المعرّف الرقمي DOI
https://doi.org/10.31430/VIFV3059

بيتر كونيغز | Peter Königs*
ترجمة عمر المغربي | Omar Almagharebi**

أزمة التشاؤمية في حقل أخلاقيات الذكاء الاصطناعي

The Negativity Crisis of AI Ethics

ملخص: على الرغم من الإمكانات الإيجابية الهائلة التي ينطوي عليها الذكاء الاصطناعي، فقد قدَّم مجتمع أخلاقيات الذكاء الاصطناعي صورة قاتمة إلى حدٍّ بعيد عن تداعياته الأخلاقية. تفحص هذه الدراسة النزعة السلبية السائدة داخل هذا الحقل من منظور فلسفة العلم. وتتأتّى هذه السلبية الطاغية من الطريقة الخاصَّة التي يُنظَّم بها مؤسَّسيًا، والتي تُجبر المشتغلين به على تصويره وفق منظور سلبي. وغدت الصورة العامة التي تقدمها أدبيات أخلاقيات الذكاء الاصطناعي سلبيّة وذات منظور أحادي. بناءً عليه، ينبغي التشكيك في السردية التشاؤمية المتداولة، والبحث عن طرائق لإصلاح المنظومة التي أفرزتها.

كلمات مفتاحية: الذكاء الاصطناعي، فلسفة العلم، الإبستيمولوجيا الاجتماعية، الذكاء الماندیفي، التحيز، أخلاقيات الذكاء الاصطناعي.

Abstract: Despite the great positive potential of AI, the AI ethics community has presented a rather gloomy picture of AI's ethical implications. This paper examines the negativity within AI ethics through a philosophy of science lens. The prevailing negativity is a result of the particular way the discipline is institutionally organized, which pressures AI ethicists to portray AI in a critical light. As a consequence, the over all picture of AI offered by the AI ethics community is one–sided and negatively biased. We should be skeptical about the negative narrative promoted by AI ethics and explore ways of reforming the system.

Keywords: Artificial Intelligence, Philosophy of Science, Social Epistemology, Mandevillian Intelligence, Bias, Ethics of AI.

* قسم الفلسفة والعلوم السياسية، جامعة دورتموند التقنية، دورتموند، ألمانيا.
Department of Philosophy and Political Science, TU Dortmund University, Dortmund, Germany.
Email: peter.koenigs@tu–dortmund.de

** باحث مساعد في المركز العربي للأبحاث ودراسة السياسات.
Assistant Researcher, Arab Center for Research and Policy Studies.
Email: omar.almagharebi@dohainstitute.edu.qa

Peter Königs, "The Negativity Crisis of AI Ethics," *Synthese*, vol. 206, no. 277 (2025).
https://doi.org/10.1007/s11229–025–05378–9

مقدمة

يتزايد المزاجُ التشاؤمي وسط المنتسبين إلى الذكاء الاصطناعي اطرادًا مع ازدهار المبحث المتعلق بأخلاقيّاته. يكفي أن يُقلِّب المرء أعداد دوريات "أخلاقيات التقنية" حتى يَخلُصَ إلى أن تطور الذكاء الاصطناعي يبدو، قبل كل شيء، مدعاةً للقلق. فالنقاشاتُ الكبرى تدور، تقريبًا بلا استثناء، حول قائمةٍ مطولة من المشكلات، نذكر منها تمثيلًا لا حصرًا: فجواتِ المسؤولية Responsibility Gaps[1]، أو مسائل الشفافية والثقة، أو الإحلال الوظيفي، أو ضمور الملكة الأخلاقية Moral Deskilling[2]، أو صور الظلم الجديدة، أو اعتبارات الخصوصيّة، أو إمكانيات التلاعب، أو معضلةَ المواءمة The Alignment Problem[3]. أمّا المنظورات الإيجابية فتبدو قليلة بل شحيحة.

على الرغم من أن الذكاء الاصطناعي ينطوي على ممكناتٍ هائلة ونافعة للإنسان، فإنّ عشرات السبل التي في وسعها جعل العالم أعدل وأفضل تظلّ، في أدبيات الحقل، عباراتٍ عابرة تُستدرك عادةً بـ "لكن" قلقة، تمثّل مهادًا لحديثٍ مطوّل عن المخاطر وضرورات الضبط والتشريع الأخلاقي والقانوني. ربما كان من الحري بوعود التقنية إن لم يُلهب حماسًا خالصًا، فعلى الأقل، أن يُلهم قدرًا من التفاؤل والثقة الحذرين، بيد أن المزاج السائد في حقل أخلاقيات الذكاء الاصطناعي يوصف على لسان أحد أبرز الباحثين بأنه "مَدٌّ متصاعدٌ من الذعر"[4]، وذلك تبعًا لما يستظهره عنوان معبّر لمقدمة عدد خاص لأحد الدوريّات، نُشر مؤخرًا، حيث يقدِّم مراجعة لحالة هذا الحقل: "أخلاقيات الذكاء الاصطناعي: إشكاليات متفاقمة، إشكاليات متجدِّدة، إشكاليات غير

(1) مصطلح صاغه الفيلسوف أندرياس ماتياس في سياق الإشكالات التي تطرحها أنظمة التعلُّم الآلي، حيث لا يستطيع المُبرمج التنبؤ بسلوكها المستقبلي. ويشير إلى إشكالية عدم إمكانية إسناد المسؤولية الأخلاقية أو القانونية إلى أي شخص عن الأضرار التي تسببها الأنظمة ذاتية التعلّم. (المترجم). ينظر:

Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics Inf Technol*, vol. 6 (2004), pp. 175–183.

(2) مصطلح طورته الفيلسوفة شانون فالور، مُستلهمة من النقاشات الاجتماعية حول ضمور الملكات الاقتصادية في القرن العشرين، ويشير المصطلح إلى تراجع القدرة على اتخاذ القرارات الأخلاقية بسبب قلة الممارسة والتجربة، وذلك نتيجة تفويض عمليات صنع القرار لتقنيات الذكاء الاصطناعي. (المترجم)، ينظر:

Shannon Vallor, "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character," *Philos Technol*, vol. 28 (2015), pp. 107–124.

(3) أحد أبرز الإشكالات التي تطرحها عمليات الأتمتة في مجال أبحاث سلامة الذكاء الاصطناعي، ويُقصد بها التحدي المُتمثِّل في ضمان عمل أنظمة الذكاء الاصطناعي وفقًا للأهداف والقيم والمبادئ الإنسانية، وكيف يُمكن الحفاظ عليها في حدود السيطرة الإنسانية [المترجم]. اعتمدنا في ترجمة جميع المصطلحات التقنية على: **معجم البيانات والذكاء الاصطناعي إنجليزي – عربي**، الهيئة السعودية للبيانات والذكاء الاصطناعي (2024)، ط 2، شوهد في 2026/1/2، في: https://acr.ps/1L9F2ur

(4) John Danaher, "The Rise of the Robots and the Crisis of Moral Patiency," *AI & Society*, vol. 34 (2019), p. 129.

مسبوقة"[5]. وباختصار، فإنّ أخلاقيّات الذكاء الاصطناعي تتمحور بكليّتها حول الإشكاليّات المنبثقة منه[6].

بطبيعة الحال، هذا لا يعني غياب مقارباتٍ متفائلة، فثمة كتابات عدة، أكاديمية وغير أكاديمية، تُدافع عن الوجه المشرق للذكاء الاصطناعي، نذكر على سبيل المثال: "آلات مفعمة بالنعمة والمحبة" لداريو أمودي[7]، و"مانفيستو التفاؤل التقني" لمارك أندريسن[8]، و**من التشاؤم إلى الوعد** لبيال أرورا[9]، و**الفاعلية الفائقة** لرايد هوفمان وكريغ بيتو[10]، و**آلة المساواة العادلة** لأورلي لوفل[11]. غير أن أيًا من هؤلاء ليس فيلسوفًا أكاديميًا، ومن يفتّش عن أسبابٍ وجيهة للتفاؤل بالذكاء الاصطناعي سيعثر عليها، في غالب الأمر، خارج الفلسفة الأكاديمية. ففكرةُ أخلاقيات إيجابية للذكاء الاصطناعي ترد أحيانًا، لكنها تظلّ حاضرةً على استحياء، ولا تكاد تحتلّ موقعًا بؤرويًّا في الحقل[12].

يبتغي هذا البحث مساءلة النزعة السلبية التي تخيم على جماعة الباحثين في أخلاقيات الذكاء الاصطناعي. وسأسلك لتحقيق ذلك طريقةً غير مباشرة. فلن أتبنى، على طريقة "المتفائلين بالتقنية"،

(5) لستُ هنا بصدد انتقاد لوتشيانو فلوريدي؛ إذ هو محقٌّ في تلخيص أدبيات أخلاقيات الذكاء الاصطناعي بأنها تدور في معظمها حول المشكلات؛

Luciano Floridi, "Introduction to The Special Issues: The Ethics of Artificial Intelligence," *American Philosophical Quarterly*, vol. 61 (2024).

(6) يُمكن الإحساس بالنزعة السلبية في كثير من النقاشات المابعدية Meta–Discussions حول أخلاقيات الذكاء الاصطناعي بوصفها مشروعًا فلسفيًا، حيث تميل إلى إبراز المخاطر والجوانب السلبية، ينظر:

Thilo Hagendorff, "Blind Spots in AI Ethics," *AI and Ethics*, vol. 2 (2022); Luke Munn, "The Uselessness of AI Ethics," *AI and Ethics*, vol. 3 (2023);

وبطريقةٍ كاشفة، يحدّد ثيلو هاغندورف "القوّة الفعلية" لأخلاقيات الذكاء الاصطناعي في "حساسيتها تجاه الأذى والمعاناة، وقدرتها على رصد الآثار الخارجية [السلبية]"، ينظر: Hagendorff, p. 862.

(7) Dario Amodei, "Machines of Loving Gracem," 13/10/2024, accessed on 19/1/2025, at: https://acr.ps/1L9F2Pk

(8) Marc Andreessen, "The Techno–Optimist Manifesto," *Andreessen Horowitz*, 16/10/2023, at: https://acr.ps/1L9F2sv

(9) Payal Arora, *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech* (Cambridge, MA: MIT Press, 2024).

(10) Reid Hoffman & Greg Beato, *Superagency: What Could Possibly Go Right with Our AI Future* (New York: Authors Equity, 2025).

(11) Orly Lobel, *The Equality Machine: Harnessing Digital Technology for a Brighter, more Inclusive Future* (London: Hachette UK, 2022).

(12) حول إمكان صوغ أخلاقيات إيجابية للذكاء الاصطناعي، ينظر:

Sven Nyholm, "What is This Thing Called the Ethics of AI and What Calls for It?" in: David J. Gunkel (ed.), *Handbook on the Ethics of Artificial Intelligence* (Cheltenham: Edward Elgar Publishing, 2024);

وحتى الكتبُ الفلسفية التي تتضمّن كلمة "يوتوبيا" في عناوينها وتطرح تصوّراتٍ طوباوية للتقنية، تظُّل قراءتُها ملتبسةً ومزدوجة الانطباع، ينظر:

Nick Bostrom, *Deep Utopia: Life and Meaning in a Solved World* (Ideapress, 2024); John Danaher, *Automation and Utopia: Human Flourishing in a World without Work* (Cambridge, MA: Harvard University Press, 2019); John Danaher, "Techno–optimism: An Analysis, an Evaluation and a Modest Defense," *Philosophy & Technology*, vol. 35 (2022).

قائمة الوعود التي يمكن للذكاء الاصطناعي أن يحققها، ولن أطيل الوقوف عند كلّ هاجسٍ أخلاقيٍ محددٍ في الأدبيات. سأنظر، عوض ذلك، إلى المزاج التشاؤمي من منظور فلسفة العلم، مما يقتضي خطوةً إلى الوراء لفحص البنية المؤسسية لحقل أخلاقيات الذكاء الاصطناعي باعتباره تخصّصًا أكاديميًا. سأقترح أن ثمة قيودًا مؤسسية تدفع العاملين في الحقل إلى الإلحاح على الجانب السلبي؛ إذ يتعين عليهم إبراز مشكلات التقنية، وإلّا عرّض نفسه لخطر التهميش. ومن ثمّ، جاءت الصورة التي يرسمها الحقل عن التداعيات الأخلاقية للذكاء الاصطناعي مُختلّة على مستويين: أحادية الوجه (إذ نادرًا ما تُناقش الجوانب الإيجابية) ومنحازة سلبيًا (حيث يُبالغ في تضخيم السلبيات). بناءً عليه، يجدر بنا أن نتشكّك في السردية التشاؤمية السائدة، وأن نفكّر جديًّا في سُبل إصلاح هذا الحقل حتى يستعيد توازنه.

ينقسم البحث على النحو الآتي: أبدأ بتشخيص ثلاث سمات مؤسسية في حقل أخلاقيات الذكاء الاصطناعي تفسر مجتمعة المزاج السلبي السائد فيه (القسم الثاني). ثم أبيّن أن هذا التفسير يمنحنا مسوغًا للقول إن الصورة القاتمة التي يرسمها أصحاب الاختصاص أحادية الجانب ومنحازة سلبيًا (القسم الثالث). ولإيضاح الحجة وإحاطتها بسياق أوسع، أقارن أزمة السلبية في أخلاقيات الذكاء الاصطناعي بأزمة التكرار التي هزت حقولاً علمية أخرى (القسم الرابع). ويعالج القسمان المواليان اعتراضين محتملين يشيران، كلٌّ بطريقته، إلى أن النزعة السلبية قد تكون حميدة (القسمان الخامس والسادس). أمّا الخاتمة فتستعرض الدلالة الأوسع للطرح الذي أقدمه (القسم السابع).

يتَّسق مشروعي في هذه الورقة مع المشروع الأوسع في فلسفة العلم؛ الذي يسعى إلى استجلاء الكيفية التي ينبغي أن يُنظَّم بها البحث العلمي اجتماعيًا كي يُكلَّل بالنجاح. يتعامل هذا البرنامج البحثي مع البحث العلمي بما هو مسعى جماعي يتوقَّف تقدُّمه على جملة من المعايير والمؤسَّسات والسياسات، الظاهرة منها والمستضمرة، التي تشكّله وتصوغه. وبعبارة فيليب كيتشر البرامجية Programmatic، فإن السؤال المطروح هو: "ما السبيل الأمثل لتصميم المؤسَّسات الاجتماعية بصورة تُعزِّز تقدُّم المعرفة؟"[13]، وقد دارت الإسهامات في هذا البرنامج حول مسائل شتّى، نذكر منها تمثيلاً لا تحديدًا: التوزيع الأمثل للجهود البحثية، وأنظمة المكافأة داخل المبحث العلمي[14]، والبنى

(13) Philip Kitcher, "The Division of Cognitive Labor," *The Journal of Philosophy,* vol. 87 (1990), p. 22.

(14) ينظر:

Ibid.; Miriam Solomon, *Social Empiricism* (Cambridge, MA: MIT Press, 2001); Michael Strevens, "The Role of the Priority Rule in Science," *The Journal of Philosophy,* vol. 100 (2003); Michael Weisberg & Ryan Muldoon, "Epistemic Landscapes and the Division of Cognitive Labor," *Philosophy of Science*, vol. 76 (2009); Kevin J. S. Zollman, "The Epistemic Benefit of Transient Diversity," *Erkenntnis*, vol. 72 (2010).

التواصليّة داخل الجماعات المعرفية[15]، والمعايير التي يتعين أن تحكم البحث العلمي[16]، أو الطريقة التي ينبغي أن تُنظَّم بها عملية مراجعة الأقران Peer Review وتمويل البحث[17].

ومع أن الكثير من الإسهامات في هذا البرنامج البحثي تتسم بالعمومية؛ إذ تبحث في الكيفية التي يتعين أن يُنظَّم بها البحث العلمي عامةً لبلوغ غاياته، فإن مشروع هذا المقال أكثرُ تحديدًا، فهو يحصر النظر في ميدانٍ فلسفي بعينه هو "أخلاقيات الذكاء الاصطناعي"، ومردُّ ذلك أنه يُحلّل الديناميات المؤسَّسية والإشكاليات المرتبطة بها التي أحسبها خاصَّة بهذا الحقل. ومع ذلك، فمن الجائز أن تعمم بعض الإشكاليات المحددة في أخلاقيات الذكاء الاصطناعي على أخلاقيات التقنية بمعناها الأوسع.

الديناميكيات المؤسسية في أخلاقيات الذكاء الاصطناعي

يمكن تفسير النزعة السلبية في أخلاقيات الذكاء الاصطناعي بسماتٍ مؤسسية مخصوصة تميز هذا التخصّص الأكاديمي، ويرجع ذلك إلى حد بعيد إلى القيود الداخلية للحقل التي تُكره الباحثين على تسليط الضوء على المساوئ المحتملة. وعلى نحوٍ أدقّ، ينشأ هذا الانحياز من تفاعل ثلاث سمات مؤسسية أسمّيها: "موضوع البحث" Subject–Matter، و"التأثير الإيجابي" Positive Impact، و"الحوافز" Incentives.

أوّلًا، فلنلتفت إلى "موضوع البحث" وبسبب الطبيعة الخاصة لأخلاقيات الذكاء الاصطناعي، يختلف مسارُ البحث فيها عمّا نجده في سائر الحقول الفلسفية. ففي تلك الحقول تكون الدافعيةُ التي تُشرّع مسعًى فلسفيًا، في الغالب، سؤالًا فلسفيًا لم يُحسَم بعد، وغالبًا ما يكون لغزًا عتيقًا دارت حوله الجدالات قرونًا عديدة. ففي الأخلاق، مثلًا، قد يكون السؤال: ما الذي يجعل فعلًا ما صائبًا أو خاطئًا من الناحية الأخلاقية؟ وفي الفلسفة السياسية قد يكون السؤال: ما العدالة؟ ومتى يكون تدخّل الدولة مشروعًا؟ أمّا نظرية المعرفة فتسأل عن ماهية المعرفة والكيفيّات المتعلقة بإمكان تحصيلها، بينما تتمحور نظرية الفعل Action Theory حول سؤال إمكانية تحقق الإرادة الحرّة. وفي حقل الميتافيزيقا قد يُطرح تساؤل عن مبدأ العلّية أو عن ماهية الزمن، أمّا فلاسفةُ العقل فيسعون إلى فهم علاقة العقل بالمادة. وهكذا يتحدّد القاسمُ المشترك بين هذه المباحث في أن دافع البحث الفلسفي فيها سؤالٌ مفتوح يُلزم الفلاسفة بالتماس جواب ما. والمحصلةُ المتوخّاة، في أحسن الأحوال، تتمثّل

(15) Kevin J. S. Zollman, "The Communication Structure of Epistemic Communities," *Philosophy of Science*, vol. 74 (2007).

(16) Helen E. Longino, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (Princeton: Princeton University Press, 1990), Ch. 4; Helen E. Longino, *The Fate of Knowledge* (Princeton: Princeton University Press, 2002), Ch 6; Liam Kofi Bright & Remco Heesen, "To Be Scientific Is to Be Communist," *Social Epistemology*, vol. 37 (2023).

(17) Marcus Arvan, Liam Kofi Bright & Remco Heesen, "Jury Theorems for Peer Review," *British Journal for the Philosophy of Science*. vol. 76 (2025); Shahar Avin, "Centralized Funding and Epistemic Exploration," *The British Journal for the Philosophy of Science*, vol. 70 (2019); Remco Heesen & Liam Kofi Bright, "Is Peer Review a Good Idea?" *The British Journal for the Philosophy of Science*, vol. 7 (2021); Carole J. Lee, "Commensuration Bias in Peer Review," *Philosophy of Science*, vol. 82 (2015).

في صياغة جواب [مُحدد] عن هذا السؤال: فيقترح فلاسفة الأخلاق نظرياتٍ للحكم الخُلقي، ويضع فلاسفة السياسة نظريات حول العدالة أو السلطة، ويطوّر فلاسفة العلم نماذج للمعرفة والتبرير، ويشرح منظّرو حقل نظرية الفعل إمكان الإرادة الحرّة أو ينقضونه، إلى غير ذلك من الأمثلة. لكن ثمّة استثناءات، فالحقول أقلُّ تجانسًا ممّا تُوحي به هذه الصورة المبسَّطة، ومع ذلك تظل الصورة التي قدمناها وافيةً بالنمط الغالب. بل إن الأعمال التي لا تنضوي ظاهرًا تحت هذا الوصف، مثل الاشتغالات المنهجية أو المفهومية، يُقصد بها، في العادة، إسهامٌ غير مباشر في الإجابة عن تِلكم الأسئلة[18].

يختلف المشهد في حقل أخلاقيات الذكاء الاصطناعي، وعموم حقل أخلاقيات التقنية عن غيره، وذلك لجهة أن موضوع الحقل، بداهةً، هو التقنيات ذاتها. هنا لا ينبثق الدافع إلى البحث من سؤالٍ فلسفي معلَّق، بل من حدثٍ أو تطوّر محدَّد، متمثلًا بظهور (أو توقّع ظهور) تقنية ذكاء اصطناعي جديدة. لذلك لا يُدعى فلاسفة أخلاق الذكاء الاصطناعي إلى حلّ معضلاتٍ فلسفية قائمة، بل إلى التأمّل في تطوُّر تقنيات جديدة والتعليق عليها. فإذا كان "الناتج" المتوقع للبحث الفلسفي هو الإجابة عن سؤالٍ فلسفي مفتوح في التخصصات الفلسفية الأخرى، فإن "الناتج" المتوقع في حقل أخلاقيات الذكاء الاصطناعي هو تأملٌ أو تعليقٌ أخلاقي حول ظهور تقنية جديدة.

تجعل السمةُ المؤسسيةُ الثانية؛ وهي "التأثير الإيجابي" Positive Impact، من العسير على فلاسفة الأخلاق في حقل الذكاء الاصطناعي أن يُعلِّقوا على التقنيات بروح متفائلة. يشير مصطلح "التأثير الإيجابي" إلى قاعدة غير مكتوبة تحكم التنظير الأخلاقي وتُملي نوع المخرجات التي يُفترض بالباحثين إنتاجها. يبدو، على وجه التحديد، أن هناك قاعدة تمنع من ملاحظة أن حدثًا أو تطورًا بعينه قد أفضى، أو سيُفضي، إلى آثار محمودة أخلاقيًا. وللتوضيح دعونا ننتقل من "أخلاقيات الذكاء الاصطناعي" إلى "الفلسفة السياسية"، بادئين بالملاحظة الآتية: إن أحوال المعيشة عالميًا تتحسن بوتيرة لافتة، فالفقر ووفيات الأطفال في أدنى مستوياتها التاريخية، ومعدَّلات القراءة والتعليم النظامي في أعلى مستوياتهما، بل إن مظاهر انعدام المساواة العالمي نفسه يتراجع[19]. هذه التطورات مرغوبة أخلاقيًا. ومع ذلك أزعم أن الحجج الفلسفيّة التي تُصاغ بالطريقة التالية ستبدو مُحرجة إلى حد بعيد:

- أُحَاجُّ في هذه الورقة بأن التراجع العالمي في معدلات الفقر يُفضي إلى زيادةٍ ملحوظة في العدالة.

(18) مع ذلك، أحسب أن أحد المراجعين كان محقًّا في تنبيهه إلى أن بعض صور الاشتغال، كالنقد المفاهيمي، وبناء الأُطر، والتحليل الجينيالوجي، قد لا تنسجم مع هذا النمط. وعلى الرغم من أن كثيرًا من هذا الضرب من الاشتغال لا يزال يرمي، ولو بصورة غير مباشرة، إلى معالجة أنماط الأسئلة المذكورة آنفًا، فإن بعضه لا يفعل ذلك. ولا أعتقد أن هذا يُقوِّض حجَّتي؛ إذ لا تقتضي الأخيرة سوى وجود نزعة عامَّة قوية من هذا القبيل، لا قاعدة مُطَّردة لا استثناء فيها.

(19) Max Roser, "The Short History of Global Living Conditions and Why It Matters That We Know It," *Our World in Data* (2016), accessed on 19/1/2026, at: https://acr.ps/1L9F2D7; Max Roser, "The History of Global Economic Inequality," *Our World in Data* (2017), accessed on 19/1/2026, at: https://acr.ps/1L9F2xo

- أزعم في هذه الورقة أن تزايد التحاق الأفراد بالتعليم النظامي يترجم إلى ارتفاعٍ كبير في العدالة التعليمية Educational Justice.
- أجادل في هذه الورقة بأن الانخفاض العالمي في مستويات اللامساواة يعني أن العالم أصبح أكثر عدلاً.

على افتراض أن الفرضيات التجريبية صحيحة، فإن كلّ واحدة من هذه المزاعم صحيحة ومهمّة. ومع ذلك، فإن مقالاتٍ تشيد حججها على هذا النحو ستبدو، فيما أزعم، شاذة إلى حدّ ما، ومن غير المرجّح أن تتجاوز التحكيم الأكاديمي. فثمة، على ما يبدو، قاعدةٌ عامّةٌ تقضي بأن الباحثين في الأخلاقيات لا ينبغي أن يكتفوا بمجرد تسجيل الآثار المرغوبة أخلاقيًا لحدثٍ أو تطوّرٍ بعينه. وينطبق هذا أيضًا على أخلاقيات الذكاء الاصطناعي؛ إذ ستغدو مقالاتٌ تتقدّم بحججٍ من النوع الآتي خرقاء في أعين المراجعين:

- أُجادل في هذه الورقة بأن "تشات جي بي تي" ChatGPT، من خلال إتاحة الموارد التعليمية بتكلفة زهيدة للجميع، يُعَدّ أمرًا محمودًا من منظور العدالة التعليمية.
- أُجادل في هذه الورقة بأن السيارات ذاتية القيادة مفيدة من زاوية العدالة الاجتماعية؛ لأن الأشخاص الأقلّ دخلاً سيجنون الفائدة الأكبر من خفض تكاليف التنقّل.

يبدو أن ثمة خللاً ما في المقالات الفلسفية التي تشيد حججها على هذا النحو، فبصرف النظر عن صحة هذه المزاعم، لا تُعدّ من الطراز الذي يُفترض بالباحث الأخلاقي أن يتقدّم بها. صحيحٌ أن مثل هذه الادعاءات قد تظهر جزءًا من حجج أوسع، مثل الاعتراض على التحكم في تلك التقنيات أو تأييد دعمها ماليًّا، ويمكن كذلك الدفعُ بها ردًا على حجج مضادة تزعم أن تشات جي بي تي أو القيادة الذاتية غير عادلة. غير أن الاكتفاء بملاحظة أن حدثًا أو تطورًا، كظهور تقنية جديدة، له آثار أخلاقية مستحسنة لا يندرج ضمن ما هو متوقّع من باحثي الأخلاق فعله، فمقال من طراز "ثمّة تطبيق ذكاء اصطناعي جديد، وهو ممتاز" لا يكون مقبولاً، بينما يُعَدّ من المقبول أن يُعلن مقال "ثمّة تطبيق جديد، وهو مثيرٌ للقلق الشديد".

تأمّل الافتتاح النموذجي لمقال صدر حديثًا عن "التَعسُّف الهيرومنطيقي" المرتبط بالذكاء الاصطناعي: "كشفت أدبيات أخلاقيات الذكاء الاصطناعي عن أشكال كثيرة من الأذى الذي تسببه أو تعززه هذه التقنية [...] غير أن شكلاً واحدًا فاته الرصد"[20]. لاحظ مدى سلاسة هذا التأطير. في المقابل، سيكون

(20) Andrew P. Rebera, Lode Lauwaert & Ann–Katrien Oimann, "Hidden Risks: Artificial Intelligence and Hermeneutic Harm," *Minds & Machines*, vol. 35, no. 33 (2025(, p. 2;

في هذا السياق، ثمة مقالة أخرى في الموضوع نفسه وبالتأطير عينه تقول: "بات أمرًا ثابتًا أن الخوارزميات قد تكون أدوات للظلم، غير أن ما يُناقَش على نحو أقلّ بكثير هو أن طرائق نشر الذكاء الاصطناعي الراهنة تجعل اكتشاف الظلم نفسه أمرًا عسيرًا، إنْ لم يكن مستحيلاً. [...] نُبين كيف يمكن للاستخلاص الخوارزمي Algorithmic Profiling أن يُولّد ظلمًا إبستيميا"، ينظر:

Silvia Milano & Carina Prunkl, "Algorithmic Profiling as a Source of Hermeneutical Injustice," *Philosophical Studies*, vol. 182 (2025), p. 186;

وأودّ التنويه إلى أنني لستُ هنا في معرض نقد هؤلاء المؤلّفين.

غريبًا أن نصادف ورقة في أخلاقيات الذكاء الاصطناعي ذات نزعة إثباتية بدلاً من نقدية، تعترف بفوائد التقنية المتعددة وتمضي إلى توثيق فائدة لم يُلتفت إليها من قبل. ومع أن الذكاء الاصطناعي يجعل العالم أكثر عدلاً بطرائق لا حصر لها، لا يُنتظر من الباحثين في حقل أخلاقيات الذكاء الاصطناعي توثيق هذه المنافع، بل يُنتظر منهم التركيز على السبل التي يهدّد بها الذكاء الاصطناعي القيم الأخلاقية، فضلاً عن العناية بكيفيّة درء تلك التهديدات.

إن القول بوجود افتراضٍ مُسبقٍ يعترض طريق المقالات الإيجابية التي تقتصر على الملاحظة يظلّ، في جوهره، طرحًا ظنيًا. غير أن هناك مسوّغًا إضافيًا لاعتقاد وجود مثل هذا الافتراض، يتجاوز مجرّد الانطباع بوجود عوار في تلك المقالات، ذلك أن الافتراض في ذاته معقول. فملاحظة أن حدثًا أو تطوّرًا بعينه محمودٌ أخلاقيًا قد ينطوي على قيمةٍ معرفية بحد ذاته، أو على الأقل تقدير، عندما لا تكون الملاحظات بديهية. لكن هذه الملاحظات، في الغالب، لا تحمل أثرًا عمليًا كبيرًا، فهي لا تمنحنا سوى باعث على الابتهاج. قارن ذلك بمقالات تُشير إلى أن حدثًا أو تطوّرًا ما يشتمل على جوانب إشكالية، مثل هذه المقالات لا تكون مبعثًا للانزعاج وحسب، بل تمنحنا سببًا للتدّخل Intervene، فهي، بصورة مباشرة أو غير مباشرة، دعواتٌ إلى الفعل. واستصحابًا تنشأ لا تماثليةٌ Asymmetry في القوّة الإلزامية بين التقديرات الإيجابية والسلبية للذكاء الاصطناعي. وبما أن توقّع كتابة أوراقٍ ذات صلةٍ إجرائية ليس أمرًا غير معقول، فمن غير المستبعد وجودُ قاعدةٍ تردّ هذه المقالات الإيجابية الخالصة ذات الحمولة التقريرية الضئيلة.

إذًا، نعني بـ"موضوع البحث" أن متخصصي أخلاقيات الذكاء الاصطناعي منشغلون أساسًا بالتعليق على التقنيات الجديدة، لا بحلّ الألغاز الفلسفية القائمة. ونعني بـ"الأثر الإيجابي" أنه يُحظر عليهم الاشتغال بالجوانب الإيجابية له. أمّا العامل المؤسسي الثالث "الحوافز"، فيشير إلى حقيقة أن الأكاديميين مضطرّون إلى النشر للحفاظ على مسارهم المهني، فالتوقّف عن كتابة أوراق حوله ليس خيارًا متاحًا لباحثي هذا الحقل[21].

يُفسّر اقتران هذه العوامل الثلاثة تفسيرًا جيدًا لهذا المدّ الصاعد من الذعر. فباحثو أخلاقيات الذكاء الاصطناعي مضطرّون إلى النشر، والطريقُ المتاح أمامهم هو تضخيمُ الهواجس الأخلاقية المتصلة بالتقنية. وينسحبُ الأمرُ عينُه على تحصيل المنح البحثية الخارجية؛ إذ لا يستطيع الباحث التقدّمَ لمنحة، وهي كثيرًا ما تكون حاسمةً في الترقية الأكاديمية، إلّا إذا أطَّرَ الذكاء الاصطناعي بوصفه مولدًا لمشكلاتٍ أخلاقية تستوجب البحث. ولنقل بوجيز العبارة: الباحثون في هذا المجال ملزمون بالتأطير السلبي من أجل المحافظة على مسارهم الوظيفي، أمّا أولئك الذين يؤمنون بغير ذلك، فمهددون بفقدان وظائفهم سريعًا.

(21) للاطلاع على مناقشات ذات صلة بهياكل الحوافز الإشكالية في العلم، ينظر:

Wesley Buckwalter, "The Replication Crisis and Philosophy," *Philosophy and the Mind Sciences*, vol. 3 (2022); Remco Heesen, "Why the Reward Structure of Science Makes Reproducibility Problems Inevitable," *The Journal of Philosophy*, vol. 115 (2018).

يحمل هذا العرض شيئًا من المبالغة؛ فلا ينحصر اشتغال هؤلاء المتخصّصين في إثارة المخاوف الأخلاقية؛ لذلك ينبغي تقييد هذا الزَعم بوجوه ثلاثة: أوّلًا، في وسع المتخصصين أن يقترحوا حلولاً للمشكلات التي يطرحونها، مثل معالجة فجوات المسؤولية، أو البطالة التي تُحدثها التقنيات، أو ضمور الملكة الأخلاقية. ثانيًا، يُمكن أن يناقش الباحثون في مقالاتهم السجالية Response Pieces[22] وجود هذه المشكلات المزعومة أو مدى خطورتها، فيذهب أحدهم إلى أن الذكاء الاصطناعي لا يفضي فعليًا إلى حدوث فجوات المسؤولية أو البطالة أو ضمور الملكة الأخلاقية، معتبرًا تلك الهواجس من قبيل المبالغات أو أنها زائفة. وأخيرًا، هناك موضوعات في الحقل لا تتمحور أساسًا حول كونه مصدر قلق أخلاقي، مثل الجدل حول الوضعية الأخلاقية والوكالية للنُظم الاصطناعية the Moral and Agential Status of Artificial Systems، أو السؤال عمّا إذا كان في إمكان هذه النُظم أن تكون أصدقاء أو شركاء عاطفيين، فضلاً عن التأمّلات الأوسع عن الظرف الإنساني Condition Humaine في عصر الذكاء الاصطناعي.

مع ذلك، لا تكاد هذه الاستثناءات تخفف كثيرًا من الديناميات الموصوفة آنفًا. فاقتراح حلول لمشكلات مرتبطة بالذكاء الاصطناعي يفترض سلفًا وجود تلك المشكلات، ومن ثم يؤكد السردية السلبية بصورة غير مباشرة. أمّا المقالات السجاليّة، فتقلّ عادةً من حيث الحضور والتأثير، فإذا صحّ أن بعضها يحقق رواجًا لافتًا، فإن ذلك واقع في خانة الاستثناء لا القاعدة[23]. وعلى الرغم من وجود عدد محدود من القضايا في حقل أخلاقيات الذكاء الاصطناعي التي لا تتمحور حول المخاوف أو المعضلات، فمن باب المبالغة القول إن الباحثين المشتغلين في هذا الحقل لا خيار لهم سوى رسم صورة قاتمة للتقنية، إلا أن النقطة الأعم، هي وجود إكراهات مؤسسية قوية تدفع في هذا الاتجاه. يقترح هذا المقال، إذًا، ميلاً عامًا واضحًا نحو النظرة السلبية، مع الإقرار بوجود بعض الاستثناءات.

يبدو أن كلَّ عاملٍ من العوامل الثلاثة السابقة ضروري لتفسير النزعة التشاؤمية. فلو كانت أخلاقيات الذكاء الاصطناعي، من حيث موضوع البحث، أقرب إلى سائر الحقول الفلسفية، لما انحصر عملُ المتخصّصين فيها في التعليق على الأحداث أو التطوّرات[24]. ولو كان من المقبول نشرُ مقالاتٍ

(22) مقالات أكاديمية قصيرة غايتها المباشرة الاشتباك مع دراسة سابقة؛ فهي تأتي على صورة تعليق أو نقد أو توضيح أو تطوير لحجج سبق طرحها، من دون أن تُقدِّم أطروحة جديدة. هذا اللون حاضر في التقليد البحثي الغربي ويكاد يكون مفقودًا في الكتابة الأكاديمية العربية. (المترجم)

(23) وتبلغ فرص نجاحها أقصاها حين تُفضي المشكلة إلى نقاش واسع، غير أن هذا لا يصدق إلا على شريحة ضئيلة من القضايا المتّصلة بالذكاء الاصطناعي التي تتناولها أدبيات الحقل.

(24) وقد ذهب مايكل هيومر على أسس مماثلة إلى أن المشتغلين بالأخلاق عامةً، لا المشتغلين بأخلاقيات الذكاء الاصطناعي وحدهم، يُضخِّمون المشكلات الأخلاقية. وقد يصحُّ هذا القول، غير أن المشكلة تبدو أكثر حدةً داخل أخلاقيات الذكاء الاصطناعي بسبب بنيته المؤسسية الخاصة. إذ يُتوقَّع من المشتغلين في هذا الحقل أن يُعلِّقوا على التطوُّرات التقنية، بينما يُحظر عليهم إلى حدٍّ بعيد قول أشياء إيجابية عنها. وهذا ما يُولِّد ضغطًا للعثور على مثالب في هذه التطوُّرات. وقد يستشعر المشتغلون بالأخلاق في حقول أخرى، أحيانًا، جذبًا مماثلاً نحو تقديم أمور بريئة بوصفها إشكالية، كما أشار هيومر. بيد أن لديهم أيضًا خيار الكتابة حول الأسئلة الكبرى للفلسفة الأخلاقية (هل ينبغي لنا دائمًا تعظيم الخير؟ هل الأخلاق موضوعية؟ ما الفضائل الموجودة؟ هل ثمة مبادئ أخلاقية عامة؟ ... إلخ)، عوضًا عن التعليق على الأحداث أو التطوُّرات الراهنة. وهذا ما يمنحهم فرصًا أوسع لإجراء بحوث أخلاقية دون اضطرارهم إلى تقديم الأشياء (التطوُّرات التقنية والاجتماعية والسياسية ... إلخ) بوصفها مُقلقة أو إشكالية، ينظر:

Michael Huemer, "The Ethicist's Vice," *Fake Nous* (2023), accessed on 19/1/2026, at: https://acr.ps/1L9F2tE

تُشيد بروعة الذكاء الاصطناعي، لاستطاع الباحثون تجنب تغذية السردية القاتمة عبر ضربٍ من المقالات المتفائلة. ولو لم يكن إنتاجُ مخرجات بحثية محدّدة شرطًا مهنيًا، لكان في مقدورهم تقليصُ عدد الكتابات النقدية أو الكفُّ عنها كليّة. غير أن هذه المقدمات غير موجودة كلها. ونتيجةً لذلك، لا يكاد يبقى أمام الباحثين المشتغلين في هذا الحقل سوى الكتابة عن الإشكالات الأخلاقية التي يطرحها الذكاء الاصطناعي. إن اجتماعَ العوامل الثلاثة يفسّر سلبية المجال، ومع رسوخ هذه الديناميات المؤسسية وازدهار الحقل المذكور، فإن موجة الذعر مرشَّحة لا محالة لمزيدٍ من الارتفاع. ومن العسير أن نتصوّر كيف يمكن أن تُصبح أخلاقيات الذكاء الاصطناعي يومًا ما حقلاً ذا نبرةٍ متفائلة تجاه التقنية.

لا أفترض بهذا المقترح التفسيري تحقق هذه السمات الثلاث في صيغ مثالية؛ إذ أعدها أنماطًا عامة وميولاً سائدة، لا قوانين مطلقة لا تعرف استثناء، لكنها في تضافرها وشيوعها كافية لتفسير النزعة التشاؤمية السائدة في الحقل.

من السلبية إلى الأحادية والتحيز

توفّر الديناميات المؤسسية الموصوفة آنفًا تفسيرًا للنبرة التشاؤمية الملحوظة في حقل أخلاقيات الذكاء الاصطناعي. وقد التزمتُ، حتى هذه اللحظة، الحيادَ إزاء ما إذا كان هذا الأمر يشكّل مأزقًا. وسأبين الآن أنه بالفعل أمرٌ إشكالي. فالصورة التي ترسمها الجماعة العلمية لآثار الذكاء الاصطناعي الأخلاقية تبدو لي مفرطة في السلبية، وذلك من وجهين: الأول، لأنها أحاديةُ الجانب One–Sided؛ إذ تميل إلى الجوانب السلبية فحسب، والثاني، لأنها منحازةٌ Biased سلبيًا.

أولاً، بإيجاز، لنتأمّل مشكلة الأحادية السلبية. تُبرز الديناميات المؤسسية، التي ذكرناها، الجوانب السلبية على نحو لافت، في حين نادرًا ما تُناقَش الجوانب الإيجابية. والنتيجة المتأدّية عن ذلك تتمثّل في أن الصورة العامة التي تتشكّل عن الذكاء الاصطناعي تصبح أحاديةً بصورة فادحة. وهي أحادية مضللة؛ إذ تجعل التقنية تبدو، في المحصّلة، أشدَّ إشكاليةً مما هي عليه فعلاً. وليست المشكلة في أن القضايا المشخَّصة متخيلة، بل المشكلة في التركيز المبالغ فيه على الجوانب السلبية بينما يجري تجاهل الجوانب الإيجابية إلى حدٍ بعيد. حتى لو صَدَقَ أن كلّ قضية ذُكرت في أدبيات أخلاقيات هذا الحقل قضية حقيقية، فستظلّ الصورة الناتجة من ذلك خادعة؛ إذ يقتضي عرضٌ دقيقٌ غير مضلّل لأي تقنية موازنةً بين سلبياتها وإيجابياتها معًا. ولا عجب أن يشعر الناس بالذعر إزاء الذكاء الاصطناعي إذا واجهوا في الأساس عيوبه، فأي تقنية ستبدو مقلقة إذا اقتصر النظر على جوانبها المظلمة.

لفهم السبب في أن الأحادية قد تكون إشكالاً، تخيّل مسافرةً عادت من إنكلترا ولم تُفصِح إلاّ عن انطباعاتها السلبية، كأن تشكو من سوء الطقس، وارتفاع أسعار الفنادق في لندن، وانحطاط المدن ما بعد الصناعية، ورداءة شبكة السكك الحديدية، إلى آخره. فلا شيء مما ذكرته باطل من حيث المبدأ؛ فكلّ ملاحظةٍ صحيحةٌ في ذاتها. غير أن حذف كلّ الانطباعات الإيجابية يجعل السرد مُضلّلاً؛

إذ بالرغم من الصحة الوصفية للملاحظات، يتشكّل لدى المستمع انطباعٌ مغلوط عن البلد. إن حقل أخلاقيات الذكاء الاصطناعي، حين يحصر عدسته في المشكلات والمخاطر، يماثل إلى حدّ بعيد رواية تلك المسافرة، ولو بدرجة أقلّ حدّة، فحتى لو افترضنا صحةَ كلِّ هاجسٍ أخلاقي حول التقنية، ستظلّ الصورةُ العامة معيبةً بأحاديتها، بالضبط كما اختلّت رواية المسافرة.

علاوةً على ذلك، هناك ما يدعو إلى افتراض أن الصورة العامة للذكاء الاصطناعي، بفعل الديناميات المؤسسية الموصوفة، ليست أحادية الوجه فحسب، بل مشوَّهة على نحو أعمق؛ فالمسألة لا تقتصر على إهمال الجوانب الإيجابية في كثير من الأحيان، بل يُحتمل أيضًا أن تُضخَّم الجوانب السلبية. وبعبارة أخرى، ينبغي التسليم بأن هذه الديناميات تُنشئ في الحقل انحيازًا تجاه السلبية والتشاؤم. يُمكن فهم التحيُّز على أنه ميلٌ منهجيٌ إلى الانحراف عن الحقيقة في اتجاهٍ بعينه[25]؛ إذ يميل خطاب أخلاقيات الذكاء الاصطناعي إلى الابتعاد عن الدقة عبر المبالغة في حجم المشكلات الأخلاقية التي تثيرها التقنية. وافتراضُ أن الديناميات المؤسسية تولّد مثل هذا الانحياز السلبي يبدو وجيهًا لثلاثة أسباب.

يرتبط السبب الأول بما يسمى "قانون الأداة" Law of the Instrument، الذي يفيد أن من لا يملك في عُدته سوى مطرقة يرى في كل شيء مسمارًا، فيرى المسامير حتى في الأماكن التي لا توجد فيها. ومن الراجحّ أن قانونًا مشابهًا يعمل في أخلاقيات الذكاء الاصطناعي، فالعدة الرئيسة، وإن لم تكن الوحيدة، لدى باحث هذا الحقل هي تشخيص المشكلات الأخلاقية. لذلك يُتوقع أن يرى الباحثون مشكلات حيث لا توجد، أو أن يعدّوها أفدح مما هي عليه. ينشأ هذا الأثر من العاملين الأول والثاني اللذين سبقت مناقشتهما: فالتفاعل بين "موضوع البحث" و"التأثير الإيجابي" يدفع الباحث إلى تبني زاوية نقدية تلقائيًا؛ إذ يستبعد الاستجابات الإيجابية من صندوق أدواته ويجعل النقد الأخلاقي أداته الافتراضية. وحين يتأمل الباحثون تقنيةً ما يميلون إلى النظر إليها من زاوية سؤال: "ما الذي يمكن أن يكون مقلقًا فيها؟". ومن المعقول افتراض أن زاوية النظر النقدية تدفعهم أحيانًا إلى رؤية مشكلات لا وجود لها أو إلى تضخيم الموجود منها بغير مسوّغ. وحتى مع استبعاد العامل الثالث المتصل بـ"الحوافز"، يبقى الاعتقاد بأن ثمة إفراطًا في تشخيص المشكلات أمرًا مسوغًا.

يرتبط السبب الثاني بالأوّل غير أنه يُدخِل عامل "الحوافز" في الحساب. فالنقد الأخلاقي لا يُعَدّ الأداة الرئيسة في عُدّة باحث أخلاقيات الذكاء الاصطناعي وحسب، بل إن هؤلاء يُشجَّعون على استعماله؛ إذ يتعين على الباحثين الانخراط في النقد الأخلاقي إن أرادوا النشر والترقّي في السلّم الأكاديمي. ويبدو المبدأ الآتي وجيهًا: متى وُجد حافزٌ يدفع المرء إلى تبني موقفٍ بعينه، مال إلى الدفاع عنه بوتيرة تفوق ما تُجيزه المعطيات المعرفية. فإذا كانت لديك حوافز قوية للدفاع عن قضية ما، فستفعل ذلك غالبًا حتى عندما لا تُسعف الأدلّة تلك القضية. وبإسقاط هذا على أخلاقيات الذكاء الاصطناعي، يُرجَّح أن الباحث يُشجَّع على الزعم أن التقنية مُقلقة أخلاقيًا، وسيكتب أوراقًا تثير هذه المخاوف، ولو لم تسندها الشواهد بما يكفي.

(25) Heather Douglas & Kevin C. Elliott, "Addressing the Reproducibility Crisis: A Response to Hudson," *Journal for General Philosophy of Science*, vol. 53 (2022), p. 202.

تجدر الإشارة أيضًا إلى أن مختصّي أخلاقيات الذكاء الاصطناعي، وأخلاقيات التقنية عمومًا، يُشاركون كثيرًا في مشاريع كونسورتيوم Consortium[26] التي تضمّ علماء حاسوب ومهندسين، حيث يُعهد إليهم تقديم التوجيه الأخلاقي للمشروع البحثي. ويجد هؤلاء الباحثون أنفسهم، في مثل هذه المشروعات، مضطرّين إلى إثارة هموم وقضايا أخلاقية تتعلّق بالعمل القائم، فليس من الممكن أن يخاطبوا المهندسين بالقول: "ما تقومون به أمرٌ مذهلٌ حقًا، ولا نرى أي مشكلات [أخلاقية] في مشروعكم، فحظًا طيبًا!". مهما كان المشروع رائعًا أو بريئًا، فإن وظيفة باحث الأخلاقيات تفرض عليه أشكلَتها.

أما السبب الثالث الذي يسوّغ الظن بأن الديناميات، التي تقدَّم ذكرها، قد تفضي إلى تضخيم المشكلات الأخلاقية المتعلّقة بالذكاء الاصطناعي، فإنه يرتبط بما تولده من آثار انتقائية محتملة. فالأشخاص المتفائلون بالتقنية قد يُحجمون عن الانخراط في مجال يلزمهم التركيز على الجوانب السلبية، الحقيقية منها أو المتوهّمة، في حين يجد المتشككون في الذكاء الاصطناعي هذا المجال أكثر جاذبية. من شأن هذا الأثر الانتقائي أن يَزيد اختلال المنظور العام في الاتجاه السلبي على نحوٍ مُقلق؛ إذ يفضي إلى تركيبة بحثية غير متوازنة ينتج منها تقييمات منحازة إلى إمكانات التقنية الأخلاقية، كما يصرف الفئة القادرة على موازنة القوتين السابقتين الدافعتين إلى التشاؤم عن الخوض في هذا المجال دون إسهام.

ثمّة، إذًا، ثلاثة أسباب تدعونا إلى الظن أن العوامل المؤسسية تدفع جماعة أخلاقيات الذكاء الاصطناعي إلى تضخيم المخاوف الأخلاقية المتعلّقة بالتقنية إجمالاً. وكما أشير آنفًا، يمكن أن يتخذ هذا التضخيم شكلين مختلفين: فإمّا أن يأتي معظّمًا لمخاوف أخلاقية حقيقة لكنها ليست بالقدر نفسه من الجسامة، وإمّا أن يأخذ صورة تعيين مشكلات أخلاقية لا وجود لها أصلاً. ولأغراض الإيضاح سأغامر، مع ما ينطوي عليه ذلك من جدلٍ محتوم، بطرح مثالٍ واحد لكلٍ من النمطين السالفين.

أما المثال على النمط الأول؛ أي المبالغة في وسم قضايا أخلاقية حقيقية بالخطورة، فيتمثّل في عَتَامة خوارزميات التعلّم العميق the Opacity of Deep Learning Algorithms في السياق الطبي. فكثيرًا ما يُفترض أن معضلة "الصندوق الأسود" Black Box[27] تُعقّد الاستخدام الأخلاقي للذكاء الاصطناعي في مجال الطب؛ إذ تدور جَمهرة مُعتبرة من الأدبيات حول كيفية بلوغ "القابلية للتفسير" Explainability أو التعويض عن غيابها. في حين يذهب عدد قليل من الأدبيات إلى أن القيمة

(26) يشير هذا المصطلح إلى ائتلاف يجمع عدة أشخاص أو مؤسسات تعمل معًا من أجل غاية مشتركة وتقتسم الموارد والخبرات. (المترجم)

(27) يُستخدم هذا المُصطلح لوصف نماذج التعلم الآلي، وخاصةً نماذج التعلم العميق، التي يصعب فهم آليات عملها الداخلية؛ إذ يُمكن اطلاع المستخدمين على المُدخلات والمُخرجات، ولكنهم لا يستطيعون تتبع المنطق الداخلي الذي يُسفر عن النتائج. نجم عن التخوفات التي يثيرها هذا الغموض تطور حقل "الذكاء الاصطناعي القابل للتفسير" Explainable AI – XAI. (المترجم) لاطلاع أوسع ينظر:

Feiyu Xu, et al., "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," in: *Natural Language Processing and Chinese Computing: NLPCC 2019*, Lecture Notes in Computer Science, vol. 11839, Jie Tang, M.–Y. Kan, Dongyan Zhao, S. Li & Jun Zhu (eds.) (Cham: Springer International Publishing, 2019), pp. 563–574.

الأخلاقية لقضية القابلية للتفسير مبالغ فيها، فما يَهمُّ ببساطة هو إذا ما كانت تؤدي وظيفتها أم لا. ومن ثم، قد يكون السؤال الحاسم، هنا، ما إذا كان العلاج أو التشخيص الطبي المدعوم بالذكاء الاصطناعي ناجعًا؛ إذ يتضاءل الشأن المعرفي المتعلق بالكيف أو السبب إذا قورن بالأثر. حتى الأطباء من البشر كثيرًا ما يعجزون عن تقديم تفسيرات مماثلة، ولا يُعَدّ ذلك أمرًا مريبًا[28]. من ثمّ قد يكون هناك مشكلة حقيقية هنا، إلا أن حجم القلق الذي يُثار حولها يبدو غير متناسبٍ مع خطورتها الفعلية[29].

وأمّا النمط الثاني؛ أي الإشكالات المتوهَّمة تمامًا، فأسوق، بتحفُّظ، مثالَ ما يُدعى "فجوات المسؤولية"؛ إذ صدرت العديد من الأدبيات عن الفكرة القائلة إن استخدام أنظمة الذكاء الاصطناعي ذاتية التحكم يخلّف فراغًا مقلقًا في إسناد المسؤولية، إلا أن وجود هذه الفجوات غير محسوم على نحوٍ لافت[30]، ونادرًا ما نجد محاولاتٍ منهجية لإثباتها، فغالبية الكتابات تفترض مسبقًا وجودها وتنصرف إلى البحث عن طرائق لردمها، [علمًا] أن وجودها يبدو موضعًا للريبة ابتداءً. خذ مثلاً الحالة النموذجية لمنظومات الأسلحة ذاتية التحكُّم: يصعب التسليم بأن استقلالية طائرةٍ مسيرة تُعفي قائدها من اللوم إذا أصابت مدنيين بدلاً من مقاتلين، فمجرد نشر تقنيةٍ ذاتية التحكُّم لا يجعل المساءلة أقلّ إمكانية مما هي عليه عند اعتماد تقنيات تقليدية لا يمكن التنبؤ بسلوكها وتنطوي هي الأخرى على ضرب من المخاطر. وهذا، على ما يبدو، ما تشير إليه أيضًا حدوس غير المختصين حيال هذه الحالات[31].

(28) للاطلاع على تنويعات لهذا المنظور المتساهل، ينظر:

Boris Babic et al., "Beware Explanations from AI in Health Care," *Science*, vol. 373 (2021); Alex John London, "Artificial Intelligence and Black–box Medical Decisions: Accuracy versus Explainability," *Hastings Center Report*, vol. 49 (2019); John Zerilli et al., "Transparency in Algorithmic and Human Decision–Making: Is There a Double Standard?" *Philosophy & Technology*, vol. 32 (2019).

(29) أركّز هنا تحديدًا على السياق الطبي؛ إذ قد تختلف المخاطر والرهانات في سياقات أخرى. وقد قدَّم بعض الباحثين أسبابًا عامّة تدعو للاهتمام بعَتَامة الخوارزميات، ينظر على سبيل المثال:

Jocelyn Maclure, "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind," *Minds and Machines*. vol. 31 (2021); Sophie Dishaw, "The Right to a Justification," *Political Philosophy*, vol. 2, no. 4 (2025); Kate Vredenburgh, "The Right to Explanation," *The Journal of Political Philosophy*, vol. 30 (2022);

لا يسعني معالجة تلك الهموم على نحو وافٍ في هذا السياق. ومع ذلك، حتى لو كانت هناك أسباب عامة تدعو إلى القلق، فقد تتباين درجةُ حدّتِها. للنظر في عدد من المقالات التي انتقدت هذا الموقف:

Sophie Dishaw, "The Right to a Justification," *Political Philosophy*, vol. 2, no. 4 (2025); John Zerilli et al., "Transparency in Algorithmic and Human Decision–Making: Is There a Double Standard?" *Philosophy & Technology*, vol. 32 (2019).

(30) ينظر:

Johannes Himmelreich, "Responsibility for Killer Robots," *Ethical Theory and Moral Practice*, vol. 22 (2019); Sebastian Köhler et al., "Technologically Blurred Accountability," in: C. Ulbert et al. (eds.), *Moral Agency and the Politics of Responsibility* (London: Routledge, 2018); Peter Königs, "Artificial Intelligence and Responsibility Gaps: What is the Problem?" *Ethics and Information Technology*, vol. 24 (2022).

(31) Philip Robbins, "Of Machines and Men: Attributions of Moral Responsibility in AI–Assisted Warfare," *Ethics and Information Technology*, vol. 27 (2025).

وبما أنني، انسجامًا مع أطروحة هذه الورقة، أخالف الرأيَ الغالب في هذه المسائل، فلا أتوقّع أن تُعَدّ هذه الملاحظات الموجزة مقنعةً أو كافية[32]. ومع ذلك، فهي تكسو أطروحة البحث المجرّدة شيئًا من اللحم، وقد تؤدّي دورًا إيضاحيًا. أكتفي بهذا القدر، لأعود في بقية الدراسة إلى نهجي غير المباشر القائم على تحليل العوامل المؤسسية لا الاعتبارات الأولية.

خلاصةُ الحجّة حتى هذه اللحظة: بفعل الديناميات المؤسسية داخل حقل أخلاقيات الذكاء الاصطناعي يُدفَع الباحثون، إلى حدٍ ما، إلى رسم صورة سلبية حول التقنية. وهذا لا يفضي إلى عرضٍ أحادي لآثارها الأخلاقية فحسب (إذ نادرًا ما يُلتفت إلى الجوانب الإيجابية)، بل يُرجَّح أيضًا أن يقود إلى تشويه الجوانب السلبية ذاتها (وذلك حين تُضخَّم وتُصوَّر أكبر ممّا هي عليه فعلاً).

الحجة في سياق: أزمة التكرار

من المفيد عقد مقارنة بين أزمة السلبية في حقل أخلاقيات الذكاء الاصطناعي وأزمة التكرار Replication Crisis المستمرّة في عدد من الحقول التجريبية؛ إذ إن أوجه الشبه والاختلاف بين هاتين الظاهرتين تمنح فهمًا مقارنًا وافيًا للديناميات الإشكالية داخل هذا الحقل، فضلاً عن كشفها لما أطرحه هنا عن طبيعة القلق وحدَّته.

تشير أزمة التكرار إلى أن عددًا كبيرًا من النتائج المنشورة في علم النفس والعلوم الطبية وغيرها من الحقول التجريبية لا يُمكن إعادة تكرارها. وإن ما يفسّر، أفضل تفسير، ازدحام الدوريات العلمية بالدراسات المشفوعة بنتائج يتعذّر إعادة إنتاجها كائن في طريقة التنظيم الاجتماعي لهذه الفروع المعرفيّة. يقدّم روبرت هادسن التحليل الآتي لأسباب الأزمة: "يتنافس العلماء على الوظائف والتمويل والمكانة الاجتماعية، وتعتمد قدرتهم في نَيْل هذه الأمور على قدرتهم على نشر أعمالهم. علاوة على ذلك، تتسابق الدوريات العلمية على نشر النتائج الأكثر إثارةً وسدادًا، ولأجل ذلك تتجنب، في العادة، نشر تكرارات للنتائج البارزة التي سبق نشرُها أو تأكيد الفرضيات الصفرية Null Hypotheses[33]. والحصيلة أن البحث العلمي المنشور يُظهر ما يسمّى "انحياز النشر"، بحيث يُعلن العلماء عن نتائج قابلة للنشر؛ أي نتائج ذات أثر دلالي إحصائي كبير، لا بالضرورة عن نتائج صادقة أو مبرَّرة على نحو كافٍ بالضرورة[34].

إحدى علل أزمة التكرار ترجع إلى تفاعل عاملَين: الأول، تميل الدوريات المرموقة إلى نشر الدراسات التي تُفضي إلى نتائج دالّة ذات تأثير كبير، بدلاً من نشر دراسات التكرار [والتحقق] أو النتائج الصفرية

(32) سلامةُ حجّتي الجوهرية لا تعتمد على مدى وجاهة الأمثلة المذكورة. أذكر ذلك لا لتحصين الحجّة من النقد؛ إذ قد تكون عرضة له بوجوه أخرى، بل للتشديد على طابعها غير المباشر.

(33) فرضية إحصائية تفترض عدم وجود أثر أو علاقة بين المتغيرات المدروسة، وتُستخدم نقطة انطلاق لاختبار الدلالة الإحصائية. ويُعد "اختبار دلالة الفرضية الصفرية" Null Hypothesis Significance Testing الإطار الإحصائي السائد في العلوم الاجتماعية والسلوكية والطبية الحيوية. (المترجم)

(34) Robert Hudson, "Should We Strive to Make Science Bias–free? A Philosophical Assessment of the Reproducibility Crisis," *Journal for General Philosophy of Science*, vol. 52 (2021), p. 396.

Null Results[35]. الثاني، الباحثون، الذين يتوقف مسارهم المهني على سجل نشرهم، يملكون حوافز قوية لنشر أعمالهم في تلك الدوريات. هذا التضافر يُغري الباحثين "باستحداث" نتائج عظيمة، فيلجأ بعضهم، أحيانًا، إلى ممارسات بحثية جدلية، مثل التلاعب بالبيانات P–Hacking[36] أو حتى التزوير الصريح. فتتكدّس حصيلة وافرة من النتائج "المهمة" التي لا يمكن تكرارها لاحقًا؛ لأن الآثار [المُفترضة] لتلك النتائج لا وجود لها في الواقع.

تتشابه أزمة التكرار، في بعض الوجوه، مع أزمة السلبية في أخلاقيات الذكاء الاصطناعي. ففي الحالتين تؤدّي حوافز التقدّم المهني دورًا حاسمًا. كذلك يظهر، في كلتيهما، ميلٌ إلى تفضيل نمط بعينه من المخرجات البحثية على سواه؛ إذ تُؤثر الدوريات العلمية (وكذلك لجان التوظيف) في الحقول التي تعاني "أزمة التكرار" الدراسات التي تُعلنُ السَبقَ إلى نَتائجَ ذات آثار جسيمة Significant Results، وذلك على حساب دراسات التحقق أو الدراسات التي لا تقدم نتائج جديدة. أمّا في أخلاقيات الذكاء الاصطناعي، فيبدو أن هناك توقّعًا بألّا تكتفي المقالات بوصف الآثار الإيجابية للتقنية (عامل "الأثر الإيجابي" الذي سبق ذكره)، ثم إن مقالات السجال البحتة القادرة على تفنيد المخاوف الأخلاقية التي يُثيرها الآخرون تُعَدّ هي الأخرى أقل أهمية.

غير أن هنالك أيضًا فروقًا، أحدها أن أزمة التكرار تنبع، في جزءٍ منها، من سلوكٍ غير أخلاقي يَصدر عن الباحثين؛ يشمل ممارساتٍ بحثيةً ملتوية، وحتى تزييفًا للبيانات. ولا أرى أن هذا ينطبق على أزمة السلبية في أخلاقيات الذكاء الاصطناعي؛ إذ إن زعمي أن الدينامية المؤسسية في هذا الحقل تولّد انحيازًا، لا أنها تدفع أعضاءه إلى التصرّف على نحوٍ غير نزيه. وثمة فرقٌ آخر هو أن لدينا، في سياق أزمة التكرار، سببَين للتشكيك في موثوقية المنظومة المعرفية: أوّلًا، يوجد مسوّغ قوي للارتياب فيها بالنظر إلى هياكل الحوافز في بعض الحقول التجريبية، وهذا مسلكٌ غير مباشر، أو قبلي، لإثارة الشكوك حول الموثوقية. ثانيًا، يتوافر دليلٌ مباشرٌ على اختلال المنظومة؛ إذ تَبين أن كثيرًا من النتائج المنشورة يُخفق فعليًا عند التحقق والتكرار. أمّا في أخلاقيات الذكاء الاصطناعي، فقد اكتفيت بتقديم أسبابٍ غير مباشرة للتشكيك في موثوقيتها؛ إذ بدلًا من أن أبيّن، على نحو مباشر، سبب المبالغات المتصلة بالمخاوف الأخلاقية المتعلقة بالذكاء الاصطناعي، فإني أعمد إلى ضرب من التكهّن مؤدّاه أن هذه المخاوف حاضرة، وذلك استنادًا، حصرًا، إلى تحليل الكيفية التي ينظّم بها هذا الحقل مؤسّسيًّا.

(35) تُشير النتائج الصفرية إلى نتائج البحث التي لا تُثبت وجود أثر ذي دلالة إحصائية؛ أي التي لا ترفض الفرضية الصفرية. وتُعرف مشكلة عدم نشر هذه النتائج بـ"مشكلة الدُّرج" File Drawer Problem، وهو مصطلح صاغه روبرت روزنتال عام 1979 للإشارة إلى أن الدراسات ذات النتائج غير الدالة إحصائيًا تبقى حبيسة أدراج الباحثين دون نشر. [المترجم]، ينظر:

Robert Rosenthal, "The File Drawer Problem and Tolerance for Null Results," *Psychological Bulletin*, vol. 86, no. 3 (1979), pp. 638–641.

(36) ويقصد بها لجوء الباحث إلى حِيَلٍ إحصائية، مثل استبعاد بعض البيانات، أو اختبار متغيّرات عديدة، أو إيقاف جمع العيّنة مبكرًا، بُغية انتزاع قيمة إحصائية دالّة من البيانات على نحو تعسفي. (المترجم)

ليس واضحًا إذا ما كان في الإمكان إيراد دليلٍ مباشر، مماثلٍ لفشل أزمة التكرار في الحقول التجريبية، داخل أخلاقيات الذكاء الاصطناعي. فصحيحٌ أن الدعاوى الأخلاقية (مثل التمسّك بوجود مشكلةٍ مرتبطة بالتقنية) قابلةٌ للطعن، غير أن هذه الطعون محكومٌ عليها بأن تظلّ موضع خلاف وجدل، فلا نجد في حقل الأخلاقيات ما يكافئ سلسلة الفشل في تكرار البحوث التي تشهدها العلوم التجريبية.

إن الطابع غير المباشر للنهج الذي اتبعته يطرح قيودًا على ما أحاول تقديمه، وذلك على الرغم من شيوعه في فلسفة العلم. ففي وضعٍ أمثل، كنا سنمتلك دلائل أكثر دقة ومباشرة على انحياز سلبي ممنهج داخل الحقل، غير أن تجاهل القرائن غير المباشرة التي تشي بهذا الانحياز لن يكون تصرّفًا حكيمًا أيضًا.

على هامشٍ متصل، انتهج ويسلي بكوالتر مقاربةً غير مباشرة مشابهة استلهم فيها أزمة التكرار. فهو يرى أن تحليل العوامل البنيوية التي تقف وراء أزمة التكرار في العلوم الحيوية والعلوم الاجتماعية يمكن أن يقدّم، بطريقٍ غير مباشر، رؤيةً عمّا إذا كانت الفلسفة هي الأخرى معرّضة لأزمة مماثلة؛ ذلك أن الفلسفة الأكاديمية، ولا سيما حين تعتمد منهج "دراسة الحالة"، فإنها تغيِّب كثيرًا من العوامل البنيوية المُسببة لأزمة التكرار في العلوم الحيوية والاجتماعية. ولذا يخامر بكوالتر الاعتقاد أن الفلسفة قد تواجه مشكلات شبيهة[37]. أمّا مشروعي هنا فأضيقُ نطاقًا من مشروعه؛ إذ أركّز على مجموعةٍ من العوامل الخاصة بأخلاقيات الذكاء الاصطناعي تؤثّر في موثوقية هذا الحقل تحديدًا. ومع ذلك، يجمع مقاربتينا افتراضٌ أساس فحواه أن التأمّل في البِنية المؤسسية للفلسفة، بوصفها تخصّصًا جامعيًا، قادرٌ على أن يكشف لنا عن صدقيّتها.

الاعتراض الأول: انحياز السلبية هو "ماندفيلي"

بعد عرض الحجة وتوضيح مفاصلها، أنتقل إلى مناقشة اعتراضين يُوجَّهان إلى الفكرة القائلة بوجود انحياز سلبي مُقلق في أخلاقيات الذكاء الاصطناعي.

قد يُقال إنني تعجّلت في الحكم على الحقل استنادًا إلى وجود هذا الانحياز، فمن المهم أن نميز بين اللاعقلانية على مستوى الفرد واللاعقلانية على مستوى المنظومة؛ أي مستوى الجماعة العلمية ذاتها. هذا التمييز ظلّ، حتى الآن، مُغفَلًا. فعند تناول مسألة الانحياز في سياق فلسفة العلم، يتركّز اهتمامنا أولًا وأساسًا في الرشد المعرفي للنظام كله: هل تنتج الجماعة العلمية، بوصفها جماعة منظَّمة من الباحثين، معرفة يُعتدّ بها؟ أما عقلانية الباحثين الأفراد، المكوِّنين لهذه الجماعة، فلا تمتلك أهمية إلا بصورة غير مباشرة؛ أي بقدر ما تؤثّر في عقلانية المنظومة برمّتها. ويصدق الأمر نفسه على أخلاقيات الذكاء الاصطناعي؛ إذ ما يعنينا، في المقام الأخير، هو صوابيّة هذا الحقل بوصفه نظامًا

(37) Wesley Buckwalter, "The Replication Crisis and Philosophy," *Philosophy and the Mind Sciences*, vol. 3 (2022).

إبستيميًا، لا صوابيَّة أفراده على نحوٍ مستقل[38]. وبناءً عليه، فإن الهمّ الرئيس في هذه الورقة يتمحور حول موثوقية مخرجات أخلاقيات الذكاء الاصطناعي بوصفها منظومةً معرفية.

يجمع المستويَين الفردي والمنظومي ارتباطٌ وثيق، فمن المنطقي أن يؤدي انحيازٌ، أو شكلٌ آخر من اللاعقلانية، على مستوى الفرد إلى خللٍ على مستوى المنظومة، لكن هذا ليس أمرًا ضروريًا. فاستنادًا إلى الأطروحة البارزة "حكاية النحل الرمزية" The Fable of the Bees لبرنارد ماندفيل Bernard Mandeville (1677–1733)، والتي تقضي بأن الرذائل الخاصّة يمكن أن تُفضي إلى آثار اجتماعية حميدة، أشار فلاسفةُ العلم إلى أن الرذائل الإدراكية Cognitive Vices الفردية قد تُخلّف آثارًا معرفية نافعة على مستوى الجماعة. وبالنظر إلى موثوقية مخرجات نظام إبستيمي، قد يكون من المُستحب أن يتصف الباحثون المُنتسبون إلى نظام إبستيمي ما ببعض النقائص الإدراكية. ومن بين هذه النقائص الماندفيلية المحتملة انحيازاتٌ تُعَدُّ لا عقلانيةً إبستيميًا على مستوى الفرد، لكنها قد تغدو مفيدةً على مستوى المنظومة[39]. وبما أنني أفترض أن انحياز السلبية، محلّ النقاش، يُضعف العقلانية على المستوى المنظومي، فقد يُعترَض بأننا بصدد حالة "ذكاء ماندفيلي" Mandevillian Intelligence، وعندئذٍ يصير هذا الانحياز غير ضارّ، على خلاف ما ذهبتُ إليه.

غير أنني أرى أن ثمة سببًا وجيهًا للافتراض بأن انحياز السلبية ليس حالةً من "الذكاء الماندفيلي"، بل لا يوجد في الواقع مسوّغٌ وجيهٌ لافتراض مقلوب الأمر. فالأصل أن نفترض قاعدةً مسبقةً ضدّ هذه الأطروحة؛ أي إن اللاعقلانية على المستوى الفردي تُترجَم، بوجهٍ عام، إلى لا عقلانية على مستوى المنظومة، ما لم يقدَّم دليلٌ إيجابيٌ يُثبت أثرًا ماندفيليًا معاكسًا. ولا يلوح، فيما يتعلّق بانحياز السلبية موضع النقاش، أي سببٍ إيجابي يدعونا إلى تبني هذا الافتراض.

أكثرُ الأسباب وجاهةً للاعتقاد بأن انحياز السلبية قد يكون ماندفيليًا هو أن انحيازًا آخر، هو "انحياز التأييد" Confirmation Bias، طُرح بوصفه حالةً ماندفيلية. ويُقصد بانحياز التأييد ميلُ المرء إلى

(38) Finnur Dellsén, "The Epistemic Impact of Theorizing: Generation Bias Implies Evaluation Bias," *Philosophical Studies*, vol. 177 (2020), p. 3665; David L. Hull, *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science* (Chicago: University of Chicago Press, 1988); H. E. Longino, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (1990), Chap 4; Miriam Solomon, "Scientific Rationality and Human Reasoning," *Philosophy of Science*, vol. 59 (1992), p. 452; Kevin J. S. Zollman, "The Epistemic Benefit of Transient Diversity," *Erkenntnis*, vol. 72 (2010), p. 33;

أما بشأن استقلال هذين المستويين فيمكن مراجعة:

Conor Mayo–Wilson et al., "The Independence Thesis: When Individual and Social Epistemology Diverge," *Philosophy of Science*, vol. 78 (2011).

(39) Joshua May, "Bias in Science: Natural and Social," *Synthese*, vol. 199 (2021), pp. 33–53; Uwe Peters, "An Argument for Egalitarian Confirmation Bias and Against Political Diversity in Academia," *Synthese*, vol. 198 (2021); Uwe Peters, "Illegitimate Values, Confirmation Bias, and Mandevillian Cognition in Science," *The British Journal for the Philosophy of Science*, vol. 72 (2021); Paul R. Smart, "Mandevillian Intelligence," *Synthese*, vol. 195 (2018).

البحث عن البينات أو تفسيرها بما يوافق معتقداته السابقة[40]. هذا الانحياز، على الرغم من لا عقلانيته على مستوى الفرد، عُدَّ ضربًا من ضروب "الذكاء الماندفيلي"؛ إذ يحفِّز الباحثين المتأثرين به إلى تفحُّص طيفٍ واسع من الفرضيات على نحو أكثر صرامة. ويوضح أوي بيترز "أن انحياز التأييد على مستوى الفرد، يدفع كلَّ عالم إلى استثمار جهدٍ كبير في أمرين: جمع بيانات تدعم أطروحته، والردّ على الأدلّة المضادّة والاعتراضات؛ ما يُفضي إلى استكشافٍ دقيقٍ للأطروحة وتنميتها، لا رفضها السريع"[41]. وبفضل التفاعل النقدي المتبادل الذي يُلغي لا عقلانيات الأفراد، قد يكون الأثر النهائي لهذا السلوك، على مستوى المنظومة، إيجابيًا[42].

سواء نجح هذا المسوِّغ في تبرير اعتبار انحياز التأييد ماندفيليًا أم لا، فإنه لا ينطبق على انحياز السلبية الذي أنا بصدده هنا. فالفرق الجوهري أن انحياز التأييد يدفع الباحثين إلى تأييد قناعاتهم المُسبقة، وهي قناعات تتباين من باحث إلى آخر، فيتوجَّه كلّ فرد نحو مسارٍ مختلفٍ عن الآخر؛ ما يُعزّز "تقسيم العمل المعرفي" Cognitive Division of Labor، وقد يفضي هذا إلى آثارٍ نافعة على مستوى المنظومة. أمّا انحياز السلبية، فيدفع الباحثين جميعًا إلى الاتجاه نفسه؛ أي نحو التشاؤم والسلبية. وبسبب هذه الطبيعة الأحادية الاتجاه للانحياز، يصبح من غير الجلي أي آلية يمكن أن تحوّل اللاعقلانية الفردية إلى عقلانيةٍ جماعية. ويلاحظ فينور ديلسين أنه قد تكون الانحيازات الفردية غير ضارّة حين تجذب الباحثين في اتجاهات متعاكسة؛ إذ يمكنها أن تلغي بعضها بعضًا، فلا يبقى انحيازٌ على مستوى المنظومة. أمّا الإشكال، فيظهر حين تنحاز مجموعة فرعية كبيرة بالقدر نفسه وفي اتجاهٍ واحد"، ففي هذه الحال ينتقل الانحياز إلى مستوى النظام المعرفي[43]. وهذا، مع الأسف، هو ما يحدث فعلًا في حقل أخلاقيات الذكاء الاصطناعي.

وبناءً عليه، تبقى قرينة الرفض المبدئي لفكرة "الذكاء الماندفيلي" قائمة بلا نقض. فخلافًا لانحياز التأييد، يبدو أن انحياز السلبية على مستوى الفرد ينتقل، وبصورة مباشرة تقريبًا، إلى انحياز مماثل على مستوى المنظومة. فإذا صحّ ما اقترحته من وجود ميل سلبي إشكالي لدى كثير من الباحثين على نحو فردي، وجب أن نفترض أن هذا الانحياز يخلّ بموثوقية حقل أخلاقيات الذكاء الاصطناعي برمّته.

---

(40) اعتمدنا في ترجمة "انحياز التأييد" بدلًا من "انحياز التأكيد" اختيار عادل مصطفى، ينظر: عادل مصطفى، **المغالطات المنطقية: فصول في المنطق غير الصوري** (القاهرة: المجلس الأعلى للثقافة، 2007)، ص 179–185. (المترجم)

(41) Uwe Peters, "Illegitimate Values, Confirmation Bias, and Mandevillian Cognition in Science," *The British Journal for the Philosophy of Science*, vol. 72 (2021), p. 1070.

(42) Ibid.; Paul R. Smart, "Mandevillian Intelligence," *Synthese*, vol. 195 (2018), pp. 4188–4191.

(43) Finnur Dellsén, "The Epistemic Impact of Theorizing: Generation Bias Implies Evaluation Bias," *Philosophical Studies*, vol. 177 (2020), pp. 3664–3665;

وبالمثل تتطلب مقاربة مريم سولومان التجريبية الاجتماعية أن "تُوزَّع المتّجهات اللاّأمبيريقية توزيعًا متكافئًا". ينظر:

Miriam Solomon, "Scientific Rationality and Human Reasoning," *Philosophy of Science*, vol. 59 (1992). p. 77;

ينظر أيضًا:

David L. Hull, *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science* (Chicago: University of Chicago Press, 1988), p. 22.

لا تنهض حجّتي ضدّ اعتبار انحياز السلبية "ذكاءً ماندفيليًا" على افتراض أن كلّ باحثٍ في أخلاقيات الذكاء الاصطناعي يرزح تحت هذا الانحياز، فذلك غير صحيحٍ على وجهٍ بيِّن. وحجّتي لا تفترض هذا أكثر مما يفترض المدافعون عن ماندفيلية انحياز التأييد أن جميع أفراد الجماعة متأثرون بذلك الانحياز. المقصود، بالأحرى، أن الانحياز متى وُجد دفع المتأثّرين به في الاتّجاه نفسه، ومن ثمّ ينبغي افتراض أنه ينعكس سلبيًا على العقلانية في مستوى المنظومة.

ختامًا، يجدر التنبيه إلى أن السبب الثالث من الأسباب الثلاثة التي سبقت مناقشتها في افتراض انحياز السلبية يتصل اتصالاً مباشرًا بتركيبة جماعة أخلاقيات الذكاء الاصطناعي. وبناءً عليه، وعلى خلاف مَصدَري الانحياز الآخرين، قد لا يكون من الأنسب فهمه بوصفه انحيازًا فرديًا يُترجَم لاحقًا إلى مستوى المنظومة؛ ذلك أن منشأه يبدو، من الأصل، كامنًا في مستوى المنظومة نفسها.

الاعتراض الثاني: انحياز السلبية على مستوى النظام مرغوبٌ فيه

يَعتبر الاعتراض الأول أن الانحياز غير مؤذٍ لأنه يَظهر على مستوى الأفراد فقط ولا ينتقل إلى مستوى المنظومة. أمّا الاعتراض الذي نحن في صدده، فيُقِرّ بأن الانحياز الفردي يتسرّب فعلاً إلى مستوى المنظومة، إلّا أنه يَعُدُّ ذلك الانحياز الجمعي مرغوبًا. ينبغي التنبيه إلى أن هذا الاعتراض لا يقول فحسب إنه من الجيد أن يدفع التنظيمُ المؤسّسي باحثي أخلاقيات الذكاء الاصطناعي إلى إخضاع تقنياته لفحصٍ أخلاقي صارم، فهذا هدفٌ محمود ولا تعارضَ فيه مع أطروحتي. بل إن الاعتراض يُفصح عن أن جَدوى التنظيم تكمن في أنه يُوَلّد تحيزًا سلبيًا في أثناء هذا الفحص؛ أي إننا قد نرغب، عمدًا، في أن تُبالغ منظومةُ أخلاقيات الذكاء الاصطناعي في نقدها.

قد ينبع هذا التصوّر من ملاحظةٍ مفادها أن فئاتٍ أخرى في المجتمع، وعلى رأسها المنشغلون بقطاع التقنية، يبالغون في التفاؤل ويُبدون قدرًا من التراخي حيال الإشكالات الأخلاقية المُحتملة للذكاء الاصطناعي. ومن ثمّ يبدو أننا نحتاج إلى فئةٍ من باحثي الأخلاقيات لتكبح هذا التفاؤل الساذج. وبناءً عليه، قد يذهب الاعتراض إلى القول إنه من المُستحسن تنظيم حقل أخلاقيات الذكاء الاصطناعي على نحوٍ يفضي إلى "فرط إنتاج" للنزعة السلبية؛ إذ إن التشاؤم المفرط داخل الحقل، والتفاؤل المفرط في دوائر أخرى، يُلغي أحدهما الآخر، فتتكوّن صورةٌ إجمالية أكثر اعتدالاً للتقنية.

وفي وسعنا فهم هذه الفكرة مرةً أخرى عبر منظور "الذكاء الماندفيلي"، مع اختلافٍ جوهري؛ هو أن جماعة أخلاقيات الذكاء الاصطناعي تُعامل هنا بوصفها جزءًا من منظومةٍ أوسع، المجتمع بأسره مثلاً، تَضم جماعات أخرى كالمشتغلين في قطاع التقنية. فالانحياز السلبي المفترض لا يُعد ماندفيليًا داخل الجماعة نفسها، بل يصبح ماندفيليًا على المستوى الأعلى؛ أي مستوى المجتمع كله. ولكي تعمل هذه المنظومة الكبرى كما ينبغي، قد يلزم أن تتحيز مكوناتها الفرعية، مثل جماعتي الأخلاقيين والتقنيين، بصورة منهجية في اتجاهين متقابلين.

تنطوي هذه الفكرة على قدرٍ من الوجاهة؛ إذ يُحتمل أن القطاع التقني، ولأسباب بنيويةٍ مشابهة، يبالغ في التفاؤل عند تصويره تقنيات الذكاء الاصطناعي ويميل إلى كَنس الهواجس الأخلاقية المشروعة تحت البساط[44]. ومع ذلك، لا أرى أن من الحكمة تَنظيم حقل أخلاقيات الذكاء الاصطناعي، عن سابق تصور وتصميم، على نحوٍ يُتوقع منه إنتاج سرديةٍ مفرطة السلبية حول تبعاته الأخلاقية.

بادئ ذي بدء، تبدو فكرة تصميم جماعةٍ فلسفيةٍ بحيث تكون مُتحيّزة، عن سابق تصور وتصميم، أمرًا مُريبًا في حدّ ذاته؛ إذ لو رَفع معهدٌ لأخلاقيات الذكاء الاصطناعي شعارًا يقول: "نُبالغ في تصوير إشكاليات الذكاء الاصطناعي بأكثر مما تقتضيه حاله"، لَاستوقفنا ذلك لا محالة، وإذا كان من الصحيح أن هذه الملاحظة لا تُفند الاستدلال السابق على نحوٍ حاسم، فإنها تكفي، على أقل تقدير، لحثّنا على التوقف والتأمل.

ثمة سببٌ آخر هو أن يوجد قيمة جوهرية Intrinsic في الإجابات الصائبة عن الأسئلة الفلسفية. فإن تحقّق انحيازٌ منهجي على مستوى المنظومة، فلن نجد أنفسنا إزاء وضعية مُستغربة فحسب، بل إننا سنفقد قسطًا من تلك القيمة الجوهرية أيضًا. على الرغم من أن المنظومة القائمة لا تفتقر إلى الموثوقية، فإنها ليست، في الوقت نفسه، مُهيأةً على الوجه الأمثل لإنتاج تَبصُّرات فلسفية صادقة وتجنب الأباطيل. ومهما يكن ما يُكسبه هذا الانحياز من قيمة أداتية، فإنه يشتمل من دون شك على خسارة قيمة جوهرية[45].

غير أن أبرز ما يُضعف الاعتراض الثاني أن التصوُّر الذي يفترضه عن الديناميات الاجتماعية ناقصٌ على نحوٍ مُضلِّل، فهو يفترض أن المجتمع، في غياب حقل أخلاقيات الذكاء الاصطناعي، سيُبالغ في التفاؤل أو التساهل حيال الذكاء الاصطناعي؛ ما يستلزم تصديًّا من باحثي الأخلاقيات. غير أن المُشتغلين في قطاع التقنية ليسوا الفاعل الاجتماعي الوحيد الجدير بالاعتبار، فصحيح أن هذا القطاع يعرض الذكاء الاصطناعي عرضًا مُغرقا في الإيجابية، لكن هناك قوى أخرى تميل، في المقابل، إلى صبغ التقنية بصبغةٍ تشاؤمية مفرطة أو إلى كبح تطوّرها واستثمارها النافع. ومن بين هؤلاء الفواعل مشرّعو سياسات وبيروقراطيون يتحرّقون لاستخدام السلطة التشريعية، فضلًا عن وجود نقاباتٍ مهنية يهمّها إعاقة التقنيات التي تقلل الاعتماد على العمالة البشرية، فضلًا عن قصورٍ عام في إدراك عموم الناس للأهمية الاقتصادية لهذه التقنيات[46]، إضافةً إلى التحيُّز للوضع القائم Status Quo Bias وانحياز "بُغض الخسارة" Loss Aversion، اللذين يجعلَان الناس أكثر خشية من الخسائر المحتملة من تبنّي التقنيات الجديدة المحفوفة بالمخاطر مقارنة بما ينبني على عدم تبنيها من تكلفة وفرص

(44) Jan–Christoph Heilinger, "The Ethics of AI Ethics. A Constructive Critique," *Philosophy & Technology*, vol. 35 (2022), pp. 11–12.

(45) لرأي مماثل ينظر مايكل هُيمر. ففكرة أن الحقيقة ذات قيمة جوهرية فكرةٌ مثيرة للجدل، فالبراغماتيون سيعترضون على هذه الفكرة على سبيل المثال: Michael Huemer, "The Ethicist's Vic," *Fake Nous,* 2023, at: https://acr.ps/1L9F2tE

(46) Bryan Caplan, *The Myth of the Rational Voter: Why Democracies Choose Bad Policies* (Princeton: Princeton University Press, 2007), pp. 40–43.

ضائعة[47]. ومن ثم، قد تتضافر هذه القوى مع قوى اجتماعية ونفسية أخرى لتُسهم في موقفٍ شديد النقد من الذكاء الاصطناعي.

يظلّ من غير الواضح إذا ما كان تعامل المجتمع مع الذكاء الاصطناعي، قبل إسهام باحثي الأخلاقيات، يتسم، في المحصّلة، بقدرٍ مفرطٍ من الإيجابية أو السلبية، بل يمكن القول، مع قدرٍ من المعقولية، إن المجتمع قد ينتفع بجرعةٍ إضافية من التفاؤل بالتقنية. وعلى أي حال، لا يتوافر ما يكفي من المسوغات للاعتقاد بأنه من المرغوب أن تتجه جماعة أخلاقيات الذكاء الاصطناعي نحو السلبيّة[48]. وإذا انعدم هذا المسوّغ، وجب على جماعة الباحثين في هذا الحقل أن تسعى لتكون حَكَمًا محايدًا، لا إحدى القوى المنحازة. ومن المُرجح أن إحدى الثمار الإيجابية لهذا الحياد هي تعزيز ثقة الفاعلين الاجتماعيين الآخرين بجماعة باحثي هذا الحقل؛ إذ ستجد صعوبةً في كسب هذه الثقة إذا صُممت بنيتها على انحيازٍ سلبي، أو إذا تصورت نفسها صراحةً في موقع مواجهةٍ مع القوى المؤيدة للتقنية.

ما يترتب وما لا يترتب على ذلك

إذا كان تحليلي يسير، عامة، في الاتجاه الصحيح، فثمّة ما يدعو إلى افتراض أن التنظيم المؤسسي لأخلاقيات الذكاء الاصطناعي، بوصفها تخصصًا جامعيًا يمسّ موثوقيتها مسًا سلبيًا. فبحكم الحوافز التي تفرضها المنظومة، يُدفَع الباحثون إلى عرض الذكاء الاصطناعي في ضوء نقدي سلبي، ومن ثم يُرجَّح أن تنشأ صورة إجمالية لآثاره الأخلاقية لا تبدو أحادية الوجه فحسب، بل منحازة كذلك إلى السلبيات. فما الاستنتاجات التي ينبغي لنا استخلاصها من هذا؟

يتبدّى الدرس الأهم في ضرورة التقليل من شأن الصورة السلبية التي يرسمها مبحث أخلاقيّات الذكاء الاصطناعي حول هذا الأخير؛ إذ لا يعني ما يقدّمه هذا المبحث من رؤية قاتمة حول الآثار السلبية للذكاء الاصطناعي، بالضرورة، أنها صورة قاتمة إلى هذه الدرجة، ولنقل، بصورة أدق، يوفِّر انحياز السلبية حجة "مُلغٍ تقويضي" Undercutting Defeater[49] للرأي القائم على التقييم السلبي للجماعة

(47) Cass R. Sunstein, *Laws of Fear: Beyond the Precautionary Principle* (Cambridge: Cambridge University Press, 2005), pp. 41–43.

(48) ينسحب هذا أيضًا على التقنية بعامة، فليس بينًا أن المجتمع سيستفيد من انحياز جماعة أخلاقيات التقنية نحو السلبية والدعوة إلى مزيد من الحيطة والتنظيم.

(49) يحتل مصطلح المُلغِي Defeater والقابلية للإلغاء Defeasibility مكانًا بارزًا في الإبستيمولوجيا المعاصرة، ينظر: صلاح إسماعيل، **نظرية المعرفة: مقدمة معاصرة**، ط 2 (القاهرة: الدار المصرية اللبنانية، 2022)، ص 61–62؛ تميز الأدبيات الإبستيمولوجية بين نوعين رئيسَين من المُلغيات: 1. المُلغي الدحضي Rebutting Defeater، وهو سبب لاعتقاد نقيض القضية (ق) أو قضية أخرى تتعارض معها. مثال ذلك: إذا رأت مريم من بعيد ما يبدو خروفًا في الحقل فاعتقدت وجود خروف، ثم أخبرها صاحب الحقل أنه لا توجد خراف فيه، فقد اكتسبت مُلغيًا دحضيًا لمعتقدها. 2. المُلغي التقويضي Undercutting Defeater، وهو سبب للكفّ عن اعتقاد (ق) دون أن يكون سببًا لاعتقاد نقيضها، وذلك بمهاجمة الصلة بين الدليل والاستنتاج. ومثال ذلك: إذا رأى شخص قطعًا تبدو حمراء، ثم عَلِم أنها مُضاءة بأضواء حمراء، فقد فَقَد سببه لاعتقاد أنها حمراء، لكنه لم يكتسب سببًا لاعتقاد أنها ليست حمراء. يُستخدم هذا المفهوم في الورقة للإشارة إلى أن التحيز السلبي المؤسسي في أخلاقيات الذكاء الاصطناعي يُقدم مُلغيًا تقويضيًا للتقييم السلبي الذي تطرحه الجماعة العلمية [المترجم]، للمزيد ينظر أيضًا:

Michael Sudduth, "Defeaters in Epistemology," in: *Internet Encyclopedia of Philosophy*, accessed on 19/1/2026, at: https://acr.ps/1L9F2Qp

العلمية؛ ومفاده أن الذكاء الاصطناعي إشكالي بدرجة كبيرة. ومن المفترض، في الأحوال العادية، أن يُشكّل هذا التقييم السلبي دليلًا قويًا على أن الذكاء الاصطناعي يتسم بوضع إشكالي بدرجة كبيرة، وذلك تبعًا للطرائق الكثيرة والانتقادات التي دفع بها مجتمع أخلاقيات الذكاء الاصطناعي، أمّا مع وجود الانحياز، فإن هذا الدليل لا يدعم تلك النظرة القاتمة، أو أنه يدعمها بدرجة أوهى وأضعف. ومن الممكن صوغ هذا الدرس من منظور جماعة أخلاقيات الذكاء الاصطناعي نفسها، ونتعامل، إذّاك، مع ما يُمكن تسميته المُلغي من الدرجة العليا Higher–Order Defeater[50]، وهو يُعَدّ غالبًا نوعًا خاصًا من المُلغي التقويضي. وكما تُبين ماريا لاسونين–آرنيو، فإن "المُلغي" من الدرجة العليا "يعمل عبر إثارة الشك في أن الحالة الاعتقادية Doxastic التي يتبناها المرء ناتجة من عملية قاصرة"[51].

وقدّم ديفيد كريستنسن مثالًا نموذجيًا على ذلك في المثال التالي: "أنا طبيبٌ مقيمٌ أُشخِّص حالات المرضى وأصفُ العلاج المناسب. بعد أن شخصت حالة مريضٍ معين ووصفت له أدويةً محددة، أخبرتني الممرضة بأني مستيقظٌ منذ ستٍ وثلاثين ساعة. وبما أعرفه عن ميل الناس إلى ارتكاب أخطاء معرفية عندما يُحرَمون من النوم (وربما أيضًا لما أعلمه عن ضعف سجلي التشخيصي في مثل هذه الظروف)، خفضتُ درجة ثقتي بتشخيصي ووصفتي ريثما أعيد فحص رأيي بعناية"[52]. هنا يمتلك الطبيب دليلًا من الدرجة العليا Higher–Order Evidence يدعوه إلى الارتياب في تقييمه أدلة الدرجة الأولى First–Order الخاصة بحالة المريض والأدوية اللازمة. وعادةً يُناقَش "الدليل من الدرجة العليا" في إطار العقلانية الفردية، كما في هذه الحالة، غير أن بسطه ليشمل الجماعات العلمية ليس بعيدًا، فكما توجد عقلانية فردية تتصل بها أدلة من الدرجة العليا، يمكن أن يكون للمنظومة المعرفية ذاتها عقلانيّة بدرجة كبرت أم صغرت (أو موثوقة، وموضوعية)، متوافرة على أدلة من الدرجة العليا. فكما يرى الطبيب أن الأدلة من الدرجة الأولى تدعم تشخيصًا وعلاجًا معينين، تستشكل جماعة أخلاقيّات الذكاء الاصطناعي هذا الأخير، وذلك بما ينطوي عليه من مشكلات أخلاقية عديدة. وبهذا، فإن الأدلة من الدرجة الأولى تُقيّم بوصفها داعمة لنظرة سلبية شاملة لآثار الذكاء الاصطناعي الأخلاقية. وكما كان من المُتعيّن على الطبيب أن يقلل من موثوقية تشخيصه، يجدر بهذه الجماعة أن تخفف من يقينها في تقديرها القاتم لهذه التقنية.

---

(50) في نظرية المعرفة، تُعدّ "الأدلة من الدرجة الأولى" First–order Evidence هي تلك التي تتعلق مباشرةً بصدق قضية معينة، بينما "الأدلة من الدرجة العليا" Higher–order Evidence هي أدلة حول طبيعة أدلتنا ذاتها، أو حول قدراتنا المعرفية واستعداداتنا للاستجابة العقلانية لها. على سبيل المثال، إذا كان دليل الأرصاد الجوية دليلًا من الدرجة الأولى على أن المطر سيهطل غدًا، فإن معرفة أن عالِم الأرصاد كان في حالة إرهاق شديد عند تقييمه للبيانات تُعدّ دليلًا من الدرجة العليا. وقد أصبح هذا التمييز محوريًا في الأدبيات الإبستيمولوجية المعاصرة. (المترجم). للمزيد، ينظر:

Sophie Horowitz, "Higher–Order Evidence," *The Stanford Encyclopedia of Philosophy* (Spring 2025), Edward N. Zalta & Uri Nodelman (eds.), at: https://plato.stanford.edu/cgi–bin/encyclopedia/archinfo.cgi?entry=higher–order–evidence

(51) Maria Lasonen–Aarnio, "Higher–Order Evidence and the Limits of Defeat," *Philosophy and Phenomenological Research*, vol. 88 (2014), p. 314.

(52) David Christensen, "Higher–Order Evidence." *Philosophy and Phenomenological Research*, vol. 81, no. 1 (2010), p. 186.

مع ذلك، ثمّة أمور لا تترتّب على ما سبق، فلا يجب أن يُفهم من هذه الأطروحة أن استخدام الذكاء الاصطناعي لا يثير إشكالاتٍ أخلاقيةً جسيمة، فمن الواضح أن مثل هذه الإشكالات قائمة، وتظلّ دراستُها من باحثي الأخلاق مهمّةً بالغة الأهمية. وعلى سبيل المثال، يبدو لي أن احتمالاً، ولو ضئيلاً، لظهور ذكاءٍ اصطناعي خارق Superintelligence قادرٍ على إحداث دمار كارثي يُعَدّ مسألةً مهمة[53].

ولا يترتب على ما سبق أن كل إشكالية أخلاقية يُزعَم ارتباطُها بالذكاء الاصطناعي تعدّ وهمية أو مبالَغٌ فيها، فحتى لو وُجد ميلٌ داخل جماعة الحقل إلى الإكثار غير المبرَّر من هذه الإشكالات أو تضخيمها، يبقى من غير الجائز إقصاء أي ادعاءٍ محدَّد حول خللٍ أخلاقي يتعلق بالذكاء الاصطناعي. فالإلمام بالديناميات المؤسسية التي قادت إلى أزمة التكرار في العلوم التجريبية لا يبيح لنا رفض النتائج الإيجابية المُبهرة من حيث المبدأ، وبالمثل، فإن الوعي بما يعتور الحقل من اختلال لا يبرّر ردّ مزاعم الإشكال الأخلاقي في الذكاء الاصطناعي جملةً وتفصيلاً. وبناءً عليه، ينبغي التعاملُ مع كل دعوى على حدة وبجدية، لا بالطعن فيها استنادًا إلى اعتباراتٍ تخصُّ النظام كله[54]. فلا تهدف هذه الورقة إلى دحض مسألة أخلاقية بعينها، وإنما يسعى لتقييم موثوقية المنظومة الإبستيمية من منظور فلسفة العلم؛ أي على مستوى النظام لا على مستوى القضايا الجزئية داخل حقل أخلاقيات الذكاء الاصطناعي.

ختامًا، لا يعني ما تقدّم أن لدينا سببًا يدعونا إلى تفاؤلٍ إيجابيٍ خالص حيال الحصيلة الأخلاقية للذكاء الاصطناعي. فمجرد ملاحظة الصورة التي تُنظّم فيها المنظومة، كما تقدَّم الوصف، يوفر لنا "مُلغيًا تقويضيًا"، بدرجة غير محسومة، وذلك نتيجة الدليل الذي يحمله التقييم السلبي الصادر عن جماعة الحقل، غير أنه لا يُزودنا بحجة إيجابية تدفعنا إلى تبني موقفٍ محابٍ للذكاء الاصطناعي. فحتى لو ثبت أن حقل أخلاقيات الذكاء الاصطناعي يتسم بالخلل الذي أشرتُ إليه، يظل من الوارد، على المستوى النظري، أن يكون الذكاء الاصطناعي في المُجمل شرًا أخلاقيًا. ومن ثم، فإن نقدي للسلبية المتفشّية في هذا المبحث لا يرقى إلى بناء قضية إيجابية للتفاؤل بالذكاء الاصطناعي[55].

(53) ينظر على سبيل المثال:

Leopold Aschenbrenner, "Situational Awareness," (June 2024), accessed on 19/1/2026, at: https://acr.ps/1L9F2lw; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); Leonard Dung, "The Argument for Near–term Human Disempowerment through AI," *AI & Society*, vol. 50 (2025); Michael Huemer, "I, for One, Welcome Our Robot Overlords," *AI & Society* (2025); Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (London: Bloomsbury, 2020), Ch. 5; Eliezer Yudkowsky & Nate Soares, *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All* (New York: Little, Brown and Company, 2025).

(54) حول القيود على توظيف الدليل من المرتبة الأعلى في الجدل الفلسفي، ينظر:

Zach Barnett, "Philosophy without Belief," *Mind*, vol. 128 (2019); Peter Königs, "Bracketing Higher–order Evidence in Scholarly Philosophical Argumentation: Why and Which?" *Synthese*, vol. 205 (2025).

(55) حول صعوبة بناء حجّة إيجابية للتفاؤل بالتقنية، ينظر:

John Danaher, "Techno–optimism: An Analysis, an Evaluation and a Modest Defense," *Philosophy & Technology*, vol. 35 (2022).

خلاصة أخيرة، يتعين علينا التفكير في سبلٍ لإصلاح المشكلة، وإن لم يكن ذلك يسيرًا. ففي الحقول التي عصفت بها أزمة التكرار أمكن لِبضع تعديلاتٍ مؤسسية، كالتسجيل المسبق للدراسات، وإمكانية نشر النتائج الصفرية، والبيانات المفتوحة المصدر، أن تقطع شوطًا لا بأس به في معالجة هذه الأزمة. أمّا في أخلاقيات الذكاء الاصطناعي، فليس واضحًا أن حلولًا مماثلة متاحة؛ إذ لا يمكن، ولا ينبغي، العبثُ عقلانيًا بأي من العوامل المؤسسية الثلاثة. فموضوع الحقل سيظل، حتمًا، هو ذاته، ومن المنطقي أن يقدّم المشتغلون في المبحث الأخلاقي تعليقًا قيميًا على التطورات التقنية الجديدة. قد يُتَصوَّر، نظريًا، التخلي عن القاعدة التي تُثني عن المقاربات الإيجابية ذات الطابع الوصفي، لكن ليس من الواضح أن ذلك تصرّفٌ حصيف، فكما سبق لنا الذكر، يُمكن تَفهُّم الدافع لتثبيط باحثي أخلاقيات الذكاء الاصطناعي، وثنيهم عن إنتاج مقالات ضعيفة الطابع التقريري. ومن ثم، ثمة وجاهةٌ للإبقاء على هذه القاعدة. أما فيما يتعلق ببُنى الحوافز داخل الأكاديمية فإن تغييرها أمرٌ عسيرٌ، كما أن تقديم أي منظومة حوافز بديلة لمعالجة هذه المعضلة قد يستجلب مشكلاتٍ أخرى، وتنتهي بحصيلة أسوأ من سابقتها. فصحيح أن المقالات السجالية يُمكنها أداء دورٍ قيم في ترشيد التقييمات السلبية للذكاء الاصطناعي، لكننا قد نندم لو شجّعنا فيضًا من المقالات السجالية التي لا تُقدم سوى تحفظات طفيفة. يفضي هذا كلّه إلى معضلة: إن مشكلة السلبية تبدو أثرًا جانبيًا مزعجًا لتضافر عوامل مؤسسية، يُمكن تعقُّل كلٍ منها على حدة.

إذا كان إحداث تغييرات في الإطار المؤسسي أمرًا متعسِّرًا، فإن ثمَّة تدابير أخرى قد تعِدُ بتغييرات طفيفة. وتتضمن هذه التدابير تغيير طريقة تفكيرنا وممارساتنا البحثية. ومن المتعيَّن علينا، بصفتنا كُتاب مقالات حول أخلاقيات الذكاء الاصطناعي، أن نكون محيطين بالديناميَّات المؤسَّسية التي تقتادنا إلى ضرب من السلبية. ولا بُدَّ لنا من بذل جهدٍ واعٍ لتجنُّب الإفراط في سوق المظاهر السلبية. ويتعين علينا، بوصفنا مراجعين ومحرِّرين، إخضاع المزاعم المتصلة بالمشاكل الأخلاقية للذكاء الاصطناعي لتدقيق وفحص كافيين. ومن واجبنا، أيضًا، أن نضاعف جهودنا لإشراك غير المتخصصين في الأخلاقيات، ولا سيما مطوِّري الذكاء الاصطناعي وعلماء الكومبيوتر والزجّ بهم في مشاريعنا البحثية بهدف التحقق الخارجي من صحة المخرجات. فالسعي الحثيث للحصول على المزيد من الآراء الإيجابية من خارج الحقل الفلسفي لا يعني، بالضرورة، الموافقة عليها وتزكيتها، وإنما لاكتساب منظور أشمل ربما يعينُ في مواجهة تحيزات المباحث المعرفية.

ستبقى هذه الإجراءات حلاًّ مؤقتًا؛ إذ إنها تترك الديناميات المؤسسية من دون علاج. ومن المرجَّح أن تواصل أمواج الفزع ارتفاعها ما دامت أخلاقيات الذكاء الاصطناعي تشهد ازدهارًا. سيحدث ذلك، إلى حدٍ بعيد، بمعزل عن مدى الإشكاليات الأخلاقية الواقعية للذكاء الاصطناعي. أمّا "الخيار النووي"، فيتمثل في تقليص جحم هذا الحقل عن سبق إصرارٍ وتصميم، حرصًا على مصلحته. فقد أشار غير مؤلف إلى أن الحقل تضخّم على نحو غير ملائم[56]. فحجم التمويل المخصص له لا يثير

(56) Jan–Christoph Heilinger, "The Ethics of AI Ethics. A Constructive Critique," *Philosophy & Technology*, vol. 35 (2022), p. 16; Luke Munn, "The Uselessness of AI Ethics," *AI and Ethics*, vol. 3 (2023), p. 873; Felicitas Lambrecht & Marina Moreno, "What is AI Ethics?" *American Philosophical Quarterly*, vol. 61 (2024), pp. 397–398.

التساؤل فحسب قياسًا على تكلفة الفرص الضائعة أو العوائد الحدية المتناقصة Marginal Returns، بل قد يقوض أيضًا موثوقية المنظومة الإبستيمية عبر توليد فجوة بين حجم الجماعة البحثية في حقل أخلاقيات الذكاء الاصطناعي وعدد الإشكالات الأخلاقية المطلوب بحثها، فضلاً عن شدّتها.

اعترافٌ وتقدير

أود أن أتقدم بالشكر الجزيل للملاحظات القيمة والحوارات المُثرية لكل من المُحكمَين من مجلة Synthese، ولجمهور المحاضرات في كريت Crete، ودورتموند Dortmund، ودوسلدورف Düsseldorf، وأيندهوفن Eindhoven، وإرلنغن Erlangen، ومانهايم Mannheim، وللمشاركين في حلقة "أوبسالا–فيينا للذكاء الاصطناعي" Uppsala–Vienna AI Colloquium.

التمويل

إتاحة الوصول المفتوح جرى تنظيمها وتمويلها عبر Projekt DEAL.

المراجع References

Arora, Payal. *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech*. Cambridge, MA: MIT Press, 2024.

Arvan, Marcus, Liam Kofi Bright & Remco Heesen. "Jury Theorems for Peer Review." *British Journal for the Philosophy of Science*. vol. 76 (2025).

Avin, Shahar. "Centralized Funding and Epistemic Exploration." *The British Journal for the Philosophy of Science*. vol. 70 (2019).

Babic, Boris et al. "Beware Explanations from AI in Health Care." *Science*. vol. 373 (2021).

Barnett, Zach. "Philosophy without Belief." *Mind*. vol. 128 (2019).

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

_____. *Deep Utopia: Life and Meaning in a Solved World*. Idea, 2024.

Bright, Liam Kofi & Remco Heesen. "To Be Scientific Is to Be Communist." *Social Epistemology*. vol. 37 (2023).

Buckwalter, Wesley. "The Replication Crisis and Philosophy." *Philosophy and the Mind Sciences*. vol. 3 (2022).

Caplan, Bryan. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*. Princeton: Princeton University Press, 2007.

Christensen, David. "Higher–Order Evidence." *Philosophy and Phenomenological Research*. vol. 81, no. 1 (2010).

Danaher, John. *Automation and Utopia: Human Flourishing in a World without Work*. Cambridge, MA: Harvard University Press, 2019.

_____. "The Rise of the Robots and the Crisis of Moral Patiency." *AI & Society*. vol. 34 (2019).

_____. "Techno–optimism: An Analysis, an Evaluation and a Modest Defence." *Philosophy & Technology*. vol. 35 (2022).

Dellsén, Finnur. "The Epistemic Impact of Theorizing: Generation Bias Implies Evaluation Bias." *Philosophical Studies*. vol. 177 (2020).

Dishaw, Sophie. "The Right to a Justification." *Political Philosophy*. vol. 2, no. 4 (2025).

Douglas, Heather & Kevin C. Elliott. "Addressing the Reproducibility Crisis: A Response to Hudson." *Journal for General Philosophy of Science*. vol. 53 (2022).

Dung, Leonard. "The Argument for Near–term Human Disempowerment through AI." *AI & Society*. vol. 50 (2025).

Floridi, Luciano. "Introduction to the Special Issues: The Ethics of Artificial Intelligence." *American Philosophical Quarterly*. vol. 61 (2024).

Gunkel, David J. (ed.). *Handbook on the Ethics of Artificial Intelligence*. Cheltenham: Edward Elgar Publishing, 2024.

Hagendorff, Thilo. "Blind Spots in AI Ethics." *AI and Ethics*. vol. 2 (2022).

Heesen, Remco. "Why the Reward Structure of Science Makes Reproducibility Problems Inevitable." *The Journal of Philosophy*. vol. 115 (2018).

Heesen, Remco & Liam Kofi Bright. "Is Peer Review a Good Idea?" *The British Journal for the Philosophy of Science*. vol. 7 (2021).

Heilinger, Jan–Christoph. "The Ethics of AI Ethics. A Constructive Critique." *Philosophy & Technology*. vol. 35 (2022).

Himmelreich, Johannes. "Responsibility for Killer Robots." *Ethical Theory and Moral Practice*. vol. 22 (2019).

Hoffmann, Reid & Greg Beato. *Superagency: What Could Possibly Go Right with Our AI Future*. Authors Equity, 2025.

Hudson, Robert. "Should We Strive to Make Science Bias–free? A Philosophical Assessment of the Reproducibility Crisis." *Journal for General Philosophy of Science*. vol. 52 (2021).

Huemer, Michael. "I, for One, Welcome Our Robot Overlords." *AI & Society*. (2025).

Hull, David L. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago: University of Chicago Press, 1988.

Kitcher, Philip. "The Division of Cognitive Labor." *The Journal of Philosophy*. vol. 87 (1990).

Königs, Peter. "Artificial Intelligence and Responsibility Gaps: What is the Problem?" *Ethics and Information Technology*. vol. 24 (2022).

_____. "Bracketing Higher–order Evidence in Scholarly Philosophical Argumentation: Why and Which?" *Synthese*. vol. 205 (2025).

Lambrecht, Felicitas & Marina Moreno. "What Is AI Ethics?" *American Philosophical Quarterly*. vol. 61 (2024).

Lasonen–Aarnio, Maria. "Higher–Order Evidence and the Limits of Defeat." *Philosophy and Phenomenological Research*. vol. 88 (2014).

Lee, Carole J. "Commensuration Bias in Peer Review." *Philosophy of Science*. vol. 82 (2015).

Lobel, Orly. *The Equality Machine: Harnessing Digital Technology for a Brighter, more Inclusive Future*. London: Hachette UK, 2022.

London, Alex John. "Artificial Intelligence and Black–box Medical Decisions: Accuracy versus Explainability." *Hastings Center Report*. vol. 49 (2019).

Longino, Helen E. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press, 1990.

_____. *The Fate of Knowledge*. Princeton: Princeton University Press, 2002.

Maclure, Jocelyn. "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind." *Minds and Machines*. vol. 31 (2021).

May, Joshua. "Bias in Science: Natural and Social." *Synthese*. vol. 199 (2021).

Mayo–Wilson, Conor et al. "The Independence Thesis: When Individual and Social Epistemology Diverge." *Philosophy of Science*. vol. 78 (2011).

Milano, Silvia, Carina Prunkl. "Algorithmic Profiling as a Source of Hermeneutical Injustice." *Philosophical Studies.* vol. 182 (2025).

Munn, Luke. "The Uselessness of AI Ethics." *AI and Ethics*. vol. 3 (2023).

Ord, Toby. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury, 2020.

Peters, Uwe. "An Argument for Egalitarian Confirmation Bias and Against Political Diversity in Academia." *Synthese*. vol. 198 (2021).

_____. "Illegitimate Values, Confirmation Bias, and Mandevillian Cognition in Science." *The British Journal for the Philosophy of Science*. vol. 72 (2021).

Rebera, Andrew P., Lode Lauwaert & Ann–Katrien Oimann. "Hidden Risks: Artificial Intelligence and Hermeneutic Harm." *Minds & Machines*. vol. 35, no. 33 (2025).

Robbins, Philip. "Of Machines and Men: Attributions of Moral Responsibility in AI–Assisted Warfare." *Ethics and Information Technology*. vol. 27 (2025).

Smart, Paul R. "Mandevillian Intelligence." *Synthese*. vol. 195 (2018).

Solomon, Miriam. "Scientific Rationality and Human Reasoning." *Philosophy of Science*. vol. 59 (1992).

_____. *Social Empiricism*. Cambridge, MA: MIT Press, 2001.

Strevens, Michael. "The Role of the Priority Rule in Science." *The Journal of Philosophy*. vol. 100 (2003).

Sunstein, Cass R. *Laws of Fear: Beyond the Precautionary Principle*. Cambridge: Cambridge University Press, 2005.

Ulbert, Cornelia. et al. (eds.). *Moral Agency and the Politics of Responsibility*. London: Routledge, 2018.

Vredenburgh, Kate. "The Right to Explanation." *The Journal of Political Philosophy*. vol. 30 (2022).

Weisberg, Michael & Ryan Muldoon. "Epistemic Landscapes and the Division of Cognitive Labor." *Philosophy of Science*. vol. 76 (2009).

Yudkowsky, Eliezer & Nate Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. New York: Little, Brown and Company, 2025.

Zerilli, John et al. "Transparency in Algorithmic and Human Decision–Making: Is There a Double Standard?" *Philosophy & Technology*. vol. 32 (2019).

Zollman, Kevin J. S. "The Communication Structure of Epistemic Communities." *Philosophy of Science*. vol. 74 (2007).

_____. "The Epistemic Benefit of Transient Diversity." *Erkenntnis*. vol. 72 (2010).