire, plus aisés à valider, pour un avantage en fonds propres désormais comparable. Il ne s'agit pas d'une rupture, mais d'un réajustement structurel des incitations à l'intérieur du système bancaire.

L'AI Act et les exigences prudentielles opèrent sur des périmètres distincts : l'AI Act cible les systèmes d'évaluation intervenant dans la décision d'octroi de crédit, non les modèles de calcul prudentiel. Les deux régimes ne se recoupent qu'en hypothèse. Le point critique est le use-test : si un modèle IRB ne sert pas exclusivement à calculer des RWA mais contribue également à la décision d'octroi, la qualification au titre de l'AI Act redevient ouverte. Ce n'est pas une conséquence automatique, mais c'est précisément le type d'interaction que les banques auraient tort de négliger, à mesure que les superviseurs s'intéressent à l'usage réel des modèles.

5.3.4 La conformité prudentielle comme amplificateur de l'exigence d'explicabilité

C'est ici que les exigences prudentielles rendent le trilemme de ce mémoire pleinement contraignant. Un modèle de machine learning peut surpasser la régression logistique sur tous les indicateurs de performance et ne pas passer la validation prudentielle, non parce qu'il est incorrect, mais parce que le superviseur juge l'explication incomplète, instable dans le temps ou mal fondée. Le rejet n'est pas administrativement neutre : il entraîne le retour forcé à l'approche standard, avec des conséquences en cascade sur les RWA, le besoin en fonds propres, la rentabilité des portefeuilles et la stratégie commerciale. Pour les grandes banques européennes, ces scénarios se chiffrent en milliards. La sophistication technique ne vaut rien si elle ne franchit pas le filtre prudentiel.

Le RGPD et l'AI Act exigent une explication. La réglementation prudentielle exige autre chose : la preuve que la banque maîtrise réellement son modèle, et pas seulement qu'elle sait en parler. Une explication post-hoc bien formulée ne satisfait pas cette exigence. Ce qui est attendu, c'est une maîtrise institutionnelle effective, compréhension interne de la logique du modèle, validation par une fonction indépendante, contrôlabilité dans le temps. La barre est nettement plus haute, et elle impose de repenser la conception des modèles dès l'origine.

Trois régimes, trois logiques, mais une pression qui converge vers le même point. Le RGPD part du droit individuel : la personne concernée peut contester une décision automatisée, exiger une explication et demander une intervention humaine, un périmètre précisé par les arrêts Schufa et Dun & Bradstreet. L'AI Act prend le modèle comme objet et l'encadre sur l'ensemble de son cycle de vie, de la conception au retrait. Les règles prudentielles court-circuitent les deux : indépendamment de ce que prescrivent les règlements, si le superviseur juge le modèle insuffisamment explicable, la banque en paie le prix en fonds propres. C'est cette dernière contrainte qui rend les deux premières concrètes.

Texte	**Champ d'application**	**Obligations principales**	**Acteur concerné**	**Sanctions / conséquences**	**Impact sur le credit scoring**
RGPD	Données personnelles, décisions automatisées	Information, accès, contestation, intervention humaine, DPIA	Responsable de traitement	Jusqu'à 20 M€ ou 4 % CA mondial	Explication individuelle requise
Directive 2023/2225	Contrats de crédit aux consommateurs	Explication de l'évaluation automatisée, intervention humaine, révision	Créancier	Sanctions nationales	Solvabilité du consommateur
AI Act	Systèmes IA à haut risque (personnes physiques)	Gestion des risques, données, documentation, logs, transparence, contrôle humain, FRIA (art. 27)	Fournisseur / déployeur	Jusqu'à 35 M€ ou 7 % CA mondial	Gouvernance ex ante du système

EBA Guidelines	Octroi et suivi des crédits	Gouvernance, données, suivi, documentation sur tout le cycle de vie	Banque	Risque de supervision	Cycle de vie du modèle
CRR/CRD / IRB	Modèles internes prudentiels	Validation, backtesting, stress testing, compréhension effective du modèle	Banque IRB	Hausse des RWA, remédiation, retour à l'approche standard	Validabilité prudentielle

Performance prédictive, explicabilité, équité algorithmique : les trois ne sont plus des critères d'évaluation optionnels. Ce sont des obligations légales, avec des sanctions financières, des risques prudentiels et des responsabilités juridiques à la clé en cas de manquement. Les solutions techniques ne peuvent donc pas être jugées uniquement sur leur capacité à prédire. Elles doivent aussi tenir face à ce cadre normatif. C'est précisément ce que ce mémoire cherche à évaluer.

Chapitre 6 : Stratégies de déploiement responsable

6.1 Réduire les biais : stratégies techniques et organisationnelles

6.1.1 Stratégies techniques de débiaisage : une intervention à trois niveaux du pipeline

Dans la littérature sur le fairness-aware machine learning, les techniques de correction du biais se distinguent par leur moment d'intervention, ce qui détermine ce qu'elles peuvent corriger, à quel coût, et avec quelles garanties sur le résultat final (Mehrabi et al., 2021 ; Barocas, Hardt et Narayanan, 2019).

a) Pré-traitement des données

L'approche la plus instinctive consiste à retirer les variables sensibles. Elle échoue en pratique : dans un espace à grande dimension, le modèle reconstitue ces informations via des proxys. Le code postal se substitue à l'origine, le type d'emploi à la catégorie sociale, l'historique de consommation à ce que les variables évincées auraient dit. La suppression des variables sensibles constitue une protection essentiellement cosmétique (Barocas et Selbst, 2016 ; Prince et Schwarcz, 2020).

Le rééchantillonnage, suréchantillonnage des groupes sous-représentés, sous-échantillonnage des autres, ou génération d'observations synthétiques via SMOTE (Chawla et al., 2002), n'est efficace que si le biais provient d'un déséquilibre de représentation. Lorsqu'il est inscrit dans les étiquettes historiques, la définition de la variable cible ou la structure des corrélations, le rééquilibrage ne change rien au fond.

Feldman et al. (2015) proposent une approche plus ciblée : plutôt que de supprimer les variables sensibles, transformer leurs distributions pour rompre leur relation avec l'appartenance au groupe protégé tout en préservant l'information prédictive. Cette neutralisation des proxys peut toutefois dégrader les performances et soulève une tension réglementaire : en modifiant les données pour les éloigner de la réalité historique, on corrige le biais algorithmique mais on introduit un écart entre le modèle et le monde, ce qu'un cadre prudentiel exigeant la fidélité de la donnée ne saurait admettre sans justification.

b) In-processing

Les techniques d'in-processing intègrent la fairness directement dans la fonction objective du modèle, sous forme de contrainte ou d'objectif multi-critères. Cette reformulation impose de choisir explicitement un critère d'équité, parité démographique, equal opportunity ou equalized odds, ce qui est loin d'être neutre sur le plan technique et normatif.

L'adversarial debiasing (Zhang, Lemoine et Mitchell, 2018) entraîne en parallèle un réseau adversarial chargé de deviner l'appartenance au groupe protégé à partir des prédictions. Lorsqu'il y parvient, le modèle principal est pénalisé, l'incitant à produire des prédictions non révélatrices du groupe. La méthode est techniquement ambitieuse, mais instable et limitée aux formes de biais que l'adversaire peut détecter.

Kozodoi et al. (2022) quantifient ce compromis sur un portefeuille de crédit réel : les contraintes de fairness réduisent effectivement les biais, mais produisent une détérioration mesurable de la performance prédictive et de la profitabilité. Ce résultat constitue une validation empirique du trilemme central de ce mémoire : la tension entre performance, conformité et acceptabilité éthique n'est pas une construction théorique, elle est observable dans les données.

c) Post-traitement

Le post-traitement consiste à fixer des seuils de décision distincts par groupe pour atteindre la parité visée, en abaissant par exemple le seuil d'acceptation pour les groupes structurellement désavantagés par le modèle (Hardt, Price et Srebro, 2016). L'approche présente l'avantage d'être applicable à tout modèle existant sans le modifier. Elle soulève néanmoins deux obstacles majeurs. D'une part, l'instauration de règles explicitement différenciées selon le groupe se heurte au principe d'égalité de traitement et aux directives anti-discrimination (2000/43/CE ; 2004/113/CE). D'autre part, le droit européen encadrant strictement l'usage des données révélant l'origine raciale ou ethnique, l'évaluation directe de la fairness sur cette dimension est difficile, voire impraticable dans la plupart des contextes opérationnels.

6.1.2 Stratégies organisationnelles

Les dispositifs techniques ne suffisent pas. Quatre stratégies organisationnelles s'y ajoutent.

Une équipe homogène a des angles morts qu'elle ne voit pas, par définition. Diversifier les profils, en genre, en origine, en formation, en parcours, c'est multiplier les points de vue capables d'identifier des biais que personne n'aurait songé à chercher.

Certaines banques ont structuré cette supervision autour de comités d'éthique dédiés, chargés d'examiner les modèles avant déploiement. L'idée : faire la preuve que les risques de biais ont bien été identifiés, mesurés et atténués (et pas seulement documentés en surface). ING pousse la démarche à l'extrême : processus de revue en 20 étapes dans lequel chacun systématiquement des systèmes d'IA est soumis à 140 risques potentiels.

Avant chaque mise en production (et à intervalles réguliers par la suite), le modèle est soumis à toute une batterie de tests qui touchent à la plupart des métriques de fairness : démographic parity, equal opportunity, equalized odds, calibration by group. Les résultats sont documentés et archivés, ce qui permet de tracer l'évolution du comportement du modèle dans le temps.

Le red-teaming, c'est une équipe dont le travail est de chercher ce que le modèle fait mal, pas de valider ce qu'il fait bien. Elle traque les proxies discriminants, génère des demandeurs fictifs, examine des profils atypiques, et documente tout ce qui produit un résultat difficile à justifier.

6.1.3 Les limites irréductibles

Chouldechova (2017) et Kleinberg et al. (2016) l'ont démontré de manière mathématique : lorsque les taux de base diffèrent entre groupes, aucune combinaison des différentes manières possibles d'évaluer les inégalités d'accès aux droits ne peut être satisfaite simultanément, et il faut alors faire un choix. Or ce choix n'est pas technique, privilégier la parité démographique plutôt que l'equal opportunity (l'égalité d'accès aux droits) décide de ceux qui supportent le coût des erreurs du modèle, et en quelle proportion. C'est donc un jugement de valeur, non un réglage de paramètre. D'ailleurs, la correction des biais peut au final dégrader la performance prédictive, donc les banques ne peuvent pas contourner ou fuir cet arbitrage. Elles doivent l'expliquer, l'assumer, le soumettre à une gouvernance explicite, plutôt que de le laisser se régler dans le silence des choix techniques.

Restent le paradoxe que le cadre européen n'a pas résolu, car c'est une condition nécessaire pour apprécier la fairness raciale du modèle ayant nécessité d'avoir accès à des données d'origine raciale ou ethnique. Or ce type de données est soumis en Europe, et particulièrement en France, à un régime juridique très restrictif. Dans la plupart des contextes opérationnels, cette évaluation est donc difficile, voire impraticable. On se retrouve dans une position inconfortable : censé garantir l'équité d'un modèle sur une dimension qu'on n'a légalement pas le droit de mesurer (Hacker, 2018 ; Chouldechova et Roth, 2020).

6.2 Répondre à l'opacité : déploiement de l'IA explicable en contexte bancaire

6.2.1 Une matrice de déploiement XAI par audience

Un client refusé, un régulateur en audit et un data scientist qui débogue le modèle n'ont pas besoin du même type d'explication. L'XAI ne peut pas répondre à ces trois attentes avec un dispositif unique.

Audience	**Besoin principal**	**Outil XAI**	**Exigence réglementaire**

Emprunteur refusé	Comprendre pourquoi et comment améliorer sa situation	Explications contrefactuelles (Wachter et al., 2018)	RGPD art. 22 ; art. 13-15 sur la logique sous-jacente ; considérant 71 ; droit à l'intervention humaine, à la contestation et à l'explication selon les cas
Analyste crédit	Vérifier la cohérence de la recommandation	Valeurs SHAP individuelles, feature importance locale	Gouvernance interne, human-in-the-loop
Régulateur / Auditeur	Valider le modèle, vérifier l'absence de biais	SHAP global, partial dependence plots, model cards	AI Act art. 11, 13 ; EBA guidelines IRB
Direction générale	Comprendre les risques et arbitrer	Rapports de synthèse, KPIs performance et fairness	Gouvernance d'entreprise, model risk management

Le « droit à l'explication » est souvent perçu en tant que garantie claire du RGPD. La réalité juridique est plus nuancée. L'article 22 consacre essentiellement le droit à ne pas être soumis à une décision reposant exclusivement sur un traitement automatisé ayant pour effet de produire des conséquences juridiques ou significatives. Il garantit la possibilité de demander une intervention humaine, d'exprimer son point de vue et de contester la décision. L'explication de la logique sous-jacente découle, elle, des articles 13 à 15 et du considérant 71, sans former un droit général et autonome à l'explication.

6.2.2 L'XAI comme composant natif du pipeline

L'XAI ne peut pas être greffée après coup. Ce serait en contradiction directe avec la transparency by design qu'exige l'article 13 de l'AI Act. Elle doit faire partie du pipeline dès le départ, et produire pour chaque décision : le score de probabilité de défaut, les contributions SHAP, une explication contrefactuelle en cas de refus, et un enregistrement horodaté conforme à l'article 12. Ce monitoring continu a aussi un bénéfice opérationnel : il rend visible le concept drift avant qu'il ne force une recalibration non planifiée (Žliobaitė et al., 2016 ; EBA, 2023b).

6.2.3 Les model cards

Les model cards, proposées par Mitchell et al. (2019, ACM FAccT), sont des fiches standardisées dans lesquelles sont documentés la description du modèle, les données d'entraînement, les métriques de performance agrégées et par sous-groupe, les indicateurs de fairness, les biais connus, les cas d'usage prévus et exclus, ainsi que les recommandations de suivi. Le contenu

recoupe point par point les exigences de l'article 11 de l'AI Act. Google, Microsoft et Hugging Face les intègrent déjà. Pour les banques, ce n'est pas une innovation à expérimenter, c'est un standard à adopter.

6.2.4 Les limites persistantes de l'XAI

L'XAI n'est pas un gage de fairness. Un modèle peut être parfaitement explicable et fondamentalement biaisé : les deux ne s'excluent pas. Le risque de fairwashing est documenté : Slack et al. (2020) montrent que les outils post hoc peuvent être manipulés pour produire des explications qui dissimulent les biais réels plutôt que de les exposer. Et même sans manipulation délibérée, ces explications restent des approximations. Rudin (2019) le rappelle sans ambiguïté : rien ne certifie qu'elles correspondent au raisonnement effectif du modèle.

6.3 Assurer la conformité : compliance by design et gouvernance des modèles

Ce que les praticiens interrogés disent, en substance, c'est que le problème n'est pas algorithmique. Le choix du modèle n'est pas ce qui résiste. Ce qui résiste, c'est la capacité de l'organisation à documenter, valider, expliquer et surveiller dans la durée. Les textes européens récents poussent exactement dans cette direction : l'enjeu n'est plus seulement la performance prédictive, c'est la gouvernance du cycle de vie complet des systèmes d'IA.

6.3.1 Le compliance by design

Le compliance by design prolonge la logique du privacy by design inscrit à l'article 25 du RGPD et théorisé par Ann Cavoukian. À l'instar de la transparence, de l'équité et de la conformité prudentielle qui ne viennent pas se rajouter à un modèle déjà conçu, mais qui sont intégrées dès les premières lignes de code, les équipes data science interagissent dès le départ avec les équipes juridiques, de conformité et de risques pour traduire en spécifications techniques les exigences réglementaires (RGPD, AI Act, CRR/CRD, recommandations EBA/BCE), choisir l'algorithme proportionné aux contraintes, déterminer en amont les métriques de fairness, intégrer des outils d'XAI dès le prototypage et documenter chaque étape comme prévu à l'article 11 de l'AI Act.

L'éthique par conception, cela ne renvoie pas seulement à la conformité. Le Groupe d'experts de haut niveau de la Commission européenne (AI HLEG, 2019) a inventorié sept exigences : « le contrôle humain, la robustesse technique, la vie privée, la transparence, la non-discrimination, le bien-être social et environnemental et la responsabilité ». Ce catalogue sera pris en compte pour rédiger l'AI Act. Concernant les banques, il leur offre ce que le droit n'ose proposer, à savoir un cadre pour bâtir des systèmes de scoring qui ne soient pas seulement conformes, mais justifiables.

6.3.2 Les trois lignes de défense du Model Risk Management

La gouvernance des modèles s'organise autour du Model Risk Management (MRM), dont la référence fondatrice est la lettre SR 11-7 de la Réserve fédérale (2011), largement intégrée par les superviseurs européens. Le cadre articule trois composantes, développement rigoureux, validation indépendante et gouvernance d'ensemble, mises en oeuvre selon la logique des trois lignes de défense.

La première ligne recouvre le travail documentaire : traçabilité de chaque étape du développement, justification des choix méthodologiques, tests de performance et de fairness. La documentation technique requise par l'AI Act en découle directement.

La deuxième ligne, assurée par la fonction de validation indépendante (MVT), soumet le modèle à un examen externe portant sur sa qualité, sa robustesse, sa conformité et ses biais potentiels. Cette exigence est explicitement posée par le cadre prudentiel de Bâle et les guidelines EBA sur les modèles IRB (EBA, 2021) : un modèle opaque rend cet examen sérieux difficile, ce qui renforce l'argument en faveur de l'explicabilité.

La troisième ligne, l'audit interne, n'a pas vocation à examiner le modèle en détail, contrairement à la MVT. Son rôle est de vérifier que le dispositif de gouvernance respecte les politiques internes et les exigences réglementaires. Une maîtrise technique approfondie n'y est pas requise.

6.3.3 Obligations réglementaires du déployeur : article 26, registre interne, FRIA et articulation avec le RGPD

Pour les personnes physiques, les systèmes de scoring de crédit sont classés à haut risque au titre de l'annexe III, point 5(b) de l'AI Act, à l'exception des systèmes dédiés à la détection de fraude. Les fournisseurs sont soumis aux obligations de documentation (art. 11), de transparence (art. 13) et d'enregistrement dans la base de données européenne (art. 49).

Les banques agissant en qualité de déployeurs relèvent de l'article 26, qui leur impose notamment de respecter les instructions du fournisseur, mettre en oeuvre des mesures techniques et organisationnelles adéquates, assurer la supervision humaine, vérifier la pertinence des données d'entrée, conserver les logs, informer les personnes concernées et coopérer avec les autorités compétentes. L'obligation d'enregistrement dans la base de données européenne incombe au fournisseur, non au déployeur. La tenue d'un registre interne des systèmes d'IA déployés est toutefois recommandée aux déployeurs privés afin de réduire le risque lié au modèle et de faciliter la démonstration de conformité lors d'un contrôle.

Avant tout déploiement, l'article 27 impose la réalisation d'un Fundamental Rights Impact Assessment (FRIA), destiné à identifier les droits fondamentaux concernés, non-discrimination, protection des données, accès au crédit, à évaluer la probabilité et la gravité des atteintes potentielles, et à décrire les mesures d'atténuation adoptées.

La FRIA s'articule nécessairement avec l'analyse d'impact sur la protection des données (AIPD) requise par le RGPD, mais leurs objets diffèrent : l'AIPD porte sur les risques liés au traitement des données personnelles, tandis que la FRIA couvre les risques afférents à l'usage du système au regard des libertés fondamentales. Ces deux exercices doivent faire l'objet d'une démarche intégrée de validation pré-déploiement, et non être traités comme deux formalités distinctes.

6.4 Architecture en trois niveaux : du modèle à la stratégie

6.4.1 Niveau 1 : Le modèle

Le choix de l'algorithme repose sur une matrice de décision qui met en regard la complexité du modèle et les enjeux auxquels il répond.

Type de crédit	Montant typique	Risque réglementaire	Algorithme recommandé
Crédit renouvelable / BNPL	< 1 000 €	Modéré à élevé selon volume et population cible	ML complexe (XGBoost) acceptable avec XAI
Crédit consommation standard	1 000 - 15 000 €	Élevé	ML avec explications contrefactuelles obligatoires
Crédit consommation élevé	15 000 - 75 000 €	Très élevé	Modèle interprétable privilégié ou hybride
Crédit immobilier	> 75 000 €	Maximal	Modèle intrinsèquement interprétable + human-in-the-loop

Le niveau de risque ne se mesure pas au montant unitaire : un produit à fort volume comme le BNPL peut produire des effets discriminatoires massifs à l'échelle du portefeuille, y compris sur des populations qui n'auraient pas déclenché d'alerte au niveau individuel.

L'approche recommandée est modulaire. Un modèle interprétable, régression logistique ou Explainable Boosting Machine, occupe le poste de champion. Un challenger de machine learning plus performant peut lui être opposé, sous réserve de satisfaire simultanément trois conditions : performance, explicabilité et conformité. En l'absence de l'une d'entre elles, le champion demeure en place.

En amont de l'entraînement, les données font l'objet d'un audit portant sur leur représentativité, les biais potentiels dans les variables comme dans les labels, et la traçabilité de leur origine et de leurs transformations. Cette dernière exigence est désormais fonctionnellement consacrée par l'article 10 de l'AI Act comme obligation documentaire.

La validation pré-déploiement s'opère à l'intersection de la FRIA (art. 27 AI Act) et, lorsque les conditions sont réunies, de l'AIPD (art. 35 RGPD). Elle intègre un test de fairness obligatoire : les métriques sont définies préalablement à l'évaluation du modèle, tout seuil non atteint est bloquant pour le déploiement, et la décision finale relève d'une instance de go/no-go transversale réunissant plusieurs métiers.

6.4.2 Niveau 2 : La gouvernance

Les comités modèles sont pluridisciplinaires par construction, réunissant data scientists, juristes, risk managers et spécialistes de l'éthique, chacun disposant d'un droit de veto sur le déploiement. La fréquence de révision s'adapte au profil du modèle : annuelle pour les modèles stables, trimestrielle pour les modèles à volume élevé ou récemment mis en production.

La documentation s'organise sur plusieurs niveaux, des model cards au registre interne des systèmes d'IA, en assurant une traçabilité complète des décisions conforme à l'article 12 de l'AI Act. Les éléments produits doivent être conservés suffisamment longtemps pour permettre un audit ex post exploitable.

La supervision humaine n'est utile que si elle est effective. Le phénomène de rubber stamping, documenté par Dietvorst, Simmons et Massey (2015), dans lequel l'analyste valide mécaniquement la recommandation algorithmique sans en examiner la pertinence, ne satisfait ni aux exigences de l'article 22 du RGPD ni à celles de l'article 14 de l'AI Act. Le prévenir suppose trois conditions : une formation adéquate des analystes, un accès aux explications sous formes SHAP et contrefactuelles, et des incitations orientées vers l'exercice du jugement critique plutôt que vers la validation mécanique.

6.4.3 Niveau 3 : La stratégie

La conformité anticipée constitue un avantage concurrentiel concret à partir du 2 août 2026, date d'entrée en application de la majorité des obligations de l'AI Act. L'exposition en cas de manquement est significative : jusqu'à 4 % du chiffre d'affaires mondial au titre du RGPD, jusqu'à 15 millions d'euros ou 3 % du chiffre d'affaires pour les manquements aux obligations relatives aux systèmes à haut risque, et jusqu'à 35 millions d'euros ou 7 % pour les pratiques prohibées (art. 5 AI Act). Au-delà du risque financier, les conséquences potentielles incluent le retrait de la labellisation des modèles IRB, une atteinte réputationnelle, un renforcement de la surveillance prudentielle, voire une remise en cause du modèle d'activité de l'établissement.

Les grandes banques disposent d'atouts structurels : historique de données étendu, infrastructure de conformité établie, relation de long terme avec le régulateur. Leur handicap est la vitesse d'exécution. Les fintechs déploient plus rapidement, mais avec une expérience réglementaire limitée et des moyens de conformité souvent insuffisants. L'AI Act rééquilibre partiellement ce rapport en imposant les mêmes obligations à tous les opérateurs visés par l'Annexe III, neutralisant ainsi l'avantage de l'agilité.

Ce qui demeure différenciant, c'est la capacité à transformer la conformité en signal de confiance. Un système d'IA explicable, audité et équitable n'offre pas seulement une sécurité juridique renforcée : il constitue un signal crédible adressé aux consommateurs, aux régulateurs et aux investisseurs, dans un secteur où la confiance se construit lentement et se perd rapidement.

6.5 Institutionnaliser le trilemme plutôt que le résoudre

Le trilemme entre performance prédictive, conformité réglementaire et acceptabilité éthique ne se résout pas : il se gère. Chouldechova (2017) et Kleinberg et al. (2016) l'établissent formellement : lorsque les taux de base diffèrent entre groupes, aucun système ne peut satisfaire simultanément toutes les définitions concurrentes de l'équité. Ce n'est pas une contrainte technique à contourner, mais une contrainte structurelle. La question n'est pas de l'éliminer, mais de choisir explicitement la façon d'arbitrer et de gouverner ces arbitrages de manière transparente.

Institutionnaliser le trilemme, ce n'est pas le résoudre : c'est construire les conditions pour le gérer sérieusement dans la durée. Cela suppose des structures organisationnelles dédiées, comités d'éthique, validation indépendante, processus de revue périodique, ainsi que des outils techniques rendant les arbitrages visibles et auditables : XAI intégrée, model cards, FRIA articulée avec l'AIPD, tests de fairness systématiques et monitoring continu. Ces arbitrages ne sont jamais définitifs : lorsque le contexte évolue, ils doivent être reformulés.

Ce mémoire ne propose pas une solution au trilemme, mais un cadre pour le penser et le gouverner. La promesse d'un algorithme à la fois performant, transparent et équitable n'existe pas. Ce qui existe, en revanche, est la possibilité de construire patiemment les institutions, les outils et la culture qui permettent de s'en approcher. Cette démarche est moins spectaculaire qu'une solution miracle, mais elle est plus rigoureuse. C'est, en définitive, ce que la recherche académique peut offrir de plus utile.

CONCLUSION GÉNÉRALE

Ce mémoire part d'une question simple, dont l'urgence n'a fait que croître à mesure que l'intelligence artificielle s'est imposée dans le secteur bancaire européen :

« Dans quelle mesure les banques de détail européennes peuvent-elles concilier performance prédictive et conformité réglementaire lors du déploiement de modèles de machine learning pour l'octroi de crédit à la consommation aux particuliers ? »

Pour y répondre, l'analyse s'est déployée en trois parties, en articulant dimensions technique, éthique et réglementaire autour d'un cadre unique : le trilemme performance-conformité-éthique. Nommer ce trilemme, en montrer la structure et les implications, c'est l'une des contributions que ce travail entend apporter.

I. Synthèse des enseignements issus de l'analyse

Premier enseignement (Partie 1) : L'IA transforme profondément le credit scoring, mais cette transformation crée de nouvelles vulnérabilités

La première partie a cartographié la mutation technologique en cours dans le credit scoring européen. Le passage des systèmes experts et de la régression logistique aux algorithmes de machine learning contemporains ne constitue pas un simple changement d'outil : c'est un changement de logique, dans lequel le rapport entre prêteur, emprunteur et risque se trouve fondamentalement reconfiguré. Les gains de performance sont documentés et, à l'échelle de

portefeuilles de plusieurs milliards d'euros, se traduisent par des économies concrètes sur les provisions. L'expansion des sources de données, Open Banking, PSD2 et données alternatives, ouvre par ailleurs des perspectives réelles d'inclusion financière pour les populations thin-file.

Ces avancées s'accompagnent toutefois de vulnérabilités structurelles : modèles plus opaques, biais plus difficiles à isoler, et réduction de la place laissée à l'intervention humaine. Il ne s'agit pas d'effets secondaires accessoires, mais de contreparties inhérentes à la transformation elle-même, analysées tout au long de cette première partie.

Maîtriser cette transformation suppose de satisfaire trois exigences qui ne convergent pas spontanément : la compréhension des modèles, la mesure des effets éthiques de leurs applications, et la connaissance du cadre réglementaire qui leur est applicable. C'est l'articulation de ces trois dimensions qui structure la suite de ce travail.

Deuxième enseignement (Partie 2) : Les biais algorithmiques et l'opacité constituent des risques structurels multidimensionnels

La deuxième partie a établi que les biais, qu'ils soient présents dans les données d'entraînement, de sélection, historiques, de mesure ou par proxy, ne constituent pas des défauts techniques isolés. Ils reflètent des inégalités sociales que les données historiques ont enregistrées et que les modèles reproduisent mécaniquement. L'arrêt Schufa (CJUE, 2023) confirme que ces questions sont désormais portées devant les juridictions européennes et y sont prises au sérieux.

Le théorème d'impossibilité joue un rôle central : aucun système ne peut satisfaire simultanément toutes les définitions concurrentes de l'équité algorithmique. Choisir entre elles ne relève pas d'une décision technique, mais d'un choix normatif, politique et éthique qui ne peut être délégué à un algorithme.

Les conséquences sont loin d'être neutres. Les amendes encourues au titre du RGPD peuvent atteindre 4 % du chiffre d'affaires mondial. À cela s'ajoutent l'invalidation potentielle des modèles IRB, le risque réputationnel et le phénomène d'algorithm aversion. Ces risques, structurels et potentiellement systémiques, pèsent simultanément sur les droits fondamentaux des emprunteurs et sur la position concurrentielle des banques européennes.

Troisième enseignement (Partie 3) : Le cadre normatif constitue une base solide mais insatisfaisante, il faut des stratégies anticipatrices à trois niveaux

La partie 3 a dressé l’inventaire des réponses mobilisables. Le cadre normatif européen est ample. L’article 22 du RGPD, grâce à une interprétation de la C.J.U.E dans l’arrêt SCHUFA (C-634/21, 2023) restreint considérablement les décisions automatisées alors même que la portée réelle des droits qu’il installe reste discutée en doctrine. L’AI Act classe le credit scoring comme un système à haut risque, impose des obligations substantielles aux articles 8 à 15 et impose de s’y conformer au 2 août 2026, sauf évolution en cours dans le cadre du Digital Omnibus on AI. Le cadre prudentiel issu de Bâle III et de sa transposition européenne s'y ajoute. L'Union européenne dispose là d'un arsenal réglementaire sans équivalent à l'échelle mondiale.

Cet arsenal laisse pourtant des zones grises considérables sur les modalités concrètes de mise en œuvre. Se contenter d'une conformité formelle serait une erreur de calcul.

Le chapitre 6 a proposé une architecture opérationnelle en trois niveaux pour y répondre. Au niveau du modèle : des choix algorithmiques adaptés à l'importance des enjeux, des tests de fairness réalisés a priori au déploiement. Au niveau de la gouvernance : des comités pluriels dotés d'un droit de veto, des model cards, un registre IA, une supervision humaine non tournant au théâtre. Au niveau stratégique : une conformité considérée comme un actif, et une Trustworthy AI placée comme un argument de différenciation vis-à-vis des consommateurs, des régulateurs et des investisseurs.

L'arbitrage entre performance, conformité et éthique ne se résout pas. Il se gouverne, par des choix explicites, documentés, et régulièrement remis sur la table.

II. Réponse synthétique à la problématique

Les banques de détail européennes peuvent concilier performance prédictive et conformité réglementaire. Mais pas sans effort, pas une fois pour toutes. Cela suppose un investissement réel dans la gouvernance des modèles, l'IA explicable, le débiaisage algorithmique et la formation des équipes. Cela suppose aussi des arbitrages assumés entre performance et transparence, entre granularité de la prédiction et équité des décisions. Et cela suppose, peut-être surtout, de renoncer à l'idée du modèle parfait. Ce que ce mémoire propose à la place, c'est une gestion dynamique et institutionnalisée du trilemme : imparfaite par construction, mais honnête et durable.

Les modèles gradient boosting, XGBoost ou LightGBM, couplés à des outils d'explicabilité post hoc comme SHAP ou les explications contrefactuelles, et encadrés par une gouvernance solide, représentent aujourd'hui l'un des équilibres les plus praticables entre les trois pôles du trilemme. La condition est double : un déploiement proportionné aux enjeux, et un monitoring qui ne s'arrête pas à la mise en production.

Pour les décisions où la transparence n'est pas négociable, les modèles intrinsèquement interprétables, régression logistique avancée ou Explainable Boosting Machines, restent la référence. Moins performants sur le papier, ils sont plus faciles à auditer, à expliquer et à défendre devant un régulateur.

III. Limites de la recherche et perspectives

Ce mémoire a des limites, et il serait malhonnête de ne pas les nommer. La première est méthodologique : l'approche est entièrement théorique et analytique. Aucune modélisation empirique sur données réelles n'a été conduite, faute d'accès aux données bancaires. C'est une contrainte réelle, pas un choix de confort.

La deuxième est contextuelle. L'AI Act est entré en vigueur en août 2024 avec des normes techniques encore incomplètes et un calendrier susceptible d'évoluer. Ce mémoire photographie un paysage réglementaire en mouvement : certaines analyses devront être révisées à mesure que les textes d'application se précisent.

La troisième est de périmètre. L'analyse se limite au credit scoring à la consommation dans l'Union européenne. Les échanges conduits avec les praticiens du secteur ont apporté un éclairage qualitatif bienvenu, mais leur nombre restreint et les contraintes de confidentialité propres au

secteur bancaire interdisent d'en faire une base empirique solide. Ils éclairent l'analyse théorique. Ils ne la valident pas.

Ces limites indiquent ce qui reste à faire. La piste la plus directe est quantitative : mesurer le trade-off performance-fairness-explicabilité sur des données européennes réelles, avec des portefeuilles et des populations concrètes. C'est faisable, mais cela suppose un accès aux données que la recherche académique obtient rarement sans partenariat institutionnel.

Dès 2026, les autorités nationales suivent la jurisprudence après SCHUFA, l'implémentation effective et significative de l'AI Act. Les premières décisions de supervision révéleront bien des choses, sur ce que les textes veulent dire, vraiment, dans la pratique.

Des études de cas permettraient d'illustrer concrètement les dispositifs de déploiement responsable, là où la littérature fourmille de prescriptions théoriques.

Le problème le plus difficile reste peut-être le paradoxe européen de la fairness sans données démographiques. Comment détecter et corriger des biais potentiels quand la collecte de données sensibles est fortement encadrée, avec des règles qui varient selon les finalités et les États membres ? C'est une question méthodologique ouverte, et sa résolution conditionne en partie la crédibilité de toute l'architecture de gouvernance proposée ici.

IV. Mot de fin

Le machine learning dans le credit scoring n'est pas une fatalité à subir, pas une menace à conjurer, pas une promesse à accueillir les yeux fermés. C'est une transformation qui appelle une réponse à la hauteur : rigoureuse techniquement, lucide sur les risques éthiques, et portée par une ambition réglementaire que l'Union européenne a eu le mérite de formaliser.

Le cadre est posé. Il est imparfait, encore incomplet par endroits, mais il existe, et il prend au sérieux à la fois les droits fondamentaux et la légitimité de l'innovation. Ce qui reste à faire est du côté des banques : transformer ce cadre en pratiques réelles, pas pour cocher des cases, mais parce qu'une IA de crédit qui ne peut pas être expliquée, auditée et contestée n'est pas seulement un risque juridique. C'est une contradiction dans les termes d'une finance qui se prétend responsable.

Ce trilemme entre performance, conformité et éthique ne se résout pas définitivement. Il se gouverne, par des choix explicites et documentés, soumis à une réévaluation régulière. La qualité de ces choix déterminera la capacité des banques de détail européennes à concilier ces trois exigences dans la durée, face à leurs clients, face aux régulateurs, et face à leurs propres standards de responsabilité.

BIBLIOGRAPHIE GÉNÉRALE

I. Ouvrages académiques

Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. https://fairmlbook.org

Molnar, C. (2019). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards* (2nd ed.). John Wiley & Sons.

II. Articles publiés dans des revues académiques, actes de conférences et communications scientifiques

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405-417.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671-732.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30-56.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the Rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*(7), 2845-2897.

Berger, A. N., & Udell, G. F. (2002). Small business credit availability and relationship lending: The importance of bank organisational structure. *The Economic Journal, 112*(477), F32-F53.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine learning, 106*(7), 1039-1082.

Bholat, D., Lastra, R., Markose, S., Miglionico, A., & Sen, K. (2020). Non-performing loans at the dawn of IFRS 9: Regulatory and accounting treatment of asset quality. *Journal of Banking Regulation, 21*, 33-54.

Boot, A., Hoffmann, P., Laeven, L., & Ratnovski, L. (2021). Fintech: what's old, what's new? *Journal of Financial Stability, 53*, 100836.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*, 203-216.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM, 63*(5), 82-89.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1-33.

Degryse, H., & Ongena, S. (2005). Distance, lending relationships, and competition. *The Journal of Finance, 60*(1), 231-266.

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law & Technology Review, 16*(1), 18-84.

Elsas, R., & Krahnen, J. P. (1998). Is relationship lending special? Evidence from credit-file data in Germany. *Journal of Banking & Finance, 22*(10-11), 1283-1316.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268). ACM.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), 5-47.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine, 38*(3), 50-57.

Gunnarsson, B. R., vanden Broucke, S., & Baesens, B. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research, 295*(1), 292-305.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A, 160*(3), 523-541.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (Vol. 29, pp. 3315-3323).

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology, 18*(1), 148-216.

Jünger, M., & Mietzner, M. (2020). Banking goes digital: The adoption of FinTech services by German households. *Finance Research Letters, 34*, 101260.

Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Technology Law Journal, 34*(1), 189-218.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767-2787.

Kozodoi, N., Jacob, J., Lorenz, S., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research, 297*(3), 1083-1094.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124-136.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36-43.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765-4774).

Mester, L. J. (1997). What's the point of credit scoring? *Federal Reserve Bank of Philadelphia Business Review, 3*, 3-16.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165*, 113986.

Prince, A. E., & Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review, 105*(3), 1257-1318.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233-242.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76-99.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 842-887.

III. Working papers et prépublications

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807.

Siegmann, C., & Anderljung, M. (2022). The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market. arXiv:2208.12645.

IV. Chapitres d'ouvrages collectifs

Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In Big Data Analysis: New Algorithms for a New Society (pp. 91-114). Springer.

V. Textes juridiques et réglementaires européens

Union européenne. (2000). Directive 2000/43/CE du Conseil du 29 juin 2000 relative à la mise en œuvre du principe de l'égalité de traitement entre les personnes sans distinction de race ou d'origine ethnique. Journal officiel des Communautés européennes, L 180, 19 juillet 2000.

Union européenne. (2004). Directive 2004/113/CE du Conseil du 13 décembre 2004 mettant en œuvre le principe de l'égalité de traitement entre les femmes et les hommes dans l'accès à des biens et services et la fourniture de biens et services. Journal officiel de l'Union européenne, L 373, 21 décembre 2004.

Union européenne. (2013). Règlement (UE) n° 575/2013 du Parlement européen et du Conseil du 26 juin 2013 concernant les exigences prudentielles applicables aux établissements de crédit et aux entreprises d'investissement et modifiant le règlement (UE) n° 648/2012. Journal officiel de l'Union européenne, L 176, 27 juin 2013.

Union européenne. (2013). Directive 2013/36/UE du Parlement européen et du Conseil du 26 juin 2013 concernant l'accès à l'activité des établissements de crédit et la surveillance prudentielle des établissements de crédit et des entreprises d'investissement. Journal officiel de l'Union européenne, L 176, 27 juin 2013.

Union européenne. (2015). Directive (UE) 2015/2366 du Parlement européen et du Conseil du 25 novembre 2015 concernant les services de paiement dans le marché intérieur, modifiant les directives 2002/65/CE, 2009/110/CE, 2013/36/UE et le règlement (UE) n° 1093/2010, et abrogeant la directive 2007/64/CE. Journal officiel de l'Union européenne, L 337, 23 décembre 2015.

Union européenne. (2016). Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE - Règlement général sur la protection des données. Journal officiel de l'Union européenne, L 119, 4 mai 2016.

Union européenne. (2022). Règlement (UE) 2022/2554 du Parlement européen et du Conseil du 14 décembre 2022 sur la résilience opérationnelle numérique du secteur financier et modifiant les règlements (CE) n° 1060/2009, (UE) n° 648/2012, (UE) n° 600/2014, (UE) n° 909/2014 et (UE) 2016/1011 - Digital Operational Resilience Act, DORA. Journal officiel de l'Union européenne, L 333, 27 décembre 2022.

Union européenne. (2022). Directive (UE) 2022/2555 du Parlement européen et du Conseil du 14 décembre 2022 concernant des mesures destinées à assurer un niveau élevé commun de cybersécurité dans l'ensemble de l'Union, modifiant le règlement (UE) n° 910/2014 et la directive (UE) 2018/1972, et abrogeant la directive (UE) 2016/1148 - Directive NIS2. Journal officiel de l'Union européenne, L 333, 27 décembre 2022.

Union européenne. (2023). Directive (UE) 2023/2225 du Parlement européen et du Conseil du 18 octobre 2023 relative aux contrats de crédit aux consommateurs et abrogeant la directive 2008/48/CE. Journal officiel de l'Union européenne, L 2023/2225, 30 octobre 2023.

Union européenne. (2024). Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant

plusieurs règlements et directives de l'Union - Artificial Intelligence Act. Journal officiel de l'Union européenne, L 2024/1689, 12 juillet 2024.

Union européenne. (2024). Règlement (UE) 2024/1623 du Parlement européen et du Conseil du 31 mai 2024 modifiant le règlement (UE) n° 575/2013 en ce qui concerne les exigences relatives au risque de crédit, au risque d'ajustement de l'évaluation de crédit, au risque opérationnel, au risque de marché et au plancher de fonds propres - CRR3. Journal officiel de l'Union européenne, L 2024/1623, 19 juin 2024.

Union européenne. (2024). Directive (UE) 2024/1619 du Parlement européen et du Conseil du 31 mai 2024 modifiant la directive 2013/36/UE en ce qui concerne les pouvoirs de surveillance, les sanctions, les succursales de pays tiers et les risques environnementaux, sociaux et de gouvernance - CRD6. Journal officiel de l'Union européenne, L 2024/1619, 19 juin 2024.

VI. Jurisprudence

Cour de justice de l'Union européenne. (2023). Arrêt du 7 décembre 2023, SCHUFA Holding AG, C-634/21, ECLI:EU:C:2023:957.

Cour de justice de l'Union européenne. (2025). Arrêt du 27 février 2025, CK c. Dun & Bradstreet Austria GmbH et Magistrat der Stadt Wien, C-203/22, ECLI:EU:C:2025:117.

Tribunal de district de La Haye. (2020). Jugement du 5 février 2020, NJCM c. État des Pays-Bas, affaire SyRI, ECLI:NL:RBDHA:2020:865.

VII. Rapports institutionnels, rapports de supervision et rapports sectoriels

AI HLEG - High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. Commission européenne.

Banque centrale européenne. (2020). *Guide to Internal Models*. European Central Bank.

BNP Paribas. (2024). *Universal Registration Document and Annual Financial Report 2024*. BNP Paribas.

Comité de Bâle sur le contrôle bancaire. (2017). *Basel III: Finalising post-crisis reforms*. Banque des règlements internationaux.

Deutsche Bank. (2024). *Annual Report 2024*. Deutsche Bank AG.

European Banking Authority. (2020). *Report on Big Data and Advanced Analytics*. EBA/REP/2020/01.

European Banking Authority. (2021). *Discussion Paper on Machine learning for IRB Models*. EBA/DP/2021/04.

Evident. (2024). *Evident AI Index: Banking sector rankings and analysis*. Evident Insights.

Financial Stability Board. (2023). *The Financial Stability Implications of Artificial Intelligence*.

ING Group. (2024). *Annual Report 2024*. ING Group.