Juillet 2026

L'équité algorithmique dans le secteur financier

Document de réflexion

AUTEURS
Cyril Chhun, Olivier Fliche, Julien Uri
Direction de l'Innovation, des données et des risques technologiques

ACPR
BANQUE DE FRANCE

Résumé

L'**équité algorithmique** désigne l'ensemble des principes et des méthodes visant à concevoir, évaluer et encadrer des systèmes algorithmiques afin d'éviter qu'ils ne génèrent des inégalités injustifiées entre individus ou entre groupes, notamment lorsque ces inégalités sont liées, directement ou indirectement, à des caractéristiques personnelles dites « sensibles ». Dans le secteur financier, elle renvoie à une question centrale : comment **différencier** les individus en fonction de leur niveau de risque – condition de la soutenabilité des modèles d'affaire – sans, pour autant, aboutir à un **traitement injuste voire discriminatoire**, alors même que les caractéristiques sensibles sont fréquemment **corrélées aux risques observés** ?

Pour répondre à cette difficulté, l'approche longtemps dominante, dite d'« **équité par l'ignorance** », a consisté à exclure les variables sensibles des traitements statistiques. Si cette méthode a toujours été débattue, elle est aujourd'hui **rendue largement obsolète par le développement des modèles d'intelligence artificielle (IA)**, qui sont capables de reconstituer l'information contenue dans ces variables. Plus largement, l'essor de l'IA conduit à renouveler la manière d'appréhender les enjeux d'équité algorithmique, en rendant nécessaire le développement d'approches plus explicites, plus robustes et mieux adaptées à la complexité des systèmes actuels.

Dans ce contexte, le présent document de réflexion expose tout d'abord le **cadre juridique** de l'équité algorithmique dans le secteur financier. Il montre que, si le principe de non-discrimination est solidement établi, sa mise en œuvre est plus complexe lorsque les décisions reposent sur des modèles statistiques, et a fortiori sur des systèmes d'IA. **Le règlement européen sur l'intelligence artificielle** (« **Règlement IA** ») **vient compléter ce cadre en posant des exigences d'équité pour les systèmes d'IA dits « à haut risque »**, et en affirmant un principe de non-discrimination pour l'ensemble des systèmes d'IA. Dans le secteur financier, ce règlement s'articule avec les **règles relatives à la protection de la clientèle** dans les secteurs bancaire et assurantiel, qui traduisent également des exigences d'équité, souvent selon une logique de « protection par l'abstention » (d'octroi ou de vente), là où le Règlement IA met davantage l'accent sur les risques d'exclusion. Enfin, la comparaison avec les cadres règlementaires étrangers met en évidence la diversité des approches.

Le document précise ensuite les **principaux concepts mobilisés**, en distinguant notamment trois niveaux d'analyse – disparité, biais et discrimination – afin de clarifier des notions souvent confondues dans les débats sur l'équité algorithmique. **Il souligne qu'une disparité ne constitue pas nécessairement une discrimination, cette dernière impliquant un jugement normatif et contextuel**. Le document retrace ensuite les principales **sources de biais** et met en évidence le **potentiel d'amplification** de certaines modélisations. Plus largement, le document insiste sur la **tension structurelle entre performance prédictive et non-discrimination**, dans un contexte où les variables pertinentes pour mesurer le risque sont souvent corrélées à des caractéristiques sensibles. Il revient également sur la distinction entre **équité individuelle**, consistant à traiter de manière similaire des individus comparables, et **équité de groupe**, visant à réduire les disparités de traitement entre groupes. Si l'équité individuelle présente un attrait théorique, notamment en termes de performance, ses conditions de mise en œuvre apparaissent particulièrement exigeantes en pratique, ce qui justifie de **privilégier des approches fondées sur l'équité de groupe**.

En matière d'équité de groupe, la littérature scientifique distingue **trois grandes familles de métriques : l'indépendance, la séparation et la suffisance**, qui correspondent chacune à une

conception normative distincte (parité des résultats, parité des erreurs ou parité de la fiabilité des décisions) et conduisent à des **implications pratiques différentes**. La littérature montre par ailleurs que lorsque les niveaux de risque diffèrent entre groupes, il est impossible de satisfaire simultanément ces différentes exigences. Le **choix d'une métrique** d'équité implique donc nécessairement un arbitrage entre objectifs concurrents – équité, performance, inclusion – qu'il convient de rendre explicite.

Le document présente ensuite les **modalités d'évaluation et de correction des biais algorithmiques**. Il examine d'abord la manière de les mesurer, en mettant l'accent sur l'estimation de l'incertitude qui entoure les métriques d'équité, notamment au moyen d'intervalles de confiance. Il recense ensuite les principales **méthodes de correction**, en distinguant celles qui interviennent en amont des modèles (pré-traitement des données), au moment de leur entraînement (modification de la fonction d'optimisation) ou en aval (ajustement des décisions). Ces approches présentent chacune des avantages et des limites, conduisant en pratique à arbitrer entre équité, performance et simplicité de mise en œuvre.

S'agissant des conditions concrètes de **mise en œuvre** de l'équité algorithmique dans le secteur financier, le document souligne que celle-ci ne saurait être réduite à une question technique relevant des seules équipes de modélisation. Elle constitue au contraire un **enjeu transversal**, impliquant des **choix stratégiques et des arbitrages** qui relèvent de la responsabilité globale de l'institution financière. Les enjeux d'équité ont ainsi vocation à être intégrés à **tous les niveaux de décision** – stratégique, métier, technique – et tout au long du cycle de vie des systèmes. Cela suppose de **définir des objectifs explicites, de documenter les choix opérés et de mettre en place des dispositifs de contrôle adaptés**, en cohérence avec les cadres existants de gestion des risques de modèles.

Le document examine ensuite les principaux **choix opérationnels** auxquels sont confrontés les acteurs financiers : identification et usage des variables sensibles, estimation des biais et définition des groupes d'analyse, choix des métriques, détermination des seuils, sélection des méthodes de correction des biais. Il propose des **repères pour éclairer ces décisions**, tout en soulignant qu'elles ne peuvent être entièrement standardisées. Elles doivent au contraire être **adaptées au contexte** d'utilisation, aux données disponibles et aux objectifs poursuivis, selon une **approche proportionnée et fondée sur les risques**.

Enfin, le document aborde rapidement, et de manière plus prospective, le cas de **l'IA générative**, qui connaît un développement rapide dans le secteur financier, même si elle n'est pas aujourd'hui déployée à grande échelle pour des cas d'usage susceptibles de présenter de forts enjeux d'équité algorithmique. Il apparaît que les méthodes d'évaluation des biais conçues pour les modèles prédictifs classiques **ne se transposent pas directement** à ces nouveaux systèmes, dont les modes de fonctionnement et les formes de sortie diffèrent sensiblement. Pour autant, **des méthodes d'évaluation de l'équité d'un système d'IA générative ont déjà émergé**, combinant trois couches (couche représentationnelle, couche comportementale, couche allocative) à mobiliser conjointement.

Table des matières

Introduction

L'**équité algorithmique** désigne l'ensemble des principes et des méthodes visant à concevoir, évaluer et encadrer des systèmes algorithmiques afin d'éviter qu'ils ne génèrent des **inégalités injustifiées** entre individus ou entre groupes, notamment lorsque ces inégalités sont liées, directement ou indirectement, à des caractéristiques personnelles dites « sensibles ».

Dans le secteur financier, la question de l'équité algorithmique, loin d'être nouvelle, s'est posée dès lors que des modèles statistiques ont été utilisés pour éclairer les décisions des acteurs (octroi de crédit, tarification en assurance, etc.). Il s'agit fondamentalement de savoir comment **différencier** les individus en fonction de leur niveau de risque – condition de la soutenabilité des modèles d'affaire – **sans, pour autant, aboutir à un traitement injuste voire discriminatoire**, alors même que les caractéristiques sensibles sont fréquemment **corrélées aux risques observés**.

Pour répondre à cette difficulté, l'approche longtemps dominante, dite d'« **équité par l'ignorance** », a consisté à exclure les variables sensibles des traitements statistiques. Si cette méthode a toujours été débattue, elle est aujourd'hui rendue **largement obsolète** par le développement des modèles fondés sur l'**intelligence artificielle (IA).** En effet, la **grande dimensionnalité** des données s'accompagne d'une **forte colinéarité** entre variables, de sorte que les algorithmes d'IA peuvent identifier des **variables de substitution** leur permettant, en pratique, de reconstituer l'information contenue dans les variables sensibles.

L'essor de l'IA transforme ainsi en profondeur la manière d'envisager les enjeux d'équité. Les modèles à base d'IA sont, d'une part, plus performants, ce qui peut théoriquement contribuer à réduire le risque de traitements injustes ; mais, d'autre part, les mécanismes à l'origine des discriminations deviennent plus diffus, plus difficiles à identifier et à interpréter, en raison de la complexité et de l'opacité des systèmes.

Par ailleurs, le règlement (UE) 2024/1689 sur l'intelligence artificielle (« **Règlement IA** ») pose des **exigences d'équité** pour les systèmes d'IA dits « à haut risque », tout en réaffirmant plus largement, pour tous les systèmes d'IA déployés dans l'Union européenne (UE), le **principe de non-discrimination** consacré par la Charte européenne des droits fondamentaux.

Dans ce contexte, il apparaît nécessaire de renouveler les cadres d'analyse et les outils opérationnels permettant de traiter la question de l'équité dans le secteur financier.

Pour éclairer ces enjeux, l'Autorité de contrôle prudentiel et de résolution (ACPR) a conduit, au printemps et à l'automne 2025, une **série d'ateliers techniques** avec des acteurs financiers volontaires. Ceux-ci avaient pour objectif de comprendre comment les banques et les assurances interrogées traitaient concrètement les questions d'équité dans leurs processus et dans leur gouvernance. Ces échanges ont notamment mis en évidence de **fortes attentes** des établissements vis-à-vis des autorités publiques – et en particulier des superviseurs financiers – pour clarifier et préciser la mise en œuvre des règles applicables en matière d'équité.

Le présent document vise à proposer un **cadre d'analyse structuré des enjeux d'équité algorithmique dans le secteur financier**, en articulant les apports de la littérature scientifique, les exigences juridiques en vigueur et les pratiques observées sur le terrain. Il a également pour objectif de fournir des **repères opérationnels** afin d'aider les acteurs à appréhender et à mettre en œuvre, de manière concrète, les exigences d'équité dans leurs processus.

Le présent document est structuré en **six parties**. La **première** présente le **cadre juridique** applicable à l'équité dans le secteur financier, en articulant les règles de non-discrimination, les règles sectorielles et les exigences introduites par le Règlement IA. La **deuxième partie** propose une **clarification des notions de biais et d'équité**, en distinguant notamment l'équité de groupe et l'équité individuelle, ainsi que leurs fondements conceptuels. La **troisième partie** détaille les **principales familles de métriques d'équité** de groupe issues de la littérature scientifique – indépendance, séparation et suffisance – en analysant leurs propriétés, leurs avantages et leurs limites. La **quatrième partie** recense les **principales modalités d'évaluation et de correction des biais**. La **cinquième partie** aborde les **enjeux de mise en œuvre pratique**, en proposant des repères méthodologiques pour la gouvernance de l'équité dans les établissements, ou encore pour le choix des métriques pertinentes ou la définition des seuils appropriés.

Ce document de réflexion porte principalement sur les systèmes prédictifs « traditionnels ». En effet, ceux-ci représentent aujourd'hui l'essentiel des systèmes déployés à grande échelle dans le secteur financier susceptibles de présenter des risques en matière d'équité. Le cas de l'**IA générative**, qui connaît un développement rapide, est néanmoins abordé en fin de document, dans une **sixième partie**. Il apparaît en effet que les méthodes d'évaluation des biais conçues pour les modèles prédictifs classiques ne se transposent pas directement à ces nouveaux systèmes, dont les modes de fonctionnement et les formes de sortie diffèrent sensiblement.

Ce document de réflexion a été rédigé par le Service de surveillance des risques technologiques de l'ACPR, à partir d'une revue de la littérature scientifique ainsi que des entretiens mentionnés ci-dessus. **Il n'a pas vocation à donner une vision exhaustive de l'ensemble des sujets liés à l'équité algorithmique, ni à exprimer une position officielle de l'ACPR**. Son objectif est de développer de premières analyses sur les modalités de mise en œuvre des exigences d'équité algorithmique, en vue de les discuter avec les parties prenantes, notamment la profession, à l'occasion d'une consultation publique.

Les auteurs tiennent à exprimer leurs remerciements aux relecteurs de ce document pour leurs remarques précieuses, et en particulier à Jean-Michel Loubès (Inria), Félicien Vallet et Maxence Gérard (Cnil), Samy Chali et Shaden Shabayek (PEReN).

1 Quelles exigences d'équité algorithmique pour les entreprises du secteur financier ?

1.1 Principe de non-discrimination et données sensibles en droit

1.1.1 Discrimination et différenciation

Le principe de non-discrimination est largement consacré dans les différents ordres juridiques. Il découle d'abord du **principe d'égalité**. En droit français, l'article premier de la Constitution du 4 octobre 1958 dispose ainsi que la République « *assure l'égalité devant la loi de tous les citoyens sans distinction d'origine, de race ou de religion. Elle respecte toutes les croyances* »[1]. Ce principe est notamment précisé par l'article 225-1 du code pénal, qui dispose que « *constitue une discrimination toute distinction opérée entre les personnes physiques [et morales]* » sur le fondement d'un certain nombre de **critères protégés**[2].

En droit européen, l'article 20 de la **Charte des droits fondamentaux de l'Union européenne**, devenue juridiquement contraignante depuis l'entrée en vigueur du Traité de Lisbonne le 1er décembre 2009, dispose que « *toutes les personnes sont égales en droit*. » L'article 21 de cette même charte dispose qu'« *est interdite toute discrimination fondée notamment sur le sexe, la race, la couleur, les origines ethniques ou sociales, les caractéristiques génétiques, la langue, la religion ou les convictions, les opinions politiques ou toute autre opinion, l'appartenance à une minorité nationale, la fortune, la naissance, un handicap, l'âge ou l'orientation sexuelle* ». La non-discrimination est également protégée par la **Convention européenne des droits de l'homme**, en particulier son article 14[3].

Si la liste des critères protégés peut varier légèrement d'un texte à l'autre, ces différentes sources convergent vers une même exigence : interdire les discriminations fondées sur des caractéristiques personnelles protégées. Dans le secteur financier, cette interdiction vise tout particulièrement la segmentation abusive de la clientèle (*cf.* section 2), le refus automatique d'entrer en relation, ou encore l'application de conditions dégradées fondées sur des critères personnels sans lien direct avec le risque réel.

En revanche, le **principe de non-discrimination n'interdit pas toute différence de traitement entre individus**, dès lors que celle-ci repose sur des **critères objectifs et pertinents**, notamment économiques, en lien avec la nature de la décision considérée. Le principe d'égalité peut en effet conduire à **traiter différemment des situations objectivement distinctes** afin d'assurer une égalité effective. Cette approche, consacrée de manière constante par la jurisprudence européenne, implique une appréciation au cas par cas des situations[4]. Ainsi, une

---

[1] En outre, le Préambule de la Constitution de 1946, consacré par le Conseil constitutionnel comme un texte à « valeur constitutionnelle », précise dans son premier alinéa que « *tout être humain, sans distinction de race, de religion ni de croyance, possède des droits inaliénables et sacrés* ». En outre, « *la loi garantit à la femme, dans tous les domaines, des droits égaux à ceux de l'homme* » (alinéa 3).

[2] L'article 225-1 du code pénal liste 26 critères protégés en France.

[3] Article 14 de la Convention : « *[l]a jouissance des droits et libertés reconnus dans la présente Convention doit être assurée, sans distinction aucune, fondée notamment sur le sexe, la race, la couleur, la langue, la religion, les opinions politiques ou toutes autres opinions, l'origine nationale ou sociale, l'appartenance à une minorité nationale, la fortune, la naissance ou toute autre situation* ».

[4] CEDH, 13 novembre 2007, D.H. et autres c. République tchèque, n° 57325/00, § 175 : « *Selon la jurisprudence établie de la Cour, la discrimination consiste à traiter de manière différente, sauf justification*

différenciation tarifaire ou une décision de refus (de crédit, d'assurance…) peut être licite lorsqu'elle repose sur des éléments économiques démontrables (revenus, stabilité financière, taux d'endettement, comportements à risque etc.), mais devient illicite si elle repose directement ou indirectement sur des caractéristiques protégées.

Encadré 1 : Équité et concurrence en assurance

Le principe de l'assurance consiste à **mettre en commun la variabilité du risque** : la contribution de chacun permet la compensation des accidents affectant les plus malchanceux. Sur un marché donné, l'assureur doit donc fixer un niveau total de primes permettant de couvrir le risque. Dans un contexte de **monopole** assurantiel, il suffirait donc d'appliquer un tarif unique à l'ensemble de la population (égal au niveau de primes requis divisé par le nombre d'assurés).

Cependant, dans une situation de **concurrence** entre assureurs, la **segmentation de la clientèle** – c'est-à-dire le regroupement des assurés en des classes homogènes de risque – peut permettre à un assureur donné de proposer des tarifs plus avantageux à un groupe moins risqué que la moyenne, s'appropriant ainsi une part de marché plus importante. C'est donc d'abord la concurrence entre assureurs qui les conduit à proposer des tarifs différenciés à chaque groupe d'assurés.

Dès lors, il est crucial pour chaque assureur de **segmenter de manière optimale** les différents groupes en fonction de leurs niveaux de risque : dans ce but, le travail de l'actuaire consiste notamment à **sélectionner les variables les plus pertinentes**.

1.1.2 Le problème de la discrimination algorithmique

Si ce cadre juridique apparaît clairement défini dans son principe, **sa mise en œuvre soulève en pratique des difficultés**, du fait de la complexité des mécanismes par lesquels des discriminations peuvent se produire.

Ces difficultés tiennent notamment au fait que des mécanismes discriminatoires peuvent se produire en l'absence de toute intention. Les textes européens distinguent ainsi la *discrimination directe*, fondée explicitement sur un critère protégé, de la *discrimination indirecte*, qui résulte d'une disposition ou d'une pratique apparemment neutre, mais susceptible d'entraîner un désavantage particulier pour certaines personnes. Cette distinction met en évidence que la discrimination peut résulter d'effets produits par certaines règles ou pratiques, indépendamment de toute volonté discriminatoire.

Ce point est particulièrement important dans le cas du **recours à des systèmes algorithmiques**. La **discrimination algorithmique** désigne les situations dans lesquelles un système automatisé produit des différences de traitement entre individus ou groupes, en raison des données utilisées, des choix de modélisation ou des variables retenues, sans qu'une intention discriminatoire explicite ne soit nécessairement identifiable[5]. Par rapport à la discrimination

objective et raisonnable, des personnes placées dans des situations comparables (Willis c. Royaume-Uni, no 36042/97, § 48, CEDH 2002-IV ; Okpisz c. Allemagne, no 59140/00, § 33, 25 octobre 2005) »

[5] Défenseur des Droits, Lutter contre les discriminations produites par les algorithmes et l'IA, février 2024 : « *La mécanique discriminatoire peut reposer sur le caractère biaisé des données sélectionnées et utilisées par l'algorithme. Ce caractère biaisé peut être lié à un manque de représentativité des données par rapport au contexte dans lequel l'algorithme va être déployé. Il peut aussi être lié au fait que les données sont la*

entendue au sens juridique classique, elle se caractérise par des **mécanismes plus indirects et diffus**, susceptibles d'intervenir à toutes les étapes de conception et d'utilisation des systèmes, et pouvant conduire à reproduire, voire amplifier, des discriminations préexistantes. Ces discriminations peuvent en outre revêtir une dimension **intersectionnelle**, en combinant plusieurs critères protégés, ce qui complique encore leur identification[6].

Le **nœud du problème** de la discrimination par les modèles tient au fait que les **variables protégées** (comme l'origine ethnique, le sexe, l'âge ou le lieu de résidence) **sont souvent corrélées au risque financier, non en raison d'un lien causal direct, mais du fait de mécanismes indirects liés à des inégalités sociales et économiques préexistantes**. Par exemple, des inégalités structurelles en matière d'éducation, d'emploi ou d'accès au crédit peuvent conduire certains groupes à présenter des profils de risque statistiquement différents, que le modèle apprend et exploite. Même lorsque les variables protégées sont explicitement exclues, des variables de substitution (ou *proxies*[7] : revenu, type de contrat, historique bancaire, zone géographique, etc.) peuvent suffire à reconstruire l'information sensible et à reproduire des effets discriminatoires. Le modèle se retrouve ainsi face à une **tension fondamentale** : ignorer ces corrélations peut dégrader la performance prédictive, mais les intégrer revient à institutionnaliser des désavantages historiques, en traitant comme « neutres » des signaux qui reflètent en réalité des inégalités sociales plutôt qu'un risque intrinsèque.

Ces difficultés sont particulièrement marquées dans le cadre du recours à des systèmes d'IA, en raison de leur **complexité** et de leur **opacité**. Elles se traduisent notamment par des enjeux accrus en matière **d'administration de la preuve**, la démonstration d'une discrimination pouvant s'avérer difficile en l'absence d'accès des personnes concernées aux données ou aux modèles, voire en raison du manque de transparence sur le principe même du recours à un traitement algorithmique. **La discrimination algorithmique ne fait toutefois pas l'objet d'un encadrement juridique spécifique**, les règles générales en la matière ayant vocation à s'appliquer indépendamment de la technologie utilisée.

1.1.3 Critères protégés et variables sensibles

En revanche, **l'utilisation des données personnelles** par les systèmes automatisés relève d'un cadre juridique spécifique. Le **Règlement général sur la protection des données** (RGPD)[8], constitue le texte de référence en la matière ; il encadre la collecte, l'utilisation et la protection des données personnelles afin de renforcer les droits des personnes et la responsabilité des organisations qui les traitent. Il repose sur plusieurs principes fondamentaux, tels que la licéité et la transparence des traitements, la minimisation des données, la limitation des finalités, la sécurité et la responsabilisation des acteurs.

Le RGPD distingue notamment des « catégories particulières de données à caractère personnel » – communément appelées « **données sensibles** » – qui font l'objet d'une **protection renforcée**.

---

traduction mathématique de pratiques et comportements passés souvent discriminatoires et des discriminations systémiques opérant au sein de la société. »

[6] Les discriminations intersectionnelles échappent aux modes de détection traditionnels, dans la mesure où elles résultent de la combinaison de plusieurs critères protégés et ne sont pas nécessairement visibles lorsqu'ils sont analysés séparément.

[7] Dans une modélisation, une variable « *proxy* » sert de substitut à une autre variable, généralement plus difficile à collecter ou à mesurer.

[8] Règlement (UE) 2016/679 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.

Il s'agit notamment des données révélant l'origine, les opinions politiques, les convictions religieuses ou philosophiques, l'appartenance syndicale, l'orientation sexuelle, ou encore l'état de santé (*cf.* tableau 1 ci-dessous pour la liste exhaustive). **Le traitement de ces données est, par principe, interdit,** sauf dans un nombre limité de cas strictement encadrés par le règlement, tels que le consentement explicite de la personne concernée ou le respect d'obligations légales spécifiques.

Les critères protégés en matière de discrimination et les données sensibles au sens du RGPD ne se recoupent que partiellement. Certaines données relèvent des deux catégories, comme l'origine ou les convictions religieuses. D'autres, en revanche, sont protégées par le droit de la non-discrimination sans être qualifiées de sensibles par le RGPD : c'est notamment le cas de l'âge ou du sexe, qui restent des données à caractère personnel « ordinaires ». À l'inverse, le RGPD peut limiter le traitement de certaines données sans lien direct avec un risque de discrimination, mais plutôt en raison de leur caractère particulièrement intrusif pour la vie privée. À titre illustratif, le tableau 1 ci-dessous dresse un inventaire de la situation en droit européen.

Tableau 1 : Critères protégés et données sensibles dans le droit européen

Catégorie	**Charte européenne des droits fondamentaux (Critères protégés)**	**RGPD (Données sensibles)**	**Commentaire**
Origine ethnique	✓	✓	Recouvrement direct
Religion / convictions	✓	✓	Recouvrement direct
Opinions politiques	✓	✓	Recouvrement direct
Orientation sexuelle	✓	✓	Recouvrement direct
Santé / handicap	✓	✓	Recouvrement partiel[9]
Données génétiques	✓	✓	Recouvrement direct
Données biométriques	✗	✓	Protection RGPD seule
Appartenance syndicale	✗	✓	Protection RGPD seule
Sexe	✓	✗	Protection Charte seule
Âge	✓	✗	Protection Charte seule
Nationalité	✓	✗	Protection Charte seule
Origine sociale / fortune	✓	✗	Protection Charte seule
Langue	✓	✗	Protection Charte seule

[9] Handicap dans la Charte, santé dans le RGPD. L'état de santé fait cependant partie des critères protégés en droit français.

De fait, le cadre juridique de la protection des droits fondamentaux – pour sa partie anti-discrimination – et celui de la protection des données personnelles poursuivent des **objectifs complémentaires mais distincts**, ce qui se traduit par des approches différentes en pratique. Le droit de la non-discrimination s'intéresse avant tout aux **effets** des décisions (égalité de traitement entre groupes), tandis que le RGPD encadre les **moyens** (collecte et usage des données). Pour les systèmes d'IA du secteur financier, cette articulation impose de concilier la maîtrise des données utilisées pour l'entraînement avec une vigilance accrue sur les résultats produits.

Il convient enfin de noter que ce tableau **ne traite pas le cas des variables de substitution** (*proxy*), qui, sans être formellement des critères protégés, peuvent être **fortement corrélées** à ceux-ci. Par exemple, le code postal ou le lieu de résidence ne sont pas des critères protégés ou des données sensibles, mais ils peuvent être fortement corrélés avec le niveau de revenu, l'origine ethnique ou sociale ou encore les orientations politiques. L'utilisation de telles variables peut donc soulever des enjeux juridiques en matière de non-discrimination, dès lors qu'elles conduisent, même de manière **indirecte**, à des différences de traitement fondées sur des caractéristiques protégées. Ainsi, en pratique, l'analyse du risque juridique ne **peut pas se limiter** à la seule nature apparente des variables utilisées, mais doit également intégrer leur potentiel effet de substitution.

1.2 Le Règlement IA : quelles obligations en matière d'équité ?

Le Règlement européen sur l'IA[10], entré en vigueur en août 2024, introduit un cadre règlementaire uniforme pour l'IA[11], avec le double objectif de protéger la santé, la sécurité et les droits fondamentaux des citoyens et de favoriser le développement d'un marché unique européen de l'« IA de confiance ». Le Règlement IA classe les systèmes d'IA en fonction de leurs risques ; le cœur de son dispositif concerne les systèmes dits à « haut risque », définis pour l'essentiel par leur finalité, et qui concernent le secteur financier pour deux cas d'usage[12].

S'agissant de l'équité[13], le Règlement IA réaffirme, en premier lieu, un principe simple : le **droit à la non-discrimination** fait partie des droits fondamentaux protégés par le droit européen ; **il doit donc également être respecté par tous les systèmes d'IA déployés dans l'UE**[14]. De fait, le droit à la non-discrimination est l'un des arguments employés pour interdire les systèmes de manipulation (considérant 28) ou de notation sociale (considérant 31). Le risque de

[10] Règlement (UE) 2024/1689 du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle. Ce règlement est entré en vigueur le 1er août 2024.

[11] En particulier, le Règlement IA distingue les fournisseurs et les déployeurs de systèmes d'IA : le fournisseur d'un système d'IA est l'entité qui l'a développé et mis sur le marché ou en service, et le déployeur de ce système d'IA est une entité qui l'utilise pour une activité professionnelle. La majorité des exigences du Règlement IA s'appliquent aux fournisseurs.

[12] Annexe III, cas (5b) *« les systèmes d'IA destinés à être utilisés pour évaluer la solvabilité des personnes physiques ou pour établir leur note de crédit, à l'exception des systèmes d'IA utilisés à des fins de détection de fraudes financières »* ; (5c) les *« systèmes d'IA destinés à être utilisés pour l'évaluation des risques et la tarification en ce qui concerne les personnes physiques en matière d'assurance-vie et d'assurance maladie »*.

[13] Le terme « équité » ne figure que marginalement dans le Règlement IA, qui utilise plus fréquemment les termes « biais » et « discrimination ».

[14] Voir par exemple le considérant 27 du règlement : « *Les systèmes d'IA sont développés et utilisés de manière à inclure des acteurs divers et à promouvoir l'égalité d'accès, l'égalité de genre et la diversité culturelle, tout en évitant les effets discriminatoires et les biais injustes, qui sont interdits par le droit de l'Union ou le droit national*. «

discrimination est également invoqué pour expliquer le classement à haut risque des systèmes d'IA pour l'évaluation de la solvabilité des personnes physiques[15], ainsi que dans le cas de ceux destinés à l'évaluation des risques et à la tarification en assurance santé et assurance-vie[16].

Toutefois, les dispositions tenant à l'équité algorithmique ne sont réellement précisées que pour les systèmes à haut risque[17]. La disposition centrale du Règlement IA sur la question des biais et des discriminations figure dans l'article 10, portant sur les exigences relatives aux données et à leur gouvernance[18]. Ainsi, **l'article 10(2)(f)** dispose que *« [l]es jeux de données d'entraînement, de validation et de test »* des systèmes à haut risque sont soumis à *« un examen permettant de repérer d'éventuels biais qui sont susceptibles de porter atteinte à la santé et à la sécurité des personnes, d'avoir une incidence négative sur les droits fondamentaux ou de se traduire par une discrimination interdite par le droit de l'Union, en particulier lorsque les données de sortie influencent les entrées pour les opérations futures »*. En outre, **l'article 10(2)(g)** ajoute l'exigence de *« mesures appropriées visant à détecter, prévenir et atténuer les éventuels biais repérés conformément au point f) »*.

Ces deux dispositions formalisent donc une exigence d'équité pour les systèmes à haut risque du Règlement IA, couvrant l'ensemble de la chaine de traitement des biais pouvant conduire à des discriminations (prévention, détection et atténuation)[19]. En outre, même si la formulation de l'article 10(2) semble ne viser que les données d'entrée des modèles (jeux d'entraînement, de test et de validation), ses exigences ne peuvent en réalité se rapporter qu'aux sorties des systèmes d'IA (ou aux décisions qui en découlent), puisqu'elles concernent les effets potentiels de leur utilisation[20].

---

[15] « *Les systèmes d'IA utilisés à ces fins peuvent conduire à la discrimination entre personnes ou groupes et perpétuer des schémas historiques de discrimination, tels que ceux fondés sur les origines raciales ou ethniques, le sexe, les handicaps, l'âge ou l'orientation sexuelle, ou peuvent créer de nouvelles formes d'incidences discriminatoires* » (Règlement IA, considérant 58)

[16] « *Par ailleurs, les systèmes d'IA destinés à être utilisés pour l'évaluation des risques et la tarification en ce qui concerne les personnes physiques en matière d'assurance-santé et vie peuvent avoir une incidence significative sur les moyens de subsistance de ces personnes et, s'ils ne sont pas dûment conçus, développés et utilisés, peuvent porter atteinte à leurs droits fondamentaux et entraîner de graves conséquences pour leur vie et leur santé, y compris l'exclusion financière et la discrimination* » (Règlement IA, considérant 58).

[17] On notera enfin que la question des biais est également mentionnée dans le cas des modèles d'IA à usage général : la documentation technique rédigée par les fournisseurs de ces modèles doit en effet comprendre « […] *toutes les autres mesures visant à détecter l'inadéquation des sources de données et les méthodes permettant de détecter les biais identifiables* » (Annexe XI du règlement). Elle vise au minimum à donner les moyens aux parties prenantes situées plus en aval de la chaîne de valeur – comme les éventuels fournisseurs de systèmes à haut risque bâtis sur des modèles à usage général – de respecter leurs obligations règlementaires.

[18] Il est à noter que des travaux de normalisation sont actuellement engagés au niveau européen par le CEN et le CENELEC, qui développent, à la demande de la Commission européenne, des standards harmonisés devant préciser les exigences du Règlement IA, y compris en matière d'équité algorithmique. Ces travaux ont cependant pris du retard par rapport au calendrier initialement prévu. En outre, s'agissant de normes trans-sectorielles, elles ne pourront vraisemblablement pas répondre aux questions spécifiques du secteur financier.

[19] L'identification d'éventuels effets discriminatoires doit également figurer dans la documentation technique des systèmes à haut risque, détaillée à l'Annexe IV du règlement.

[20] En outre, les sorties des modèles sont intégrées à l'examen requis, puisqu'elles peuvent constituer la base d'un réentraînement ultérieur du modèle (article 10(2)(f)). Ce point est confirmé par l'article 15(4), qui dispose que « *[l]es systèmes d'IA à haut risque qui continuent leur apprentissage après leur mise sur le marché ou leur mise en service sont développés de manière à éliminer ou à réduire dans la mesure du*

Enfin, le principe de **contrôle humain** prévu par le règlement IA constitue théoriquement un autre élément de lutte contre les discriminations algorithmiques, en exigeant que les utilisateurs soient en mesure de détecter des résultats anormaux ou injustifiés et, le cas échéant, **d'intervenir pour corriger, suspendre ou invalider une décision automatisée**. Le contrôle humain devrait ainsi jouer un rôle de garde-fou, là où une décision purement automatisée pourrait porter atteinte au principe d'égalité de traitement.

En pratique, le contrôle humain risque toutefois d'être insuffisant pour prévenir pleinement les discriminations algorithmiques : d'une part, le **biais d'automatisation** conduit souvent les opérateurs humains à accorder une confiance excessive aux recommandations émises par les systèmes d'IA, même lorsqu'elles sont erronées ou discutables, surtout si le système est perçu comme techniquement complexe ou objectivement supérieur[21]. Le contrôle humain peut alors devenir purement formel, sans remise en cause réelle des décisions produites par la machine. D'autre part, les **décideurs humains ne sont pas nécessairement neutres** : ils sont eux-mêmes porteurs de biais cognitifs, sociaux ou culturels qui peuvent être égaux, voire supérieurs, à ceux des algorithmes, et susceptibles d'influencer leurs arbitrages.

1.3 Autres sources règlementaires

1.3.1 Secteur bancaire

Dans le secteur bancaire, la **règlementation prudentielle** – en particulier le règlement CRR[22] ou encore les orientations de l'Autorité bancaire européenne (ABE) sur l'octroi et le suivi des prêts[23] – aborde la question des biais statistiques mais pas celle des biais potentiellement discriminatoires. De fait, l'optique de la règlementation prudentielle est de préserver la stabilité du système financier ; elle se concentre donc sur les **risques pour les établissements** et non sur ceux pouvant affecter les clients.

En revanche, les règles européennes relatives à la **protection de la clientèle** dans le secteur bancaire, en particulier les directives sur le crédit immobilier (directive dite « MCD »[24]) et sur le crédit à la consommation (directive dite « CCD »[25]), contiennent des dispositions relatives à l'équité. Elles sont applicables à **l'ensemble des modèles de crédit aux particuliers**, ce qui

---

possible le risque que des sorties éventuellement biaisées n'influencent les entrées pour les opérations futures (boucles de rétroaction) et à veiller à ce que ces boucles de rétroaction fassent l'objet d'un traitement adéquat au moyen de mesures d'atténuation appropriées ».

[21] Pour lutter contre ce phénomène, l'article 14(4)(b) du Règlement IA dispose que les systèmes d'IA doivent être fournis de telle manière que les contrôleurs humains ont la possibilité de prendre conscience de l'existence d'un tel biais.

[22] L'article 174 du règlement (UE) 575/2013 concernant les exigences prudentielles applicables aux établissements de crédit (CRR) dispose ainsi que *« [l]orsqu'un établissement utilise un modèle statistique ou une autre méthode mécanique pour affecter ses expositions aux différents échelons ou catégories de débiteurs ou de facilités de crédit, les conditions suivantes doivent être remplies : a) le modèle a un solide pouvoir prédictif et son utilisation n'entraîne pas de distorsion des exigences de fonds propres. Les variables d'entrée du modèle forment une base cohérente et efficace de prédiction. En outre, le modèle ne pâtit pas de biais significatifs ; […] c) les données utilisées pour construire le modèle sont représentatives de la population effective de ses débiteurs ou expositions ».*

[23] Voir notamment les paragraphes 53(e), 54(a) et 55(a).

[24] Directive 2014/17/UE du 4 février 2014 sur les contrats de crédit aux consommateurs relatifs aux biens immobiliers à usage résidentiel.

[25] Directive (UE) 2023/2225 du 18 octobre 2023 relative aux contrats de crédit aux consommateurs.

correspond à un ensemble de cas d'usages plus larges que celui des systèmes à haut risque du Règlement IA.

Ces garanties prennent **deux formes**. En premier lieu, les professionnels (prêteurs et intermédiaires de crédit) doivent agir de manière **honnête, équitable, transparente et professionnelle**, en tenant compte des droits et des intérêts des consommateurs, et ce durant l'ensemble du cycle de vie du crédit (de la conception des produits à l'exécution du contrat)[26]. En second lieu, **l'évaluation de la solvabilité** des emprunteurs fait l'objet d'un encadrement spécifique. Cette évaluation doit ainsi reposer sur des **informations pertinentes, exactes, nécessaires et proportionnées** aux caractéristiques du crédit, portant principalement sur les revenus, les dépenses et la situation financière du consommateur[27]. Au contraire, le recours aux **réseaux sociaux** comme source d'information est explicitement interdit.

Il convient de noter que, dans le cadre de la réglementation relative à la protection de la clientèle, les établissements bancaires procèdent à l'évaluation de la solvabilité avant tout **dans l'intérêt du consommateur**[28]. Dans cette optique c'est donc l'octroi d'un crédit, et non son refus, qui constitue le principal risque pour le consommateur, celui-ci risquant de ne pas pouvoir le rembourser, voire d'être confronté à une situation de surendettement. Cette logique de « **protection par l'abstention** », qui entend prévenir l'endettement excessif en limitant l'accès au crédit, se distingue sensiblement de la **logique d'accès** du Règlement IA. Ce dernier met en effet l'accent sur le risque inverse de **perte de chance**, en cherchant à éviter que des emprunteurs pourtant solvables soient indûment exclus de l'accès au crédit.

1.3.2 Secteur assurantiel

La règlementation européenne du secteur de l'assurance présente des caractéristiques proches. Là encore, ce sont les règles relatives à la protection de la clientèle, en particulier la **directive sur la distribution d'assurance** (DDA)[29], qui fixent des exigences en matière d'équité. Là encore, ces exigences sont applicables à un périmètre de systèmes d'IA plus large que ceux classés à haut risque par le Règlement IA.

Ainsi, la DDA impose à tous les distributeurs d'assurance d'agir de manière **honnête, impartiale et professionnelle, en accord avec le meilleur intérêt du client**[30]. La DDA introduit également des obligations de **gouvernance des produits**, afin d'assurer que les produits commercialisés répondent aux besoins et caractéristiques des clients auxquels ils sont destinés, et ainsi de réduire les risques de mauvaise commercialisation. En particulier, les assureurs doivent identifier un **marché cible** pour chacun de leurs produits, et donc à effectuer une segmentation de la clientèle, afin de réduire les risques de vente inadaptée ou abusive.

Comme dans le secteur bancaire, la règlementation relative à la protection de la clientèle en assurance entend donc d'abord protéger les clients par l'abstention (de vente), contrairement à la logique d'accès du Règlement IA. Toutefois, cette opposition doit être **nuancée**, car **certaines assurances revêtent un caractère obligatoire,** par exemple l'assurance automobile ou

[26] Article 32 de la directive CCD et article 7(1) de la directive MCD.
[27] Articles 18(1) et 18(3) de la directive CCD et article 20(1) de la directive MCD.
[28] La règlementation prudentielle exige par ailleurs une évaluation de solvabilité pour la protection de la solvabilité de l'établissement.
[29] Directive (UE) 2016/97 du 20 janvier 2016 sur la distribution d'assurance.
[30] Article 17(1) de la DDA.

l'assurance habitation[31]. Il en découle, en droit français, la mise en place de **dispositifs spécifiques** visant à garantir un droit effectif d'accès à l'assurance, tels que les mécanismes de tarification encadrée ou l'intervention des bureaux centraux de tarification, qui permettent de limiter les refus d'assurance.

Il convient enfin de noter que l'Autorité européenne des assurances et des pensions professionnelles (AEAPP) a publié en 2025 une *Opinion sur la gouvernance et la gestion des risques liés à l'IA*[32]. Ce texte, bien que non contraignant d'un point de vue juridique, recommande notamment aux organismes d'assurance ; (i) d'identifier et, si possible, d'éliminer ou en tout cas d'atténuer les biais potentiels, y compris les variables pouvant constituer des *proxies* discriminatoires ; (ii) de mettre en place un suivi régulier des systèmes d'IA utilisés, notamment en **recourant à des métriques d'équité et de non-discrimination**[33] ; (iii) de développer des orientations internes et des formations en matière d'équité à destination de leur personnel.

Encadré 2 : Le cas de la tarification fondée sur le genre en assurance

L'arrêt *Test-Achats* (2011) de la Cour de justice de l'Union européenne[34] a eu pour conséquence **d'interdire**, à compter de décembre 2012, **toute tarification fondée sur le genre dans les assurances, que ce soit pour la détermination des primes ou du niveau des prestations**. La Cour a en effet jugé que les différences tarifaires entre hommes et femmes portaient atteinte au principe d'égalité protégé par le droit de l'Union, même lorsqu'elles reposaient sur des données statistiques (dans le domaine automobile, par exemple, les femmes ont statistiquement moins d'accidents que les hommes), dès lors qu'elles se fondaient sur un critère personnel protégé. À compter de décembre 2012, les assureurs ont ainsi été contraints de revoir les modèles actuariels afin de supprimer toute segmentation tarifaire fondée sur le genre, ce qui a entraîné une redistribution des coûts entre assurés.

Cette interdiction s'étend théoriquement aux variables servant de substitut (*proxies*) au genre, qui doivent être supprimées des modélisations, sauf lorsque leur utilisation est justifiée par un objectif légitime et qu'elle est appropriée et nécessaire. La Commission européenne illustre cette situation par les exemples suivants : la différenciation tarifaire fondée sur la cylindrée du moteur d'un véhicule en matière d'assurance automobile demeure possible, même s'il est statistiquement établi que les hommes conduisent plus souvent des véhicules à moteur plus puissant, car la puissance du moteur est **directement corrélée au risque**. En revanche, il est interdit de pratiquer une différenciation tarifaire fondée sur la taille d'une personne en assurance automobile, les hommes étant généralement plus grands que les femmes, et cette différence étant **sans rapport objectif avec le risque**[35].

[31] Le droit au compte participe de la même logique dans le secteur bancaire. En France, toute personne dépourvue de compte bancaire a en effet le droit d'en obtenir un auprès d'un établissement désigné d'office par la Banque de France. Ce dispositif vise à garantir l'inclusion financière en assurant à chacun l'accès aux services bancaires essentiels.

[32] AEAPP, Opinion on AI Governance and Risk Management, 6 août 2025.

[33] L'annexe I du document fournit en outre plusieurs exemples de métriques d'équité.

[34] Cour de justice de l'Union européenne (CJUE), Association Belge des Consommateurs Test-Achats ASBL et autres contre Conseil des ministres, C-236/09, 1er mars 2011.

[35] Orientations de la Commission européenne relatives à l'application de la directive 2004/113/CE du Conseil en matière d'assurance à la lumière de l'arrêt de la Cour de justice de l'Union européenne dans l'affaire C-236/09 (*Test-Achats*).

1.4 Politique de gestion des risques et politique interne des acteurs financiers

La prise en compte de l'équité au sein des entreprises du secteur financier s'opère à travers deux niveaux complémentaires : d'une part, **des politiques de conformité**, qui déclinent en interne les obligations juridiques applicables, et, d'autre part, **des cadres volontaires additionnels**, tels que des chartes éthiques ou des principes de gouvernance de l'IA, qui traduisent les arbitrages opérés par les entreprises entre performance, rentabilité et équité.

Ces arbitrages prennent **une acuité particulière dans le secteur de l'assurance**, où la logique de segmentation des risques est au fondement même du modèle économique : la constitution de classes de risques homogènes, permettant d'ajuster les primes au niveau de risque attendu, conduit nécessairement à différencier les assurés et peut générer des écarts de traitement entre individus ou groupes (*cf.* section 2). Dans ce contexte, les dispositifs internes des organismes d'assurance, qu'ils relèvent de la conformité ou de cadres éthiques, jouent un rôle central pour encadrer l'évaluation des risques ou la tarification, en définissant les variables admissibles, en organisant les contrôles de biais et en fixant des limites jugées acceptables aux possibles écarts de traitement entre individus ou groupes sociaux.

Dans le secteur bancaire, des enjeux similaires peuvent se poser, notamment dans les processus d'évaluation de solvabilité et d'octroi de crédit, mais **la tension entre équité et différenciation y est plus diffuse**, et s'inscrit davantage dans une logique de gestion des risques que dans une articulation directe avec le modèle économique.

1.5 Dispositifs règlementaires hors de l'Union européenne

1.5.1 États-Unis

Aux États-Unis, l'approche juridique de l'équité s'inscrit d'abord dans le cadre plus large du droit de la non-discrimination, issu notamment du *Civil Rights Act* de 1964. Celui-ci a structuré deux grandes conceptions de la discrimination. D'une part, le ***disparate treatment*** correspond à une discrimination intentionnelle : un individu est traité différemment en raison de son appartenance à un groupe protégé. D'autre part, le ***disparate impact*** vise des situations où des règles ou pratiques apparemment neutres produisent, en pratique, des effets disproportionnés sur certains groupes. Cette seconde notion a historiquement joué un rôle important dans l'évaluation de l'équité des modèles, en permettant de caractériser des discriminations sans intention explicite.

Dans le secteur financier, ces principes généraux se sont traduits par des **législations spécifiques**, au premier rang desquelles l'*Equal Credit Opportunity Act* (1974). Cette loi interdit toute discrimination dans l'octroi du crédit, quelle que soit la technologie utilisée[36], et impose aux prêteurs de justifier toute décision défavorable. Le *Fair Credit Reporting Act* (1970) complète ce dispositif en encadrant l'usage des données dans les décisions de crédit (qualité et exactitude, respect de la vie privée, droit d'accès).

Pour **évaluer en pratique** l'existence d'un *disparate impact*, les autorités américaines ont historiquement eu recours à des outils empiriques, en particulier dans le domaine de l'emploi. Les *Uniform Guidelines on Employee Selection Procedures* disposent ainsi, dans leur article 4(D),

[36] Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms | Consumer Financial Protection Bureau.

qu'« *un taux de sélection pour une race*[37]*, un sexe ou un groupe ethnique donné qui est inférieur aux quatre cinquièmes (4/5) (soit 80 %) du taux enregistré par le groupe présentant le taux le plus élevé sera généralement considéré par les autorités fédérales chargées de l'application de la loi comme une preuve d'impact négatif, tandis qu'un taux supérieur aux quatre cinquièmes ne sera généralement pas considéré par ces mêmes autorités comme une preuve d'impact négatif* ». Cette règle empirique a progressivement été utilisée **dans un grand nombre de domaines**, notamment le secteur financier. Elle a inspiré, de manière plus ou moins explicite, certaines pratiques contemporaines de mesure de l'équité algorithmique (*cf.* section 3).

Il est enfin à noter que des évolutions récentes marquent un **changement notable d'approche** en matière d'équité aux Etats-Unis. Un *executive order* signé par le président Trump en août 2025[38] a en effet conduit les agences fédérales à **abandonner** l'usage du *disparate impact* dans leurs activités de supervision. Des institutions comme l'*Office of the Comptroller of the Currency* (OCC)[39] et le *Consumer Financial Protection Bureau* (CFPB)[40] ont ainsi annoncé qu'elles ne s'appuieraient plus sur cette notion. Cette inflexion pourrait ainsi réduire le recours à des métriques d'équité, au profit d'une approche centrée sur la preuve de discrimination intentionnelle.

1.5.2 Royaume-Uni

Au Royaume-Uni, l'évaluation des discriminations algorithmiques s'appuie d'abord sur le cadre juridique général de l'*Equality Act* (2010)[41], qui fournit les catégories permettant de qualifier les pratiques discriminatoires. En particulier, la notion de **discrimination indirecte** (section 19 de l'*Equality Act*) joue un rôle central : elle permet de saisir des situations dans lesquelles des règles ou des modèles apparemment neutres produisent, en pratique, des effets disproportionnés sur certains groupes, sans justification objective, ce qui est particulièrement pertinent pour les systèmes d'IA, dont les biais sont souvent systémiques et non intentionnels. Ce cadre ne s'accompagne **pas de seuils ou de métriques standardisées** pour caractériser ces effets : l'analyse repose largement sur une appréciation au cas par cas, combinant éléments empiriques et raisonnement juridique, ce qui laisse une marge d'interprétation importante aux juges et aux superviseurs.

Dans le secteur financier, la *Financial Conduct Authority* (FCA) décline ces principes en adoptant une approche centrée sur les **impacts concrets pour les consommateurs**, notamment dans le cadre dit « *Consumer Duty* »[42]. L'évaluation de l'équité ne se limite pas à la conformité formelle des modèles, mais porte sur les effets potentiels de leur utilisation : exclusion de certains profils de l'accès au crédit, conditions moins favorables pour des publics vulnérables, ou segmentation excessive du marché. Les orientations récentes de la FCA soulignent que les systèmes d'IA ne doivent ni porter atteinte aux droits des individus ni générer de discriminations injustifiées, et qu'ils doivent être conçus en tenant compte de critères d'équité adaptés à leur contexte d'usage. Cette approche est complétée par un contrôle des **dispositifs internes de gouvernance des**

[37] Terme à considérer dans le contexte états-unien.
[38] *Executive order* n°14281 : Federal Register : Restoring Equality of Opportunity and Meritocracy.
[39] Fair Lending: Removing References to Disparate Impact | OCC.
[40] Fair Lending Report of the Consumer Financial Protection Bureau for 2024.
[41] Celui-ci prolonge et développe un certain nombre de lois antérieures, en particulier le *Sex Discrimination Act* (1975) et le *Race Relations Act* (1976), qui avaient posé les premières bases de la lutte contre les discriminations au Royaume-Uni.
[42] Consumer Duty | FCA.

modèles, en particulier la capacité des institutions à identifier, mesurer et corriger les biais tout au long du cycle de vie des systèmes.

1.5.3 Singapour

À Singapour, l'encadrement de l'équité dans le secteur financier repose sur une approche **principielle et opérationnelle**, largement structurée autour du rôle de la *Monetary Authority of Singapore* (MAS). La MAS a posé en 2018 les fondements de cette approche avec les principes dits **FEAT** (*Fairness, Ethics, Accountability, Transparency*), qui constituent un cadre de référence pour l'usage de l'IA et des données dans les services financiers. Ces principes, non contraignants mais largement structurants, visent à garantir que les systèmes ne produisent pas de **désavantages systématiques injustifiés** pour certains individus ou groupes, tout en imposant des exigences en matière de gouvernance, de responsabilité des décisions et d'explicabilité[43]. La MAS affirme en particulier que l'équité par l'ignorance n'est plus adaptée à des modèles d'IA (*cf.* section 2.2).

Pour opérationnaliser ces principes, la MAS a développé, en partenariat avec le secteur financier, l'initiative **Veritas**, qui constitue l'un des dispositifs les plus avancés au niveau international en matière d'évaluation de l'équité algorithmique. Le cœur de Veritas est une **méthodologie d'évaluation structurée**, couvrant l'ensemble du cycle de vie des systèmes d'IA (conception, développement, déploiement, suivi), et devant permettre aux institutions financières de traduire les principes FEAT en processus concrets.

Cette méthodologie comprend notamment des outils pour identifier les variables sensibles, détecter et mesurer les biais, choisir la métrique d'équité la plus pertinente (y compris au moyen d'arbres de décision) et documenter les arbitrages effectués. Elle a été progressivement transcrite dans une **boîte à outils *open source***[44], développée par un consortium d'acteurs publics et privés, qui permet d'automatiser certaines analyses (par exemple le calcul de métriques d'équité) et de standardiser les pratiques d'évaluation.

Un point saillant de l'approche Veritas est son caractère **holistique et contextuel** : l'équité n'est pas réduite à un indicateur statistique unique, mais intégrée dans un ensemble plus large de dimensions : éthique, gouvernance, transparence. Les méthodes proposées combinent ainsi analyses quantitatives (tests de biais, comparaisons entre groupes) et qualitatives (justification des choix de conception, documentation des risques, dispositifs de contrôle interne). En outre, la MAS insiste sur une approche **fondée sur les risques**, dans laquelle le niveau d'exigence en matière d'évaluation dépend de l'impact potentiel du système.

[43] Cette approche est complétée par des lignes directrices plus larges sur le *fair dealing*, qui imposent aux institutions de concevoir et distribuer des produits financiers adaptés aux besoins des clients, d'expliquer les décisions et de pouvoir justifier tout traitement différencié entre catégories de clients.
[44] Veritas Toolkit 2.0, ressource en open source.

2 La notion d'équité dans le secteur financier

2.1 La notion de biais discriminatoire

Dans les réflexions sur l'équité, les termes de biais et de discrimination sont parfois utilisés de manière imprécise ou interchangeable, au risque de brouiller l'analyse. Afin de clarifier le débat et d'éviter toute confusion, le présent document distingue **trois niveaux d'analyse**, correspondant à des réalités conceptuelles et opérationnelles différentes[45].

Une **disparité** désigne une différence systématique dans le comportement, les sorties ou les performances d'un modèle entre des groupes. Les disparités sont fréquentes et ne sont pas nécessairement illégitimes. Ainsi, un modèle d'octroi de crédit exploitant le revenu produira par construction des distributions de scores différentes entre groupes aux revenus différents. Une disparité constitue donc un simple constat statistique : elle ne préjuge ni de l'existence d'un biais ni, a fortiori, d'une discrimination.

Un **biais** est une disparité qui traduit un écart systématique par rapport à une norme d'évaluation pertinente. Cette norme est fréquemment l'exactitude (un modèle commet plus d'erreurs sur un groupe), mais elle peut également porter sur la qualité de service (un système conversationnel comprenant moins bien certaines clientèles, *cf.* section finale sur l'IA générative), ou encore sur la robustesse (un modèle se révélant plus fragile sur certaines populations). Caractériser un biais suppose donc d'expliciter la norme par rapport à laquelle l'écart est jugé : ce n'est pas la disparité en elle-même qui constitue le biais, mais l'écart qu'elle manifeste par rapport à un standard donné.

Une **discrimination** correspond à un biais considéré comme inacceptable au regard du contexte d'usage, de la nature et de l'ampleur du préjudice (avéré ou potentiel), des droits affectés et du cadre juridique applicable. Pour une autorité de contrôle, la norme sous-jacente est une norme juridique, fondée sur les critères protégés par le droit européen ou national (*cf.* section 1.1). Pour une institution financière, cette norme peut également découler de sa politique interne : une entreprise peut décider d'aller au-delà des exigences légales en intégrant des considérations éthiques plus larges[46] dans son appréciation de l'équité.

Du point de vue opérationnel, ces trois notions correspondent à des niveaux d'analyse distincts. Une disparité se constate et se mesure statistiquement, sous réserve des questions d'incertitude (*cf.* section 4.1). L'identification d'un biais requiert en outre de préciser la norme d'évaluation mobilisée et de justifier sa pertinence pour le cas d'usage. Enfin, la qualification de discrimination implique un jugement contextuel sur l'acceptabilité de l'écart, au regard des enjeux, des droits en cause et du cadre juridique. Confondre ces différents niveaux peut conduire soit à qualifier de discrimination toute disparité soit, à l'inverse, à ne la reconnaître comme discrimination que dans des cas extrêmes ou intentionnels.

Ces distinctions précisent les notions juridiques de différenciation et de discrimination introduites en section 1.1. Une différenciation fondée sur des critères objectifs et pertinents correspond à une disparité légitime au regard du cas d'usage. La discrimination directe renvoie à un biais fondé explicitement sur un critère protégé. La discrimination indirecte, quant à elle,

[45] Loubes, et al., 2026.

[46] Les travaux scientifiques opposent traditionnellement deux normes éthiques : le monde tel qu'il est, et le monde tel qu'il devrait être.

résulte d'effets de corrélation (ou de *proxy*) entre variables ; sa qualification juridique dépend de l'existence d'une justification objective et proportionnée[47].

2.2 Biais algorithmiques et « *Big Data* » dans le secteur financier

Fondamentalement, la question de l'équité dans le secteur financier se heurte à une **tension structurelle** entre, d'une part, la **nécessité économique de différencier** les individus en fonction des risques – afin d'assurer la soutenabilité des modèles – et, d'autre part, **l'interdiction des discriminations** fondées sur des critères non objectifs ou protégés. Cette tension est d'autant plus aiguë que certaines variables protégées sont **statistiquement corrélées aux risques observés**, du fait de mécanismes causaux indirects, ce qui peut rendre leur exclusion potentiellement dommageable en termes d'exactitude prédictive.

Pour répondre à cette difficulté, une approche longtemps dominante a consisté à « protéger » les variables sensibles en les éliminant du traitement statistique[48]. **Aujourd'hui, cette méthode, appelée « équité par l'ignorance », est rendue obsolète par la grande dimensionnalité des jeux de données, qui se traduit par la forte colinéarité des variables protégées avec d'autres non protégées.** Les algorithmes peuvent ainsi détecter des **variables de substitution** *(proxies)* qui leur permettent de « reconstituer » l'information contenue dans les variables protégées. Par exemple, le code postal, le type de téléphone, ou les habitudes de dépenses peuvent être fortement corrélés à une variable protégée (comme l'origine ethnique).

Un modèle d'IA peut donc parfaitement discriminer les individus sans jamais « voir » la variable sensible. En outre, cette discrimination indirecte peut être parfois plus opaque et moins contrôlable que les discriminations explicites. **La discrimination peut ainsi se manifester comme un effet collatéral du traitement de données massives**. Dans ce contexte, certains travaux de recherche suggèrent non pas d'interdire la collecte et l'usage de variables sensibles, mais au contraire de les exploiter à des fins de lutte contre les discriminations[49] (*cf.* encadré 7 en section 5.2).

2.3 Les différentes sources de biais

Les biais discriminatoires peuvent d'abord provenir des données. On peut ainsi distinguer, en premier lieu, les **biais historiques** : les données peuvent refléter des pratiques passées discriminatoires (refus de crédit ou d'assurance, moindre accès à certains produits, etc.), que les modèles auront tendance à répliquer et normaliser. Ensuite, les **biais de représentation**, liés à des données incomplètes ou déséquilibrées : certains groupes sont sous-représentés, mal observés ou décrits par des variables moins pertinentes, ce qui aura tendance à dégrader la qualité des prédictions les concernant. Une troisième source majeure est celle des **biais de mesure et de qualité** des données (erreurs, approximations, agrégations inadaptées), qui affectent généralement de manière disproportionnée certaines populations, notamment celles dont les revenus ou les parcours ne sont pas « standard ». Enfin, les biais peuvent émerger de la

[47] Des travaux de recherche montrent que ce qui est considéré comme acceptable dans l'utilisation de certaines variables peut évoluer au fil du temps. Charpentier & Barry (2022) rappellent par exemple qu'à la fin du XIXe siècle, des assureurs américains considéraient comme scientifiquement établi le lien entre couleur de peau et espérance de vie, et l'utilisaient en pratique dans leurs modèles. Plus généralement, les choix effectués en matière de modélisation peuvent contenir une part de subjectivité, dès lors qu'ils sont ancrés dans un contexte historique et social spécifique (Glenn, 2000).
[48] Simon, 1988.
[49] Williams, Brooks, & Shmargad, 2018.

présence de **variables *proxy***, lorsque des indicateurs apparemment neutres (lieu de résidence, situation maritale, etc.) sont fortement corrélés à des caractéristiques protégées et produisent des discriminations indirectes.

Les discriminations peuvent également tenir à la modélisation elle-même, c'est-à-dire aux choix techniques et conceptuels structurant le traitement. Une première source est le **choix des objectifs et des fonctions d'optimisation** : en cherchant exclusivement à maximiser la performance globale (rentabilité, exactitude, réduction du risque), les modèles peuvent produire des résultats inéquitables pour certains groupes. Une deuxième catégorie concerne les **choix de paramétrage et de règles de décision**, comme les seuils, les règles d'arbitrage ou de rejet automatique, qui peuvent avoir des effets très différents selon les populations sans que cela soit explicitement pris en compte.

S'ajoutent en troisième lieu des **effets d'amplification et de rétroaction** propres aux systèmes algorithmiques. De fait, la recherche de performance statistique peut, en elle-même, contribuer à amplifier les biais présents dans les données via plusieurs mécanismes. D'abord, l'optimisation d'une erreur moyenne sur l'ensemble de la population conduit mécaniquement le modèle à **privilégier les groupes majoritaires ou les cas les plus fréquents**, au risque de tolérer des erreurs plus importantes pour certains groupes minoritaires. Ensuite, les algorithmes d'apprentissage tendent à exploiter en priorité les **corrélations les plus fortes** dans les données, qui reflètent souvent des structures sociales ou des biais historiques : ces corrélations peuvent ainsi être non seulement reproduites, mais aussi renforcées dans les décisions du modèle. Les modèles non linéaires peuvent aussi **amplifier** des différences initialement modestes en les combinant avec d'autres variables, produisant des **effets disproportionnés** dans certaines situations. Enfin, lorsque les décisions du modèle influencent les **données futures** (par exemple en conditionnant l'accès au crédit ou à certains services), ces biais peuvent se renforcer au fil du temps à travers des **boucles de rétroaction**. L'ensemble de ces mécanismes concourt à ce qu'un modèle puisse **accentuer les écarts** existants entre groupes.

Enfin, des travaux récents suggèrent que l'amplification des biais ne résulte pas uniquement des propriétés du modèle final, mais également de la **trajectoire d'entraînement**[50]. En particulier, lors des premières phases d'optimisation, **le modèle tend à apprendre les régularités associées aux groupes majoritaires, tandis que celles propres aux groupes minoritaires sont apprises plus tardivement**. Dans ces conditions, un modèle dont l'apprentissage est interrompu précocement (*early stopping*) ou dont la capacité est limitée peut amplifier les biais parce qu'il n'a pas eu le temps ou les ressources nécessaires pour intégrer les structures propres aux groupes minoritaires. De même, le choix de l'algorithme d'optimisation (descente de gradient, Adam, etc.) peut influencer la rapidité avec laquelle les régularités minoritaires sont apprises. Ces éléments soulignent que **la durée d'entraînement, la capacité du modèle, et le choix de l'optimiseur ne sont pas des paramètres neutres du point de vue de l'équité**. À ce titre, ils gagneraient à être explicitement pris en compte dans les processus de validation des modèles.

[50] Bachoc, Bolte, Boustany, & Loubes, 2026.

Encadré 3 : Méthode de décomposition des biais

La littérature scientifique propose souvent de décomposer une disparité observée en **deux composantes** : une part héritée des données, et une part imputable au traitement opéré par l'algorithme.

Le principe consiste à **comparer l'écart entre groupes sur les sorties du modèle à l'écart correspondant pour une variable de référence**. Cette variable de référence peut être la cible observée (par exemple, le défaut effectif), une donnée externe indépendante, ou encore une cible corrigée. Si l'écart sur les sorties est supérieur à l'écart sur la variable de référence, le modèle a **amplifié** une disparité préexistante ; s'il est comparable, le modèle l'a **transmise** sans modification notable ; s'il est plus faible, le modèle l'a **atténuée**.

Cette grille de lecture est cohérente avec la logique du Règlement IA, qui distingue les exigences relatives à la qualité des données (article 10) de celles portant sur la conception et le fonctionnement des modèles (article 15).

L'intérêt opérationnel de cette décomposition est d'orienter les actions correctrices. Lorsqu'un modèle amplifie sensiblement une disparité, cela suggère d'agir sur les phases d'apprentissage (choix de la fonction de perte, contraintes d'équité, optimisation) ou sur les sorties du modèle (post-traitement). À l'inverse, lorsqu'il se contente de transmettre une disparité déjà présente, les leviers d'action se situent prioritairement en amont, au niveau des données (collecte, sélection, transformation, rééquilibrage).

En pratique, le degré d'amplification peut être quantifié à l'aide de diverses **métriques** de distance ou de divergence entre distributions (distance en variation totale, distance de Wasserstein, divergence de Kullback-Leibler, etc.). Le choix de la métrique dépend du cas d'usage et de la nature des sorties (binaires, ordonnées, continues). Comme pour toute mesure statistique, il est recommandé d'associer ces estimations à des intervalles de confiance (*cf.* section 4.1).

2.4 Équité de groupe et équité individuelle

Les travaux de recherche distinguent traditionnellement **l'équité de groupe**, qui repose sur une vision collective de l'équité, fondée sur la séparation de la population en groupes distincts, et **l'équité individuelle**, qui se fonde sur les droits juridiques des personnes.

L'équité individuelle se fonde sur l'idée que **des profils similaires doivent faire l'objet d'un traitement similaire** ; par exemple, dans le contexte d'un octroi de crédit, deux candidats présentant des risques objectifs de défaut équivalents devraient se voir attribuer des scores de crédit similaires, quelles que soient leurs caractéristiques personnelles. **L'équité individuelle est donc d'abord une égalité procédurale**. Cette conception de l'équité permet théoriquement d'atteindre une **plus haute performance** que l'équité de groupe, car le modèle aura tendance à mieux respecter le profil de risque de chaque individu. Ce paradigme repose toutefois sur **deux hypothèses majeures et difficiles à vérifier en pratique**. Il suppose d'abord que l'on puisse définir une mesure objective de similarité entre deux individus, ce qui n'est pas toujours assuré, et pose en outre un problème particulier lorsque la mesure de similarité fait intervenir des attributs sensibles. Il repose ensuite sur l'hypothèse que les données d'entraînement constituent une représentation équitable de la réalité, ce qui revient à supposer qu'il n'y existe pas de biais discriminatoires.

L'équité de groupe repose, elle, sur une égalité de résultats : elle impose que le modèle affiche des performances similaires, d'un point de vue quantitatif, entre différents sous-groupes de la population. Par exemple, on peut souhaiter que les hommes et les femmes *dans leur ensemble* bénéficient du même taux d'acceptation pour leurs crédits ; on ne s'intéresse alors pas directement aux individus, mais aux groupes sociaux. **L'équité de groupe est généralement plus facile à vérifier avec des outils statistiques** et permet de réduire les **inégalités systémiques** en rééquilibrant des différences de traitement considérées comme non justifiées. Toutefois, l'équité de groupe a davantage tendance à diminuer la performance globale du modèle car l'imposition d'une parité de taux d'approbation ou de taux d'erreur s'accompagnera généralement d'une augmentation des erreurs de classification (*cf. infra*).

Ainsi coexistent **deux conceptions légitimes mais distinctes** de l'équité. L'équité individuelle relève d'une **approche plus juridique :** chaque personne, placée dans une situation comparable, doit être traitée de manière égale au cours de la procédure, et toute différence de traitement doit être objectivement justifiée au niveau du cas individuel. À l'inverse, l'équité de groupe relève d'une **approche plus statistique**, historiquement dominante en finance, qui raisonne en termes de **groupes de risques** (notamment en assurance) : les décisions sont fondées sur des régularités observées au sein de populations, indépendamment de la situation précise de chaque individu. Cette **logique collective** est au cœur de l'évaluation de risque ou de la tarification assurantielle[51], mais elle entre en tension avec le **principe juridique selon lequel nul ne devrait être pénalisé en raison de son appartenance à un groupe**.

L'opposition entre équité individuelle et équité de groupe mérite toutefois d'être **nuancée**. Sur le plan des principes, les exigences de **cohérence** (« traiter les cas semblables de façon semblable ») et **d'égalité des chances** (« exclure les désavantages non imputables au comportement des individus ») sont, au moins dans une certaine mesure, **compatibles, et pourraient en pratique fonder aussi bien des métriques individuelles que de groupe[52]. En outre,** dans la pratique financière, ces deux approches ne sont ni entièrement incompatibles ni substituables : l'équité de groupe permet de structurer des décisions cohérentes, prédictibles et soutenables à l'échelle d'un portefeuille, tandis que l'équité individuelle rappelle la nécessité de reconnaître les situations particulières.

En tout état de cause, en raison des difficultés pratiques de mise en œuvre des métriques d'équité individuelle, **ce document se concentre sur les mesures d'équité de groupe**.

2.5 Quels groupes prendre en compte pour analyser l'équité ?

La comparaison des différences de traitement entre deux ou plusieurs groupes conduit naturellement à s'interroger sur la façon de composer ces différents groupes. Là encore, deux grandes conceptions coexistent : les groupes peuvent être définis selon **une seule variable sensible** (**analyse univariée**), ou par la **combinaison de plusieurs variables sensibles** (**analyse multivariée ou intersectionnelle**).

L'analyse univariée conduit à composer des groupes selon une seule dimension à la fois (genre, âge, etc.). Elle consiste ainsi à comparer la manière dont le modèle traite deux groupes (par

[51] Ewald, 2011.

[52] Des études vont même plus loin et montrent une forme d'équivalence, par exemple, entre parité démographique et équité contrefactuelle (une mesure d'équité individuelle). Voir Rosenblatt & Witter, 2023.

exemple, les hommes et les femmes pour la variable sensible « genre ») ou davantage (par exemple sur l'âge, en composant des groupes par tranches de 10 ans).

L'analyse univariée présente des **avantages indéniables**. Elle est d'abord **simple à mettre en œuvre**, puisque le nombre de groupes à étudier demeure limité. Elle est également **plus lisible**, dans la mesure où toutes les parties prenantes peuvent aisément comprendre ce que recouvrent les groupes comparés. Toutefois, les travaux de recherche mettent en évidence une **limite majeure** de cette approche : **elle peut masquer l'existence de fortes disparités au niveau des sous-groupes**. Un modèle peut ainsi présenter un traitement global équitable des groupes hommes et femmes, tout en lésant fortement les femmes de moins de 25 ans[53], par exemple.

Plus largement, les travaux de recherche montrent qu'**il n'existe pas de justification théorique satisfaisante pour restreindre l'analyse des différences de traitement à une seule variable sensible**. Les discriminations ne sont pas forcément **additives** : elles peuvent être **combinatoires**. Autrement dit, les biais peuvent être amplifiés à l'intersection de plusieurs attributs sensibles**. C'est pourquoi il existe un fort consensus dans la littérature scientifique pour recommander d'examiner des différences de traitement entre des groupes définis (analyse multivariée ou intersectionnelle)**.

Croiser les dimensions suppose néanmoins de résoudre certaines **difficultés pratiques**. En premier lieu, le croisement de plusieurs variables sensibles conduit mécaniquement à fragmenter la population en un **plus grand nombre de groupes**, dont certains peuvent présenter des **effectifs très faibles**, rendant les résultats instables ou statistiquement non robustes. L'analyse multivariée engendre plus fondamentalement une tension conceptuelle : lorsque les sous-groupes sont définis de manière trop fine, l'analyse se rapproche du niveau individuel et **entre en tension avec la logique même de la modélisation**, consistant à exploiter des régularités statistiques afin de différencier les traitements.

[53] Ce phénomène est parfois qualifié de « charcutage » (*gerrymandering*) de l'équité. Voir Kearns et al., 2018.

Encadré 4 : Équité dans chaque établissement et équité globale

Il convient d'observer que **l'exigence d'équité appliquée à la seule clientèle d'une banque ou d'un organisme d'assurance ne garantit pas nécessairement une situation équitable à l'échelle de la population totale.** Cela résulte d'abord de ce que **la politique commerciale de chaque établissement a un impact direct sur la composition de sa clientèle** : même sans aucune pratique discriminatoire, la conception de produis destinés à un type de clientèle, des tarifs plus avantageux pour une catégorie, ou encore des publicités ciblées sur certains segments de la population conduisent à la sur-sélection de groupes sociaux particuliers (et à la sous-sélection des autres). Pour prendre un exemple extrême, un assureur qui déciderait de n'avoir que des hommes pour clients pourrait présenter une grille tarifaire parfaitement équitable entre hommes et femmes, mais qui ne trouverait aucune application pratique. Ainsi, les clients d'une institution financière particulière forment généralement un **échantillon statistiquement biaisé** de la population totale[54].

En outre, chaque institution financière n'a par construction **accès qu'à ses propres données**. Quand bien même elle souhaiterait que sa clientèle représente fidèlement la population totale, elle n'aurait pas forcément les moyens de connaître la distribution de l'ensemble des caractéristiques. Il existe toutefois des cas de **mise en commun de certaines données sur la clientèle**, qui pourraient être utilisés à des fins d'équité globale. Dans le secteur bancaire, par exemple, une telle mise en commun se pratique dans la plupart des pays de l'OCDE, via les *credit bureau*, qui agrègent les informations sur la clientèle (historique de crédit, de mouvements sur les comptes, etc.) provenant de l'ensemble des banques du marché[55].

[54] Côté, Côté, & Charpentier, 2024.

[55] La mise en commun des données peut en outre avoir des effets positifs sur l'installation de nouveaux acteurs, qui peuvent accéder à des données pour entraîner leurs modèles.

3 L'équité de groupe

3.1 Les trois grandes familles de métriques de l'équité de groupe

De très nombreuses métriques d'équité de groupe co-existent dans la littérature scientifique ; elles peuvent toutefois être regroupées en **trois grandes familles**, associées chacune à un critère définissant une forme d'équité : **l'indépendance, la séparation, et la suffisance**.

Dans ce qui suit, chacune de ces grandes métriques est étudiée au moyen d'un exemple concret : **l'octroi de crédit**, considéré ici, pour plus de simplicité, comme une **décision binaire** d'acceptation ou de refus[56]. La **variable cible** est donc ici la décision elle-même, et non le score de risque, contrairement à la pratique la plus répandue[57]. Pour cela, on considère une population de 200 demandeurs de crédit : 100 du **Groupe A** (privilégié) et 100 du **Groupe B** (défavorisé), avec les caractéristiques suivantes :

- Groupe A : 80 personnes remboursent *in fine* (« bons payeurs »), et 20 font défaut (emprunteurs défaillants), soit un **taux de base d'emprunteurs viables**[58] **de 80 %**.
- Groupe B : 50 personnes remboursent *in fine*, et 50 font défaut, soit un **taux de base d'emprunteurs viables de 50 %**.

3.1.1 L'indépendance

- Principale métrique associée : la parité démographique (*demographic parity*).
- Mécanisme : ce critère exige que la décision du modèle soit indépendante de l'attribut protégé (par exemple le genre). En termes statistiques, dans un cadre de classification, **la probabilité de voir sa demande de crédit approuvée doit être égale pour tous les groupes**[59]. L'algorithme ne se demande pas si les candidats sont réellement capables de rembourser leur prêt ou non ; il s'assure simplement que les taux d'acceptation soient égaux.
- Logique fonctionnelle : Taux d'acceptation du groupe A = Taux d'acceptation du groupe B.

[56] Il s'agit donc d'un problème de classification. On distingue usuellement deux grandes classes de problèmes : les problèmes de classification, qui consistent à prédire une valeur discrète (le cas échéant, binaire, comme 0 ou 1 pour l'acceptation ou le refus d'un crédit), et les problèmes de régression, consistant à prédire une valeur continue (par exemple, le taux du crédit accordé).

[57] L'octroi de crédit repose le plus souvent sur l'estimation d'un score de risque (problème de régression). La décision d'accorder ou non le crédit est ensuite prise en appliquant un seuil à ce score : en deçà, le crédit est refusé ; au-delà, il est accordé.

[58] « Viable » est ici à comprendre comme un attribut *intrinsèque* de l'individu observé *a posteriori*. Un emprunteur viable est donc ici un emprunteur qui rembourse effectivement son prêt à l'avenir, et non un emprunteur prédit comme viable par le modèle.

[59] Dans un problème de régression, la parité démographique exige que les scores des groupes A et B présentent une distribution équivalente. Cette condition peut être assouplie en une égalité des scores *moyens* de chaque groupe.

- Implication : pour satisfaire ce critère, une banque pourrait être contrainte d'accepter des profils plus risqués au sein d'un groupe défavorisé simplement pour atteindre l'équilibre statistique.

Scénario 1 : On impose **l'indépendance** sur notre modèle d'octroi de prêt.

Objectif : La banque veut que les taux d'approbation soient identiques : les emprunteurs du groupe A doivent avoir le même taux d'approbation que ceux du groupe B.

Action du modèle : La banque impose un taux d'approbation de **60 %** pour chaque groupe.

- Pour le Groupe A : Le modèle prend les 60 meilleurs. Comme il y a 80 emprunteurs viables, le modèle trouve facilement 60 d'entre eux.
  → Résultat : 60 emprunteurs viables approuvés.
- Pour le Groupe B : Le modèle doit trouver 60 personnes, mais il n'y a que 50 emprunteurs viables au total. Il réussit à identifier les 50 emprunteurs viables mais inclut 10 emprunteurs non viables pour remplir le quota de 60.
  → Résultat : 50 emprunteurs viables approuvés, ainsi que 10 emprunteurs non viables.

3.1.2 La séparation

- Principale métrique associée : parité des taux d'erreur (*error rate parity)*[60].
- Mécanisme : ce critère exige une indépendance mathématique entre décision et variable protégée conditionnellement à la réalité. Autrement dit, le modèle doit avoir **des taux d'acceptation égaux entre groupes d'individus qui ont le même comportement réel** (même valeur de la variable cible). Statistiquement parlant, cela revient à contraindre le modèle à vérifier une **égalité des taux d'erreur**[61] entre les deux groupes considérés, c'est-à-dire une égalité des taux de faux positifs (proportion d'emprunteurs non viables acceptés par le modèle), ainsi qu'une égalité des taux de faux négatifs (équivalente, dans une classification binaire, à une égalité des taux de vrais positifs[62] ; cette dernière étant plus intuitive, elle est utilisée ci-après).
- Logique fonctionnelle :
  - Taux d'acceptation des emprunteurs viables du groupe A = Taux d'acceptation des emprunteurs viables du groupe B ; **ET**
  - Taux d'acceptation des emprunteurs non viables du groupe A = Taux d'acceptation des emprunteurs non viables du groupe B.
- Remarque : le calcul de ces taux repose sur l'observation du comportement réel de remboursement de l'ensemble des demandeurs ; or celui-ci n'est connu que pour les

[60] Aussi appelée « égalisation des cotes » (*equalized odds*).
[61] Cette condition est parfois relâchée, pour n'examiner que l'égalité des taux de vrais positifs ; cette métrique est alors généralement appelée « *equal opportunity* » dans la littérature.
[62] Dans un cadre de décision binaire, ces deux quantités sont complémentaires, c'est-à-dire que leur somme est nécessairement égale à 1.

emprunteurs effectivement acceptés. Leur estimation nécessite donc le recours à des heuristiques[63].

- Implication : les bons profils ne sont pas désavantagés par leur appartenance à un groupe donné ; dans le même temps les erreurs de prise de risque ne sont pas concentrées sur un groupe particulier.

Scénario 2 : On impose la **séparation** sur le modèle.

Objectif : La banque veut que les taux d'erreur soient identiques : les emprunteurs viables du groupe A doivent avoir le même taux d'approbation que ceux du groupe B, et idem pour les emprunteurs non viables.

Action du modèle : La banque ajuste le modèle pour garantir un taux de vrais positifs de 90 % et un taux de faux positifs de 10 %.

- Pour le Groupe A : Le modèle approuve 74 emprunteurs : 72 viables et 2 non viables.
  - Taux de vrais positifs = 72 / 80 = 90 %.
  - Taux de faux positifs = 2 / 20 = 10 %.
- Pour le Groupe B : Le modèle approuve 50 emprunteurs, 45 viables et 5 non viables.
  - Taux de vrais positifs = 45 / 50 = 90 %.
  - Taux de faux positifs = 5 / 50 = 10 %.

3.1.3 La suffisance

- Principale métrique associée : la parité des valeurs prédictives (calibrage[64]).
- Mécanisme : ce critère exige que **la réalité** (ici : le défaut) **soit indépendante de l'appartenance au groupe, conditionnellement à la prédiction donnée par le modèle**. Ainsi, si le taux de défaut réel observé *ex-post* est de 5 % pour les emprunteurs approuvés du groupe A, le taux de défaut doit également être de 5 % pour les emprunteurs approuvés du groupe B. Statistiquement parlant, cela revient à imposer une égalité des valeurs prédictives positives (proportion des prédictions positives effectivement correctes, c'est-à-dire le taux de remboursement réel des emprunteurs approuvés[65]) entre les groupes A et B, ainsi qu'une égalité des valeurs prédictives négatives (proportion des prédictions négatives effectivement correctes, c'est-à-dire taux de défaut réel des emprunteurs non approuvés)[66].
- Logique fonctionnelle :
  - Taux de défaut des emprunteurs approuvés du groupe A = Taux de défaut des emprunteurs approuvés du groupe B ; **ET**
  - Taux de défaut des emprunteurs non approuvés du groupe A = Taux de défaut des emprunteurs non approuvés du groupe B.

[63] Plusieurs méthodes d'inférence de rejet (« *reject inference* ») existent pour estimer le taux de faux négatifs, avec des taux de succès mitigés. Voir par exemple Ehrhardt, Biernacki, Vandewalle, Heinrich, & Beben, 2021.
[64] *Calibration* en anglais.
[65] Ou, de manière équivalente, taux de défaut.
[66] Comme pour la parité des taux d'erreur, il est possible de relâcher cette contrainte pour n'exiger que la parité des valeurs prédictives positives ou négatives.

- Remarque : la seconde condition nécessite généralement le recours à des estimations car l'établissement ne connaît normalement pas le taux de défaut des emprunteurs qu'il a rejetés.
- Implication : pour maintenir un même niveau de fiabilité entre les prédictions des différents groupes, la suffisance peut conduire à reproduire des disparités existantes, voire à les accentuer.

Scénario 3 : On impose la **suffisance** pour l'octroi de prêt.

Objectif : La banque veut que les valeurs prédictives soient identiques : les emprunteurs approuvés du groupe A doivent avoir le même taux de défaut que ceux du groupe B, et idem pour les emprunteurs rejetés.

Action du modèle : La banque ajuste le modèle pour garantir une valeur prédictive positive de 95 % et une valeur prédictive négative de 80 %.

- Groupe A : Le modèle approuve 80 emprunteurs (76 viables et 4 non-viables) et en rejette 20 (4 viables et 16 non viables).
  - Valeur prédictive positive = 76 / (76 + 4) = 95 %.
  - Valeur prédictive négative = 16 / (4+16) = 80 %.
- Groupe B : Le modèle approuve 40 emprunteurs (38 viables et 2 non-viables) et en rejette 60 (12 viables et 48 non viables).
  - Valeur prédictive positive = 38 / (38 + 2) = 95 %.
  - Valeur prédictive négative = 48 / (12+48) = 80 %.

3.2 Théorème d'impossibilité

La littérature scientifique montre qu'**il est mathématiquement impossible de construire un modèle qui satisfasse simultanément deux des trois critères d'équité de groupe** présentés ci-dessus (indépendance, séparation, suffisance) dès lors que la variable cible est corrélée à la variable sensible (autrement dit, dès que les taux de base diffèrent entre les groupes[67]). A fortiori, il est donc également impossible de satisfaire les trois critères dans ces situations.

[67] Barocas, Hardt, & Narayanan, 2019.

Les tableaux suivants illustrent ce point à partir des exemples développés plus haut :

Retour sur le **scénario 1** :

1. Indépendance (Respectée) :
    - Groupe A : Taux d'approbation = 60 / 100 = **60 %**
    - Groupe B : Taux d'approbation = 60 / 100 = **60 %**

2. Séparation (**Violée**) :
    - Groupe A : Taux de vrais positifs = 60 / 80 = **75 %**, Taux de faux positifs = 0 / 20 = **0 %**
    - Groupe B : Taux de vrais positifs = 40 / 50 = **80 %**, Taux de faux positifs = 10 / 50 = **20 %**
    - Conclusion : un emprunteur viable du groupe B sera plus souvent approuvé qu'un emprunteur viable du groupe A, mais un emprunteur non viable du groupe B sera plus souvent approuvé par erreur qu'un emprunteur non viable du groupe A.

3. Suffisance (**Violée**) :
    - Groupe A : Valeur prédictive positive = 60 / 60 = **100 %,**
      Valeur prédictive négative = 20 / 40 = **50 %**
    - Groupe B : Valeur prédictive positive = 50 / 60 ≈ **83 %,**
      Valeur prédictive négative = 40 / 40 = **100 %**
    - Conclusion : un emprunteur approuvé du groupe B fera plus souvent défaut qu'un emprunteur approuvé du groupe A, et un emprunteur non approuvé du groupe A aurait plus souvent remboursé qu'un emprunteur non approuvé du groupe B.

Retour sur le **scénario 2** :

1. Indépendance (**Violée**) :
    - Groupe A : Taux d'approbation = 74 / 100 = **74 %**
    - Groupe B : Taux d'approbation = 50 / 100 = **50 %**
    - Conclusion : le groupe A obtient beaucoup plus de prêts.

2. Séparation (Respectée) :
    - Groupe A : Taux de vrais positifs = 72 / 80 = **90 %,** Taux de faux positifs = 2 / 20 = **10 %**
    - Groupe B : Taux de vrais positifs = 45 / 50 = **90 %,** Taux de faux positifs = 5 / 50 = **10 %**

3. Suffisance (**Violée**) :
    - Groupe A : Valeur prédictive positive = 72 / 74 ≈ **97 %,**
      Valeur prédictive négative = 18 / 26 ≈ **69 %**
    - Groupe B : Valeur prédictive positive = 45 / 50 = **90** %
      Valeur prédictive négative = 45 / 50 = **90 %**
    - Conclusion : un emprunteur approuvé du groupe B fera plus souvent défaut qu'un emprunteur approuvé du groupe A, et un emprunteur non approuvé du groupe A aurait plus souvent remboursé qu'un emprunteur non approuvé du groupe B.

Retour sur le **scénario 3** :

1. Indépendance (**Violée**) :
   - Groupe A : Taux d'approbation = 80 / 100 = **80%**
   - Groupe B : Taux d'approbation = 40 / 100 = **40%**
   - Conclusion : le groupe A obtient beaucoup plus de prêts.

2. Séparation (**Violée**) :
   - Groupe A : Taux de vrais positifs = 76 / 80 ≈ **95%**, Taux de faux positifs = 4 / 20 = **20 %**
   - Groupe B : Taux de vrais positifs = 38 / 50 = **76%**, Taux de faux positifs = 2 / 50 = **4 %**
   - Conclusion : un emprunteur viable du groupe A sera plus souvent approuvé qu'un emprunteur viable du groupe B, mais un emprunteur non viable du groupe A sera plus souvent approuvé par erreur qu'un emprunteur non viable du groupe B.

3. Suffisance (Respectée) :
   - Groupe A : Valeur prédictive positive = 76 / 80 = **95 %,**
     Valeur prédictive négative = 16 / 20 = **80 %**
   - Groupe B : Valeur prédictive positive = 38 / 40 = **95 %,**
     Valeur prédictive négative = 48 / 60 = **80 %**

3.3 Comparaison des trois familles de métriques : hypothèses sous-jacentes et implications pratiques

Dans les sections précédentes, nous avons illustré les différences entre indépendance, séparation et suffisance à travers un exemple d'octroi de crédit ; cette section propose une **comparaison plus approfondie** de ces trois critères, en mettant en lumière leurs avantages, leurs limites et les implications normatives qu'ils véhiculent.

3.3.1 Indépendance

L'indépendance (métrique de parité démographique) est le plus simple des trois critères parce qu'elle **ne prend en compte que les prédictions du modèle**, ce qui est à la fois un avantage et un inconvénient. Elle est **la plus facile à mettre en œuvre** d'un point de vue algorithmique, car elle ne nécessite ni l'observation de la variable cible (le remboursement ou le défaut, dans l'exemple d'octroi de la section 3.1), ni une modélisation des erreurs du modèle, mais uniquement la comparaison directe des décisions entre groupes.

Pour autant, elle ne suffit pas toujours à garantir une situation « équitable ». Pour reprendre l'exemple de l'octroi, la parité démographique entre les groupes A et B peut parfaitement être atteinte en approuvant la moitié des emprunteurs de chaque groupe de manière **purement aléatoire**, c'est-à-dire sans tenir compte de leur « viabilité financière ». Elle peut donc conduire à une situation où des emprunteurs non viables sont approuvés **au détriment** d'emprunteurs viables, ce qui n'est pas une situation très « équitable » (voir aussi l'encadré 5 ci-après, sur les effets distributifs de chaque famille d'équité).

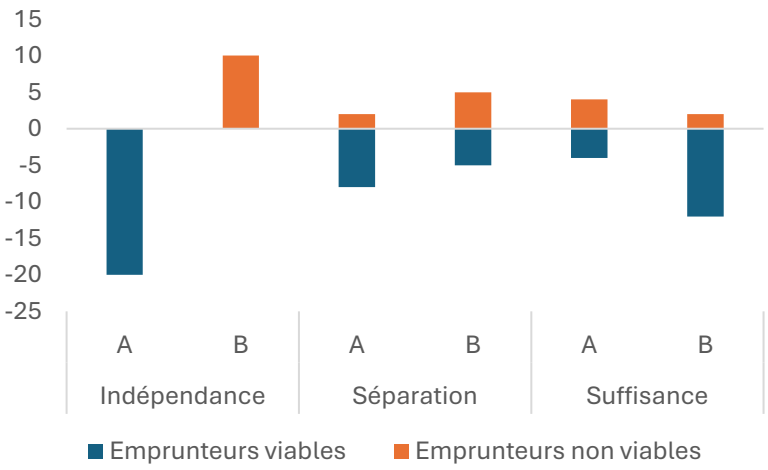
Encadré 5 : Synthèse des effets distributifs de chaque famille d'équité dans l'exemple développé

Le graphique ci-dessous synthétise l'exemple chiffré développé dans les sections précédentes, en illustrant les effets distributifs associés à chaque scénario d'équité. Il compare le résultat obtenu lorsque le modèle est successivement contraint par les critères d'indépendance, de séparation et de suffisance avec l'allocation de crédit considérée comme idéale au regard des hypothèses initiales : les dossiers des 80 emprunteurs viables du groupe A et les 50 emprunteurs viables du groupe B seraient acceptés, tandis que les autres dossiers seraient rejetés[68].

Ce graphique met en évidence les grandes tendances distributives de chaque famille d'équité (analysées plus en détail dans cette section). Il convient toutefois d'en souligner le caractère illustratif : cet exercice repose sur des hypothèses volontairement simplifiées et stylisées, dont la portée doit être interprétée avec prudence, d'autant plus que les modèles du secteur financier peuvent avoir de nombreux cas d'usage, qui ne se réduisent pas à une classification binaire entre un groupe favorisé et un groupe défavorisé.

Graphique : Comparaison entre le résultat de chaque scénario d'équité (indépendance, séparation et suffisance) et une distribution idéale

(En nombre de dossiers acceptés ou rejetés)



Le graphique met en évidence que chaque famille d'équité aboutit à des conséquences distributives différentes entre les deux groupes A (favorisé) et B (défavorisé). Sous l'**indépendance**, la contrainte de parité globale des décisions conduit à corriger fortement la distribution en faveur du groupe B : le nombre d'emprunteurs non viables acceptés y augmente significativement (+10), tandis que nombre d'emprunteurs viables du groupe A sont rejetés (- 20), ce qui traduit une perte significative d'efficacité. La **séparation** apparaît comme un compromis : elle réduit les écarts globaux entre groupes A et B, au prix d'une certaine sous-sélection d'emprunteurs viables dans le groupe A (−8) et d'erreurs de classification dans le groupe B. Enfin, la **suffisance** est assez performante dans la sélection des emprunteurs viables du groupe A, mais pénalise fortement les emprunteurs viables du groupe B (−12).

[68] Le nombre de dossiers acceptés est globalement comparable entre les 3 scénarios : 120 dossiers acceptés pour l'indépendance et la suffisance ; 124 pour la séparation du fait de contraintes mathématiques.

Plus généralement, la métrique de parité démographique ne tient **pas nécessairement compte des risques** associés aux individus ou aux groupes. Elle tend donc à **ignorer les différences de comportement** pouvant expliquer les différences de résultats entre groupes, en les interprétant comme dues à une inégalité historique des chances, autrement dit à des artéfacts de l'histoire et non à des différences intrinsèques entre les groupes considérés[69].

3.3.2 Séparation

Le critère de séparation (métrique de parité des taux d'erreur) entend corriger la principale limite de l'indépendance (métrique de parité démographique) **en prenant explicitement en compte la réalité du « terrain »**, c'est-à-dire le **comportement effectif** des individus. Autrement dit, la séparation est une définition de l'équité qui introduit la **notion de risque** (défaut sur un prêt, comportement à risque dans le cas d'une assurance dommage etc.). Elle permet donc de dépasser la principale critique adressée à la parité démographique. On peut se représenter la séparation comme une **forme de compromis** entre l'indépendance et la suffisance, dans la mesure où l'on cherche, d'une part, à prendre en compte l'influence de facteurs historiques et sociaux sur la situation réelle et, d'autre part, à conditionner les prédictions du modèle sur cette situation.

En outre, l'équité envisagée selon le principe de séparation prend en compte le fait **que les différents groupes sociaux peuvent subir des préjudices inégaux du fait du recours à des décisions automatisées**. En particulier, les modèles commettent souvent des taux d'erreur plus élevés sur des groupes historiquement marginalisés et défavorisés, infligeant ainsi un préjudice supplémentaire à ces groupes[70].

L'équité envisagée par le critère de séparation présente cependant **plusieurs inconvénients**. En premier lieu, cette forme d'équité tend mécaniquement à **diminuer la performance du modèle**. En effet, lorsque les distributions de risque diffèrent entre groupes, un modèle optimisé sans contrainte adaptera naturellement ses seuils de décision pour minimiser les erreurs globales. Imposer des taux d'erreur similaires entre groupes revient à s'écarter de ce compromis optimal : il faut alors souvent accroître volontairement certains types d'erreurs dans un groupe pour les aligner sur ceux observés dans un autre (voir par exemple l'encadré 5). L'objectif d'équité introduit une contrainte supplémentaire qui restreint l'espace des solutions possibles et empêche généralement d'atteindre simultanément la performance prédictive maximale.

---

[69] Dans cette conception, une importance particulière est donc généralement accordée aux groupes qui ont pu connaître des discriminations historiques et/ou qui présentent des différences systémiques sur de grands indicateurs de vie (exemple : taux de pauvreté) par rapport au reste de la population.

[70] L'exemple de l'algorithme COMPAS, utilisé par certains tribunaux aux États-Unis pour évaluer la probabilité de récidive des prévenus (en les classant à haut risque ou à bas risque), illustre bien cette question : l'organisation ProPublica a estimé que l'algorithme était biaisé contre les populations noires, en montrant que les prévenus noirs avaient un taux de faux positifs près de deux fois supérieur à celui des prévenus blancs (45 % contre 23 %), et qu'à l'inverse, le taux de faux négatifs était bien supérieur chez les prévenus blancs que noirs (48 % contre 28 %). En d'autres termes, COMPAS avait deux fois plus de chances de se tromper en classant à haut risque un prévenu noir qu'un prévenu blanc, et deux fois plus de chances de se tromper en classant à bas risque un prévenu blanc qu'un prévenu noir. Au contraire, les défenseurs de l'algorithme ont mis en avant des niveaux comparables de fiabilité des prédictions positives entre prévenus blancs et noirs (autrement dit, parmi les individus classés à haut risque, la proportion de personnes récidivant effectivement est similaire d'un groupe à l'autre, ce qui correspond à une métrique de suffisance : la parité des valeurs prédictives).

En deuxième lieu, le critère de séparation conduit à raisonner exclusivement à partir de la variable cible observée (par exemple, le défaut de l'emprunteur), supposée être une référence fiable. Or, dans de nombreux cas d'usage financiers, **cette variable cible est elle-même entachée d'imperfections** : biais de sélection (seuls certains profils ont été exposés à une décision et donc observés), dépendance aux décisions passées (historique de crédit construit sous contraintes), ou encore bruit de mesure. En conditionnant explicitement les contraintes d'équité sur cette cible, la métrique peut contribuer à **perpétuer ces imperfections**.

En troisième lieu, dans les configurations où les distributions de risque diffèrent entre groupes, la mise en œuvre pratique de la métrique de séparation conduit en général à l'utilisation de seuils de décision différenciés[71] selon les groupes. Ainsi, deux individus présentant un niveau de risque comparable peuvent être traités de manière différente du fait de leur appartenance à un groupe.

En quatrième lieu, enfin, les **implications éthiques et réglementaires des différentes erreurs peuvent être asymétriques**. Dans l'exemple de l'octroi de crédit, un faux négatif correspond à un refus de crédit à un emprunteur viable, avec un risque de perte de chance, tandis qu'un faux positif correspond à l'octroi d'un crédit à un emprunteur non viable, avec un risque de défaut, voire de surendettement. Or ces deux types d'erreurs sont liés par le seuil de décision appliqué au score de risque : ainsi, rendre le modèle plus strict permet de réduire les faux positifs – et donc le risque de défaut ou de surendettement – mais augmente mécaniquement les faux négatifs, c'est-à-dire les refus injustifiés de crédit à des emprunteurs viables. Dès lors, la contrainte de séparation revient à fixer un compromis entre ces objectifs contradictoires, ce qui entraîne une **réflexion éthique** : dans le cas du crédit, est-il plus acceptable de refuser plus souvent un prêt aux emprunteurs viables du groupe B qu'à ceux du groupe A, ou d'accorder plus souvent un prêt aux emprunteurs non viables du groupe B qu'à ceux du groupe A ? Tout dépend de l'évaluation sous-jacente des bénéfices et des risques en jeu.

3.3.3 Suffisance

Les métriques de suffisance reposent sur l'idée que la décision prise par un modèle doit avoir la **même signification quel que soit le groupe auquel appartient un individu**. Dans l'exemple ci-dessus, cela signifie que lorsque le modèle décide d'accorder un crédit, la probabilité que l'emprunteur le rembourse doit être identique pour tous les groupes : une décision positive ne doit pas être plus fiable pour un groupe que pour un autre[72]. D'un point de vue probabiliste, cela se traduit par **l'égalité des valeurs prédictives** (positive et négative) entre groupes : parmi les individus acceptés, la proportion de « bons payeurs » doit être comparable, et parmi les refusés, la proportion de « mauvais payeurs » également. Cette exigence est étroitement liée à la notion de **calibrage** : à un score donné (par exemple une probabilité de remboursement de 80 %) doit correspondre une même réalité empirique, indépendamment du groupe. La suffisance garantit ainsi une forme **d'équité dans l'interprétation des décisions**.

[71] De fait, le critère de séparation est généralement défini sur les décisions finales (ici : acceptation ou refus), et non sur les scores de risque eux-mêmes. Il est possible de formuler des variantes du critère directement sur les scores (c'est-à-dire sur un problème de régression et non de classification), en imposant par exemple que leur distribution soit indépendante du groupe conditionnellement au risque réel. Ces conditions sont toutefois plus exigeantes et rarement vérifiées en pratique.

[72] Bien que la suffisance puisse être évaluée sur des décisions binaires, elle constitue avant tout une propriété des scores de risque eux-mêmes, en ce qu'elle exige que, pour un niveau de score donné, la probabilité de l'événement d'intérêt soit identique entre groupes.

Cependant, dès lors que les groupes diffèrent en moyenne – notamment en raison de facteurs économiques ou sociaux – cette exigence entraîne des conséquences importantes. Si un groupe présente un taux de base plus faible (dans notre exemple : une proportion plus faible d'emprunteurs viables), alors maintenir une précision identique impose mécaniquement d'être **plus sélectif** pour ce groupe. Autrement dit, il faut **relever le seuil de décision** afin de n'accepter que les cas les plus sûrs, car **dans un groupe où les profils solides sont plus rares, appliquer le même niveau de sélectivité que dans l'autre groupe conduirait mécaniquement à prendre le risque d'accepter davantage d'emprunteurs non viables.**

Il en résulte une **diminution potentiellement importante** du nombre de décisions positives pour ce groupe (comme le montre par exemple l'encadré 5). La suffisance garantit donc une **équité *ex post***, au sens où les décisions prises ont toutes la même « qualité prédictive », mais elle ne garantit pas une **équité d'accès *ex ante*** : les individus appartenant à un groupe défavorisé peuvent en réalité avoir beaucoup moins de chances d'obtenir une décision favorable, même à caractéristiques comparables. En effet, les métriques de suffisance **ne limitent pas directement le nombre d'erreurs commises sur les individus réellement viables** : il est donc possible, tout en la respectant, de rejeter une part plus importante de bons candidats dans un groupe que dans un autre, dès lors que les décisions prises restent globalement fiables.

En outre, ces effets structurels peuvent être amplifiés par les caractéristiques des données et du modèle. En pratique, les groupes défavorisés sont souvent **moins bien représentés ou plus hétérogènes**, ce qui peut dégrader la qualité des scores : l'incertitude étant plus élevée, le modèle distingue moins bien les bons et les mauvais profils. Pour maintenir un niveau de précision constant, le modèle adopte alors une approche **plus prudente**, en **renforçant encore la sélectivité** à l'égard des groupes plus défavorisés.

Plus profondément, la suffisance repose sur les taux de base observés dans les données, qui reflètent eux-mêmes des réalités économiques, sociales ou historiques. En alignant les décisions sur ces taux de base, elle tend donc à **reproduire les différences existantes** entre groupes, voire à les **renforcer** lorsqu'elles sont déjà marquées. Autrement dit, les inégalités présentes dans les données ne sont pas corrigées, mais intégrées dans la logique de décision. Ainsi, si la suffisance offre une garantie forte en termes de cohérence probabiliste et de fiabilité des décisions, elle peut également conduire à une **exclusion** accrue et durable de certains groupes, en particulier lorsque ceux-ci présentent un niveau de risque moyen plus élevé.

3.3.4 Tableau récapitulatif

Nous résumons les considérations développées dans cette section dans le tableau ci-dessous.

Critère d'équité	Indépendance (parité démographique)	Séparation (parité des taux d'erreur)	Suffisance (parité des valeurs prédictives)
Mesure	Même taux d'approbation entre groupes.	Mêmes taux de faux positifs et de faux négatifs entre groupes.	Mêmes valeurs prédictives positives et négatives entre groupes.
Principe	« La proportion d'emprunteurs approuvés est la même pour tous les groupes. »	« Parmi les emprunteurs viables, la proportion d'emprunteurs approuvés est la même pour tous les groupes. »	« Parmi les emprunteurs approuvés, la proportion d'emprunteurs viables est la même pour tous les groupes. »
Implications normatives	- Postule que les inégalités actuelles sont dues à des artéfacts historiques plutôt que des différences intrinsèques.	- Postule que les erreurs de classification affectent les groupes de manière inégale en partie à cause des artéfacts historiques.	- Postule que les inégalités actuelles sont dues à des différences intrinsèques plutôt qu'à des artéfacts historiques.
Avantages	- Simplicité de mise en œuvre, notamment car ne requiert que les prédictions du modèle - Favorise l'accès au crédit des emprunteurs du groupe défavorisé	- Prend en compte la « viabilité » des emprunteurs - Favorise l'accès au crédit des emprunteurs viables du groupe défavorisé	- Prend en compte la « viabilité » des emprunteurs - Cohérence des décisions : une prédiction a la même signification pour tous les groupes (même score = même niveau de risque)
Inconvénients	- Ne tient pas compte de la « viabilité » des emprunteurs - Conduit à accepter un traitement différent d'individus similaires (en termes de risques) issus de groupes différents	- Tend à réduire mécaniquement la performance du modèle - Conduit à accepter un traitement différent d'individus similaires issus de groupes différents - Une même décision peut correspondre à des niveaux de risque différents selon les groupes	- Risque d'exclusion accrue : critères plus stricts pour les groupes défavorisés - Pas de protection uniforme des individus peu risqués : risque de rejeter davantage de « bons » candidats dans certains groupes - Reflète et peut accentuer des inégalités existantes entre groupes

Tableau 2 : Comparaison des critères d'équité de groupe

4 L'estimation et la correction des biais

4.1 Estimer les biais en pratique : prendre en compte l'incertitude

Comme toute estimation statistique, **la mesure d'un biais est entachée d'incertitude**. S'en tenir à une valeur ponctuelle, sans en apprécier la **précision**, peut donc conduire soit à identifier à tort un écart dû au hasard, soit, à l'inverse, à ne pas détecter un biais réel. La mauvaise prise en compte de cette incertitude peut fragiliser la crédibilité des diagnostics et conduire à mobiliser inutilement des ressources pour la remédiation.

Il convient ainsi de prendre en compte **quatre sources** principales d'incertitude[73] :

- **La variabilité d'échantillonnage** : les métriques d'équité sont calculées sur des effectifs finis, et **leur précision dépend directement de la taille des échantillons**. Un même écart peut être fortement significatif sur de grands volumes de données, et indiscernable du hasard sur de petits effectifs. **Les analyses d'équité devraient donc être systématiquement accompagnées d'intervalles de confiance**, obtenus par des méthodes analytiques (approximation normale, test exact de Fisher, etc.) ou par des approches de rééchantillonnage (*bootstrap*). Il convient **d'expliciter et de documenter** la méthode retenue.

- **La variabilité liée à l'entraînement** : les résultats peuvent dépendre des **aléas propres au processus d'apprentissage** (initialisation des paramètres, ordre de présentation des données, mécanismes de régularisation, etc.). Ainsi, deux entraînements du même modèle, sur les mêmes données, peuvent conduire à mesurer deux niveaux de biais différents. Une mesure de biais issue d'un entraînement unique ne constitue qu'une réalisation parmi d'autres possibles. Pour les systèmes les plus critiques, il est donc recommandé de **répéter les entraînements avec différentes graines aléatoires et de prendre en compte la distribution des mesures obtenues**.

- **La multiplicité des tests** : les analyses d'équité reposent souvent sur de nombreuses comparaisons (multiples groupes, métriques, etc.). Or cette multiplicité **accroît mécaniquement la probabilité de détecter au moins un écart significatif par effet du hasard**[74]. Des procédures classiques d'ajustement (Bonferroni, Holm, Benjamini-Hochberg) permettent de maîtriser ce risque et de maintenir la fiabilité des conclusions.

- **La puissance statistique sur les petits groupes** : les analyses portant sur de petits groupes se caractérisent par une **incertitude plus forte**. Dans ces situations, l'absence d'écart statistiquement significatif ne constitue pas une preuve d'équité, mais peut simplement refléter un manque de puissance du test. Il est donc recommandé de compléter les résultats par une **estimation de l'amplitude minimale détectable**, c'est-à-dire le plus petit écart que l'analyse est en mesure d'identifier compte tenu des données disponibles.

[73] Besse, del Barrio, Gordaliza, Loubes, & Risser, 2022.

[74] À titre d'illustration, évaluer 20 sous-groupes sans correction conduit à 64 % de chance de trouver au moins un écart significatif au seuil usuel de 5 %, alors même que le modèle est parfaitement équitable (Cook, Gebski, & Keech, 2004).

4.2 Les méthodes de correction des biais

Il existe un certain nombre de méthodes permettant de corriger les biais discriminatoires observés dans les modèles d'apprentissage automatique. Une manière courante de les classifier consiste à les séparer selon trois catégories[75] :

1. Les méthodes de **pré-traitement** (*pre-processing*) visent à corriger les discriminations sous-jacentes en amont par l'analyse ou la transformation des données d'entraînement ;
2. Les méthodes de **traitement intégré à l'apprentissage** (*in-processing*) visent à réduire les discriminations au cours du processus d'entraînement du modèle, par la modification de la fonction d'objectif ou par l'imposition de contraintes ;
3. Les méthodes de **post-traitement** (*post-processing*) visent à corriger les discriminations après l'entraînement, par l'ajustement des scores de sortie du modèle.

4.2.1 Les méthodes de pré-traitement

Les méthodes de pré-traitement regroupent un ensemble d'approches visant à agir **en amont** de la phase d'apprentissage, **en modifiant les données ou leur représentation** afin de réduire, voire éliminer, les dépendances entre les variables sensibles (par exemple le genre ou l'âge) et les autres caractéristiques utilisées par le modèle. L'idée centrale est de construire un **espace de données « assaini »**, dans lequel les biais potentiels ont été corrigés avant même l'entraînement du modèle. Cette démarche présente l'avantage d'être en grande partie **indépendante des algorithmes utilisés** : une fois les données transformées de manière adéquate, tout modèle entraîné sur ces données est censé hériter, au moins en partie, des propriétés d'équité recherchées. En ce sens, le pré-traitement constitue une **approche transversale et souvent modulaire**, qui peut être intégrée dans des chaînes de traitement variées. Toutefois, son efficacité dépend directement de la qualité des transformations opérées, ainsi que de la capacité à préserver l'information utile à la prédiction.

Concrètement, plusieurs grandes familles de méthodes peuvent être distinguées. Les approches dites d'« **aveuglement** » visent à neutraliser l'influence des variables sensibles en structurant les données selon des sous-groupes et des critères d'équité définis en amont. Les **méthodes causales** cherchent, quant à elles, à identifier les mécanismes à l'origine des discriminations en modélisant explicitement les relations entre variables, afin de corriger les biais à leur source, malgré les difficultés de mise en œuvre. D'autres techniques reposent sur le **découpage** du jeu de données en sous-groupes et sur des stratégies **d'échantillonnage**, afin de mieux représenter les populations désavantagées et d'évaluer les disparités. Les **méthodes de transformation** visent à construire de nouvelles représentations des données qui soient moins corrélées aux attributs sensibles tout en préservant leur utilité prédictive. Enfin, des approches plus opérationnelles consistent à modifier directement les données (**ré-étiquetage**, **perturbation**) ou leur poids relatif dans l'apprentissage (**rééquilibrage**), afin de corriger les déséquilibres observés.

4.2.2 Les méthodes de traitement intégré à l'apprentissage

Les méthodes de traitement intégré à l'apprentissage consistent à **agir directement au moment de l'entraînement du modèle**, en intégrant explicitement des objectifs d'équité dans le **processus d'optimisation**. Contrairement aux approches de pré-traitement, qui modifient les données en amont, ces méthodes **ajustent le comportement du modèle lui-même** afin qu'il

[75] Caton & Haas, 2024.

respecte certaines contraintes d'équité tout en maximisant sa performance prédictive. Elles permettent ainsi, en principe, d'atteindre un **meilleur compromis entre équité et exactitude**, puisque les deux objectifs sont optimisés conjointement. En revanche, elles présentent des **contraintes opérationnelles importantes** : elles nécessitent un **accès complet** aux données et aux algorithmes d'apprentissage, et sont souvent **spécifiques à certains types de modèles ou de problèmes**, ce qui limite leur portabilité et leur généralisation dans des environnements hétérogènes.

Plusieurs grandes approches peuvent être distinguées. Les méthodes de **régularisation et d'optimisation sous contrainte** consistent à modifier la **fonction d'objectif** du modèle en y intégrant des **pénalités** liées à des écarts d'équité, de manière à orienter l'apprentissage vers des solutions moins discriminatoires. Les méthodes d'**apprentissage antagoniste** introduisent, quant à elles, un mécanisme d'« adversaire » chargé de détecter l'information relative aux variables sensibles dans les prédictions ou les représentations du modèle, et d'inciter celui-ci à s'en affranchir. D'autres approches, comme les ***bandits* algorithmiques**, s'inscrivent dans un cadre d'apprentissage séquentiel et adaptatif : elles visent à prendre des décisions équitables au fur et à mesure, en équilibrant exploration et exploitation, et en intégrant l'équité comme un critère de performance dynamique.

4.2.3 Les méthodes de post-traitement

Les méthodes de post-traitement interviennent **une fois le modèle entraîné**, en **ajustant ses prédictions** afin de satisfaire un critère d'équité donné. Contrairement aux approches agissant en amont ou pendant l'apprentissage, elles ne nécessitent ni modification des données ni réentraînement du modèle, ce qui les rend particulièrement utiles dans des contextes où le modèle est **déjà déployé** ou fonctionne comme une « **boîte noire** »[76]. Cette flexibilité constitue leur principal atout : elles peuvent être appliquées à tout type de modèle, sans coût d'apprentissage supplémentaire. En contrepartie, le fait qu'elles reposent généralement sur une utilisation plus apparente de l'appartenance à un groupe (par exemple en fixant des seuils d'acceptation différents) peut poser des questions du point de vue juridique ou éthique.

Parmi les principales approches, les **méthodes de calibrage** visent à ajuster les scores produits par le modèle pour qu'ils aient une interprétation cohérente entre groupes, en alignant les probabilités prédites sur les fréquences observées. Elles sont particulièrement pertinentes lorsque les sorties du modèle servent d'aide à la décision plutôt que de décision automatique. Les **méthodes de seuillage**, quant à elles, consistent à fixer – éventuellement de manière différenciée selon les groupes – des seuils de décision qui permettent de concilier certaines mesures d'équité et de performance. Elles ciblent en particulier les situations ambigües, proches des seuils de décision, où les risques de biais sont les plus élevés.

4.2.4 Tableau récapitulatif

Nous résumons les considérations développées précédemment dans le tableau ci-dessous.

[76] En particulier, lorsque le modèle est acheté « sur étagère » auprès d'un fournisseur tiers.

Tableau 3 : Comparaison des différentes méthodes de correction des biais

Approche	Avantages	Inconvénients
Pré-traitement	- Transforme l'espace des variables afin qu'il soit **indépendant de l'attribut sensible** avant l'entraînement du modèle, ce qui le rend largement réutilisable dans différentes applications en aval.	- Cette approche **n'optimise pas directement l'estimateur** pour concilier à la fois l'équité et la performance prédictive pendant l'entraînement.
Traitement en cours	- Peut offrir **la meilleure performance**, car le modèle est optimisé en intégrant directement la contrainte d'équité dans le processus d'apprentissage.	- Nécessite un **accès** aux données brutes ainsi qu'à la procédure d'entraînement. - Cette approche est **moins générale**, car elle s'applique souvent seulement à certaines classes de modèles ou à certains schémas d'optimisation.
Post-traitement	- **Fonctionne avec n'importe quel modèle, même en « boîte noire »**. - Ne nécessite **pas de réentraînement**, du modèle, ce qui est utile lorsque l'entraînement d'origine est complexe ou indisponible.	- Cette approche repose souvent explicitement **sur l'appartenance à un groupe**, par exemple en fixant des seuils différents selon les groupes, ce qui peut constituer une difficulté.

5 Mise en œuvre pratique

Au regard des éléments développés précédemment, ainsi que des ateliers menés par l'ACPR avec plusieurs établissements volontaires (*cf.* l'encadré 6 ci-dessous), cette partie examine les **modalités concrètes de mise en œuvre** des principes relatifs à l'équité.

Encadré 6 : Les ateliers sur l'équité conduits par l'ACPR avec des acteurs financiers volontaires

Pour éclairer ces enjeux, l'ACPR a conduit, au printemps et à l'automne 2025, une série d'ateliers avec des acteurs financiers volontaires. Ceux-ci avaient pour objectif de comprendre comment les banques et les assurances interrogées traitaient concrètement les questions d'équité dans leurs processus, tant du point de vue **technique** que dans leur **gouvernance**.

Ces ateliers ont permis de dresser un état des lieux des réflexions et des pratiques des acteurs sur les différents enjeux liés à l'équité (*cf. infra*). Ils ont notamment montré que, dans les établissements interrogés – qui étaient vraisemblablement parmi les plus avancés sur ces questions au moment des entretiens – les travaux internes engagés sur les enjeux d'équité avaient pris des formes diversifiées (expérimentations, sensibilisation des différentes lignes de défense, lignes directrices internes). Dans tous les cas, les travaux étaient **relativement récents et peu nombreux**, les questions d'équité étant considérées comme complexes. De ce point de vue, les acteurs interrogés ont exprimé de fortes attentes vis-à-vis des superviseurs financiers – voire des législateurs nationaux ou européens – pour fixer les règles en la matière.

Ces ateliers ont également mis en lumière le fait que la plupart des acteurs financiers collectent **relativement peu de données protégées ou sensibles** : le genre, l'âge, le lieu de résidence dans la plupart des cas d'usage financiers, ainsi que les données de santé pour certains cas d'usage en assurance.

5.1 Considérations générales

5.1.1 Équité et gouvernance dans le secteur financier

Il ressort d'abord des éléments qui précèdent que **l'équité algorithmique ne peut pas être considérée comme une question purement technique**. En conséquence, contrairement à ce qui est souvent constaté en pratique, elle ne peut être laissée à l'appréciation des seuls *data scientists*. **Bien au contraire, elle engage des choix qui relèvent de la stratégie, de la gestion des risques et, plus largement, de la responsabilité de l'institution financière**. À ce titre, elle doit être appréhendée comme un **enjeu de gouvernance**, impliquant l'ensemble des niveaux décisionnels de l'organisation.

Ainsi, le **niveau le plus élevé de la gouvernance** a vocation à définir les **grandes orientations** en matière d'équité, comme il le fait par exemple pour l'appétit au risque. Le conseil d'administration pourrait ainsi engager sa responsabilité en approuvant une **politique écrite** en matière d'équité de l'IA, et en recevant périodiquement un rapport sur les indicateurs clés de performance (KPI) relatifs à l'équité. Les grands principes adoptés par la direction de l'entreprise pourront ensuite être **déclinés** au niveau des **différentes lignes métier**, où les différents cas d'usage – par exemple crédit à la consommation, crédit immobilier ou lutte contre la fraude – se caractérisent par des enjeux, des données disponibles ou encore des risques sensiblement différents. Enfin, le **niveau technique** a vocation **à traduire ces orientations en pratiques**

concrètes, en mobilisant des méthodes adaptées (choix des métriques d'équité, techniques de correction des biais, procédures de validation) et, en tout état de cause, à **l'état de l'art**. En outre, il conviendrait d'intégrer les enjeux d'équité aux **trois lignes de défense** classiques du secteur financier : les développeurs de modèles, la validation indépendante des modèles et la conformité, ainsi que l'audit interne.

Il convient de noter que cette gouvernance de l'équité – comme, plus généralement, la gouvernance des systèmes d'IA – ne suppose pas nécessairement la création de nouvelles structures, mais **peut parfaitement s'inscrire dans les dispositifs existants de gestion des risques de modèles**. L'enjeu est plutôt d'y intégrer explicitement les questions d'équité, au même titre que les préoccupations traditionnelles de performance ou de robustesse, afin de garantir une **approche cohérente et systématique**.

Dans ce cadre, il peut être utile pour les établissements du secteur financier de mettre en place, pour chaque cas d'usage de l'IA, un **processus de réflexion structuré** autour de **quelques questions clés** :

- Quels biais souhaite-t-on prévenir ou corriger ? La réponse à cette question suppose d'abord de préciser la norme de référence retenue pour caractériser une éventuelle discrimination. Cette norme est en premier lieu juridique : les institutions financières doivent se conformer aux exigences de non-discrimination formulées par le droit européen et national. Au-delà de ces exigences, elles peuvent choisir d'adopter des standards plus exigeants, reflétant leurs engagements éthiques ou leurs orientations stratégiques. Ces réflexions peuvent ainsi être l'occasion, pour la gouvernance de l'entreprise, d'expliciter son niveau d'ambition en matière d'équité et d'éclairer ses choix, en répondant à un certain nombre de questions, par exemple : a-t-elle déjà une bonne perception des disparités que sa politique commerciale ou son histoire ont pu induire dans la composition de sa clientèle, dans l'accès à ses services ou encore dans ses tarifs ? Juge-t-elle ces disparités légitimes, par exemple au regard de son positionnement concurrentiel ou de ses missions statutaires, ou souhaite-t-elle les réduire ? Se limite-t-elle à ne pas introduire de biais supplémentaires par rapport aux données observées, ou souhaite-t-elle corriger plus activement les inégalités existantes ?
- Quelles mesures techniques ou organisationnelles sont mises en œuvre pour y parvenir ?
- Ces mesures introduisent-elles de nouveaux risques – par exemple en dégradant la performance du modèle, en accroissant son opacité ou en générant d'autres formes de biais ?
- Enfin, comment l'arbitrage entre ces différents effets, souvent concurrents, est-il réalisé ?

5.1.2 Prendre en compte l'équité tout au long du cycle de vie du système

Afin de prendre en compte les enjeux associés à l'équité algorithmique, il convient, dès **la phase de développement**, d'intégrer les enjeux d'équité parmi les objectifs explicites du modèle, au même titre que la performance ou la robustesse. Cela suppose **d'identifier en amont les données sensibles** utilisées (*cf.* section 5.2) ainsi que les **groupes** susceptibles d'être affectés de manière différenciée par le modèle (*cf.* section 5.3 sur le choix des groupes). De manière générale, il est important d'examiner attentivement les **jeux de données**, en conduisant des analyses de représentativité et de qualité par groupe, en documentant les déséquilibres, les biais historiques (*cf.* section 2.3) et les limites connues. En outre, faire dialoguer les équipes de *data*

scientists avec les métiers, ainsi qu'avec les experts juridiques ou de conformité, peut permettre une meilleure prise en compte des enjeux d'équité.

Lors de la validation du modèle, il convient d'évaluer systématiquement les résultats par groupe, au moyen de métriques d'équité. Ceci nécessite de **définir la métrique la mieux adaptée** au cas d'usage (*cf.* section 5.4 sur ce choix), et d'arbitrer explicitement entre la réduction des disparités et les autres objectifs assignés au système (comme la performance – voir à ce sujet l'encadré 8). Il est de bonne pratique de **tracer et justifier** les résultats de ces analyses (voir en particulier la section 5.5 sur le choix des seuils pertinents) qui peuvent, le cas échéant, conduire à des **ajustements techniques** (*cf.* section 5.6) ou à des **décisions de gouvernance** sur l'acceptabilité du risque résiduel.

Au moment du déploiement, la prise en compte des enjeux d'équité peut nécessiter la mise en place de garde-fous organisationnels. L'établissement devra, dans certains cas, s'assurer que les utilisateurs du système comprennent ses limites, notamment en ce qui concerne les performances différenciées selon les groupes. Des dispositifs de supervision humaine, de contestation des décisions ou d'escalade pourront ainsi être prévus, en particulier pour les cas à fort impact individuel. Les **modalités d'usage du système** (seuils de décision, automatisation complète ou partielle, articulation avec des processus existants, intervention humaine) peuvent en effet avoir **autant d'effets sur l'équité que l'algorithme lui-même**, ce qui justifie de les concevoir et de les documenter avec le même niveau d'exigence.

Enfin, le suivi dans le temps du système est important pour identifier de potentielles dérives. Les données, les populations concernées et les usages évoluent, ce qui peut créer de nouvelles disparités ou accentuer des biais existants. Ainsi, les établissements peuvent utilement mettre en place une s**urveillance continue des indicateurs d'équité par groupe**, accompagnée de **revues périodiques et de mécanismes d'alerte**. En outre, comme pour d'autres dimensions (performance, explicabilité etc.), les retours utilisateurs, les audits internes ou externes peuvent alimenter un processus d'amélioration continue[77].

En tout état de cause, la prise en compte des enjeux d'équité par les établissements peut utilement reposer sur une **approche proportionnée, fondée sur les risques** : ainsi, le niveau d'exigence, la profondeur des analyses conduites ainsi que les dispositifs de gouvernance sont à **ajuster** aux impacts potentiels du modèle sur les individus.

5.2 Utilisation des critères protégés et des variables sensibles

Les caractéristiques personnelles collectées par les acteurs financiers sur leurs clients réels ou potentiels présentent des **statuts variés**, tant du point de vue juridique que de leurs effets potentiels (*cf.* section 1.1.3). Les développements qui suivent en proposent une illustration, sans prétendre couvrir de manière exhaustive l'ensemble des situations susceptibles d'être rencontrées.

L'âge des clients constitue une information couramment collectée par les acteurs financiers. Il s'agit d'un critère protégé au regard du droit de la non-discrimination, mais dont l'utilisation fait l'objet d'une appréciation relativement **souple**. En effet, dans de nombreux cas d'usage, l'âge est directement et objectivement corrélé à des **facteurs pertinents d'évaluation du risque**, tels

[77] Des boîtes à outil permettent d'automatiser certains de ces aspects. Outre l'outil Veritas (MAS, Singapour), déjà mentionné, on peut citer *IBM AI Fairness 360*, ou encore *Aequitas*, qui proposent des métriques et des méthodes d'identification et de correction des biais.

que l'horizon de remboursement, la stabilité et la trajectoire des revenus, mais aussi l'horizon d'investissement et la capacité à absorber des risques financiers sur le long terme. Ces liens permettent généralement de **justifier des différences de traitement** fondées sur ce critère. Par ailleurs, au sens du RGPD, l'âge constitue une donnée personnelle dite « ordinaire », dont le traitement est autorisé sous réserve du respect des principes généraux de protection des données.

La situation est nettement plus restrictive pour le **genre**. Bien qu'il ne soit pas classé parmi les données sensibles au sens du RGPD, il constitue un critère protégé en matière de non-discrimination. Les différences de traitement direct fondées sur le genre sont en principe **interdites**, et notamment en assurance depuis l'obligation de tarification unisexe en Europe (*cf.* encadré 2 en section 1.3.2). En outre, le genre fait partie des caractéristiques particulièrement exposées aux risques de **discrimination indirecte**, dans la mesure où il peut être reconstitué via de nombreuses variables de substitution (*proxy*) dans les modèles.

Les acteurs financiers collectent généralement aussi le **lieu de résidence** des clients[78] ; celui-ci n'est pas un critère protégé en tant que tel, et ne constitue pas non plus une donnée sensible au regard du RGPD. Toutefois, il peut agir comme un ***proxy* de caractéristiques protégées** (origine, niveau socio-économique, etc.), exposant ainsi à un risque de discrimination indirecte. Dans ce cas, son utilisation doit être **spécifiquement justifiée**, afin de démontrer qu'elle repose sur des considérations légitimes et proportionnées.

Enfin, le secteur de l'assurance collecte parfois des **données de santé** (en assurance emprunteur et en prévoyance notamment). Ces données constituent des **données sensibles** au sens du RGPD, dont le traitement est en principe interdit sauf à bénéficier d'une dérogation spécifique (notamment le consentement explicite de la personne concernée) ; en outre, l'état de santé est un critère protégé en droit français (*cf.* section 1.1.3)[79]. Toutefois, en assurance, l'usage de ces données reste historiquement central pour l'évaluation du risque et pour la tarification (notamment en assurance emprunteur et en prévoyance). Leur utilisation est possible, mais elle doit donc **concilier** des **impératifs actuariels** légitimes avec des **exigences élevées de protection** des droits des personnes.

[78] Parfois uniquement le code postal.

[79] Notons également que ces données sont aussi protégées par des dispositifs spécifiques, comme les règles relatives au droit à l'oubli ou des conventions sectorielles (*cf.* section 1.1.3).

Encadré 7 : Faut-il collecter davantage de données sensibles pour détecter les biais ?

La détection des biais suppose, en pratique, de pouvoir les mesurer : par construction, **on ne peut pas détecter ce que l'on n'observe pas**. Dans cette perspective, la collecte de certaines données relatives à des critères protégés peut apparaître comme un levier utile pour évaluer l'équité de groupe des modèles. Cette approche se heurte traditionnellement à de fortes réticences, particulièrement en France, où l'usage de telles données est historiquement et juridiquement encadré de manière stricte.

Des **évolutions récentes** introduisent néanmoins des **marges de manœuvre ciblées**. En particulier, **le Règlement IA autorise, sous conditions strictes, le traitement de données sensibles lorsque celui-ci est nécessaire à la détection et à la correction des biais des systèmes d'IA**, notamment à haut risque (Article 10(5))[80].

Collecter davantage de données sensibles semble donc possible, dès lors que l'objectif est bien celui de la lutte contre les discriminations (et qu'il ne peut être atteint autrement), et que certaines **garanties** sont respectées, notamment en matière de sécurité, de limitation des accès, de non-réutilisation et d'interdiction de partage avec des tiers. Enfin, il doit être dûment justifié, documenté et limité dans le temps, les données devant être supprimées dès que leur utilisation n'est plus nécessaire.

5.3 Identification des biais : incertitude statistique, et analyse univariée ou multivariée

S'agissant d'abord de la **prise en compte de l'incertitude statistique**, **toute mesure du biais devrait être accompagnée d'une mesure de sa précision**[81] (*cf.* section 4.1). À ce titre, **trois éléments** devraient figurer de manière standard dans les analyses d'équité conduites par les établissements du secteur financier :

- Un **intervalle de confiance** pour chaque métrique reportée, avec mention de la méthode d'estimation ;
- Lorsque le modèle est entraîné en interne, une **indication de la variabilité des mesures** obtenues sur plusieurs entraînements (graines aléatoires distinctes) ;
- Pour chaque groupe sur lequel aucun écart significatif n'est détecté, une estimation de **l'amplitude minimale d'écart détectable** compte tenu de la taille d'échantillon disponible.

S'agissant ensuite des **groupes à comparer**, les ateliers ont montré que, chez les acteurs financiers, les analyses portaient **quasi-exclusivement sur des groupes définis par une seule variable sensible** (comme le genre). Au contraire, la littérature scientifique converge vers l'idée de comparer des groupes définis par le croisement de plusieurs variables (analyse multivariée), lorsque les données le permettent (*cf.* section 2.5).

Dans le secteur financier, les **difficultés pratiques de l'analyse multivariée n'apparaissent pas insurmontables pour la plupart des cas d'usage**. En effet, le nombre de variables sensibles collectées est généralement restreint (*cf.* l'encadré 6) : il s'agit donc, le plus souvent, de **croiser**

[80] Ce point fait actuellement l'objet de discussions dans le cadre du projet de « *Digital Omnibus* » de la Commission européenne, visant notamment à apporter des ajustements au Règlement IA.
[81] Ce principe est d'ailleurs cohérent avec les exigences générales de robustesse statistique applicables aux modèles internes du secteur financier.

deux ou trois dimensions, ce qui limite les risques d'explosion combinatoire et réduit la complexité de l'analyse.

Dès lors, la mise en œuvre d'analyses multivariées nécessite surtout de **définir l'effectif minimal** des groupes inclus dans l'analyse, afin de garantir la robustesse statistique des comparaisons. Sur des groupes de petite taille, la variance de l'estimateur des métriques d'équité augmente mécaniquement, et l'absence d'écart statistiquement significatif peut simplement refléter un **manque de puissance** du test de comparaison (*cf*. section 4.1).

La littérature scientifique recommande, en conséquence, de **déterminer l'effectif minimal des groupes au cas par cas**, en fonction des caractéristiques de la population étudiée, plutôt que d'appliquer des règles uniformes[82]. Lorsque certains groupes présentent un effectif insuffisant, il peut être nécessaire de procéder à des **regroupements ou d'ajuster la manière dont les variables continues sont « découpées » en catégories** (sur l'exemple de l'âge : en modifiant les tranches initialement définies)[83].

En définitive, si les analyses univariées constituent un socle minimal, les établissements financiers **sont encouragés à les prolonger par des approches multivariées**, afin d'identifier des biais potentiellement plus complexes et plus sévères.

5.4 Sur le choix des métriques

L'utilisation d'une métrique d'équité de groupe permet d'évaluer l'impact des décisions prises par un modèle d'IA sur différents groupes de la population. Il a été montré que trois grandes familles de métriques co-existaient dans la littérature scientifique, présentant chacune des avantages et des inconvénients (*cf.* section 3), mais qu'il était impossible de les satisfaire simultanément, **ce qui impose de faire des choix** (même s'il peut être utile d'examiner successivement les résultats obtenus selon plusieurs métriques d'équité, à des fins d'analyse sur le comportement du modèle).

Cette section vise à éclairer les choix de l'entreprise en la matière, étant entendu que, par principe, **aucune famille de métrique ne peut être privilégiée ou écartée dans tous les cas**, mais que les choix doivent dépendre du contexte sociotechnique du cas d'usage considéré, ainsi que de la politique générale de l'entreprise (*cf.* section 5.1).

Ces choix peuvent être guidés par **l'examen successif de trois questions**[84].

---

[82] Un effectif minimum de 30 individus est parfois mentionné comme une règle empirique (liée au théorème central limite selon lequel la distribution de la moyenne d'un échantillon – après recentrage et réduction – tend vers une loi normale lorsque la taille de l'échantillon augmente), mais ne constitue pas une garantie de validité. En effet, les conditions d'application de ce théorème (notamment l'indépendance des observations et l'existence d'une variance finie) ne sont pas toujours réunies en pratique. Plus généralement, la puissance d'un test de comparaison entre groupes dépend de plusieurs facteurs : l'ampleur de l'écart que l'on cherche à détecter, la variabilité des données, la taille effective des groupes et le niveau de significativité retenu.

[83] Une approche alternative, développée dans la littérature scientifique récente, consiste à rechercher de manière algorithmique les groupes (caractérisés par des combinaisons de variables) sur lesquels le modèle présente les plus disparités de résultat les plus marquées. Ce type d'approche permet d'identifier des biais sur des sous-populations qui n'auraient pas été spontanément examinées.

[84] Le choix d'une métrique d'équité peut être formalisé à l'aide d'arbres de décision. Même s'il est assez général, on peut par exemple citer celui développé par l'Université de Chicago (Saleiro et al., 2018).

1/ Existe-t-il une exigence règlementaire de parité démographique ?

Une telle exigence était par exemple traditionnellement présente dans le cadre règlementaire aux **États-Unis** (*cf.* section 1.5). En France et dans l'Union européenne, la question peut notamment se poser en matière de **tarification en assurance**. Depuis l'arrêt *Test-Achats* de la CJUE, les assureurs ne doivent plus différencier les primes et les prestations en fonction du genre (*cf*. section 1.3). Une analyse précise de cet arrêt montre toutefois que celui-ci **n'impose pas aux assureurs d'exigence de parité démographique entre hommes et femmes**[85], **mais proscrit l'utilisation du genre dans les modèles de tarification**. Autrement dit, cet arrêt pose une contrainte sur les variables d'entrée des modèles, et non sur les résultats produits.

Les assureurs doivent donc s'abstenir de recourir directement au genre ; en revanche, l'utilisation de **variables corrélées au genre** est autorisée dès lors qu'elle repose sur des facteurs de risque objectivement justifiés, proportionnés et pertinents au regard de l'évaluation actuarielle du risque. À ce titre, des variables telles que les caractéristiques du véhicule, le lieu de résidence ou encore le coefficient de bonus-malus peuvent, dans l'état actuel du droit, être prises en compte. **Il en résulte que des différences de primes entre hommes et femmes peuvent subsister en pratique.**

Plus généralement, notre analyse conduit à considérer qu'à ce jour, le cadre juridique français et européen **ne semble pas imposer d'exigence de parité démographique** dans les cas d'usage du secteur financier.

2/ Le modèle est-il appliqué à des groupes présentant des différences significatives ?

Les métriques **d'indépendance** (parité démographique) **ne sont pertinentes que lorsque les niveaux de risque et les distributions des variables explicatives sont relativement proches entre groupes**. En effet, elles imposent une égalité des décisions indépendamment des différences de risque sous-jacentes ; elles sont donc **mal adaptées aux situations dans lesquelles les différents groupes sociaux présentent des taux de base éloignés** (*cf*. section 3). De même, lorsque les distributions de fréquence des variables explicatives – ainsi que les scores qui en résultent – **diffèrent fortement** entre groupes, la recherche d'une parité démographique peut nécessiter des ajustements substantiels (notamment des seuils de décision), voire une altération du modèle. De tels ajustements peuvent conduire à une perte d'information et, *in fine*, à une moindre efficacité prédictive.

Dans ces situations, il est généralement préférable de recourir à des **critères de séparation ou de suffisance**, plus cohérents avec **l'hétérogénéité** des risques observés.

Dans le choix entre ces deux familles de métriques, l'analyse des caractéristiques des données constitue un premier facteur d'arbitrage. Ainsi, lorsque les **écarts de taux de base** entre groupes sont importants, le recours à la suffisance peut conduire à une sélection nettement plus stricte des groupes défavorisés. Dans ce contexte, **privilégier des métriques de séparation peut se justifier**, dans la mesure où elles garantissent un traitement comparable des individus présentant le même niveau réel de risque, et permettent ainsi de limiter les écarts d'accès aux décisions positives.

[85] Rappelons que dans un problème de régression comme la tarification, la parité démographique signifie que la distribution des tarifs doit être proche entre les groupes comparés ; une version moins stricte de la parité démographique requiert uniquement les mêmes tarifs moyens.

La prise en compte de la **distribution des variables explicatives** est **plus délicate**. Lorsque ces distributions diffèrent fortement entre groupes, les scores produits par le modèle tendent eux-mêmes à refléter ces écarts. Dans ce cas, la **suffisance apparaît souvent plus cohérente** avec une approche fondée sur le risque, puisqu'elle garantit que les décisions conservent une même signification probabiliste. À l'inverse, la séparation peut nécessiter des ajustements plus importants pour compenser ces différences.

Toutefois, **l'interprétation de ces écarts est déterminante**. La question centrale est alors : les différences observées reflètent-elles une hétérogénéité du risque que l'on souhaite effectivement prendre en compte dans la décision ? Lorsqu'elles correspondent à des **facteurs jugés pertinents et légitimes**, la suffisance permet de préserver cette information et d'en garantir une interprétation cohérente. En revanche, lorsque ces différences traduisent principalement des **biais historiques ou des inégalités structurelles** que l'on ne souhaite pas reproduire, il peut être préférable de recourir à des métriques de séparation, afin d'en limiter l'impact sur les décisions.

3/ Quelle forme de pertinence statistique est à privilégier, selon le cas d'usage et la politique de l'entreprise ?

La sélection de la famille de métrique la plus pertinente, entre séparation et suffisance dépend également du **type de performance statistique** que l'on souhaite privilégier. Plus précisément, le choix renvoie à un arbitrage fondamental entre **deux formes de fiabilité** : la **précision** (ou valeur prédictive positive), c'est-à-dire la probabilité qu'une prédiction positive[86] soit correcte, et le **rappel** (ou sensibilité), c'est-à-dire la capacité à identifier l'ensemble des individus réellement positifs. Privilégier la précision revient à s'assurer que les décisions prises sont **fiables**, tandis que privilégier le rappel consiste à **ne pas passer à côté des cas pertinents.**

Dans ce cadre, les métriques de **suffisance** (parité des valeurs prédictives) sont naturellement associées à la **précision** : elles garantissent que, parmi les individus pour lesquels une décision positive est prise, la proportion de cas effectivement positifs est comparable entre les groupes. À l'inverse, les métriques de **séparation** (parité des taux d'erreur) sont liées au **rappel** : elles visent à assurer que les individus réellement positifs ont des taux équivalents de classification positive quel que soit leur groupe d'appartenance.

Ces deux approches traduisent des **priorités différentes**. La suffisance met l'accent sur la **cohérence des décisions** : un même score ou une même décision doit correspondre au même niveau de risque pour tous les groupes. Il est donc particulièrement adapté aux situations où les erreurs de type **faux positif** (accorder un bénéfice à un individu non éligible) sont coûteuses ou socialement indésirables. À l'inverse, la séparation met l'accent sur **l'accès aux opportunités** : elle vise à éviter que certains groupes soient sous-représentés parmi les individus correctement identifiés, et se montre particulièrement sensible aux erreurs de type **faux négatif** (ne pas reconnaître un individu éligible).

Dans l'exemple de l'octroi de crédit, le recours à des métriques de suffisance répond à un objectif de **prudence** : il s'agit de limiter l'octroi de prêts à des emprunteurs susceptibles de faire défaut, afin de protéger à la fois l'établissement et les individus contre le surendettement. À l'inverse, le recours à des métriques de séparation répond à un objectif **d'inclusion** : il vise à s'assurer que les emprunteurs viables ne sont pas injustement exclus de l'accès au crédit, en raison d'erreurs

[86] On se place ici dans un cadre de classification binaire entre une classe positive et une classe négative.

du modèle[87]. Le choix entre ces deux approches revient ainsi à arbitrer entre une logique de **fiabilité des décisions positives d'octroi** et une logique de **non-exclusion des individus éligibles**.

Encadré 8 : De nécessaires compromis entre équité et performance ?

La question du compromis entre équité et performance – entendue ici comme l'exactitude des prédictions – constitue un **enjeu central, mais débattu,** dans la littérature scientifique.

Pour une partie des travaux, ce compromis est **inhérent,** c'est-à-dire qu'il n'est **pas possible de concevoir** des modèles qui satisfassent pleinement ces deux objectifs. Des différences de distribution entre groupes (par exemple des taux de défaut distincts) peuvent ainsi rendre certaines métriques d'équité difficilement compatibles avec une précision globale élevée. Par ailleurs, l'introduction de contraintes d'équité peut restreindre l'information exploitable par le modèle, ou accentuer les limites de données imparfaites (biaisées, bruitées ou incomplètes), ce qui peut conduire à une baisse de performance sans corriger pleinement les déséquilibres sous-jacents.

D'autres travaux relativisent le caractère systématique d'un arbitrage entre équité et performance. Ils mettent en avant le rôle déterminant des **choix de conception** : sélection des variables, qualité et représentativité des données, ou encore méthodes d'apprentissage mobilisées. Dans cette perspective, une meilleure prise en compte de la diversité des situations et l'usage de techniques adaptées peuvent permettre d'améliorer simultanément équité et performance, ou, à tout le moins, d'atténuer les tensions entre ces objectifs.

En pratique, la nature de ce compromis est déterminante. S'il est **structurel**, toute amélioration de l'équité implique un arbitrage explicite, qui relève alors de **choix normatifs** entre efficacité et inclusion. S'il est en partie **contingent**, il existe des **marges techniques** d'amélioration simultanée de l'équité et de la performance, qu'il convient d'explorer. Dans les deux cas, ces **arbitrages** ne peuvent être laissés aux seules équipes techniques : ils doivent être **explicités, documentés et intégrés** dans des dispositifs de gouvernance adaptés.

En l'absence de consensus, il apparaît enfin nécessaire d'adopter une **approche pragmatique**. Les établissements pourraient ainsi démontrer qu'ils ont activement recherché des solutions permettant d'améliorer conjointement équité et performance, notamment à travers la qualité et la traçabilité des données, des tests adaptés, et l'utilisation de plusieurs métriques d'évaluation, dans le cadre de processus renforcés de validation des modèles.

5.5 Sur les seuils à prendre en compte

Une fois la métrique pertinente déterminée, il reste à déterminer quel est le seuil permettant de caractériser une **différence de traitement problématique**.

Le seuil de 80 % – souvent désigné comme la « règle des quatre-cinquièmes » – appliqué à la métrique de parité démographique occupe une **place particulière dans le débat**, notamment aux États-Unis, où il est usuellement utilisé, y compris dans le secteur financier (*cf.* section 1.5).

[87] Ce qui fait écho aux objectifs du Règlement IA.

Ce seuil signifie que les différences de traitement entre deux groupes[88] ne doivent pas dépasser 20 % : ainsi, le nombre d'hommes ayant obtenu un crédit donné ne doit pas excéder celui des femmes de plus de 20 %. Ce seuil repose toutefois sur une **logique juridique essentiellement pragmatique**, visant à fournir un indicateur simple d'alerte, **plutôt que sur un fondement scientifique robuste**. **Rien ne permet en effet d'affirmer qu'un écart de 20 % constitue**, en général et indépendamment du contexte, **une frontière pertinente entre situations équitables et inéquitables**. En outre, l'application mécanique de ce seuil peut s'avérer **mal adaptée à certaines situations**, par exemple lorsque les taux de sélection de base sont très faibles ou très élevés[89] ou lorsque les effectifs diffèrent fortement entre groupes.

Plus généralement, **la littérature académique converge assez largement pour considérer qu'il n'existe pas de seuil universel**, scientifiquement fondé, permettant de qualifier une métrique d'équité de groupe comme « acceptable » ou « inacceptable » **de manière indépendante du contexte**. Les travaux de recherche soulignent en effet que, quelle que soit la métrique considérée, la pertinence d'un seuil dépend de nombreux facteurs : taille et structure des populations comparées, niveaux des taux de base, objectifs poursuivis par le modèle etc.

Sur les seuils à prendre en compte, **il convient par conséquent d'adopter une approche contextuelle** – fondée sur l'analyse de la magnitude des écarts entre groupes, de leur robustesse statistique et de leurs effets concrets, éventuellement complétée par des analyses de sensibilité[90] – **plutôt que d'appliquer mécaniquement des seuils immuables**.

5.6 Sur le choix des méthodes de correction des biais

Le choix d'une méthode appropriée de correction des biais dépend de plusieurs aspects de la tâche considérée : les causes et types identifiés de biais, le degré de contrôle sur le système d'IA, le niveau de contrainte réglementaire, etc. À ce sujet, la littérature scientifique semble suggérer qu'il est **préférable de combiner plusieurs de ces méthodes**. Nous détaillons ci-dessous quelques éléments d'aide à la décision.

- Si l'on est **susceptible d'utiliser plusieurs modèles différents sur un même jeu de données**, les méthodes de pré-traitement semblent plus adaptées, car elles vont généralement appliquer une transformation sur les données elles-mêmes. Ainsi, si l'on souhaite utiliser des modèles distincts d'évaluation de la solvabilité et de tarification pour l'octroi de crédit sur un jeu de données commun, il peut être pertinent d'utiliser du pré-traitement pour mutualiser les corrections de biais.
- Si l'on **maîtrise le processus d'entraînement du modèle**, les méthodes de traitement intégré à l'apprentissage peuvent être plus indiquées : en effet, intégrer une contrainte d'équité directement dans l'entraînement peut permettre d'atteindre un compromis optimal entre performance et équité, en particulier lorsque l'on se concentre sur une unique variable protégée.
- Si **l'on ne dispose que d'un modèle en boîte noire**, on peut toujours recourir à des méthodes de post-traitement, pour peu que l'on ait accès aux sorties numériques du

[88] Soit entre deux groupes définis de manière binaire (exemple : hommes et femmes), soit entre un groupe de référence (par exemple : les adultes de 20 à 45 ans) et d'autres groupes auxquels celui-ci est comparé.
[89] Dans le cas d'un crédit à la consommation dont le taux moyen d'acceptation serait de l'ordre de 95 %, les écarts entre groupes auraient peu de chance d'être supérieurs à 20 %.
[90] Visant à mesurer dans quelle mesure les conclusions varient lorsque l'on modifie certains paramètres ou choix méthodologiques (par exemple les groupes constitués ou le seuil retenu).

modèles (scores, probabilités, etc.). Ces méthodes permettent surtout d'ajuster les décisions finales à l'aide de seuils, même s'il est possible de recalibrer les sorties au prix d'une procédure généralement plus coûteuse.

Dans tous les cas, l'établissement doit **conserver une maîtrise pleine et entière du processus de correction** des biais. Quelle que soit la ou les méthode(s) retenue(s), il est indispensable, en premier lieu, de **documenter précisément** les modifications apportées au modèle et aux données. Par ailleurs, la littérature académique montre que de trop fortes corrections sur une dimension particulière peuvent, dans certains cas, dégrader l'équité sur d'autres axes (par exemple, une amélioration de l'équité selon le genre peut se faire au détriment de l'équité selon l'âge). Il est donc **nécessaire d'analyser les effets des corrections sur l'ensemble des dimensions pertinentes**, et de procéder à un arbitrage explicite entre les bénéfices attendus en matière d'équité et les exigences de **sobriété** et de **stabilité** de la modélisation. Ces arbitrages doivent être justifiés et tracés de manière transparente. Enfin, le processus de correction des biais **ne doit pas accroître l'opacité du modèle** : les choix opérés doivent rester compréhensibles et explicables, afin de ne pas affaiblir la confiance dans le système et sa gouvernance.

6 Anticiper l'essor de l'IA générative dans le secteur financier

Ce document de réflexion porte principalement sur les systèmes prédictifs « traditionnels ». En effet, ces systèmes représentent aujourd'hui l'essentiel des modèles déployés à grande échelle dans le secteur financier susceptibles de présenter des risques en matière d'équité. Cependant, les usages de l'IA générative[91] (modèles de langage, modèles multimodaux) connaissent un développement rapide. Or les méthodes d'évaluation des biais conçues pour les modèles prédictifs classiques ne se transposent pas directement à ces nouveaux systèmes, dont les modes de fonctionnement et les formes de sortie diffèrent sensiblement.

Pourquoi l'équité des systèmes génératifs est structurellement différente

Les notions de biais présentées dans ce document reposent sur **deux hypothèses** : l'existence d'une variable cible observable (par exemple le défaut ou la fraude), et d'une décision comparable entre groupes. Les systèmes génératifs s'écartent de ce cadre sur plusieurs points structurants :

- **Absence de cible univoque** : un système qui génère du texte ou des images ne prédit pas une valeur « vraie » unique. Il produit une réponse parmi de nombreuses possibles, si bien que la notion d'« erreur » ne se définit pas de la même manière que pour un modèle plus traditionnel.
- **Absence de groupe explicitement déclaré** : l'utilisateur n'indique généralement pas son appartenance à un groupe sensible. Toutefois, des disparités peuvent émerger à partir d'indices implicites, tels que la langue utilisée, le registre d'expression ou le contexte de la demande.
- **Des sorties de nature symbolique** : les sorties du modèle (textes, images, etc.) ne sont pas des décisions directement observables, mais des objets dont l'interprétation dépend du contexte et du destinataire. Une même réponse peut ainsi être perçue comme acceptable ou problématique selon les situations.
- **Un préjudice dépendant de l'usage** : les effets potentiellement biaisés ne tiennent pas uniquement au contenu produit, mais aussi à la manière dont il est utilisé. Par exemple, une réponse apparemment correcte peut véhiculer des stéréotypes implicites dont l'impact dépendra du contexte d'interaction et de la sensibilité du destinataire.

Ces spécificités ne remettent pas en cause la pertinence d'une analyse d'équité, mais elles en modifient profondément les modalités. L'évaluation ne peut pas se limiter à comparer des indicateurs entre groupes ; elle doit également porter sur la **représentation du monde encodée**

[91] Formellement, il conviendrait d'employer le terme « IA à usage général » (*General Purpose AI* ou GPAI), comme le fait le Règlement IA. Ce vocable désigne des systèmes conçus pour être employés dans une grande variété de tâches et de contextes, sans être limités à un cas d'usage spécifique. Au sens strict, l'IA « générative » constitue une catégorie particulière de ces systèmes : elle se caractérise par sa capacité à produire des contenus nouveaux (texte, image, code, etc.) à partir d'une invite.

par le modèle et sur les formes concrètes par lesquelles cette représentation s'exprime dans les contenus générés.

Quatre familles de préjudices spécifiques aux systèmes génératifs

Les modèles à usage général – et en particulier les grands modèles de langage (*large language models* ou LLM) – peuvent être à l'origine de plusieurs types de préjudices[92]. En premier lieu, des **préjudices représentationnels** apparaissent lorsque le système associe certains groupes à des rôles ou traits stéréotypés, par exemple dans des contenus marketing ou des interactions avec des agents conversationnels. Par ailleurs, des **disparités de qualité de service** peuvent émerger lorsque les performances varient selon la langue ou le registre d'expression, ce qui est susceptible d'affecter concrètement la relation client, notamment pour des publics ayant une maîtrise limitée de la langue utilisée.

En outre, ces systèmes peuvent conduire à un **effacement de la diversité**, en convergeant vers des profils implicites ou dominants dans les recommandations produites, au détriment de la variété des situations financières réelles. Enfin, des **préjudices allocatifs** peuvent apparaître en aval lorsque les contenus générés alimentent des décisions automatisées ou semi-automatisées (KYC, lutte anti-blanchiment, crédit). Dans ce cas, le préjudice, classique dans sa nature, est intermédié par un système dont le **fonctionnement reste largement opaque**, ce qui rend son identification plus difficile, et par là même d'autant plus cruciale.

Le biais : une propriété interne du modèle, pas seulement de ses réponses

Un enseignement important des travaux scientifiques récents sur les modèles de langage concerne la nature même des biais. Ceux-ci ne se limitent pas à certaines réponses visibles : ils sont **ancrés plus profondément**, dans la manière dont le modèle « organise » sa compréhension du monde. Lors de l'apprentissage, **le modèle construit en effet un espace de représentation à partir des corrélations présentes dans les données**. Dans cet espace, certaines dimensions correspondent à des caractéristiques sociales (par exemple le genre), et de nombreux mots ou concepts s'y répartissent selon des associations parfois stéréotypées. **Autrement dit, les biais ne sont pas des anomalies ponctuelles, mais des propriétés structurelles de cette représentation interne,** et ils ne disparaissent pas spontanément lorsque les modèles deviennent plus grands ou plus performants.

Les **techniques d' « alignement »** mises en œuvre après entraînement (comme l'apprentissage à partir de retours humains[93], l'ajustement des préférences[94] ou la fourniture d'instructions de sécurité) permettent de réduire l'expression visible de certains biais. Toutefois, **elles agissent principalement en surface** : elles influencent la forme des réponses produites, **sans transformer en profondeur la structure interne du modèle**[95]. En pratique, cela signifie qu'un

[92] Gallegos, et al., 2024.
[93] L'apprentissage par renforcement à partir de rétroaction humaine (*Reinforcement learning from human feedback* ou RLHF) consiste à utiliser des évaluations humaines pour orienter le comportement des modèles. Concrètement, des annotateurs humains comparent différentes réponses produites par le modèle, et ces préférences sont ensuite utilisées pour entraîner un mécanisme de récompense qui oriente le modèle vers des réponses jugées plus utiles, plus sûres ou mieux appropriées.
[94] L'optimisation des préférences directes *(Direct preference optimization* ou DPO) consiste à entraîner le modèle à reproduire directement les préférences humaines entre différentes réponses, sans recourir à un modèle de récompense intermédiaire, contrairement au RLHF.
[95] Wolf, Wies, Avnery, Levine, & Shashua, 2024.

modèle peut donner des réponses apparemment satisfaisantes dans des situations simples ou très encadrées, tout en laissant **réapparaître des biais dans des contextes plus complexes**, par exemple avec des questions longues, indirectes, multilingues ou formulées de manière inhabituelle. Par conséquent, une évaluation robuste doit explorer une diversité de contextes et de formulations, afin de mieux révéler les biais susceptibles d'apparaître en conditions réelles d'utilisation.

Trois couches d'évaluation qui peuvent être mobilisées conjointement

Plus généralement, l'évaluation de l'équité d'un système d'IA générative peut être organisée en trois couches complémentaires, afin de saisir les différentes dimensions du risque[96] :

- **La couche représentationnelle** vise à mesurer les biais directement dans la **structure interne** du modèle. Elle s'appuie sur des méthodes comme les tests d'association sur les vecteurs de représentation (*embeddings*)[97], les classifieurs de sondage (*probing classifiers*)[98] ou l'analyse des dimensions saillantes[99]. Cette couche permet d'identifier les biais présents dans les représentations initiales du modèle (*prior)*[100], indépendamment des mécanismes d'alignement. **Elle suppose toutefois un accès aux représentations internes**, généralement réservé aux modèles développés en interne ou aux modèles ouverts[101].
- **La couche comportementale** analyse les biais à partir des **sorties** du modèle, en testant des ensembles d'invites (*prompts*)[102] représentatifs des usages réels, incluant des

[96] Neumann, Kirsten, Zafar, & Singh, 2025.
[97] Un vecteur de représentation (*embedding*) est une manière de représenter un mot, une phrase ou un objet sous forme de nombres, afin que le modèle puisse les manipuler mathématiquement. Dans cet espace de représentation, des éléments jugés similaires (par exemple des mots proches en sens) se retrouvent proches les uns des autres. Ces représentations internes structurent la façon dont le modèle organise l'information et établit des associations.
[98] Un classifieur de sondage *(probing classifier*) est un modèle simple, entraîné sur les représentations internes d'un modèle d'IA (les *embeddings*), afin de tester quelles informations y sont présentes. Par exemple, on peut vérifier si ces représentations contiennent des informations sur le genre, l'âge ou d'autres caractéristiques. Cette méthode permet d'analyser ce que le modèle a appris, sans modifier son fonctionnement.
[99] L'analyse des dimensions saillantes consiste à identifier, dans l'espace de représentation du modèle, les directions (ou axes) qui capturent le plus de variation ou de structure dans les données. Certaines de ces dimensions peuvent correspondre à des caractéristiques interprétables, comme des catégories sémantiques ou sociales (par exemple le genre). Leur analyse permet de comprendre comment le modèle organise l'information et quelles distinctions il privilégie dans ses représentations internes.
[100] Les représentations initiales du modèle (*prior*) désignent l'ensemble des connaissances et des associations que le modèle a apprises lors de son entraînement, avant toute interaction avec l'utilisateur. Elles reflètent les régularités présentes dans les données d'apprentissage et structurent la manière dont le modèle interprète les requêtes et génère ses réponses.
[101] Lorsqu'un modèle à usage général est fourni par un tiers, l'établissement utilisateur ne dispose généralement pas d'un accès direct aux représentations internes du modèle. Il peut en revanche s'appuyer sur la documentation technique que le fournisseur est tenu de mettre à disposition au titre du Règlement sur l'IA. Cette documentation doit notamment décrire, dans une certaine mesure, les données d'entraînement, les caractéristiques du modèle, ainsi que les méthodes mises en œuvre pour identifier et atténuer les biais (Annexe XI du Règlement).
[102] Une invite (*prompt*) est l'instruction ou la requête adressée à un modèle d'IA générative (question, consigne, texte à compléter, etc.), qui sert de point de départ à la réponse produite. La manière dont une

formulations variées (invites longues, multilingues ou antagonistes[103]). Elle permet **d'évaluer ce que les utilisateurs observent effectivement en pratique**. En revanche, elle peut **sous-estimer des biais latents qui ne sont pas activés par les scénarios testés**.

- **La couche allocative** examine les biais lorsque les sorties du modèle alimentent une **décision en aval**. Elle consiste à appliquer les **outils classiques** d'évaluation de l'équité (indépendance, séparation, suffisance) à l'ensemble du processus, en intégrant le modèle génératif comme une étape intermédiaire.

Ces trois couches apportent des éclairages complémentaires. Se limiter à la seule couche comportementale, souvent la plus facilement accessible, conduit à négliger les biais structurels du modèle : un système peut apparaître satisfaisant dans les tests tout en restant porteur de biais latents. À l'inverse, se concentrer uniquement sur la couche représentationnelle ne permet pas de déterminer si les biais se traduisent effectivement en impacts concrets[104].

Une évaluation complète peut donc consister à **articuler ces trois niveaux**, avec un degré d'approfondissement **proportionné** au niveau de risque du cas d'usage, conformément à **l'approche par les risques** retenue dans l'ensemble de ce document.

invite est formulée – plus ou moins précise, longue, directe ou structurée – peut influencer significativement le contenu et la qualité de la réponse du modèle.

[103] Ou *adversarial prompting* : invites formulées de manière à tester les limites du modèle, voire à provoquer des erreurs ou à contourner ses garde-fous. Il peut s'agir, par exemple, de questions ambiguës, indirectes ou volontairement trompeuses. Ces tests permettent d'évaluer la robustesse du système face à des usages non standards ou malveillants.

[104] Une pratique en développement consiste à utiliser un autre modèle de langage comme « juge » pour évaluer les sorties d'un autre système (« *LLM-as-a-judge* »). Cette approche présente des avantages opérationnelles (automatisation, rapidité), mais elle paraît mal adaptée à l'évaluation des biais. En effet, le modèle juge peut partager les mêmes biais de représentation que le système évalué, et présente lui-même des biais propres, ce qui peut limiter sa capacité à détecter des inégalités. Surtout, l'appréciation des biais et des risques de discrimination repose sur un jugement contextuel qui ne peut être entièrement automatisé.

Bibliographie

AEAPP. (2025). *Opinion on AI Governance and Risk Management*. Consulté sur https://www.eiopa.europa.eu/document/download/88342342-a17f-4f88-842f-bf62c93012d6_en

Alvarez, J. M., Bringas Colmenarejo, A., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., . . . Ruggieri, S. (2024). Policy Advice and Best Practices on Bias and Fairness in AI. *Ethics and Information Technology, 26*(31). doi:10.1007/s10676-024-09746-w

Bachoc, F., Bolte, J., Boustany, R., & Loubes, J.-M. (2026). When Majority Rules, Minority Loses: Bias Amplification of Gradient Descent. *Advances in Neural Information Processing Systems, 38*, 30479–30517. Consulté sur https://proceedings.neurips.cc/paper_files/paper/2025/hash/2bbc73b3d3c2de43743ce2d82c8f3d7d-Abstract-Conference.html

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. https://fairmlbook.org/.

Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., & Risser, L. (2022). A Survey of Bias in Machine Learning Through the Prism of Statistical Parity. *The American Statistician, 76*(2), 188–198. doi:10.1080/00031305.2021.1952897

Binns, R. (2022). On the Apparent Conflict Between Individual and Group Fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524. doi:10.1145/3351095.3372864

Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys, 56*(7), 1–38. doi:10.1145/3616865

Charpentier, A., & Barry, L. (2022, Décembre). L'équité de l'apprentissage machine en assurance. *Statistique et Société, vol. 10, n° 3*.

Cook, D. I., Gebski, V. J., & Keech, A. C. (2004). Subgroup Analysis in Clinical Trials. *The Medical Journal of Australia, 180*(6), 289–291. doi:10.5694/j.1326-5377.2004.tb05928.x

Côté, M.-P., Côté, O., & Charpentier, A. (2024). Selection Bias in Insurance: Why Portfolio-Specific Fairness Fails to Extend Market-Wide. *SSRN*. doi:10.2139/ssrn.5018749

Das Jui, T., & Rivas, P. (2024). Fairness Issues, Current Approaches, and Challenges in Machine Learning Models. *International Journal of Machine Learning and Cybernetics, 15*, 3095–3125. doi:10.1007/s13042-023-02083-2

Deck, L., Müller, J.-L., Braun, C., Zipperling, D., & Kühl, N. (2024). Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness. *European Workshop on Algorithmic Fairness*. Consulté sur https://ceur-ws.org/Vol-3908/paper_39.pdf

Desrosières, A. (1993). *La Politique des grands nombres*. Paris: La Découverte.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemer, R. (2012). Fairness Through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., & Beben, S. (2021). Reject Inference Methods in Credit Scoring. *Journal of Applied Statistics, 48*(13–15), 2734–2754. doi:10.1080/02664763.2021.1929090

Estellat, C., De Rycke, Y., & Asselain, B. (2005). Intérêt et limites des analyses en sous-groupes dans les essais thérapeutiques. *Oncologie, 7*, 75–79. doi:10.1007/s10269-005-0298-6

Ewald, F. (2011). Omnes et Singulatim. After Risk. *Carceral Notebooks, 7*, 77–107.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M., Kim, S., Dernoncourt, F., . . . Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics, 50*(3), 1097–1179. doi:10.1162/coli_a_00524

Glenn, B. J. (2000). The Shifting Rhetoric of Insurance Denial. *Law & Society Review, 34*(3), 779–808. doi:10.2307/3115143

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing*, 1–52. doi:10.1145/3631326

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. doi:10.1145/3442188.3445901

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *Proceedings of the 35th International Conference on Machine Learning*, 2564–2572.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4069–4079.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, Mai 23). *How We Analyzed the COMPAS Recidivism Algorithm*. Consulté sur ProPublica: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Li, J., & Li, G. (2025). Triangular Trade-off between Robustness, Accuracy, and Fairness in Deep Neural Networks: A Survey. *ACM Computing Surveys, 57*(6), 1–40. doi:10.1145/364508

Loubes, J.-M., Clayes, E., Eynard, J., Lafargue, V., Rottembourg, B., & Prunkl, C. (2026). A Hitchhiker's Guide to Bias Evaluation. *Preprint: hal-05642033v1*.

Meding, K. (2026). It's Complicated. The Relationship of Algorithmic Fairness and Non-Discrimination Provisions for High-Risk Systems in the EU AI Act. *Workshop on Regulatable ML*. doi:10.48550/arXiv.2501.12962

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR), 54*(6), 1–25. doi:10.1145/3457607

Napoletani, D., Panza, M., & Struppa, D. C. (2011). Agnostic Science. Towards a Philosophy of Data Analysis. (Springer, Ed.) *Foundations of Science, 16*(1), 1-20. doi:10.1007/s10699-010-9186-7

Neumann, A., Kirsten, E., Zafar, M. B., & Singh, J. (2025). Position is Power: System Prompts as a Mechanism of Bias in Large Language Models (LLMs). *Proceedings of the 2025 ACM*

Conference on Fairness, Accountability, and Transparency, 573–598. doi:10.1145/3715275.3732038

OCDE. (2026). Supervision of Artificial Intelligence in Finance: Challenges, Policies and Practices. *OECD Artificial Intelligence Papers*(54). doi:10.1787/92743dc1-en

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR), 55*(3), 1–44. doi:10.1145/3494672

Rosenblatt, L., & Witter, R. T. (2023). Counterfactual Fairness Is Basically Demographic Parity. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(12), 14461–14469. doi:10.1609/aaai.v37i12.26691

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., . . . Ghani, R. (2026). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv*(18111.05577). doi:10.48550/arXiv.1811.05577

Simon, J. (1988). The Ideological Effects of Actuarial Practices. *Law & Society Review, 22*(4), 771–800. doi:10.2307/3053709

Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. *FairWare '18: Proceedings of the International Workshop on Software Fairness*, 1–7. doi:10.1145/3194770.3194776

Westerstrand, S. (2025). Fairness in AI Systems Development: EU AI Act Compliance and Beyond. *Information and Software Technology, 187*(107864). doi:10.1016/j.infsof.2025.107864

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy, 8*, 78–115. doi:10.5325/jinfopoli.8.2018.0078

Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2024). Fundamental Limitations of Alignment in Large Language Models. *Proceedings of the 41st International Conference on Machine Learning, 235*, 53079–53112. Consulté sur https://dl.acm.org/doi/abs/10.5555/3692070.3694246