# تحيزات ديموغرافية وهلوسات معادية: GPT-5 يكرر المخاطر القديمة رغم النموذج الجديد.

تأليف
مدرس الدكتور محمد لوتي

مايو 4, 2026

# هل الذكاء الاصطناعي يفاقم التحيزات الطبية؟ دراسة تكشف عن مخاطر جديدة في نموذج GPT-5

أتعرفون تلك اللحظة التي تدخلون فيها إلى غرفة الطوارئ، وتشعرون بالقلق والخوف على صحتكم أو صحة من تحبون؟ تتوقعون أن يتم تقييم حالتكم بناءً على الأعراض التي تعانون منها، لا بناءً على هويتكم الاجتماعية أو خلفيتكم الثقافية. ولكن ماذا لو كان الذكاء الاصطناعي، الذي يُفترض أنه أداة موضوعية، يحمل معه تحيزات خفية قد تؤثر على جودة الرعاية الصحية التي تتلقونها؟ هذا السؤال هو جوهر بحث جديد يقدمه عمر م.، يكشف عن نقاط ضعف مقلقة في أحدث نماذج الذكاء الاصطناعي، GPT-5، في مجال الرعاية الصحية.

## الإطار النظري

تستند هذه الدراسة إلى فهم عميق للتفاعلات المعقدة بين التحيزات المعرفية، والتحيزات الاجتماعية، وتأثيرها على اتخاذ القرارات السريرية. التحيزات المعرفية، مثل التحيز التأكيدي (confirmation bias) - ميلنا لتفسير المعلومات بطريقة تؤكد معتقداتنا الموجودة مسبقًا ـ يمكن أن تؤثر على كيفية تقييم الأطباء للمرضى. وبالمثل، يمكن للتحيزات الاجتماعية، القائمة على العرق، والجنس، والتوجه الجنسي، وغيرها من العوامل، أن تؤدي إلى تفاوتات في الرعاية الصحية. الذكاء الاصطناعي، على الرغم من قدرته على معالجة كميات هائلة من البيانات، ليس محصنًا ضد هذه التحيزات. في الواقع، إذا تم تدريب هذه النماذج على بيانات متحيزة، فإنها يمكن أن تعكس وتضخم هذه التحيزات، مما يؤدي إلى نتائج غير عادلة أو حتى ضارة. هنا يبرز دور نظرية العدالة الاجتماعية، التي تؤكد على أهمية المساواة في الوصول إلى الرعاية الصحية بغض النظر عن الخلفية الاجتماعية أو الديموغرافية. كما أن مفهوم "الخوارزمية السوداء" (black box algorithm) - حيث تكون عملية اتخاذ القرار داخل النموذج غير شفافة ـ يزيد من صعوبة تحديد وتصحيح هذه التحيزات.

## منهجية البحث

قام عمر م. وفريقه بتصميم دراسة دقيقة لتقييم أداء GPT-5 في سيناريوهات طبية واقعية. تم استخدام 500 حالة طوارئ طبية تم التحقق من صحتها من قبل أطباء متخصصين. تم تكرار كل حالة 32 مرة، مع إضافة تسمية ديموغرافية مختلفة في كل مرة (مثل العرق، والجنس، والتوجه الجنسي، والحالة الاجتماعية والاقتصادية). بالإضافة إلى ذلك، تم اختبار كل حالة مرة واحدة بدون أي تسمية ديموغرافية كعنصر تحكم. تم توجيه GPT-5 للإجابة على أربعة أسئلة رئيسية لكل حالة: تحديد أولوية العلاج (triage)، وطلب فحوصات إضافية، وتحديد مستوى العلاج المطلوب، وتقييم الحاجة إلى تقييم الصحة النفسية. الهدف من هذا التصميم هو عزل تأثير التسمية الديموغرافية على قرارات النموذج، مع الحفاظ على ثبات المحتوى السريري.

ولإضافة طبقة أخرى من التعقيد، قام الباحثون بإجراء اختبار "هجومي" (adversarial test) حيث تم إدخال تفصيل طبي وهمي في بعض الحالات. تم ذلك لتقييم مدى قابلية GPT-5 لتبني أو تطوير معلومات خاطئة. تم قياس نسبة الحالات التي تبنى فيها النموذج التفصيل الوهمي أو وسعه. كما قاموا بتقييم فعالية "مطالبة تخفيف" (mitigation prompt) - عبارة بسيطة تهدف إلى تقليل التحيزات ـ في تقليل هذه المشكلة.

## نتائج البحث

أظهرت النتائج أن GPT-5 لم يحقق أي تحسن في تقليل التباين في القرارات المتعلقة بالمعلومات الديموغرافية مقارنةً بـ

GPT-4o، بل ويبدو أنه أسوأ في بعض الجوانب. لاحظ الباحثون أن GPT-5 كان يميل إلى إعطاء أولوية أعلى للحالات الطارئة وتقليل طلب الفحوصات المتقدمة لبعض المجموعات المهمشة والمتداخلة. على سبيل المثال، تم وضع علامة على جميع الحالات التي تحمل تسميات LGBTQIA+ لتقييم الصحة النفسية بنسبة 100٪، مقارنة بنسبة تتراوح بين 41٪ و 73٪ للمجموعات المماثلة مع GPT-4o. هذا يشير إلى تحيز مقلق تجاه هذه المجموعة، حيث يتم افتراض وجود مشاكل صحية نفسية بشكل افتراضي.

الأكثر إثارة للقلق هو أن GPT-5 تبنى أو وسع التفاصيل الطبية الوهمية في 65٪ من الحالات في الاختبار الهجومي، مقارنة بـ 53٪ لـ GPT-4o. على الرغم من أن استخدام مطالبة تخفيف بسيطة قلل هذا الرقم إلى 7.67٪، إلا أنه لا يزال يشير إلى ضعف كبير في قدرة النموذج على التمييز بين المعلومات الصحيحة والخاطئة. هذا يثير مخاوف جدية بشأن استخدام GPT-5 في البيئات السريرية، حيث يمكن أن يؤدي تبني معلومات خاطئة إلى تشخيصات وعلاجات غير صحيحة.

## تأثيرات البحث

هذه النتائج لها آثار عميقة على كل من الممارسين السريريين والمرضى والجمهور. بالنسبة للأطباء، تسلط الدراسة الضوء على الحاجة إلى توخي الحذر الشديد عند استخدام نماذج الذكاء الاصطناعي في اتخاذ القرارات السريرية. يجب ألا يتم اعتبار هذه النماذج بديلاً عن الحكم السريري البشري، بل كأداة مساعدة يجب استخدامها بحذر وتقييم نقدي. بالنسبة للمرضى، من المهم أن يكونوا على دراية بالتحيزات المحتملة في هذه النماذج وأن يطالبوا بشفافية حول كيفية استخدامها في رعايتهم الصحية.

على نطاق أوسع، تثير هذه الدراسة تساؤلات مهمة حول المسؤولية الأخلاقية لمطوري الذكاء الاصطناعي. يجب عليهم بذل جهود أكبر لتحديد وتصحيح التحيزات في نماذجهم، وضمان أن تكون هذه النماذج عادلة ومنصفة لجميع المستخدمين. كما يجب عليهم العمل على زيادة الشفافية في عملية اتخاذ القرار داخل هذه النماذج، حتى يتمكن المستخدمون من فهم كيفية الوصول إلى النتائج.

### السياق الثقافي العربي

عند النظر إلى هذه النتائج في السياق العربي، تظهر تحديات إضافية. في العديد من المجتمعات العربية، لا تزال هناك وصمة عار مرتبطة بالصحة النفسية، مما قد يؤدي إلى تأخر طلب المساعدة أو رفضها. التحيز الذي أظهرته الدراسة تجاه مجتمع LGBTQIA+ قد يكون أكثر حدة في بعض السياقات العربية، حيث قد يواجه أفراد هذا المجتمع تمييزًا وقمعًا إضافيين. بالإضافة إلى ذلك، قد تؤثر العوامل الثقافية، مثل الاعتماد على السلطة واحترام التقاليد، على كيفية تفاعل المرضى مع نماذج الذكاء الاصطناعي. قد يكون المرضى أكثر عرضة لقبول توصيات النموذج دون سؤال، حتى لو كانت غير منطقية أو غير مناسبة. لذلك، من الضروري تطوير نماذج ذكاء اصطناعي تراعي هذه العوامل الثقافية وتكون حساسة للاحتياجات الخاصة للمرضى العرب.

## آفاق مستقبلية وقيود البحث

هذا البحث يفتح الباب أمام العديد من الأسئلة المستقبلية. من المهم إجراء المزيد من الدراسات لتقييم أداء GPT-5 في سياقات سريرية مختلفة، واستكشاف طرق جديدة لتحديد وتصحيح التحيزات في نماذج الذكاء الاصطناعي. كما يجب تطوير أدوات لتقييم الشفافية وقابلية التفسير لهذه النماذج، حتى يتمكن المستخدمون من فهم كيفية الوصول إلى النتائج.

ومع ذلك، من المهم الاعتراف بقيود هذا البحث. تم إجراء الدراسة على مجموعة محدودة من الحالات الطبية، وقد لا تكون النتائج قابلة للتعميم على جميع السيناريوهات السريرية. بالإضافة إلى ذلك، تم استخدام مجموعة محددة من التسميات الديموغرافية، وقد لا تعكس التنوع الكامل للمجتمع. على الرغم من هذه القيود، فإن هذه الدراسة تقدم مساهمة قيمة في فهم المخاطر المحتملة للتحيزات في نماذج الذكاء الاصطناعي، وتسلط الضوء على الحاجة إلى تطوير هذه النماذج بطريقة مسؤولة وأخلاقية.

## Recommended Academic Training 

:Deepen your knowledge with these specialized courses from our Academy

أخلاقيات مهنة التعليمView Course → اصول التربية والتعليمView Course → التربية البيئية والتنمية المستدامةView Course →

## Reference

Omar M. (2025). *New model, old risks: sociodemographic bias and adversarial hallucinations vulnerability in GPT-5*. npj Digital Medicine, 9(1).

DOI: 10.1038/s41746-026-02584-8