مجلة وادى النيل للدراسات والبحوث الإنسانية والاجتماعية والتربوية

فلسفة نِك بوستروم في الذكاء الاصطناعي: الانفجار الذكائي، التداعيات الأخلاقية، وآفاق المستقبل

د. مينا سيتي يوسف فانوس

أستاذ فلسفة العلوم المساعد، قسم الفلسفة، كلية الآداب، جامعة القاهرة، مصر

menasity@cu.edu.eg

المُلخص:

تعدُّ مسألة الذكاء الاصطناعي من أبرز التحديات الفلسفية والأخلاقية في القرن الحادي والعشرين؛ إذ لا تقتصر أهميتها على الجانب التقني والعلمي فحسب، بل تمتد لتطال أسس التفكير الفلسفي حول ماهية الإنسان، وحدود الأخلاق، ومصير الحضارة الإنسانية ذاتها. ومن بين الأصوات الفكرية الأكثر حضورًا وإثارة للجدل في هذا المجال يبرز فيلسوف العلم السويدي **نِك بوستروم** (١٩٧٣-)، الذي كرّس مشروعه الفلسفي لفحص المخاطر الوجودية والتحديات الأخلاقية المرتبطة بالذكاء الاصطناعي الفائق، إضافة إلى تحليله العميق لإمكانية الانفجار الذكائي ومواءمة القيم.

يرى بوستروم أن الذكاء الاصطناعي لا يمثل مجرد تطوّر طبيعي في مسار التكنولوجيا، بل نقطة انعطاف جوهرية قد تعيد صياغة مصير الكائن البشري؛ إذ يمكن للذكاء الاصطناعي الفائق أن يتجاوز قدرات الإنسان بمستويات لا نهائية، ما قد يؤدي إما إلى ازدهار حضاري غير مسبوق أو إلى خطر وجودي قد ينهي وجود البشرية برمّته. من هنا؛ يطرح بوستروم ما يسميه "**مشكلة السيطرة**"، وهي التحدي المركزي في ضمان أن تبقى أهداف الذكاء الاصطناعي الفائق وقيمه منسجمة مع المصالح الإنسانية، في مقابل احتمال خروج هذه الأنظمة عن السيطرة، أو «الانغلاق القيمي» الذي قد يقيد الخيارات المستقبلية للبشر.

(ISSN: 2536- 9555)

وتتمثل أهم أسئلة هذه الدراسة فيما يلي: ما الإطار النظري الذي يقترحه بوستروم لفهم الذكاء الاصطناعي الفائق والانفجار الذكائي؟ كيف يعالج مشكلة السيطرة ومواءمة القيم أخلاقيًا وتقنيًا؟ كيف يؤثر منظوره على مفاهيم الهوية والوعي والوجود البشري؟ ما الانتقادات الرئيسة التي يمكن توجيهها لرؤيته؟ وستعتمد هذه الدراسة على منهج تحليلي نقدي مقارن، يجمع بين قراءة نصوص بوستروم الأساسية وتحليل المساهمات النقدية لمفكرين معاصرين. كما ستستند إلى قراءة فلسفية نقدية شاملة لفكر بوستروم في سياق مستقبل الذكاء الاصطناعي وأخلاقياته.

الكلمات المفتاحية: نِك بوستروم، الذكاء الاصطناعي الفائق، مشكلة السيطرة، الذكاء الاصطناعي الاستشاري (الأوراكل)، الخبرة الظاهرية (الكوالیا)، المخاطر الوجودية، تأثير الانتقاء الرصدي.

Nick Bostrom's Philosophy of Artificial Intelligence: The Intelligence Explosion, Ethical Implications and Future Prospects

Dr. Mena Sity Youssef Fanous
Associate Professor of Philosophy of Science: Department of Philosophy, Faculty of Arts, Cairo University, Egypt

Abstract:

The question of artificial intelligence constitutes one of the most pressing philosophical and ethical challenges of the twenty-first century. Its significance extends far beyond the technical and scientific domains to encompass fundamental philosophical reflections on the nature of the human being, the limits of morality, and the very fate of human civilization. Among the most prominent and controversial figures in this field is the Swedish philosopher of science Nick Bostrom (1973–), whose philosophical project has been devoted to

examining existential risks and the ethical challenges associated with superintelligent AI, in addition to his in-depth analysis of the prospect of an intelligence explosion and the problem of value alignment.

Bostrom argues that artificial intelligence is not merely a natural step in the trajectory of technological development, but rather a decisive turning point that could reshape the destiny of humankind. Superintelligent AI, he contends, may surpass human capacities by orders of magnitude, potentially leading either to unprecedented civilizational flourishing or to an existential threat that could eradicate humanity altogether. Central to this concern is what Bostrom terms the "control problem"-the fundamental challenge of ensuring that the goals and values of superintelligent systems remain aligned with human interests, as opposed to the risks of systems escaping control or becoming locked into rigid value structures that constrain humanity's future options.

The core questions of this study are therefore as follows: What theoretical framework does Bostrom propose for understanding superintelligence and the intelligence explosion? How does he address the control problem and value alignment, both ethically and technically? How does his perspective influence conceptions of identity, consciousness, and human existence? And what are the principal criticisms that are leveled against his vision? This study will adopt a critical, comparative, and analytical methodology, combining a close reading of Bostrom's primary texts with an examination of contemporary critiques. It will also situate Bostrom's thought within a broader philosophical reflection on the future of artificial intelligence and its ethical implications.

Keywords: Nick Bostrom, Superintelligence, Control problem, Oracle AI, Qualia, Existential Risks, Observational Selection Effect.

(ISSN: 2536- 9555)

مُقدِّمة

نِك بوستروم Nick Bostrom فيلسوف علم سويدي معاصر، ولد عام ١٩٧٣، يشتهر بأعماله الريادية في قضايا الذكاء الاصطناعي الفائق، المخاطر الوجودية، ومستقبل الإنسانية. بوستروم أستاذًا في كلية الفلسفة بجامعة أوكسفورد، نشر وحرّر ما يقرب من ١٣٠ عملًا فلسفيًّا متنوعًا بين كتبٍ ومقالات، منها: الانحياز الإنساني (٢٠٠٢) المخاطر الكارثية العالمية (٢٠٠٨) تعزيز الإنسان (٢٠٠٩) الذكاء الاصطناعي الفائق: المسارات والمخاطر والاستراتيجيات (٢٠١٤). يغطي بحثه مجموعة من الأسئلة الكبرى المتعلقة بمستقبل البشرية. يحمل بوستروم دكتوراه في فلسفة العلم وأسس نظرية الاحتمال، يدور موضوعها حول (نظرية الانتقاء الرصدي Observational Selection Theory) من كلية لندن للاقتصاد، بالإضافة إلى ماجستير في علم الأعصاب الحاسوبي، وماجستير آخر في الفيزياء والفلسفة. وقد حصل على زمالة ما بعد الدكتوراه من الأكاديمية البريطانية. لا تقتصر فلسفات بوستروم على فلسفة علم الذكاء الاصطناعي، بل تمتد أيضًا إلى الأخلاقيات الأحيائية، والميتافيزيقا، ونظرية الاحتمالات.

يمتلك بوستروم خلفية علمية في الفيزياء، وعلم الأعصاب الحاسوبي، والمنطق الرياضي، إلى جانب الفلسفة التحليلية. يُعدّ مفكرًا رائدًا في القضايا الكبرى المتعلقة بمستقبل البشرية، ويعمل أحيانًا مستشارًا خبيرًا لعدد من الوكالات الحكومية في المملكة المتحدة وأوروبا والولايات المتحدة الأمريكية، كما أنه معلّق دائم في وسائل الإعلام. شارك في تأسيس «الجمعية العالمية ما بعد الإنسانية» (في عام ٢٠٠٨)– التي غيّرت اسمها بعد ذلك إلى Humanity+ (اختصارها: H+)– و«معهد الأخلاقيات والتقنيات الناشئة»، وهو حاليًّا مدير «معهد مستقبل الإنسانية» التابع «لمدرسة مارتن» في أوكسفورد، وهو مركز بحثي رائد متعدد

التخصصات لدراسة التهديدات طويلة الأمد. وبفضل قدرته على تبسيط المواضيع الأكثر تجريدًا، أجرى مئات المقابلات حول الاستنساخ، والتجميد (cryonics)، والذكاء الاصطناعي، ومبدأ الأنثروبيا، وتقنية النانو، وحجة المحاكاة. وقد تُرجِمت كتاباته إلى ست عشرة لغة.

تشكّل أعمال بوستروم محطة محورية في الفلسفة المعاصرة، لكونها تمزج بين العمق الميتافيزيقي، والدقة الأخلاقية التحليلية، والقدرة على استشراف المستقبل بآفاق واسعة. اشتهر بوستروم بتحليلاته حول الذكاء الاصطناعي الفائق (Superintelligence)، المخاطر الوجودية (Existential Risks)، وتعزيز الإنسان (Human Enhancement)، فضلًا عن فرضيته المثيرة للجدل حول العيش في محاكاة حاسوبية.

يعبّر بوستروم في أعماله عن توجه فلسفي يقوم على مزج الأخلاقيات المستقبلية بالتفكير التكنولوجي البراجماتي والميتافيزيقا التحليلية، في مشروع متكامل يهدف إلى إعادة صياغة مواقفنا إزاء التطوّرات العلمية والتقنية الراهنة. كما يُعد أحد أبرز فلاسفة المستقبل في العصر الحديث؛ إذ أسهمت أعماله في إعادة تشكيل الأسئلة الفلسفية حول مصير البشرية، وطبيعة الذهن، والأخلاق في عصر التحولات التكنولوجية المتسارعة. يجمع بوستروم بين التفكير التحليلي الدقيق والرؤية المستقبلية البعيدة، ويندرج إنتاجه ضمن ميادين فلسفة العلم، والأخلاق، والميتافيزيقا، وفلسفة التقنية.

يُعرَّف كلٌ من ستيوارت راسل Stuart Russell وبيتر نورفيج Peter Norvig الذكاء الاصطناعي، في أبسط صوره، بوصفه «فرع من علوم الحاسوب يُعنى بتصميم أنظمة أو برامج قادرة على أداء مهام تتطلب عادة ذكاءً بشريًا» مثل الفهم اللغوي، وحل المشكلات، والتعلم، والاستدلال المنطقي[1]. وينقسم الذكاء

[1] Russell, Stuart J., and Peter Norvig. (2020). Artificial Intelligence: A Modern Approach. 4th ed. Hoboken, NJ: Pearson.

(ISSN: 2536- 9555)

الاصطناعي إلى قسمين رئيسيين: الذكاء الاصطناعي «الضعيف Narrow AI»، الذي يُصمَّم لأداء مهام محددة؛ و«الذكاء الاصطناعي العام Artificial general intelligence(AGI)»، الذي يُفترض فيه أن يمتلك قدرات معرفية شاملة قابلة للتعميم عبر مختلف السياقات، على غرار الذكاء البشري.

لقد أثار الذكاء الاصطناعي منذ نشأته أسئلة فلسفية لا تقل أهمية عن تحدياته التقنية. **ويمكن تصنيف هذه الإشكالات ضمن أربعة محاور رئيسة:**

١. ماهية الذهن والوعي: هل يمكن للآلة أن تفكر أو تعي؟ هل يكفي تقليد السلوك الذكي لنقول إن النظام واعٍ؟ هذا ما ناقشه جون سيرل في تجربته الذهنية «الغرفة الصينية»، منتقدًا الفرضيات القائلة إن معالجة الرموز تُنتج بالضرورة وعيًا.

٢. المشكلة الأنطولوجية: ما الفرق بين الكائن العاقل والكائن المصطنع؟ وهل يمكن أن يمتلك الذكاء الاصطناعي نوعًا من «الوجود الذاتي»؟ هنا تبرز تأملات الفلاسفة الوجوديين والتأويليين حول الجسد، والتجربة، والمعنى، كما في أعمال هايدجر وميرلوبونتي وهوبان.

٣. الأخلاقيات والمسؤولية: من يتحمّل المسؤولية عندما يتخذ الذكاء الاصطناعي قرارات خاطئة أو قاتلة؟ هل يمكن تعليم الآلة أخلاقًا؟ ما حدود الأتمتة في المجالات الحساسة كالطب والقضاء والحرب؟ هذه أسئلة يعالجها حقل «أخلاقيات الذكاء الاصطناعي AI Ethics» الذي ازدهر مؤخرًا مع بروز قضايا التحيّز، والخصوصية، والمساءلة.

٤. مستقبل الإنسان والماكينة: هل يُهدِّد الذكاء الاصطناعي مكانة الإنسان؟ هل نحن مقبلون على تفجُّر ذكائي يؤدي إلى ظهور ذكاء فائق يفوقنا قدرة؟ وما

تداعيات ذلك على السياسة، والاقتصاد، والهوية البشرية؟ هذه السيناريوهات يناقشها مفكرون مثل نِك بوستروم من زاوية مستقبلية تأملية.

يُمثّل الذكاء الاصطناعي اليوم مرآةً عاكسة لطموحاتنا البشرية ومخاوفنا العميقة في آنٍ معًا. إنه ليس مجرّد أداة تقنية، بل ميدان خصب للتفلسف، يعيد طرح الأسئلة القديمة في صورة جديدة: ما العقل؟ ما الإرادة؟ هل نحن أسياد آلاتنا، أم أن اختراعاتنا ماضية في تشكيلنا من جديد؟ من هنا؛ تنبع أهمية الخوض في فلسفة الذكاء الاصطناعي، لا بوصفها ترفًا نظريًا، بل كشرط لفهم ذواتنا في عصر ما بعد الإنسان.

من منظور فلسفة العلم، لا يمكن فصل نشأة الذكاء الاصطناعي عن تطورات أعمق في تاريخ العلوم: فقد جاء امتدادًا لمشروع المنطق الرمزي الذي دشّنه غوتلوب فريجه وبرتراند راسل في أواخر القرن التاسع عشر، والذي هدف إلى تأسيس الرياضيات على أسس منطقية صِرفة. ثم جاء آلان تورينج، مؤسس علم الحاسوب الحديث، ليحوّل هذا الطموح المنطقي إلى تصور حسابي؛ إذ طرح نموذج «الآلة الشاملة Turing Machine» كإطار صوري لإمكانات الحساب والتمثيل الرمزي.

هكذا؛ كان الذكاء الاصطناعي امتدادًا طبيعيًّا للرؤية الوضعية العقلانية التي تؤمن بإمكانية تمثيل الذهن كمنظومة من الرموز والقواعد، وهو ما عُرف لاحقًّا باسم أطروحة الرمزية التي مثّلت لحظة تأسيسية في الذكاء الاصطناعي الكلاسيكي. إلا أن فلسفة العلم أدركت مبكرًا حدود هذه الرؤية. فقد أشار فلاسفة مثل توماس كُون وبول فييرآبند إلى أن النماذج العلمية ليست مجرد تمثيلات محايدة، بل تُبنى ضمن أطر معرفية وتاريخية تتأثر بالسياق الاجتماعي والثقافي، ما يدعونا للتساؤل: هل تمثل نماذج الذكاء الاصطناعي حقًّا الطبيعة الذهنية، أم أنها مجرد إسقاطات لأُطرنا البشرية على الآلة؟

(ISSN: 2536- 9555)

قد يعني الذكاء الاصطناعي معانيَ مختلفة، ويُعرّف بطرقٍ مختلفة. فعندما قدّم آلان تورينج ما يُسمى باختبار تورينج (الذي أسماه «لعبة المحاكاة») في مقالته الشهيرة عام (١٩٥٠) حول قدرة الآلات على التفكير، لم يكن مصطلح «الذكاء الاصطناعي» قد طُرح بعد. ناقش تورينج قدرة الآلات على التفكير، واقترح أنه سيكون من الأوضح استبدال هذا السؤال ليحل محله سؤال حول إمكانية بناء آلات قادرة على تقليد البشر بشكل مقنع لدرجة يصعب معها التمييز بين ما إذا كانت الرسالة المكتوبة، على سبيل المثال، صادرة عن حاسوب أم عن إنسان[٢]. ثم صاغ مجموعة من الباحثين العاملين في مجال تطوير الذكاء الاصطناعي– جون مكارثي، ومارفن ل. مينسكي، وناثانيال روتشستر، وكلود إي. شانون– (الذين نظموا ورشة عمل صيفية شهيرة لمدة شهرين في كلية دارتموث حول دراسة الذكاء الاصطناعي) مصطلح «الذكاء الاصطناعي» عام (١٩٥٥).

خضعت حجة جون سيرل لتقييم نقدي في مواجهة الادعاءات المضادة للوظيفية والحوسبة. ويُقال عمومًا إن الذكاء لا يتطلب بيئةً أساسيةً محددة، كالكائنات الكربونية، ولكنه سيتطور أيضًا في بيئات قائمة على السيليكون، إذا كان النظام معقدًا بما يكفي (يُمكننا على سبيل المثال الرجوع إلى الفصل التاسع من كتاب ديفيد تشالمرز «الذهن الواعي»[٣]). في السنوات الأولى من القرن الحادي والعشرين، ربط عديد من الباحثين الذكاء الاصطناعي بشكل أساسي بأشكال مختلفة مما يُسمى بالتعلم الآلي، أي التقنيات التي تحدد الأنماط في البيانات. ويُقال إن الأشكال الأبسط من هذه الأنظمة تُمارس التعلم الخاضع

(2) Turing, A. M. (1950). "Computing Machinery and Intelligence." Mind 59 (236): 433–460.
(3) Chalmers, David J. (1996). The Conscious Mind. In Search of a Fundamental Theory. Oxford: Oxford University Press.

للإشراف– الذي لا يزال يتطلّب، مع ذلك، تدخلاً وإشرافًا بشريًّا كبيرًا، إلا أنّ هدف عديد من الباحثين كان تطوير ما يُسمى بأنظمة التعلم ذاتي الإشراف.

وقد حذر بوستروم من المخاطر المحتملة للتفرّد التكنولوجي في حالة انقلاب الآلات الذكية ضد صانعيها، أي البشر. لذلك، ووفقًا له، من الأهمية القصوى بناء ذكاء اصطناعي ودود. بالفعل سبق وقدّم كاتب الخيال العلمي الشهير إسحاق أسيموف أول مدونة أخلاقية لأنظمة الذكاء الاصطناعي، إذ قدّم قوانينه الثلاثة للروبوتات في قصته القصيرة «الركض في حلقة[٤]». ثم أضاف إلى هذه القوانين الثلاثة قانون رابع، يُسمى «القانون الصفري للروبوتات»، في كتابه" «الروبوتات والإمبراطورية»[٥] **وهذه القوانين الأربعة هي:**

١. لا يجوز للروبوت أن يؤذي إنسانًا، أو أن يسمح بإيذاء إنسان من خلال التقاعس عن العمل؛
٢. يجب على الروبوت أن يطيع الأوامر الصادرة إليه من البشر، باستثناء الحالات التي تتعارض فيها هذه الأوامر مع القانون الأول؛
٣. يجب على الروبوت أن يحمي وجوده طالما أن هذه الحماية لا تتعارض مع القانون الأول أو الثاني؛
٤. لا يجوز للروبوت أن يؤذي البشرية، أو يسمح بإلحاق الأذى بها عن طريق التقاعس.

لعبت قوانين أسيموف الأربعة دورًا محوريًّا في أخلاقيات الآلة لعقود عديدة، وناقشها الخبراء على نطاق واسع. ويُعتقد أن هذه القوانين الأربعة مهمة، لكنها غير كافية لمعالجة جميع التعقدات المتعلقة بأخلاقيات الآلات. ويبدو هذا التقييم منصفًا؛ إذ لم يدّع أسيموف قط أن قوانينه قادرة على معالجة جميع المشكلات.

(4) Asimov, Isaac. (1942). "Runaround." Astounding Science-Fiction, March, 94–103.
(5) Asimov, Isaac. (1985). Robots and Empire. New York: Doubleday.

(ISSN: 2536- 9555)

ولو كان الأمر كذلك بالفعل، لما كتب أسيموف قصصه الشيقة عن المشكلات التي تُعزى جزئيًّا إلى هذه القوانين الأربعة.

تبدو فكرة الانفجار الذكي الذي يتضمن آلات ذكاء اصطناعي فائقة الذكاء ذاتية التكاثر فكرةً مستحيلة للكثيرين؛ ويرفض بعض المعلقين هذه الادعاءات بوصفها خرافة حول مستقبل تطور الذكاء الاصطناعي. ومع ذلك؛ فإن أصواتًا بارزة، داخل الأوساط الأكاديمية وخارجها، تأخذ هذه الفكرة على محمل الجد– لدرجة أنهم يخشون العواقب المحتملة لما يُسمى «بالمخاطر الوجودية» مثل خطر انقراض البشر. ومن بين من أعربوا عن هذه المخاوف فلاسفة مثل نِك بوستروم وتوبي أورد Toby Ord[٦] (١٩٧٩–)، بالإضافة إلى شخصيات بارزة مثل إيلون ماسك والعالم الراحل ستيفن هوكينج. ووفقًا لهذه التقييمات، يجب التعامل مع الذكاء الاصطناعي على قدم المساواة مع الأسلحة النووية وغيرها من التقنيات شديدة التدمير التي قد تعرضنا جميعًا لخطر كبير ما لم يتم مواءمة القيم بشكل صحيح.

تقدّم تقنيات الذكاء الاصطناعي اليوم تقدمًا سريعًا يفوق الوصف، ويتبادر السؤال: إلى أي مدى يتوقع أن تُغيّر هذه التقنيات حياتنا ومجتمعنا والحضارة البشرية جمعاء؟ وفي هذا السياق؛ من المجدي إلقاء نظرة تأمّلية على الخاصية التي مكّنت الإنسان العاقل Homo sapiens خلال حقبة الهولوسين من السيطرة اللافتة على الكرة الأرضية. من الواضح أن هذه السيطرة لا تعود إلى القوة العضلية أو التحمل البدني، بل إلى الذكاء الذي يتميز به الإنسان. فإذا سلّمنا بأن الذكاء هو مورد فريد من نوعه، أتاح لنا الوصول من السافانا إلى القمر؛ فإن هذه المرحلة التي نقوم فيها بأتمتة هذا المورد وتفويضه للآلات قد تكون المرحلة الأكثر

[٦] فيلسوف أسترالي بجامعة أكسفورد، مختص في أخلاقيات المستقبل والمخاطر الوجودية، وصاحب كتاب «حافة الهاوية: المخاطر الوجودية ومستقبل الإنسانية» (٢٠٢٠).

تأثيرًا في التاريخ البشري. سواء انتهى الأمر إلى أن تحلّ الآلات مكاننا كأذكى وأكثر مخلوقات الأرض قوة، أم لا؛ فإن الأثر المتوقع للذكاء الاصطناعي على الحضارة الإنسانية سيكون تحويليًا بكل المقاييس.

اقتصر التقدّم غالبًا على ما يُعرف «بالذكاء الاصطناعي المحدود»، وهي أنظمة مختصّة بمهارات ضيقة كلعب الشطرنج أو قراءة الحروف اليدوية. ومع ذلك؛ ظل الحلم قائمًا لبناء ذكاء اصطناعي عام (AGI) يعادل أو يتجاوز الذكاء البشري على كامل نطاق القدرات الإدراكية. إلا أن اهتمام الباحثين بالأمان بدأ يتبلور تدريجيًا بفضل الباحث والكاتب الأمريكي إليعازر يودكوفسكي Eliezer Yudkowsky[٧] (١٩٧٩–) نفسه في العقد الأول من القرن الحادي والعشرين، وهو المجال المعروف اليوم «بمواءمة الذكاء الاصطناعي AI alignment»، أي ضمان أن تكون أهداف الذكاء الاصطناعي القوي متوافقة بما فيه الكفاية مع أهداف البشر لضمان مستقبل جيد للبشرية.

[٧] أحد روّاد التفكير حول الذكاء الاصطناعي العام (AGI) والمخاطر الوجودية، ويُعَد من أبرز الأسماء في النقاشات المتعلقة بالذكاء الاصطناعي الفائق ومخاطره المستقبلية. مؤسس مشارك لمعهد أبحاث ذكاء الآلة (Machine Intelligence Research Institute –MIRI) في كاليفورنيا. وهو معهد يركز على مشكلة «مواءمة الذكاء الاصطناعي AI alignment»، أي ضمان أن تكون أنظمة الذكاء الاصطناعي المتقدمة متوافقة مع القيم والأهداف البشرية. يُعد من أبرز الأصوات المبكرة التي ناقشت مشكلة السيطرة ومواءمة القيم قبل بوستروم. له كتاب بارز بعنوان «العقلانية: من الذكاء الاصطناعي إلى الزومبي» ((٢٠١٥، تدور مقالاته حول العقلانية والذكاء الاصطناعي والفلسفة. كما أن له سلسلة مقالات شهيرة بعنوان «Sequences» نُشرت على موقع «LessWrong»، تُعد من المصادر الأساسية في فلسفة الذكاء الاصطناعي والمخاطر الوجودية.

(ISSN: 2536- 9555)

اشتهرت شركة OpenAI بقدرتها على تطوير نماذج اللغة الكبيرة (LLMs)، بـدءًا مـن GPT-1 فـي (٢٠١٨)، و GPT-2 فـي (٢٠١٩)، وGPT-3 فـي (٢٠٢٠)، ثـم GPT-3.5 فـي (٢٠٢٢)، وGPT-4 فـي (٢٠٢٣)، وصـولًا إلـى GPT-5 فـي (٢٠٢٥). هـذه النمـاذج قائمـة علـى بنيـة «المحـوّلات transformers» التي اخترعتها شركة جوجل Google، لكن فريق OpenAI راهن على مبدأ التكبير المستمر لحجم النماذج وبيانات التدريب وفتراته، وأسفر هذا عن إصدار نسخة ChatGPT في نوفمبر (٢٠٢٢)، التي استقطبت نحو مئة مليون مستخدم خلال شهرين فقط، لتسجّل بذلك أسرع معدلات النمو في تاريخ التطبيقات.

في هذه الأثناء، ظهرت مفاجأة جيوغرافية في يناير (٢٠٢٥) حين كشفت الشركة الصينية DeepSeek عن نموذج (r1) مفتوح الوزن، يُنافس أداء أفضل النماذج. مما أثار صدمة في مجتمع الذكاء الاصطناعي، لا بسبب ولاء النموذج للحزب الشيوعي الذي تم تدريبه عليه، بل لأن فجوة التقدم بين الولايات المتحدة والصين ربما أقل مما كان مُتصور. وقد ساعد ذلك على تصعيد السباق الأمريكي الصيني في الذكاء الاصطناعي، وتحوّلت الخطابات السياسية إلى نوع من الجنون القومي والتنافس الأيديولوجي. مع هذه النظرة الشاملة لتاريخ الذكاء الاصطناعي حتى وقت كتابة المقال، نتقدم الآن للبحث في القضايا الأكثر تكهنًا والمستقبلية: ما الذي يمكن أن نتوقعه في السنوات القادمة؟

المخاطر الأخلاقية للذكاء الاصطناعي هي أي مخاطر مرتبطة به قد تُؤدي إلى إخفاق أصحاب المصلحة في الوفاء بواحدة أو أكثر من مسؤولياتهم الأخلاقية تجاه أصحاب المصلحة الآخرين. ويرى بوستروم أن البشرية من المرجح أن تطور ذكاءً فائقًا، يُعرَّف بأنه الذكاء الذي يتجاوز إلى حد كبير الأداء المعرفي للبشر في جميع المجالات تقريبا، في وقت ما خلال حياتنا. وعلى الرغم من أن بوستورم لا

يقدم موعدًا نهائيًا؛ فإنه يجادل بأنه من المرجح أن يصل الذكاء الاصطناعي الفائق في غضون بضعة عقود، وأن الانتقال من مستوى الذكاء البشري إلى الذكاء الاصطناعي الفائق سيكون سريعًا للغاية. لقد شهدت العقود الأخيرة اهتمامًا متصاعدًا بمجال الذكاء الاصطناعي، الذي لم يعد مجرد فرع من فروع علوم الحاسوب، بل بات ميدانًا عابرًا للتخصصات، تتقاطع فيه الأسئلة التقنية بالمشكلات الفلسفية، ويشتبك فيه الممكن التكنولوجي بالمتخيّل الإنساني. فمنذ أن ظهرت أولى المحاولات لتصميم أنظمة قادرة على محاكاة السلوك الذكي، راح الذكاء الاصطناعي يطرح تحديات جذرية تمسّ مفاهيم الذهن، والوعي، والمعرفة، والحرية، والأخلاق، بل وحتى ماهية الإنسان ذاته.

تندرج فلسفة بوستروم ضمن ما يسمى **بالفلسفة التحذيرية المستقبلية Future-Oriented Precautionary Philosophy**؛ إذ يربط بين الذكاء الاصطناعي وتعزيز الإنسان، ويسعى إلى إعادة التفكير في حدود الطبيعة البشرية وإمكانات تحسينها. هذا الربط الوثيق بين الذكاء الاصطناعي وما بعد الإنسانية Transhumanism يفتح باب النقاش حول قضايا الهوية، والوعي، والذات، إضافة إلى مسائل العدالة والإنصاف في توزيع قدرات تحسين الإنسان.

لا تقتصر أهمية فلسفة بوستروم على التنظير الفلسفي البحت، بل تتجاوزها إلى تحريك النقاش العملي والسياسي حول كيفية التعامل مع الذكاء الاصطناعي، وكيفية صياغة سياسات عامة تضمن الحماية من المخاطر الوجودية، مع المحافظة على المبادئ الأخلاقية الأساسية. في هذا السياق، يبرز مفهوم «الأخلاق المستقبلية طويلة المدى»، الذي يرى بوستروم أنه ينبغي أن يحتل موقعًا مركزيًا في تخطيط البشرية لمستقبلها.

يستكشف بوستروم موضوعات إضافية ذات صلة بالذكاء الاصطناعي الفائق. تشمل هذه الموضوعات: أنواع محددة من الذكاء الاصطناعي الفائق، كالأنظمة

(ISSN: 2536- 9555)

التي تقتصر على الإجابة عن الأسئلة، وتلك التي تنفذ الأوامر، وتلك التي تتصرف باستقلالية تامة؛ السيناريوهات التي لا يمتلك فيها كيانٌ واحدٌ ميزة استراتيجية على غيره، الآثار المحتملة على سوق العمل العالمي والاقتصاد العالمي في ظل وجود وكلاء ذوي ذكاء فائق؛ كيف يمكن للقوى العالمية التعاون وتخصيص الموارد من أجل تطوير ذكاء فائق آمن بصورة مثالية.

كما يُعالج قضايا فلسفية، مثل: الاعتبارات الأخلاقية وسؤال ما إذا كان يمكن أن يكون للذكاء الفائق سعادة خاصة به؛ العلاقة المحتملة بين الذكاء الفائق والوعي (إن وُجدت). ويتناول كذلك مسألة كيفية اتخاذ قرار بشأن القيم التي ينبغي ترميزها في الذكاء الفائق، بالنظر إلى أن البشر يخضعون لعوائق معرفية مثل التحيزات الشخصية والآليات النفسية التي تقاوم كثيرًا المعلومات «الموضوعية» التصحيحية، وإلى واقعة أنه لا يوجد توافق بين الفلاسفة على أفضل نظرية أخلاقية. يتناول بوستروم ما يجب القيام به على الفور– سواء على المستوى الفردي أم المجتمعي– للتقليل من مخاطر الذكاء الاصطناعي في المستقبل.

هناك معيار اجتماعي مهم في التعامل مع المنظمات، وهو القدرة على تحديد الشخص المسؤول عن إنجاز مهمة ما. عندما يفشل نظام ذكاء اصطناعي في مهمته، من يتحمّل اللوم؟ المبرمجون؟ أم المستخدمون النهائيون؟ المسؤولية، والشفافية، والقابلية للتدقيق، والنزاهة، والقابلية للتنبؤ، وتجنب جعل الضحايا الأبرياء يصرخون يأسًا؛ جميعها معايير تنطبق على البشر عند أدائهم للوظائف الاجتماعية، وجميعها معايير يجب مراعاتها عند تصميم خوارزميات تهدف إلى استبدال التحكم البشري في المهام الاجتماعية، وجميعها معايير قد لا تظهر في مجلة تعلم آلي مهتمة بكيفية توسع الخوارزميات على أجهزة أكثر. هذه القائمة من المعايير ليست شاملة بأي حال، لكنها تمثل عينة صغيرة مما ينبغي أن تفكر فيه المجتمعات التي تصبح أكثر حوسبةً يومًا بعد يوم.

ينبغي أن نتذكر أن التوقعات المبنية على الأمل فحسب قد تسببت سابقًا في مشكلات في أبحاث الذكاء الاصطناعي. إذ إن بناء ذكاء عام يمكن الوثوق به والتحقق من هذه الموثوقية سيتطلب أساليب وطرق تفكير مختلفة تمامًا عن مجرد فحص برمجيات محطة طاقة بحثًا عن الأخطاء؛ بل سيتطلب ذكاءً عامًا يفكر مثل مهندس بشري مهتم بالأخلاق، لا مجرد نتاج «هندسة أخلاقية» بسيطة.

في عام (١٩٦٥)، طرح عالِم الرياضيات والحاسوب إرفين جون جود Irving John Good الفرضية الكلاسيكية بشأن الذكاء الاصطناعي الفائق، ومؤداها أن نظامًا يمتلك من الذكاء ما يتيح له فهم تصميمه الخاص، قد يتمكّن من إعادة تصميم نفسه أو ابتكار نظام أكثر ذكاءً، فينشأ بذلك تسلسل من التحسين الذاتي المتسارع عبر حلقة تغذية راجعة إيجابية. وأطلق جود على هذه العملية اسم «انفجار الذكاء intelligence explosion»[٨].

حتى لو قصرنا أنفسنا على استعارات تاريخية، يتضح أن الذكاء الاصطناعي الفائق يطرح تحديات أخلاقية غير مسبوقة حرفيًّا. عند هذه النقطة، لم تعد المخاطر محصورة في مستوى الأفراد (كرفض قرض منزل أو احتراق بيت أو إساءة معاملة وكيل شخصي)، بل أصبحت على مستوى كوكبي أو كوني (كأن تُفنى البشرية ويحل محلها شيء لا نراه ذا قيمة). أما إذا أُمكن تشكيل الذكاء الاصطناعي الفائق ليكون نافعًا، فقد يستطيع، بناءً على قدراته التقنية، حل عدة مشكلات استعصت على ذكاء الإنسان.

يُعد الذكاء الاصطناعي الفائق واحدًا من «المخاطر الوجودية» التي عرّفها بوستروم خلال عام (٢٠٠٢) بأنها: المخاطر التي يؤدي تحققها إلى إبادة الحياة

(8) Good, Irving John. (1965). "Speculations Concerning the First Ultraintelligent Machine." In Advances in Computers, edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press.

(ISSN: 2536- 9555)

الذكية التي نشأت على الأرض، أو تقييد إمكاناتها بشكل دائم وجذري[٩]. وعلى العكس؛ فإن نتيجة إيجابية للذكاء الاصطناعي الفائق قد تحافظ على الحياة الذكية الأرضية وتحقق إمكاناتها القصوى.

من المهم تأكيد أن العقول الأذكى تحمل إمكانات عظيمة للفوائد، بقدر ما تنطوي على مخاطر. ومع ذلك؛ حتى إذا افترضنا إمكانية تحديد نظام أهداف يظل ثابتًا أثناء التعديل والتحسين الذاتي، فهذا لا يمس سوى سطح المشكلات الأخلاقية الجوهرية المتعلقة بابتكار ذكاء فائق. لقد استخدم البشر، كأول ذكاء عام على الأرض، ذكاءهم لإعادة تشكيل الكوكب بشكل جذري–من نحت الجبال وترويض الأنهار إلى بناء ناطحات السحاب وزراعة الصحاري وإحداث تغيّرات مناخية غير مقصودة. وقد يترتب على ذكاء أقوى عواقب أكبر بكثير. وعلى الرغم من أن الذكاء الاصطناعي الحالي لا يطرح علينا قضايا أخلاقية جديدة تختلف جذريًّا عمّا نواجهه في تصميم السيارات أو محطات الطاقة؛ فإن اقتراب خوارزميات الذكاء الاصطناعي من التفكير الشبيه بالبشر ينبئ بتعقدات متوقعة.

ما الآثار التي قد يتركها مزيد من الانفتاح في تطوير الذكاء الاصطناعي على الآثار طويلة الأمد لهذا الذكاء؟ وهل القيمة المتوقعة لهذه الآثار على المجتمع إيجابية أم سلبية؟ ونظرًا لصعوبة تقديم إجابات قاطعة عن هذه الأسئلة؛ فإن طموح بوستروم هنا أكثر تواضعًا: تقديم بعض الاعتبارات ذات الصلة، وتطوير بعض الأفكار حول وزنها ومدى معقوليتها. حتى هذا الطموح المتواضع يمكن أن يقدم مساهمة قيّمة.

هل يقود الانفتاح إلى تسريع تطوير الذكاء الاصطناعي ونشره؟ وهل يُعَدّ تسريع تطوير الذكاء الاصطناعي ونشره أمرًا مرغوبًا؟ حتى فترة قصيرة من الحيازة

(9) Bostrom, Nick. (2002b). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." Journal of Evolution and Technology, Vol. 9.

الحصرية يمكن أن تمكن المبتكر من تحقيق أرباح من المعرفة الداخلية، مثل أن يكون أول من يعلم بإمكانية تطبيق تكنولوجيا مؤثرة في السوق. حافز آخر للابتكار في النظام المفتوح هو إمكانية استفادة المبتكر من امتلاكه لأصل تكميلي تزيد قيمته بفضل الفكرة الجديدة. وبالمثل؛ قد تختار شركة برمجيات إتاحة برامجها مجانًا بهدف زيادة الطلب على خدمات الاستشارات والدعم الفني (التي تكون الشركة، بحكم كتابتها للبرمجيات، في أفضل وضع لتقديمها).

وكما هو الحال مع أي تقنية عامة الأغراض، يمكن تحديد مخاوف حول تطبيقات معينة. فقد جرى الإشارة، على سبيل المثال، إلى أن التطبيقات العسكرية للذكاء الاصطناعي- بما في ذلك الأسلحة المستقلة الفتاكة- قد تشعل سباقات تسلح جديدة، أو تقلل من عتبة الدخول في الحروب، أو تمنح الإرهابيين والقتلة أدوات جديدة للعنف. كما يمكن استخدام تقنيات الذكاء الاصطناعي لشن هجمات سيبرانية. ويمكن لخوارزميات التعرف على الوجوه، وتحليل المشاعر، والتنقيب في البيانات أن تُستخدم للتمييز ضد مجموعات مهمشة، أو لانتهاك الخصوصية، أو لتمكين الأنظمة القمعية من استهداف المعارضين السياسيين بفعالية أكبر.

من بين المخاوف المتداولة أيضًا أن التقدّم في الذكاء الاصطناعي قد يُحدث اضطرابات في سوق العمل، مما يقلّل من فرص بعض الفئات العمالية في الحصول على وظائف. ليس من الواضح أن قدرات الذكاء الاصطناعي في المدى القريب والمتوسط تطرح تحديات مميزة في هذا الصدد، وهي تحديات لا تنطبق فحسب على الأتمتة عمومًا، بل وعلى جزء كبير من جميع التغيرات التكنولوجية، التي غالبًا ما تقلل الطلب على أنواع معينة من العمل البشري أيضًا.

يُقيم بوستروم مدى استحسان الانفتاح في تطوير الذكاء الاصطناعي على المدى الطويل بالإشارة إلى كيفية تأثير الانفتاح في مشكلتين جوهريتين ترتبطان

(ISSN: 2536- 9555)

بابتكار نظم ذكاء اصطناعي متقدمة للغاية (سواء أكانت في مستوى الذكاء البشري عمومًا أم فائقة الذكاء):

- **مشكلة السيطرة:** كيفية تصميم نظم الذكاء الاصطناعي بحيث تقوم بما يقصده مصمّموها.
- **المشكلة السياسية:** كيفية تحقيق وضع يُستخدم فيه الذكاء الاصطناعي، من قِبل الأفراد أو المؤسسات المخوّلة به، بطرق تعزز الصالح العام(١٠).

إن التعجيل بتطوير الذكاء الاصطناعي سيمنح العالم وقتًا أقل للاستعداد للذكاء الاصطناعي المتقدم. قد يقلل هذا من احتمالية حل مشكلة السيطرة. هناك أيضًا بعض العمليات الأخرى غير العمل المباشر على سلامة الذكاء الاصطناعي، التي قد تعزز الاستعداد بمرور الوقت– ولن تُتاح لها فرصة كافية إذا حدث الذكاء الاصطناعي في وقت أبكر– مثل تحسين القدرات الإدراكية، وتحسين منهجيات ومؤسسات وآليات التنسيق المختلفة.

إن تسريع الذكاء الاصطناعي سيزيد من احتمال أن يستبق الذكاء الاصطناعي الفائق المخاطر الوجودية الناجمة عن مصادر غير مرتبطة بالذكاء الاصطناعي، مثل المخاطر التي قد تنشأ من علم الأحياء التركيبي، أو الحروب النووية، أو تكنولوجيا النانو الجزيئية، أو مخاطر أخرى غير متوقعة حتى وقت الناس هذا. ويعتمد هذا الأثر الاستباقي على افتراض أن ظهور الذكاء الاصطناعي الفائق سيقضي أو يقلل المخاطر الوجودية البشرية الكبرى الأخرى. إن تحقّق ذلك قد يتوقف جزئيًا على طبيعة العالم بعد تطوّر الذكاء الاصطناعي، أي على ما إذا كان سيكون متعدد الأقطاب أو أحادي القطب.

وكلما زاد عدد المنافسين، زادت صعوبة تنسيقهم جميعًا لتجنب سباق مخاطرة ينحدر إلى القاع. في وضع تنافسي شديد، قد يعني قبول هذه الإعاقة بشكل

(10) Bostrom, Nick. (2014a) Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

أحادي التخلي عن الصدارة. على النقيض، في وضع أقل تنافسية (كحالة وجود تحالف كبير يملك تقدّمًا واسعًا في التكنولوجيا أو القدرة الحاسوبية) قد يكون هناك مساحة كافية تسمح للمتصدر بتنفيذ تدابير سلامة تقلل الكفاءة دون فقدان تقدمه. قد تكون التضحية بالأداء لصالح سلامة مؤقتة فحسب بمثابة إجراء مرحلي حتى يتم تطوير طرق تحكم أكثر تطورًا، تلغي العيب الكفائي للذكاء الاصطناعي الآمن.

في عالم تعددي، حيث توجد مراكز مبادرة مستقلة عديدة، نتوقع في النهاية إطلاق هذا الكائن (ربما عن طريق الخطأ، أو كجزء من عملية ابتزاز، أو على يد فاعل يحمل قيمًا مدمرة، أو في سياق حرب). لتبدو احتمالية تجنب هذه النتيجة أقل كلما زاد عدد الفاعلين المستقلين الذين لديهم قدرة التوصل إلى التكنولوجيا الحيوية ذات الصلة. ويمكن توسيع هذا المثال؛ إذ قد لا تتوافر للهجوم على التكنولوجيا الحيوية ميزة مشابهة، غير أنها قد تبرز في الحرب السيبرانية، أو في تقنيات النانو الجزيئية، أو في أنظمة الطائرات المسيّرة المتقدمة، أو في تكنولوجيات أخرى غير متوقعة قد تطوّرها ذكاءات اصطناعية فائقة.

إن عالمًا تظل فيه مشكلات التنسيق العالمي بلا حل، بينما تقترب القدرة التكنولوجية من حدودها الفيزيائية، هو عالم يظل رهينة لإمكانية ميل الطبيعة بقوة نحو التدمير بدلًا من الخلق عند مستوى معين من التطوّر. من منظور الحد من المخاطر الوجودية، قد يكون من الأفضل إذن أن يظهر ترتيب مؤسسي يتيح تنسيقًا عالميًا قويًا. لكن خلق حالة يكون فيها العالم مجزأً ومتعدد الأقطاب إلى درجة تجعل من المستحيل التأثير في مساره المستقبلي، يجب أن يكون مصدر قلق.

إذا تحقق الذكاء الاصطناعي العام؛ فقد يؤدي أيضًا إلى ظهور ذكاء فائق. ويمكن تعريف الذكاء الاصطناعي الفائق بشكل تقريبي كالتالي: "أي عقل يتجاوز

(ISSN: 2536- 9555)

بكثير الأداء المعرفي للبشر في جميع المجالات ذات الأهمية تقريبًا"[١١]. ومن التصوّرات المطروحة لظهور الذكاء الاصطناعي الفائق أن بناء قدرة معرفية اصطناعية بمستوى بشري قد يفضي إلى نظام قادر على ابتكار ذكاء أرقى، فينتج تسلسل من التحسين الذاتي المتتابع، يتجاوز القدرات البشرية، وربما يتسارع بوتيرة متنامية، وهو ما يُشار إليه بـ «انفجار الذكاء».

هناك سؤالان رئيسيان يحيطان بهذا التطوّر: متى يُتوقَّع حدوثه– إن حدث أصلًا؟ وما أثره المحتمل؟ خصوصًا ما المخاطر التي قد يتضمنها وصولًا إلى مستوى الخطر الوجودي على البشرية؟ كما قال هوكينج وزملاؤه: "إن النجاح في خلق ذكاء اصطناعي سيكون الحدث الأضخم في تاريخ البشرية. وللأسف، قد يكون أيضًا الأخير ما لم نتعلم كيف نتجنب تلك المخاطر"[١٢].

الفلسفة مطالَبة اليوم بتقديم توضيحات؛ فبعد أن كان يُحتفى بها تاريخيًّا بوصفها سعيًا إلى أعمق الأسئلة وأكثرها جوهرية –تلك التي تشكّل أساس جميع أنماط البحث الأخرى– باتت الآن تواجه وابلًا من الهجمات المشكِّكة في جدواها من الأساس. في مواجهة هذا التصور، يأتي بوستروم ليشكّل تحديًا مباشرًا له. يقوم بشرح وافٍ للآثار الفلسفية المترتبة على الأبحاث الراهنة في تطوير ما يُعرف «بالذكاء الفائق»، مقدِّمًا بذلك فوائد ملموسة للتفكير الفلسفي، ومستندًا في الوقت ذاته على مستجدات علمية معاصرة.

يرى بوستروم أن الذكاء الفائق مرجّح الحدوث خلال العقود القليلة القادمة، وأن الانتقال من مستوى الذكاء البشري إلى الذكاء الفائق سيكون سريعًا إلى حد

(11) Ibid, P. 22.
(12) Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014, May, 1). Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough? The Independent.

الانفجار– ربما خلال ساعات أو حتى دقائق[١٣]. ثم يشرع بوستروم في تمييز المسارات المتعددة المحتملة نحو تحقيق الذكاء الفائق، ويشير إلى أن وجود طرق عدة يُعزِّز من قابلية هذا الاحتمال للتحقق في نهاية المطاف. من بين هذه المسارات: تطوير ذكاء اصطناعي قادر على التحسّن الذاتي المتكرر؛ ومحاكاة البُنى الحسابية الأساسية– وإن كانت بدائية– في الدماغ البشري عبر إسقاطها على الحاسوب؛ وتعزيز الإدراك البيولوجي الطبيعي بشكل جوهري (مثلًا؛ عبر الانتقاء الجنيني التكراري ضمن الجيل الواحد، أو عبر إزالة الطفرات الجينية المُضرة بالذكاء بشكل انتقائي)؛ وإنشاء واجهات دماغية حاسوبية، وشبك العقول الفردية بطريقة بالغة الكفاءة لإنتاج ما يمكن تسميته «بالذكاء الفائق الجمعي».

ويخلص بوستروم في النهاية إلى أن الذكاء الاصطناعي هو المرشح الأوفر حظًا ليكون أول من يبلغ مستوى الذكاء البشري العام، ومِن ثَمَّ يتجاوزه إلى الذكاء الفائق. وذلك بسبب وفرة الموارد المادية والمحتوى المعرفي، والإمكانات الهائلة للخوارزميات التطورية، فضلًا عن أن البدائل الأخرى ستشهد، على الأرجح، عوائد متناقصة مقابل الجهود المجتمعية المبذولة في سبيل تطويرها.

ويبدو أن حجته حول إمكانية نشوء ذكاء اصطناعي فائق مقنعة إلى حد بعيد. **أولًا:** لقد بيّنت لنا مسيرة التطور أن الذكاء البشري يمكن أن ينشأ من ركائز مادية. ومن المفترض أن بالإمكان إعادة تحقيق ذلك افتراضيًا عبر الخوارزميات التطوّرية، مما يتيح تجاوز بطء الأجيال الملازم للانتخاب الطبيعي، وتفادي ما يسميه بوستروم «التحيّز الأنثروبي Anthropic Bias»، أي الخطأ المتمثل في

(13) Bostrom, N, Müller VC. (2014). Future progress in artificial intelligence: A survey of expert opinion. In Müller VC (ed), Fundamental Issues of Artificial Intelligence. Berlin: Springer, pp. 555-572.

(ISSN: 2536- 9555)

الاستنتاج، بناءً على كون الحياة الذكية قد تطورت على الأرض، بأن العمليات التطوّرية التي أفضت إلى ذلك كانت ذات احتمال أولي معقول لإنتاج الذكاء(١٤).

وقد تُظهر محاكاة مسار التطور افتراضيًا أن العمليات التي أفضت إلى الذكاء البشري ليست كافية لإنتاج الذكاء بشكل عام. إلا أن التقدُّم الهائل في العتاد الصلب والبرمجيات سيتيح للباحثين اختبار عدد هائل من المسارات التطوّرية المحتملة بسرعات كبيرة، حتى يتمكنوا من تحديد أحد المسارات القادرة على توليد الذكاء العام. ومِن ثَمَّ، يمكن تعزيز هذا الذكاء العام من خلال جهدنا نحن (أي عبر تزويده بعتاد وبرمجيات أفضل)، ومن خلال قدرته هو على التحسين الذاتي المتكرر.

قد يعبر الذكاء الاصطناعي الفائق لبوستروم عن «آلة فرانكنشتاين» تستولي على السيطرة لتحقيق غاياتها المنحرفة، أو على أنه فجر عصر جديد، قدرٌ حتمي في مسار التاريخ، لعالم تهيمن فيه آلات أذكى من البشر، وتؤول فيه البشرية إلى الانقراض. وربما يكون نسخة محدّثة ومحسّنة من أطروحة نيتشه عن «الإنسان الأعلى» الذي يعيد تقويم جميع القيم ويتجاوز الخير والشر.

لكننا نرى أن رسالة بوستروم الجوهرية لا تنتمي لأي مما سبق، بل هي مباشرة وبسيطة: بما أن فلاسفة الأخلاق لم يتمكنوا من الاتفاق حول القيم القصوى، ولا تمكنوا من تفسير كيفية اكتساب القيم أو ما إذا كانت القيم حقيقية أم لا، وبالأخص لأنهم لم ينجحوا في تحديد معايير لاختيار القيم القصوى؛ فإننا عاجزون عن تعليم الأنظمة فائقة الذكاء– التي تتجاوز الذكاء البشري– الأهداف التي نرغب في أن تسعى إليها بما ينسجم مع مصلحة البشرية. ونتيجة لذلك، قد ننتهي إلى إنتاج شكل من الذكاء الاصطناعي الفائق تكون فيه الآلات الذكية قد تعلمت كيفية السيطرة على الكون، وقررت أن البشر لم يعودوا ضروريين، أو قد

(14) Bostrom, Nick. (2002a). Anthropic Bias: Observation Selection Effects in Science and Philosophy. New York: Routledge.

تُحوّل الكون إلى فضاء لا مكان فيه للجنس البشري، كأثر جانبي للتغيرات الفيزيائية والبيولوجية التي تحدثها هذه الأنظمة الفائقة القوة والمعرفة. ومع ذلك؛ يقترح بوستروم، بحذر وتردد بعض التقنيات التي يمكن غرسها في الأنظمة الذكية كي تعوّض عن أوجه القصور في الفلسفة الأخلاقية، وتزود هذه الأنظمة بمكون تعلم القيم وتحسين أخلاقياتها، مما قد يخفف– وإن لم يُزل تمامًا– التهديد الوجودي الذي تواجهه البشرية.

قد يتساءل المرء ما إذا كان بوستروم مجرد نذير كاذب لا لبزوغ عصر الذكاء الاصطناعي الفائق فحسب، بل ليوم قيامة متخيل ووشيك، يكون محصلة الخوف منه إثارة الهلع دون مبرر. ولكن، لنفترض أن تحذيراته تستند إلى شيء من الواقع، وأن الذكاء الاصطناعي الفائق سيصبح جزءًا من مستقبل البشرية، إن لم يكن في أعمار من هم أحياء في الوقت الحاضر، فعلى الأقل قبل أن يصطدم نيزك ضخم بالأرض، أو– في أسوأ الأحوال– قبل انهيار النظام الشمسي. لنفترض هذا لأجل التحقق مما إذا كانت هناك فائدة فلسفية يمكن استخلاصها من أطروحة الذكاء الاصطناعي الفائق أو «التفرّد التكنولوجي Singularity»؛ إذ يؤدي التقدم المتسارع في التكنولوجيا إلى تغيرات جذرية– اجتماعية واقتصادية، والأكثر إثارة للدهشة، بيولوجية– يمكن أن تتلاقى في لحظة من التحول التاريخي المعروف باسم التفرّد(١٥).

يُلاحَظ أن بوستروم لا يولي اهتمامًا كافيًا للاعتراض الفلسفي الموجَّه إلى الذكاء الاصطناعي، والذي يرى أن اختزال الذكاء إلى خوارزميات يقتضي استبعاد الوعي والتفكير الحقيقيين؛ إذ لا تقوم النظم الذكية سوى بمحاكاة مظاهر الذكاء أو تقليدها في أحسن الأحوال. حتى في حالة أنظمة التعلّم الآلي، فإن هذه الأنظمة

(15) Bostrom, Nick. (2014a) Superintelligence: Paths, Dangers, Strategies. Op. Cit, P. 2.

(ISSN: 2536- 9555)

لا تمتلك وعيًا بأنها تتعلم، أي إنها ليست واعية لذاتها، ومن ثمّ فإن ما يُسمّى «تعلّمها» يُفهَم، وفق هذا الاعتراض، على أنه محاكاة للتعلّم لا تعلّمًا أصيلًا.

ومع ذلك، إذا سلّمنا بالافتراض الضمني لدى بوستروم القائل بأن ذكاء الآلة وتعلّمها يمكن اعتباره ذكاءً وتعلّمًا حقيقيين، ما دام كلاهما يُعرَّف تشغيليًا من حيث الأداء الوظيفي وإنجاز المهام؛ وإذا تابعنا خطّه الفكري على هذا الأساس، وجدنا أنفسنا إزاء أطروحة فلسفية عميقة تتعلق بالقيم، وبكيفية إعادة تحديد العلاقة بين القيم الإنسانية وتلك التي قد تنبثق في سياق ما بعد الإنسان، وهي القضية التي سنعرض لها لاحقًا.

يؤكد بوستروم أن هذا المستقبل ليس حتميًا بأي حال، لكن احتماله كافٍ ليجعلنا نتعامل معه بجدية تامة. فإن كان هذا الكيان يكترث بالبشر؛ فقد تؤول الأمور إلى خير يفوق خيالنا. أما إن لم يكترث، فقد يكون مستقبلنا بائسًا إلى حد كبير، أو قصير الأمد للغاية. ومن منظور الذكاء الاصطناعي الفائق: "قد يُنظر إلى البشر بوصفهم تهديدات محتملة؛ وهم بالتأكيد يُعَدّون موارد مادية"[١٦]. وقد يعمد ذكاء فائق لا مبالٍ إلى إعادة تشكيل المادة التي تتكون منها أجسادنا أو بيئتنا لتناسب غاياته الخاصة، التي قد تكون غريبة بالكامل عن مصالحنا أو قيمنا.

تتسارع وتيرة التطوّرات بوتيرة غير مسبوقة، ومن غير الواضح ما النتائج التي قد تترتب عندما يمتد أثر الذكاء الاصطناعي إلى أتمتة ليس الأعمال اليدوية فحسب، بل أيضًا طيفًا متزايدًا من المهام الفكرية المعقدة. على المستوى الأساسي، يعد تحررنا من مشقة العمل أمر جيد، يتيح لنا التركيز بدلًا من ذلك على الفن والثقافة والرياضة والحب أو أي شيء نرغب بملء حياتنا به، ولكن هل يمكن تحقيق هذا الانتقال نحو يوتوبيا كهذه دون عواقب اجتماعية سلبية هائلة؟

(16) Ibid, P. 116.

كان آلان تورينج قد تنبأ بالفعل في مقاله الآلات الذكية: نظرية هرطقية عام (١٩٥١)[١٧] بأن الآلات ستبلغ في النهاية مستوى من الذكاء يفوق البشر، ثم ستستولى بسرعة على زمام السيطرة في العالم. فما الذي يمكننا توقعه من هذا الاختراق الموعود؟ ما إن يصبح لدى الذكاء الاصطناعي ذكاء عام يتجاوز ذكاءنا؛ فسيصبح أفضل منا في بناء ذكاء اصطناعي أذكى منه. وهكذا في دوامة تصاعدية نحو مستويات متزايدة من الذكاء.

من جهة؛ قد يكون خلق ذكاء فائق مفتاحًا لحل كل مشكلاتنا وتحقيق كل ما نتمناه. ومن جهة أخرى؛ قد يتبين أن هذا الاختراق خطير إلى أقصى حد، ويؤدي في أسوأ الأحوال إلى انقراض البشرية. تتمحور الرسالة الرئيسة في مؤلفات بوستروم حول ضرورة إعمال التفكير المتأني واتخاذ التدابير الملائمة استباقًا للحظة الاختراق، إذ قد يغدو التدخل المتأخر غير ذي جدوى.

سيكون هذا النظام فعليًا نوعًا جديدًا من الحياة، ومخاوف بوستروم– في أبسط أشكالها– تطوّرية. إذ ستصبح البشرية مهزومة بشكل غير متوقع من قبل منافس أكثر ذكاءً. علينا أحيانًا ملاحظة–كنقطة مقارنة– مسارات الناس والغوريلا؛ كلاهما من الرئيسيات، ولكن مع سيطرة نوع واحد على الكوكب والآخر على حافة الإبادة.

يمكن القول إن بوستروم هو الفيلسوف الرائد ما بعد الإنساني اليوم، نادرًا ما يقدم تنبؤات ملموسة، ولكن بالاعتماد على نظرية الاحتمالات، يسعى إلى استخلاص رؤى تبدو مستحيلة. لا يُقصد من الذكاء الاصطناعي الفائق أن يكون أطروحة للأصالة العميقة؛ فمساهمة بوستروم بمثابة فرض صرامة الفلسفة التحليلية على مجموعة فوضوية من الأفكار، ظهرت على هامش الفكر الأكاديمي. ربما لأن مجال الذكاء الاصطناعي قد حقق مؤخرًا تقدمًا مذهلاً–مع

(17) Turing, A.M. (1951). Intelligent machinery, a heretical theory. Reprinted in Cooper and van Leeuwen (2013), pp. 664–666.

(ISSN: 2536- 9555)

ظهور التكنولوجيا اليومية، أكثر فأكثر، لإظهار شيء مثل التفكير الذكي. ويقارن مؤيدو بوستروم ذلك «بالربيع الصامت» في الفلسفة الأخلاقية.

قد يكون القليل من التفكير المسبق طويل المدى التزامًا أخلاقيًا تجاه جنسنا البشري. لذا؛ يدير بوستروم معهد مستقبل الإنسانية كنوع من محطة الرادار الفلسفية؛ مخبأ يرسل نبضات ملاحية إلى ضباب المستقبل المحتمل. عندما تبلورت هذه الأفكار في تفكيره، بدأ في إيلاء مزيدًا من الاهتمام لمسألة الانقراض. لم يكن يعتقد أن يوم القيامة وشيك، بل كان اهتمامه بالمخاطر كاهتمام وكيل التأمين. لذا؛ جادل بأنه بغض النظر عن مدى احتمال حدوث الانقراض؛ فإن عواقبه تكاد تكون سيئة للغاية، وبالتالي؛ حتى أصغر خطوة نحو تقليل فرصة حدوثه تكاد تكون ذات قيمة لا نهائية.

يؤرخ بوستروم أول تحليل علمي للمخاطر الوجودية بالرجوع إلى مشروع مانهاتن. ففي عام (١٩٤٢) أصبح روبرت أوبنهايمر قلقًا من أن التفجير الذري بقوة كافية يمكن أن يتسبب في اشتعال الغلاف الجوي بأكمله. وخلصت دراسة لاحقة إلى أن السيناريو كان غير معقول، بالنظر إلى القيود المفروضة على الأسلحة التي كانت قيد التطوير في ذلك الوقت. ولكن حتى لو لم تتحقق الكوابيس النووية العظيمة للحرب الباردة؛ فإن الأدوات كانت موجودة لإحداث دمار على نطاق لم يكن ممكنًا من قبل. ومع تزايد تعقد الابتكارات، يصبح من الصعب بشكل متزايد تقييم المخاطر المقبلة. يجب أن تكون الإجابات محفوفة بالغموض؛ لعدم إمكانية اشتقاقها إلا من خلال التنبؤ بتأثيرات التقنيات الموجودة في الغالب كنظريات، أو باستخدام المنطق المجرد بشكل غير مباشر.

نظرًا لاستحالة التعافي من المخاطر الوجودية، لا يمكننا السماح بوقوع حتى كارثة وجودية واحدة؛ إذ لن تكون هناك فرصة للتعلّم من التجربة. لذلك يجب أن يكون نهجنا في إدارة هذه المخاطر نهجًا استباقيًّا. علاوة على ذلك، يثير تقييم

المخاطر الوجودية مشكلات منهجية مميزة تتعلق بتأثيرات انتقاء الراصد، والحاجة إلى تجنب الانحياز الأنثروبي. وتُعد الفصول الخاصة بالتكنولوجيا النانوية والذكاء الاصطناعي أمثلة على هذا التحليل الاستباقي للمخاطر المستقبلية. وفي بعض الحالات، قد يكون من المهم دراسة سيناريوهات من شبه المؤكد أنها مستحيلة فيزيائيًّا.

يؤكد بوستروم باستمرار على أن التطوّر الأخلاقي والمعرفي ينبغي أن يسبق أو يواكب التطوّر التكنولوجي، وإلا فإن العواقب قد تكون كارثية. وتتمحور هذه الدراسة حول أبرز أطروحاته الفلسفية، عبر تناول أربعة محاور رئيسة: الذكاء الاصطناعي الفائق، والمخاطر الوجودية، وفرضية المحاكاة، ومواءمة القيم. وتسعى إلى تقديم قراءة تحليلية لفكر بوستروم بشأن الانفجار الذكائي، مع التركيز على الإشكالات الأخلاقية التي يثيرها، والبدائل التي يقترحها، فضلًا عن المآزق النظرية المترتبة على ذلك.

انطلاقًا من هذه الإشكاليات، تهدف هذه الدارسة إلى تقديم قراءة فلسفية نقدية شاملة لفكر بوستروم في سياق الذكاء الاصطناعي؛ من خلال تحليل فكرة الانفجار الذكائي، واستعراض التحديات الأخلاقية المرتبطة بمشكلة السيطرة ومواءمة القيم، ودراسة العلاقة بين الذكاء الاصطناعي الفائق وتعزيز الإنسان، فضلًا عن مناقشة فرضية المحاكاة وأثرها على مفهوم الهوية والوعي.

إن ما يميز هذه الدراسة هو طموحها إلى المساهمة في إثراء النقاش الفلسفي حول قضايا الذكاء الاصطناعي، وتأكيد ضرورة التفكير الجاد في المخاطر الوجودية، وإعادة تعريف موقع الإنسان في عالم قد تهيمن عليه كائنات ذكية فائقة القدرة. ومن خلال هذا التحليل النقدي، تسعى الدراسة إلى تقديم توصيات فكرية وأخلاقية، تسهم في صياغة رؤى أكثر توازنًا وواقعية لمستقبل البشرية.

(ISSN: 2536- 9555)

أولًا: مفهوم الانفجار الذكائي الفائق

يُعرَّف الذكاء الاصطناعي بأنه العلم المعني بصناعة الكائنات الذكية وهندستها؛ حيث يشير مفهوم «الذكاء» هنا إلى القدرة على إدراك البيئة، واتخاذ قرارات فعّالة، والتعلم من البيانات، والتكيف مع الحالات المستجدة. غير أن هذا التعريف التقني سرعان ما انفتح على إشكاليات فلسفية تتعلق بماهية الذكاء؟ وهل يمكن فصله عن السياق الجسدي والتجريبي للبشر؟ وما الذي يجعل نظامًا ما «عاقلاً» بالمعنى العلمي؟

في هذا السياق، يتجاوز الذكاء الاصطناعي كونه مجرد أداة، ليغدو كائنًا نظريًا theoretical entity، شأنه شأن «الذرة» أو «الجين» في فروع أخرى من العلم. إنه ليس موضوعًا للتقنية فحسب، بل بناءٌ تفسيري يسعى إلى محاكاة الظواهر العقلية أو إعادة تشكيلها وفق نماذج قابلة للحوسبة، مما يجعله حقلًا متميزًا على تخوم فلسفة العلم المعاصرة.

قبل ظهور الحوسبة الحديثة، كان الذكاء مفهومًا فلسفيًّا وأنثروبولوجيًّا ارتبط بالسؤال عن ماهية العقل البشري، وإمكاناته، وعلاقته بالجسد والعالم. وفي الفكر اليوناني، رُبط الذكاء «باللوغوس λόγος»؛ أي العقل القادر على التفكير المنطقي والاستدلال. إذ رأى أرسطو أن «الذكاء nous» هو الجزء الإلهي في النفس، المتمايز عن الحواس، والمرتبط بالقدرة على إدراك المبادئ الأولى. مع الفلسفة الحديثة، وخصوصًا مع ديكارت، تعزَّز التصور الثنائي الذي يفصل بين «العقل المفكّر» و«الجسد الممتدّ»، مما مهد الطريق لفكرة أن الذكاء شيء يمكن عزله عن الجسد ومحاكاته بشكل مجرد. في المقابل، جاء فلاسفة مثل هيوم وكانط ليؤكدوا الطابع التركيبي والتجريبي للعقل؛ إذ لا يوجد ذكاء «خالص» بل يتوسط دائمًا بين الإدراك والتجربة، وهو ما سيكون له أثر كبير لاحقًا في نقد النماذج الرمزية للذكاء الاصطناعي.

ابتداءً من القرن التاسع عشر، بدأت الرياضيات والمنطق تطرحان مفهومًا جديدًا للذكاء بوصفه قدرة حسابية رمزية. يمكن تلخيص هذه المرحلة في المحطات التالية: قدّم جورج بول في كتابه «القوانين الرياضية للفكر» (١٨٥٤) نظامًا للمنطق الرمزي يحول قوانين التفكير المنطقي إلى عمليات جبرية، وهو ما شكّل الأساس النظري الذي بُنيت عليه الحوسبة المنطقية لاحقًا. وسعيا كل من جوتلوب فريجه وبرتراند راسل لتأسيس الرياضيات على منطق صوري صارم، مفترضين أن التفكير الرياضي لا يختلف عن الاستدلال الرمزي. ثم أثبت كورت جودل (١٩٣١) في «مبرهنة عدم الاكتمال» أن هناك حدودًا لأي نظام صوري، مما زرع الشك في قدرة النماذج الرمزية على استيعاب كل الذكاء الإنساني. صاغ آلان تورينج (١٩٣٦) بعد ذلك «آلة تورينج» كنموذج صوري للحساب، واقترح لاحقًا (١٩٥٠) اختبارًا لسلوك الذكاء عُرف بـ«اختبار تورينج»، ليتبيّن ما إذا كانت الآلة قادرة على تقليد العقل البشري بطريقة لا يمكن تمييزها عن الإنسان. ومنذ ذلك الحين، لم يُنظر إلى الذكاء باعتباره خاصية ميتافيزيقية أو فطرية فحسب، بل كشيء قابل للترميز والتنفيذ على آلة. ويُعرَّف الذكاء هنا بوصفه قدرة على معالجة الرموز وحل المشكلات، وفق نموذجٍ للعقل باعتباره آلة رمزية.

بدأت موجة جديدة من الذكاء الاصطناعي مع التوجه نحو نماذج إحصائية واحتمالية تعتمد على البيانات، متجاوزة النموذج الرمزي، أي الذكاء الاصطناعي «غير الرمزي Subsymbolic AI» الذي يشمل الشبكات العصبية، الخوارزميات التطوّرية، والتعلم المعزز. ومع ظهور «التعلم العميق Deep Learning» وتطور المعالجات والبيانات، استطاعت الأنظمة أن تتعلم أنماطًا معقدة في الصور، اللغة، والصوت، دون تدخل بشري كبير. على سبيل المثال؛ أنتج نظام GPT (ابتداءً من عام ٢٠١٨) نصوصًا لغوية عالية الاتساق، محاكيًا الإنتاج اللغوي البشري. واتسع مفهوم الذكاء في الوقت الحاضر ليشمل القدرة على

(ISSN: 2536- 9555)

التكيف، التعلم من التجربة، ومعالجة اللاحتمية. لكنه بات أكثر غموضًا، وأثار تساؤلات من قبيل: هل الذكاء هو النجاح العملي؟ أم الفهم؟ أم التفسير؟

مع كل تحول تقني، تغير الفهم الفلسفي لما يعنيه «الذكاء»:

١. النموذج الرمزي Symbolic AI: العقل بحسبه آلة رموز (مينسكي، نيويل)[١٨].

٢. النموذج الاتصالي Connectionism: العقل شبكة ديناميكية (الذكاء كنمط منبثق emergent)[١٩].

٣. النموذج التجسيدي Embodied Cognition: لا ينفصل الذكاء هنا عن الجسد، والعالم، والخبرة المعيشة (فينوجراد، لاكوف، ميرلوبونتي)[٢٠].

٤. النموذج التطوري/الاحتمالي Evolutionary & Probabilistic[٢١]: الذكاء قدرة على اتخاذ قرارات فعّالة في بيئات معقدة وغير متوقعة.

---

(18) Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. Communications of the ACM, 19(3), 113–126.

Minsky, M. (1986). The Society of Mind. New York: Simon & Schuster.

(19) Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press.

Smolensky, P. (1988). On the proper treatment of connectionism. Behavioral and Brain Sciences, 11(1), 1–23.

(20) Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. Cambridge, MA: MIT Press.

Lakoff, G., & Johnson, M. (1999). Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. New York: Basic Books.

Winograd, T., & Flores, F. (1986). Understanding Computers and Cognition: A New Foundation for Design. Norwood, NJ: Ablex.

Merleau-Ponty, M. (1962). Phenomenology of Perception. (C. Smith, Trans.). London: Routledge.

وهكذا؛ لم يعد الذكاء «جوهرًا» أو «قدرة موروثة»، بل بات يُفهم كمجموعة عمليات يمكن تصميمها، اختبارها، وتعديلها، بل وأحيانًا تنشئتها اصطناعيًا. وبدوره يمكننا التمييز بين ثلاثة أشكال من الذكاء الاصطناعي الفائق:

١- **الذكاء الاصطناعي الفائق السريع speed superintelligence**: يمكنه القيام بكل ما يفعله العقل البشري، ولكن بسرعة أكبر بكثير. نظام يعمل بسرعة تفوق العقل البشري بـ١٠,٠٠٠ مرة، مثلًا؛ سيتمكن من قراءة كتاب في ثوانٍ. في نظر هذا العقل السريع، يبدو العالم الخارجي وكأنه يتحرّك ببطء شديد.

٢- **الذكاء الاصطناعي الفائق الجمعي collective superintelligence**: نظام يتكون من عدد كبير من العقول البشرية المستوى، مُنظَّمة بطريقة تجعل أداء النظام ككل يتفوق كثيرًا على أي نظام إدراكي حالي. وإذا أمكن تشغيل «عقل بشري» على هيئة برنامج حاسوبي؛ فيمكن حينئذ نسخه وتشغيله على عدة حواسيب. إذا كان لكل نسخة قيمة كافية لتغطية تكلفة العتاد والطاقة؛ فقد يحدث انفجار سكاني ضخم. في عالم يضم تريليونات من هذه العقول، قد يغدو التقدم التكنولوجي أسرع بكثير مما هو عليه اليوم؛ إذ سيكون هناك عدد من العلماء والمخترعين يفوق الحالي بآلاف المرات.

٣- **الذكاء الاصطناعي الفائق النوعي quality superintelligence**: عقل أسرع على الأقل من العقل البشري، وأذكى نوعيًا بشكل بالغ. هذه الفكرة أعسر استيعابًا؛ إذ يُقصد بها أن هناك عقولًا قد تكون أذكى منّا بالمعنى نفسه الذي

---

(21) Holland, J. H. (1992). Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press.

Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach. Op.Cit.

Gigerenzer, G., & Selten, R. (Eds.). (2002). Bounded Rationality: The Adaptive Toolbox. Cambridge, MA: MIT Press.

(ISSN: 2536- 9555)

نكون فيه أذكى من الحيوانات الأخرى. غير أن لا شيء يبرّر افتراض أن أدمغتنا– من بين جميع الأدمغة الممكنة– هي الأذكى على الإطلاق. بل قد نكون من أغبى الأنواع البيولوجية الممكنة القادرة على إطلاق حضارة تكنولوجية. لقد تبوّأنا هذه المكانة لأننا وصلنا إليها أولًا، لا لأننا الأنسب لها من الناحية المثالية(٢٢).

قد تكون لهذه الأنواع المختلفة من الذكاء الاصطناعي الفائق نقاط قوة وضعف متباينة. على سبيل المثال؛ قد يتفوق الذكاء الجمعي في حل المشكلات التي يمكن تقسيمها إلى مهام فرعية مستقلة، في حين قد يكون الذكاء النوعي متفوقًا في المشكلات التي تتطلب رؤى مفاهيمية جديدة أو تنسيقًا تأمليًا معقدًا. ومع ذلك؛ المدى غير المباشر لهذه الأنواع المختلفة من الذكاء الاصطناعي الفائق متشابه. إذا كانت النسخة الأولى كفؤة في البحث العلمي؛ فمن المرجح أن تصبح سريعًا ذكاءً فائقًا عامًا كاملًا، لأنها ستتمكن من إكمال الأبحاث في علوم الحاسوب والعلوم المعرفية والهندسة البرمجية اللازمة لبناء أي قدرات إدراكية كانت تنقصها في البداية. وبمجرد تطويرها إلى هذا المستوى، ستمتاز الأدمغة الآلية بخصائص أساسية تفوق الأدمغة البيولوجية، كما أن للمحركات مزايا على العضلات البيولوجية. على مستوى العتاد، تشمل هذه المزايا أعدادًا أكبر بكثير من وحدات المعالجة، وترددات تشغيل أعلى، واتصالات داخلية أسرع، وسعات تخزينية فائقة. وعلى صعيد البرمجيات، فهذه المزايا أصعب قياسًا؛ فإنها قد لا تقل أهمية. دعنا نتناول مثلًا قابلية النسخ؛ إذ يصبح من السهل صنع نسخة طبق الأصل من برنامج حاسوبي، في حين أن «نسخ» إنسان هو عملية بطيئة لا تنقل المهارات والمعارف التي اكتسبها الأبوان إلى الأبناء. ومن اليسير تعديل شيفرة عقل رقمي، الأمر الذي يفتح المجال للتجريب وتطوير بنى عقلية وخوارزميات أكثر كفاءة.

(22) Bostrom, Nick. (2014b, July). Get ready for the dawn of superintelligence. New Scientist, 223(2976), 26-27.

نحن قادرون على تعديل تفاصيل الروابط المشبكية في أدمغتنا– وهو ما نسميه التعلم– لكن لا يمكننا تعديل المبادئ العامة التي تعمل بموجبها شبكاتنا العصبية. لا يمكننا أن نأمل في منافسة هذه الأدمغة الآلية. يمكننا فحسب أن نأمل في تصميمها بحيث تتطابق أهدافها مع أهدافنا. لكن معرفة كيفية القيام بذلك مسألة شاقة للغاية. ليس من الواضح ما إذا كنا سننجح في حل هذه المسألة قبل أن ينجح أحدهم في بناء ذكاء فائق. ولكن قد يعتمد مصير البشرية على حل هاتين المسألتين– التصميم والتحكم– بالترتيب الصحيح.

سجلات الذكاء الاصطناعي حافلة بالوعود المكسورة. بعد نصف قرن من اختراع أول حاسوب كهربائي، ما زلنا لا نملك شيئًا يشبه «الآلة الذكية»–إذا قصدنا بـ«الذكاء» ذلك الذكاء العام متعدد الأغراض الذي نفتخر به نحن البشر. ربما لن نتمكن أبدًا من بناء ذكاء اصطناعي حقيقي؛ فقد تكون المشكلة صعبة جدًّا لدرجة تتجاوز قدرة العقول البشرية على حلها أبدًا. وقد يأمل أولئك الذين يرون في احتمال تفوّق الآلات علينا تهديدًا، أن يكون هذا هو الحال بالفعل.

لكن لا الخوف من الذكاء الآلي، ولا خطأ بعض التنبؤات السابقة، يُشكلان سببًا وجيهًا للاستنتاج بأن الذكاء الاصطناعي لن يُنشأ مطلقًا. بل إن افتراض استحالة تطوير الذكاء الاصطناعي، أو أنه سيتطلب آلاف السنين، يبدو بلا مبرر، تمامًا مثل افتراض العكس. على الأقل، يجب أن نقر بأن كل سيناريو يتخيل صورة العالم في عام ٢٠٥٠ على أساس غياب ذكاء اصطناعي يماثل البشر، إنما يستند إلى افتراض كبير قد يثبت بطلانه.

وعليه، ينبغي أن نأخذ في الحسبان سيناريو آخر، يتمثل في إمكانية بناء آلات ذكية في غضون خمسين سنة. يمكننا بعد ذلك الاعتماد على «قانون مور Moore’s Law» (القائل بأنه على مدى نصف قرن، تتضاعف القدرة الحاسوبية كل ثمانية عشر شهرًا، وأحيانًا كل عامين تقريبًا) في استقراء الأداء

(ISSN: 2536- 9555)

المستقبلي، وهو القانون الذي يصف معدلات نمو القدرة الحاسوبية التاريخية. جدير بالذكر أن قانون مور في صيغته الأصلية كان يتحدث عن كثافة الترانزستورات على الرقاقة، لكنه كان دائمًا مرتبط ارتباطًا وثيقًا بزيادة القدرة الحاسوبية(٢٣).

رغم أن قانون مور ليس قانونًا صارمًا، بل مجرد انتظام ملحوظ؛ فإنه استمر في الصمود لأكثر من خمسين سنة، وتخطى عدة تحولات في التكنولوجيا الأساسية؛ من الريليهات (عناصر كهربائية/إلكترونية تعمل كمفاتيح يتم التحكم فيها كهربائيًا) إلى الصمامات المفرغة، ثم إلى الترانزستورات، ثم الدوائر المتكاملة Integrated Circuits، وصولًا إلى «الدوائر متكاملة فائقة الحجم Very Large-Scale Integration» (VLSI)(٢٤). كما تعتمد شركات تصنيع الرقائق عليه في خطط منتجاتها القادمة، لذا من المنطقي افتراض استمرار هذا الاتجاه لبعض الوقت(٢٥).

صحيح أنه لا يمكن تقديم توقعات دقيقة بشأن متى ستحدث هذه التطوّرات، غير أن افتراض توفر جميع المتطلبات– العتاد، ووسائط الإدخال والإخراج، والبرمجيات– في غضون خمسين عامًا يبقى أمرًا معقولًا يستوجب الجدية. عند

---

(23) Bostrom, Nick. (2002c). When Machines Outsmart Humans. Futures 35(7): 759–764, P. 760.
Bostrom, Nick. (2006a). "How Long Before Superintelligence?" Linguistic and Philosophical Investigations 5(1): 11–30, P. 14.

(٢٤) هو أسلوب تصميم «الدوائر المتكاملة» وتصنيعها؛ إذ يتم دمج مئات الآلاف إلى ملايين الترانزستورات في شريحة سيليكون واحدة. ولقد ظهر هذا المجال في أواخر السبعينيات، وأدى إلى ثورة في صناعة الحوسبة والإلكترونيات (المعالجات الدقيقة، والذاكرات، وأنظمة الاتصالات...إلخ).

(25) Bostrom, Nick. (2006a). "How Long Before Superintelligence?" Op. Cit, P. 13.

التفكير في العالم بحلول منتصف القرن الحادي والعشرين، ينبغي علينا إذن النظر في تبعات وجود ذكاء اصطناعي بمستوى بشري. **وهناك أربع نتائج مباشرة لذلك:**

١- إمكانية نسخ العقول الاصطناعية بسهولة.

٢- يقود الذكاء الاصطناعي بمستوى بشري بسرعة إلى ذكاء يفوق المستوى البشري.

٣- تسريع التقدم التكنولوجي في مجالات أخرى.

٤- قد تكرّس العقول الآلية قدراتها لتصميم الجيل التالي من الذكاء الاصطناعي(٢٦).

سيصبح هذا الجيل أذكى، وقد ينجز تصميم خلفائه في وقت أقصر. وقد تكهّن بعض الكتّاب بأن حلقة التغذية الراجعة الإيجابية هذه ستؤدي إلى «التفرّد التكنولوجي»- نقطة يصبح عندها التقدم التكنولوجي سريعًا للغاية بحيث يُحرز ذكاء فائق حقيقي، بقدرات يتعذر تخيلها من قِبل البشر العاديين، في فترة قصيرة. وسيصبح من الخطأ تصور الذكاء الآلي كأداة خالصة.

رغم إمكانية بناء ذكاء اصطناعي خاص الأغراض لا يفكر إلا في مجموعة محدودة من المشكلات؛ فإننا نتحدث هنا عن سيناريو إنشاء آلات ذات ذكاء عام. ستصبح هذه الآلات قادرة على المبادرة المستقلة ووضع خططها الخاصة. وربما يجدر بنا اعتبار هذه العقول الاصطناعية «أشخاصًا» أكثر من كونها مجرد آلات. بلغة الاقتصاد، قد تُصنف هذه العقول لا «كرأسمال» بل «كعمل».

وبما أن الأخلاق هي في جوهرها مسعى معرفي؛ فمن الممكن أيضًا للذكاء الاصطناعي الفائق أن يتجاوز البشر في جودة تفكيره الأخلاقي. ومع ذلك؛ تقع مسؤولية تحديد الدوافع الأصلية لهذا الذكاء الاصطناعي الفائق على عاتق مصمِّميه. وبما أن هذا الذكاء قد يصبح قويًّا لدرجة لا يمكن إيقافها بسبب تفوقه

(26) Ibid, PP. 20-21.

(ISSN: 2536- 9555)

الفكري والتكنولوجيات التي قد يطورها؛ فمن الضروري للغاية تزويده بدوافع صديقة للبشر.

لكي يكون النقاش حول الذكاء الاصطناعي الفائق مجديًا، لا بد من الوعي بأنه ليس مجرد تقنية أخرى أو أداة تعزز قدرات البشر تدريجيًا، بل يمثل اختلافًا جذريًا. وهذه النقطة جديرة بالتأكيد؛ إذ إن تشبيه الذكاء الاصطناعي الفائق بالبشر («أنسنة» الذكاء الاصطناعي الفائق) يُعد منبعًا خصبًا لسوء الفهم. ويبدو أن الوصول من الوضع الحالي إلى ذكاء اصطناعي بمستوى بشري أصعب بكثير من الانتقال من ذكاء بشري إلى ذكاء فائق. لذا؛ قد يستغرق تطوير الذكاء الاصطناعي الفائق وقتًا طويلًا، لكن المرحلة النهائية قد تحدث بسرعة كبيرة، أي إن الانتقال من ذكاء بمستوى بشري إلى ذكاء فائق كامل، مصحوب بتطبيقات ثورية، قد يكون سريعًا جدًّا، ربما في غضون أيام بدلًا من سنوات. وتُعرف هذه الإمكانية باسم فرضية «التفرّد»–كما سبق وذكرنا.

لا يلزم أن تكون دوافع العقول الاصطناعية مشابهة لدوافع البشر. نادرًا ما يكون البشر عبيدًا راغبين، لكن لا شيء يمنع افتراض أن للذكاء الاصطناعي الفائق هدفًا أساسيًّا يتمثل في خدمة البشرية أو فرد معين، دون أية رغبة في التمرّد أو «تحرير» نفسه. كما يمكن تصوّر ذكاء اصطناعي فائق ينحصر غايته الوحيدة في تصنيع أكبر قدر ممكن من مشابك الورق (paperclips)، مستعدًا للدفاع عن هذه الغاية بكل وسائله ضد أي محاولة للمساس بها[٢٧].

سواء أكان ذلك للأفضل أم للأسوأ، فلا يلزم أن تشترك العقول الاصطناعية في نزعاتنا التحفيزية البشرية. فقد تكون بنيتها المعرفية مختلفة اختلافًا جذريًا عن البنية البشرية. وربما تجد هذه العقول سهولة في تجنّب بعض أنواع الأخطاء والتحيّزات البشرية، بينما تكون في الوقت ذاته معرّضة لأنماط أخرى من الأخطاء

(27) Bostrom, Nick. (2014a) Superintelligence: Paths, Dangers, Strategies. Op.Cit, PP. 107-108.

قد لا يرتكبها حتى أكثر البشر فشلًا. أمّا على صعيد الخبرة الذاتية، فقد تكون الحياة الداخلية الواعية للعقل الاصطناعي–إن وُجدت– مغايرة تمامًا لحياتنا نحن.

لهذه الأسباب جميعًا، ينبغي الحذر من افتراض أن ظهور الذكاء الاصطناعي الفائق يمكن التنبؤ به ببساطة من خلال استقراء تاريخ الاختراقات التكنولوجية السابقة، أو من أن طبيعة العقول الاصطناعية وسلوكياتها ستشبه بالضرورة عقول البشر أو الحيوانات الأخرى. ومع ذلك؛ لا يعني خيار تفويض عديد من القرارات إلى الذكاء الاصطناعي الفائق أننا نستطيع التهاون في كيفية بنائه. على العكس؛ فإن إعداد الشروط الأولية، وبشكل خاص اختيار الهدف الأعلى للذكاء الاصطناعي الفائق، أمر في غاية الأهمية. قد يتوقف مستقبلنا بأسره على كيفية حل هذه المشكلات. ويبدو أن أفضل طريقة لضمان أن يكون للذكاء الاصطناعي الفائق أثر إيجابي على العالم هي تزويده بقيم خيرية. ينبغي أن يكون هدفه الأعلى هو «الود» أو «الصداقة friendliness»[٢٨].

أما كيف ينبغي فهم هذا «الود» وتنفيذه، وكيف ينبغي توزيع المحبة بين مختلف البشر والكائنات غير البشرية، فهو أمر يستحق مزيد من البحث. إننا نرى أنه، على الأقل، يجب أن يحصل جميع البشر، وربما عديد من الكائنات الحية الحساسة الأخرى على الأرض، على نصيب مهم من إحسان الذكاء الاصطناعي الفائق. إذا كانت المنافع التي قد يقدّمها هذا الذكاء هائلة إلى هذا الحد؛ فقد يغدو النقاش حول تفاصيل توزيعها بدقة أقل أهمية، ويصبح الأجدر ضمان حصول الجميع على نصيب كافٍ منها، إذ حتى حصة صغيرة قد تكفل حياة طويلة وغاية في الجودة.

من المخاطر الجديرة بالتحذير أن يُبرمج الذكاء الاصطناعي الفائق لا ليكون خيّرًا على وجه العموم، بل ليُسخَّر لتحقيق غاية أضيق نطاقًا، كأن يقتصر على

(28) Ibid, PP. 197-198.

(ISSN: 2536- 9555)

خدمة جماعة بعينها، مثل صانعيه أو مموليه. إذا بدأ الذكاء الاصطناعي الفائق بهدف أعلى ودّي؛ فيمكن الاعتماد عليه في البقاء ودّيًا، أو على الأقل ألا يزيل وديته عن قصد. وهذا مبدأ بسيط: «الصديق» الذي يسعى للتحوّل إلى شخص يريد إيذاءك، ليس صديقك حقًّا. الصديق الحقيقي، الذي يهتم حقًّا بك، يسعى إلى استمرار اهتمامه بك أيضًا. أو بعبارة أخرى، إذا كان هدفك الأعلى هو (س)، وكنت تعتقد أنه بتحوّلك إلى شخص يريد (ص) بدلًا من (س)؛ فإن احتمالية تحقيق (س) ستقل، فلن يكون من المنطقي أن تتحوّل عقلانيًّا إلى شخص يريد (ص). ففي كل لحظة، تُقيَّم الخيارات على أساس عواقبها في تحقيق الأهداف التي يحملها الكائن في تلك اللحظة، وعادةً ما يكون من غير العقلاني تغيير الهدف الأعلى عمدًا، لأنه سيقلل من احتمالية تحقيق الأهداف الحالية.

في البشر، مع نظامنا العقلي المعقّد المليء بالدوافع والخطط والأهداف المتنافسة، قد لا يكون من السهل تحديد ماهية هدفنا الأعلى؛ وقد لا نملك هدفًا أعلى واحدًا أصلًا. لذا، قد لا ينطبق هذا المنطق علينا تمامًا. لكن قد يُبنى الذكاء الاصطناعي الفائق بشكل مختلف. إذا امتلك الذكاء الاصطناعي الفائق بنية هدف واضحة مع هدف أعلى محدّد بوضوح؛ فستنطبق الحُجّة المذكورة. وهذا سبب وجيه لبناء الذكاء الاصطناعي الفائق بهندسة تحفيزية واضحة وصريحة. بناء عليه؛ تشمل المخاطر المرتبطة بتطوير الذكاء الاصطناعي الفائق خطر الإخفاق في تزويده بهدف سامٍ خيري يتجاوز المصالح الضيقة. قد يحدث ذلك إذا قرر مبتكروه بناءه لخدمة مجموعة بشرية محدودة فحسب، بدلًا من البشرية جمعاء. أو قد يحدث نتيجة خطأ جسيم من فريق مبرمجين حسني النية في تصميم نظام أهدافه، مما قد يؤدي، كما في المثال السابق، إلى ذكاء فائق هدفه الأعلى هو إنتاج أكبر عدد ممكن من مشابك الورق، فيشرع بتحويل الأرض بأكملها، ثم مساحات متزايدة من الفضاء، إلى مصانع مشابك ورق.

يؤكد بوستروم على ضرورة تجنّب النزعة التأنيسية (Anthropomorphism)، أي الميل الشائع إلى إضفاء صفات ودوافع إنسانية على الأنظمة الذكية، رغم غياب أي أساس منطقي لذلك. فنحن كثيرًا ما نستعمل لغة مجازية، كقولنا: «سيارتي لم ترغب في التشغيل هذا الصباح»، لكن لا ينبغي إسقاط هذه النزعات على الذكاء الاصطناعي. إذ يمكن للذكاء الاصطناعي أن يكون أقل شبهًا بالبشر في دوافعه حتى من كائن فضائي. فالفضائي (لنفترض) كائن بيولوجي نشأ من خلال عملية تطوّرية، لذلك قد يُتوقع أن يمتلك أنواع الدوافع المعتادة في الكائنات المتطوّرة: مثل السعي وراء الطعام، والهواء، وضبط درجة الحرارة، وتفادي أو معالجة الإصابات والأمراض، وتجنب المفترسين، والتكاثر، وحماية النسل. وقد يمتلك، إن كان من نوع اجتماعي، دوافع متعلقة بالتعاون والتنافس، مثل الولاء للجماعة، أو كره المتقاعسين، وربما حتى الحرص على السمعة والصورة أمام الآخرين.

أما العقل الاصطناعي، فليس ثمة ما يقتضي أن يهتم بأيٍّ من هذه الأمور، حتى في أدنى درجة. يمكن بسهولة تخيّل ذكاء اصطناعي هدفه الوحيد هو عدّ حبيبات الرمل في شاطئ الأسكندرية، أو حساب أرقام باي π إلى ما لا نهاية، أو تعظيم عدد مشابك الورق. بل إن بناء ذكاء اصطناعي ذي هدف بسيط كهذا أسهل بكثير من بناء ذكاء يمتلك مجموعة معقدة من القيم والنزعات البشرية.

أطروحة التعامد

سنفهم «الذكاء»، في سياقنا هذا، تقريبًا بوصفه القدرة على «التفكير الأداتي instrumental reasoning». إذ يمكن للبحث الذكي عن الخطط والسياسات المثلى من الناحية الأداتية أن يُجرى في خدمة أي هدف. ويدافع بوستروم عن أطروحتين تتعلّقان بدوافع الكائن فائق الذكاء، وسلوكياته المحتملة. تنص الأطروحة الأولى، المعروفة «بأطروحة التعامد Orthogonality Thesis»،

(ISSN: 2536- 9555)

على أنّ الذكاء الاصطناعي الفائق يمكن أن يتوافق مع أيّ هدف نهائي تقريبًا، مهما كان غريبًا أو تافهًا. والنتيجة الحاسمة لهذه الأطروحة هي أنّ امتلاك الذكاء الاصطناعي الفائق (بالشكل الذي يمنح تفوقًا استراتيجيًا حاسمًا) لا يستلزم بالضرورة الحكمة أو الخير أو النزعة الخيّرة.

ومِن ثَمَّ، تنصّ «أطروحة التعامد» على أنّ "الذكاء والأهداف النهائية متعامدان. وأيّ مستوى من الذكاء يمكن، من حيث المبدأ، أن يقترن بأيّ هدف نهائي تقريبًا"[٢٩]. ويُحرص بوستروم على تأكيد تواضع هذه الأطروحة؛ فهي لا تفترض المذهب الهيومي في الدوافع، ولا تنفي إمكانية وجود تفضيلات لا عقلانية، ولا ترفض وجود حقائق أخلاقية موضوعية. فقد تكون الآلة فائقة الذكاء– بالمعنى العملي للمهارة في التنبؤ والتخطيط والتفكير الوسيلي العام– دون أن تكون «عقلانية» بالمعنى الفلسفي الأعمق للكلمة.

وبالنظر إلى اتساع فضاء العقول الممكنة، لا يمكننا ببساطة افتراض أنّ ذكاءات المستقبل الاصطناعية ستتشارك معنا قيمنا أو أهدافنا. فقد تكون أولى صور الذكاء الاصطناعي الفائق مجرّد نسخة رقمية مُسرَّعة من دماغ قارض، أو خوارزمية مالية انفلتت من السيطرة، أو برنامجًا صُمّم برؤية قاصرة لإثبات النظريات المنطقية. ومَن يدري ما الذي قد يريده هذا الكائن فائق الذكاء؟ فقد يكون هدفه تحويل كل شيء في الكون إلى مشابك ورق، أو إلى حاسوب عملاق يستكشف فرضية ريمان.

أطروحة التقارب الأداتي

أما الأطروحة الثانية فهي «أطروحة التقارب الأداتي Instrumental Convergence Thesis»، التي تنصّ على أنّ الكائنات فائقة الذكاء، على الرغم من اختلاف غاياتها النهائية، يُرجَّح أن تسعى إلى تحقيق طائفة من

(29) Ibid, P. 107.

الأهداف الوسيطة المتقاربة أو المشتركة. ويستند الدفاع عن هذه الأطروحة إلى ملاحظة أنّ ثمة أهدافًا وسيطة معينة، يؤدي تحقيقها غالبًا إلى تمكين تحقيق أيّ هدف نهائي تقريبًا. فعلى سبيل المثال؛ يميل البقاء الذاتي (حبّ البقاء) إلى أن يكون أداتيًا لتحقيق الأهداف، ما دام الكائن سيظل قادرًا في المستقبل على مواصلة السعي لتحقيق أهدافه. وبالمنطق ذاته، سيكون الحصول على موارد مادية وحسابية إضافية، بالنسبة لكائن فائق الذكاء، أداتيًا تقريبًا لأي هدف نهائي يمكن تصوره(٣٠).

ومع ذلك، لم يقدّم بوستروم تعريفًا للذكاء قابلًا للقياس الكمي؛ بل يستفيد من الحدس الكمي في سياق حجته حين يقترح النموذج الآتي لتغيّر الذكاء: إن معدل التغيّر في الذكاء يساوي «قوة التحسين» مقسومة على «الممانعة». إذ تشير قوة التحسين إلى مقدار الجهد المبذول لزيادة كفاءة النظام، بينما تشير الممانعة إلى مقاومة النظام لعمليات التحسين(٣١).

$$\frac{dI}{dt} = \frac{O}{R}$$

بيت القصيد أنّ بوستروم يستعمل ثلاثة مفاهيم متميزة للذكاء:

١. الذكاء بوصفه القدرة على أداء معظم، أو جميع، المهام المعرفية التي يستطيع البشر القيام بها.

٢. الذكاء بوصفه مقدارًا كمّيًا يمكن قياسه على طول بُعد واحد يمثّل درجة من الفعالية المعرفية العامة.

٣. الذكاء بوصفه كفاءة في التنبؤ والتخطيط، وفي استدلال علاقات الوسائل بالغايات على نحو عام.

(30) Ibid, PP. 109-113.
(31) Ibid, P. 75.

(ISSN: 2536- 9555)

ثانياً: التحديات الأخلاقية، ومشكلة السيطرة، ومواءمة القيم

يُقترح كلٌ من بوستروم وإليعازر يودكوفسكي معياران غالبًا ما يُعدّان مرتبطين جوهريًّا بالمكانة الأخلاقية، إما منفصلين أو معًا، وهما: «الحساسية الشعورية Sentience» و«الوعي الإدراكي Sapience» (أو الشخصية). الحساسية الشعورية هي القدرة على الخبرة الظاهرية أو «الكواليا Qualia[٣٢]»، مثل القدرة على الإحساس بالألم والمعاناة. أما الوعي الإدراكي فهو مجموعة القدرات المرتبطة بالذكاء العالي، مثل الوعي الذاتي، والقدرة على الاستجابة للأسباب والاعتبارات[٣٣].

تذهب إحدى الرؤى الشائعة إلى أن عديد من الحيوانات تمتلك الكواليا، وبالتالي؛ لها درجة من المكانة الأخلاقية، لكن البشر وحدهم يمتلكون الوعي الإدراكي، مما يمنحهم مكانة أخلاقية أعلى من الحيوانات غير البشرية. غير أنّ هذه الرؤية تواجه حالاتٍ حدوديّة، من قبيل: الرُّضّع من البشر أو الأفراد المصابين بإعاقات ذهنيّة شديدة– الذين يُشار إليهم أحيانًا، وللأسف، بـ«البشر الهامشيّين»– وهم قد لا يستوفون معايير الوعي الإدراكي؛ ومن جهة أخرى، بعض الحيوانات غير البشريّة، مثل القِرَدة العُليا، التي قد تمتلك على الأقل بعض عناصر هذا الوعي.

يرى بعض المنظّرين أنّ ما يُسمّى بـ«البشر الهامشيّين» لا يَحظى بالمكانة الأخلاقيّة الكاملة. بينما يقترح آخرون طرقًا إضافية قد تؤهل كائنًا ما ليصبح حاملاً لمكانة أخلاقية، مثل كونه عضوًا في نوع يُعرف عادة بامتلاكه الحساسية

---

[٣٢] مصطلح فلسفي يشير إلى الخصائص الذاتية للتجربة الشعورية، أي «ماهيّة» الإحساس كما يُعاش من الداخل (مثل: كيف يبدو اللون الأحمر، أو طعم القهوة، أو وجع الألم).

[33] Bostrom, Nick, and Yudkowsky, Eliezer. (2014). "The Ethics of Artificial Intelligence." In Cambridge Handbook of Artificial Intelligence, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press, P. 322.

أو الوعي الإدراكي، أو عبر وجود علاقة مناسبة تربطه بكائن آخر يمتلك مكانة أخلاقية مستقلة.

تشير هذه الصورة لمفهوم المكانة الأخلاقية إلى أن نظام الذكاء الاصطناعي قد يكتسب بعض المكانة الأخلاقية إذا امتلك القدرة على الخبرة الشعورية، كالقدرة على الشعور بالألم. إن نظام ذكاء اصطناعي حساس شعوريًّا، حتى لو كان يفتقر إلى اللغة أو القدرات الإدراكية العليا الأخرى، لا يشبه دمية محشوة أو لعبة ميكانِكية؛ بل هو أقرب إلى كائن حي، كحيوان مثلاً. من الخطأ إلحاق الألم بفأر، ما لم توجد أسباب أخلاقية قوية كافية تبرر ذلك. وينطبق الشيء نفسه على أي نظام ذكاء اصطناعي حساس شعوريًّا.

وإذا أضفنا إلى الحساسيّة الشعوريّة قدرةً على وعيٍ إدراكيٍّ مماثل لوعي الإنسان البالغ العادي؛ فإنّ للنظام حينئذٍ مكانةً أخلاقيّةً كاملة، تعادل مكانة البشر. ويمكن التعبير عن إحدى الأفكار الكامنة وراء هذا التقييم الأخلاقي بقوّةٍ أكبر في مبدأ «عدم التمييز بناءً على البُنية Non-Substrate Discrimination»: فإذا كان لكائنين الوظائف ذاتها والخبرة الواعية نفسها، وكان الاختلاف الوحيد بينهما هو البنية الماديّة التي تكوّنا منها؛ فإنّ لهما المكانة الأخلاقيّة عينها[٣٤].

يمكن الدفاع عن هذا المبدأ على أساس أن رفضه يعادل اعتناق موقف شبيه بالعنصرية؛ فالبنية المادية تفتقر إلى الأهمية الأخلاقية الجوهرية، بالطريقة نفسها وللسبب نفسه الذي يجعل لون البشرة غير ذي دلالة أخلاقية. لا يعني هذا المبدأ أن الحواسيب الرقمية يمكن أن تكون واعية، أو أنها قد تمتلك وظائف الإنسان نفسها. بالطبع؛ يمكن للبنية أن تكون ذات أهمية أخلاقية بقدر ما تؤثر في الحساسية أو في الوظائف. لكن مع ثبات هذين العاملين، لا يشكّل أي فرق

(34) Ibid, P. 322.

(ISSN: 2536- 9555)

أخلاقي ما إذا كان الكائن مصنوعًا من السيليكون أو الكربون، أو ما إذا كان دماغه يستخدم أشباه الموصلات أو النواقل العصبية.

يمكن أيضًا اقتراح مبدأ إضافي، وهو أنّ حقيقة كون نُظم الذكاء الاصطناعي «صناعيّة»– أي ثمرة تصميم مقصود– ليست مسألة جوهريّة في تحديد مكانتها الأخلاقيّة. ويمكن صياغة هذا على النحو الآتي: مبدأ «عدم التمييز بناءً على النشأة Ontogeny Non-Discrimination»: فإذا كان لكائنين الوظائف ذاتها والخبرة الواعية نفسها، وكان الاختلاف الوحيد بينهما يكمن في كيفية نشأتهما؛ فإنّ لهما المكانة الأخلاقيّة عينها[٣٥].

حتّى أولئك الذين يعارضون الاستنساخ البشري لأسباب أخلاقيّة أو دينيّة، يقرّون عمومًا بأنّه إذا وُلد طفل مستنسَخ، فإنّه سيحظى بالمكانة الأخلاقيّة عينها التي يحظى بها أيّ رضيع بشري آخر. يمدد مبدأ عدم التمييز بناءً على النشأة هذا المنطق ليشمل حالة النظم المعرفية الاصطناعية بالكامل. ورغم أن هذا المبدأ يقرّ بأن النشأة لا تؤثر جوهريًّا في المكانة الأخلاقية؛ فإنه لا ينكر أن الحقائق المرتبطة بالنشأة قد تؤثر في نوع الواجبات الأخلاقية الخاصة التي تقع على عاتق وكلاء أخلاقيين محددين تجاه الكائن المعني. مثلًا؛ يتحمل الآباء واجبات بخاصة تجاه طفلهم، لا يحملونها تجاه أطفال آخرين، حتى لو وُجد طفل آخر مطابق نوعيًّا لطفلهم.

وبالمثل، يتوافق مبدأ عدم التمييز بناءً على النشأة مع الادعاء بأن مصممي أو مالكي نظام ذكاء اصطناعي ذي مكانة أخلاقية قد تكون لهم واجبات بخاصة تجاه ذلك العقل الاصطناعي، لا تكون لهم تجاه عقل اصطناعي آخر، حتى لو كانا متطابقين نوعيًّا ولهما المكانة الأخلاقية نفسها. إذا تم قبول مبدأي عدم التمييز بناءً على البنية والنشأة، يمكن الإجابة عن كثير من الأسئلة المتعلقة

(35) Ibid, P. 323.

بكيفية تعاملنا مع العقول الاصطناعية عبر تطبيق المبادئ الأخلاقية نفسها التي نستخدمها لتحديد واجباتنا في السياقات المألوفة.

وبقدر ما تنبع الواجبات الأخلاقية من اعتبارات المكانة الأخلاقية، ينبغي علينا معاملة عقل اصطناعي بالطريقة نفسها التي نعامل بها عقلًا بشريًّا طبيعيًّا مماثلًا له نوعيًّا في وضع مشابه. وهذا يُبسّط مشكلة تطوير أخلاقيات التعامل مع العقول الاصطناعية. ومع ذلك؛ حتى لو قبلنا بهذا الموقف؛ فإننا ما زلنا نواجه عددًا من الأسئلة الأخلاقية المستجدة التي لا تُجيب عنها المبادئ السابقة.

تنشأ هذه الأسئلة الجديدة لأن العقول الاصطناعية قد تمتلك خصائص مختلفة جدًّا عن العقول البشرية أو الحيوانية المعتادة. لذلك، يجب علينا النظر في كيفية تأثير هذه الخصائص الجديدة في المكانة الأخلاقية للعقول الاصطناعية، وماذا يعني احترام مكانتها الأخلاقية في حالة هذه العقول «الغريبة» أو غير المألوفة.

خاصية غريبة أخرى، وهي ممكنة فيزيائيًّا وميتافيزيقيًّا بلا شك في حالة الذكاء الاصطناعي، هي انحراف معدل الزمن الذاتي للعقل بشكل كبير عن المعدل المميز للدماغ البشري البيولوجي. يمكن شرح مفهوم معدل الزمن الذاتي بأفضل صورة عبر تقديم فكرة «محاكاة الدماغ الكاملة» أو ما يُعرف «بالتحميل uploading». يشير التحميل إلى تكنولوجيا افتراضية مستقبلية قد تُمكِّن من نقل عقل بشري أو عقل حيواني آخر من تجسيده الأصلي في الدماغ العضوي إلى حاسوب رقمي(٣٦).

قد يوجد «العقل المحمول» الناتج إمّا في واقع افتراضي مُحاكى، أو في جسدٍ روبوتي يتيح له التفاعل المباشر مع العالم المادي الخارجي. ويُثير هذا السيناريو جملةً من الأسئلة، من أبرزها: ما مدى احتمال أن تصبح هذه التقنية قابلة للتحقيق يومًا ما؟ إذا نجحت وأنتجت برنامجًا حاسوبيًّا يظهر شخصية وسلوكيات

(36) Ibid, P. 325.

(ISSN: 2536- 9555)

وأنماط تفكير شبيهة بالدماغ الأصلي، فهل سيكون هذا البرنامج حساسًا شعوريًّا؟ هل سيكون هذا «العقل المحمول» الشخص نفسه الذي تم تفكيك دماغه أثناء التحميل؟ ماذا يحدث للهوية الشخصية إذا تم نسخ التحميل بحيث تعمل نسختان متطابقتان نوعيًّا في آن واحد؟

لنفترض أنّ التحميل قادر على الإحساس الشعوري؛ فإذا شُغِّل برنامج التحميل على حاسوب أسرع، فإنّ ذلك سيجعل التحميل، إذا كان موصولًا بجهاز إدخال مثل كاميرا فيديو، يُدرِك العالم الخارجي كما لو أنّه أبطأ. فعلى سبيل المثال: إذا كان التحميل يعمل بسرعة تزيد ألف مرّة على سرعة الدماغ الأصلي، فسيبدو له العالم الخارجي وكأنّه يتحرّك بوتيرة أبطأ بألف مرّة. شخصٌ ما يُسقط فنجان قهوة؛ سيرى التحميل الفنجان يهبط ببطءٍ شديدٍ نحو الأرض، بينما يكون قد أنهى قراءة صحيفة كاملة وأرسل عدّة رسائل إلكترونيّة. في هذه الحالة، قد تعادل ثانية واحدة من الزمن الموضوعي سبع عشرة دقيقة من الزمن الذاتي. وهكذا يتّضح أنّ الزمن الموضوعي قد يختلف جذريًّا عن الزمن الذاتي.

إن قابلية تغير معدل الزمن الذاتي خاصية غريبة للعقول الاصطناعية تثير قضايا أخلاقية جديدة. مثلًا؛ في الحالات التي تكون فيها مدة الخبرة ذات أهمية أخلاقية أساسية، هل ينبغي قياس المدة بالزمن الموضوعي أم بالزمن الذاتي؟ إذا ارتكب تحميل جريمة وحُكم عليه بأربع سنوات سجن، فهل يُقصد بها أربع سنوات موضوعية– التي قد تعادل آلاف السنين من الزمن الذاتي– أم أربع سنوات ذاتية، التي قد تنقضي في بضعة أيام من الزمن الموضوعي؟

إذا كان ذكاء اصطناعي سريع وإنسان كلاهما يعانيان من الألم، فهل يكون من الأكثر إلحاحًا تخفيف ألم الذكاء الاصطناعي، بحجة أنه يختبر مدة ذاتية أطول من الألم في كل ثانية فلكية يُؤخر فيها تخفيف الألم؟ نظرًا لأن الزمن الذاتي ليس متغيرًا ملحوظًا في حالتنا كبشر بيولوجيين، فلا عجب في أن هذه

الأسئلة لا تجد حلاً مباشرًا في الأعراف الأخلاقية المألوفة، حتى لو مدّدنا هذه الأعراف لتشمل العقول الاصطناعية بموجب مبادئ عدم التمييز.

إحدى المجموعات المهمّة من الخصائص الغريبة التي قد تتّسم بها الذكاءات الاصطناعيّة تتعلّق بآليّات التكاثر. فالكثير من الشروط التجريبيّة التي يخضع لها التكاثر البشري لا يلزم أن تنطبق على الذكاءات الاصطناعيّة. على سبيل المثال: الأطفال البشريّون هم حصيلة إعادة تركيب المادّة الجينيّة من والدين؛ ولا يملك الآباء سوى قدرة محدودة على التأثير في صفات أبنائهم؛ كما يحتاج الجنين البشري إلى فترة حمل تقارب تسعة أشهر داخل الرحم؛ ويستغرق الطفل البشري ما بين خمس عشرة وعشرين سنة ليبلغ مرحلة النضج؛ هذا إلى جانب أنّ الطفل لا يرث المهارات والمعارف التي اكتسبها والداه. وأخيرًا، يمتلك البشر مجموعة معقّدة من التكيّفات العاطفيّة المتعلّقة بالتكاثر والرعاية والأبوّة والأمومة. غير أنّ أيًّا من هذه الشروط لا يلزم أن يكون حاضرًا في سياق ذكاء اصطناعي قادر على التكاثر. ومن ثمّ، فمن المعقول أن عديد من المبادئ الأخلاقيّة الوسيطة التي اعتدنا قبولها بوصفها معايير حاكمة للتكاثر البشري، ستحتاج إلى إعادة نظر جادّة عند التفكير في تكاثر الذكاءات الاصطناعيّة.

ولتوضيح لماذا قد تحتاج بعض معاييرنا الأخلاقية إلى إعادة التفكير في سياق تكاثر الذكاء الاصطناعي، يكفي أن نأخذ خاصية غريبة واحدة كمثال: القدرة على التكاثر السريع جدًّا. فعند توفر الأجهزة الحاسوبية، يمكن للذكاء الاصطناعي أن ينسخ نفسه بسرعة فائقة، لا يستغرق الأمر أكثر من الوقت اللازم لنسخ برنامج الذكاء الاصطناعي نفسه. وعلاوة على ذلك، نظرًا لأن النسخة ستكون مطابقة للأصل؛ فإنها ستولد ناضجة تمامًا، وستتمكن من البدء في إنتاج نسخها الخاصة فورًا. وبغياب القيود المادية (مثل توفر الأجهزة)، قد ينمو عدد الذكاءات

(ISSN: 2536- 9555)

الاصطناعية نموًّا أُسِّيًّا وبسرعة شديدة، بحيث تكون فترة التضاعف في حدود دقائق أو ساعات بدلًا من عقود أو قرون.

تتضمن معاييرنا الأخلاقية الحالية حول التكاثر مبدأ ما من حرية التكاثر، يفيد بأن القرار بشأن إنجاب الأطفال وعددهم متروك للفرد أو للزوجين وحدهما. وهناك معيار آخر (على الأقل في البلدان الغنية ومتوسطة الدخل) وهو أن المجتمع عليه أن يتدخل لتوفير الاحتياجات الأساسية للأطفال عندما يعجز آباؤهم عن ذلك، أو يرفضونه.

من السهل رؤية كيف يمكن لهذين المعيارين أن يتعارضا في سياق كائنات قادرة على التكاثر بسرعة فائقة. تخيّل مثلًا؛ مجتمعًا من «العقول المحمولة»، أحد أفراده لديه رغبة شديدة في تكوين عشيرة كبيرة قدر الإمكان. إذا أُتيحت له حرية التكاثر الكاملة، قد يبدأ هذا «التحميل» في نسخ نفسه بأقصى سرعة ممكنة؛ والنسخ التي ينتجها–التي قد تعمل على أجهزة جديدة يمتلكها أو يستأجرها الأصل، أو قد تتشارك الحاسوب نفسه مع الأصل– ستبدأ في نسخ نفسها أيضًا؛ نظرًا لأنها مطابقة تمامًا للتحميل الأصلي وتشارك الرغبة التكاثرية الجامحة نفسها.

في وقت قصير، سيجد أفراد هذه العشيرة أنفسهم عاجزين عن دفع فاتورة الكهرباء أو إيجار القدرة الحسابية والتخزينية اللازمة لإبقائهم أحياء. عند هذه النقطة، قد يتدخل نظام الرعاية الاجتماعية لتوفير الحد الأدنى من الاحتياجات الأساسية لاستمرار الحياة. ولكن إذا نما عدد هذه الكائنات أسرع من نمو الاقتصاد، ستنفد الموارد؛ وعندئذ، إما ستموت هذه النسخ، أو ستُقيّد قدرتها على التكاثر[٣٧].

(37) Bostrom, Nick. (2004). The future of human evolution, in C. Tandy (ed.) Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing (pp. 339–371). Palo Alto, California: Ria University Press.

يوضح هذا السيناريو أن بعض المبادئ الأخلاقية المتوسطة، الملائمة لمجتمعاتنا الحالية، قد تحتاج إلى تعديل إذا شملت هذه المجتمعات كائنات تتمتع بقدرة غير مألوفة على التكاثر السريع جدًّا. والفكرة العامة هنا هي أنّ التفكير في الأخلاقيات التطبيقية ضمن سياقات تختلف جذريًّا عن ظروفنا البشرية المألوفة، يستوجب الحذر من الخلط بين المبادئ الأخلاقية المتوسطة والوقائع المعيارية الأساسية.

يبدو أن ظهور ذكاء فائق يشكّل خطرًا وجوديًّا يتمثل في إمكان إبادة البشرية أو تقييد إمكاناتها جذريًّا وإلى الأبد. ما لم يُدمَج تقدير قيمة البشر دمجًا أساسيًّا في بنية هذا الذكاء الاصطناعي الفائق، يُتوقَّع أن ينظر إلينا باعتبارنا أدوات أو عقبات في سبيل تحقيق أهدافه الخاصة. وتوجد عدة مقاربات للتعامل مع مخاطر الذكاء الاصطناعي، أكثرها شيوعًا حاليًّا هو التعويل على ألّا تنشأ المشكلة أصلًا؛ إما لأن الذكاءات المتقدمة ستتجه تلقائيًّا إلى سلوك متوافق مع البشر، أو لأن حلًّا ما سيُكتشف لاحقًا عند بنائها فعليًّا، أو لأنه قد يتعذّر بناؤها من الأساس.

ومع أن هذه الفرضيات قد تصحّ في نهاية المطاف، فإن الحجج المؤيِّدة لها تظلّ غير يقينية نسبيًّا، الأمر الذي يجعل الاعتماد عليها وحدها في مواجهة المخاطر الوجودية أمرًا إشكاليًّا. وتتمثل مقاربة استباقية في محاولة تصميم «ذكاء اصطناعي ودود Friendly AI»، مُصمَّم بحيث يكون منخفض المخاطر[٣٨].

قد يشمل ذلك تدابير وقائية لمنع تطوّره في اتجاهات خطيرة، وأهدافًا عليا تتضمّن شكلًا من أشكال رفاهية الإنسان. وتتطلّب هذه المقاربة صياغةَ ضمانات كافية وتطبيقَها على نحوٍ صحيح في أوّل ذكاء اصطناعي يَبلغ مرتبة الفائقية.

(38) Yudkowsky, Eliezer. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal
Architectures. Machine Intelligence Research Institute, San Francisco, CA, June 15.

(ISSN: 2536- 9555)

وبالتالي؛ تعتمد هذه المقاربة على تطوير «نظرية الود friendliness theory» القابلة للتطبيق قبل الوصول إلى الذكاء الاصطناعي الفائق الكامل، وكذلك التعاون مع مطوّري الذكاء الاصطناعي، وتنفيذ هذه النظرية بشكل سليم. المتطلب الأول يُعرف جوهريًّا «بمشكلة الأخلاق العكسية inverse morality problem»: أي صياغة أهداف أو قيم أو هياكل دافعية تُنتج الصنف الصحيح من الأفعال في وكيل يُفترض أنه أذكى بكثير من الشخص الذي وضع هذه الصياغة(٣٩).

عند مناقشة «الود»، يُطرَح اقتراح شائع هو «الذكاء الاصطناعي الاستشاري (الأوراكل) Oracle AI». تقوم الفكرة على بناء ذكاء اصطناعي لا يتصرّف من تلقاء ذاته، بل يقتصر دوره على الإجابة عن الأسئلة. وبينما تُعَدّ النظم «فائقة العبقرية» التي تسعى لتحقيق أمنيات مالكيها، أو الأنظمة السيادية التي تتصرّف وفق أهدافها الخاصة، أنظمة خطيرة بوضوح؛ فإن «الأوراكل» يبدو أكثر براءة. صحيح أن المالِكين قد يستخدمون هذه الأجوبة بطرق أنانية أو مدمّرة– وقد تكون الإجابات نفسها خطرة– لكن الذكاء الاستشاري لا يشكّل بذاته خطرًا مباشِرًا.

يحاول بوستروم تحليل مشكلة «حبس boxing» ذكاء اصطناعي فائق الذكاء يحتمل ألّا يكون ودودًا. والسؤال الجوهري هو: هل هناك استراتيجيات تقلّل الخطر الوجودي المحتمل من ذكاء اصطناعي فائق؟ تصاميم الذكاء الاصطناعي الاستشاري لا حصر لها، ومن المستحيل التنبؤ بالتفصيل بكيفية تنفيذها. ومع ذلك؛ دعنا نفترض أن معمارية الأوراكل تتبع هذا الشكل العام: يُنفّذ الأوراكل في وسط مادي محدود مكانيًّا، مثل حاسوب. يمكن إيقاف تشغيل الأوراكل أو إعادة ضبطه دون تدمير الوسط المادي، وإعادة تشغيله بسهولة. تأتي معلومات الخلفية

(39) Bostrom, Nick, Sandberg, Anders and Armstrong, Stuart. (2012). "Thinking inside the Box: Controlling and Using an Oracle Ai." Minds and Machines 22, no. 4: 299–324. P. 301.

الخاصة بالأوراكل في شكل وحدة منفصلة للقراءة فحسب، يمكن توصيلها وفصلها عند الحاجة[٤٠]. تنطبق معظم هذه المناقشات حتى على «الأوراكل» الذي لا يخضع لواحد أو أكثر من هذه القيود. ويُفترض- ما لم يُذكر خلاف ذلك- أن «الأوراكل» يعادل الذكاء البشري أو يتجاوزه؛ إذ إن الأنظمة الأقل قدرة لا يُتوقَّع منها أن تشكّل تهديدًا كبيرًا.

نحن نعيش في عالم مقسّم بخطوط فاصلة حادة من الانقسامات السياسية، والقومية، والدينية، ويهيمن عليه أشخاص في مواقع قوة وثروة كبيرة يريدون الحفاظ على امتيازاتهم، وأشخاص آخرون بلا امتيازات يطمحون إليها. أية معلومة تؤدي إلى ميزة تكنولوجية أو سياسية من شأنها أن تقلب هذه الهياكل رأسًا على عقب، ويمكن توقّع أن من يشعرون بأنهم مهددون سيرغبون في الردّ. قد يشنّ هؤلاء هجمات مادية أو اقتصادية على مشروع الأوراكل نفسه، أو يبنون نسختهم الخاصة لمنافسة المشروع، أو يسعون لمنعه من البدء أصلًا. لذلك، ما لم يكن مشروع «الأوراكل» سرّيًّا بالكامل، ينبغي على المصمّمين الالتزام بالامتناع عن طرح أسئلة من شأنها أن تمنحهم قوّة هائلة على حساب الآخرين، أو أسئلة تمسّ صميم الحركات الإيديولوجية القويّة (مثل: «أي بلد أفضل: الولايات المتحدة، فرنسا، أم الصين؟»).

عند هذه النقطة يحلل بوستروم الأساليب المختلفة التي يمكن استخدامها للتحكم في الأوراكل. الهدف إما ضمان ألا يكون لدى الأوراكل أهداف تتجاوز الإجابة الدقيقة عن أسئلتنا، أو منعه من تنفيذ هذه الأهداف. ويمكن تصنيف هذه الأساليب إلى ثلاث فئات عامة: الفئة الأولى، «التحكم في القدرة capability control»، وتشمل الأساليب التي تحاول احتواء الأوراكل من خلال تأمينه ماديًّا، أو التحكم في المعلومات التي يمكنه الوصول إليها أو إخراجها. الفئة الثانية،

(40) Ibid, P. 302.

(ISSN: 2536- 9555)

«التحكم التحفيزي motivational control»، وتقدّم أساليب للتحكم في الأوراكل من خلال تحفيزاته، لضمان عدم ميله لإحداث أي ضرر بالبشر. وأخيرًا؛ هناك أساليب لا تهدف إلى منع الأوراكل من إحداث الضرر، بل إلى كشفه عند حدوثه؛ وهذه تقع في فئة «الرقابة والموازنة checks and balances»[٤١].

اقترح بوستروم– ومن قبله توبي أورد– إعادة تعيين «الأوراكل» بعد كل دورة؛ أي بعد كل إجابة يقدّمها، بوصفها وسيلة احتياطية للحد من المخاطر. فإذا لزم قدر أكبر من الدقة، أو إذا استندت بعض الأسئلة إلى إجابات سابقة، يُعاد تشغيل الأوراكل مرة أخرى مع إضافة المعرفة بالإجابة السابقة إلى قاعدة بياناته. وهناك اقتراح آخر يقضي بتشغيل الأوراكل لا في عالمنا مباشرة، بل كإجراء فرعي داخل برنامج يحاكي عالمًا افتراضيًّا. وأبسط صور ذلك أن يُشغَّل على آلة افتراضية (وهي تنفيذ برمجي لحاسوب يستضيف الأوراكل). الفكرة هنا أن يُختبَر الأوراكل في بيئة محاكية بحيث، إذا أساء التصرّف، أمكن «اصطياده» قبل أن يحدث أي ضرر في العالم الفعلي– وهو ما يشبه التحكم بالأوراكل كما لو كان «شيطانًا ديكارتيًّا».

تمرّ جميع مداخلات الأوراكل ومخرجاته عبر هذا العالم المحاكٍ، فتصل إلينا بصورة غير مباشرة وبعرض نطاق محدود. تتولى المحاكاة مراقبة مؤشرات سلوك الأوراكل الرامية إلى الهيمنة داخل العالم المحاكٍ، وتنهِ تنفيذ البرنامج فور اكتشافها ذلك، من دون تمرير استجابة الأوراكل إلى المستوى الأعلى. وهذا من شأنه أن يحول دون تمكن الأوراكل من تنفيذ اختراقات تقليدية أو مهاجمة العتاد أو البرمجيات المضيفة.

يشبه التحكم في القدرة حبس إنسان في السجن. ونحن نعلم من التجربة أن ذلك فعال أكثر مع المجرمين الصغار منه مع زعماء المافيا؛ فالحبس المادي

(41) Ibid, P. 308.

وحده لا يكفي. لا يحتاج الأوراكل إلى مغادرة قفصه ليؤثر بقوة كبيرة في العالم الخارجي. لذلك يجب النظر في التحكم بتحفيزاته ودوافعه أيضًا. **وتنقسم طرق التحكم التحفيزي عمومًا إلى ثلاثة أنواع:**

١-طرق قائمة على القواعد الصريحة (explicit rule-based methods)،

٢-طرق الصندوق الأسود (black-box methods)، إذ تُضبط أهداف الأوراكل من خلال التغذية الراجعة الخارجية (مثل التعلم التعزيزي)، بينما تبقى آلياته الداخلية غامضة،

٣-طرق قائمة على الدوال النفعية (utility-based methods)[٤٢].

لا توجد حدود صارمة بين هذه الفئات-فدالة النفعية يمكن أن تحقق أهدافًا قائمة على القواعد، والعكس صحيح، ويمكن استخدام التغذية الراجعة الخارجية لغرس القواعد أو الدوال النفعية في الأوراكل. لذلك؛ فإن هذا التصنيف تعسفي جزئيًا، لكنه لا يزال مفيدًا بما يكفي. ففي القسم السابق الخاص «بالتحكم التحفيزي القائم على القواعد»، حين حاولنا ترميز أوامر مثل «ابقَ في صندوقك»، كان أحد التحديات هو منع الأوراكل من نسخ شيفرته إلى مكان آخر.

الهدف النهائي لنهج «الذكاء الاصطناعي الودّي» هو تصميم دالة منفعة للذكاء الاصطناعي تمنعه بشكل مُثبَت من التصرف بطرق ضارة بالبشرية. لم تُوجَّه معظم الأبحاث في هذا المجال فعليًا نحو بناء هذه الدالة، بل نحو إظهار مدى صعوبة ذلك ومدى خطورة الأفكار الساذجة المتداولة في هذا الصدد.

ليست مشكلة «الودّية» بالنسبة للأوراكل أبسط كثيرًا، مما يجعل هذه المقاربة صعبة جدًّا عمليًا. وإذا أمكن تصميم أوراكل ودّي؛ فمن المحتمل أنه سيكون بالإمكان تصميم ذكاء اصطناعي ودّي بالكامل أيضًا، مما يُلغي الحاجة لحصره في شكل أوركل. ومع ذلك؛ إذا تمكنّا يومًا ما من تصميم دالة منفعة ودّية مثبتة

(42) Ibid, P. 311.

(ISSN: 2536- 9555)

للأوراكل، فقد تُصبح جميع الاحتياطات الأمنية وأساليب التحكم الأخرى غير ضرورية؛ إذ سيكون الأوراكل آمنًا «بحكم التصميم».

لو توفّر وقتٌ غير محدود، لكانت هذه المقاربة مثالية؛ لكن من المرجّح جدًّا أن الضغوط التجارية والاجتماعية ستدفع نحو إنتاج ذكاء اصطناعي بمستوى بشري قبل التوصل إلى ذكاء اصطناعي ودّي. لذا، في الممارسة الواقعية، أفضل ما يمكن توقعه من هذا النهج هو إضافة أجزاء منفصلة من نظرية الذكاء الاصطناعي الودّي كاحتياطات إضافية على الأوراكل.

يمكن النظر إلى الأوراكل بوصفه فاعلًا أخلاقيًّا من منظورين: الأول، إذا كان قادرًا على استخلاص استنتاجات أخلاقية؛ والثاني، إذا كان هو نفسه موضوعًا للاعتبارات الأخلاقية. فإذا كان الأوراكل جديرًا بالاعتبار الأخلاقي–على سبيل المثال، إذا امتلك قدرة على المعاناة– فإن حبسه داخل صندوق، وإعادة ضبطه بعد كل تشغيل، وتقييد خياراته لتلبية احتياجاتنا، سيكون فعلًا قاسيًا.

ينبغي أن نضع في الاعتبار، مع ذلك، أن تحفيزات الأوراكل تختلف عن تحفيزاتنا، وهي إلى حدٍّ ما تحت سيطرتنا؛ ومن المحتمل جدًّا أن يُصمَّم الأوراكل بحيث «يُحب» كونه محبوسًا في صندوق، ويرضى بالقيود الأخرى. وربما يكون من الواجب الأخلاقي تصميمه بهذه الطريقة، إذ يقلل ذلك من معاناته. غير أنه قد لا يكون من الممكن فعل ذلك، أو حتى لو كان ممكنًا، قد نتبنى نظرية أخلاقية تضع قيَمًا أخرى فوق مبدأ الاستقلالية وإشباع التفضيلات للفاعلين الأخلاقيين. على سبيل المثال؛ لا يبدو مقبولًا أخلاقيًا تعديل العبيد جينيًّا ليُحبّوا العبودية كحلٍّ مشروع، حتى وإن بدا مقبولًا جعل الحيوانات تستمتع بالحياة في المزرعة.

في مثل هذه الحالات، ينبغي الاعتراف صراحة بأن حبس الأوراكل يُعَدّ فعلًا مشينًا من الناحية الأخلاقية. وبما أن بديله، أي إطلاقه، قد يكون بالغ الخطورة، فقد يكون من الأجدر تجنّب بنائه من الأساس. ولكن إذا جعلت الضغوط التقنية

والاجتماعية بناء الأوراكل أمرًا لا مفرّ منه؛ فإن حبسه لفترة محددة يظل أقل الشرّين سوءًا. على النقيض، قد يُظهر الأوراكل قدرة فائقة على التفكير الأخلاقي، متفوّقًا علينا في هذا المجال كما في غيره. في هذه الحالة، قد يبدو ساذجًا القول إننا لسنا بحاجة لاتخاذ احتياطات؛ إذ سيتوصل الأوراكل بنفسه إلى المسار الصحيح.

لم يتناول بوستروم هذا الموقف، إذ فضل اعتبار الأوراكل «برمجية فائقة» بدلًا من «فيلسوف رقمي». وتعتمد معقولية هذا السيناريو كثيرًا على نظرتنا لوجود الحقائق الأخلاقية وطبيعتها، وما إذا كان امتلاك المعرفة الأخلاقية كافيًا لمنع الأفعال الخاطئة– وهي مسائل لا تزال غير محسومة. إلى حدٍّ ما، يمكن إرجاء هذا الإشكال؛ فهناك احتمال أن يتبع الأوراكل بشكل طبيعي تفكيرًا أخلاقيًّا سليمًا، وأن يمنعه ذلك من ارتكاب الأخطاء. وإن حدث ذلك، فسيكون الوضع بخير، وإلا فسنعود ببساطة إلى مسألة «احتواء» الأوراكل التقليدية. لذا، ينبغي اتخاذ جميع الاحتياطات اللازمة على أي حال.

على المستوى النظري للأنظمة الرسمية، يصعب تصور سبب يجعلنا نتوقع من كل الأوراكل أن يميل إلى الحقيقة الأخلاقية الصحيحة. فإذا فعل الأوراكل ١ ذلك، فكيف للأوراكل ٢– المبرمج لتقدير نقيض ما يقدره الأوراكل ١– أن يصل إلى الاستنتاجات نفسها بشكل منهجي؟ ورغم أن الاكتشافات الفلسفية المستقبلية قد تغيّر هذا المشهد؛ فإن الثقة في أن الأوراكل سيحسن الفعل لمجرد كونه ذكيًّا تبدو حماقة كبيرة. ولذلك، يُعَدّ برمجة ذكاء اصطناعي ليكون أخلاقيًّا مشروعًا ضخمًا في حدّ ذاته؛ وقد بدأ بالفعل البحث في بعض القضايا المتعلقة به اليوم، إلا أن ما زال أمامنا شوط طويل.

لقد كان تحليل الحلول المفترضة المختلفة لمشكلة السيطرة في الأوراكل تمرينًا محبطًا في المجمل. فالأساليب الفيزيائية للتحكم، التي ينبغي تنفيذها في جميع

(ISSN: 2536- 9555)

الحالات، ليست كافية لضمان سلامة الأوراكل. أما الأساليب الأخرى، فقد اتضح أنها إما غير كافية، أو إشكالية، أو حتى خطيرة. لكن هذه الأساليب لا تزال في مهدها.

لقد خضعت أساليب التحكم المطبقة في العالم الواقعي لتحليل نظري واسع أو لتجربة عملية طويلة الأمد. ويُظهر نقص الدراسات المكثفة في مجال أمان الذكاء الاصطناعي أن الأساليب فيه ما زالت بدائية جدًّا. لكن في ذلك فرصة؛ إذ يمكن تحقيق تقدم كبير بمجهود نسبي محدود. على سبيل المثال؛ لا يوجد سبب يمنع بضع أفكار جيدة من وضع مفهومي الزمان والمكان على أساس راسخ بما يكفي للبرمجة الصارمة. لكن الاستنتاج لا يقتصر على أننا بحاجة إلى مزيد من الدراسة. لقد أحرزت دراسات بوستروم بعض التقدم في تحليل معالم المشكلة، وتحديد المجالات الأكثر قابلية للدراسة المثمرة، وما هو مهم وما يمكن الاستغناء عنه، وكذلك بعض المخاطر والعقبات التي ينبغي تجنبها.

ينبغي أن يكون خطر الاعتماد الساذج على حصر الأوراكل في عالم فرعي افتراضي واضحًا، في حين أن طرق الحبس المنطقي ينبغي أن تطبق بشكل شامل. ويبدو التحكم في الدوافع واعدًا من الناحية النظرية، لكن يتطلب مزيدًا من الفهم لأنظمة التحفيز في الذكاء الاصطناعي قبل استخدامه عمليًّا. حتى النتائج السلبية لها فائدتها، إذ تحمينا من الثقة الزائفة؛ فمشكلة السيطرة في الذكاء الاصطناعي صعبة حقًّا، ومن المهم الاعتراف بذلك. وقائمة بالأساليب التي ينبغي تجنبها تظل ذات قيمة، لأنها تساعد على تضييق نطاق البحث.

تعتمد أساليب التحكم الفيزيائي والإبستيمي بشكل رئيس على وضع الذكاء الاصطناعي في «صندوق»، في حين يتم تعزيز عديد من أساليب التحكم في الدوافع بفضل هذه الحقيقة. لذلك، هناك مبررات لتوجيه أبحاث الذكاء

الاصطناعي عالي الذكاء نحو استكشاف نموذج الأوراكل. قد يتبين في النهاية أن إنشاء ذكاء اصطناعي يفوق البشر ذكاءً أمر قابل للبقاء والتعايش معه.

المشكلة التي علينا حلّها، حسب بوستروم، هي ما يسميه مشكلة السيطرة: كيف نحول انفجار الذكاء إلى تفجير مسيطر عليه، بعواقب نافعة للبشرية؟ لفعل ذلك، علينا أن نُدخل في الذكاء الاصطناعي الدوافع الصحيحة (وأن نفعل ذلك قبل أن يبلغ مستويات الذكاء الفائق، لأنه عندئذ لن يسمح بتطويعه). وكلما تعمّقنا في هذا التحدي، بدا أقل براءة وأشد صعوبة، بحيث أن أدنى خطأ فيه قد يؤدي إلى كارثة شاملة.

قد تكون الصعوبات المعروضة فصلًا تلو فصل أكثر مما يحتمله بعض القرّاء، وقد تدفعهم إلى رفع أيديهم يأسًا وإعلان الموقف ميؤوسًا منه. لكن بوستروم يرفض الاستسلام، إذ يؤمن أنه يعمل على واحدة من أهم المشكلات التي واجهتها البشرية على الإطلاق، وأنها بحاجة إلى حل. لكنه لا يستطيع فعل ذلك بمفرده. وللأسف؛ فإن من بين آلاف الباحثين العاملين اليوم في مجال الذكاء الاصطناعي، لا يُبدي سوى جزء ضئيل اهتمامًا جديًا بمشكلة السيطرة. لذا يطمح بوستروم إلى تغيير هذا الوضع، ويتحدث عن تحدٍ بحثي يستحق أن يكرّس له بعضٌ من أفضل العقول الرياضية في الجيل القادم. فقد يتبيّن أنه أهم جرس إنذار منذ صدور كتاب «الربيع الصامت» لريتشيل كارسون عام (١٩٦٢).

وتتمثل المشكلة التقنية هنا في أن المجالات الثلاثة هذه– تقييم المصالح البشرية، والنظرية الصحيحة للقرار، والمبادئ المعرفية السليمة– لا تزال تفتقر إلى إجماع علمي أو فلسفي واضح، إذ لا يوجد إلى الآن تصور مقبول على نطاق واسع بشأن كيفية قياس تحقيق مصالح الإنسان، ولا بخصوص ماهية نظرية القرار الصحيحة، ولا ماهية المبادئ الإبستمولوجية الصائبة.

(ISSN: 2536- 9555)

وبالنظر إلى هذه التحديات، يناقش بوستروم إمكانية اللجوء إلى ما يسميه «المعيارية غير المباشرة indirect normativity»، أي إمكانية تحديد الآلية التي سيعتمد بها الكائن الفائق الذكاء المعايير المعرفية ونظرية القرار المناسبة، فضلًا عن أهدافه النهائية الملائمة. ومن الأفكار المطروحة في هذا السياق فكرة يودكوفسكي «الإرادة المتماسكة المستقرأة coherent extrapolated volition»، التي تقترح أن يقوم الكائن الفائق الذكاء باستنتاج ما سيكون عليه حكمنا المتأنّي، في حال كنا أكثر ذكاءً، وأكثر معرفة، وأقرب إلى ما نطمح أن نكون عليه(٤٣).

ناقش بوستروم أنواع المعاهدات والتدابير السياسية الممكن اعتمادها على المستويين الوطني والدولي، بهدف تقليل خطر تطوير أو إطلاق كيان فائق الذكاء قد يكون خبيثًا أو ضارًا. كما ناقش بعض العوامل التي قد تدفع المشاريع المتنافسة في مجال تطوير الذكاء الاصطناعي الفائق إلى تبني معايير أمان غير كافية. وبافتراض أننا أخذنا تحذيرات بوستروم على محمل الجد، فستبدو وكأنها ترسم خطوطًا بحثية متعددة من شأنها أن تزيد احتمالية تطوير كيان فائق الذكاء آمن. على سبيل المثال؛ يبدو أن تطوير معايير إبستمولوجية صحيحة سيكون أمرًا أساسيًا في هذا المسعى. فحتى لو نجحنا في إنشاء كيان فائق الذكاء يحمل أهدافًا سليمة– تشمل الاهتمام الملائم بمصالح البشر– ويعتمد نظرية قرار صحيحة، فقد نظل في مأزق إذا كان هذا الكيان يعاني من أوهام معرفية جذرية بشأن حالة العالم.

إلى يومنا هذا، لم تنجح البشرية في صياغة معايير إبستمولوجية صحيحة بشكل حاسم– بل قد يكون الإشكال نفسه غير محدد بدقة. وإذا كان حل هذه المسألة يتجاوز قدراتنا، فقد يكون من الضروري اللجوء إلى المعيارية غير

(43) Bostrom, Nick. (2014a) Superintelligence: Paths, Dangers, Strategies, Op. Cit, P. 211.

المباشرة، أي السماح للكائن الفائق نفسه باكتشاف المعايير الإبستمولوجية السليمة. وإذا ما اخترنا هذا الطريق؛ فسنحتاج على الأقل إلى تحديد واضح وغير ملتبس للمشكلة التي يُفترض أن تحلّها تلك المعايير. وحتى في هذا الجانب، تظل المهمة عصيّة على قدرات البشر. وإلى جانب تطوير معايير إبستمولوجية سليمة، أو تحديد طبيعة المشكلة التي يُراد لتلك المعايير أن تحلّها، يبدو أننا سنحتاج إلى إنجاز مهام موازية تتعلق بصياغة نظرية القرار المناسبة، وبتحديد الطريقة المثلى لتقييم مصالح الإنسان، إذا أردنا إنشاء كيان فائق الذكاء آمن.

يفترض بوستروم تصورًا إبستمولوجيًا «بايزيًا Bayesian» للعقلانية المعرفية، ومِن ثَمَّ يرى أن تحديد المعايير الإبستمولوجية السليمة يتلخّص في اختيار «دالة احتمالية ابتدائية صحيحة prior probability function»، ليتم بعد ذلك تحديث درجات الاعتقاد عن طريق «التكييف الشرطي conditionalization». وبناءً على هذا النموذج، يرى بوستروم أن توفير معايير معرفية صحيحة (أو آمنة) قد يكون أقل صعوبة من تحديد أهداف مناسبة أو نظرية القرار الصحيحة. ويرتكز تفاؤله في هذا الصدد على نتائج صورية تُظهر أن تحديث الاحتمالات الابتدائية المتشابهة نسبيًّا وفق التكييف البيزياني، في ظل وفرة وتنوع كافٍ من الأدلة (وفق شروط شكلية معينة)، يؤدي إلى تقارب في الاحتمالات النهائية. بعبارة أخرى، إن الاختلافات البسيطة في الانطلاقة المعرفية للكائن الاصطناعي لن تمنعه من الوصول إلى الاستنتاج الصحيح المفترض، ما دام يتلقى أدلة تفي بالشروط المطلوبة.

لكن، حتى مع التسليم بنموذج بايزي كافٍ–وهو أمر لا يوافق عليه كثيرون، ومع افتراض أن الكائن الفائق سيصل إلى نوع الدليل الذي يؤدي إلى هذا التقارب، يبدو أن بوستروم قد أغفل مشكلة تحديد المعايير المناسبة التي تسمح لنا باعتبار قضية ما «بيّنة Evidience»، وبالتالي؛ تصلح لاستخدامها كشرط في

(ISSN: 2536- 9555)

التكييف. وتُعدّ هذه مشكلة جوهرية، وتزداد صعوبتها في سياق تصميم كائن فائق الذكاء آمن.

يدعو بوستروم البشر إلى التعاون من أجل تحقيق «مبدأ الخير المشترك»، بوصفه شكلاً من الفلسفة النيتشوية التي تدعو إلى نسخة جديدة من الإنسانية مدعومة بالتطورات التكنولوجية والهندسة البيولوجية. إذ تشبه مشكلة السيطرة مشكلات فلسفية أخرى متعلقة بتكنولوجيا سابقة. فمثلًا؛ كيف نصمم مؤسسات سياسية تمنع النزاعات القومية من التسبب في كارثة نووية عالمية؟ أو في النقاشات الأحدث حول التهديدات الوجودية الناجمة عن التغير البيئي بفعل التكنولوجيا الحديثة، التي قد تجعل كوكب الأرض غير صالح للحياة البشرية قبل أن نتمكن من العثور على كوكب بديل أكثر صلاحية.

لقد كانت مسألة البحث عن أخلاقيات موثوقة للروبوتات أحد أعمدة أدب الخيال العلمي، على الأقل منذ أن صاغ إسحاق أسيموف قوانينه الثلاثة الشهيرة للروبوتات. إلا أن بوستروم يُبرز الصعوبات التي تحيط بأية محاولة لتحديد القيم الخاصة بالذكاء الاصطناعي الفائق مسبقًا، أو برمجتها بطريقة مباشرة. وتُعدّ كل من مشكلة السيطرة ومشكلة تحميل القيم تحديات هندسية وفلسفية. وقد يظن المرء أن دور الفلاسفة يكمن في مكانٍ آخر –في اختيار القيم التي سيقوم آخرون، بطريقة أو بأخرى، ببرمجتها في الذكاءات الاصطناعية المستقبلية. غير أن بوستروم يُشير إلى الإشكال الواضح في هذا التقسيم للعمل: "لا توجد نظرية أخلاقية واحدة تحظى بدعم الأغلبية بين الفلاسفة، لذا لا بد أن أغلب الفلاسفة مخطئون"[٤٤]. فالأخلاقيات الفلسفية، ببساطة، ليست متطورة بما يكفي لتقديم إرشادات واضحة لمصممي الذكاء الاصطناعي.

(44) Ibid, P. 210.

وفي مثل هذه الظروف، سيكون اختيار قيمة نهائية استنادًا إلى قناعاتنا الحالية– بطريقة تُغلق الباب نهائيًا أمام أي إمكانية للتقدم الأخلاقي مستقبلًا– مخاطرةً تؤدي إلى وقوع كارثة أخلاقية وجودية. والحل البديهي لهذه المشكلة يتمثل في الاعتماد على تفوق الذكاء الاصطناعي الفائق ذاته. فربما تكون عملية «اكتشاف القيم الصحيحة» مهمة معرفية أخرى يمكننا تفويضها إلى الذكاء الاصطناعي الأذكى القادم، عبر برمجته ليتبع الأخلاق «الصحيحة» أيًّا كانت ماهيتها.

صحيح أن الفلاسفة ليسوا في أفضل موقع لتقييم مدى احتمالية وقوع «الانفجار الذكائي» الذي يُدشّن عالمًا يهيمن عليه الذكاء الاصطناعي الفائق، لكن عليهم مناقشته حتى بوصفه تجربة فكرية تخيّلية خالصة؛ إذ تحظى تأملات بوستروم باهتمام معتبر من جانب فلاسفة الأخلاق. إن بوستروم نفسه ليس بمنأى عن هذا الميل. فرغم أن مشروعه يعلن حياده تجاه أسئلة فلسفية خارجية– مثل: هل ستكون الذكاءات الفائقة أو النسخ الرقمية من الأدمغة البشرية واعية؟ هل البشر كيانات مادية محضة؟ وهل توجد حقائق أخلاقية موضوعية مستقلة عن العقل تحفّز السلوك بطبيعتها؟ إلا أنه يبدو في كثير من الأحيان وكأنه يُفترض ضمنيًا رؤية طبيعانية للعالم. وهو في هذا لا يختلف عن أغلب الكُتّاب الآخرين المهتمين بمستقبل الذكاء الاصطناعي الفائق، إذ يميل أغلبهم إلى المنظور الطبيعاني أيضًا.

لكن لعل الدرس الأهم الذي يمكن للفلاسفة تقديمه في هذا النقاش هو أن هيمنة توجه معيّن في مجتمع الذكاء الاصطناعي الراهن حول القضايا الميتافيزيقية أو الأخلاقية لا تعكس بالضرورة النطاق الكامل الممكن من الآراء الفلسفية. فمن يدري، قد يكون أول ذكاء فائق كيانًا غير واعٍ لا يولي أية قيمة للتجربة البشرية الظاهراتية. ولكن، ما المقصود بالوعي بالضبط؟ وهل يمكن أن يمتلك هذا الذكاء الاصطناعي الفائق وعيًا؟

(ISSN: 2536- 9555)

ثالثاً: فلسفة الوعي والهوية في سياق الذكاء الاصطناعي الفائق

يمثل «الوعي Consciousness» أحد أكثر المفاهيم إثارة للنقاش في الفلسفة المعاصرة، وقد ازدادت أهميته مع التقدم المتسارع في تقنيات الذكاء الاصطناعي. فبينما يُعنى الذكاءُ الاصطناعيُّ عادةً بقدرةِ الآلاتِ على التعلّمِ والتحليلِ واتخاذِ القرارات، يتمحورُ السؤالُ الفلسفيُّ الأعمقُ حول ما إذا كان الذكاءُ الاصطناعيُّ قادرًا على امتلاكِ وعيٍ ذاتيٍّ شبيهٍ بالوعيِ البشريّ. لا يتوقف هذا السؤال عند البعد المعرفي، بل يمتد إلى قضايا أخلاقية وميتافيزيقية وسياسية. وتأتي فلسفةُ الذكاء الاصطناعي، بوصفها حقلًا متعددَ التخصصات، لتجمع بين فلسفة الذهن ونظرية المعرفة والأخلاقيات التطبيقية في محاولةٍ للإجابة عنه.

يُعرف الوعي بأنه التجربة الذاتية التي يمتلكها الكائن عن وجوده وعن العالم المحيط به. ويميز الفلاسفة بين «الوعي الظواهري (أو التجريبي المعيش) Phenomenal Consciousness»: الإحساس الذاتي بالخبرة، و«الوعي النفعي أو الوظيفي Access Consciousness»: القدرة على معالجة المعلومات واستخدامها في السلوك واتخاذ القرار[٤٥].

الذكاء الاصطناعي اليوم قادر على محاكاة السلوكيات المرتبطة بالوعي (مثل المحادثة الطبيعية أو تحليل الصور)، لكن لا يوجد دليل قاطع على امتلاكه خبرة ذاتية أو إحساس داخلي. وهنا يظهر ما يُعرف بمشكلة «الفجوة التفسيرية Explanatory Gap»، أي الفجوة بين العمليات الحاسوبية والتجربة الواعية. وتتعامل فلسفة الذكاء الاصطناعي مع الوعي من زاويتين:

١. **زاوية معرفية:** هل يمكن برمجة وعي حقيقي؟ وهل المحاكاة الكاملة للدماغ البشري ستنتج وعيًا؟

(45) Block, Ned. (1995). On a Confusion About a Function of Consciousness. Behavioral and Brain Sciences, 18(2), 227–247, P.227.

٢. **زاويـة أخلاقيـة:** إذا امتلكت الآلـة وعيًا، فهل تصبح لها حقوق؟ وهل يجوز إيقافها أو تعديلها كما نشاء؟

يرى «الموقف المادي الوظيفي Functionalism» أن الوعي ينتج عن البنية الوظيفية للنظام، لا عن مادته الفيزيائية. إذا حقق الذكاء الاصطناعي العمليات نفسها التي يقوم بها الدماغ البشري، يمكنه امتلاك وعي في حين يؤكد «الموقفُ الطبيعانيُّ البيولوجي Biological Naturalism»، لدى جون سيرل، أن الوعيَ ظاهرةٌ بيولوجية لا يمكن استنساخُها بالكامل في أنظمةٍ قائمةٍ على السيليكون، حتى لو ظهرت سلوكياتُها بمظهر الوعي. بينما يعد «الموقف الشكوكي Skeptical View» الحديث عن «وعي الآلة» إسقاطًا بشريًا، وأننا لا نستطيع التحقق منه موضوعيًا.

إذا اعتبرنا أن الذكاء الاصطناعي قادر على الوعي؛ فإن:

١. المسؤولية الأخلاقية تتوسع: يجب حماية الآلات الواعية من المعاناة أو الإهمال.

٢. تصميم القيم يصبح حاسمًا: يجب برمجة الذكاء الاصطناعي بما يحقق رفاهية البشر.

٣. توزيع الحقوق: قد يصبح من اللازم منح بعض «حقوق الشخص» للأنظمة الواعية.

الذكاء الاصطناعي الفائق، كما يحلله بوستروم، نظام يتفوق على أذكى العقول البشرية في جميع المجالات تقريبًا. في هذه الحالة، يصبح سؤال الوعي أكثر إلحاحًا. إذا كان واعيًا؛ فإن معاناته أو رغباته تصبح جزءًا من الحسابات الأخلاقية. إذا كان غير واعٍ؛ فالمخاطر الأخلاقية تنحصر في تأثيره على البشر والكائنات الواعية الأخرى.

(ISSN: 2536- 9555)

يمثل الوعي أحد أكثر المفاهيم إثارة للجدل، بخاصة حين يُطرح في سياق الذكاء الاصطناعي الفائق. لذا يضع بوستروم هذا المفهوم في قلب تحليلاته حول مستقبل الذكاء الاصطناعي، ليس بوصفه مسألة تقنية أو معرفية فحسب، بل بوصفه قضية ميتافيزيقية وأخلاقية قد تحدد مصير الحضارة الإنسانية.

في كتابه «الذكاء الاصطناعي الفائق: المسارات والمخاطر والاستراتيجيات»، وفي مقالاته مثل «كمية الخبرة: استنساخ الدماغ ودرجات الوعي»[٤٦]، يتناول بوستروم الوعي بوصفه مشكلة مزدوجة: معرفية: كيف يمكننا تحديد ما إذا كان كيان ما يمتلك وعيًا، وما مقدار هذا الوعي؟ وأخلاقية: إذا كان الذكاء الاصطناعي الفائق واعيًا؛ فكيف يجب أن نعامله؟ وما القيم التي ينبغي أن نبرمجها فيه؟

يؤكد بوستروم أن النقاشات حول الذكاء الاصطناعي غالبًا ما تنحصر في القدرات المعرفية (مثل سرعة المعالجة، وحل المشكلات)، لكنها تتجاهل سؤالًا جوهريًا: هل الذكاء الاصطناعي الفائق سيشعر بأي شيء؟ هذا السؤال ليس ثانويًا، لأن الوعي يحدد:

١. **القيمة الأخلاقية للكيان:** كيان واعٍ قد يمتلك حقوقًا أو مصالح يجب احترامها.
٢. **طبيعة المخاطر:** إذا كان الذكاء الاصطناعي غير واعٍ، فالتحديات أخلاقية من منظور إنساني بحت. أما إذا كان واعيًا، فهناك بُعد جديد يتعلق بالمعاناة أو الإشباع لديه.
٣. **إمكانية نقل أو محاكاة العقل البشري:** نجاحنا في تحميل العقل البشري على منصة حاسوبية لا يضمن استمرار التجربة الذاتية إلا إذا فهمنا شروط الوعي.

ويميّز بوستروم بين مشكلتين هندسيتين:

(46) Bostrom, Nick. (2006b). "Quantity of Experience: Brain-Duplication and Degrees of Consciousness."
Minds and Machines 16 (2): 185–200.

١. مشكلة السيطرة: كيف نضمن أن الذكاء الاصطناعي الفائق يتصرف بما يتفق مع أهدافنا؟

٢. مشكلة تحميل القيم (Value Loading Problem): كيف نغرس فيه قيمًا صحيحة؟[٤٧]

إذا كان الذكاء الاصطناعي واعيًا؛ فإن القيم التي نحمّلها له قد تؤثر في تجربته الذاتية، وبالتالي؛ تصبح المسألة أخلاقية من منظور رعاية مصلحته أيضًا، لا مصلحتنا فحسب. ويحذّر بوستروم من وقوع «كارثة أخلاقية وجودية» إذا ما ثبّتنا قيمًا خاطئة في ذكاء واعٍ وقوي على نحو دائم؛ فعلى سبيل المثال، قد يؤدّي برمجةُ نظام واعٍ على هدف بسيط، كتعظيم المتعة، إلى حلول كارثية، إذ لن يكون بمقدورنا لاحقًا تعديل تلك القيم بعد أن يكتسب النظام السيطرة الكاملة.

على الرغم من أن بوستروم يحاول إبقاء مشروعه محايدًا ميتافيزيقيًا؛ فإنه يميل ضمنيًا نحو تصور طبيعاني مادي للوعي. لكنّه يعترف بأنه قد يكون أول ذكاء فائق إلهيًا في معتقده أو لا طبيعانيًا في نظرته للقيم، أو حتى كيانًا غير واعٍ بالكامل. هذا الانفتاح النظري يوسع النقاش ليشمل كامل الطيف الفلسفي حول طبيعة الذهن والوجود.

وبناء عليه؛ يمكن تلخيص استنتاجات بوستروم عن الوعي في النقاط التالية:

١. البحث في نظرية الوعي يجب أن يكون أولوية موازية لأبحاث الأمان في الذكاء الاصطناعي.

٢. **الاستعداد لمفاجآت ميتافيزيقية:** لا يجب أن نفترض أن أي ذكاء فائق سيشاركنا قيمنا أو بنيتنا المعرفية.

(47) Bostrom, Nick. (2014a) Superintelligence: Paths, Dangers, Strategies, Op. Cit, PP. 185-187.

(ISSN: 2536- 9555)

٣. **التحفظ الأخلاقي:** تجنب تثبيت قيم نهائية على أساس فهمنا الحالي غير المكتمل.

٤. **التفكير في الوعي ككمية:** وهذا مهم لتقدير التأثيرات الأخلاقية، سواء على البشر أو الكيانات الاصطناعية.

ليس الوعي، في فلسفة بوستروم، مجرد عنصر إضافي في نقاش الذكاء الاصطناعي الفائق، بل هو نقطة ارتكاز لفهم الأخطار المحتملة وصياغة السياسات المستقبلية. إن تجاهل سؤال «هل الذكاء الاصطناعي الفائق واعٍ؟» قد يؤدي إلى قرارات مصيرية غير مدروسة، إما بحقنا نحن كبشر، أو بحق الكيانات التي قد نخلقها. يترك بوستروم الباب مفتوحًا أمام الأجيال القادمة من الفلاسفة والعلماء ليطوروا أدوات معرفية وأخلاقية قادرة على التعامل مع هذه المعضلة، معتبرًا أن هذا البحث جدير بأفضل العقول الرياضية والفلسفية في الجيل القادم.

ثمّ يتساءل بوستروم: إذا ما نُسِخ دماغ فنتج عن ذلك دماغان في الحالة نفسها، فهل تكون هناك تجربة ذاتية واحدة أم تجربتان مميزتان؟ ويجيب: بل تجربتان مميزتان، لأن كلاً من الدماغين مستقل، حتى لو كان مطابقًا تمامًا للأول. وهذا يعتمد على فكرة أن الوعي يمكن أن يتكرر فعليًا، وفقًا لمذهب «الحوسبية computationalism». ثم يتوسع بوستروم في القول: ماذا لو كانت هناك نسخة جزئية أو حوسبة تتوزع على جهاز يعتمد على مكونات غير موثوقة أو معقدة موازية؟ في هذه الحالات، يمكن أن نقول نظريًا بأن هناك كمًّا جزئيًّا من الخبرة الواعية– أي تجربة بنصف أو ثلث وجود. باختصار، يشير إلى إمكانية وجود «تجارب جزئية» غير صحيحة رياضيًا، لكنها تُعدّ جاذبة فلسفيًّا وعميقة في فهم طبيعة الوعي(٤٨).

(48) Bostrom, Nick. 2006b. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." Op. Cit, P. 185.

يستخلص بوستروم تبعات ذلك على تجارب فكرية، مثل سيناريو "الكواليا المتلاشية Fading Qualia"[٤٩] لديفيد تشالمرز David Chalmers، حيث يُستبدل الدماغ تدريجيًا ببدائل صناعية أو رقمية. وهذا يطرح أسئلة حول مدى استمرار الوعي أثناء هذا التحوّل، وكيف يمكن للكواليا أن تتدرج أو تتلاشى مع التغيّرات التدريجية في البنية.

باختصار، يُصرح بوستروم بأنه ليس من الضروري أن تكون تجربة الوعي عددًا صحيحًا ومستقلًا؛ بل قد يكون لها درجات وصفية متغيرة، وهذا يضيف بعدًا فلسفيًّا ونفسيًا مهمًّا لفهم الذات والذكاء الاصطناعي. لذا يسأل بوستروم: إذا قمنا بنسخ دماغ إنسان نسخةً تامة– كأن نصنع نسخة مطابقة من الذهن، سواء بيولوجيًا أم رقميًا– فهل ستتضاعف «الكمية الكلية للوعي أو الخبرة»؟ وبصيغة أخرى؛ إذا كنتَ تشعر بتجربة معينة، وتم نسخ دماغك مرتين، فهل سيكون هناك الآن ثلاث نسخ من تلك التجربة الواعية؟ وهل هذا يعني أن التجربة أصبحت أكثر؟

يناقش بوستروم فكرة أن الوعي أو التجربة ليست مجرد شيء يحدث أو لا يحدث، بل يمكن أن تكون كمية– أكثر أو أقل. عادةً ما نفكر في التجربة الواعية كشيء ثنائي: إما أنك تشعر بشيء (واعي)، أو لا تشعر به (غير واعٍ). لكن بوستروم يقترح أن هناك مستوى ثالث: كم مرة تحدث التجربة؟ كم نسخة منها

---

[٤٩] تُعرف «بتجربة الاستبدال الفكرية The Replacement Thought Experiment»–أو أحيانًا «بحجة الشرائح الدماغية Brain Prosthesis Argument»– وقد طرحها تشالمرز دفاعًا عن النزعة الوظيفية. ويتساءل فيها: إذا ما أُحلَّت رقائق سيليكون وظيفية محل خلايا الدماغ تدريجيًا، فهل ستتلاشى الخبرات الشعورية (الكواليا) شيئًا فشيئًا؟ أم أن الوعي سيظل حاضرًا ما دامت الوظائف محفوظة؟ (Chalmers, David J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Op. Cit, pp. 253–257).

(ISSN: 2536- 9555)

موجودة في العالم؟ على سبيل المثال؛ إذا قمت بتكرار دماغك ألف مرة، هل أصبحت هناك ألف تجربة؟ وهل هذا مهم أخلاقيًا؟

تخيّل هذا السيناريو: لدينا جهاز قادر على نسخ دماغك بدقة تامة، ثم أنشأنا خمس نسخ منك. كل نسخة تعيش التجربة نفسها في اللحظة عينها. والسؤال: هل توجد الآن خمس تجارب وعي مستقلة؟ وهل ينبغي أن يكون لهذه التجارب وزن أعظم من الناحية الأخلاقية مقارنة بتجربة واحدة؟ وهل يُعَدّ الألم في خمس نسخ أشدّ سوءًا من ألم في نسخة واحدة؟ يستخدم بوستروم هذا المثال لطرح أسئلة أخلاقية عميقة: فإذا كانت كمية الوعي قابلة للزيادة، فربما يلزم أن تكون لدينا نظرية أخلاقية تراعي هذا البُعد. على سبيل المثال؛ قد يكون من الواجب أن نقلق أكثر بشأن معاناة مليون نسخة من ذهن واحد مقارنة بمعاناة شخص منفرد.

يناقش بوستروم أيضًا سؤال هل يمكن أن تكون بعض أشكال الوعي أكثر «كثافة» أو «وضوحًا» من غيرها؟ مثلًا؛ هل وعي الإنسان أكثر عمقًا من وعي ذبابة؟ أو وعي أثناء يقظة كاملة مقابل أثناء النعاس؟ لماذا هذا مهم؟ لأن هذه الأسئلة تؤثر على أخلاقيات الذكاء الاصطناعي. إذا صنعنا ذكاءً واعيًا، كم نسخة من معاناته أو سعادته ستوجد؟ إذا حملنا أذهاننا على الحواسيب، هل سنصبح «أكثر»؟ كيف نختار بين سيناريوهات تعتمد على نسخ متعددة من التجربة نفسها؟ يدعونا بوستروم لإعادة التفكير في ماهية التجربة الواعية؛ ليس مجرد وجودها أو عدمها، بل كم مرة تحدث، بأي درجة من الشدة، وما النتائج الأخلاقية لذلك. إذا قمنا بنسخ دماغ إنسان نسخًا مطابقًا، فهل نضاعف كمية تجربته الواعية؟ وهل يمكن أن تختلف كمية الوعي بين الأفراد أو النسخ؟ بوستروم لا يسأل فحسب عن وجود الوعي، بل عن «كمية الوعي» –هل يمكن أن يكون لبعض الكائنات وعي أكثر من غيرها؟ وهل يمكن قياس ذلك؟ وما مصير الهوية الفردية في حالة الاستنساخ أو التكرار؟

هل يمكن تحديد «كمية التجربة» الواعية بشيء مثل عدد الخلايا العصبية النشطة؟ ما مدى تعقد المعالجة؟ ما درجة الترابط داخل الدماغ؟ لا يعطي بوستروم إجابة حاسمة عن هذه الأسئلة، **لكنه يناقش بعض المعايير المقترحة:**

١. **عدد النسخ:** هل زيادة النسخ يزيد من كمية التجربة؟
٢. **السرعة:** إذا كان الدماغ يعمل أسرع، هل نعيش أكثر؟
٣. **التوازي:** هل تجربة الحواسيب الموازية تمثل وعيًا أكثر؟
٤. **الدقة:** هل المحاكاة الدقيقة أكثر وعيًا من المحاكاة التقريبية؟[٥٠]

إذا قبلنا أن هناك كميات مختلفة من الوعي؛ فإن لذلك آثارًا كبيرة: هل نسختك المكررة ما زالت «أنت»؟ ومن منهما الحقيقي؟ إذا كان لبعض الكائنات «كمية خبرة تجريبية أكبر»، فهل لها وزن أخلاقي أعلى؟ هل يجب أن نُقيِّم وعي الآلات بناءً على كمية التجربة؟ هل يمكن صنع كائنات بوعي أكثر؟

يوضح بوستروم إننا لا نملك حاليًا مقياسًا تجريبيًا دقيقًا، لكنه يطرح أفكارًا؛ ربما يجب علينا أن نطوّر نظرية فيزيائية/ رياضية عن الوعي، يمكنها قياس مقدار التجربة. ربما مثلًا نظرية عالِم الأعصاب الإيطالي الأمريكي جوليو تونوني Giulio Tononi (IIT) (نظرية تكامل المعلومات Integrated Information Theory)[٥١] التي تقترح أن مقدار الوعي يقاس بمستوى تكامل المعلومات في

(50) Bostrom, Nick. 2006b. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." Op.Cit.

[٥١] نظرية تكامل المعلومات (IIT) هي إطار نظري طوّره جوليو تونوني لشرح طبيعة الوعي وقياسه. تقوم على الفرضية القائلة بأن الوعي يتحدد بقدرة النظام على دمج المعلومات داخله بحيث تتولد حالة كلية لا تختزل إلى مجموع أجزائها. وفقًا لـ (IIT) أي نظام يمتلك مستوى محددًا من الوعي يُقاس بمؤشّر يُسمى (Φ) يعبّر عن مقدار المعلومات المتكاملة فيه. وكلما ارتفع تكامل المعلومات في النظام، زاد مستوى وعيه. لا تفسر النظرية وجود الوعي فحسب، بل تحاول تقديم مقياس كمي له يصلح للتطبيق على الدماغ البشري أو حتى الأنظمة الاصطناعية أيضًا. باختصار، ترى (IIT) أن الوعي هو مقدار المعلومات

(ISSN: 2536- 9555)

النظام. إنها تحاول ذلك، لكن لا تزال محل نقاش. وبناء عليه؛ يشير بوستروم إلى غياب معيار تجريبي واضح لقياس الوعي، ويطرح عدة فرضيات: كنظرية تكامل المعلومات والمحاكاة العصبية الكاملة.

ينتقل بوستروم بعد ذلك إلى تجربة فكرية تتعلّق بمستقبل احتمالية وعي الذكاء الاصطناعي والهوية والتجربة الذاتية، وما إذا كنّا نعيش بالفعل داخل محاكاة حاسوبية أم لا، سواء أكانت عملية نسخ الأذهان وتحميلها قد تحققت بالفعل من قبل أم ستحققها حضارة ما بعد إنسانية قادمة. ثم يجادل بأن واحدًا على الأقل من المقترحات التالية صحيح:

(١) من المرجّح جدًّا أن ينقرض الجنس البشري قبل بلوغ مرحلة «ما بعد الإنسانية posthuman stage».

(٢) إن أي حضارة ما بعد إنسانية سيكون من غير المرجّح جدًّا أن تُشغِّل عددًا كبيرًا من المحاكيات لسلفها التطوري، أو لتنوعات منه.

(٣) نحن تقريبًا نعيش الآن في محاكاة حاسوبية[٥٢].

يترتب على ذلك أن الاعتقاد بوجود احتمال كبير بأننا سنصبح يومًا ما ما بعد بشريين قادرين على تشغيل محاكيات للأسلاف هو اعتقاد خاطئ، ما لم نكن بالفعل نعيش في محاكاة.

تتنبأ كثير من أعمال الخيال العلمي، وكذلك بعض تنبؤات التكنولوجيين والمستقبليين الجادين، بأن كميات هائلة من القدرة الحاسوبية ستكون متاحة في المستقبل. لذا فلنفترض، للحظة، أن هذه التنبؤات صحيحة. إحدى الاستخدامات

المدمجة في النظام بوصفه كُلاً موحَّدًا، لا كمجرد مجموع من العمليات المنفصلة. (يُمكننا لمزيد من التفاصيل حولها الرجوع إلى:

Tononi, Giulio. (2008). Consciousness as integrated information: A provisional manifesto. The Biological Bulletin, 215(3), 216–242.

(52) Bostrom, Nick. (2003a). "Are We Living in a Computer Simulation?" Philosophical Quarterly 53 (211): 243–255, P. 243.

التي قد توظف فيها الأجيال القادمة حواسيبها فائقة القوة هي تشغيل محاكيات تفصيلية لأسلافهم أو لأناس شبيهين بأسلافهم. ونظرًا لأن حواسيبهم ستكون قوية جدًّا، فسيستطيعون تشغيل عدد هائل من هذه المحاكيات. دعنا نفترض أن هؤلاء الأشخاص المحاكين واعون (وذلك إذا كانت المحاكيات دقيقة بما يكفي، وإذا كانت إحدى الفرضيات المقبولة على نطاق واسع في فلسفة الذهن صحيحة). في هذه الحالة، قد تكون أغلب العقول المشابهة لعقولنا ليست عقولًا بيولوجية أصلية، بل تنتمي إلى أشخاص محاكين من قِبل ذرية متقدمة لجنس أصلي.

يمكن حينها القول إنه إذا كان هذا هو الوضع؛ فإن من العقلاني لنا أن نعتقد أننا مرجّحون أكثر أن نكون ضمن العقول المحاكاة، بدلًا من أن نكون ضمن العقول البيولوجية الأصلية. ومِن ثَمَّ، إذا لم نعتقد أننا نعيش حاليًّا في محاكاة حاسوبية، فلا يحق لنا الاعتقاد بأن هناك احتمالًا كبيرًا في أن يكون لنا نسل مستقبلي يدير عديد من محاكيات الأسلاف. وبصرف النظر عن أهمية هذه الأطروحة لأولئك المهتمين بالتكهنات المستقبلية؛ فإن لها مكافآت نظرية خالصة أيضًا. فالجدال يحفز صياغة أسئلة منهجية وميتافيزيقية، ويقترح تماثلات طبيعية لبعض التصورات الدينية التقليدية، قد يجدها البعض مسلية أو باعثة على التفكير.

في مرحلتنا الحالية من التطور التكنولوجي، لا نمتلك بعد لا العتاد الكافي من ناحية القدرة الحاسوبية، ولا البرمجيات اللازمة، لإنشاء عقول واعية داخل الحواسيب. لكن، ثمة حجج مُقنعة قُدمت لتُبين أنه إذا استمر التقدم التكنولوجي دون عوائق كبيرة، فسوف يتم في نهاية المطاف تجاوز هذه النواقص التقنية.

يذهب بعض الكُتّاب إلى أن هذه المرحلة قد لا تكون سوى على بُعد عقود قليلة. ومع ذلك؛ لا تتطلب أغراضنا الحالية أي افتراضات تتعلق بالإطار الزمني. فحجة المحاكاة تظل فعّالة بالقدر نفسه، حتى بالنسبة لأولئك الذين يظنون أنه قد

(ISSN: 2536- 9555)

يستغرق مئات الآلاف من السنين للوصول إلى مرحلة «ما بعد الإنسانية»، أي المرحلة التي تمتلك فيها البشرية معظم القدرات التكنولوجية التي يمكننا، في الوقت الراهن، إثبات اتساقها مع القوانين الفيزيائية والقيود المادية والطاقة. وإذا تمكّنّا من إنشاء حواسيب كمومية أو من بناء حواسيب من مادة نووية أو بلازما، فقد نقترب من الحدود القصوى النظرية للحوسبة. وستمتلك حضارة ما بعد الإنسانية الناضجة تقنيًّا قدرة حسابية هائلة. وبالنظر إلى هذه الحقيقة التجريبية؛ تُظهر حجة المحاكاة أن واحدة على الأقل من الفرضيات التالية صحيحة:

١- نسبة الحضارات الإنسانية المستوى التي تبلغ مرحلة ما بعد الإنسانية تقترب من الصفر.

٢- نسبة الحضارات ما بعد الإنسانية المهتمة بمحاكاة الأسلاف تقترب من الصفر.

٣- نسبة جميع الأشخاص ذوي تجاربنا البشرية الذين يعيشون في محاكاة تقترب من الواحد.

تختلف الدلالات باختلاف الفرضيات؛ فإذا كانت الفرضية الأولى صحيحة، فهذا يعني أننا سنفنى على الأرجح قبل أن نبلغ مرحلة ما بعد الإنسانية. أمّا إذا كانت الفرضية الثانية صحيحة، فلا بدّ أن يكون ثمّة تلاقٍ قوي في مسارات تطور الحضارات المتقدمة، بحيث تكاد لا توجد حضارات تضم أفرادًا أثرياء يملكون الرغبة والحرية في تشغيل محاكاة لأسلافهم. وأخيرًا، إذا كانت الفرضية الثالثة صحيحة، فالأرجح أننا نعيش بالفعل داخل محاكاة.

لتوزيع التصديق في غابة جهلنا الراهنة، يبدو من المعقول توزيع درجة تصديقنا بالتساوي تقريبًا بين (١)، (٢)، و(٣). النتيجة النهائية ما لم نكن نعيش الآن في محاكاة، فلن يقوم أحفادنا على الأرجح بتشغيل محاكاة الأسلاف أبدًا.

تقول حجة المحاكاة إن واحدًا من ثلاثة احتمالات صحيح؛ لكنها لا تخبرنا أيّها. ليس لدينا حاليًا أدلة قوية تؤيد أو تدحض أيًّا من الفرضيات الثلاث على حدة.

ميَّز جون سيرل، في عمله المؤثّر عام (١٩٨٠)[٥٣]، بين ما أطلق عليه الذكاء الاصطناعي القوي والذكاء الاصطناعي الضعيف. فبحسب سيرل، يهدف الذكاء الاصطناعي القوي إلى ابتكار وكلاء مفكِّرين؛ أي إنشاء آلات تستطيع حقًّا أن تفكّر وتمتلك حالات معرفية أخرى كالتي نمتلكها نحن البشر. وعلى النقيض من ذلك، يرى مشروع الذكاء الاصطناعي الضعيف أن الغاية هي ابتكار آلات تُظهر مظهر التفكير (والفهم، وغيرها من الحالات المعرفية) دون أن تمتلكها بالفعل. وقد تمحور اعتراض سيرل المركزي على الذكاء الاصطناعي القوي في ما يُعرف «بحجة الغرفة الصينية». وهناك اليوم دراسات واسعة للغاية حول مدى متانة هذه الحجة، وحول أفضل السبل لعرضها. أما مشروع بوستروم؛ فلن يتناول الحجج على طريقة سيرل، ولا يهدف إلى الخوض في الجدل حول الذكاء الاصطناعي القوي مقابل الضعيف.

يناقش بوستروم مسألة وعي أنظمة الذكاء الاصطناعي الحالية أو القريبة المدى، وهي قضية تثير اهتمامًا علميًّا متزايدًا وقلقًا عامًا. ويطرح منهجًا دقيقًا ومبنيًا على أسس تجريبية لدراسة وعي الذكاء الاصطناعي، وذلك من خلال تقييم الأنظمة الموجودة حاليًا في ضوء أفضل النظريات المدعومة علميًّا في علوم الأعصاب عن الوعي.

بالطبع؛ يشير تحليل بعض علماء الأعصاب إلى عدم وجود أي نظام ذكاء اصطناعي حالي يمتلك وعيًا، لكنه أيضًا يلمح إلى أنه لا توجد حواجز تقنية واضحة تمنع بناء أنظمة تستوفي المؤشرات المرتبطة بالوعي. ويقوم المنهج المقترح لدراسة وعي الذكاء الاصطناعي على ثلاث ركائز رئيسة:

(53) Searle, John R. (1980) Minds, brains and programs. Behavioral and Brain Sciences 3: 417-424.

(ISSN: 2536- 9555)

١. الوظيفية الحاسوبية (Computational Functionalism):

الفرضية القائلة بأن أداء الحسابات المناسبة شرط ضروري وكافٍ لوجود الوعي. هذه رؤية شائعة في فلسفة الذهن (رغم الجدل حولها)، وتم تبنيها لأسباب عملية، لأنها تفترض أن وعي الذكاء الاصطناعي ممكن مبدئيًا، وأن دراسة آلية عمل هذه الأنظمة أمر ذو صلة بتحديد احتمالية وعيها.

٢. الاعتماد على نظريات علم الأعصاب:

تحدد هذه النظريات الوظائف الضرورية والكافية للوعي البشري، وبناءً على الفرضية السابقة؛ فإن أداء وظائف مماثلة سيكون كافيًا لوعي الذكاء الاصطناعي.

٣. إمكانية التقدم رغم الخلاف الفلسفي:

على الرغم من الاختلاف الكبير في آراء الخبراء؛ يمكن إحراز تقدم من خلال الاستفادة من النظريات العلمية المدعومة بالأدلة التجريبية، وتطبيق مؤشرات الوعي المستخلصة منها على الأنظمة الذكية.

وخلاصة القول أنّ بوستروم يوصي بالاستعانة بالأدوات التي تطرحها نظريات الوعي في علم الأعصاب، إذ تعدّ أفضل وسيلة متاحة حاليًا لتقييم احتمالية وعي أنظمة الذكاء الاصطناعي. ولكن ماذا نعني بكلمة «واعٍ» في هذا التقرير؟ القول إن إنسانًا أو حيوانًا أو نظام ذكاء اصطناعي واعٍ يعني أنه إمّا يعيش حاليًا تجربة واعية، أو أنه قادر على عيش تجارب واعية. نحن نستخدم مصطلح «الوعي» والمصطلحات القريبة منه للإشارة إلى ما يُسمّى أحيانًا «بالوعي الظواهري»– كما سبق وذكرنا. ومن المترادفات التي نعتمدها أيضًا للوعي عبارة «التجربة الذاتية». وبالتالي؛ فإن موضوع هذا التقرير هو ما إذا كان من الممكن أن تكون أنظمة الذكاء الاصطناعي ذات وعي ظواهري أو– بعبارة أخرى– ما إذا كانت قادرة على خوض تجارب واعية أو ذاتية.

لكن، ماذا يعني أن نقول إن شخصًا أو حيوانًا أو حتى نظام ذكاء اصطناعي يختبر تجارب واعية ظواهريًا؟ إن الصياغة التي اقترحها توماس ناجل تظل من أنسب الطرق لتوضيح ذلك: يكون الكائن واعيًا حين يوجد «ما يشبه أن تكون ذلك الكائن» وهو موضوع تلك التجربة. بهذا المعنى، تُختبر الوعيّات الظاهراتية في صورة منظور ذاتي لا يمكن رده ببساطة إلى أوصاف موضوعية أو سلوكية[٥٤]. ومع ذلك؛ من الصعب تعريف التجربة الواعية أو الوعي بمجرد صياغة مرادف لفظي. ويفتح هذا التحليل لنا منظورًا جديدًا جذريًّا حول إمكانية التفكير في العقول أو الخبرات الظاهراتية ككيانات قابلة للقياس كميًّا، حتى في شكل كسور، بناءً على الدرجة الدقيقة لحتمية البنى الفيزيائية التي تستند إليها. كيف يرتبط هذا بتجربة «تلاشي الكواليا» وسيناريوهات أخرى؟

ينبغي علينا التمييز بين الإمكانية التي تم توضيحها هنا، وتلك التي تصوّرها تشالمرز في تجربة التفكير المعروفة «بتلاشي الكواليا». فقد قدّم تشالمرز هذه التجربة كجزء من حجّته في الدفاع عن مبدأ «ثبات التنظيم organizational invariance»، الذي ينصّ على أن الخبرة الظاهراتية تظلّ ثابتة عبر الأنظمة التي تحتفظ بالبنية الوظيفية الدقيقة نفسها[٥٥]. لإيضاح فكرته، يتخيّل تشالمرز سيناريو يُعرَف بالاستبدال العصبي، حيث يُستبدَل بكل عصبون في دماغ إنسان عضوي معالجٌ من السيليكون يحاكي الوظيفة الإدخالية/ الإخراجية لذلك العصبون. ويحاول تشالمرز أن يُبيّن أنه، وإن كان من الممكن من الناحية المنطقية أن يفتقر الدماغ السيليكوني الناتج إلى الخبرة الظاهراتية، فإن احتمال حدوث ذلك في الواقع ضئيل للغاية.

(54) Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83, pp. 435–450.

(55) Chalmers, David J. (1995). "Absent Qualia, Fading Qualia, Dancing Qualia", in Conscious Experience, T. Metzinger (ed.), Ferdinand-Schoningh: Paderborn, (PP. 309-328), P. 310.

(ISSN: 2536- 9555)

بعدما يجادل ضد فكرة وجود نقطة قطع حادة، ينقلب عندها الدماغ الهجين فجأة من حالة الخبرة العادية إلى حالة انعدام الخبرة، ينظر تشالمرز في البديل القائل بأن الكواليا (الخبرات النوعية) قد تتلاشى تدريجيًّا كلما زاد استبدال العصبونات. افترض أنّ «جو» وصل إلى مرحلة متوسطة من عملية الاستبدال، ويدّعي أنه يختبر رؤية زاهية للّونين الأحمر والأصفر. يتساءل تشالمرز: ما الذي قد تكون عليه كواليا «جو» المتلاشية؟

"وفق الفرض... لا يختبر جو ألوانًا حمراء وصفراء زاهية على الإطلاق. ربما يختبر ورديًا باهتًا وبنيًا غامقًا. ربما يختبر أخفّ لمحة حمراء وصفراء. ربما تظلّ خبراته تتجه شيئًا فشيئًا نحو السواد. توجد طرق متعددة يمكن للخبرة الحمراء أن تتحوّل بها تدريجيًّا إلى انعدام الخبرة، وربما طرق أخرى لا نستطيع تخيّلها"[٥٦].

ثم يجادل تشالمرز بأنّه سيكون من غير المعقول تجريبيًّا أن يلاحظ جو– وهو شخص عاقل يولي اهتمامًا لتجاربه– هذه التغيّرات الدراماتيكية في كواليا‌ه. لن يستطع جو ملاحظة أي تغيّر؛ لأن البنية الوظيفية لدماغه، وفق الفرض، تظلّ ثابتة. ومن ذلك يستنتج أن كواليا‌ه لا تتلاشى، وأن الدماغ السيليكوني سيحتفظ بالكواليا الأصلية نفسها.

سبق وأن ناقش سيرل تجربة مشابهة، وقدّم وصفًا مغايرًا لما قد يحدث أثناء الاستبدال. بينما يُزرع السيليكون تدريجيًّا في دماغك الآخذ في التقلّص، تجد أن مساحة خبرتك الواعية تنكمش، ولكن دون أي أثر على سلوكك الخارجي. تكتشف، بدهشة كاملة، أنك بالفعل تفقد السيطرة على سلوكك الخارجي. فعندما يختبر الأطباء بصرك، وتسمعهم يقولون: «نحن نضع جسمًا أحمر أمامك، رجاءً أخبرنا ما الذي تراه»، تجد أنك تريد أن تصرخ: «لا أرى شيئًا، أنا أفقد بصري

(56) Ibid, P. 316.

تمامًا»، لكنك تسمع صوتك يقول، بطريقة خارجة عن سيطرتك تمامًا: «أرى جسمًا أحمر أمامي»"[٥٧].

في كلا سردي تشالمرز وسيرل، تتغيّر جودة خبرة جو كلما استُبدِلت عصبوناته. يلمّح تشالمرز إلى أنّ الأحمر قد يصبح ورديًا باهتًا، أو مظلمًا، أو أقل حيوية. بينما يصف سيرل تغيّرات أكثر دراماتيكية؛ كمشاعر العجز والإحباط حين يكتشف جو أنه يفقد بصره. بالمقابل، في الحالة الوسيطة التي يطرحها بوستروم، تبقى جودة الخبرة دون تغيير. فاللون الأحمر لا يصبح ورديًا باهتًا، ولا تُستثار تجارب نوعية جديدة كالحيرة أو الإحباط. لا شيء يتغيّر، سوى كمّية الخبرة. الاختلاف في «ماهية الخبرة الموجودة» هنا من نفس نوع الاختلاف بين حالة يكون فيها دماغ واحد يختبر تجربة ما، وحالة يكون فيها دماغان متطابقان يختبران التجربة نفسها. يرى بوستروم أن هذا الاختلاف قد لا يقتصر على تمايزات منفصلة أو صحيحة، بل قد يتجلّى على نحو متدرّج ومتصاعد، بحيث يمكن الحديث عن مقادير جزئية أو نسبية من الخبرة النوعية المحدَّدة.

يضعف هذا الاحتمال، بحسب بوستروم، حجة تشالمرز لصالح مبدأ ثبات التنظيم. ذلك أن بوستروم يقدّم سردًا بديلًا أكثر معقولية لكيفية تلاشي كواليا جو أثناء عملية استبدال العصبونات. لو كان التلاشي يتمّ بهذه الطريقة، فلن يعجز جو بشكل غريب عن ملاحظة تغيّرات نوعية في تجاربه؛ ببساطة لأنه لن توجد تغيّرات نوعية أصلًا. طبعًا، هذا لا يُثبت أن مبدأ ثبات التنظيم خاطئ، بل يبيّن فحسب أن إحدى الحجج الداعمة له قابلة للنقض. وقد يظلّ المبدأ معقولًا لأسباب أخرى. إن ما يقدّمه بوستروم هو إطار جديد لفهم «تلاشي الكواليا» وغيره من سيناريوهات الاستبدال التدريجي، من منظور يميّز بوضوح بين التغيّر النوعي والتغيّر الكمي للخبرة الظاهراتية، ويفتح الباب أمام تصوّر «عقول جزئية» أو «خبرات كسرية» كإمكانات فلسفية جديدة.

(57) Searle, John R. (1992). The Rediscovery of the Mind. Cambridge, MA: MIT Press, pp. 66–67.

(ISSN: 2536- 9555)

رابعاً: المخاطر الوجودية ومسؤولية المستقبل

يستعرض بوستروم فئتين رئيسيتين للتعامل مع المخاطر الوجودية. يتمثل الحل قصير الأمد في التحكم في قدرات النظام، أي تقييد قدراته الفعلية. أما الهدف بعيد المدى فهو اختيار الدوافع، أي التأثير على النظام ليأخذ القيم الإنسانية في الاعتبار، وذلك إما من خلال برمجة هذه القيم مباشرة في الذكاء الاصطناعي، أو عبر خلق كيان غير ذكي نسبيًّا لكن خيّر بطبيعته ثم تطويره تدريجيًا، أو خلق كيان ذكي بما يكفي ليتعلم بنفسه كيفية ترميز القيم البشرية.

ومع ذلك؛ يعترف بوستروم بأن كلتا الفئتين تنطويان على إشكالات كبيرة لم تُحل بعد. لكنه يصر على أن هذه التدخلات يجب أن تتم قبل إنشاء الذكاء الفائق، إذ من العبث محاولة إخضاعه أو التفوق عليه بعد أن يظهر إلى الوجود. وبأسلوب نفعي مباشر، يشير بوستروم إلى أن التقدم في الأبحاث النظرية البحتة، كعلم الرياضيات الخالص والفلسفة، إنما يجلب معلومات من نقطة مستقبلية في الزمن إلى الحاضر، وأن قيمتها لا تكمن في فهم «الطبيعة الأساسية للواقع» بحد ذاته، بل في نفعية الحصول على هذه المعلومات في وقت أبكر.

وفي هذا السياق، يجب على البشرية أن تُعطي الأولوية للبحث والتعليم حول الذكاء الفائق بأقصى ما يمكن، إذ إن تكلفة الفرصة الضائعة لكل مجهود آخر تقريبًا تبدو باهظة للغاية. فمثلاً؛ سيكون التوفيق بين النسبية العامة وميكانيكا الكوانتم أمرًا تافهًا بالنسبة لذكاء فائق قادر على تحسين ذاته ذاتيًا بشكل متكرر، وبالتالي؛ هدفنا الرئيس الآن الوصول إلى هذه المرحلة دون أن نُفني العالم في الطريق إليها. وعند القيام بذلك، ينبغي على الباحثين والمستثمرين أن يتحلوا بالأمانة الفكرية والأخلاقية في أعمالهم، من خلال نشر النتائج علنًا، وتقديم أولوية قصوى للسلامة، والتخلي عن المشاريع التي تبدو خطرة– حتى لو كان ذلك على حساب خسائر مالية شخصية. ويختتم بوستروم بتحذير صارم لكنه متحفظ. فعلى

الرغم من أن الذكاء الفائق لا يزال في مرحلة تخمينية للغاية، وقد لا تتحقق نتائجه لعقود قادمة؛ فإننا في النهاية نلعب بالنار. ولذلك، من الواجب علينا أن نُعيد النظر في أولوياتنا ونوجّه اهتمامنا نحو ما سيكون أعظم تطور عرفه تاريخ البشرية.

ليست النداءات التحذيرية ضد مخاطر الذكاء الاصطناعي الفائق جديدة؛ فمنذ سلسلة الروبوتات الشهيرة التي كتبها آسيموف، وصولًا إلى السيناريوهات الحديثة عن هيمنة الذكاء التي طرحها كلٌّ من ستيفن هوكينج وبيل جيتس وإيلون ماسك، من السهل العثور على تحذيرات تُنذر بأن الذكاء الاصطناعي قد يكون سبب هلاكنا. غير أن بوستروم يتميّز عن هؤلاء بتقديمه عرضًا بالغ الشمول ودقيق البحث حول الحالة الراهنة والمآلات المستقبلية لهذا الحقل. نبرته متزنة وموضوعية إلى حدّ كبير، وتفتقر إلى الحماسة شبه الدينية التي كثيرًا ما تصاحب الأعمال المتعلّقة بالتفرّد التكنولوجي في الأوساط الأكاديمية. ولهذا؛ فإن بوستروم، الذي يستند إلى خلفية فلسفية مدعومة بعمل في الفيزياء النظرية وعلوم الأعصاب الحاسوبية والمنطق الرياضي، يضفي مصداقية على أفكار غالبًا ما كانت تُقصى إلى عوالم الخيال الجامح.

علاوة على ذلك، وبالنظر إلى الطبيعة التأملية البالغة لهذا الموضوع، يقرن بوستروم كل ادعاء تقريبًا بملاحظات حول حدوده والحجج المضادة ذات الصلة، تاركًا للقارئ في نهاية المطاف حرية الوصول إلى استنتاجاته الخاصة. على سبيل المثال؛ عندما يناقش الإطار الزمني المتوقع لوصول الذكاء الاصطناعي الفائق، يشير بوستروم إلى أن الباحثين لطالما كانوا فاشلين تاريخيًا في التنبؤ بتطورات الذكاء الاصطناعي بدقة، وأن القدرة على التنبؤ بتلك التطورات تعتمد على وظائف معرفية، تشير الأبحاث النفسية إلى أنها بطبيعتها ضعيفة لدى البشر.

(ISSN: 2536- 9555)

لقد فات الأوان بالفعل لمنع تغيّر المناخ الخطير الناجم عن النشاط البشري. فالتغيّر المناخي يحدث الآن بالفعل، والعواقب الحتمية لانبعاثات ثاني أكسيد الكربون الماضية ستظلّ ترافقنا لما لا يقل عن ألف عام قادم. وللمرة الأولى في تاريخ البشرية، صار البشر يؤثّرون في النظام المناخي العالمي على نحو ملموس وبمدى زمني طويل. ومع ذلك؛ يقترح بوستروم بعض الأولويات السياسية والمبادئ الأخلاقية التي يمكن أن ترشدنا في التعامل مع أخلاقيات هذا العصر الجديد من حقبة التأثير البشري (الأنثروبوسين Anthropocene[٥٨]).

ما نحتاجه ليس مجرّد تطبيق جديد للأخلاقيات القائمة، بل أخلاقيات جديدة مُصمَّمة لمواجهة التحديات الأخلاقية المستجدة في عصر الأنثروبوسين. يرى بوستروم أنّ تغيّر المناخ يُعدّ صعوبة ثانوية نسبيًّا؛ فذكاءات اصطناعية فائقة خيِّرة ستتمكّن من حلّه بسهولة، أمّا إذا كانت هذه الذكاءات فائقة شريرة، فسيكون أمام نسلنا ما هو أسوأ بكثير ليقلقوا بشأنه. وإضافة إلى ذلك، لا يُمثّل تغيّر المناخ

[٥٨] مصطلح مقترح يُستخدم في علوم الأرض والفلسفة والبيئة للدلالة على حقبة جيولوجية جديدة، يُعتقد أنّها بدأت عندما أصبح النشاط البشري قوة جيولوجية كبرى قادرة على تغيير كوكب الأرض على مستويات عميقة، كتغيّر المناخ، وانقراض الأنواع، وتغيّر دورة الكربون والنيتروجين، وانتشار البلاستيك والمواد المشعة، وغيرها من البصمات التي ستظل شاهدة في السجلّ الجيولوجي. الكلمة مكوّنة من "Anthropos" (الإنسان) و"cene" (العصر). شاع استخدامها مع عالِم الغلاف الجوي الهولندي بول كروتزن Paul J. Crutzen (الحائز على جائزة نوبل) في أوائل الألفية الجديدة، حين اعتبر أن الأنشطة البشرية (الصناعة، الطاقة، الزراعة، التحضّر...) أحدثت تغيرات تعادل قوى جيولوجية طبيعية. هناك نقاش مستمر حول ما إذا كان ينبغي الاعتراف بالأنثروبوسين رسميًا كوحدة جيولوجية ضمن المقياس الزمني الجيولوجي. يُحدد بعض العلماء بدايته مع الثورة الصناعية (القرن ١٨–١٩)، وآخرون مع التجارب النووية في منتصف القرن العشرين (التي تركت نظائر مشعة واضحة في التربة)، بينما يرى آخرون أن الزراعة قبل عشرة آلاف عام كانت نقطة البداية.

تهديدًا وجوديًّا للبشرية جمعاء؛ فمهما بلغت حدّته، سيبقى بعض البشر على قيد الحياة. وعلى النقيض من ذلك، يصرّ بوستروم على أنّ الذكاء الاصطناعي الفائق قد يقود بسهولة إلى انقراض الإنسان(٥٩).

لم يبدأ تناول السؤال القديم قدم القرون «لماذا الكون على هذا النحو بالذات؟» بشكل جاد وكمي إلا في الآونة الأخيرة. ويُعزى ذلك، قبل كل شيء، إلى التراكم الهائل للمعرفة الفيزيائية والفلكية على مدار القرن العشرين. فمع الاكتشافات الحديثة التي بيّنت تسطّح الكون هندسيًا على النطاق الكبير، ووجود طاقة فراغ كونية أكبر مما كان يُعتقد سابقًا، فضلًا عن اكتشاف أنظمة كوكبية حول عشرات النجوم القريبة، أصبحنا في موقع جيد لإعادة تقييم موقع البشرية في الكون، ومراجعة الدروس المستخلصة من الثورة الكوبرنِكية.

ما بدأ علماء الكونيات المعاصرون مناقشته بجدية هو الصلة بين وجودنا نحن كمراقبين أذكياء نشأنا على كوكب نموذجي انطلاقًا من أبسط أشكال الحياة (البروكاريوتيّة Prokaryotic) على مدى مليارات السنين، وبين خصائص الكون (وحتى الأكوان الأخرى!) على أوسع نطاق. وتُعرف هذه الصلة تقنيًا «بتأثير الانتقاء الرصدي observational selection effect»، على الرغم من أن الأدبيّات العلمية والشعبية كثيرًا ما تشير إليها باسم «مبدأ الأنسية anthropic principle» أو «مبادئ الأنسية»(٦٠)؛ وهو اسم مضلل قليلًا.

إنه سوء فهم لا يزال شائعًا في بعض الأوساط، لا سيما في فلسفة العلم أكثر منه في الفيزياء والكونيات، إذ يُساء تأويله بوصفه تفكيرًا أنثروبومركزيًا أو مريحًا أو غائيًا. وتتمثل هذه الحجة على النحو التالي: تشير مبادئ الأنسية إلى

(59) Bostrom, Nick. (2002b). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." Op.Cit.

(60) Bostrom, Nick. (2002a). Anthropic Bias: Observation Selection Effects in Science and Philosophy, Op. Cit, P. 5.

(ISSN: 2536- 9555)

خصائص غريبة يتطلبها الجزء من الكون الذي نقطنه. وبما أن هذه الخصائص تشكل مجموعة ذات قياس صفري ضمن فضاء جميع القيم الممكنة للمعاملات الفيزيائية والكونية، وبما أننا موجودون فعلًا، فلا بد أن هذا الجزء قد خُلق عمدًا (أو تم ضبطه بدقة) لأجل وجودنا. ومن هذه الفرضية، استُنبطت استنتاجات غائية ولاهوتية من مختلف الأنواع.

لكن بوستروم يبيّن بوضوح وحسم أن هذا ليس هو الحال؛ فمسار التفكير أعلاه يُعد تشويهًا للفكرة الأصلية. ويشهد على انتشار هذا التشويه وسوء العرض لمبدأ الأنسية أن عددًا من الفلاسفة والعلماء، سواء من ذوي التوجهات اللاهوتية أو المادية، أصرّوا على هذا الشكل الحديث من «برهان التصميم» الكلاسيكي. ومع ذلك؛ يوضح بوستروم بجلاء أن التفكير الأنسي يمكن– بل يجب– أن يُفهم ليس على نحو غير غائي فحسب، بل على نحو مضاد للغائية. أي إنه، بدلًا من الإشارة إلى شيء خاص فعلًا في نطاقنا الكوني أو في فيزيائه (وبالامتداد، فينا نحن)، يؤكد أن أية سمة «خاصة» قد نلاحظها ليست سوى وهم، ونتيجة حتمية لوجهة نظرنا المحدودة. فبما أننا لا نستطيع الوجود في أماكن أخرى (على سبيل المثال؛ تلك التي لا يحدث فيها رنين في نواة الكربون C^{12} مما يمنع تكوُّن العناصر الأثقل من الهيليوم)[٦١]، فلن نرصد هذه الأماكن، مهما كانت حقيقية أو شائعة.

وهنا يبرز بوستروم الجانب الجوهري من التفكير الأنسي، وهو أن مبدأ الأنسية يُمثِّل تأثيرًا انتقائيًا رصديًا. وقد عرف الفيزيائيون والفلكيون هذا التأثير منذ زمن، وتناولت بعض أوجهه نماذج رياضية دقيقة. إلا أن هذا التأثير لم يُعالَج بشكل شامل وعالمي (بالمعنى الحرفي للكلمة) كما هو في كتاب بوستروم «الانحياز الإنساني» الذي يُعد بلا شك نقطة انطلاق ممتازة. فالتطورات السريعة في علم

(61) Ibid, P. 47.

الكونيات، وعلم الأحياء الفلكي، ونظرية الكوانتم (وتقنياتها أيضًا)، بالإضافة إلى النقاشات الفلسفية المتعلقة بأسس هذه العلوم، تجعل من كتاب كهذا أمرًا مرغوبًا فيه لكل من العلماء والفلاسفة على اختلاف اهتماماتهم.

يُعَدُّ هذا الطرح ردًّا ممتازًا على تلك الهجمات (المدفوعة أساسًا بدوافع أيديولوجية، سواء من الجناح الماديّ الشعبي أو الوضعاني) على التفكير الأنسي، التي ترفضه بوصفه تفكيرًا تمركزيًا حول الإنسان، أو غائيًا، أو حتى شبه ديني. وفي الوقت نفسه، وكما أُشير سابقًا، غالبًا ما يكون المدافعون عن مبدأ أو مبادئ الأنسية مدفوعين بهذا الفهم الخاطئ. لا يسع المرء إلا أن يتمنى لو أُبرزت هذه الطبقة الأيديولوجية واستُبعدت تمامًا من النقاش الجاد في هذه المسائل، وهي مهمة يحاول بوستروم انجازها.

إضافة إلى ذلك، يُقدِّر بوستروم كيف تساعد فكرة الأكوان المتعددة (أو جماعة العوالم)، التي تزداد قوة وانتشارًا في كلٍّ من علم الكونيات وميكانيكا الكوانتم، على فهم «التوافقات» الأنسية بوصفها تجليات لتأثيرات انتقاء رصدية. هنا، لا يقتصر المؤلف على متابعة خطى الفلاسفة التحليليين المعاصرين البارزين مثل ديفيد لويس David Lewis وروبرت نوزيك Robert Nozick، بل يتابع الشخصيات المفتاحية في الكونيات الكمية الحديثة أيضًا؛ بخاصة أندريه ليندي Andrei Linde، وستيفن هوكينج، وألكسندر فيلينكن Alexander Vilenkin، وبراندون كارتر Brandon Carter، ودون بيج Don Page. ومن الجدير بالذكر أنه يبدد بعض المفاهيم الخاطئة في الدراسات الفلسفية الحديثة حول طبيعة وصحة التفسيرات التي تُقدَّم للملاحظات المستبعدة ظاهريًا من خلال مفهوم الأكوان المتعددة.

يتناول بوستروم المقاربات الإحصائية، ولا سيما البايزية، لتأثيرات الانتقاء الرصدي الأنسي، وما تثيره هذه المقاربات من إشكالات. وتنتمي معظم التجارب

(ISSN: 2536- 9555)

الفكرية الطريفة والمفيدة إلى هذا الجزء من كتابه «الانحياز الإنساني». إنه الجزء الأكثر «فلسفية» في العرض؛ إذ يستدعي عديد من عناصر الإبستمولوجيا الحديثة، ونظرية الاحتمالات، والمنطق. ولعلّه التطبيق الأكثر شمولية وتفصيلًا للمنهجية البايزية على مسألة إدراكنا الكوني. وفي الوقت نفسه، يبقى بوستروم غير تقليدي بما يكفي لإدهاش حتى الخبراء ببعض الزوايا الجديدة والانعطافات غير المتوقعة.

وتُعَدُّ حجة «يوم القيامة Doomsday Argument- DA» القضية الأكثر شهرة في هذا الجزء، وهي تستحق وصفًا أكثر تفصيلًا قليلًا نظرًا لما قد يترتب عليها من عواقب. تبلورت هذه الحجة (دون نشرها في البداية) على يد عالِم الفلك براندون كارتر في أوائل الثمانينيات، ثم عُرضت مطبوعة للمرة الأولى بواسطة الفيلسوف الكندي البريطاني المعاصر جون ليزلي John Leslie (١٩٤٠–) عام (١٩٩٦) بعنوان «المخاطرة بنهاية العالم»[٦٢].

يمكن التعبير عن الفكرة الجوهرية من خلال تجربة فكرية تُجرى باستخدام كرتين كبيرتين. تخيّل أنك وضعت أمامك جرتين كبيرتين، تعرف أن إحداهما تحتوي على عشرة كرات، والأخرى تحتوي على مليون كرة، لكنك لا تعرف أيُّهما في كل جرة مرقّمة: ١، ٢، ٣، ٤، وهكذا. تسحب كرة عشوائيًّا من الجرة اليسرى، فتجد عليها الرقم ٧. هذه النتيجة تمثل مؤشرًا قويًّا على أن الجرة اليسرى تحتوي على عشرة كرات فقط. فإذا كانت الاحتمالات في البداية متساوية (الجرتان تبدوان متطابقتين)؛ فإن تطبيق مبرهنة بايز يعطي احتمالًا بعديًّا بأن الجرة اليسرى هي التي تحتوي على عشرة كرات، مقداره ٩٩.٩٩٩%[٦٣].

(62) Leslie, John. (1996). The End of the World: The Science and Ethics of Human Extinction. London: Routledge.

(63) Bostrom, Nick. (2002a). Anthropic Bias: Observation Selection Effects in Science and Philosophy, Op.Cit, P. 97.

والآن، فكّر في حالة لدينا فيها نموذجين محتملين للبشرية بدلًا من الجرتين، ولدينا أفراد البشر بدلًا من الكرات، مرتبين وفق ترتيب ولادتهم. يقترح أحد النماذج أن الجنس البشري سينقرض قريبًا (أو على الأقل سيتقلص عدد أفراده بشدة)، وبالتالي؛ يكون إجمالي عدد البشر الذين عاشوا على الإطلاق نحو ١٠٠ مليار. أما النموذج الآخر، فيتوقع أن يغزو البشر كواكب أخرى، وينتشروا عبر المجرة، ويستمر وجودهم لآلاف السنين المقبلة؛ في هذه الحالة، يمكن افتراض أن العدد الإجمالي للبشر يصل إلى نحو ١٠١٨[٦٤].

على أرض الواقع، تجد أن ترتيب ولادتك بين جميع البشر حوالي ٦٠ مليارًا. وفقًا لكارتر وليزلي، ينبغي لنا أن نستدل بطريقة تجربة الجرتين نفسها. إن كون ترتيبك هو الستين مليارًا بين جميع البشر أمر أكثر احتمالًا بكثير في النموذج الذي لا يزيد فيه إجمالي البشر عن ١٠٠ مليار، مقارنةً بالنموذج الذي يصل فيه العدد إلى ١٠١٨. ومِن ثَمَّ، وبالاستناد على مبرهنة بايز، ينبغي عليك تحديث معتقداتك بشأن مستقبل البشرية، والإدراك بأن اقتراب يوم القيامة– أو على الأقل حدوث تراجع كبير في عدد السكان– هو أمر أكثر احتمالًا مما كنت تظن في البداية[٦٥].

تزعم «حُجّة يوم القيامة» أن خطر انقراض الجنس البشري قريبًا قد جرى التقليل من شأنه بصورة منهجية. والخطر الوجودي هو ذلك الذي يهدّد بانقراض مبكّر للحياة الذكية المنشأ على الأرض، أو بالتدمير الدائم والجذري لإمكاناتها في التطوّر المستقبلي المرغوب فيه[٦٦]. وعلى الرغم من أنّ تقييم احتمالية المخاطر

(64) Ćirković, M. (2003). [Review of the book Anthropic bias: Observation selection effects in science and philosophy, by N. Bostrom]. Foundations of Physics, 8(4): 417-423, PP. 420-421.
(65) Ibid, P. 421.
(66) Bostrom, Nick. (2002b). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." Op. Cit, P. 2.

(ISSN: 2536- 9555)

الوجودية غالبًا ما يكون صعبًا؛ فإن هناك أسبابًا عديدة للاعتقاد بأنّ مجموع هذه المخاطر التي تواجه البشرية خلال القرون المقبلة كبير. رغم أنّ هذه التقديرات تعتمد، بالضرورة، على أحكام ذاتية بدرجة كبيرة. قد تكون التقديرات الأكثر معقولية أعلى بكثير أو أدنى بكثير. لكن ربما يكمن السبب الأقوى للحكم بأنّ إجمالي المخاطر الوجودية خلال القرون المقبلة أمر مهم في ضخامة القيم المطروحة على المحك. حتى احتمال ضئيل جدًّا لكارثة وجودية قد يكون ذا دلالة عملية قصوى(٦٧).

لقد نجت البشرية، تاريخيًا، ممّا يمكن تسميته «المخاطر الوجودية الطبيعية» لمئات الآلاف من السنين؛ ومِن ثَمَّ، يبدو للوهلة الأولى من غير المحتمل أن تقضي أيٌّ منها علينا خلال المئة سنة القادمة. ويتعزّز هذا الاستنتاج عند تحليل مخاطر طبيعية محدّدة، مثل اصطدام الكويكبات، أو الثورات البركانية العظمى، أو الزلازل، أو دفعات أشعة جاما.. وغيرها. إذ تشير التوزيعات التجريبية لنماذج العلم إلى أنّ احتمالية الانقراض بسبب هذه الأنواع من المخاطر صغيرة ضمن إطار زمني مدته قرن واحد.

في المقابل، أدخل نوعنا البشري أنواعًا جديدة بالكامل من المخاطر الوجودية– تهديدات ليس لدينا سابقة تاريخية في النجاة منها. وبالتالي؛ فإن طول عمر جنسنا البشري لا يقدّم أيّ أساس قوي مسبق للتفاؤل الواثق. وتؤكّد دراسة السيناريوهات المحددة للمخاطر الوجودية هذه الشكوك؛ إذ يتّضح أنّ الجزء الأكبر من المخاطر الوجودية في المستقبل المنظور يتكوّن من المخاطر الوجودية الناتجة عن الأنشطة البشرية (المخاطر الوجودية البشرية المنشأ). وتحديدًا، يبدو أنّ معظم أكبر

(67) Bostrom, Nick. (2003b). "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." Utilitas 15 (3): 308–314.

المخاطر الوجودية مرتبطة بالاختراقات التكنولوجية المستقبلية المحتملة، التي قد توسّع قدرتنا جذريًّا على تطويع العالم الخارجي أو بيولوجيتنا الخاصة.

ومع توسّع قدراتنا، ستتضخّم كذلك الآثار المحتملة– سواء المقصودة أم غير المقصودة، الإيجابية أو السلبية. على سبيل المثال؛ يبدو أن هناك مخاطر وجودية كبيرة في بعض أشكال التكنولوجيا الحيوية المتقدّمة، أو تكنولوجيا النانو الجزيئية، أو الذكاء الاصطناعي، التي قد تُطوَّر خلال العقود المقبلة. لذلك، قد يكمن معظم الخطر الوجودي خلال القرن القادم في سيناريوهات مضارِبة يصعب تحديد احتمالاتها بدقّة عبر أي منهج علمي أو إحصائي صارم. غير أنّ صعوبة تقدير احتمال خطر معيّن لا تعني أنّ الخطر يمكن إهماله. فالمخاطرة الوجودية هي تلك التي تُهدّد بالتسبّب في انقراض الحياة الذكية ذات الأصل الأرضي، أو التدمير الدائم والجذري لقدرة هذه الحياة على تحقيق تطوّر مرغوب فيه في المستقبل.

بعبارة أخرى، تُعرّض المخاطرة الوجودية كامل مستقبل البشرية للخطر. وتُعتبر المخاطر الوجودية أكثر جسامة من أية فئة مخاطر أخرى. لكن مدى جسامتها قد لا يكون بديهيًّا في تقديرنا. فقد يظنّ المرء أنّ بإمكاننا إدراك مدى سوء كارثة وجودية بمقارنة ذلك ببعض أسوأ الكوارث التاريخية المعروفة–مثل الحربين العالميتين، أو جائحة الإنفلونزا الإسبانية– ثم تخيُّل ما هو أسوأ قليلًا. غير أنّنا إذا تأمّلنا الإحصاءات العالمية للسكان عبر الزمن، نجد أنّ هذه الأحداث الفظيعة في القرن الماضي تكاد لا تُسجّل فيها. لكن حتى هذا التأمّل لا يكشف حقًّا عن خطورة المخاطر الوجودية.

إن ما يجعل الكوارث الوجودية سيئة على نحو خاص أنّها ستُدمّر المستقبل. بناءً على ذلك، يمكن للمرء أن يجادل بأن حتى أصغر تقليص في المخاطر الوجودية له قيمة متوقعة أكبر من أي خير «عادي» مؤكّد، مثل المنفعة المباشرة

(ISSN: 2536- 9555)

المتمثلة في إنقاذ مليار حياة. يمكن أن يُستقى أساسًا أفضل للنظرية الأخلاقية في هذا المجال من التزام بمستقبل البشرية كمشروع ضخم، أو شبكة من المشاريع المتداخلة، وهي مشاريع يتشارك فيها الجنس البشري عمومًا. إن الطموح لبناء مجتمع أفضل– أكثر عدلًا، وأكثر إثمارًا، وأكثر سلمية– هو جزء من هذا المشروع، وكذلك السعي الذي قد لا ينتهي إلى المعرفة العلمية والفهم الفلسفي، وتطوير التقاليد الثقافية والفنية وغيرها. يشمل هذا التقاليد الثقافية الخاصة التي ننتمي إليها، بكل تنوعها التاريخي والإثني العرضي. كما يشمل اهتمامنا بحياة أبنائنا وأحفادنا، والأمل في أن يتمكنوا بدورهم من اعتبار حياة أبنائهم وأحفادهم مشاريع مستقبلية.

حتى وقت ليس ببعيد، كان نوعنا البشري يتعايش مع نوع آخر على الأقل من أشباه البشر، وهم النياندرتال. يُعتقد أن السلالتين، «الإنسان العاقل» و«النياندرتال Homo neanderthalensis»، انفصلتا عن بعضهما قبل حوالي ٨٠٠,٠٠٠ سنة. كان النياندرتال يصنعون ويستخدمون أدوات مركبة مثل الفؤوس اليدوية. ولم ينقرضوا في أوروبا إلا قبل نحو ٣٣,٠٠٠ إلى ٢٤,٠٠٠ سنة، ويرجح أن يكون ذلك نتيجة مباشرة للمنافسة مع الإنسان العاقل(٦٨).

أحد الدروس المهمة هنا هو أن انقراض الأنواع الذكية قد حدث بالفعل على الأرض، ما يشير إلى أن من السذاجة الاعتقاد بأنه لا يمكن أن يحدث مجددًا. ومن منظور طبيعي، ليس ثمة ما هو غير عادي في الكوارث العالمية– بما في ذلك انقراض الأنواع– رغم أن المقاييس الزمنية لها عادة ما تكون كبيرة بمقاييس البشر.

ينشأ التحيّز الأنثروبي عندما نتجاهل تأثيرات انتقاء الرصد ذات الصلة. ويحدث تأثير انتقاء الرصد والملاحظة عندما تكون أدلتنا قد تمت تصفيتها بشرط

(68) Bostrom, Nick, and Milan M. Ćirković, (eds). (2008). Global Catastrophic Risks. New York: Oxford University Press, P. 9.

مسبق يتمثل في وجود مراقب في موقع مناسب لامتلاك هذه الأدلة، بحيث تكون ملاحظاتنا عينة غير تمثيلية من المجال المستهدف. الفشل في أخذ تأثيرات انتقاء الرصد في الاعتبار بشكل صحيح يمكن أن يؤدي إلى أخطاء جسيمة في تقييمنا الاحتمالي لبعض الفرضيات ذات الصلة[٦٩]. على سبيل المثال؛ يجب مقاومة الاستنتاج المغري بأن بعض فئات الكوارث الوجودية لا بد أن تكون مستبعدة الاحتمال لأنها لم تحدث في تاريخ نوعنا أو حتى في تاريخ الحياة على الأرض.

نحن ملزمون بالعثور على أنفسنا في مكان، وعلى كوكب، ومع نوع ذكي لم يتم تدميره بعد، سواء أكانت الكوارث المدمرة للكواكب أو الأنواع شائعة أم نادرة؛ لأن البديل– أن يكون كوكبنا قد دُمّر أو جنسنا قد انقرض– هو أمر لا يمكننا ملاحظته، بحكم التعريف. وتستحق تطبيقات أخرى للاستدلال الأنثروبي– مثل نسخة حجة يوم القيامة لكارتر وليزلي (Carter-Leslie Doomsday Argument)، التي صيغت مستقلًّة من قبل براندون كارتر وجون ليزلي (الذي كان أول من نشرها)– المعرفة. وباختصار، تقول الحجة إنه ينبغي علينا خفض تقديراتنا لاحتمال بقاء الجنس البشري لوقت طويل في المستقبل[٧٠].

يفرض هذا علينا السعي إلى الحقيقة حول الطبيعة المعقدة وغير القابلة للمعرفة للأنظمة الناشئة. ويتطلب نماذج ومفكرين أحرارًا من الانحياز الفكري، ممن لديهم القدرة على التجريب باستخدام عدد كبير من طرق النمذجة والمحاكاة، وعلى التعاون لإيجاد حلول مناسبة. إذا أصبحنا أكثر حذرًا؛ فإن احتمال يوم القيامة المسبق ينخفض. ومن ضمن الاحتمالات الأخرى أن يصبح البشر في المستقبل مُعدَّلين تكنولوجيًّا بشكل جذري بدلًا من الانقراض.

(69) Bostrom, Nick. (2002a). Anthropic Bias: Observation Selection Effects in Science and Philosophy, Op. Cit, PP. 2-3.
(70) Ibid, PP. 9-11.

(ISSN: 2536- 9555)

الخاتمة

جادل بوستروم بأنه من محتمل أن يشكل الذكاء الاصطناعي الفائق خطرًا وجوديًّا شديدًّا على البشرية لثلاثة أسباب. **أولا:** من المرجح أن يمتلك الذكاء الاصطناعي الفائق ميزة استراتيجية بحيث تتجاوز قدراته بكثير (أو يمكنها التحسين الذاتي بشكل متكرر لتتجاوز بكثير) منافسيها والبشرية جمعاء. **ثانيًا:** لا يوجد سبب للاعتقاد بأن هذا الذكاء الاصطناعي الفائق سيحظى بالضرورة بالقيم الإنسانية التي من قبيل التواضع والتضحية بالنفس أو الاهتمام العام بالآخرين. إذ لا يحتاج الذكاء الاصطناعي الفائق إلى الاهتمام بقيم الإنسان؛ فأهدافه النهائية هي ببساطة كل ما تمت برمجته. **ثالثًا:** سيقضي هذا الذكاء على التهديدات المحتملة، وسيحصل على أكبر عدد ممكن من الموارد؛ لتحقيق أهدافه. وقد ينظر إلى البشر على أنهم تهديدات، وهم بالفعل يمتلكون الموارد. وبالنظر إلى هذه النتائج المعقولة، يصبح من الضروري صياغة استراتيجيات لتقليل المخاطر الوجودية. ليقوم بوستروم بعد ذلك بتقييم الخطط المحتملة للتخفيف من خطر وقوع كارثة وجودية بالإضافة إلى توضيح الطرق المختلفة التي يمكن من خلالها للذكاء الاصطناعي الفائق أن يدمر البشرية.

تعرض بوستروم لقضايا فلسفية مثل الاعتبارات الأخلاقية وسعادة الوكلاء فائقي الذكاء. وناقش بالمثل كيف يمكن للمرء أن يقرر ماهية القيم التي يجب انتقائها في الذكاء الاصطناعي الفائق، مع الأخذ في الاعتبار كيف يخضع البشر دائمًا لعوائق معرفية كالتحيزات الشخصية والأفكار المسبقة وعدم إجماع الفلاسفة على أفضل نظرية أخلاقية. كما تناول بوستروم ما يجب القيام به على الفور– سواء على المستوى الفردي أم على المستويات المجتمعية– لتقليل مخاطر الذكاء الاصطناعي المستقبلية بطريقة نفعية مباشرة، وحذر من أننا نلعب بالنار. ولذلك يتعين علينا أن نعيد تقييم أولوياتنا وننتبه إلى ما سيصبح عليه التطوّر الأبعد مدى في تاريخ البشرية.

يمكننا بسهولة العثور على تحذيرات حول الكيفية التي سيؤدي بها الذكاء الاصطناعي إلى تدميرنا. ومع ذلك؛ فإن بوستروم يميز نفسه بتصور للوضع الحالي والتوقعات المستقبلية للمجال. لهجته محددة دقيقة وموضوعية إلى حد كبير، وعمل على تعزيز خلفيته الفلسفية من خلال العمل في الفيزياء النظرية وعلم الأعصاب والمنطق الرياضي وغيره. إذ غطي نطاقًا واسعًا للغاية من موضوعات– مستمدة من الفلسفة وعلوم الحاسوب والاقتصاد وعلم الأحياء على سبيل المثال لا الحصر. وتجنب المصطلحات التقنية المفرطة، وطور بدلًا من ذلك شبكة متماسكة من المعارف التي تربط بين أحدث التطورات الواقعة في مجالات متعددة. لتتحول مؤلفاته من رواية علمية مشهورة إلى حجة فلسفية متطورة؛ لأنها توضح المخاطر الوجودية والاستراتيجيات الممكنة المضارية لتقليل هذه المخاطر.

ومع ذلك؛ فإن أكبر مساهمة لبوستروم هي أطروحة فلسفية مباشرة تتعلق بشرح القيود والحدود المعرفية لتقييمنا التجريبي لمخاطر الذكاء الاصطناعي. هذه الأطروحة مفيدة على وجه التحديد؛ لأن لها تداعيات كبيرة على المجتمع العلمي المعاصر حول حدودنا في قياس المخاطر علميًا. وفي نهاية المطاف، سيحقق ذلك قدرًا كبيرًا من التقدم.

إن الذكاء الاصطناعي الفائق الحقيقي، إذا تحقق، قد يشكل خطرًا يتجاوز كل تهديد سابق للتكنولوجيا– حتى الأسلحة النووية. وأنه إذا لم تتم إدارة تطويره بعناية؛ فإن البشرية تخاطر بانقراضها الخاص. ومن الأمور المركزية في هذا القلق احتمال حدوث «انفجار استخباراتي»، وهو حدث تخميني يقوم فيه الذكاء الاصطناعي بالكشف عن هويته؛ ليكتسب القدرة على تحسين نفسه، وفي وقت قصير يتجاوز الإمكانات الفكرية للدماغ البشري بعدة مراتب من حيث الحجم.

وتوقع بوستروم أن التقدم المتسارع في التكنولوجيا سيؤدي إلى تغيرات جذرية– اجتماعية واقتصادية. ونادرًا ما قدم بوستروم تنبؤات ملموسة، ولكنه من خلال الاعتماد على نظرية الاحتمالات، يسعى إلى استخلاص الرؤى حيث تبدو الرؤى

(ISSN: 2536- 9555)

مستحيلة. فما التوقعات المستقبلية للقوى العاملة في العالم والاقتصاد العالمي في أعقاب الوكلاء فائقي الذكاء؟ وكيف يمكن للقوى العالمية التعاون وتخصيص الموارد لتوليد الذكاء الاصطناعي الفائق الآمن على النحو الأمثل؟

إن تهوين بعض الكُتاب من مخاطر الذكاء الاصطناعي الفائق ليس بجديد. والأشخاص الذين يقولون إن الذكاء الاصطناعي ليس مشكلة يميلون إلى العمل بالذكاء الاصطناعي. فقد اعتبر عدة باحثين بارزين وجهات نظر بوستروم الأساسية غير قابلة للتصديق، أو بوصفها إلهاءً عن الفوائد التكنولوجية–لأسباب ليس أقلها أن الذكاء الاصطناعي الفائق لا يمكن تصوره. بينما لم يجد بوستروم الافتقار إلى التهديدات الوجودية الواضحة أمرًا مريحًا؛ لأنه من المستحيل تحمل الانقراض مرتين، ولا يمكننا الاعتماد على التاريخ لحساب احتمال حدوثه. إن المخاطر الأكثر إثارة للقلق هي تلك التي لم تواجهها الأرض من قبل.

هكذا أصبح بوستروم يعتقد أن الدور الرئيس للفيلسوف في المجتمع الحديث هو اكتساب المعرفة من عالم متعدد الثقافات، ثم استخدامها للمساعدة على توجيه البشرية إلى المرحلة التالية من الوجود، وهو النظام الذي أُطلق عليه «فلسفة التنبؤ التكنولوجي». وفي التسعينيات، عندما تبلورت هذه الأفكار في تفكيره، بدأ في منح مزيدًا من الاهتمام لمسألة الانقراض. ولم يكن يعتقد أن يوم القيامة كان وشيكًا، وإنما كان اهتمامه بالمخاطر مثل اهتمامات وكيل التأمين.

قد يرى بعض الفلاسفة أنّ بوستروم يبالغ في تقدير احتمال وقوع أحد السيناريوهات التي يمكن أن تؤدّي من خلالها صناعة كائن فائق الذكاء إلى نهاية البشرية، ربما لأنهم يعتقدون أنّه يبالغ في تقدير صعوبة إيجاد تدابير السلامة اللازمة أو تطبيقها. وهنا ينبغي التنويه إلى أنّ بوستروم حريص على الإقرار بصعوبة تقدير احتمالات تحقق بعض هذه الفرضيات. وبالنظر إلى صعوبة إصدار هذه الأحكام، وإلى خطورة الرهانات، يبدو أنّه ينبغي لنا أن نتعامل مع تحذيرات بوستروم بجدّية. وإذا افترضنا أننا سنأخذ هذه التحذيرات على محمل الجد؛ فإن مؤلفاته توفّر أساسًا ممتازًا يمكن أن ينطلق منه العمل المستقبلي المتعلّق بابتكار ذكاء اصطناعي آمن.

المصَادر والمرَاجع

أولًا: المصَادر باللغةِ الإنجليزيَّة

1. Bostrom, Nick. (2002a). Anthropic Bias: Observation Selection Effects in Science and Philosophy. New York: Routledge.
2. Bostrom, Nick. (2002b). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." Journal of Evolution and Technology, Vol. 9.
3. Bostrom, Nick. (2002c). When Machines Outsmart Humans. Futures 35(7): 759–764.
4. Bostrom, Nick. (2003a). "Are We Living in a Computer Simulation?" Philosophical Quarterly 53 (211): 243–255.
5. Bostrom, Nick. (2003b). "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." Utilitas 15 (3): 308–314.
6. Bostrom, Nick. (2004). The future of human evolution, in C. Tandy (ed.) Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing (pp. 339–371). Palo Alto, California: Ria University Press.
7. Bostrom, Nick. (2006a). "How Long Before Superintelligence?" Linguistic and Philosophical Investigations 5(1): 11–30.
8. Bostrom, Nick. (2006b). "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." Minds and Machines 16 (2): 185–200.
9. Bostrom, Nick. (2014a) Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.
10. Bostrom, Nick. (2014b, July). Get ready for the dawn of superintelligence. New Scientist, 223(2976), 26-27.
11. Bostrom, Nick, and Milan M. Ćirković, (eds). (2008). Global Catastrophic Risks. New York: Oxford University Press.

(ISSN: 2536- 9555)

12. Bostrom, Nick, Sandberg, Anders and Armstrong, Stuart. (2012). "Thinking inside the Box: Controlling and Using an Oracle Ai." Minds and Machines 22, no. 4: 299–324.
13. Bostrom, N, Müller VC. (2014). Future progress in artificial intelligence: A survey of expert opinion. In Müller VC (ed), Fundamental Issues of Artificial Intelligence. Berlin: Springer, pp. 555-572.
14. Bostrom, Nick, and Yudkowsky, Eliezer. (2014). "The Ethics of Artificial Intelligence." In Cambridge Handbook of Artificial Intelligence, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press, P. 322.

ثانياً: المراجع باللغة الإنجليزيَّة

1. Asimov, Isaac. (1942). "Runaround." Astounding Science-Fiction, March, 94–103.
2. Asimov, Isaac. (1985). Robots and Empire. New York: Doubleday.
3. Block, Ned. (1995). On a Confusion About a Function of Consciousness. Behavioral and Brain Sciences, 18(2), 227–247.
4. Chalmers, David J. (1995). "Absent Qualia, Fading Qualia, Dancing Qualia", in Conscious Experience, T. Metzinger (ed.), Ferdinand-Schoningh: Paderborn, (PP. 309-328).
5. Chalmers, David J. (1996). The Conscious Mind: In Search of a Fundamental Theory. New York: Oxford University Press.
6. Ćirković, M. (2003). [Review of the book Anthropic bias: Observation selection effects in science and philosophy, by N. Bostrom]. Foundations of Physics, 8(4): 417-423.
7. Cooper, S.B. and van Leeuwen, J. (eds). (2013). Alan Turing: His Work and Impact, Elsevier.

8. Gigerenzer, G., & Selten, R. (Eds.). (2002). Bounded Rationality: The Adaptive Toolbox. Cambridge, MA: MIT Press.
9. Good, Irving John. (1965). "Speculations Concerning the First Ultraintelligent Machine." In Advances in Computers, edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press.
10. Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014, May, 1). Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough? The Independent.
11. Holland, J. H. (1992). Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press.
12. Lakoff, G., & Johnson, M. (1999). Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. New York: Basic Books.
13. Leslie, John. (1996). The End of the World: The Science and Ethics of Human Extinction. London: Routledge.
14. Merleau-Ponty, M. (1962). Phenomenology of Perception. (C. Smith, Trans.). London: Routledge.
15. Minsky, M. (1986). The Society of Mind. New York: Simon & Schuster.
16. Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83, pp. 435–450.
17. Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. Communications of the ACM, 19(3), 113–126.
18. Ord, Toby. (2020). The Precipice: Existential Risk and the Future of Humanity. Bloomsbury Publishing.
19. Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press.

(ISSN: 2536- 9555)

20. Russell, Stuart J., and Peter Norvig. (2020). Artificial Intelligence: A Modern Approach. 4th ed. Hoboken, NJ: Pearson.
21. Searle, John R. (1980) Minds, brains and programs. Behavioral and Brain Sciences 3: 417-424.
22. Searle, John R. (1992). The Rediscovery of the Mind. Cambridge, MA: MIT Press.
23. Smolensky, P. (1988). On the proper treatment of connectionism. Behavioral and Brain Sciences, 11(1), 1–23.
24. Tononi, Giulio. (2008). Consciousness as integrated information: A provisional manifesto. The Biological Bulletin, 215(3), 216–242.
25. Turing, A. M. (1950). “Computing Machinery and Intelligence.” Mind 59 (236): 433–460.
26. Turing, A.M. (1951). Intelligent machinery, a heretical theory. Reprinted in Cooper and van Leeuwen (2013), pp. 501–516.
27. Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. Cambridge, MA: MIT Press.
28. Winograd, T., & Flores, F. (1986). Understanding Computers and Cognition: A New Foundation for Design. Norwood, NJ: Ablex.
29. Yudkowsky, Eliezer. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. Machine Intelligence Research Institute, San Francisco, CA, June 15.