

L'IA est un juge impartial de la performance académique

Richard, O., & Havran, P.

Dans Berger, J.-L., Jan, D. et da Silva, S. (dir.). (2026). *Mythologie pédagogique : l'intelligence artificielle en éducation*. Amazon Kindle Direct Publishing.

Editeur

Amazon Kindle Direct Publishing

URI

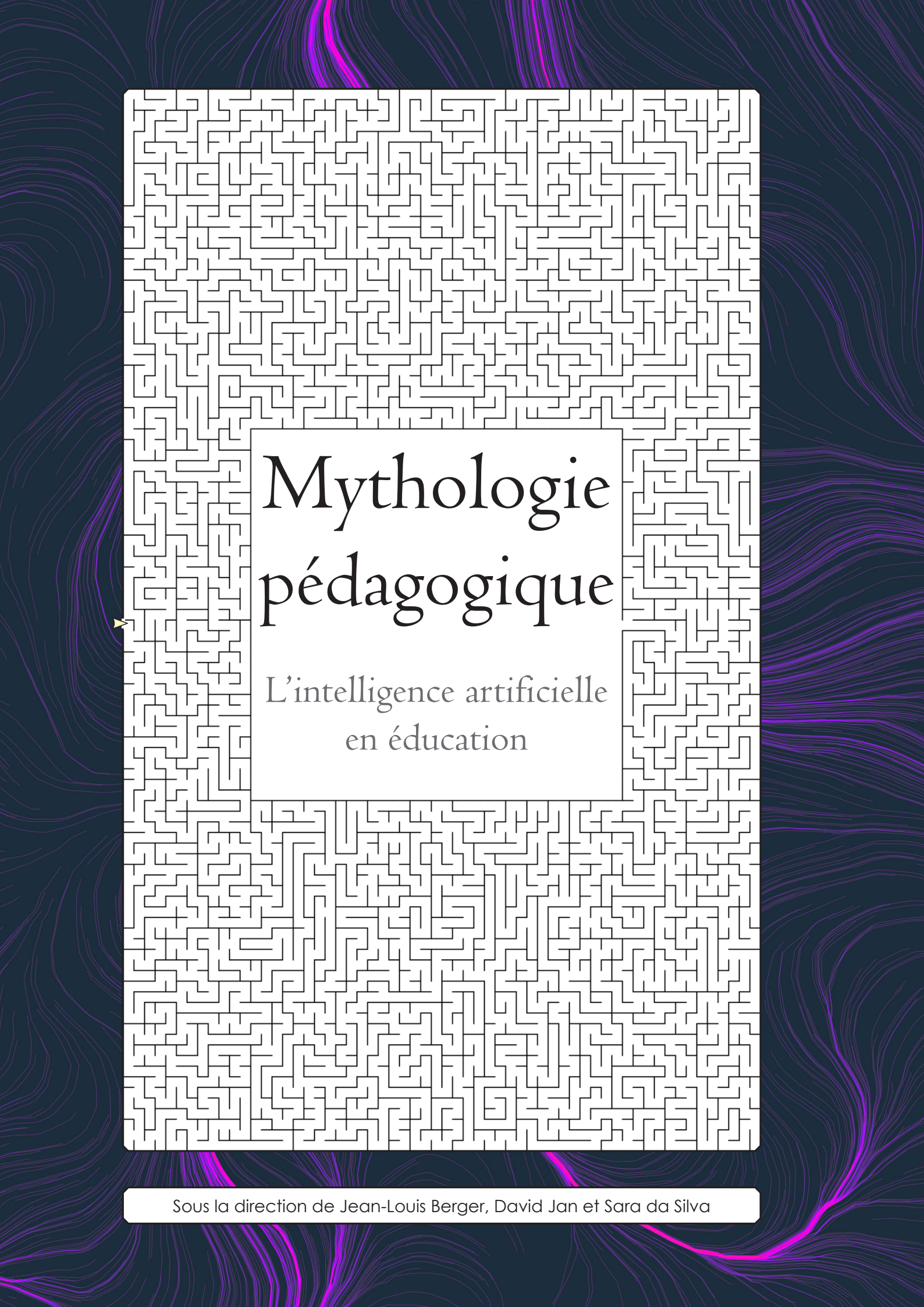
<https://plemos.education/wp-content/uploads/mp2/mp2-04.pdf>

Résumé

Ce travail déconstruit le mythe selon lequel l'IA constituerait un juge impartial de la performance académique. La littérature scientifique permet d'identifier trois raisons expliquant la persistance de ce mythe : la croyance en l'objectivité des machines, la plausibilité de surface des outils génératifs, et le désir d'un substitut au jugement humain faillible. Les recherches disponibles, principalement issues de l'enseignement supérieur en Europe, aux États-Unis et en Asie de l'Est, convergent vers un même constat : l'accord entre l'IA et les évaluateur·rice·s humain·e·s reste faible, l'IA tend à surévaluer et à compresser l'échelle de notation, peine à détecter les textes hors sujet, et reproduit les inégalités inscrites dans ses données d'entraînement. La fiabilité algorithmique, souvent confondue avec l'impartialité, ne garantit donc pas une évaluation juste : un biais systématique peut tout à fait être appliqué de manière parfaitement reproductible. L'évaluation par IA pourrait néanmoins être envisagée si les algorithmes étaient rendus transparents, si des grilles structurées encadraient leur usage, et si une supervision humaine était maintenue pour toute évaluation à enjeux élevés.

Citer ce chapitre

Richard, O., & Havran, P. (2026). L'IA est un juge impartial de la performance académique. Dans J.-L. Berger, D. Jan et S. da Silva (dir.), *Mythologie pédagogique : l'intelligence artificielle en éducation* (p. 35-45). Amazon Kindle Direct Publishing.
<https://plemos.education/wp-content/uploads/mp2/mp2-04.pdf>



Mythologie pédagogique

L'intelligence artificielle
en éducation

Sous la direction de Jean-Louis Berger, David Jan et Sara da Silva

Mythologie pédagogique :
l'intelligence artificielle en
éducation

Sous la direction de
Jean-Louis Berger, David Jan et Sara da Silva

© 2026, Jean-Louis Berger, David Jan et Sara da Silva

Cet ouvrage est mis à disposition selon les termes de la licence Creative Commons Attribution, Pas d'Utilisation Commerciale 4.0 International (CC BY-NC 4.0). Cette licence autorise le partage et l'adaptation de l'ouvrage à des fins non commerciales, à condition d'en créditer les auteurs. <https://creativecommons.org/licenses/by-nc/4.0/deed.fr>

ISBN : 9798175269940

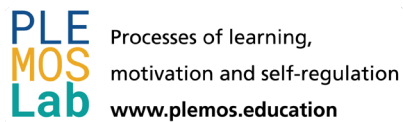
Couverture : David Jan

Révision des textes et contrôle de la qualité scientifique : Jean-Louis Berger et Sara da Silva

Mise en page : Sara da Silva

Publication: Amazon Kindle Direct Publishing (KDP)

Contact : info@plemos.education



Comment citer cet ouvrage

Berger, J.-L., Jan, D., & Da Silva, S. (2026). *Mythologie pédagogique : l'intelligence artificielle en éducation*. Amazon Kindle Direct Publishing.

Remerciements

Les étudiantes et les étudiants qui ont rédigé les chapitres composant cet ouvrage, dont le travail témoigne d'un engagement rigoureux et d'une contribution substantielle à la réflexion qui y est développée.

Sarah Chardonnens, Pierre-François Coen et Lionel Alvarez pour leurs chapitres de discussion, qui reflètent avec justesse la profondeur de leur expertise respective et qui constituent un apport précieux à l'ensemble de l'ouvrage.

Gibreel Mursal pour la relecture linguistique attentive de cet ouvrage ainsi que pour le soin apporté à sa correction.

Table des matières

INTRODUCTION

De l'importance de discuter de l'IA en éducation sous forme de mythes 1

PAR JEAN-LOUIS BERGER, DAVID JAN ET SARA DA SILVA

SECTION 1

MYTHES PÉDAGOGIQUES

CHAPITRE 1

L'IA va remplacer les enseignant·e·s 13

PAR ALBANE DAVRON ET SANDY LAFOSSE

CHAPITRE 2

L'IA favorise la réflexion autonome de l'apprenant·e 24

PAR SAMUEL FAVEZ

CHAPITRE 3

L'IA est un juge impartial de la performance académique 35

PAR ORIANE RICHARD ET PAULINA HAVRAN

CHAPITRE 4

L'IA supprime la créativité des élèves 46

PAR JULIEN GRANGE ET GIBREEL MURSAL

CHAPITRE 5

L'IA déshumanise l'éducation 56

PAR TIFFANY FUMEAUX ET LÉA ZÜRCHER

CHAPITRE 6

L'IA motive les élèves à apprendre 66

PAR GIORGIA CASOLI ET AURÉLIE BRODARD

CHAPITRE 7

L'IA comprend les émotions des élèves 75

PAR LÉA BOU CHACRA ET FLORENT DROËL

CHAPITRE 8

L'IA est inadaptée au métier d'enseignant 86

PAR NATHAN NAGALINGAM

SECTION 2

DISCUTANT·E·S

CHAPITRE 9

L'IA en éducation : prendre les mythes au sérieux pour mieux les dépasser..... 97

PAR SARAH CHARDONNENS LEHMANN

CHAPITRE 10

Le/la colocataire docile..... 105

PAR PIERRE-FRANÇOIS COEN

CHAPITRE 11

L'urgence de dépasser le technocentrisme..... 114

PAR LIONEL ALVAREZ

Glossaire 122

Chapitre 3

L'IA est un juge impartial de la performance académique

Oriane Richard et Paulina Havran

Résumé

Ce travail déconstruit le mythe selon lequel l'IA constituerait un juge impartial de la performance académique. La littérature scientifique permet d'identifier trois raisons expliquant la persistance de ce mythe : la croyance en l'objectivité des machines, la plausibilité de surface des outils génératifs, et le désir d'un substitut au jugement humain faillible. Les recherches disponibles, principalement issues de l'enseignement supérieur en Europe, aux États-Unis et en Asie de l'Est, convergent vers un même constat : l'accord entre l'IA et les évaluateur·rice·s humain·e·s reste faible, l'IA tend à surévaluer et à compresser l'échelle de notation, peine à détecter les textes hors sujet, et reproduit les inégalités inscrites dans ses données d'entraînement. La fiabilité algorithmique, souvent confondue avec l'impartialité, ne garantit donc pas une évaluation juste : un biais systématique peut tout à fait être appliqué de manière parfaitement reproductible. L'évaluation par IA pourrait néanmoins être envisagée si les algorithmes étaient rendus transparents, si des grilles structurées encadraient leur usage, et si une supervision humaine était maintenue pour toute évaluation à enjeux élevés.

Description du mythe

L'essor des grands modèles de langage⁶ a profondément transformé nos pratiques dans tous les domaines, et le secteur de l'éducation n'échappe pas à ce mouvement. L'IA y est régulièrement présentée comme une solution miracle : instantanée, économique, et susceptible d'améliorer l'évaluation de la performance académique. Face aux biais d'évaluation et au manque d'impartialité reprochés

⁶ Dans cet ouvrage, nous utiliserons l'expression « grands modèles de langage » pour désigner les *Large Language Models* (LLM), une définition détaillée de ce terme est fournie dans l'index.

aux évaluateur·rice·s humain·e·s, les entreprises d'EdTech répondent que l'IA serait capable d'évaluer la performance académique avec une impartialité que le jugement humain ne peut atteindre. Cette promesse plaît aussi bien aux enseignant·e·s, qui y voient un moyen de se décharger d'une tâche fastidieuse, qu'aux défenseur·se·s d'une évaluation plus juste.

Cette idée que l'algorithme traiterait chaque copie de manière identique, sans émotion ni favoritisme, séduit de plus en plus. Certain·e·s défenseur·se·s avancent ainsi que les grands modèles de langage évaluent en l'absence d'affects, de fatigue et de biais (Weber et Hubbertz, 2025), et seraient donc meilleurs que les enseignant·e·s dans cette tâche. Ce discours s'inscrit dans ce que Jones-Jang et al. (2025) désignent comme les machines heuristiques, soit la tendance à attribuer aux machines une objectivité idéalisée qui ne correspond pas nécessairement à leurs propriétés réelles. Du côté des étudiant·e·s, le mythe se nourrit également de la frustration envers le système existant. Dans une étude expérimentale menée auprès de 228 étudiant·e·s d'universités sud-coréennes, Jones-Jang et al. (2025) montrent que les personnes se sentant désavantagées par les systèmes de notation traditionnels accordent davantage confiance à l'IA. Cette confiance n'est pas le fruit d'une évaluation rationnelle des performances de l'outil, mais l'expression d'une réaction émotionnelle face à un système perçu comme injuste.

À cela s'ajoute une lacune de la recherche : dans une revue systématique de 146 publications sur l'IA dans l'enseignement supérieur, Zawacki-Richter et al. (2019) constatent que la majorité des études sont conduites sans impliquer les enseignant·e·s, laissant le champ libre à des discours techno-optimistes au détriment d'une analyse critique ancrée dans les réalités de classe.

Pour quelles raisons ce mythe existe-t-il ?

La réalité des biais humains dans l'évaluation constitue le socle de ce mythe. La méta-analyse de Malouff et Thorsteinsson (2016), qui regroupe vingt études expérimentales pour un total de 1'935 correcteur·rice·s, établit un effet global modéré ($g = 0,36$), lié au genre, à l'origine ethnique, à l'attractivité physique ou aux performances passées de l'étudiant·e. Face à ce constat, l'idée d'un algorithme qui traiterait toutes les copies selon les mêmes paramètres devient séduisante. La plausibilité de surface alimente également cette croyance. Flodén (2025) rapporte qu'à première vue, certain·e·s des enseignant·e·s interviewé·e·s ont été impressionné·e·s par la correction de l'IA, au point d'estimer qu'elle aurait pu passer pour le travail d'un·e collègue.

Daly et Deglaire (2025) confirment ainsi que, sur des questions à réponse factuelle, la correction de l'IA s'aligne complètement avec celle des expert·e·s humain·e·s ; un résultat probant pour ce type d'évaluations factuelles et automatisables, mais qui ne saurait prédire les performances de l'IA face à la complexité de l'évaluation académique. Ce genre de cas entretient pourtant la croyance que l'IA pourrait réellement révolutionner l'évaluation.

Pourquoi ce mythe est-il erroné ?

Les études mobilisées utilisent majoritairement des grands modèles de langage grand public, et parfois des outils de notation à visée éducative. Elles reposent par ailleurs presque exclusivement sur des contextes d'enseignement supérieur, dans des pays occidentaux et est-asiatiques (voir Tableau 1). Ces résultats ne peuvent donc prétendre à une validité universelle, applicable à tout contexte et à toute forme d'enseignement. Leur convergence sur les points évoqués ci-dessous est en revanche suffisamment marquée pour étayer notre propos.

Un accord insuffisant avec les expert·e·s humain·e·s

Le premier argument porte sur la faiblesse de la concordance entre les notes attribuées par l'IA et celles des évaluateur·rice·s humain·e·s. Les études disponibles dressent un constat convergent : dans les contextes d'enseignement supérieur examinés, l'accord exact dépasse rarement la moitié des cas. Zacharis et Papadakis (2025) ne l'observent ainsi que dans 40,7 % des situations, alors même que la fiabilité entre les deux correcteur·rice·s humain·e·s était jugée excellente. Flodén (2025) aboutit à un résultat similaire : l'IA et l'enseignant·e ne s'accordent sur la note finale que dans 30 % des copies. De tels niveaux restent insuffisants pour des évaluations à enjeux élevés, dont les résultats impliquent des décisions déterminantes pour le parcours de l'étudiant·e.

Ce constat n'est pas propre aux outils actuels. Dès 2014, Reilly et al. documentaient déjà, dans le cadre de deux MOOCs, des taux d'accord exacts entre systèmes automatisés et correcteur·rice·s expert·e·s de 17,24 % et 13,28 % selon les cours, des chiffres si bas que les auteur·rice·s eux-mêmes déconseillaient leur usage en contexte évaluatif à enjeux. Ces résultats, antérieurs à l'émergence des grands modèles de langage, suggèrent que la difficulté est inhérente à la tâche elle-même plutôt qu'à une limite technologique surmontable : évaluer des productions écrites complexes mobilise un jugement qui résiste, par nature, à toute forme d'automatisation.

Tableau 1*Contextes des principales études empiriques mobilisées*

Auteurs·rices	Contexte (pays, niveau)	Type de productions évaluées et N	Type d'IA	Type d'étude
Daly et Deglaire (2025)	France, école de commerce	144 réponses à des questions ouvertes d'examen écrit	LLM grand public	Auto-ethnographie qualitative longitudinale
Flodén (2025)	Suède, master en logistique	463 réponses à des questions ouvertes d'examen écrit	LLM grand public	Empirique quantitative comparative (analyse comparative IA/humains, analyse de biais proportionnel)
Jones-Jang et al. (2025)	Corée du Sud, université	228 étudiant·e·s	LLM grand public	Étude empirique quantitative expérimentale (scénario fictif avec note positive/négative ; analyse de la perception d'équité via un questionnaire)
Morjaria et al. (2024)	Canada, bachelor en médecine	60 réponses à des vignettes cliniques	LLM grand public	Étude empirique quantitative comparative (corrélations IA/humains avec 4 conditions expérimentales)
Reilly et al. (2014)	États-Unis, MOOCs	423 dissertations de philosophie / 206 réponses courtes en pharmacie	Outil de notation automatisé (avant LLM)	Étude empirique quantitative (deux études comparatives de notation automatique vs enseignant·e·s ; corrélations, kappa pondéré)
Weber et Hubbertz (2025)	Allemagne, secondaire	10 textes d'élèves	IA éducative	Empirique quantitative à petite échelle (réplication partielle de Mühlhoff et Henningsen (2024) ; design test-retest)
Zacharis et Papadakis (2025)	Grèce, master en éducation	9 dissertations	LLM grand public	Empirique quantitative (notation IA vs notation humaine, accord inter-juges)

Note. « LLM grand public » désigne les grands modèles de langage accessibles au grand public (p. ex., ChatGPT, Gemini) ; « IA éducative » désigne les outils spécifiquement conçus pour la notation (p. ex., Fobizz, e-rater). Les Massive Open Online Courses (MOOCs) sont des cours en ligne ouverts à très grande échelle.

Une surévaluation systématique et une échelle compressée

Au-delà du désaccord, les études révèlent un biais systématique : l'IA tend à surévaluer les copies tout en compressant l'échelle de notation. Zacharis et Papadakis (2025), Flodén (2025), Morjaria et al. (2024) ainsi que Weber et Hubbertz (2025) confirment tous cette tendance à l'indulgence. Zacharis et Papadakis (2025) documentent ainsi une inflation moyenne de 2 à 3 points selon les copies, l'IA remontant les copies faibles vers la moyenne sans jamais atteindre les notes hautes, ce qui réduit l'écart séparant les meilleures productions. Morjaria et al. (2024) observent dans le même sens que, sans rubrique détaillée, l'IA tend à attribuer des scores plus élevés, avec un effet de taille notable ($\eta^2 = 0,33$). Concrètement, là où les évaluateur·rice·s humain·e·s mobilisent l'ensemble de l'échelle, l'IA évite les notes basses et attribue la plupart de ses notes autour de la moyenne. Flodén (2025) observe ainsi que l'IA n'attribue jamais la note minimale: un système incapable de signaler une production très insuffisante perd dès lors une partie de sa fonction évaluative. Ceci contredit directement l'idée selon laquelle l'IA évaluerait avec la même précision et la même rigueur qu'un·e correcteur·rice humain·e expérimenté·e.

Au-delà de cette compression systématique émerge cependant un problème plus fondamental : l'IA ne parvient pas à demeurer cohérente dans les jugements qu'elle porte sur une même copie.

Une incohérence interne révélatrice

L'incohérence interne constitue l'argument le plus direct contre la prétendue objectivité de l'IA. Flodén (2025) comme Daly et Deglaire (2025) observent qu'en soumettant la même copie plusieurs fois au même outil, entre un tiers et la moitié des étudiant·e·s reçoivent une note différente d'un passage à l'autre. Or l'attente envers un·e évaluateur·rice dit « objectif·ve » serait précisément qu'il ou elle attribue systématiquement la même note au même travail. Flodén (2025) le confirme : l'IA présente une cohérence interne modérée et ne s'accorde avec les notes des enseignant·e·s que dans trois cas sur dix.

Cette instabilité pose un problème de justice procédurale : deux étudiant·e·s ayant rendu un travail similaire peuvent en effet recevoir des notes différentes selon le moment où la copie est soumise, ce qui contrevient au principe d'égalité de traitement attendu d'une évaluation scolaire ou académique. Daly et Deglaire (2025) relèvent par ailleurs plusieurs défaillances qualitatives : des feedbacks portant sur des réponses inexistantes, des erreurs arithmétiques élémentaires, et des scores apparemment justes mais issus de deux erreurs qui se neutralisent. Ces

constats invalident l'image d'un·e juge fiable et cohérent·e que le mythe attribue à l'IA.

Des limites de validité : l'IA ne mesure pas ce qu'elle prétend évaluer

Au-delà de la fiabilité, c'est la validité de construit qui pose problème : l'IA mesure-t-elle vraiment ce qu'elle est censée évaluer ? Plusieurs études convergent pour répondre que non. Weber et Hubbertz (2025) rapportent un cas particulièrement révélateur : parmi dix textes soumis à trois enseignant·e·s et à une IA éducative, le texte hors sujet et incohérent a été repéré et sous-noté par l'ensemble des correcteur·rice·s humain·e·s, alors que l'IA a échoué à le détecter et a eu tendance à le surestimer. Flodén (2025) observe le même phénomène à plus grande échelle : dans les cas de grandes divergences entre l'IA et l'humain·e, environ la moitié des réponses concernées étaient hors sujet. Ces réponses étaient factuellement correctes mais sans rapport avec le contexte du cours, et l'IA les notait correctement sans repérer qu'elles ne répondaient pas à la question posée.

Pour les productions écrites, le problème se prolonge davantage : l'IA semble fonder son jugement sur des indices de surface plutôt que sur la qualité du raisonnement. Zacharis et Papadakis (2025) montrent qu'elle se concentre sur la fluidité syntaxique, la longueur et la structure. Flodén (2025) confirme un effet de longueur préoccupant, mais inversé par rapport à ce qu'on pourrait imaginer : dans son corpus, l'IA favorise les réponses courtes, alors que les enseignant·e·s valorisent les développements approfondis. Swiecki et al. (2022) ainsi que Yin et al. (2025) relèvent que ces systèmes peinent à saisir des nuances telles que la créativité, la pensée critique et l'argumentation, qui font pourtant partie des compétences valorisées dans l'enseignement supérieur. Or ces compétences supposent une forme de jugement contextuel et de prise en compte de l'intention de l'auteur·rice que les indices de surface ne capturent pas : un texte fluide et bien structuré peut être conceptuellement vide, tandis qu'un texte moins abouti dans sa formulation peut témoigner d'une réflexion réelle. L'IA ne peut donc être considérée comme un juge totalement optimal, même dans les meilleures conditions d'utilisation.

Des biais structurels invisibles mais persistants

Le dernier argument majeur concerne les biais structurels reproduits par les systèmes algorithmiques. Le mythe suppose que la machine serait insensible aux caractéristiques de la personne qui a rédigé le travail. Cette idée est pourtant trompeuse : ce n'est pas parce que les biais sont invisibles qu'ils sont inexistants. Les grands modèles de langage sont façonnés par leurs données d'entraînement, elles-mêmes imprégnées de biais humains.

Dans sa revue systématique, Idowu (2023) montre que des algorithmes prédictifs de la performance universitaire prédisaient de meilleurs résultats pour les hommes que pour les femmes, un exemple parlant des biais de genre inscrits, tant au niveau algorithmique qu'au niveau des données, dans ces systèmes.

Il serait tentant de conclure, à partir de la méta-analyse de Malouff et Thorsteinsson (2016) qui documente un biais humain global significatif dans la notation, que l'IA constituerait une alternative plus neutre. Cet argument résiste mal à l'examen : comme le montre Idowu (2023), les biais ne disparaissent pas avec l'IA, ils se déplacent et se camouflent. Par le mécanisme des indicateurs indirects, un système peut ainsi discriminer selon l'origine sociale ou le genre sans jamais accéder directement à ces variables, en s'appuyant sur des corrélats comme le code postal ou le parcours scolaire antérieur. L'objectivité apparente de l'IA ne traduit donc pas une absence de biais, mais une invisibilisation du biais.

Une simple suppression des données personnelles (p. ex., genre, origine ethnique) ne suffit pas à neutraliser ces biais. Idowu (2023) qualifie cette approche de « *fairness by unawareness*⁷ » (p.1511) et la juge insuffisante : les biais persistent à travers des corrélations avec des variables apparemment neutres. L'auteur précise même qu'il est mathématiquement impossible de satisfaire simultanément plusieurs critères formels d'objectivité. Dans leur revue systématique, Yin et al. (2025) illustrent ces biais à travers plusieurs études de cas issues de la littérature : des systèmes de notation automatisée de rédactions pénalisent systématiquement les étudiant·e·s dont l'anglais est la langue seconde, tandis que des modèles prédictifs entraînés sur des données d'écoles urbaines généralisent mal aux contextes ruraux ou défavorisés.

Les biais observés ne relèvent donc pas du hasard. Zajko (2022) rappelle que les algorithmes tendent à perpétuer l'ordre social existant plutôt qu'à le remettre en question. Le problème est que cette reproduction des inégalités peut se faire de manière parfaitement cohérente : un système peut traiter toutes les copies de façon uniforme tout en reproduisant les inégalités sociales existantes.

⁷ Objectivité par ignorance [traduction libre]

Quels sont les problèmes engendrés par ce mythe ?

La croyance en ce mythe a des conséquences bien réelles. Un·e juge impartial·e doit pouvoir justifier ses décisions en exposant le raisonnement qui les sous-tend. Or les systèmes opaques violent ce principe fondamental de l'explicabilité : on ne peut comprendre pourquoi une décision a été prise. L'opacité pose cependant un problème plus fondamental encore : sans accès aux données et aux paramètres qui structurent le jugement, on ne peut pas vérifier ni l'objectivité informationnelle du système, ni même l'équité de ses résultats (Lünich et Keller, 2024). L'absence de transparence empêche donc à la fois de justifier comment la décision a été prise et d'examiner sur quelle base elle l'a été. Dans leur étude, Daly et Deglaire (2025) montrent ainsi que l'enseignant·e déplore de ne pas avoir accès au processus interne de correction de l'IA, ce qui contraint à s'y fier aveuglément.

L'opacité ne concerne donc pas seulement la technique : Lünich et Keller (2024) montrent que les étudiant·e·s peuvent percevoir un système d'IA comme juste même quand il ne l'est pas, simplement parce qu'elles·ils font confiance à l'université qui l'implémente. La responsabilité de la transparence algorithmique incombe dès lors à plusieurs acteur·rice·s : institutions, concepteur·rice·s, chercheur·euses en éducation et enseignant·e·s.

Le deuxième problème tient à l'image de neutralité de l'IA entretenue par ce mythe. Quand un algorithme est perçu comme objectif, ses décisions semblent souvent plus fiables et plus justes (Jones-Jang et al., 2025), une autorité que le jugement humain n'a pas. Le biais algorithmique devient alors plus difficile à contester, parce qu'il se dissimule derrière une apparence de neutralité. Comme le souligne Zajko (2022), une discrimination socialement invisible est plus dangereuse qu'un biais humain visible et contestable. Zhang (2024) illustre concrètement ce mécanisme : lors du scandale des A-levels britanniques de 2020, un algorithme présenté comme objectif attribuait des notes en fonction des performances historiques de l'établissement, pénalisant ainsi systématiquement les élèves issu·e·s de milieux défavorisés, dans une parfaite opacité.

Le troisième problème concerne directement les étudiant·e·s : les erreurs de notation de l'IA ne sont pas anodines. Zacharis et Papadakis (2025) relèvent qu'environ 5 % des copies ont reçu de l'IA une note très éloignée de celle de l'humain·e, avec des écarts parfois assez importants pour faire basculer un travail de la réussite à l'échec, ou inversement. Sur une promotion de 200 personnes, ce sont ainsi une dizaine de parcours potentiellement affectés par ce type d'erreur de classification.

Enfin, Kaufman et al. (2025) montrent que l'adoption des outils d'IA varie fortement selon les ressources des établissements aux États-Unis. Cette inégalité d'accès se superpose aux biais décrits précédemment, avec le risque que les outils les plus discutables soient déployés dans les contextes les moins armés pour les contester.

Pistes de réflexion pour contrer le mythe et pour l'explorer

Déconstruire ce mythe ne revient pas à rejeter tout usage de l'IA dans l'évaluation, mais à renoncer à l'idée qu'une impartialité parfaite serait accessible grâce à la machine. Lünich et Keller (2024) soulignent que, pour qu'un système d'IA puisse évaluer, il doit d'abord être explicable : les humain·e·s doivent pouvoir comprendre sur quoi il fonde ses décisions, condition nécessaire pour identifier les biais et concevoir des algorithmes plus justes. Tang et al. (2024) approfondissent cette piste en montrant que l'IA s'aligne davantage sur les évaluateur·rice·s humain·e·s lorsqu'on lui communique des critères d'évaluation structurés et que chaque décision est justifiée. Cette approche combine explicabilité et définition claire des standards : l'IA ne décide plus en « boîte noire », mais selon des règles explicites que les humain·e·s ont fixées.

Concrètement, deux modalités de mise en œuvre méritent attention. D'une part, les grilles de notation détaillées : la méta-analyse de Malouff et Thorsteinsson (2016) montre qu'elles réduisent significativement les biais dans l'évaluation humaine. D'autre part, l'application stricte de ces mêmes critères à l'IA : Morjaria et al. (2024) confirment que la surévaluation par l'IA diminue quand les critères sont explicites et exhaustifs. Cette approche a néanmoins ses limites. Zacharis et Papadakis (2025), Reilly et al. (2014) et Flodén (2025) le rappellent : les critères structurés atténuent les biais sans les éliminer. L'IA doit donc rester subordonnée au jugement humain dans toute évaluation sommative à enjeux élevés. Meirmanova (2025) le synthétise ainsi : l'IA fonctionne comme un complément à la réflexion pédagogique, non comme son substitut.

Une vigilance institutionnelle s'impose également : avant tout déploiement, des audits systématiques des biais algorithmiques devraient être menés pour vérifier que le système ne reproduit pas les inégalités existantes. La recherche elle-même doit enfin s'élargir. La quasi-totalité des études disponibles porte sur l'enseignement supérieur en contextes occidentaux et est-asiatiques ; il serait essentiel d'examiner également l'enseignement obligatoire, des contextes culturels sous-représentés, ainsi que les effets longitudinaux d'une exposition à l'évaluation partiellement automatisée.

Edenberg et Wood (2023) soulignent par ailleurs que neutralité et justice ne sont pas équivalentes : traiter chaque copie de la même façon ne garantit pas des résultats justes pour des étudiant·e·s aux profils différents. Une piste de recherche future pourrait dès lors consister à examiner si des systèmes capables d'ajuster leurs critères selon les profils permettraient d'atteindre une forme d'équité substantive, à condition qu'une objectivité solide et auditée soit d'abord garantie.

Références

- Daly, P. et Deglaire, E. (2025). AI-enabled correction: A professor's journey. *Innovations in Education and Teaching International*, 62(4), 1241–1257. <https://doi.org/10.1080/14703297.2024.2390486>
- Edenberg, E. et Wood, A. (2023). Disambiguating algorithmic bias: From neutrality to justice. Dans *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (p. 691–704). Association for Computing Machinery. <https://doi.org/10.1145/3600211.3604695>
- Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, 51, 201–224. <https://doi.org/10.1002/berj.4069>
- Idowu, J. A. (2023). Debiasing education algorithms. *International Journal of Artificial Intelligence in Education*, 34(4), 1510–1540. <https://doi.org/10.1007/s40593-023-00389-4>
- Jones-Jang, S. M., Chung, M., Choi, J., Kim, N. et Lee, S. (2025). Fairness perceptions of AI in grading systems: Examining how discontent with the status quo and outcome favorability reduce AI reluctance. *Computers and Education: Artificial Intelligence*, 8, Article 100419. <https://doi.org/10.1016/j.caeai.2025.100419>
- Kaufman, J. H., Woo, A., Eagan, J., Lee, S. et Kassan, E. B. (2025). *Uneven adoption of artificial intelligence tools among U.S. teachers and principals in the 2023–2024 school year* (Report No. RR-A134-25). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA134-25.html
- Lünich, M. et Keller, B. (2024). Explainable artificial intelligence for academic performance prediction: An experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions. Dans *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (p. 1752–1763). Association for Computing Machinery. <https://doi.org/10.1145/3630106.3658953>
- Malouff, J. M. et Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, 60(3), 245–256. <https://doi.org/10.1177/0004944116664618>
- Meirmanova, N. (2025). Analyse comparative entre la génération automatique de questionnaires à choix multiples par ChatGPT et le corps enseignant pour l'évaluation de l'apprentissage. *Revue internationale des technologies en pédagogie universitaire*, 22(1), 1–13. <https://doi.org/10.18162/ritpu-2025-v22n1-03>
- Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M. et Sibbald, M. (2024). Examining the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, 3(1), 32–43. <https://doi.org/10.3390/ime3010004>

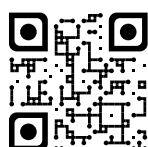
- Reilly, E., Stafford, R., Williams, K. et Corliss, S. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open and Distributed Learning*, 15(5), 83–98. <https://doi.org/10.19173/irrodl.v15i5.1857>
- Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N. et Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, Article 100075. <https://doi.org/10.1016/j.caeai.2022.100075>
- Tang, X., Chen, H., Lin, D. et Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14), Article e34262. <https://doi.org/10.1016/j.heliyon.2024.e34262>
- Weber, F. et Hubbertz, H. (2025). Humans vs. AI in grading students' texts: A pilot study with 3 teachers against one AI. Dans *Proceedings of the 22nd International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2025)* (p. 128–134). IADIS Press. <https://files.eric.ed.gov/fulltext/ED677814.pdf>
- Yin, Z., Chinta, S. V., Wang, Z., Gonzalez, M. et Zhang, W. (2025). *FairAIED: Navigating fairness, bias, and ethics in educational AI applications* [Préimpression]. arXiv. <https://doi.org/10.48550/arXiv.2407.18745>
- Zacharis, G. et Papadakis, S. (2025). Can AI grade like a human? Validity, reliability, and fairness in university coursework assessment. *Educational Process: International Journal*, 19, Article e2025591. <https://doi.org/10.22521/edupij.2025.19.591>
- Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3), Article e12962. <https://doi.org/10.1111/soc4.12962>
- Zawacki-Richter, O., Marín, V. I., Bond, M. et Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhang, X. (2024). Fairness and effectiveness in AI-driven educational assessments: Challenges and mitigation strategies. *Academic Journal of Science and Technology*, 12(1), 113–116. <https://doi.org/10.54097/adcxze34>

Dans la continuité du premier volume de la série « Mythologie pédagogique », cet ouvrage collectif poursuit son travail d'analyse critique des idées reçues en portant cette fois son regard sur les fascinations, craintes et promesses associées à l'intelligence artificielle en éducation.

Au travers de 11 chapitres, l'ouvrage explore des enjeux essentiels tels que le remplacement supposé des enseignantes et enseignants, l'impact de l'IA sur la créativité et l'autonomie des élèves, ou encore les capacités d'évaluation de l'IA et de prise en compte des émotions.

L'ouvrage s'appuie sur une démarche fondée sur des travaux scientifiques. Fruit des travaux d'étudiantes et étudiants de Master en sciences de l'éducation à l'Université de Fribourg, chaque chapitre a parcouru un protocole rigoureux de relecture et de révision par les pairs. Cette dynamique scientifique est enrichie par les contributions de trois spécialistes en sciences de l'éducation et de la formation, qui apportent des éclairages complémentaires sur ces enjeux.

Loin de l'enthousiasme aveugle et des discours alarmistes, ce second volume de la série « Mythologie pédagogique » offre aux équipes éducatives, aux chercheurs et chercheuses ainsi qu'aux étudiants et étudiantes des repères solides pour penser l'apprentissage à l'ère de l'intelligence artificielle avec recul et discernement.



**PLE
MOS
Lab** Processes of learning,
motivation and self-regulation
www.plemos.education

