Confrontation de la grille d’analyse

Critères ALTAIL vs Critère cadre théorique

Collecte de données ciblées

Critères ALTAIL	Critères		
	Approche réglementaire	Approche éthique by design	Approche interactionnelle
Analyse d’impact sur les droits fondamentaux			
Documentation des arbitrages éthiques			
Interaction avec la prise de décision humaine			
Préservation de l’autonomie décisionnelle			

Signalement des décisions algorithmiques			✅ Communication sur les biais
Identification des agents virtuels			
Répartition des tâches entre IA et humains			
Renforcement des capacités humaines			
Prévention de la dépendance excessive			
Définition du niveau de contrôle humain	✅ Contrôle humain : Interface permettant une supervision humaine continue et une intervention manuelle en cas de besoin.		
Identification de l'acteur humain responsable			
Mécanismes de supervision humaine	✅ Contrôle humain : Interface permettant une supervision humaine continue et une intervention manuelle en cas de besoin.		
Auditabilité et gouvernance de l'autonomie			
Contrôle renforcé pour les systèmes autonomes ou auto-apprenants			

Détection des anomalies et réponse aux risques			
Présence d'un mécanisme d'arrêt d'urgence			
Évaluation des vulnérabilités potentielles			
Identification des types de vulnérabilités			
Mesures de résilience et d'intégrité			
Comportement en environnement imprévu			
Gestion du double usage			
Existence de solutions de secours			
Évaluation du niveau de risque contextuel			
Processus d'évaluation des risques et de la sécurité			
Communication des risques pour l'intégrité physique			
Politique d'assurance			

Plan de gestion des usages abusifs ou détournés			
Évaluation des dommages potentiels			
Règles de responsabilité et protection des consommateurs			
Évaluation des impacts environnementaux et sur les animaux			
Analyse des menaces de cybersécurité			
Évaluation des défaillances critiques			
Définition de seuils de déclenchement			
Test des solutions de secours			
Définition du niveau de précision attendu			
Méthodes de mesure de la précision			
Qualité et actualisation des données			
Évaluation des besoins en données supplémentaires			
Évaluation des préjudices liés aux erreurs			

Suivi du taux d'erreurs			
Plan de correction des erreurs			
Stratégie de contrôle du système			☑ Contrôle utilisateur
Prise en compte des contextes spécifiques			
Méthodes de vérification de la fiabilité et de la reproductibilité			
Documentation des réglages critiques			
Application des processus de test			
Communication de la fiabilité aux utilisateurs			
Mécanisme de signalement des atteintes à la vie privée			
Évaluation du type et de la portée des données			
Minimisation de l'usage de données sensibles			
Mécanismes de consentement et de contrôle			
Mesures de protection des données			
Mobilisation du DPO (Data Protection Officer)			

Alignement sur les normes reconnues			✅ Documentation technique intégrée
Mécanismes de contrôle des données			
Évaluation de la qualité des sources externes			
Garantie de la qualité et de l'intégrité des données			
Vérification de la sécurité des ensembles de données			
Documentation des protocoles de gouvernance des données			
Analyse des accès aux données			
Qualification et légitimité des personnes ayant accès aux données			
Traçabilité des accès aux données	✅ Journalisation des événements : Enregistrement systématique des événements critiques pour assurer la traçabilité.		
Mise en place de mesures de traçabilité			

Documentation des méthodes de conception	✅ Documentation technique : 1 : Description des algorithmes, sources de données, méthodes de traitement (ex. : anonymisation, réduction des biais). 2 : Preuves des conformité aux normes de sécurité et d'éthique. 3 : Plans d'atténuation des risques et d'intervention en cas de défaillance. 4 : Mécanismes d'audit et d'explicabilité.		
Documentation de l'entraînement des modèles (IA fondée sur l'apprentissage)			
Documentation des méthodes de test et de validation			✅ Documentation technique intégrée
Documentation des résultats et des décisions	✅ Transparence envers les déployeurs : Fourniture d'une notice d'information claire sur le fonctionnement, les risques, les limites et les conditions d'utilisation du système. ✅ Documentation technique : 1 : Description des algorithmes, sources de données, méthodes de traitement (ex. : anonymisation, réduction des biais). 2 : Preuves des conformité aux normes de sécurité et d'éthique. 3 : Plans d'atténuation des risques et d'intervention en cas de défaillance. 4 : Mécanismes d'audit et d'explicabilité.		
Compréhensibilité des décisions du système			✅ Compréhension utilisateur
Explication des choix algorithmiques			✅ Explication contextualisée
Évaluation de l'influence sur les processus décisionnels			

Justification du déploiement dans un domaine spécifique			
Évaluation du modèle économique			
Conception orientée vers l'interprétabilité		✅ Design centré utilisateur (VSD) : 1 : Intégration d'experts en valeurs dans les équipes de conception. 2 : Prise en compte du contexte cognitif et social d'usage. 3 : Conception d'interfaces favorisant la compréhension et l'appropriation.	✅ Explicabilité graduée
Choix de modèles interprétables			✅ Transparence située
Analyse des données d'entraînement et de test			
Accès à la logique interne du modèle			
Identification claire du système comme étant une IA			
Explication des critères de décision			
Accessibilité de l'information pour tous les profils d'utilisateurs		✅ Design centré utilisateur (VSD) : 1 : Intégration d'experts en valeurs dans les équipes de conception. 2 : Prise en compte du contexte cognitif et social d'usage. 3 : Conception d'interfaces favorisant la compréhension et l'appropriation.	✅ Accessibilité cognitive
Prise en compte des retours utilisateurs			✅ Recours utilisateur
Communication des risques et biais			✅ Visualisation pédagogique
Communication élargie selon les publics			
Indication de la finalité du système			

Définition et explication des scénarios d'usage			
Prise en compte des limites cognitives humaines			
Communication différenciée selon les rôles	✅ Transparence envers les déployeurs : Fourniture d'une notice d'information claire sur le fonctionnement, les risques, les limites et les conditions d'utilisation du système.		✅ Feedback formateur
Stratégie de prévention des biais			
Évaluation des limites des ensembles de données			
Prise en compte de la diversité des utilisateurs			
Utilisation d'outils d'analyse des biais			
Processus de détection et de contrôle des biais			
Mécanisme de signalement des biais			
Communication sur les biais et les recours			
Prise en compte des effets indirects		✅ Responsabilité sociale : 1 : Anticipation des médiations (morales, comportementales, cognitives).	

		2 : Analyse des effets négatifs (biais, perte d'autonomie, surveillance). 3 : intégration de médiations souhaitables (durabilité, surveillance). 4 : intégration de médiations souhaitables (durabilité, transparence). 5 : Partage de la responsabilité entre concepteurs, utilisateurs et systèmes.	
Évaluation de la variabilité des décisions			
Analyse des causes de variabilité			
Évaluation de l'impact de la variabilité sur les droits fondamentaux			
Définition de l'équité appliquée			
Mesure de l'équité			
Mécanismes de garantie de l'équité			
Adaptation aux préférences et capacités individuelles			
Prise en compte des besoins spécifiques et du handicap			
Compatibilité avec les technologies d'assistance			
Consultation des communautés concernées			

Évaluation de l'impact sur les groupes d'utilisateurs			
Diversité de l'équipe de conception			
Évaluation des effets disproportionnés		✅ Responsabilité sociale : 1 : Anticipation des médiations (morales, comportementales, cognitives). 2 : Analyse des effets négatifs (biais, perte d'autonomie, surveillance). 3 : intégration de médiations souhaitables (durabilité, surveillance). 4 : intégration de médiations souhaitables (durabilité, transparence). 5 : Partage de la responsabilité entre concepteurs, utilisateurs et systèmes.	
Prise en compte des retours de groupes externes			
Participation des parties prenantes		✅ Gouvernance éthique : 1 : Cartographie des parties prenantes. 2 : Elicitation des valeurs et traduction en exigences techniques. 3 : Evaluation continue via audits et retours d'expérience	

Mobilisation des acteurs internes			
Mesure de l'impact environnemental			
Réduction de l'empreinte écologique			
Évaluation de l'attachement émotionnel		✅ Responsabilité sociale : 1 : Anticipation des médiations (morales, comportementales, cognitives). 2 : Analyse des effets négatifs (biais, perte d'autonomie, surveillance). 3 : intégration de médiations souhaitables (durabilité, surveillance). 4 : intégration de médiations souhaitables (durabilité, transparence). 5 : Partage de la responsabilité entre concepteurs, utilisateurs et systèmes.	
Clarté sur la nature simulée de l'interaction			
Analyse des incidences sociales		✅ Responsabilité sociale : 1 : Anticipation des médiations (morales, comportementales, cognitives). 2 : Analyse des effets négatifs (biais, perte d'autonomie, surveillance). 3 : intégration de médiations souhaitables (durabilité, surveillance). 4 : intégration de médiations souhaitables (durabilité, transparence).	

		✅ Partage de la responsabilité entre concepteurs, utilisateurs et systèmes.	
Évaluation de l'impact sociétal élargi			
Auditabilité du système	✅ journalisation des événements : Enregistrement systématique des événements critiques pour assurer la traçabilité.		
Analyse des risques multi-acteurs			
Mise en place de formations à la responsabilité			
Identification des publics concernés par la formation			
Intégration du cadre juridique dans les formations			
Mise en place d'un comité d'éthique ou d'un mécanisme équivalent		✅ Gouvernance éthique : 1 : Cartographie des parties prenantes. 2 : Elicitation des valeurs et traduction en exigences techniques. 3 : Evaluation continue via audits et retours d'expérience	
Recours à des audits ou orientations externes			
Mécanisme de signalement des vulnérabilités			
Mécanisme de recensement des valeurs et intérêts		✅ Méthodologie éthique intégrée : 1 : Identification explicite des valeurs humaines fondamentales. 2 : Co-construction de solutions techniques respectueuses de ces valeurs. 3 : Acceptation de compromis fonctionnels pour respecter l'éthique.	
Processus de décision pour les arbitrages		✅ Méthodologie éthique intégrée : 1 : Identification explicite des valeurs humaines fondamentales. 2 : Co-construction de solutions techniques respectueuses de ces valeurs. 3 : Acceptation de compromis fonctionnels pour respecter l'éthique.	

Documentation des arbitrages	☑ Documentation technique : 1 : Description des algorithmes, sources de données, méthodes de traitement (ex. : anonymisation, réduction des biais). 2 : Preuves des conformité aux normes de sécurité et d'éthique. 3 : Plans d'atténuation des risques et d'intervention en cas de défaillance. 4 : Mécanismes d'audit et d'explicabilité.		☑ Ethique by design
Mécanismes de recours en cas de préjudice			
Information sur les voies de recours			

Nous n'avons rien trouver concernant l'inscription au registre public européen, donc nous n'avons pas pu l'intégrer dans la grille.

Résultats

Notre grille d'analyse refléter 24 critères de transparence sur 129 critères aux total. Après confrontation avec notre grille d'analyse, nous constations que les critères de transparence serait de 27 mais qu'elle serait diviser dans toutes les dimensions proposé par l'outil ALTAI et pas seulement celle définit dans la dimension de transparence (traçabilité, explicabilité, communication). Dans le tableau suivant, ont peut voir les critères concordants avec notre grille d'analyse, ceux étant non concordant ainsi que ceux relative à la dimension de transparence de l'outil ALTAI mais que nous n'avons pas retrouver.

Critères concordants (27 critères)	Critères non concordants (hors transparence)	Critères non concordants (transparence)
Signalement des décisions algorithmiques	Analyse d'impact sur les droits fondamentaux	Mise en place de mesures de traçabilité
Définition du niveau de contrôle humain	Documentation des arbitrages éthiques	Documentation de l'entraînement des modèles (IA fondée sur l'apprentissage)
Mécanismes de supervision humaine	Interaction avec la prise de décision humaine	Évaluation de l'influence sur les processus décisionnels
Stratégie de contrôle du système	Préservation de l'autonomie décisionnelle	Justification du déploiement dans un domaine spécifique
Alignement sur les normes reconnues	Identification des agents virtuels	Évaluation du modèle économique
Traçabilité des accès aux données	Répartition des tâches entre IA et humains	Analyse des données d'entraînement et de test
Documentation des méthodes de conception	Renforcement des capacités humaines	Accès à la logique interne du modèle
Documentation des méthodes de test et de validation	Prévention de la dépendance excessive	Identification claire du système comme étant une IA
Documentation des résultats et des décisions	Identification de l'acteur humain responsable	Explication des critères de décision
Compréhensibilité des décisions du système	Auditabilité et gouvernance de l'autonomie	Communication élargie selon les publics
Explication des choix algorithmiques	Contrôle renforcé pour les systèmes autonomes ou auto-apprenants	Indication de la finalité du système
Conception orientée vers l'interprétabilité	Détection des anomalies et réponse aux risques	Définition et explication des scénarios d'usage
Choix de modèles interprétables	Présence d'un mécanisme d'arrêt d'urgence	Prise en compte des limites cognitives humaines
Accessibilité de l'information pour tous les profils d'utilisateurs	Évaluation des vulnérabilités potentielles	Évaluation de l'impact sociétal élargi
Prise en compte des retours utilisateurs	Identification des types de vulnérabilités	
Communication des risques et biais	Mesures de résilience et d'intégrité	
Communication différenciée selon les rôles	Comportement en environnement imprévu	
Prise en compte des effets indirects	Gestion du double usage	
Évaluation des effets disproportionnés	Existence de solutions de secours	
Participation des parties prenantes	Évaluation du niveau de risque contextuel	
Évaluation de l'attachement émotionnel	Processus d'évaluation des risques et de la sécurité	
Analyse des incidences sociales	Communication des risques pour l'intégrité physique	
Auditabilité du système	Politique d'assurance	
Mise en place d'un comité d'éthique ou d'un mécanisme équivalent	Plan de gestion des usages abusifs ou détournés	
Mécanisme de recensement des valeurs et intérêts	Évaluation des dommages potentiels	
Processus de décision pour les arbitrages	Règles de responsabilité et protection des consommateurs	
Documentation des arbitrages	Évaluation des impacts environnementaux et sur les animaux	
	Analyse des menaces de cybersécurité	
	Évaluation des défaillances critiques	
	Définition de seuils de déclenchement	
	Test des solutions de secours	
	Définition du niveau de précision attendu	
	Méthodes de mesure de la précision	
	Qualité et actualisation des données	
	Évaluation des besoins en données supplémentaires	
	Évaluation des préjudices liés aux erreurs	
	Suivi du taux d'erreurs	
	Plan de correction des erreurs	

	Prise en compte des contextes spécifiques Méthodes de vérification de la fiabilité et de la reproductibilité Documentation des réglages critiques Application des processus de test Communication de la fiabilité aux utilisateurs Mécanisme de signalement des atteintes à la vie privée Évaluation du type et de la portée des données Minimisation de l'usage de données sensibles Mécanismes de consentement et de contrôle Mesures de protection des données Mobilisation du DPO (Data Protection Officer) Mécanismes de contrôle des données Évaluation de la qualité des sources externes Garantie de la qualité et de l'intégrité des données Vérification de la sécurité des ensembles de données Documentation des protocoles de gouvernance des données Analyse des accès aux données Qualification et légitimité des personnes ayant accès aux données Stratégie de prévention des biais Évaluation des limites des ensembles de données Prise en compte de la diversité des utilisateurs Utilisation d'outils d'analyse des biais Processus de détection et de contrôle des biais Mécanisme de signalement des biais Communication sur les biais et les recours Évaluation de la variabilité des décisions Analyse des causes de variabilité Évaluation de l'impact de la variabilité sur les droits fondamentaux Définition de l'équité appliquée Mesure de l'équité Mécanismes de garantie de l'équité Adaptation aux préférences et capacités individuelles Prise en compte des besoins spécifiques et du handicap Compatibilité avec les technologies d'assistance Consultation des communautés concernées Évaluation de l'impact sur les groupes d'utilisateurs Diversité de l'équipe de conception Prise en compte des retours de groupes externes Mobilisation des acteurs internes Mesure de l'impact environnemental Réduction de l'empreinte écologique Clarté sur la nature simulée de l'interaction Analyse des risques multi-acteurs Mise en place de formations à la responsabilité Identification des publics concernés par la formation Intégration du cadre juridique dans les formations	

	Recours à des audits ou orientations externes Mécanisme de signalement des vulnérabilités Mécanismes de recours en cas de préjudice Information sur les voies de recours	

Critères ALTAIL vs critères de transparence des entreprises

Collecte de données ciblées

128 critères	Entreprises Salesforce	Entreprise Grammarly	Entreprise Turnitin	Copilot (lien)	Gemini (lien 1) (lien 2)	Claude AI (Anthropic) (lien)
Analyse d'impact sur les droits fondamentaux						
Documentation des arbitrages éthiques						
Interaction avec la prise de décision humaine		✅ Contrôle utilisateur sur les suggestions : « Les utilisateurs peuvent accepter ou refuser les suggestions, ajuster les types de suggestions, et reçoivent des explications pour prendre des décisions éclairées ».	✅ L'IA doit soutenir les objectifs pédagogiques et protéger l'intégrité académique : « L'IA est conçue pour renforcer l'apprentissage, non pour le remplacer ou le biaiser ».			
Préservation de l'autonomie décisionnelle		✅ Contrôle utilisateur sur les suggestions : « Les utilisateurs peuvent accepter ou refuser les suggestions, ajuster les types de suggestions, et reçoivent des explications pour prendre des décisions éclairées ».				

Signalement des décisions algorithmiques		✅ Sensibilisation des utilisateurs : « Les utilisateurs doivent savoir clairement quand ils interagissent avec l'IA, identifier le contenu généré par l'IA, et comprendre comment les décisions sont prises ».				
Identification des agents virtuels						
Répartition des tâches entre IA et humains						
Renforcement des capacités humaines			✅ L'IA doit soutenir les objectifs pédagogiques et protéger l'intégrité académique : « L'IA est conçue pour renforcer l'apprentissage, non pour le remplacer ou le biaiser ».			
Prévention de la dépendance excessive						
Définition du niveau de contrôle humain						
Identification de l'acteur humain responsable						
Mécanismes de supervision humaine						
Auditabilité et gouvernance de l'autonomie						

Contrôle renforcé pour les systèmes autonomes ou auto-apprenants						
Détection des anomalies et réponse aux risques						
Présence d'un mécanisme d'arrêt d'urgence					✅ Sécurité des données et stockage à froid : « Les clés privées sont stockées hors ligne dans des modules de sécurité matériels, jamais connectés à Internet, avec des contrôles d'accès biométriques »	
Évaluation des vulnérabilités potentielles						✅ Sécurité, auditabilité et conformité : « Anthropic décrit les méthodes de test, les résultats de sécurité, et les engagements de conformité (ex. : ASL-2)
Identification des types de vulnérabilités						
Mesures de résilience et d'intégrité						
Comportement en environnement imprévu						
Gestion du double usage						

Existence de solutions de secours						
Évaluation du niveau de risque contextuel						
Processus d'évaluation des risques et de la sécurité						
Communication des risques pour l'intégrité physique						
Politique d'assurance					✅ Règlementation et responsabilité : « Gemini est un dépositaire fiduciaire qualifié, réglementé par l'Etat de new York, avec une couverture d'assurance pour les pertes liées aux cryptomonnaies »	
Plan de gestion des usages abusifs ou détournés						
Évaluation des dommages potentiels						
Règles de responsabilité et protection des consommateurs					✅ Règlementation et responsabilité : « Gemini est un dépositaire fiduciaire qualifié, réglementé par l'Etat de new York, avec une couverture d'assurance pour les pertes liées aux cryptomonnaies »	
Évaluation des impacts environnementaux et sur les animaux						

Analyse des menaces de cybersécurité						
Évaluation des défaillances critiques						
Définition de seuils de déclenchement						
Test des solutions de secours						
Définition du niveau de précision attendu						
Méthodes de mesure de la précision						
Qualité et actualisation des données						
Évaluation des besoins en données supplémentaires						
Évaluation des préjudices liés aux erreurs						
Suivi du taux d'erreurs						
Plan de correction des erreurs						
Stratégie de contrôle du système						
Prise en compte des contextes spécifiques				✅ Importance du contexte d'utilisation (technologie, personnes, environnement) : « Microsoft souligne que l'IA ne se		

				limite pas à la technologie, mais inclut les utilisateurs et leur environnement ».		
Méthodes de vérification de la fiabilité et de la reproductibilité			✅ Respect de critères rigoureux d'analyse statistique et d'apprentissage automatique : « L'IA doit être fondée sur des méthodes robustes et vérifiables »			
Documentation des réglages critiques						
Application des processus de test			✅ Respect de critères rigoureux d'analyse statistique et d'apprentissage automatique : « L'IA doit être fondée sur des méthodes robustes et vérifiables »			
Communication de la fiabilité aux utilisateurs						
Mécanisme de signalement des atteintes à la vie privée						
Évaluation du type et de la portée des données		✅ Documentation détaillée : « Les capacités, limites, données utilisées, méthodologies, risques et directives doivent être documentées. Les utilisateurs doivent savoir comment les données sont utilisées ».				✅ Transparence sur les données d'entraînement : « Claude n'est pas entraîné sur les données des utilisateurs. Les données proviennent de sources publiques, de partenaires tiers, ou sont générées en internes »
Minimisation de l'usage de données sensibles						✅ Transparence sur les données d'entraînement : « Claude n'est pas entraîné sur les données des utilisateurs. Les données proviennent de sources publiques, de partenaires

						tiers, ou sont générées en internes »
Mécanismes de consentement et de contrôle	✅ « Les données et les modèles de nos clients leur appartiennent, c'est pourquoi nous nous assurons qu'ils en gardent toujours le contrôle »	✅ Documentation détaillée : « Les capacités, limites, données utilisées, méthodologies, risques et directives doivent être documentées. Les utilisateurs doivent savoir comment les données sont utilisées ».	✅ Conception selon des critères exigeants de confidentialité et de propriété des données : « Les données des utilisateurs doivent être protégées et rester sous leur contrôle »	✅ Traitement responsable des données : « Microsoft décrit les étapes de traitement des requêtes, incluant des vérifications de conformité à l'IA responsable »		
Mesures de protection des données		✅ Documentation détaillée : « Les capacités, limites, données utilisées, méthodologies, risques et directives doivent être documentées. Les utilisateurs doivent savoir comment les données sont utilisées ».	✅ Conception selon des critères exigeants de confidentialité et de propriété des données : « Les données des utilisateurs doivent être protégées et rester sous leur contrôle »	✅ Traitement responsable des données : « Microsoft décrit les étapes de traitement des requêtes, incluant des vérifications de conformité à l'IA responsable »	✅ Sécurité des données et stockage à froid : « Les clés privées sont stockées hors ligne dans des modules de sécurité matériels, jamais connectés à Internet, avec des contrôles d'accès biométriques »	✅ Transparence sur les données d'entraînement : « Claude n'est pas entraîné sur les données des utilisateurs. Les données proviennent de sources publiques, de partenaires tiers, ou sont générées en internes »
Mobilisation du DPO (Data Protection Officer)						
Alignement sur les normes reconnues						
Mécanismes de contrôle des données						
Évaluation de la qualité des sources externes						
Garantie de la qualité et de l'intégrité des données						

Vérification de la sécurité des ensembles de données					✅ Sécurité des données et stockage à froid : « Les clés privées sont stockées hors ligne dans des modules de sécurité matériels, jamais connectés à Internet, avec des contrôles d'accès biométriques »	
Documentation des protocoles de gouvernance des données					✅ Gouvernance des accès et séparation des fonds : « Les fonds des clients sont détenus à un ratio de 1 :1 séparés des fonds de l'entreprise, et peuvent être retirés à tout moment »	
Analyse des accès aux données					✅ Gouvernance des accès et séparation des fonds : « Les fonds des clients sont détenus à un ratio de 1 :1 séparés des fonds de l'entreprise, et peuvent être retirés à tout moment »	
Qualification et légitimité des personnes ayant accès aux données	✅ « Les données et les modèles de nos clients leur appartiennent, c'est pourquoi nous nous assurons qu'ils en gardent toujours le contrôle »		✅ Conception selon des critères exigeants de confidentialité et de propriété des données : « Les données des utilisateurs doivent être protégées et rester sous leur contrôle »		✅ Gouvernance des accès et séparation des fonds : « Les fonds des clients sont détenus à un ratio de 1 :1 séparés des fonds de l'entreprise, et peuvent être retirés à tout moment »	
Traçabilité des accès aux données	✅ « Les données et les modèles de nos clients leur appartiennent, c'est pourquoi nous nous assurons qu'ils en gardent toujours le contrôle »				✅ Conformité et transparence des actifs : « Tous les actifs des clients sont séparés à l'aide d'adresses numériques uniques, vérifiables sur la blockchain. Les auditeurs ont un accès en lecture seule aux soldes, transactions et activités »	

Mise en place de mesures de traçabilité						
Documentation des méthodes de conception	✅ « Nous publions des cartes de modèles qui décrivent la manière dont ils ont été créés, les cas d'usage potentiels, les implications éthiques ou sociétales connues ainsi que les indices de performance »	✅ Documentation détaillée : « Les capacités, limites, données utilisées, méthodologies, risques et directives doivent être documentées. Les utilisateurs doivent savoir comment les données sont utilisées ».	✅ Respect de critères rigoureux d'analyse statistique et d'apprentissage automatique : « L'IA doit être fondée sur des méthodes robustes et vérifiables »	✅ Comprendre le fonctionnement de l'IA, ses capacité et ses limites : « Microsoft insiste sur la nécessité de comprendre comment fonctionne Copilot, ce qu'il peut faire, ce qu'il ne peut pas faire, et dans quel contexte il est utilisé »		✅ Compréhension du fonctionnement du modèle (Claude 3.7 Sonnet) : « Anthropic fournit un résumé clair des capacités, des limites, des méthodes d'entraînement, et des techniques de sécurité de Claude ».
Documentation de l'entraînement des modèles (IA fondée sur l'apprentissage)						
Documentation des méthodes de test et de validation						✅ Sécurité, auditabilité et conformité : « Anthropic décrit les méthodes de test, les résultats de sécurité, et les engagements de conformité (ex. : ASL-2)
Documentation des résultats et des décisions	✅ « Nous publions des cartes de modèles qui décrivent la manière dont ils ont été créés, les cas d'usage potentiels, les implications éthiques ou sociétales connues ainsi que les indices de performance »	✅ Documentation détaillée : « Les capacités, limites, données utilisées, méthodologies, risques et directives doivent être documentées. Les utilisateurs doivent savoir comment les données sont utilisées ».		✅ Comprendre le fonctionnement de l'IA, ses capacité et ses limites : « Microsoft insiste sur la nécessité de comprendre comment fonctionne Copilot, ce qu'il peut faire, ce qu'il ne peut pas faire, et dans quel contexte il est utilisé » ✅ Grounding (ancrage contextuel des réponses) : « Copilot utilise des sources	✅ Conformité et transparence des actifs : « Tous les actifs des clients sont séparés à l'aide d'adresses numériques uniques, vérifiables sur la blockchain. Les auditeurs ont un accès en lecture seule aux soldes, transactions et activités »	✅ Compréhension du fonctionnement du modèle (Claude 3.7 Sonnet) : « Anthropic fournit un résumé clair des capacités, des limites, des méthodes d'entraînement, et des techniques de sécurité de Claude ».

				comme Microsoft Graph pour contextualiser les réponses et améliorer leur pertinence »		
Compréhensibilité des décisions du système	✅ « Nous souhaitons que nos clients comprennent les tenants et aboutissants de chaque suggestion générées par l'IA »	✅ Sensibilisation des utilisateurs « Les utilisateurs doivent savoir clairement quand ils interagissent avec l'IA, identifier le contenu généré par l'IA, et comprendre comment les décisions sont prises ».		✅ Grounding (ancrage contextuel des réponses) : « Copilot utilise des sources comme Microsoft Graph pour contextualiser les réponses et améliorer leur pertinence »		
Explication des choix algorithmiques	✅ « Nous souhaitons que nos clients comprennent les tenants et aboutissants de chaque suggestion générées par l'IA »	✅ Sensibilisation des utilisateurs « Les utilisateurs doivent savoir clairement quand ils interagissent avec l'IA, identifier le contenu généré par l'IA, et comprendre comment les décisions sont prises ».				
Évaluation de l'influence sur les processus décisionnels						
Justification du déploiement dans un domaine spécifique						
Évaluation du modèle économique	✅ « Nous publions des cartes de modèles qui décrivent la manière dont ils ont été créés, les cas d'usage potentiels, les implications éthiques ou sociétales connues	✅ Développement et limites du système : « Les utilisateurs doivent comprendre les risques, les conflits d'intérêts, et les motivations commerciales. Cela renforce la confiance dans l'IA ».				

	ainsi que les indices de performance »					
Conception orientée vers l'interprétabilité						✅ Compréhension du fonctionnement du modèle (Claude 3.7 Sonnet) : « Anthropic fournit un résumé clair des capacités, des limites, des méthodes d'entraînement, et des techniques de sécurité de Claude ».
Choix de modèles interprétables				✅ Transparence sur les choix de conception et d'implémentation : « Microsoft décrit les choix faits dans la conception du système, notamment le recours à différents modèles selon les besoins (vitesse, créativité, etc.)		
Analyse des données d'entraînement et de test						
Accès à la logique interne du modèle						
Identification claire du système comme étant une IA		✅ Sensibilisation des utilisateurs « Les utilisateurs doivent savoir clairement quand ils interagissent avec l'IA, identifier le contenu généré par l'IA, et comprendre comment les décisions sont prises ».				
Explication des critères de décision	✅ « Nous souhaitons que nos clients comprennent les tenants et aboutissants de chaque suggestion générées par l'IA »			✅ Grounding (ancrage contextuel des réponses) : « Copilot utilise des sources comme Microsoft Graph pour contextualiser les réponses et améliorer leur pertinence »		

Accessibilité de l'information pour tous les profils d'utilisateurs		✅ Contrôle utilisateur sur les suggestions : « Les utilisateurs peuvent accepter ou refuser les suggestions, ajuster les types de suggestions, et reçoivent des explications pour prendre des décisions éclairées ».	✅ Amélioration continue pour plus d'accessibilité, d'équité et de bénéfices : « L'IA doit évoluer pour mieux servir tous les profils d'utilisateurs »			
Prise en compte des retours utilisateurs						
Communication des risques et biais		✅ Développement et limites du système : « Les utilisateurs doivent comprendre les risques, les conflits d'intérêts, et les motivations commerciales. Cela renforce la confiance dans l'IA ».				
Communication élargie selon les publics						
Indication de la finalité du système	✅ Nous leurs fournissons également une charte clarifiant les conditions d'utilisation et les applications des solutions d'IA »					
Définition et explication des scénarios d'usage	✅ Nous leurs fournissons également une charte clarifiant les conditions d'utilisation et les applications des solutions d'IA »					
Prise en compte des limites cognitives humaines						

Communication différenciée selon les rôles	✅ Nous leurs fournissons également une charte clarifiant les conditions d'utilisation et les applications des solutions d'IA »			✅ Comprendre le fonctionnement de l'IA, ses capacité et ses limites : « Microsoft insiste sur la nécessité de comprendre comment fonctionne Copilot, ce qu'il peut faire, ce qu'il ne peut pas faire, et dans quel contexte il est utilisé »		✅ Compréhension du fonctionnement du modèle (Claude 3.7 Sonnet) : « Anthropic fournit un résumé clair des capacités, des limites, des méthodes d'entraînement, et des techniques de sécurité de Claude ».
Stratégie de prévention des biais			✅ L'IA doit atténuer la subjectivité : « Elle doit être conçue pour réduire les biais et favoriser l'équité »			✅ Constitutionnal AI et alignement sur des principes éthiques : « Claude est entraîné avec une méthode appelée « Conditionnal AI » pour garantir des réponses utiles, inoffensives et honnêtes ».
Évaluation des limites des ensembles de données				✅ Comprendre le fonctionnement de l'IA, ses capacité et ses limites : « Microsoft insiste sur la nécessité de comprendre comment fonctionne Copilot, ce qu'il peut faire, ce qu'il ne peut pas faire, et dans quel contexte il est utilisé »		
Prise en compte de la diversité des utilisateurs			✅ Conception avec un groupe diversifié d'étudiants et d'enseignants : « L'IA doit être codéveloppée avec les utilisateurs finaux pour refléter leurs besoins »			
Utilisation d'outils d'analyse des biais						
Processus de détection et de contrôle des biais			✅ L'IA doit atténuer la subjectivité : « Elle doit être conçue pour réduire les biais et favoriser l'équité »			

Mécanisme de signalement des biais						
Communication sur les biais et les recours						
Prise en compte des effets indirects						
Évaluation de la variabilité des décisions						
Analyse des causes de variabilité						
Évaluation de l'impact de la variabilité sur les droits fondamentaux						
Définition de l'équité appliquée						
Mesure de l'équité			✅ L'IA doit atténuer la subjectivité : « Elle doit être conçue pour réduire les biais et favoriser l'équité »			
Mécanismes de garantie de l'équité						
Adaptation aux préférences et capacités individuelles			✅ Amélioration continue pour plus d'accessibilité, d'équité et de bénéfices : « L'IA doit évoluer pour mieux servir tous les profils d'utilisateurs »			

Prise en compte des besoins spécifiques et du handicap						
Compatibilité avec les technologies d'assistance						
Consultation des communautés concernées			✅ Conception avec un groupe diversifié d'étudiants et d'enseignants : « L'IA doit être codéveloppée avec les utilisateurs finaux pour refléter leurs besoins »			
Évaluation de l'impact sur les groupes d'utilisateurs						
Diversité de l'équipe de conception						
Évaluation des effets disproportionnés			✅ Amélioration continue pour plus d'accessibilité, d'équité et de bénéfices : « L'IA doit évoluer pour mieux servir tous les profils d'utilisateurs »			
Prise en compte des retours de groupes externes						

Participation des parties prenantes			✅ Conception avec un groupe diversifié d'étudiants et d'enseignants : « L'IA doit être codéveloppée avec les utilisateurs finaux pour refléter leurs besoins »	✅ Importance du contexte d'utilisation (technologie, personnes, environnement) : « Microsoft souligne que l'IA ne se limite pas à la technologie, mais inclut les utilisateurs et leur environnement ».		
Mobilisation des acteurs internes						
Mesure de l'impact environnemental						
Réduction de l'empreinte écologique						
Évaluation de l'attachement émotionnel						
Clarté sur la nature simulée de l'interaction						

Analyse des incidences sociales	✅ « Nous publions des cartes de modèles qui décrivent la manière dont ils ont été créés, les cas d'usage potentiels, les implications éthiques ou sociétales connues ainsi que les indices de performance »	✅ Développement et limites du système : « Les utilisateurs doivent comprendre les risques, les conflits d'intérêts, et les motivations commerciales. Cela renforce la confiance dans l'IA ».	✅ L'IA doit soutenir les objectifs pédagogiques et protéger l'intégrité académique : « L'IA est conçue pour renforcer l'apprentissage, non pour le remplacer ou le biaiser ».	✅ Importance du contexte d'utilisation (technologie, personnes, environnement) : « Microsoft souligne que l'IA ne se limite pas à la technologie, mais inclut les utilisateurs et leur environnement ».		
Évaluation de l'impact sociétal élargi						
Auditabilité du système				✅ Traitement responsable des données : « Microsoft décrit les étapes de traitement des requêtes, incluant des vérifications de conformité à l'IA responsable »	✅ Conformité et transparence des actifs : « Tous les actifs des clients sont séparés à l'aide d'adresses numériques uniques, vérifiables sur la blockchain. Les auditeurs ont un accès en lecture seule aux soldes, transactions et activités »	✅ Sécurité, auditabilité et conformité : « Anthropic décrit les méthodes de test, les résultats de sécurité, et les engagements de conformité (ex. : ASL-2)
Analyse des risques multi-acteurs						
Mise en place de formations à la responsabilité						
Identification des publics concernés par la formation						
Intégration du cadre juridique dans les formations						
Mise en place d'un comité d'éthique ou d'un mécanisme équivalent						

Recours à des audits ou orientations externes					✅ Règlementation et responsabilité : « Gemini est un dépositaire fiduciaire qualifié, réglementé par l'Etat de new York, avec une couverture d'assurance pour les pertes liées aux cryptomonnaies »	
Mécanisme de signalement des vulnérabilités						
Mécanisme de recensement des valeurs et intérêts						✅ Constitutionnal AI et alignement sur des principes éthiques : « Claude est entraîné avec une méthode appelée« Conditionnal AI » pour garantir des réponses utiles, inoffensives et honnêtes ».
Processus de décision pour les arbitrages				✅ Transparence sur les choix de conception et d'implémentation : « Microsoft décrit les choix faits dans la conception du système, notamment le recours à différents modèles selon les besoins (vitesse, créativité, etc.)		
Documentation des arbitrages		✅ Développement et limites du système : « Les utilisateurs doivent comprendre les risques, les conflits d'intérêts, et les motivations commerciales. Cela renforce la confiance dans l'IA ».		✅ Transparence sur les choix de conception et d'implémentation : « Microsoft décrit les choix faits dans la conception du système, notamment le recours à différents modèles selon les besoins (vitesse, créativité, etc.)		✅ Constitutionnal AI et alignement sur des principes éthiques : « Claude est entraîné avec une méthode appelée « Conditionnal AI » pour garantir des réponses utiles, inoffensives et honnêtes ».
Mécanismes de recours en cas de préjudice						
Information sur les voies de recours						

Nous n'avons pas pu analyse l'application ChatGPT car aucune page dédié à la transparence n'était disponible sur leur site. En faisant des recherches sur internet nous avons pu constater de fort débat concernant la transparence de cette entreprise.

Résultats

Notre grille d'analyse refléter 24 critères de transparence sur 129 critères aux total. Après confrontation avec notre grille d'analyse, nous constations que les critères de transparence serait de 50 mais qu'elle serait diviser dans toutes les dimensions proposé par l'outil ALTAI et pas seulement celle définit dans la dimension de transparence (traçabilité, explicabilité, communication). Dans le tableau suivant, on peut voir les critères concordants avec notre grille d'analyse, ceux étant non concordant ainsi que ceux relative à la dimension de transparence de l'outil ALTAI mais que nous n'avons pas retrouver.

Critères concordants (50 critères)	Critères non concordants (hors transparence) (70 critères)	Critères non concordants (transparence) (9 critères)
Interaction avec la prise de décision humaine	Analyse d'impact sur les droits fondamentaux	Mise en place de mesures de traçabilité
Préservation de l'autonomie décisionnelle	Documentation des arbitrages éthiques	Documentation de l'entraînement des modèles (IA fondée sur l'apprentissage)
Signalement des décisions algorithmiques	Identification des agents virtuels	Évaluation de l'influence sur les processus décisionnels
Renforcement des capacités humaines	Répartition des tâches entre IA et humains	Justification du déploiement dans un domaine spécifique
Présence d'un mécanisme d'arrêt d'urgence	Prévention de la dépendance excessive	Analyse des données d'entraînement et de test
Évaluation des vulnérabilités potentielles	Définition du niveau de contrôle humain	Accès à la logique interne du modèle
Politique d'assurance	Identification de l'acteur humain responsable	Prise en compte des retours utilisateurs
Règles de responsabilité et protection des consommateurs	Mécanismes de supervision humaine	Communication élargie selon les publics
Prise en compte des contextes spécifiques	Auditabilité et gouvernance de l'autonomie	Prise en compte des limites cognitives humaines
Méthodes de vérification de la fiabilité et de la reproductibilité	Contrôle renforcé pour les systèmes autonomes ou auto-apprenants	
Application des processus de test	Détection des anomalies et réponse aux risques	
Évaluation du type et de la portée des données	Identification des types de vulnérabilités	
Minimisation de l'usage de données sensibles	Mesures de résilience et d'intégrité	
Mécanismes de consentement et de contrôle	Comportement en environnement imprévu	
Mesures de protection des données	Gestion du double usage	
Vérification de la sécurité des ensembles de données	Existence de solutions de secours	
Documentation des protocoles de gouvernance des données	Évaluation du niveau de risque contextuel	
Analyse des accès aux données	Processus d'évaluation des risques et de la sécurité	
Qualification et légitimité des personnes ayant accès aux données	Communication des risques pour l'intégrité physique	
Traçabilité des accès aux données	Plan de gestion des usages abusifs ou détournés	
Documentation des méthodes de conception	Évaluation des dommages potentiels	
Documentation des méthodes de test et de validation	Évaluation des impacts environnementaux et sur les animaux	
Documentation des résultats et des décisions	Analyse des menaces de cybersécurité	
Compréhensibilité des décisions du système	Évaluation des défaillances critiques	
Explication des choix algorithmiques	Définition de seuils de déclenchement	
Évaluation du modèle économique	Test des solutions de secours	
Conception orientée vers l'interprétabilité	Définition du niveau de précision attendu	
Choix de modèles interprétables	Méthodes de mesure de la précision	
Identification claire du système comme étant une IA	Qualité et actualisation des données	
Explication des critères de décision	Évaluation des besoins en données supplémentaires	
Accessibilité de l'information pour tous les profils d'utilisateurs	Évaluation des préjudices liés aux erreurs	
Communication des risques et biais	Suivi du taux d'erreurs	
Indication de la finalité du système	Plan de correction des erreurs	
Définition et explication des scénarios d'usage	Stratégie de contrôle du système	
Communication différenciée selon les rôles	Documentation des réglages critiques	
Stratégie de prévention des biais	Communication de la fiabilité aux utilisateurs	
Évaluation des limites des ensembles de données	Mécanisme de signalement des atteintes à la vie privée	
Prise en compte de la diversité des utilisateurs	Mobilisation du DPO (Data Protection Officer)	

Processus de détection et de contrôle des biais Mesure de l'équité Adaptation aux préférences et capacités individuelles Consultation des communautés concernées Évaluation des effets disproportionnés Participation des parties prenantes Analyse des incidences sociales Auditabilité du système Recours à des audits ou orientations externes Mécanisme de recensement des valeurs et intérêts Processus de décision pour les arbitrages Documentation des arbitrages	Alignement sur les normes reconnues Mécanismes de contrôle des données Évaluation de la qualité des sources externes Garantie de la qualité et de l'intégrité des données Utilisation d'outils d'analyse des biais Mécanisme de signalement des biais Communication sur les biais et les recours Prise en compte des effets indirects Évaluation de la variabilité des décisions Analyse des causes de variabilité Évaluation de l'impact de la variabilité sur les droits fondamentaux Définition de l'équité appliquée Mécanismes de garantie de l'équité Prise en compte des besoins spécifiques et du handicap Compatibilité avec les technologies d'assistance Évaluation de l'impact sur les groupes d'utilisateurs Diversité de l'équipe de conception Prise en compte des retours de groupes externes Mobilisation des acteurs internes Mesure de l'impact environnemental Réduction de l'empreinte écologique Évaluation de l'attachement émotionnel Clarté sur la nature simulée de l'interaction Évaluation de l'impact sociétal élargi Analyse des risques multi-acteurs Mise en place de formations à la responsabilité Identification des publics concernés par la formation Intégration du cadre juridique dans les formations Mise en place d'un comité d'éthique ou d'un mécanisme équivalent Mécanisme de signalement des vulnérabilités Mécanismes de recours en cas de préjudice Information sur les voies de recours	

Comparaisons critères de transparence entre : outil ALTAIL, cadre théorique, entreprises

Critères de transparence outil ALTAIL (24 critères)	Critères de transparence cadre théorique (27 critères)	Critères de transparence des entreprises (50 critères)

Mise en place de mesures de traçabilité Documentation des méthodes de conception Documentation de l'entraînement des modèles (IA fondée sur l'apprentissage) Documentation des méthodes de test et de validation Documentation des résultats et des décisions Compréhensibilité des décisions du système Explication des choix algorithmiques Évaluation de l'influence sur les processus décisionnels Justification du déploiement dans un domaine spécifique Évaluation du modèle économique Conception orientée vers l'interprétabilité Choix de modèles interprétables Analyse des données d'entraînement et de test Accès à la logique interne du modèle Identification claire du système comme étant une IA Explication des critères de décision Accessibilité de l'information pour tous les profils d'utilisateurs Prise en compte des retours utilisateurs Communication des risques et biais Communication élargie selon les publics Indication de la finalité du système Définition et explication des scénarios d'usage Prise en compte des limites cognitives humaines Communication différenciée selon les rôles	Signalement des décisions algorithmiques Définition du niveau de contrôle humain Mécanismes de supervision humaine Stratégie de contrôle du système Alignement sur les normes reconnues Traçabilité des accès aux données Documentation des méthodes de conception Documentation des méthodes de test et de validation Documentation des résultats et des décisions Compréhensibilité des décisions du système Explication des choix algorithmiques Conception orientée vers l'interprétabilité Choix de modèles interprétables Accessibilité de l'information pour tous les profils d'utilisateurs Prise en compte des retours utilisateurs Communication des risques et biais Communication différenciée selon les rôles Prise en compte des effets indirects Évaluation des effets disproportionnés Participation des parties prenantes Évaluation de l'attachement émotionnel Analyse des incidences sociales Auditabilité du système Mise en place d'un comité d'éthique ou d'un mécanisme équivalent Mécanisme de recensement des valeurs et intérêts Processus de décision pour les arbitrages Documentation des arbitrages	Interaction avec la prise de décision humaine Préservation de l'autonomie décisionnelle Signalement des décisions algorithmiques Renforcement des capacités humaines Présence d'un mécanisme d'arrêt d'urgence Évaluation des vulnérabilités potentielles Politique d'assurance Règles de responsabilité et protection des consommateurs Prise en compte des contextes spécifiques Méthodes de vérification de la fiabilité et de la reproductibilité Application des processus de test Évaluation du type et de la portée des données Minimisation de l'usage de données sensibles Mécanismes de consentement et de contrôle Mesures de protection des données Vérification de la sécurité des ensembles de données Documentation des protocoles de gouvernance des données Analyse des accès aux données Qualification et légitimité des personnes ayant accès aux données Traçabilité des accès aux données Documentation des méthodes de conception Documentation des méthodes de test et de validation Documentation des résultats et des décisions Compréhensibilité des décisions du système Explication des choix algorithmiques Évaluation du modèle économique Conception orientée vers l'interprétabilité Choix de modèles interprétables Identification claire du système comme étant une IA Explication des critères de décision Accessibilité de l'information pour tous les profils d'utilisateurs Communication des risques et biais Indication de la finalité du système Définition et explication des scénarios d'usage Communication différenciée selon les rôles Stratégie de prévention des biais Évaluation des limites des ensembles de données Prise en compte de la diversité des utilisateurs Processus de détection et de contrôle des biais Mesure de l'équité Adaptation aux préférences et capacités individuelles Consultation des communautés concernées Évaluation des effets disproportionnés Participation des parties prenantes Analyse des incidences sociales Auditabilité du système

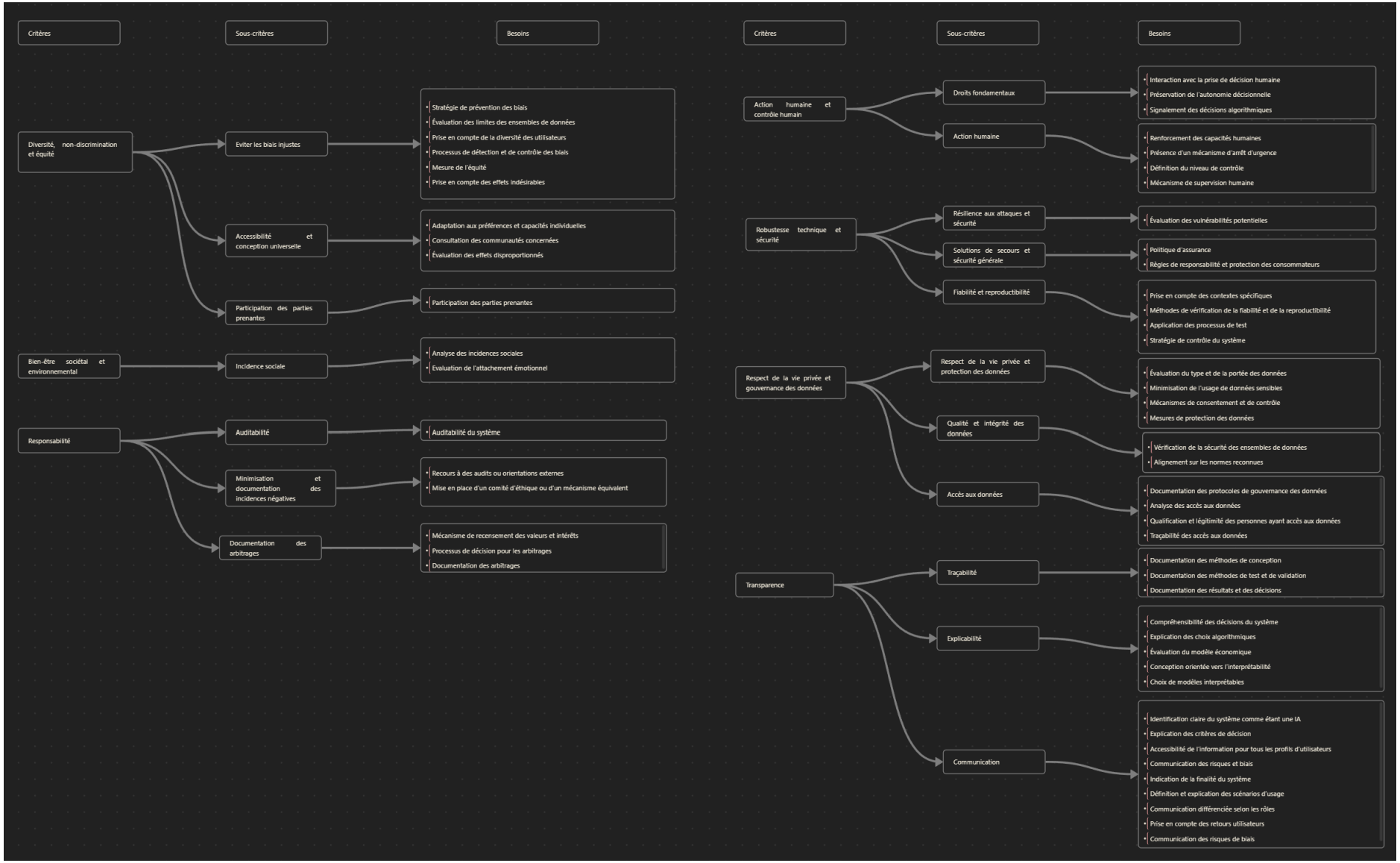
		Recours à des audits ou orientations externes Mécanisme de recensement des valeurs et intérêts Processus de décision pour les arbitrages Documentation des arbitrages

Nous pouvons analyser que dans ce tableau récapitulatif, une quantité et une richesse de critères sont présent :

- Le référentiel de l'outil ALTAI comprend 24 critères qui reflète du cadre normatif européen, structuré autour de la traçabilité, l'explicabilité et la communication.
- Le référentiel de notre cadre théorique comprend 27 critères qui enrichissement le référentiel ALTAI avec des dimensions comme l'équité, les effets indirects ou la participation.
- Le référentiel des entreprises comprend 50 critères d'une très grande granularité, mais hétérogène avec des critères très détaillées et d'autres absent ou flous.

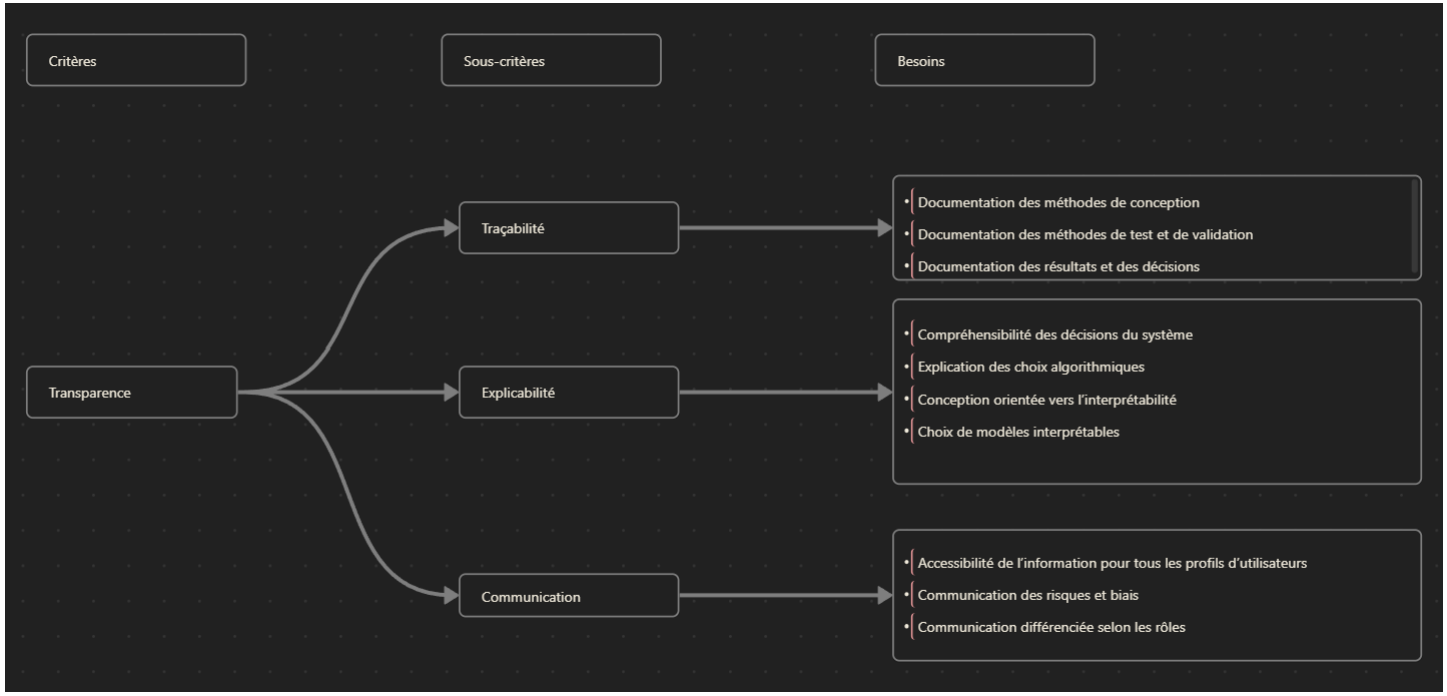
Les entreprises mobilisent un plus grand nombre de critères, mais cela ne signifie pas nécessairement une meilleure couverture. Il s'agit souvent de redondances, de reformulations ou de critères spécifiques à leur contexte commercial.

Les critères sont présents à la fois dans l'outil ALTAI, la grille d'analyse et les pratiques des entreprises sont présenté ci-dessous :


Critères
Sous-critères
Besoins
Diversité, non-discrimination et équité
Eviter les biais injustes
Stratégie de prévention des biais
Évaluation des limites des ensembles de données
Prise en compte de la diversité des utilisateurs
Processus de détection et de contrôle des biais
Mesure de l'équité
Prise en compte des effets indésirables
Accessibilité et conception universelle
Adaptation aux préférences et capacités individuelles
Consultation des communautés concernées
Évaluation des effets disproportionnés
Participation des parties prenantes
Participation des parties prenantes
Bien-être sociétal et environnemental
Incidence sociale
Analyse des incidences sociales
Evaluation de l'attachement émotionnel
Responsabilité
Auditabilité
Auditabilité du système
Minimisation et documentation des incidences négatives
Recours à des audits ou orientations externes
Mise en place d'un comité d'éthique ou d'un mécanisme équivalent
Documentation des arbitrages
Mécanisme de recensement des valeurs et intérêts
Processus de décision pour les arbitrages
Documentation des arbitrages
Critères
Sous-critères
Besoins
Action humaine et contrôle humain
Droits fondamentaux
Interaction avec la prise de décision humaine
Préservation de l'autonomie décisionnelle
Signalement des décisions algorithmiques
Action humaine
Renforcement des capacités humaines
Présence d'un mécanisme d'arrêt d'urgence
Définition du niveau de contrôle
Mécanisme de supervision humaine
Robustesse technique et sécurité
Résilience aux attaques et sécurité
Évaluation des vulnérabilités potentielles
Solutions de secours et sécurité générale
Politique d'assurance
Règles de responsabilité et protection des consommateurs
Fiabilité et reproductibilité
Prise en compte des contextes spécifiques
Méthodes de vérification de la fiabilité et de la reproductibilité
Application des processus de test
Stratégie de contrôle du système
Respect de la vie privée et gouvernance des données
Respect de la vie privée et protection des données
Évaluation du type et de la portée des données
Minimisation de l'usage de données sensibles
Mécanismes de consentement et de contrôle
Mesures de protection des données
Qualité et intégrité des données
Vérification de la sécurité des ensembles de données
Alignement sur les normes reconnues
Accès aux données
Documentation des protocoles de gouvernance des données
Analyse des accès aux données
Qualification et légitimité des personnes ayant accès aux données
Traçabilité des accès aux données
Transparence
Traçabilité
Documentation des méthodes de conception
Documentation des méthodes de test et de validation
Documentation des résultats et des décisions
Explicabilité
Compréhensibilité des décisions du système
Explication des choix algorithmiques
Évaluation du modèle économique
Conception orientée vers l'interprétabilité
Choix de modèles interprétables
Communication
Identification claire du système comme étant une IA
Explication des critères de décision
Accessibilité de l'information pour tous les profils d'utilisateurs
Communication des risques et biais
Indication de la finalité du système
Définition et explication des scénarios d'usage
Communication différenciée selon les rôles
Prise en compte des retours utilisateurs
Communication des risques de biais

Ces critères représente le socle commun de la gouvernance éthique de l'IA. Nous pouvons voir qu'il y a un équilibre entre droits fondamentaux, action humaine, résilience aux attaques et sécurité, solutions de secours et sécurité générale, fiabilité et reproductibilité, responsabilité de la vie privée et protection des données, qualité et intégrité des données, accès aux données, traçabilité, explicabilité, communication, éviter les biais injustes, accessibilité et conception universelle, participation des parties prenantes, incidences sociales, auditabilité, minimisation et documentation des incidences négatives, documentation des arbitrages.

Les critères suivants sont présents dans les trois référentiels, ce qui montre un noyau dur de la transparence :



C'est critères constituent les fondamentaux de la transparence algorithmique. Leur présence dans les trois référentiels montre leur légitimité et les applicabilité.

Certains critères sont recommandés par ALTAI et intégrés dans le cadre théorique, mais peu ou pas repris par les entreprises, tel que :

- Définition du niveau de contrôle humain
- Mécanisme de supervision humaine
- Signalement des décisions algorithmiques
- Stratégie de contrôle du système
- Alignement sur les normes reconnues.

Ces critères relèvent d'une transparence structurelle ou organisationnelle, souvent négligée par les entreprises au profit d'une transparence plus marketing ou centrée sur l'interface utilisateur.

Le cadre théorique introduit des critères absents d'ALTAI, mais essentiels pour une approche éthique approfondie :

- Prise en compte des effets indirects
- Evaluation des effets disproportionnées
- Participation des parties prenantes
- Evaluation de l'attachement émotionnel
- Analyse des incidences sociales
- Mise en place d'un comité d'éthique
- Documentation des arbitrages

Ces critères traduisent une volonté de gouvernance éthique et de responsabilité sociale, encore peu intégrée dans les pratiques industrielles.

Les entreprises introduisent des critères non présents dans ALTAI ni dans notre cadre théorique, souvent liées à la sécurité des données (ex. : stockage à froid, séparation des fonds), la conformité réglementaire (ex. : ASL-2, SOC2), la contextualisation des réponses (ex. : grounding via Microsoft Graph), la gestion des accès (ex. : journalisation blockchain). Ces critères montent une adaptation aux contraintes opérationnelles, mais peuvent aussi refléter une stratégie de communication plus qu'un engagement éthique profond.

Confrontation réglementation IA Act avec les critères de transparence des entreprises

Critères des entreprises (Approche réglementaire en gras) (5/50 critères)	**Critères de notre grille d'analyse (Approche réglementaire en gras) (7/27 critères)**
Interaction avec la prise de décision humaine	Signalement des décisions algorithmiques
Préservation de l'autonomie décisionnelle	**Définition du niveau de contrôle humain**
Signalement des décisions algorithmiques	**Mécanismes de supervision humaine**
Renforcement des capacités humaines	Stratégie de contrôle du système
Présence d'un mécanisme d'arrêt d'urgence	Alignement sur les normes reconnues
Évaluation des vulnérabilités potentielles	**Traçabilité des accès aux données**
Politique d'assurance	**Documentation des méthodes de conception**
Règles de responsabilité et protection des consommateurs	Documentation des méthodes de test et de validation
Prise en compte des contextes spécifiques	**Documentation des résultats et des décisions**
Méthodes de vérification de la fiabilité et de la reproductibilité	Compréhensibilité des décisions du système
Application des processus de test	Explication des choix algorithmiques
Évaluation du type et de la portée des données	Conception orientée vers l'interprétabilité
Minimisation de l'usage de données sensibles	Choix de modèles interprétables
Mécanismes de consentement et de contrôle	Accessibilité de l'information pour tous les profils d'utilisateurs
Mesures de protection des données	Prise en compte des retours utilisateurs
Vérification de la sécurité des ensembles de données	Communication des risques et biais
Documentation des protocoles de gouvernance des données	**Communication différenciée selon les rôles**
Analyse des accès aux données	Prise en compte des effets indirects
Qualification et légitimité des personnes ayant accès aux données	Évaluation des effets disproportionnés
Traçabilité des accès aux données	Participation des parties prenantes
Documentation des méthodes de conception	Évaluation de l'attachement émotionnel
Documentation des méthodes de test et de validation	Analyse des incidences sociales
Documentation des résultats et des décisions	Auditabilité du système
Compréhensibilité des décisions du système	Mise en place d'un comité d'éthique ou d'un mécanisme équivalent
Explication des choix algorithmiques	Mécanisme de recensement des valeurs et intérêts
Évaluation du modèle économique	Processus de décision pour les arbitrages
Conception orientée vers l'interprétabilité	**Documentation des arbitrages**
Choix de modèles interprétables	
Identification claire du système comme étant une IA	
Explication des critères de décision	
Accessibilité de l'information pour tous les profils d'utilisateurs	
Communication des risques et biais	
Indication de la finalité du système	
Définition et explication des scénarios d'usage	
Communication différenciée selon les rôles	
Stratégie de prévention des biais	
Évaluation des limites des ensembles de données	
Prise en compte de la diversité des utilisateurs	
Processus de détection et de contrôle des biais	
Mesure de l'équité	
Adaptation aux préférences et capacités individuelles	

Consultation des communautés concernées Évaluation des effets disproportionnés Participation des parties prenantes Analyse des incidences sociales Auditabilité du système Recours à des audits ou orientations externes Mécanisme de recensement des valeurs et intérêts Processus de décision pour les arbitrages **Documentation des arbitrages**	

L'analyse des pratiques des entreprises à travers la grille révèle des écarts significatifs entre les critères théoriques issus de l'outil ALTAI et leur mise en œuvre concrète. Plusieurs critères clés ont été examinés, notamment en matières de contrôle humain, de transparence, de documentation et d'audibilité.

Contrôle humain

Deux critères fondamentaux liés au contrôle humains sont absents chez l'ensemble des entreprises analysées :

- Définition du niveau de contrôle humain : aucune entreprise ne précise clairement le niveau de contrôle attendu de la part des utilisateurs humains dans les processus décisionnels.
- Mécanismes de supervision humaine : aucun dispositif explicite de supervision ou de validation humaine n'est documenté, ce qui soulève des questions sur la capacité des utilisateurs à reprendre la main en cas de dérive du système.

Traçabilité des accès aux données

Seules deux entreprises mettent en œuvre des mécanismes de journalisation des accès aux données :

- L'une affirme que « les données et les modèles de nos clients leur appartiennent », garantissant ainsi un contrôle utilisateur.
- L'autre met en place une traçabilité blockchain avec accès en lecture seule pour les auditeurs, ce qui constitue une bonne pratique en matière de transparence et de sécurité.

Documentation des méthodes de conception

Cinq entreprises se distinguent par une documentation technique approfondie :

- Elles publient des cartes de modèles, décrivent les méthodologies d'entraînement, les limites, les risques et les cas d'usage.
- Certaines insistent sur la compréhension du fonctionnement de l'IA par les utilisateurs, ce qui renforce la transparence et l'appropriation.

Documentation des résultats et décisions

Là encore, cinq entreprises fournissent une documentation claire sur les résultats produits par l'IA :

- Elles expliquent les choix algorithmiques, les sources de données, et les mécanismes de contextualisation (ex. : grounding via Microsoft Graph).
- Certaines utilisent des adresses numériques vérifiables pour assurer la transparence des transactions et décisions.

Communication différenciée selon les rôles

Trois entreprises adoptent une communication différenciée selon les profils d'utilisateurs :

- Elles fournissent des chartes d'utilisation, des notices explicatives et des résumés techniques adaptés aux administrateurs ou aux utilisateurs finaux.
- Cette approche est cohérente avec l'idée de transparence située, qui adapte l'information au contexte et au rôle de l'utilisateur.

Auditabilité du système

Trois entreprises mettent en œuvre des mécanismes d'auditabilité :

- Cela inclut la journalisation des événements, des engagements de conformité (ex. : ASL-2), et des accès en lecture seule pour les auditeurs.
- Ces pratiques permettent une vérification externe du fonctionnement du système, mais restent encore limitées à certaines entreprises.

Documentation des arbitrages

Enfin, trois entreprises documentent les arbitrages éthiques réalisés lors de la conception :

- Elles évoquent les conflits d’intérêts, les motivations commerciales, ou encore les choix de modèles selon les besoins.
- L’approche « Constitutionnal AI » d’Anthropic est un exemple de formalisation des valeurs dans la conception algorithmique.