أخلاقيات الآلة (الدوافع والمخاطر)

Machine Ethics Motives and risks

إعـــــداد

د. محمد حامد ذكي همام

مدرس بقسم الفلسفة ـ كلية الآداب

جامعة الوادي الجديد

تاريخ الاستلام : ٢٩/ ٥/ ٢٠٢٢م

تاريخ القبول : ٧ / ٦/ ٢٠٢٢م

ملخص:

يناقش هذا البحث بعض القضايا الفلسفية المحيطة بمفهوم الآلة الأخلاقية، وأهداف أخلاقيات الآلة مع تحديد المخاطر المحتملة التي يجب أخذها في الاعتبار وإدارتها، وأسباب متابعة مجال أخلاقيات الآلة ومناهضتها، والذي يُفهم على أنه مجال يهدف إلى بناء "الآلات الأخلاقية". لذلك سوف نشرح طبيعة هذا الهدف، ولماذا يستحق المتابعة، والمخاطر التي ينطوي عليها السعي لتحقيقه؟

أولاً– سنناقش ونوضح بعض القضايا الفلسفية المحيطة بمفهوم "الآلة الأخلاقية"، وأهداف أخلاقيات الآلة.

ثانيًا– سنناقش وجود الأسباب الوجيهة في البداية لمتابعة أخلاقيات الآلة، بما في ذلك إمكانية تحسين التوافق الأخلاقي لكل من البشر، والآلات؛ إلا أن هناك أيضًا بعض المخاطر المحتملة التي يجب أخذها في الاعتبار.

ثالثًا– سنناقش هذه المخاطر المحتملة، ونشير إلى المكان الذي يجب أن يكرس فيه البحث لتوضيح، المخاطر المحتملة وإدارتها. نختتم البحث بتقديم بعض التوصيات حول الأسئلة التي قد يتطرق إليها العمل في المستقبل من خلال توضيخ أخلاقيات الذكاء الاصطناعي للآلة، وما لها من حقوق وواجبات.

الكلمات المفتاحية: التوافق الأخلاقي، التفكير الأخلاقي، وكالة الآلة، أخلاقيات الآلة، أخلاقيات الذكاء الاصطناعي.

Abstract:

This paper discusses some of the philosophical issues surrounding the concept of a moral machine, the goals of machine ethics while identifying potential risks that should be considered and managed, and the reasons for, and opposition to pursuing, the field of machine ethics, which is understood as a field aimed at building "moral machines". So we will explain the nature of this goal, why it is worth pursuing, and the risks involved in pursuing it.

First, we will discuss and clarify some of the philosophical issues surrounding the concept of "the moral machine", and the goals of machine ethics.

Second, we will argue that while there are good reasons at first to pursue machine ethics, including the possibility of improving the moral fit of both humans and machines; However, there are also some potential risks that should be taken into consideration.

Third, we will discuss these potential risks and indicate where research should be devoted to clarifying and managing potential risks. We conclude the research by providing some recommendations about the questions that may be addressed in future work by clarifying the ethics of artificial intelligence for the machine, and its rights and duties.

Keywords: Moral compatibility, moral reasoning, machine agency, machine ethics, artificial intelligence ethics.

مقدمة

تعد أخلاقيات الآلة مجال بحث يدرس إنشاء "الآلات الأخلاقية"، ويهدف إلى توضيح ما يعنيه مشروع "البناء الأخلاقي للآلة"، ولماذا استحق المتابعة كهدف، والمخاطر التي ينطوي عليها. لذلك نرى أن الأسئلة حول الدوافع، والمخاطر مهمة لأي مجال من مجالات البحث.

نظرًا لوجود موارد محدودة للبحث العلمي؛ فإن استخدامها بطريقة مسؤولة أخلاقيًا يتطلب توضيح المجال الذي يقصد به بحث معين، وما إذا كان هذا هدفًا مرغوبًا وعمليًا، وما إذا كان ينطوي على أي مخاطر كبيرة سواء للباحثين، أم للمستخدمين، أم للجمهور الأوسع[1].

يهدف هذا البحث على وجه التحديد إلى تقديم ثلاثة إسهامات: نبدأ بمناقشة بعض التعقيدات الفلسفية الأساسية التي ينطوي عليها مفهوم "الآلة الأخلاقية". ثم نحدد بعض الفوائد المحتملة التي قد يجلبها إنشاء مثل هذه الآلات، والمخاطر المحتملة المرتبطة بها، وانتهينا إلى أنه على الجانب الإيجابي لإنشاء آلات أخلاقية يجب التركيز على المزيد من الأبحاث لتوضيح هذه المخاطر وإدارتها. هدفنا في مناقشة هذه الأسئلة هو في المقام الأول تحديد المشاكل أو المضاعفات المحتملة التي قد يتم تناولها من خلال البحث في المستقبل.

واقتضت طبيعة هذا البحث الاستناد والاعتماد على المناهج الآتية:

١- **المنهج التحليلي:** وقد استخدمته لتحليل رؤية بعض الفلاسفة المتناولة في البحث، وذلك للكشف عن تفاصيل أفكارهم المتعلقة بموضوع البحث.

٢- **المنهج النقدي:** وقد اعتمدت عليه في توجيه النقد بموضوعية بعيداً عن أي تحيز لبعض هذه الأفكار.

ويشمل البحث مقدمة، وثلاثة محاور رئسية، يعقبها خاتمة، وقائمة المصادر والمراجع.

المقدمة: فيها تمهيد وتعريف موجز لموضوع البحث والمنهج المستخدم.

المحور الأول: أخلاقيات الآلة.

المحور الثاني: دوافع أخلاقيات الآلة أو الماكينة.

المحور الثالث: مخاطر محاولة إنشاء آلات أخلاقية.

المحور الأول: أخلاقيات الآلة

يشمل مصطلح "آلة" بالمعنى الواسع كلًا من الآلات المادية العادية، والروبوتات المستقلة، فضلاً عن الأنظمة الخوارزمية البحتة، ويتضمن أيضًا ماكينات الصراف الآلي، وروبوتات التنظيف، وتطبيقات الهواتف الذكية، ومع ذلك سوف نركز على نوع معين من الآلات أي تلك التي يمكن أن يكون لديها القدرة على "التفكير الأخلاقي" مثل الآلات المادية، والروبوتات المستقلة، والأنظمة الخوارزمية البحتة.

لذلك نقتصر مصطلح "أخلاقيات الآلة" على البحث الذي يسهم بشكل مباشر في إنشاء آلات أخلاقية، وهذا يشمل محاولات المهندسين والعلماء لتصنيع مثل هذه الآلات فعليًا، والبحث النظري الذي يهدف إلى تسهيل أو تمكين ذلك، وليس الاستفسارات الفلسفية الواسعة حول الآثار المترتبة على هذه التكنولوجيا. التي يُطلق عليها أحيانًا "ميتا أخلاقيات الآلة"*(٢).

* الميتا– أخلاق أو "ماوراء الأخلاق" Metaethics هي محاولة فهم الافتراضات المسبقة والالتزامات الميتافيزيقية والإبستمولوجية والسيمانطيقية والسيكولوجية التي يحملها كل من التفكير والكلام والممارسة الأخلاقية، كما تتضمن الميتا – أخلاق في حدودها مجموعةً واسعةً من الأسئلة والإشكالات، مثل السؤال هل الأخلاق مسألة أذواق أكثر من كونها مسألة حقيقة؟ أو هل المعايير الأخلاقية نسبية حسب الثقافة؟ أو هل هناك حقائق أخلاقية أصلًا؟ وإن كانت هناك حقائق أخلاقية، فما مصدرها؟ و كيف تضع معيارًا مناسبًا لسلوكنا؟ وكيف ترتبط الحقائق الأخلاقية بالحقائق الأخرى (كالتي عن السيكولوجيا أو السعادة أو الأعراف الإنسانية مثلًا...)؟ وإن كانت الحقائق الأخلاقية موجودة فكيف نَعرِفها؟ هذه الأسئلة تقودنا بطبيعة الحال إلى إشكالات متعلّقة بمعنى الادّعاءات الأخلاقية وأخرى متعلّقة بالحقيقة الأخلاقية و مسوّغات التزاماتنا الأخلاقية. تبحث الميتا–أخلاق كذلك في العلاقة بين القِيَم ومسوّغات الفعل والدافع الإنساني، إذ تسأل كيف تعطينا المعايير الأخلاقية مسوّغاتٍ للفعل أو الامتناع حسب ما تأمر أو تنهى، وتعالج أيضًا كثيرًا من القضايا المتعلقة بطبيعة الحرية وأهميتها (أو عدم أهميتها) للمسؤولية الأخلاقية. انظر، موسوعة ستانفورد للفلسفة، الميتا أخلاق، ترجمة حسان عبيدات، ص ١
https://hekmah.org/wp-content/uploads/2019/02/%D8%A7%D9%84%D9%85%D9%8A%D8%AA%D8%A7-%D8%A3%D8%AE%D9%84%D8%A7% D9%82.pdf (٢٠٢٢/٥/١٧) تاريخ الدخول على الموقع

يستخدم مصطلح " الأخلاق" تقنيًا من قبل الفلاسفة ليعني دراسة فلسفيه للأخلاق، وتُفهم الأخلاق على أنها مجموعة من القواعد والمبادئ والمعايير الاجتماعية التي تهدف إلى توجيه السلوك البشري في المجتمع، وكمعتقدات حول السلوك الصواب والخطأ، وكذلك الشخصية الجيدة أو السيئة. فعلى الرغم من أن الأخلاق (Morality) هي موضوع الأخلاق (Ethics)، إلا أنها غالبًا ما تستخدم بالتبادل مع "الأخلاق المعيارية" (ethics). على الرغم من الاستفسارات أو التحليلات الفلسفيه التي قام بها فلاسفة أخلاقيون فيما يتعلق بالأخلاق (morality) (القيم الأخلاقية السائدة في المجتمع) – التحليلات التي غالبًا ما تؤدي إلى مواقف أو استنتاجات متنوعة – إلا أن السمات الأساسية، والعناصر الأساسية لأخلاقيات المجتمع، تظل هي المبادئ والقيم الأخلاقية التي توجه وتؤثر في حياة الناس كما هي إلى حد كبير. ما يحاول الفلاسفة الأخلاقيون القيام به هنا من خلال تحليلاتهم وحججهم النقدية هو شرح المفاهيم والقضايا الأخلاقية، وتوضيحها. على الرغم من أن المعتقدات والظروف الأخلاقية لمجتمعاتهم تشكل المحور المباشر لأنشطتهم الفلسفية – لأن التجربة الإنسانية محسوسة مباشرةً في سياق اجتماعي أو ثقافي معين – إلا أنهم لم يفكروا أو يقترحوا أن نتائج أنشطتهم العاكسة يجب أن تكون مرتبطة بمجتمعاتهم على هذا النحو. إنما يعتقدون على العكس من ذلك أنه في ضوء إنسانيتنا المشتركة التي تتحدث عن المشاعر، والأهدإف، والاستجابات، والآمال، والتطلعات المشتركة لجميع البشر فيما يتعلق بمواقف معينة، فإن استنتاجات تأملاتهم ستكون لها بالتأكيد آثار على المجتمع البشري الواسع، وللأسرة البشرية العالمية.[٣] وهنا نلاحظ أن المصطلحين "أخلاقي" و"أخلاقي" "moral" and "ethical" غالبًا ما يتم استخدامهما بالتبادل في الخطاب العادي، وكذلك في الكثير من الأدبيات حول أخلاقيات الآلة[٤]، لذلك ميز بعض الفلاسفة اختلافًا حادًا بين الاثنين[٥]، ولكن لا توجد طريقة واحدة غير مثيرة للجدل لرسم مثل هذا التمييز[٦]. ومن أجل البساطة؛ فإننا نتبع الاستخدام المعتاد، ونستخدم

المصطلحين بالتبادل عندما تكون هناك حاجة إلى توضيح مزيد من الفروق الدقيقة، فسيتم تقديمها بشكل صريح.

ما "الآلة الأخلاقية"؟

نهدف هنا إلى توضيح بعض المصطلحات الأساسية، وهذا بدورة سيحدد أربع قضايا بارزة ناشئة عن أخلاقيات الآلة، والتي ستنظم مناقشتنا في بقية البحث.

تم التعبير عن أهداف أخلاقيات الآلة بعدة طرق بما في ذلك القدرة على بناء الآلات وإنشائها، والتي توصف بـ "الأفراد الأخلاقيين المصنعين"، والتي يمكن أن تتبع المبادئ الأخلاقية[٧]، وتكون قادرة على اتخاذ قرارات أخلاقية، ولها "بعد أخلاقي"، أو التي تكون قادرة على فعل الأشياء التي تتطلبها الأخلاق لدى البشر[٨].

قام جيمس إتش مور (**J. H. Moor**[*]) بتمييز دقيق بين أنواع مختلفة من "الآلات الأخلاقية" التي قد تسعى إليها أخلاقيات الآلة. **أولاً**– "العوامل الأخلاقية الضمنية" بمعني تحجيم الآلات لتعزيز السلوك الأخلاقي، أو على الأقل تجنب السلوك غير الأخلاقى. **ثانيًا**– "الأفراد الأخلاقيون الصريحون" القادرون (بمعنى ما) على تمثيل الفئات أو المبادئ الأخلاقية. **ثالثًا**– الآلة هي " فاعل أخلاقي كامل" إذا كانت قابلة للمقارنة في العديد من النواحي ذات الصلة بصانعي القرار "الأخلاقيين البشريين".[٩]

[*] **جيمس إتش مور J. H. Moor** هو أستاذ دانيال بي ستون للفلسفة الفكرية والأخلاقية في كلية دارتموث. حصل على الدكتوراه. في عام ١٩٧٢ من جامعة إنديانا. وكتب مور عام ١٩٨٥ بحث بعنوان "ما هي أخلاقيات الكمبيوتر؟" جعله كواحد من المنظرين الرائدين في مجال أخلاقيات الكمبيوتر. وقد كتب أيضًا على نطاق واسع في اختبار تورينج. تشمل أبحاثه دراسة فلسفة الذكاء الاصطناعي وفلسفة العقل وفلسفة العلم والمنطق.
https://philosophy.dartmouth.edu/people/james-h-moor
تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)

على الرغم من أن هذه التعريفات توفر نقطة انطلاق جيدة؛ فإننا نجد بعض الغموض لا يزال قائمًا فيما يتعلق بكل من الكلمتين (أخلاقي" و "أخلاقي" " "ethical and“ moral”) في مصطلح "الأفراد الأخلاقيين".

أبدأ بمصطلح "أخلاقي" ethical” في مصطلح " العوامل الأخلاقية الضمنية "؛ يُستخدم مصطلح "أخلاقي" هنا ليعني "وفقًا للأخلاق": بأن الآلات الأخلاقية هي تلك التي يتماشى سلوكها بشكل صحيح مع المبدأ الأخلاقي ذي الصلة؛ لأن الأمر هنا سيكون أكثر إثارة للجدل حول تحديد المبادئ التي يجب على الآلة اتباعها كما نستكشف في ما يلي: تتعارض الآلات الأخلاقية بهذا المعنى مع الآلات غير الأخلاقية مثل أجهزة الصراف الآلي المصممة لسرقة التفاصيل المصرفية للمستخدمين.

وعلى النقيض من ذلك ففي مصطلح " الأخلاقيون الصريحون " يتم استخدام كلمة "أخلاقي" بشكل مرادف للسمات المميزة للآلات الأخلاقية، مما يعني أنهم قادرون على التفكير في الأخلاق أو المبادئ الأخلاقية، وهذا يتناقض مع الآلات المستخدمة بطريقة غير أخلاقية (amoral machines) مثل السيارة التي تحتوي على ميزات أمان مدمجة مثل حزام الأمان، ولكنها لا توضح في حد ذاتها ما يجب أن تكون عليه. للتمييز بوضوح بين هذين المعنيين "الأخلاقيين" نقترح بدلاً من ذلك التمييز بين الآلات المتوافقة أخلاقيًا مما يعكس مصطلحات المبادرة العالمية حول أخلاقيات الأنظمة المستقلة، والتصميم الذكي المتوافق أخلاقيًا (رؤية لإعطاء الأولوية لرفاهية الإنسان بأنظمة مستقلة وذكية)[١٠]. أي الآلات التي تعمل بطريقة مرغوبة أخلاقيًا، أو على الأقل مقبولة أخلاقيًا، والآلات التي يمكن أن يكون لديها القدرة على التفكير الأخلاقي(من خلال استخدام برامج أخلاقية إلكترونية تستهدف التفكير الأخلاقي من قبل المبرمجين). سنشرح هذين المعنيين "الأخلاقيين" بمزيد من التفصيل في المحاور التالية.

وبالمثل فإن مفهوم "وكالة الآلة" ومصطلح "وكيل" يحملان في طياتهما بعض الدلالات الفلسفية التي نعدها غير مفيدة لبناء أهداف أخلاقيات الآلة، وهنا يكمن جزء من القلق حول ما إذا كان تعيين "وكالة" للآلات سيؤدي إلى مواقف إشكالية حول حقوق الآلة، ومسؤولياتنا تجاهها.

أ- الآلات المتوافقة أخلاقيًا.

يعتمد وجود شئ ما متواقف من الناحية الأخلاقية على ما يعدُّ مرغوبًا ومقبولًا من الناحية الأخلاقية، وهذه بالطبع مفاهيم مثيرة للجدل في الأخلاق الفلسفية. علاوة على ذلك؛ فإن الحقيقة الأساسية للخطاب الأخلاقي والسياسي العام هي أن الناس يختلفون حول ما هو سلوك مرغوب فيه ومقبول من الناحية الأخلاقية. إذن لا يوجد سبب لافتراض أنه ستكون هناك إجابة واحدة شاملة على سؤال ما الذي يشكل آلة متوافقة أخلاقيًا؟ إنه أمر لا جدال فيه في جميع وجهات النظر الفلسفية، ناهيك عن الجمهور بشكل عام.

ومع ذلك فمن الناحية العملية، يمكن للمجتمعات أن تتوصل إلى توافق في الآراء، أو تتنازل عن المواقف التي يرغب الأفراد في قبولها باعتبارها جيدة بما يكفي للعمل الجماعي حتى لو اختلفوا من حيث المبدأ. غالبًا ما يتم التوسط في ذلك من خلال الهياكل الاجتماعية والمؤسسية (مثل المحاكم، والتصويت، والوساطة، وعمليات التشاور العامة، وما إلى ذلك) التي يحترمها الناس كطرق مشروعة لحل النزاعات. وهنا يتضح لنا أن الفلسفة السياسية تحتوي على عدة نظريات حول كيفية تحقيق التنازلات الجماعية بشكل شرعي في مواجهة الخلاف العميق والواسع النطاق[١١].

هنا أود أن أشير إلى أنني لن أقدم مراجعة شاملة للنظريات الحالية للشرعية السياسية، أو أظهر أيها ينطبق بشكل أفضل على أخلاقيات الآلة. لذلك سوف نعتمد على الاقتراح التالي كفكرة إرشادية عامة، وهي أن الآلات المتوافقة أخلاقيًا: هي تلك

التي يدعم سلوكها بشكل كافٍ، ويعزز بشكل مثالي مصالح، وقيم أصحاب المصلحة المشاركين في سياق معين.

هناك بالطبع العديد من الأسئلة الصعبة وراء تفسير هذه الصيغة مثل ما "القيمة" أو "المنفعة"؟ هل كل القيم والمصالح متساوية في الأهمية؟ من أصحاب المصلحة المعنيين؟ هل يشمل البشر فقط؟ ماذا يعني "الحفاظ بشكل مناسب" على قيم مختلف لأصحاب المصلحة بالنظر إلى أن هذه القيم غالبًا ما تتعارض؟ ستعطي النظريات الأخلاقية المختلفة إجابات مختلفة، ولا نقترح أي حل لهذه الأسئلة هنا. بدلاً من ذلك عند تبني المفهوم الإرشادي للآلات المتوافقة أخلاقيًا نأمل ببساطة في إعطاء فكرة عما قد تتضمنه "آلة متوافقة أخلاقيًا "، وتسليط الضوء على بعض القضايا الخلافية التي قد تكون ذات صلة بمناقشات التوافق الأخلاقي.

ب- التفكير الأخلاقي.

الاستدلال كما نفهمه هنا هو معالجة المعلومات من أجل إيجاد حل لمشكلة ما. لذلك من الممكن التمييز بين أنواع التفكير من حيث أنواع المشاكل التي يعالجها. هذا يسمح لنا بتعريف التفكير الأخلاقي كعمليات تدخل في حل المشاكل الأخلاقية المختلفة. بهذه الطريقة يمكن تمييز التفكير الأخلاقي عن التفكير الرياضي الذي يتعامل مع المشكلات الرياضية، أو التفكير الواقعي حول العالم التجريبي. قد يكون من الصعب رسم خط واضح بينهما، ولكن يكفي أن نكون قادرين على التمييز بشكل أساسي بين المشاكل الأخلاقية، مثل " هل يجب أن أقتل المريض لتخفيف آلامه إذا طلب مني ذلك؟ أم " هل هذا الدواء يقتل هذا المريض"؟ وهنا لا يوجد في الغالب أي نوع من الواقعية. لذلك سنناقش بعض الأسئلة الإضافية التي أثارها هذا التعريف للتفكير الأخلاقي.[١٢]

عندما نتحدث عن (القدرة) على التفكير الأخلاقي في الآلات؛ فإن السؤال الذي يطرح نفسه هو ما إذا كانت الآلات قادرة على التفكير أم لا؟ في بعض السياقات يتم استخدام "التفكير" بمعنى مطلب، والذي يتضمن واحدًا، أو أكثر من القدرات المحددة مثل التفكير الواعي، أو النية، أو "معرفة السبب"، أو "الانفتاح على العالم"، أو فهم الأهمية؛ فإذا كان بإمكان الآلة أن تمتلك مثل هذه القدرات فهذا بالطبع اعتراض معروف على برنامج ذكاء اصطناعي قوي(١٣). من ناحية أخرى يستخدم مصطلح "التفكير" أيضًا بشكل شائع خاصة في أبحاث الذكاء الاصطناعي بمعنى أوسع ليعني ببساطة أي معالجة يتم إجراؤها للوصول إلى نتيجة، ولأن أبحاث أخلاقيات الذكاء الاصطناعي هي فرع معاصر من الفلسفة، وتحديداً ضمن تخصص "أخلاقيات التكنولوجيا"، وتهتم بالقضايا الأخلاقية المرتبطة بالروبوتات وأنواع مختلفة من الذكاء الاصطناعي. فهي تستند إلى دراسة جانبين مرتبطين بالقيمة الأخلاقية المتعلقة بهذا المجال: الأول هو علاقة الآلة بالإنسان، والآخر هو علاقة الإنسان بالآلة.(١٤)

يتعلق الجانب الأول وهو العلاقة بين الإنسان والآلة بمسألة الطرق التي يمكن أن تكون بها الآلات مفيدة أو ضارة للبشر، وهل يمكن للروبوتات، أو ينبغي أن يكون لها مبرر أخلاقي؟ في هذه الحالة ما السلوك الأخلاقي الذي يجب على الروبوتات اتباعه؟ كيف يمكن استخدام الآلات لإيذاء البشر؟ ما السبل الممكنة لتجنب هذا الخطر؟ وبالعكس كيف يمكن الاستفادة منها في خدمة البشرية ونفعها؟

بينما الجانب الآخر من أخلاقيات الذكاء الاصطناعي، أي الجانب المعني بعلاقة الإنسان بالآلة، يهتم بطريقة وهدف استخدام الآلات، أي أنه يهتم بأسئلة مثل: كيف يصمم الإنسان الآلة؟ كيف يبنيها؟ كيف يتعامل معها؟ والأهم من ذلك، ما الذي تستخدمه؟ هل للآلة حقوق وكذلك هل عليها واجبات.(١٥)

هذا هو المعنى المستخدم عند إسناد الاستدلالات الضمنية، أو اللاواعية إلى البشر، أو عند الحديث عن الاستدلال الآلي بواسطة الأنظمة الحسابية. قد يجادل

البعض بأن التفكير الأخلاقي إذا تم فهمه بشكل صحيح لا يمكن إلا أن يفكر بالمعنى الأكثر تطلبًا؛ لأن التفكير الأخلاقي يتطلب فهمًا لأهمية القضايا الأخلاقية المطروحة. لذلك أري أن هذا يتعلق بمسائل الحقوق والمسؤوليات التي يثيرها جيه إتش مور (**J. H. Moor**) تحت عنوان "وكلاء أخلاقيون كاملون"، ومع ذلك لا يزال بإمكاننا التساؤل عما إذا كانت الآلات قادرة على التفكير الأخلاقي بالمعنى الثاني الأضعف. نظرًا لأن هذا هو السؤال الذي يهتم به علماء أخلاقيات الآلة؛ لأننا نستخدم مصطلح "التفكير الأخلاقي" بهذا المعنى الأضعف من الآن فصاعدًا، ما لم يُذكر خلاف ذلك.

قد يكون هناك اعتراض آخر، وهو أن أي قرار يمكن تفسيره على أنه حل لمشكلة أخلاقية تقريبًا، سيؤدى إلى تقليل أهمية فكرة التفكير الأخلاقي، وهذا يعني أنه عندما تقوم الآلة بتقديم حلول لأي مشكلة أخلاقية فهذا يعنى تهميشًا لقدرة الإنسان على التفكير الأخلاقي. للتوضيح لنفترض أن آلة مبرمجة لمراقبة المرضى عبر الكاميرا، واستخدم هذه المعلومات لاستنتاج مستويات السكر في الدم، وتزويدهم بالأنسولين إذا وصلت مستويات السكر في الدم إلى نسبة محددة مسبقًا. من الواضح أن هذا النظام يتخذ قرارات التبعية الأخلاقية، ولكن هل يُظهر التفكير الأخلاقي؟ نحن نعدُّ هذه حالة محدودة تنطوي على التفكير الأخلاقي فقط إلى درجة غير مهمة حيث تشير أهمية عملية التفكير إلى صعوبة حل مشكلة بواسطة آلة مقارنة بالموارد، والمدخلات الموجودة تحت تصرفها. وهكذا في هذا المثال تعمل الآلة على حل مشكلة مهمة في العالم الحقيقي من خلال استنتاج مستويات السكر في الدم من إدخال الفيديو، بينما "المشكلة" الأخلاقية لتطبيق قاعدة قرار واحدة لا لبس فيها، على سبيل المثال "إذا وصلت مستويات السكر في الدم لدى المريض إلى المستوى المطلوب T، ثم تم إعطاؤه الأنسولين "أمر بسيط". على النقيض من ذلك ضع في اعتبارك روبوت الرعاية الصحية، كما وصفه أندرسون وآخرون[١٦]، يتم استخدام هذا التعلم الخاضع للإشراف

لاستنتاج قاعدة عندما يكون واجب حماية صحة المريض يجب أن يفوق واجب عدم التعدي، وانتهاك استقلاليته من خلال إدارة طب الوالدين، وبالنظر إلى المدخلات؛ فإن استنتاج قاعدة القرار التي توازن بشكل كاف بين هذين الواجبين الظاهريين ضد بعضهما البعض يمثل – أمر إشكالي للغاية– مشكلة كبيرة.

يتضمن أخيرًا مثال Anderson et al. هل يجب علينا أيضًا تصنيف المناهج "من أعلى إلى أسفل" (عادةً ما تكون رمزية، أو قائمة على المعرفة) باعتبارها طرقًا للتفكير الأخلاقي لأخلاقيات الآلة، حيث يتم برمجة الآلة لاستنتاج الآثار المترتبة على المبادئ الأخلاقية الأكثر عمومية لسياقات معينة، أو لحلها تتعارض بين مبادئ الوجوه المتعددة الواضحة؟ مرة أخرى، يسمح لنا إطار العمل الخاص بنا بتمييز هذا عن مجرد تطبيق قاعدة قرار مباشر، حيث تحتاج الآلة إلى حل المشكلات الصعبة من أجل اشتقاق مبدأ عمل محدد بالكامل من عدة مبادئ عالية المستوى قد تكون متضاربة. ما إذا كان يمكن أخيرًا رسم أي حدود حادة بين مجرد تطبيق قاعدة القرار، واستنتاج الآثار المترتبة على مبدأ أكثر عمومية ليس ذا أهمية حاسمة هنا. النقطة المهمة هي ببساطة أنه يمكننا تمييز حالات من المشكلات الأخلاقية التافهة إلى المشكلات الأخلاقية الأكثر أهمية (أي صعوبة النظر إلى المدخلات)، لذلك فإن اهتمامنا هنا هو بناء آلات قادرة على حل المشكلات الأخلاقية.[١٧]

ج – "وكالة" الآلة

يُستخدم مصطلح "وكيل" في مجالات مثل الروبوتات، والتعلم الآلي، والذكاء الاصطناعي في غالب الأمر للإشارة إلى قدرته على التصرف في حل المشكلات، ومعالجة المعلومات والبيانات. فعلى سبيل المثال أحد التعريفات المؤثرة للذكاء الاصطناعي هو دراسة العوامل الذكية ذات القدرة على الإدراك والتصرف[١٨]. يعد هذا

استخدامًا أوسع من التخصصات الأخرى مثل الفلسفة، ولكن ليس كل شخص في علوم الكمبيوتر راضٍ عن هذا الاستخدام الواسع، وهناك نقاش حيوي حول الشروط الأخرى التي قد تحتاجها الآلات للوفاء بها لتكون وكالة محسوبة بمعني أنها تكون مسئولة عن أفعالها[١٩].

على الأقل في الفلسفة الغربية المعاصرة، يتطلب الحساب القياسي للفاعلية القدرة على الأفعال المتعمدة. لذلك يعد الفعل مقصودًا عندما يكون ناتجًا عن الحالات العقلية المتعمدة للفاعل، على سبيل المثال، معتقداته أو رغباته[٢٠]. تتميز الأفعال المتعمدة عن السلوكيات المجردة، والتي لا تفترض مسبقًا أي نوايا. هناك مفهومان مختلفان على الأقل عن القصدية: ١) إحساس "واقعي" أقوى، والذي يصعب نسبته إلى الآلات؛ و ٢) ضعف "العازف" الذي يسمح بنسب أكثر وضوحًا. بالمعنى الواقعي، يتطلب الفعل المتعمد بعض الخصائص التي ذكرناها فيما يتعلق بالحس المطلوب للتفكير، مثل القدرة على الفهم والوعي الهائل. على سبيل المثال من غير المرجح أن تمتلك آلة بسيطة مثل Roomba القدرة على الفعل المتعمد بهذا المعنى القوي؛ لأنها تفتقر إلى نوع المعتقدات والرغبات الواعية الحقيقية التي يمتلكها البشر. تعمل Roombas على تصنيفات تستند إلى السمات النحوية والإدراكية بدلاً من الفئات الدلالية، وبالتالي ليس لديها فهم لما يفعلونه[٢١]. من ناحية أخرى؛ فإن الإحساس الأداتي بالنية سواءُ أكانت المعتقدات أم الرغبات يمكن أن تُنسب إلى كيان ما يعتمد كليًا على مدى فائدة مثل هذه الأوصاف في تفسير سلوكها [٢٢]. وفقًا لوجهة النظر هذه إذا كان من المفيد إسناد المعتقدات والرغبات إلى Roomba من أجل شرح سلوكها، فإن هذه الفائدة تكون كافية للقيام بذلك مما يجعل هذا الرأي من المعقول نسب الكفاءة المقصودة أو عزوها إلى الآلات[٢٣].

هذه الحجة مهمة للفلاسفة بسبب التقليد طويل الأمد القائل بأن القصدية (النية) هي السمة المميزة للعقلية[٢٤]، وبالتالي ترتبط صفات الفاعلية المقصودة بأسئلة حول الحياة العقلية، بما في ذلك عمليات تفكير الفاعل ووعيه وإرادته الحرة[٢٥]. بالإضافة إلى القصد، غالبًا ما يُعتقد أن العوامل الأخلاقية قد تتطلب بعض الشروط الإضافية مثل القدرة على التصرف بطريقة توضح فهم المسؤولية تجاه الأفراد الآخرين، أو القدرة على مراقبة سلوكهم في ضوء أخلاقهم. الواجبات والأضرار المتوقعة التي قد تسببها أفعالهم[٢٦].

وطبقًا لذلك فإن ما يمكن أن تسميه الآلات بالعوامل الأخلاقية بأي معنى قوي هو سؤال فلسفي مثير للجدل. في رأينا أنه من المهم أن نسأل عما إذا كانت الآلات تمتلك هذه القدرات؛ لأنها تتعلق بسؤالين أخلاقيين واضحين، وهما:

١) ما إذا كانت الآلات لديها مسؤوليات. و ٢) ما إذا كان لديها حقوق. كل من هؤلاء لديهم روابط لمفهوم الوكالة. فيما يتعلق بالمسئولية إذا كانت الآلة قادرة على فهم واجباتها، أو الأضرار المتوقعة لأفعالها، بالمعنى الواقعي الموصوف أعلاه، فقد يكون من المغري تحميلها المسؤولية عن أي أضرار تسببها. أما فيما يتعلق بالحقوق يعتقد البعض أن أي كائن له أهداف، أو رغبات له مكانة أخلاقية يجب احترامها[٢٧]. إذا كان من الممكن تخصيص حالات مقصودة لآلة، فقد يستلزم ذلك تحملنا مسؤولية احترام حقوقها[٢٨]. لكن يبدو أن كل هذا يفترض مسبقًا وجود إحساس أقوى بالفعالية المقصودة – والتي يصعب نسبتها إلى الآلات. ما مدى صعوبة ذلك إذا كان أي نظام، أو روبوت حالي، أو مستقبلي للذكاء الاصطناعي سيتأهل على أنه يمتلك وكالة مقصودة بهذا المعنى الأقوى؟ هذا يعد أمرًا مثيرًا للجدل. ومع ذلك يتفق العديد من الفلاسفة على أن مجرد وجود القصدية بالمعنى الذرائعي (الأداتي) لا يكفي لتأسيس أي حقوق أو مسؤوليات مهمة.[٢٩]

قد يبدو من البديهي أن الآلات مستثناة من مجال الأخلاقيات. في الواقع هناك روبوتات مصممة لإدراك العواطف والتصرف وفقًا لها. لكن ما الذي يعنيه أن نقول إن "ربوتا " قادر على إدراك العواطف والإحساس بها؟ تثير مسألة "تعلم الآلة Machine * (Learning)" قضايا أخلاقية بسبب إمكانية تتبعها، تكرارها، وتنميطها لـ (سلوكيات) الأفراد. تبعًا لكل هذا، هل يمكننا أن نصف" الذكاء الاصطناعي " بكونه وكيلًا أخلاقيًا في حد ذاته؟ يقودنا الحس السليم إلى ربط الأخلاق ب" الفاعل الإنساني " الذي يملك القدرة على استخدام هذه التقنيات الجديدة من عدمه.(٣٠)

وهذا يوضح لنا أن مصطلح "العامل الأخلاقي" سيحمل حتمًا دلالات معقدة، وأن هدفنا الأساسي هنا هو إبراز التعقيدات بدلاً من حلها. في حين أن الأسئلة حول ما إذا كانت الآلات يمكن أن يكون لها مسؤوليات أو حقوق أخلاقية هي أسئلة مهمة، وستتم مناقشتها بشكل أكبر على النحو التالي؛ لأننا نعتقد أنه من غير المفيد بناء هذه الدلالات في تحديد أهداف أخلاقيات الآلة. لذلك سنتحدث عن الآلات، بدلاً من الوكلاء، وسنطرح أسئلة حول الحقوق والمسؤوليات بشكل منفصل.

حقوق الآلة

يشير مصطلح "حقوق الآلة" إلى التزام البشر الأخلاقي تجاه الآلات التي يمتلكونها. ربما يكون للروبوت، إذا تم تطويره بشكل كافٍ، الحق في الوجود تمامًا كما يحق للإنسان أن يعيش. تضمن بعض المؤسسات التي ترعى أبحاث الذكاء الاصطناعي حقوقًا أخرى للآلات أيضًا إذا وصلت إلى درجة عالية من الوعي، وهو ما لم يحدث بالطبع بعد، لكنه ممكن، مثل حرية التعبير والمساواة أمام القانون.(٣١)

* تعلم الآلة (Machine Learning) هو أحد فروع الذكاء الاصطناعي التي تهتم بتصميم وتطوير خوارزميات وتقنيات تسمح للحواسيب بامتلاك خاصية «التعلم». بشكل عام هناك مستويين من التعلم: الاستقرائي والاستنتاجي. يقوم الاستقرائي باستنتاج قواعد وأحكام عامة من البيانات الضخمة. تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)https://en.wikipedia.org/wiki/Machine_learning

يجادل بعض الباحثين بأن هذا مرتبط بواجب الروبوت الأول لخدمة البشر؛ لأن الحقوق مثل الحق في الوجود، والحق في القيام بعمل المرء يجب أن تكون أولاً متسقة مع واجباته تجاهنا كبشر، ويجب أن تتوافق حقوق مثل الحق في الحياة، والحق في العمل مع واجباته تجاه المجتمع البشري ككل.

على سبيل المثال، يعتقد عالم الكمبيوتر ومؤسس برنامج "ELISA" الشهير "جوزيف وايزنباوم*"Joseph Weizenbaum" أنه لا ينبغي أبدًا استخدام الذكاء الاصطناعي كبديل للبشر في وظائف معينة، مثل خدمة العملاء، والعلاج النفسي، ورعاية المسنين، والأمن، والقضاء والشرطة؛ لأن هذه المهن تتطلب درجة من الرعاية والاحترام. لا يستطيع الذكاء الاصطناعي توفيرها، وبالتالي لا يمكنه أداء هذه الوظائف، لأن الإحساس الجوهري بالعاطفة والإيثار يلعبان دورًا مهمًا فيها.(٣٢)

ويرتبط هذا الرأي بموقف وايزنباوم نفسه المتشكك في إمكانات الذكاء الاصطناعي. ولكن إذا كان مشروع "الذكاء الاصطناعي القوي" ممكنًا أخيرًا، واكتسبت

* جوزيف وايزنباوم Joseph Weizenbaum: ولد في برلين، ألمانيا يهودي هرب والديه إلى ألمانيا النازية في يناير ١٩٣٦، هاجر مع عائلته إلى الولايات المتحدة الأمريكية. بدأ دراسة الرياضيات عام ١٩٤١ في جامعة واين ستيت، في ديترويت، ميشيغان. في عام ١٩٤٢، توقف عن دراسته للخدمة في سلاح الجو بالجيش الأمريكي كخبير أرصاد جوية، بعد أن رُفض بسبب عمله في علم التشفير بسبب وضعه "أجنبي عدو". بعد الحرب، في عام ١٩٤٦، عاد إلى واين ستيت، وحصل على شهادة البكالوريوس. في الرياضيات عام ١٩٤٨، وماجستير في العلوم. في عام ١٩٥٠. حوالي عام ١٩٥٢، عمل فايزنباوم كمساعد باحث في واين أجهزة الكمبيوتر التناظرية وساعد في إنشاء جهاز كمبيوتر رقمي. في عام ١٩٥٦ عمل لدى جنرال إلكتريك على إرما، وهو نظام كمبيوتر أدخل استخدام الخطوط المشفرة مغناطيسيًا المطبوعة على الحد السفلي للشيكات، مما يسمح بمعالجة الشيكات تلقائيًا عبر التعرف على الحروف الحبر المغناطيسي (MICR). في عام ١٩٦٤ تولى منصبًا في معهد ماساتشوستس للتكنولوجيا. تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧) https://upwikiar.top/wiki/Joseph_Weizenbaum

الروبوتات صفة الوعي والشعور، ففي هذه الحالة يكون لديه جميع الحقوق التي يتمتع بها البشر، وهذه حجة خطأ.

إذا كان على الذكاء الاصطناعي أن يعمل وفقًا لنموذج أخلاقي معين، فما النموذج الذي سيكون؟

يرتبط الجدل الحالي حول حقوق الإنسان الآلي بالنقاش الحالي حول حقوق الحيوان، حيث إن الذكاء الاصطناعي لم يصل بعد إلى مستوى الذكاء البشري. على سبيل المثال، عندما أصدرت شركة بوسطن ديناميكس لصناعة الروبوتات مقطع فيديو يظهر الموظفين وهم يركلون بقعة لاختبار توازنها، أصدرت PETA بيانًا وصفت السلوك بأنه "غير مقبول".(٣٣)

إذن، فمسألة حقوق الآلة تتعلق بأكثر من عامل، من بينها قدرة الذكاء الاصطناعي على اكتساب صفة الوعي والشعور، وهل له حقوق، وإن كان دون درجة عالية من الوعي والشعور، كما ندافع على سبيل المثال، عن حقوق الحيوان، وأخيرًا ما إذا كانت حقوق الآلة تعتمد في النهاية على منفعه للإنسان، أم مستقلة عن خدمتها لهذه المنفعة، وهنا نجد الآراء مختلفة في الإجابة عن كل سؤال.

واجبات الآلة

منذ اندلاع الثورة الصناعية من قبل القوى العاملة التي تشغل الآلة، تم استخدام الآلات لتسهيل العمل، وهو ما يختلف عن الذكاء الاصطناعي. حيث إن العمل في هذه الحالة يتم بواسطة الآلة فقط، دون تدخل بشري، أي دون الحاجة إلى العمالة. هذا يعني أن الآلة الذكية ستكون أكثر استقلالية بشكل ملحوظ من أي آلة أخرى، وهذه الحقيقة تثير أسئلة مهمة حول مسؤولياتها الأخلاقية وقدرتها على تمييز الصواب من الخطأ.(٣٤)

فعند التفكير في إطار القيم الأخلاقية الذي يجب أن يتبناه الذكاء الاصطناعي، ويبني الروبوتات وفقًا له، للعمل ضمنه، واكتساب بعض المنطق الأخلاقي، تظهر هنا مشكلة أخرى مهمة جدًا تتعلق بمسألة الأخلاق نفسها وفلسفتها.

فلو كان الذكاء الاصطناعي يعمل وفقًا لنموذج أخلاقي معين، فما النموذج الذي سيكون؟ تختلف النماذج والقيم الأخلاقية باختلاف الثقافات والشعوب والأديان، وحتى داخل الشعب الواحد قد تختلف من وقت لآخر ومن مجموعة إلى أخرى، ناهيك عن الاختلاف بين الأفراد أنفسهم.

وهذا ما يميز القيمة الأخلاقية عن القيمة المعرفية أو العلمية. إذا كانت الأخيرة تستند إلى الواقع المباشر، فهذا يجعلها معرفة "موضوعية". أما الأخلاق فقد تكون نسبية، ولا تستند إلى حقائق موضوعية، والدليل على ذلك هو الاختلاف بين الناس فيما بينهم، واختلاف الفلاسفة الأخلاقيين في رؤيتهم للصواب والخطأ. في هذه الحالة، ما النموذج أو النظرية الأخلاقية التي يجب برمجة الذكاء الاصطناعي عليها؟

قدمت الفلسفة الأخلاقية عددًا من النظريات التي يمكن أن يستفيد منها الذكاء الاصطناعي. لكنها نظريات تختلف وتتعارض مع بعضها البعض. لدينا الواجب الأخلاقي الكانطي، الذي يرى السلوك الأخلاقي الجيد كغاية في حد ذاته، وهذا يعني أن الغاية لا تبرر الوسيلة بأي شكل من الأشكال. على النقيض من الأخلاق الكانطية، نجد العقيدة النفعية أن الغاية تبرر الوسيلة، وأن المنفعة الناتجة عن الفعل بمثابة معيار للحكم على أخلاقية هذا الفعل أم لا.[٣٥]

إذا كنا كبشر نواجه مشكلة مع وفرة النماذج والنظريات الأخلاقية، وافتقارنا إلى اتفاق على أحكام أخلاقية نهائية، فكيف يجب أن تصنع الروبوتات بطريقة تجعلها "أخلاقية"؟

وبناءً على ما سبق يمكننا إعادة صياغة القضايا الرئيسة التي أبرزها إطار عمل إتش مور **(J. H. Moor)** على النحو التالي:

١) بناء آلات متوافقة أخلاقيًا .

٢) بناء آلات يمكن أن يكون لديها القدرة على التفكير الأخلاقي.

٣) بناء آلات لها: ١) مسؤوليات أخلاقية أو ٢) حقوق.

نحن نعتقد أن هذه الأمور تستحوذ على القضايا الرئيسة في مشروع محاولة بناء "آلات أخلاقية". إلى الحد الذي يكون فيه: بناء آلات متوافقة أخلاقيًا ممكنًا، من المفترض أن يكون هذا شيئًا يجب أن يهدف إليه أي تخصص هندسي: يجب تصميم أجهزة الصراف الآلي وخوارزميات مقارنة الأسعار بحيث لا تتحايل على المستخدمين والسيارات بحيث يكون لديهم مستوى مقبول أخلاقيًا من السلامة على الطرق، وهكذا في بعض الأحيان، تُستخدم "أخلاقيات الآلة" بمعنى واسع لتعني مشروع بناء آلات مستقلة أخلاقية. أما أخلاقيات الآلة بالمعنى الضيق الذي نهتم به هنا تكون مميزة من حيث إنها تسعى إلى بناء آلات يمكن أن يكون لديها القدرة على التفكير الأخلاقي كوسيلة لتحقيق آلات متوافقة أخلاقيًا. ربما يكون روبوت الرعاية الصحية أفضل من الناحية الأخلاقية إذا كان بإمكانه استنتاج ما إذا كان تذكير المرضى بتناول أدويتهم سيكون رعاية أبوية غير ضرورية أم مجرد رعاية مناسبة ؟ تثير متابعة بناء الآلات التي يمكن أن يكون لديها القدرة على التفكير الأخلاقي بدورها أسئلة حول ما إذا كانت الآلات التي لها مسؤوليات أخلاقية أو الآلات التي لها حقوق ضرورية أو يمكن أن تكون أحد الآثار الجانبية للآلات التي لديها القدرة على التفكير الأخلاقي، ومن هذه الأسئلة هل هناك نقطة تتطلب منا القدرة على التفكير الأخلاقي المعقد أن ندرك على سبيل المثال أن روبوت الدردشة، أو الشات بوت لديه حقوقه ومسؤولياته ؟[٣٦]

وهذا يدفعنا إلى النظر في كيفية إسهام التفكير الأخلاقي في التوافق الأخلاقي. وإلي بعض المخاطر المحتملة الناشئة عن ذلك، بما في ذلك الآثار الجانبية المحتملة التي تنطوي على الآلات التي لها مسؤوليات أخلاقية أو الآلات التي لها حقوق ضرورية.

المحور الثاني: دوافع أخلاقيات الآلة أو الماكينة

إن الدوافع الأخلاقية لمتابعة أخلاقيات الآلة. تعتمد على الادعاء بأن بناء الآلات ذات القدرة على التفكير الأخلاقي سيعزز ما نسميه بالتوافق الأخلاقي، وبالتالي مصالح وقيم أصحاب المصلحة المعنيين. فكما ذكرنا سابقا هناك مشاكل لم يتم حلها في توصيف التوافق الأخلاقي نابعة من حقيقة الخلاف المنتشر. ومع ذلك سنضعها جانبًا، ونركز على تقديم حجج إيجابية لمتابعة أخلاقيات الآلة. حيث توجد عدة طرق مناسبة للتغلب على الخلافات الأخلاقية العميقة بالفعل في بعض المجالات، ويبدو أنه من الممكن تطوير هذا لأخلاقيات الآلة أيضًا.

لذلك نبدأ بتقديم تمييز لبعض الطرق التي قد يؤدي بها إعطاء الآلات إمكانية القدرة على التفكير الأخلاقي إلى تعزيز التوافق الأخلاقي.

أولًا– نميز بين طريقتين لتحسين التوافق الأخلاقي للآلة:

أ – تحسين سلوك الجهاز نفسه.

ب– تحسين سلوك متخذي القرار البشري باستخدام الآلة.

ثانيًا– نميز بين معنيين يمكننا من خلالهما تحسين سلوك صانعي القرار الآلي أو البشري:

ج – تحسين القرارات الفردية.

د– تحسين كيفية ملائمة صانعي القرار للأخلاق كنظام اجتماعي أوسع.

من ناحية تحسين التوافق الأخلاقي للآلة يمكننا النظر إلى قرار فردي لآلة، أو إنسان، ونسأل عما إذا كان يتوافق مع معايير السلوك المرغوب فيه أخلاقيًا، أو المقبول (بغض النظر عن ما نعتبره). ولكن من ناحية تحسين كيفية ملائمة صانعي القرار

للأخلاق كنظام اجتماعي أوسع يمكننا أيضًا تقييم التوافق الأخلاقي لصانع القرار من حيث كيفية ارتباطه، وتفاعله مع صانعي القرار الآخرين. فالأخلاق ليست مجرد مجموعة من المعايير لسلوك الأفراد؛ إنها أيضًا نظام اجتماعي يعتمد فيه صانعو القرار على بعضهم البعض ويثقون فيه، حيث يمكنهم تقديم تفسيرات لأي طلب، وطلب إجراءات معينة، ويُمكنهم من تقديم الاعتذارات أو التعويضات عند ارتكاب الأخطاء. كما أوضحنا في تحسين كيفية ملائمة صانعي القرار للأخلاق كنظام اجتماعي أوسع، فإن من أعمق المخاطر الناشئة عن أخلاقيات الآلة سوف تتعلق بمسألة كيف يتلاءم صانعو القرار الآليون مع الأخلاق أو يغيرونها كنظام اجتماعي.[٣٧]

يمنحنا هنا الجمع بين هذين التمييزين تصنيفًا لأربع طرق لتعزيز التوافق الأخلاقي، وضمن كل ذلك قد نسأل أيضًا عن معيار التوافق الأخلاقي الذي نهدف إليه. على سبيل المثال قد نهدف فقط إلى تأمين التوافق الأخلاقي مع معيار مقبول على المستوى البشري، أي بالنسبة لمعيار ما قد نجده مقبولًا أخلاقيًا للإنسان. لاحظ أنه بالنسبة للبشر حتى هذا المعيار ليس تافهاً ؛ غالبًا ما يقع صانعو القرار البشريون دون المعايير التي نتوقعها منهم، سواء من خلال الصدفة أم الحقد أم عدم التفكير إلي معايير غير مقبولة. لكن يمكننا أيضًا أن نهدف إلى معايير أعلى بشكل متزايد للسلوك المرغوب على المستوى البشري. علاوة على ذلك اقترح البعض أن أخلاقيات الآلة يمكن أن تحسن التوافق الأخلاقي لصانعي القرار (الآلة أو الإنسان) بما يتجاوز المعايير البشرية الحالية.[٣٨] نقول المزيد حول ما قد يعنيه هذا في ما يلي، وكيف جادل مؤيدو إنشاء آلات أخلاقية بأن هذا قد يحسن المحاذاة وفقًا لكل فئة من الفئات الأساسية الأربعة المذكورة أعلاه. سننظر أيضًا في كيفية تحقيق ذلك وفقًا لمعايير مختلفة.

أ- قرارات الجهاز الفردية.

إن أخلاقيات الآلة الأكثر شيوعًا (بمعنى بناء آلات يمكن أن يكون لديها القدرة على التفكير الأخلاقي) تكون مدفوعة بأمثلة للأنظمة المستقلة التي يتم تطويرها حاليًا- على سبيل المثال السيارات ذاتية القيادة، أو الأسلحة المستقلة، أو روبوتات الرعاية الصحية – والتي ستتخذ قرارات تبعية أخلاقية. يجادل علماء أخلاقيات الآلة بأن إعطاء هذه الأنظمة القدرة على التفكير الأخلاقي سيساعد في ضمان أن قراراتهم تكون متوافقة أخلاقيًا[٣٩]. إذا كان الأمر كذلك، وإذا كنا سنشهد حتمًا نشر المزيد والمزيد من الأنظمة المستقلة، فسوف يوفر هذا سببًا أخلاقيًا لمتابعة أخلاقيات الآلة؛ لأنه سيعزز مصالح أصحاب المصلحة المعنيين وقيمهم.

فعلى الرغم من أن هذا الدافع معقول للوهلة الأولى، إلا أنه يجب ملاحظة أنه لا توجد علاقة ضرورية بين التفكير الأخلاقي (moral reasoning) والقرارات المتوافقة أخلاقيًا(ethically aligned decisions)، وذلك لأن:

أولاً- القدرة على التفكير الأخلاقي لا تضمن في حد ذاتها قرارات متوافقة أخلاقيًا، كما يبرهن البشر في كثير من الأحيان. بأن القوة الحسابية المحدودة، وعدم الدقة في المباني، أو بيانات التدريب المقدمة لمنطق الآلة، وقيود طبيعة الأخلاق نفسها قد تمنع الآلة من اتخاذ قرارات متوافقة أخلاقيًا، حتى ولو كانت لديها القدرة على التفكير الأخلاقي[٤٠]. لذلك فمن المفترض أن يتفق معظم علماء أخلاقيات الآلة على أن التفكير الأخلاقي لا يضمن التوافق الأخلاقي الكامل، وبناء عليه يعتمد الدافع وراء متابعة أخلاقيات الآلة على الادعاء الأكثر تواضعًا بأنه يمكن تحقيق تقدم تدريجي كافٍ لأخلاقيات الآلة لتقديم إسهام إيجابي في المواءمة الأخلاقية لصنع القرار الآلي[٤١].

فلو افتراضنا أن أخلاقيات الآلة يمكن أن تقدم إسهامًا إيجابيًا في المواءمة* الأخلاقية لقرارات الآلة، فلا يزال يتعين علينا التساؤل عما إذا كانت طريقة ضرورية، أو أكثر دقة، وفعالة من حيث التكلفة، مقارنة بالخيارات الأخرى. هناك العديد من السياقات التي تعمل فيها الآلات التي لا تملك القدرة على التفكير الأخلاقي، أو الآلات التي لا تحل سوى مشاكل أخلاقية تافهة في الواقع بطرق أخلاقية غير إشكالية. تشمل الأمثلة روبوتات المصانع، وآلات صرف النقود الآلية، أو قطارات المترو الآلية (٤٢). يتم تحقيق التوافق الأخلاقي في هذه الحالات من خلال تدابير السلامة المناسبة والقيود الخارجية على تشغيل الآلات. على سبيل المثال، لردع عمليات الاحتيال ستقوم أجهزة الصراف الآلي بتوزيع الأموال النقدية فقط في ظل ظروف معينة على سبيل المثال ببطاقة صالحة ورمز PIN، علاوة على ذلك فهي مقيدة بالمبلغ النقدي الذي يمكنها صرفه. إن القرار بأن هذه القيود مناسبة للآلات النقدية يتم بالطبع على التفكير الأخلاقي البشري، ولكنه لا يتطلب من الآلة حل أي مشاكل أخلاقية مهمة؛ إنه يتبع فقط هذه القواعد المحددة مسبقًا.

ومع ذلك يجادل علماء أخلاقيات الآلة بأنه عندما يُطلب من الآلات العمل بمرونة في بيئة معقدة سببيًا، فإن منحها القدرة على التفكير الأخلاقي يصبح أمرًا مهمًا. لتقييم هذه الحجة لاحظ أن مجرد التعقيد السببي – أي البيئات التي يمكن أن تتحد فيها مجموعة واسعة من العوامل السببية ذات الصلة في عدد من الطرق المفتوحة – لا يتطلب دائمًا القدرة على التفكير الأخلاقي. على سبيل المثال، قد يتعين على النظام المستقل المطور جيدًا إدارة العمليات المعقدة للغاية داخل مصنع آلي بعناية. ومع ذلك

* **المواءمة** هي التناسق، والائتلاف، لذلك يوجد لكلمة **المواءمة** عدة مرادفات هي الاتساق، والتناسق، والمجاراة، والمطاوعة، والمطابقة، والائتلاف، والمناسبة، والتألف، والموافقة، فعلى سبيل المثال يجب على الإنسان مواءمة العصر، أي يجب على الإنسان مجاراة العصر الذي يعيش فيه، وذلك من أجل الائتلاف، والاتساق داخل المجتمع.
تاريخ الدخول على الموقع (٢٠٢٢/٥/٢٧) https://mhtwyat.com

إذا كان الشاغل الوحيد ذي الصلة من الناحية الأخلاقية هو ضمان إغلاق جميع الآلات عند دخول البشر إلى أرضية الإنتاج، فلن يحتاج النظام إلى الانخراط في أي تفكير أخلاقي مهم. أثناء تحديد ما إذا كان الإنسان موجودًا يمكن أن يكون مشكلة واقعية معقدة، لا يلزم أي تفكير أخلاقي إضافي لتحديد ما إذا كان يجب إيقاف العمل بمجرد تحديد ذلك. طالما يمكن التعرف على البشر بشكل موثوق، ويمكن للمصممين تنفيذ مبدأ عمل محدد مسبقًا ومبرمج.(٤٣)

بدلاً من ذلك عندما تنتج البيئات المعقدة سببيًا ما يمكن أن نسميه التعقيد الأخلاقي تصبح القدرة على التفكير الأخلاقي مهمة. نعني بكلمة "التعقيد الأخلاقي" الحالات التي: أ) لا يستطيع الأخلاقيون البشريون صياغة مبدأ عام محدد تمامًا للعمل و ب) حيث يمكن لصانعي القرار بسبب التعقيد السببي مواجهة مجموعة مفتوحة من المواقف المتميزة أخلاقيًا، والتي لا يمكن ببساطة تعدادها مقدمًا. يمكن أن ينشأ التعقيد الأخلاقي لعدد من الأسباب، بما في ذلك ما يلي:

الحالة الأكثر وضوحًا هي عندما تتنافس عدة واجبات ظاهرة للعيان، مثل عند اتخاذ قرار بشأن إعطاء الأولوية لواجب حماية صحة المريض على واجب احترام استقلاليته. نظرًا لأن علماء الأخلاق البشرية يطلبون أحكامًا سياقية لحل المعضلات الناشئة عن مثل هذه المواقف، فإن هذا يخلق تعقيدًا أخلاقيًا عندما لا يستطيع مصممو الآلة التنبؤ مسبقًا بمجموعات العوامل في كل هذه المواقف المحتملة.(٤٤)

يمكن أن يظهر شكل آخر من أشكال التعقيد الأخلاقي عندما يؤدي التعقيد السببي إلى عدم اليقين. في مثل هذه الحالات قد يتعين على صانعي القرار أن يوازنوا بين مخاطر النتائج السلبية الكاذبة ومخاطر الإيجابيات الكاذبة. ما لم يتم تحديد جميع المواقف المحتملة التي قد تواجهها الآلة مسبقًا، فإن هذا بدوره يتطلب مقارنة "الوزن" الأخلاقي لهذه المخاطر، سواء من خلال تخصيص التسهيلات العددية أو بعض الوسائل الأخرى، والتي تتطلب بشكل معقول بعض القدرة على التفكير الأخلاقي المهم.

ليست كل المواقف المعقدة أخلاقيًا تنطوي على مبادئ متضاربة في حد ذاتها. على سبيل المثال يتطلب تحديد المبادئ أو الإجراءات التي سيتم تطبيقها في موقف معين القدرة على تحديد الجوانب الأخلاقية ذات الصلة بالموقف، وما إذا كان يحتوي على معضلات أو مقايضات. في بعض الأحيان، قد تواجه الآلة بيئة تجعل فيها التعقيد السببي من غير المهم عزل الجوانب الأخلاقية ذات الصلة بالموقف. نظرًا لعدم وجود مبدأ عام محدد تمامًا لتحديد الجوانب ذات الصلة للموقف، فإن القدرة على هذا النوع من التفكير الأخلاقي قد تؤدي إلى التوافق الأخلاقي في هذه الحالات.[٤٥]

ب- ملاءمة الآلات مع النظام الأخلاقي.

لو تمكنا من التأكد من أن القرارات الفردية للآلة متوافقة أخلاقيًا مع المعيار البشري، فسيظل الكثيرون مترددون في السماح لها باستبدال عملية صنع القرار البشري، وذلك إذا كانت الآلة غير قادرة على تفسير قراراتها وتبريرها. لذلك سلط العديد من العلماء الضوء على مفهوم "الذكاء الاصطناعي القابل للتفسير" باعتباره مهمًا لإنشاء آلات جديرة بالثقة وخاضعة للمساءلة[٤٦].

وهذا يكون مثالًا على كيف يمكن بناء آلات لديها القدرة على شرح أسباب عملها أن تحسن علاقتها بالأخلاق كنظام اجتماعي. إنه جزء مهم من الأنظمة الاجتماعية الأخلاقية البشرية التي يمكننا تقديمها، والسؤال عن الأسباب الكامنة وراء قراراتنا، بما في ذلك على وجه الخصوص أسبابنا الأخلاقية مثل القيم التي أعطيناها الأولوية في موقف معين. على سبيل المثال قد نعطي الأولوية لمنع الضرر في سيناريوهات معينة، أو احترام الاستقلال في سيناريوهات أخرى. إن شرح سبب اعتقادنا أن القرار كان مبررًا أخلاقيًا، حتى لو كان بأثر رجعي فقط، يسمح لنا بتحدي بعضنا البعض، والاقتناع بأن الشخص الآخر كان على حق بعد كل شيء، أو المطالبة باعتذار أو تعويض عندما تكون الأسباب المقدمة غير مرضية. للمشاركة في هذه الجوانب من أنظمتنا الأخلاقية ستحتاج الآلات إلى القدرة على تمثيل وتوصيل التفكير

الأخلاقي في شكل مفهوم للبشر. علاوة على ذلك، لو أردنا لهذه التفسيرات أن تكون أكثر من مجرد قصص "فقط"، فيجب أن تعكس إلى حد ما على الأقل عمليات التفكير الفعلية وراء قرارات الآلة.[٤٧]

يمكن القول هنا إن امتلاك الآلات القدرة على شرح أسبابها لا يكفي لمشاركتها بشكل مناسب في هذه الجوانب الاجتماعية للأخلاق. نظرًا لأن الآلات تفتقر إلى القدرة على الشعور بالندم، فقد يتساءل البعض عما إذا كان بإمكانها تقديم اعتذارات حقيقية أم لا؟ لاحظ، مع ذلك أن أفراد المجموعة، مثل الدول أو الشركات، يبدو أحيانًا أنهم يعتذرون عن أفعالهم، فعلى الرغم من أنهم يفترض أنهم لا يشعرون بالندم أيضًا. إلا أنه يمكن تفسير ذلك من خلال حقيقة أن أفراد المجموعة لديهم سمات أخرى ذات صلة تفتقر إليها الآلات، فهم يمتلكون ممتلكات يمكن تقديمها كتعويض، ويتكونون من أفراد بشريين، قد يشعرون بالندم نيابة عن فرد المجموعة. لن نتابع هنا التشابه بين أفراد المجموعة والآلات (فيما يتعلق بوكالة المجموعة بشكل عام [٤٨]. النقطة الأكثر عمومية لدينا هي: على الرغم من أنها ليست كافية في حد ذاتها، فقد يكون من الممكن وضع آلات لديها القدرة على التفكير الأخلاقي (سواء من خلال الجهاز أم مالكيها أم مصمميها) ضمن إطار اجتماعي أو قانوني أوسع مما يجعل من الممكن تقديم الاعتذارات. في هذه الحالة قد تؤدي القدرة على التفكير الأخلاقي إلى تعزيز التوافق الأخلاقي للآلات بفضل تحسين ملاءمتها لجوانب الأخلاق البشرية على مستوى النظام.

بذلك تصبح القدرة على إعطاء تفسيرات تستند إلى السبب مهمة بشكل خاص إذا كنا بحاجة إلى أنظمة مستقلة للعمل في مجالات معقدة أخلاقيًا حيث لا يستطيع الأخلاقيون البشريون صياغة وسائل واضحة قائمة على النتائج لمراقبة أدائهم. على سبيل المثال، يكون تقييم نظام الذكاء الاصطناعي الطبي المكلف فقط بالتوصية بعلاج معين لمرضي السرطان سهل نسبيًا إذا أدى النظام إلى انخفاض معدل الوفيات

والمرضى من الأورام غير المعالجة، ومن العلاجات غير الضرورية، فيمكننا القول بأننا سنحصل على نتائج جيدة – أسباب مبنية على الثقة – حتى لو لم يستطع النظام شرح أسبابها. على النقيض من ذلك، سيتعين على النظام المسؤول عن إدارة جميع قرارات العلاج داخل المستشفى اتخاذ العديد من القرارات الخلافية من الناحية الأخلاقية حول مجموعات المرضى التي يجب منحها الأولوية. نظرًا لأن مصداقية الآلة في مثل هذه الحالات ستعتمد إلى حد ما على الأسباب التي يمكن أن تقدمها لأفعالها، فقد نطلب من الآلات أن تفي بمعايير أعلى من قدرة الإنسان على التفسير. إلى الحد الذي نريد أن تعمل فيه الأنظمة المستقلة في مثل هذه السياقات، يصبح من الأهمية بمكان أن تتمكن من تمثيل وتوصيل المنطق الأخلاقي الكامن وراء أفعالها بدقة.[٤٩]

ج – القرارات البشرية الفردية.

يجادل بعض مؤيدي بناء الآلات التي يمكن أن يكون لديها القدرة على التفكير الأخلاقي بأن القيام بذلك قد يؤدي أيضًا إلى تحسين التوافق الأخلاقي للبشر. الطريقة الأولى للقيام بذلك هي تحسين القرارات البشرية الفردية. يخضع التفكير البشري لعدد من العيوب التي نعتبرها بحق إخفاقات مثل قراراتنا التى تتأثر بالتحيزات، والمصالح الذاتية، والتفكير السيء. فنحن كبشر ماهرون، وبارعون بشكل مثير للقلق في خداع أنفسنا وخداع الآخرين للاعتقاد بأن أفعالنا سليمة من الناحية الأخلاقية. جادل علماء أخلاقيات الآلة بأن المنطق الأخلاقي الآلي سيكون خاليًا من العديد من هذه القيود البشرية[٥٠]. لنفترض أنه يمكننا بناء نظام للتفكير الأخلاقي قادر على حساب الآثار المترتبة على التزاماتنا الأخلاقية وإبرازها على سبيل المثال أن التزامي بالتخفيف من تغير المناخ لا يتوافق مع طلب شريحة لحم في مطعم. حتى لو لم يتم تنفيذ مثل هذا النظام في أي نظام مستقل، فقد يظل قادرًا على تحسين وتوسيع نطاق التفكير الأخلاقي البشري، على غرار الطريقة التي تعمل بها حاسبات الجيب على تحسين وتوسيع التفكير العددي البشري.

د- تكافؤ البشر مع النظام الأخلاقي.

بالإضافة إلى تحسين قراراتنا الفردية، يجادل *S.L Anderson بأن أخلاقيات الآلة قد تساعد في تحسين الأخلاق البشرية ككل، من خلال مساعدتنا في صياغة نظريات أخلاقية أكثر وضوحًا واتساقًا، وتحقيق إجماع متزايد حول المعضلات الأخلاقية. على سبيل المثال يجادل أندرسون بأنه إذا حاول فلاسفة الأخلاق صياغة نظرياتهم بصيغة يمكن حسابها بواسطة آلة؛ فإن هذا سيجبرهم على مواجهة الآثار المترتبة على نظرياتهم بشكل مباشر. يمكن القول إن تحسين التنظير الأخلاقي بهذه الطريقة يتطلب آلات قادرة على تمثيل التفكير الأخلاقي بشكل صريح، ويجب أن تكون قادرة على أن تكشف ليس فقط لفلاسفة الأخلاق ما الآثار المترتبة على نظرية أخلاقية معينة، ولكن كيف يتم التوصل إلى الاستنتاجات.(٥١)

قد تهدف بعض طرق تحسين الأخلاق البشرية إلى ضمان أن أفعالنا تفي بمعاييرنا الحالية باستمرار. على سبيل المثال تتعلم أنظمة التوجيه الأخلاقي لأندرسون حل المعضلات الأخلاقية بناءً على أمثلة تدريبية حيث يتفق علماء الأخلاق البشرية على الحل الصحيح (٥٢). قد يكون مثل هذا النظام قادرًا على رفع الأخلاق البشرية إلى مستوى الإجماع الحالي لعلماء الأخلاق البشرية "الخبراء". قد تعد التطبيقات الأخرى لأخلاقيات الآلة بتجاوز الإجماع الحالي لحل الخلافات الأخلاقية العالقة، أو الكشف عن الأماكن التي يمكن تحسين الإجماع الحالي فيها. يقترح بعض أنصار أخلاقيات الآلة أنها قد تكون بذلك قادرة على تعزيز التقدم الأخلاقي البشري بشكل فعال(٥٣).

* S.L Anderson أستاذ التطوير وعلم البيئة والسلوك، حاصل على درجة الدكتوراة من جامعة شيكاغو، الينوي، وحاصل على درجة البكالوريوس من كلية كارلتون مينسيوتا.
تاريخ الدخول على الموقع https://sib-illinois- du.translate.goog/profile/andersps
(٢٠٢٢/٥/١٧)

حتى ولو كان المنطق الأخلاقي للآلة لا يسمح للبشر بالوصول إلى إجماع متزايد، على سبيل المثال إذا كانت بعض الخلافات غير قابلة للحل بشكل أساسي فقد لا يزال بإمكانهم تحسين ملاءمة البشر داخل النظم الأخلاقية من خلال المساعدة في شرح هذه الاختلافات وجعلها مفهومة، وتحسين القدرة على فهم طبيعة خلافاتنا وشرحها لبعضنا البعض يمكن أن يحسن قدرتنا على التفاوض، أو إدارة مثل هذه النزاعات.

ففي حين أن بعض الفوائد المحتملة الموضحة أعلاه هي في الغالب وعود مضاربة في هذه المرحلة، فإن الفوائد المحتملة كبيرة بما يكفي لتوفير دافع حدسي لمحاولة بناء آلات قادرة على التفكير الأخلاقي، ولكن بشكل حاسم يجب موازنة هذه الفوائد مقابل أي مخاطر محتملة قد تكون متأصلة في تحقيق الامتثال الأخلاقي من خلال أخلاقيات الآلة، أو قد تنشأ كمنتجات ثانوية، وهذا ما سوف نتناوله في المحور الثالث المعنون ب " مخاطر إنشاء آلات أخلاقية.

المحور الثالث: مخاطر محاولة إنشاء آلات أخلاقية

في هذا المحور، سنناقش أربع فئات واسعة من المخاطر:

أ – خطر فشل الآلات المتوافقة أخلاقيًا أن تصبح غير أخلاقية.

ب– مخاطر قيام الآلات المتوافقة أخلاقيًا بتهميش أنظمة القيمة البديلة.

ج– خطر خلق مرضى أخلاقيين اصطناعيين.

د – خطر أن يؤدي استخدامنا للآلات الأخلاقية إلى تقليل قدرتنا الأخلاقية البشرية.

أ– الفشل والقابلية للفساد.

كما ذكرنا سابقًا إن امتلاك القدرة على التفكير الأخلاقي لا يضمن اتخاذ قرارات متوافقة أخلاقيًا. وبالتالي فإن تحميل الآلات بقرارات مهمة من الناحية الأخلاقية يحمل في طياته خطر الوصول إلى استنتاجات غير مقبولة أخلاقيًا، والتي كان البشر سيعترفون بها على هذا النحو الآتي:

أولاً– حتى أفضل علماء المنطق يمكنهم استخلاص استنتاجات خطأ إذا اعتمدوا على مقدمات زائفة. أبسط حالة لذلك هو إذا كانت الآلة تعتمد على معلومات خطأ حول المواقف التي تعمل فيها إذا فشلت في الكشف عن وجود بشر يجب أن تحميهم على نحو متصل. سلط البعض هنا الضوء على أن الاستعصاء الحسابي على التنبؤ بتأثيرات الفعل في المواقف الاجتماعية المعقدة قد يؤدي إلى عقل أخلاقي لا تشوبه شائبة بمعلومات كاملة إلى استنتاجات غير مقبولة أخلاقيًا. علاوة على ذلك إذا كانت المبادئ الأخلاقية، أو الأمثلة التدريبية التي قدمها مطورو النظام البشري تحتوي على عيوب فقد يؤدي ذلك إلى استنتاج الروبوتات لمبادئ غير مقبولة أخلاقيًا[٥٤].

ففي حين أنه سيكون شيئًا مهمًا لأخلاقيات الآلة أن تكون قادرة على ضمان أن الآلة تتخذ قرارات أخلاقية بأدنى حد من المعايير البشرية المقبولة، إلا أن هذا قد لا يكون جيدًا بما يكفي للآلات.

وهذا لعدة أسباب:

أولاً– قد نقبل بعض الأخطاء من البشر الأفراد، إذا تم تطبيق نظام مستقل على نطاق عالمي مثل مركبة مستقلة من شركة تصنيع كبيرة؛ فإن الأخطاء الفردية تكون بسيطة، ولكن المنهجية قد ترقى إلى مشاكل خطيرة للغاية في المجموع.

ثانيًا– قد نقبل مستويات معينة من متوسط الموثوقية من البشر لأننا طورنا طرقًا للتنبؤ بهذه الأخطاء وإدارتها. ومع ذلك كما هو موضح بأمثلة لتقنيات الخصومة في التعلم الآلي[٥٥]، غالبًا ما تفشل الآلات بطرق مختلفة عن البشر فعلى سبيل المثال قد تكون عرضة للفشل في ظل ظروف لا يفشل فيها البشر عادةً، وبالتالي فإن الخطر هو أنه عندما تفشل الآلات؛ فإنها تفعل ذلك بطرق يصعب التنبؤ بها، أو إدارتها. وبالتالي فإن مستويات الأداء التي ستكون عليه الآلة مقبولة تكون غير واضحة، ومن المحتمل أن يكون سياقها محددًا.

ينشأ خطر آخر من احتمال تلف أنظمة التفكير الأخلاقي بسهولة[٥٦]، سواء من قبل المصممين الخبيثين أم المتسللين أم أخطاء الترميز. يمكن أن يتفاقم هذا الخطر أكثر إذا كان للآلات الخبيثة – المبرمجة على أفعال غير أخلاقية – في الوقت نفسه قدرة قوية على تقديم تفسيرات خادعة أو متلاعبة لأفعالها. قد يتم استغلال الآلات التي لديها القدرة على إنتاج تفسيرات أخلاقية مقنعة لإقناع البشر بقبول الاختلافات الجادة عن التوافق الأخلاقي. من المؤكد أن العديد من الأجهزة اليوم التي تفتقر إلى القدرة على التفكير الأخلاقي معرضة للخطأ وعرضة للفساد. إن الابتعاد عن آلات البناء ذات التفكير الأخلاقي لن يحل هذه المشكلة، وقد يشمل ذلك التفكير الأخلاقي.

في الأنظمة الحالية يمكن أن تجعلها أكثر مرونة في مواجهة هذه المشاكل، ومع ذلك فإن قلقنا هنا هو أن قدرات التفكير الأخلاقي قد تكون في حد ذاتها غير معصومة وقابلة للتلف، وقد تكون ضعيفة بشكل خاص، كما يجادل فاندرلست ووينفيلد Vanderelst و Winfield[٥٧]. إذا كانت التقنية نفسها التي من شأنها أن تمنح الآلات القدرة على التفكير الأخلاقي يمكن أن تفشل بسهولة، أو تتعثر في إنتاج سلوك غير أخلاقي؛ فإن هذا من شأنه أن يوفر ثقلًا موازنًا قويًا لأي أسباب إيجابية لمتابعة أخلاقيات الآلة. على الأقل، يجب الحرص على عدم تكرار المشكلات التي كان من المفترض أن تحلها أخلاقيات الآلة.

لذلك وضع إسحاق أسيموف Isaac Asimov* (١٩٢٠-١٩٩٢) في روايته "الحلقة المفرغة" القوانين الأخلاقية الثلاثة الشهيرة. حاول كاتب الخيال العلمي الشهير "إسحاق أسيموف" وضع قوانين أخلاقية للروبوتات تنظم حقوقها وواجباتها، فضلاً عن تحديد الأساس الذي تستند إليه علاقة الإنسان بالآلة، والآلة بالإنسان، لتصبح في وئام بعيدًا عن أي شيء- مخاطر أو مشاكل- لبرمجة الذكاء الاصطناعي على أساسه، هذه القوانين هي:

١- لا يجوز للروبوت أن يؤذي الإنسان أو يسكت عما يضر به.

٢- يجب أن يطيع الروبوت الأوامر البشرية ما لم تتعارض مع القانون الأول.

* **كان إسحاق أسيموف** Isaac Asimov (٢ يناير ١٩٢٠ – ٦أبريل ١٩٩٢م) كاتبًا أمريكيًا وأستاذًا للكيمياء الحيوية في جامعة بوسطن. خلال حياته، كان أسيموف يُعتبر أحد كتاب الخيال العلمي "الثلاثة الكبار"، جنبًا إلى جنب مع روبرت أ. هاينلين وآرثر سي كلارك. كاتب غزير الإنتاج، كتب أو حرر أكثر من ٥٠٠ كتاب. كما كتب ما يقدر بـ ٩٠.٠٠٠ رسالة وبطاقة بريدية. بالإضافة إلى الكثير من الأعمال الواقعية. تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)-https://en-m-wikipedia-org.translate.goog/wiki/Isaac_Asimov?

٣- يجب أن يحافظ الروبوت على بقائه إذا لم يتعارض ذلك مع القانونين الأول والثاني.(٥٨)

كما هو واضح تم وضع هذه القوانين بسبب الخوف من خروج الروبوتات عن سيطرة الإنسان، وتحويلها إلى مدمرة كما هو موضح في أفلام الخيال العلمي. يعتبرها العديد من العاملين في مجال البرمجة الذكية قوانين أساسية وتوجيهية بالنسبة لهم في عملهم، لكن المتأمل سوف يكتشف أنها قوانين أخلاقية خيالية ومتسامية مشابهة للأوامر المطلقة لإيمانويل كانط، مما يعني أنها ملزمة تمامًا للإنسان في جميع الأحوال، وهو ما اعترض عليه كما هو معروف في تاريخ الفلسفة، بنيامين كونستان، عندما دافع عن حق الكذب من أجل إنقاذ أرواح البشر المهددين من قبل البشر الآخرين، على سبيل المثال.

وهكذا عندما نتحدث عن أخلاقيات "الأخلاق" الروبوتات، فهذا يعني أننا نتحدث عن أخلاقيات تتعلق بالتجربة وليس بالتعالي، أي الأخلاق القريبة من الحياة وتعقيدات الواقع. لا مجال هنا للحديث عن الصيغة: "يجب، لأنه يجب". بدلاً من ذلك، يجب أن يستدعي بالضرورة المرونة والتوازن بين الأمور، وترجيح كل منهما للآخر. لكل هذه الاعتبارات يبدو من الصعب إلقاء اللوم على الروبوت، وهذا بالضبط ما دفع اللجنة العالمية لأخلاقيات المعرفة العلمية والتكنولوجية التابعة لليونسكو، في عام ٢٠١٧، إلى الدعوة إلى ضرورة احترام القيم الإنسانية الأساسية، والمبادئ مثل الكرامة والاستقلال والمسؤولية.(٥٩)

قضى أسيموف في كتاباته شوطًا طويلاً في فحص حدود وثغرات هذه القوانين وعواقبها، أي عندما يمكن أن تسبب مفارقات أخلاقية وسلوكًية غير متوقعة، ورأى أخيرًا أن هذه القوانين ليست كافية تمامًا في حد ذاتها لتجنب أي الأخطار التي يسببها الذكاء الاصطناعي، لأن هناك مواقف قد تنتج عن الأذى البشري، وإن كانت تتفق مع تلك القوانين، إلا أنها كانت تحدث باستمرار في روايات أسيموف وقصصه القصيرة.

وهنا قد لا تكون قوانين Asimov نفسها عادلة – للروبوتات؛ لأن قوانين أسيموف تعامل الروبوتات بشكل فعال مثل العبيد، وهذا قد يكون مقبولاً في الوقت الحالي، لكنه قد يصبح مشكوكًا فيه أخلاقيًا (ويصعب تطبيقه) حيث تصبح الآلات أكثر ذكاءً وربما أكثر وعياً بذاتها.

كان هذا السبب نفسه أيضًا نقطة دخول لانتقاد مهندسي الذكاء الاصطناعي لقوانين أسيموف، الذين أقروا بوجود تيار فكري أخلاقي يؤمن بالحاجة إلى بناء الروبوتات على أساس تلك القواعد التي تشكلت من حولهم. أما بالنسبة لـ " أسيموف " نفسه، فقد رأى أن هذه القوانين الثلاثة على الرغم من نواقصها، إلا أنها تظل أفضل طريقة لتقليل الصراع المتوقع بين الإنسان والآلة، واعتقد أنه لا يمكن وضع قوانين قطعية كافية في حد ذاتها لمنع أي ضرر. لكن هؤلاء الثلاثة يظلون الأكثر عملية، حتى مع حدوث انتهاكات ذات عواقب كارثية محتملة في نموذج القوانين الثلاثة لأسيموف.(٦٠)

تختلف الروبوتات القتالية اختلافًا جوهريًا عن الأسلحة التي يتحكم فيها الإنسان؛ لأنها تتمتع بدرجة من استقلالية القرار.

ربما كل الحديث حتى الآن عن مستقبل بعيد، ولهذا السبب يبدو أقرب إلى الخيال العلمي من أي مشاكل حقيقية، لكن النقد الأخير لقوانين أسيموف يعيد المناقشة إلى حدود الواقع الحالي، وهذا النقد يتعلق بـ استخدام الذكاء الاصطناعي في التسلح. يتم تطوير الكثير من برامج الذكاء الاصطناعي الحالية للأغراض العسكرية، أي أنها مصممة خصيصًا لإلحاق الأذى بالبشر، وهو ما يتعارض مع قانون أسيموف الأول، ونظامه الأخلاقي بأكمله، ويؤدي بالضرورة إلى التفكير في منطق مختلف لإدارة علاقة الإنسان بآلة شديدة التعقيد، قد تكون مصممة أساسًا للقتل.(٦١)

ب- عدم قابلية القياس والتعددية والإمبريالية.

تؤكد تعددية القيم أن هناك العديد من القيم الأخلاقية المختلفة، حيث تُفهم "القيمة" على نطاق واسع لتشمل الواجبات، والسلع، والفضائل، وما إلى ذلك[٦٢]. إذا كانت تعددية القيمة صحيحة فلا يمكننا اختزال كل القيم إلى قيمة واحدة مثل السعادة، أو المتعة. ينكر مؤيدو القيمة إمكانية وجود مثل هذه الشروط، وبدلاً من ذلك يؤكدون أن هناك دائمًا حقيقة محددة حول كيفية التصرف بشكل أخلاقي. فعلي الرغم من أن كانط دافع عن الرأي القائل بوجود مبدأ أخلاقي واحد يجب على الفاعلين الأخلاقيين الالتزام به، وأن أي مبادئ أخلاقية أخرى يمكن اختزالها في هذا المبدأ[٦٣]. إلا أنه لا يتفق جميع علماء الأخلاق كمؤيدين للقيم، وهنا يعتقد "دبليو دي روس*"WD Ross أن هناك العديد من الواجبات الأخلاقية التي قد تتعارض أحيانًا. علاوة على ذلك، عندما تتعارض الواجبات، قد تكون المعضلة غير قابلة للحل على عكس الوحدويين. يميل أنصار التعددية إلى الاعتقاد بوجود بعض المعضلات الأخلاقية المعقدة على الأقل، وربما العديد منها، والتي تنتج عن تضارب بين القيم المتنافسة، وغير القابلة للقياس، والتي لا يمكن حلها.[٦٤].

لوضع التمييز في المصطلحات الرياضية، يدعي الأحاديون أن هناك ترتيبًا إجماليًا على مجموعة جميع الإجراءات الممكنة، بينما يدعي مؤيدو التعددية أنه لا يوجد سوى نظام جزئي. إذا كان هذا هو الحال فلا يمكننا أن نتوقع أن تكون الآلة قادرة

* **السير ويليام ديفيد روس** (15أبريل ١٨٧٧ – ٥ مايو ١٩٧١)، المعروف باسم **ديفيد روس** ولكن يُشار إليه عادةً باسم **WD Ross**، فيلسوفًا اسكتلنديًا معروفًا بعمله في الأخلاق. أشهر أعماله هو "الحق والخير"١٩٣٠، وربما اشتهر بتطويره لشكل تعددي، أخلاقي من الأخلاق الحدسية استجابة لشكل جنرال إلكتريك العواقبي من الحدس. قام روس أيضًا بتحرير وترجمة عدد من أعمال أرسطو بشكل نقدي، بالإضافة إلى كتابته عن الفلسفة اليونانية تشمل إنجازاته عمله مع جون ألكسندر سميث في ترجمة مؤلفة من ١٢ مجلدًا لأرسطو. https://stringfixer.com/ar/W._D._Ross تاريخ الدخول على الموقع (٢٠٢٢/٥/٢٩)

على حل مثل هذه المعضلات حيث لا يوجد مثل هذا الحل. علاوة على ذلك يرى البعض أن هناك أسبابًا للحفاظ على هذه التعددية أو التنوع. عند مواجهة معضلات أخلاقية لم يتم حلها، يتعين على الناس أحيانًا التصرف. في هذه الحالات قد يكون الإجراء أو النتيجة غير مرضية. ومع ذلك كما ناقشنا أعلاه، يتمتع معظم البشر بمجال تأثير محدود، ولكن قد لا يكون الأمر كذلك بالنسبة للآلات التي يمكن نشرها بشكل جماعي، بينما تخضع لتحكمها خوارزمية واحدة. وبالتالي فأيًا كان الأسلوب التجريبي المستخدم للتغلب على أي معضلات لا يمكن حلها حقًا يمكن أن يكون ذا تأثير كبير، ويمكن أن يؤدي هذا إلى شيء مثل قيمة الإمبريالية، أي تعميم مجموعة من القيم بطريقة تعكس نظام القيم لمجموعة واحدة (مثل المبرمجين). يمكن متابعة هذا عن قصد، أو ربما بشكل أكثر إثارة للقلق، ويمكن أيضًا أن يتم ارتكابه عن غير قصد إذا قام المبرمجون عن غير قصد بتضمين قيمهم في خوارزمية لها تأثير واسع. قد تؤثر إمبريالية القيمة هذه، أو تعطل الثقافات بشكل مختلف، أو تقلل من الاستقلالية الثقافية.[٦٥]

ج- تكوين مرضى أخلاقيين.

لقد أظهرنا سابقًا أن الآلات المصممة لتتمتع بقدرات التفكير الأخلاقي الخاصة بها يمكن أن يكون لها في الواقع سمات نربطها بالوكالة الحقيقية، وأشرنا أيضًا إلى بعض القضايا الفلسفية التي تنشأ من عزو الفعالية الحقيقية للآلات، ولاحظنا أنه أصبح من الشائع الحديث عن الآلات كعوامل من المفارقات إلى حد ما. في حين أن علماء أخلاقيات الآلة قد يتابعون الضرورة الأخلاقية لبناء آلات تعزز القرارات المتوافقة أخلاقيًا، وتحسن الأخلاق البشرية، فإن القيام بذلك قد يؤدي إلى تعاملنا مع هذه الآلات كعوامل مقصودة، والتي بدورها قد تمنحنا مكانة أخلاقية. وهنا يجادل المرضى في أن هذا يخاطر بخلق واجبات أخلاقية جديدة للبشر، واجبات قد تحجمنا بطرق مهمة وتوسع مسؤولياتنا الأخلاقية.[٦٦]

لقد لاحظنا من قبل أن البشر هم أفراد أخلاقيون ومرضى أخلاقيون. تنبع مسؤولياتهم الأخلاقية من وكالتنا: نظرًا لأننا قادرون عن علم على التصرف، أو انتهاك المعايير الأخلاقية، فإننا نتحمل المسؤولية عن أفعالنا (أو فشلنا في التصرف). في الوقت نفسه، نحن أيضًا مرضى أخلاقيًا: لدينا حقوق، ويُعتقد عادةً أن مصالحنا مهمة، ويتفق علماء الأخلاق على أنه لا ينبغي أن نُظلم أو نؤذي دون مبرر معقول.

يمكن فصل هذين المفهومين– الفاعلية الأخلاقية*، والسلطة الأخلاقية* (moral agency and moral patiency) – بشكل واضح، ولكنهما مع ذلك قد

---

* **الفاعلية الأخلاقية**" ترتكز على ثلاثة محددات تتمثل فيما يلي: ١– الملكية الأخلاقية أو "امتلاك زمام الذات الأخلاقية" moral ownership: وتتضمن الشعور بالمسئولية عن اتخاذ فعل أخلاقي عندما مواجهة قضايا أو مآزق أخلاقية. ٢– الفاعلية الأخلاقية moral efficacy: أو فاعلية الذات الأخلاقية وتتضمن اعتقادات الأفراد بقدرتهم على تنظيم المصادر وحشدها للقيام بعمل أخلاقي. ٣– الشجاعة الأخلاقية moral courage: وتعني امتلاك الأفراد للجسارة في مواجهة التهديدات والتغلب على المخاوف المتعلقة بالفعل أو العمل الأخلاقي.
Hannah, Avolio & ،Walumbwa, 2011; Hannah et al., 2009; Goud, 2005.

* **السلطة الأخلاقية** هي سلطة تقوم على مباديء أو حقائق أساسية مستقلة عن القوانين الوَضْعِيَّة أو المكتوبة. وعلى هذا الأساس، تستَوْجِب السلطة الأخلاقية وجود الحقيقة والامتثال لها. ولأن الحقيقة لا تتغير، فإن مباديء السلطة الأخلاقية غير قابلة للتغيير أيضًا، ذلك بالرغم من أنه عند تطبيقها على ظروف الأفراد، فإن تعليمات السلطة الأخلاقية للتَّصَرُّفات قد تَتَفَاوَتُ بسبب متطلبات الحياة البشرية. هذه المباديء – التي يمكن أن تكون ذات طبيعة ميتافيزيقية أو دينية– تُعْتَبَر معيارًا لأفعال الأفراد، سواء أكانت تلك الأفعال مشمولة في القوانين المكتوبة أم لا، ذلك حتى وإن كان المجتمع يتجاهل تلك المباديء أو ينتهكها. بالتالي، فإن السلطة الأخلاقية تنطبق على الضمير الخاص بكل فرد، ويكون هو حرًّا بالتصرف وفقًا لتعليمات تلك السلطة أو خِلَافًا لها. بناءً على ذلك، تُعَرَّف السلطة الأخلاقية أيضًا بأنها «المُسَلَّمَات الأساسية التي توجِّه أفكارنا عن العالم».
https://ar.wikipedia.org/wiki/%D8%B3%D9%84%D8%B7%D8%A9_%D8%A3%D8%AE%D9%84%D8%A7%D9%82%D9%8A%D8%A9 تاريخ الدخول على الموقع (٢٠٢٢/٥/٢٩)

يكونان مرتبطين في الممارسة. لذلك في حين أن الفاعلية الأخلاقية ليست ضرورية للموقف الأخلاقي (على سبيل المثال، قد نَعَدُّ الأطفال أو بعض الحيوانات مرضى أخلاقيين، لكن ليسوا أفراد أخلاقيين)، فقد يكون ذلك كافياً. وهذا يعني أن القدرات ذاتها التي تكمن وراء الفاعلية الأخلاقية قد تبرر أيضًا ادعاء المرض الأخلاقي. إذا كان الأمر كذلك، فمن خلال خلق عوامل أخلاقية مصطنعة، يمكننا (عن غير قصد) خلق مرضى أخلاقيين.

ما أساس المرض الأخلاقي (moral patiency) ؟ إنه موضوع نقاش كبير. لكن وجهة النظر الحديثة، التي دافع عنها العديد من الفلاسفة اليوم، تشير إلى القدرات المعرفية المتطورة[٦٧]. تم الدفاع عن القدرات المختلفة للمرشحين على سبيل المثال القدرة على الإرادة[٦٨] أو القدرة على الحصول على نوع من الوعي الذاتي[٦٩]. هذا التقليد في الإشارة إلى القدرات الفكرية المختلفة يعود على الأقل إلى كانط. نوع آخر من القدرات المعرفية التي غالبًا ما يُفترض أنها كافية للوضع الأخلاقي، أو على الأقل درجة معينة من الحالة الأخلاقية (moral status) لأولئك الذين يسمحون للمكانة الأخلاقية بالاعتراف بالدرجات هي القدرة على الشعور بالألم أو المعاناة.

وهنا يبدو من غير المحتمل أن تكون الآلات الحالية قد طورت وعيًا خرافيًا، وبالتالي من غير المرجح أن تشعر بالمتعة أو الألم، أو لديها القدرة على المعاناة [٧٠]، ومع ذلك من المحتمل أن تمتلك الآلات قدرات تفكير متقدمة أخرى من شأنها أن تقودنا إلى معاملتها مثل المرضى الأخلاقيين. كما ذكرنا بالنسبة للبعض؛ فإن الوعي الذاتي (أو الملاحظة الذاتية)، والقدرة على تمثيل الذات بشكل انعكاسي يؤسسان الموقف الأخلاقي. على الرغم من أن معظم الأجهزة أو جميعها تفتقر حاليًا إلى هذه السعة، إلا أن هناك بالفعل بعض الاستثناءات المعقولة. على سبيل المثال، تعمل بعض الخوارزميات مع طبقات هرمية من الشبكات العصبية، حيث تتنبأ المستويات الأعلى باحتمالية نجاح الطبقات الدنيا، وبالتالي تشارك في نوع من المراقبة الذاتية، والتمثيل

الذاتي[٧١]، وهنا سيطلب منا كانط أن نسأل ما إذا كانت الآلات قادرة على التفكير الذاتي المستقل، والعملي أم لا، وفي هذه الحالة يمكن أن يؤسس هذا كرامتهم، ويتطلب ألا نتعامل معها على أنها مجرد وسائل لتحقيق غاياتنا. لقد رأينا بالفعل أن محاولة بناء الآلات بقدرات تفكير أخلاقي (moral reasoning) مستقل هو الهدف الواضح في تكوين أخلاقيات الآلة (machines ethics) والذي يؤسس الدافع الأخلاقي (moral motivation) لها.

هناك مخاطر كبيرة في بناء الآلات التي يمكن اعتبارها مريضة من الناحية الأخلاقية. بصفتنا أفراد أخلاقيين مسؤولين، سنكون ملزمين بأخذ مصالحهم على محمل الجد، وقد يكون لهذا تكاليف باهظة: قد لا نتمكن من استخدام مثل هذه الآلات كمجرد أدوات أو عبيد، ولكن قد يتعين علينا احترام استقلاليتهم، أو حقهم في الوجود وعدم إيقاف تشغيلهم. إذا وصلنا إلى نقطة يعتمد فيها اقتصادنا، وأنظمتنا من الرعاية الصحية إلى التعلم بشكل كبير على الذكاء الاصطناعي، فقد يكون هذا مدمرًا بشكل كبير. قد نضطر أيضًا إلى مشاركة الامتيازات الخاصة بنا من خلال منح الذكاء الاصطناعي المتقدم بشكل مناسب الحق في التصويت، أو حتى الحصول على وطن خاص بهم.

وبالتالي، جادلت بريسون (**J. J. Bryson**[*])[٧٢] بأن المهندسين يتحملون

[*] جوانا جي بريسون **J. J. Bryson** أكاديمية معترف بها لخبرتها الواسعة في مجال الذكاء وطبيعته وعواقبه. تحمل درجتين في علم النفس والذكاء الاصطناعي (بكالوريوس شيكاغو، ماجستير في أدنبرة، دكتوراه في معهد ماساتشوستس للتكنولوجيا)، وهي منذ عام ٢٠٢٠ أستاذة الأخلاق والتكنولوجيا في مدرسة هيرتي للحكم في برلين، وهي الآن رائدة في أخلاقيات الذكاء الاصطناعي، حيث شاركت منذ ذلك الحين في تأليف أول سياسة أخلاقية للذكاء الاصطناعي على المستوى الوطني، وهي مبادئ الروبوتات في المملكة المتحدة (٢٠١١)، تركز أبحاثها الحالية على تأثيرات التكنولوجيا على المجتمعات البشرية، والنماذج الجديدة لحوكمة الذكاء الاصطناعي والتكنولوجيا الرقمية. وهي عضو مؤسس لمركز Hertie School للحوكمة الرقمية، وواحدة من تسعة خبراء في ألمانيا تم ترشيحهم إلى الشراكة العالمية للذكاء الاصطناعي. للمزيد انظر
تاريخ الدخول على الموقع https://www-joannajbryson-org.translate.goog/about?

مسؤولية عدم بناء روبوتات واعية، لذلك ليس لدينا أي التزامات خاصة تجاههم، وتجادل (بريسون) أيضًا بأن الروبوتات يجب أن تكون "عبيدنا" ويجب أن تخدمنا دون أن ندين لها بأي شيء (على الرغم من أن الأدوات قد تكون تشبيهًا أفضل، إلا أن معظم الناس يدركون الآن أن العبيد يُحرمون ظلماً من حالة المرض الأخلاقي).

د- تقويض المسؤولية أو تحجيمها.

الخطر الرابع المحتمل لأخلاقيات الآلة هو أنها ستقوض الفاعلية الأخلاقية للإنسان – أي أنها ستقوض قدرتنا على إصدار أحكام أخلاقية، واستعدادنا وقدرتنا على استخدام هذه القدرة، أو رغبتنا وقدرتنا على تحمل المسؤولية الأخلاقية (القرارات والنتائج).

يمكن أن تنشأ مثل هذه المواقف نتيجة لما يسمى بـ "مفارقة الأتمتة*"، وهي مشكلة شائعة تظهر في معظم الآلات الموفرة للعمالة. وهنا سوف نوضح كيف تنطبق هذه المشكلة على الآلات القادرة على التفكير الأخلاقي، ونسلط الضوء على التحديات الأخلاقية التي يثيرها ذلك.

(٢٠٢٢/٥/٢٧)

* **الأتمتة** (Automation) تسمى أيضًا " التشغيل الآلى" وفي بعض الأحيان " المكننة " وهو مصطلح حديث نسبيًا يغطي مجالًا واسعًا من التكنولوجيا التي تتطلب قدرًا ضئيلًا من التدخل البشري، ويشمل ذلك أتمتة عمليات التصنيع والتكنولوجيا والمعلومات والتسويق كما يغطي التطبيقات الشخصية مثل التشغيل الآلي للأجهزة المنزلية. بذلك يكون مفهوم الأتمتة مستوحى من كلمة "أوتوماتيكية " ولم يكن لها استخدام واسع حتى عام ١٩٤٧م عندما أنشات شركة فورد إدارة التشغيل الآلي. عمومًا يمكن تعريف الأتمتة بأنها تقنية تهتم بتنفيذ عملية ما من خلال الأوامر المبرمجة مع التحكم التلقائي في التغذية الراجعة، لضمان التنفيذ الصحيح للتعليمات، ويكون النظام الناتج قادرًا على العمل دون التدخل البشري للمزيد انظر:
https://www.ida2at.com/what-is-automation-and-how-has-it-evolved/

يحدد Harford ثلاثة عناصر لهذه المشكلة:

١- الأنظمة الآلية "تستوعب أوجه القصور" من خلال تصحيح الأخطاء تلقائيًا.

٢- حتى عندما يكون البشر على قدر كبير من المهارة، ستتضائل مهاراتهم لعدم ممارستهم لتنمية هذه المهارة وتطويرها، وذلك نظرا لاعتمادهم على الآلة.

٣- تميل الأنظمة الآلية إلى الفشل في المواقف غير العادية، أو الصعبة أو المعقدة بشكل خاص، مما يؤدي إلى أن الحاجة إلى التدخل البشري في أكثر المواقف اختبارًا، والتي قد يكون الإنسان غير مستعد لها.

إن هذه العناصر الثلاثة لها تأثير مباشر على اتخاذ القرار الأخلاقي. قد يكون العنصر الأول وثيق الصلة بالظروف التي يكون فيها هدف الأنظمة الآلية هو أن تتخذ الآلة قرارات أخلاقية من تلقاء نفسها، أو إذا كان هدفها هو مساعدة الإنسان في اتخاذ مثل هذه القرارات. في الحالة الأولى، عندما تتخذ الآلات قرارات، من الممكن ألا يطور البشر المهارات في البيئة المناسبة لأنفسهم. على سبيل المثال في حالة روبوت الرعاية الصحية، قد لا يكون الطاقم البشري قد طور الحكم، والحساسية اللازمتين لتقرير متى يكون تدخل رعاية الوالدين مطلوبًا للتأكد من أن المريض يأخذ أدويته. أما في الحالات التي يتخذ فيها الإنسان القرار، يمكن أن تضمن المساعدة الروبوتية عدم ظهور أوجه القصور في قدرات التفكير الأخلاقي البشري (في الحالات القياسية)، بالطريقة التي يضمن مصممي نظام تحديد الموقع العالمي (GPS) عدم ظهور أوجه القصور في المهارات الملاحية. للبشر – إلا عندما يفشل النظام.[٧٣]

العنصر الثاني من مفارقة الأتمتة، وهو خطر تآكل المهارات وثيق الصلة أيضًا، لا سيما في الحالات التي تكون فيها عملية صنع القرار مؤتمتة بالكامل (بما في ذلك الحالات التي يقصد فيها النظام أن يعمل على مستوى أفضل من المستوى

البشري)، والدليل على أن التفكير الأخلاقي هو في الواقع مهارة يبرز من خلال مدى ظهوره في التنشئة الاجتماعية للأطفال وتعليمهم، وحقيقة أنه جزء من التعليم المهني، على سبيل المثال في الطب. إذا كنا نعتقد أن الافتقار إلى الممارسة بسبب الأتمتة يمكن أن يؤدي إلى تآكل المهارات في بعض الأماكن – وهنا يستشهد هارفورد بقضية رحلة الخطوط الجوية الفرنسية ٤٤٧ Air France Flight 447 التي تحطمت بعد استجابة الطيار ومساعديه بشكل سيئ للتحطم – يكون بسبب واضح منذ البداية، وهو الاعتقاد بأن الأمر يتعلق باتخاذ قرار أخلاقي.

يوجد هنا توجه نحو ربط أنظمة تشغيل الطائرات بأنظمة خارجية كشريك ثالث بين المشغل والطائرة، على أن تكون مهمة هذا الشريك الثالث تحليل البيانات من قِبل خوارزميات يتم تثبيتها على عقل الآلة التي ترصد هذه البيانات مما يتيح خلال حصول هذا الشريك على البيانات الوصول إلى نتائج أكثر دقة، ثم يتم تقيمها عن طريق الأفراد، وهذا قد يطرح تحديات في المعايير المتعارضة وبالتالي يؤدي إلى عدم اليقين في عملية صنع القرار، فلدى المجتمعات آراء أخلاقية مختلفة ونظام أخلاقي مختلف! فليس الاختيار الاجتماعي للأخلاقيات في الذكاء الاصطناعى خيارًا جيدًا لأن المجتمع قد لا تكون له رؤية أخلاقية مشتركة واحدة فحسب، بل أصبح من الصعب تحديد أي نظام أخلاقي قويم من الناحية الأخلاقية لينبغي برمجته في الآلة؛ لذلك ينبغي التنسيق بين سلوكيات الأفراد والآلة، لأن البشر قد يقومون ببرمجة أخلاقيات قد تكون منحازة إلى شيء ربما تؤدي به إلى نتائج سلبية(٧٤).

العنصر الثالث يثبت الأولين، وهنا يمكننا أن نتخيل الآلات التي تتنقل بنجاح في الأسئلة الأخلاقية اليومية، والمفاضلات الخاصة ببيئتها، كما هو الحال في المستشفى أو على الطريق. يجب أن نأمل أن تتمكن هذه الآلات أيضًا من التعرف على قيودها الخاصة، وتنبيه الإنسان عندما يواجه موقفًا

يتجاوز التدريب أو البرمجة. ولكن هناك فرصة جيدة لأن تكون هذه المواقف (أو بعضها) أكثر تعقيدًا من المتوسط، وبالتالي قد تكون فقط تلك التي قد تكون أكثر صعوبة على الإنسان. من الممكن أيضًا أن يتم اتخاذ هذه القرارات بسرعة، على سبيل المثال في حالة قرارات نوع عربة الترولي للسيارات ذاتية القيادة، حيث يتسبب أي من الخيارين المحتملين في ضرر كبير، سيكون هذا جزءًا من الثانية. وهذا يثير احتمال أن البشر غير المستعدين، والذين لم يتم تطوير مهاراتهم (العنصر الأول) أو تآكلت (العنصر الثاني)، سوف يفرضون عليهم، ربما في غضون مهلة قصيرة، تلك القرارات الأخلاقية الأكثر صعوبة؛ فمن السهل أن ترى مدى سوء هذا.

يمكن أن تتفاقم هذه المشاكل إذا زادت الفاعلية الأخلاقية للآلات. كما أشرنا أعلاه، ترتبط الفاعلية ارتباطًا وثيقًا بالمسؤولية: تلك الكيانات التي نميل إلى اعتبارهم أفراد أخلاقيين مثاليين (البالغين الأصحاء) هي تلك الكيانات التي نحملها المسؤولية عن أفعالهم. إذا أصبحت الآلات أكثر كفاءة، فسنميل بشكل متزايد إلى تحميلها المسؤولية عن قراراتها وأفعالها، في حين أننا حتى هذه المرحلة ربما قمنا بتعيين المسؤولية للمطورين البشريين، أو المالكين، أو المستخدمين. قد تكون هناك أطر عمل يمكن أن تسير فيها الأمور على ما يرام. ولكن يمكننا أيضًا أن نتخيل أن إسناد المسؤولية الأخلاقية رسميًا للآلات من شأنه أن يؤدي إلى تفاقم مخاطر مفارقة الأتمتة المذكورة أعلاه، والتي يشعر فيها البشر بالفعل "بالابتعاد عن الخطر"[٧٥].

نظرًا لأن الآلات، والمواقف الأخلاقية التي تواجهها تصبح أكثر تعقيدًا وتعقيدًا؛ يمكن أن تتفاقم هذه التحديات بشكل أكبر. لقد أشرنا أعلاه إلى أن بعض القرارات يمكن أن تمتد إلى مثل هذا النطاق الواسع من القيم والمتغيرات

الأخرى (على سبيل المثال إعطاء الأولوية لموارد المستشفى)، بحيث نجد صعوبة في الاعتماد فقط على وسائل المراقبة القائمة على النتائج، وبدلاً من ذلك نعتمد أيضًا على النظام الأخلاقي الأوسع للتفسير، وإعطاء السبب بالنسبة لبعض فئات الخوارزميات، فإن القابلية للتفسير من هذا النوع تشكل بالفعل تحديات فنية كبيرة[٧٦]. لكن هذا يمكن أن يمثل تحديًا لأي نظام.

يعتمد لدينا نظام العطاء في عقلنا البشري على ما يمكن أن نفهمه نحن البشر، ولكي تكون قرارات الآلة مفهومة بالمثل، فهذا يعني أنها مفهومة لنا نحن البشر مع قدراتنا المعرفية، وحدودنا من المتصور أن تتمتع الآلات الأخلاقية بالقدرة على اتخاذ القرارات في المجالات التي يتجاوز تعقيدها قدراتنا البشرية لفهمها، على سبيل المثال، حيث تتأثر حياة العديد من الأشخاص بطرق مختلفة على مدى فترات زمنية طويلة، مما يتطلب أعدادًا كبيرة من المقايضات. في مثل هذه الحالات، يتم تحديد مفاهيم إعطاء العقل، والشفافية، والقابلية للتفسير بشدة. ربما يمكن لآلة متطورة بشكل صحيح أن تحاول إيصال أسبابها إلينا، ولكن فقط من خلال التبسيط الجسيم، بالطريقة التي قد يبسط بها الإنسان البالغ حجة أخلاقية لطفل صغير. لكن من الصعب أن نرى كيف يمكن للبشر أن يشرحوا بشكل هادف آلة من خلال نظام العطاء العقلي في مثل هذه الظروف.[٧٧]

يمكننا تخيل الحالة القصوى عند نشر الآلات الأخلاقية بشكل متزايد في البيئات اليومية من رعاية المسنين إلى المدارس، وهي تعمل بشكل جيد لتحقيق أهداف يمكن فهمها بسهولة. ثم يتم استخدام أنظمة أكثر تعقيدًا لتقديم المشورة بشأن الأمور الأكثر تعقيدًا. بعد إنشاء سجل حافل يعتمد عليه صناع القرار من ضباط الشرطة إلى مخططي المدن على هذه الأنظمة، ويزيدون من

تفويضهم للقرارات. يتم الاستمتاع على نطاق واسع بالتحسينات في العديد من جوانب الحياة الخاصة، والعامة وتنسب إلى الآلات. لذلك تم تكليفهم بالتدخل في مجالات ذات مستويات من التطور تتجاوز القدرات البشرية، سواء أكان ذلك لتحسين تدفق حركة المرور أم لتحسين الجينوم البشري. مرة أخرى يتم الاستمتاع بالمزايا، ولكن البشر لم يعد بإمكانهم فهم ما تفعله الآلات.

في مثل هذه الحالة، سيكون البشر في طريقهم للتخلي عن المسؤولية الأخلاقية عن القرارات المتخذة نيابة عنهم. قد يَعدُّ البعض أن هذا يستحق مهما كانت النتائج الجيدة التي يمكن الاستمتاع بها نتيجة تصرفات الآلات؛ بالطبع مثل هذا السيناريو سيجلب معه كل مخاطر مفارقة الأتمتة، مما يخلق مخاطر كبيرة في حالة تعطل الآلات. لكنه يجلب أيضًا قلقًا إضافيًا أكثر إثارة للقلق نظرًا لأن البشر في هذا السيناريو يتوقفون عن استخدام قدراتهم الأخلاقية، فإن الخطر يزيد أنهم لن يعرفوا حتى ما يعنيه فشل الآلات. وبالتالي فإن اتخاذ قرارات أخلاقية تبعية كافية للاعتماد على آلات معقدة للغاية بحيث لا يمكننا فهمها، وهذا يمكن أن يشكل خطرًا على نظام التفكير والمسؤولية الأخلاقية بالكامل.

الخاتمة

في هذا البحث، حاولنا توضيح مايلي:

أولًا– أهداف إنشاء آلات أخلاقية ومخاطرها، وجادلنا بأن هناك أسبابًا وجيهة للوهلة الأولى لمتابعة مخاطر إنشاء آلات أخلاقية.

ثانيًا– يمكن أن يؤدي تصميم الآلات ذات القدرة على التفكير الأخلاقي إلى تحسين التوافق الأخلاقي لكل من البشر والآلات. ومع ذلك فإن هذه الأسباب الظاهرة في حد ذاتها لا تعطي سببًا كافيًا لمتابعة أخلاقيات الماكينة ما لم تتم إدارتها بشكل صحيح، إما عن طريق تطوير حلول يمكن أن تخفف من المخاطر عند ظهورها، أو عن طريق صياغة لوائح تحجم استخدام الآلات. التفكير الأخلاقي للسياقات منخفضة المخاطر.

ثالثًا– إن الخطوة الأولى الحاسمة في تطوير الحلول لتخفيف هذه المخاطر وظهورها هي الحصول على مزيد من الوضوح حول ما إذا كان من المحتمل ظهور هذه المخاطر ومتى.

رابعًا– نستنتج من خلال هذا البحث أيضا، أننا نجد العلماء في هذا المجال منقسمين إلى قسمين؛ القسم الأول يدعي أن التكنولوجيا تشكل تهديدًا للجنس البشري، وعلي البشر تقبل ذلك نظراً لأن الفوائد تتجاوز التكاليف. أما القسم الثاني فإنه يميل إلى التركيز على الحذر والشفافية في التعامل مع هذه التكنولوجيا، وهؤلاء يطالبون بتأسيس هيئة دولية لمراقبة وتقنين البحث العلمي والابتكارات في هذا المجال، وهم أيضًا مهتمون بدراسة تحولات سوق العمل جراء الأتمتة.

خامسًا– نستنتج أيضًا أن الحالة الواقعية والتنبؤات المستقبلية لتأثير الذكاء الاصطناعي. أن البشر في صدد عيش تحول حتمي يكون فية الإنسان مستقبلًا جزءًا من نظامه، وليس مسيرًا لنظامه كما هو الحال اليوم.

سادسًا– نستنتج أيضًا أن نمط المجتمعات البشرية سيتبع منحني جديد ليتجه نحو مجتمعات جديدة متعايشة مع الآلات ومتوافقة معها، وهذا بدأ مع المدن الذكية، والمنازل الذكية، وبذلك تستبعد في المستقبل تدمير الجنس البشري بسبب الآلة كما ذهب إليها بعض المبالغيين، وتستبعد أيضًا استقلالية الآلة تمامًا جراء تعلمها وذاتيتها لنستند إلي قاعدتين أساسسيتين إحداهما نظرية تأسيس علم الحوسبة، والتي مفادها أنه من المستحيل للإنسان أن يضع خوارزمية مطلقة لأن واضعها غير مطلق بطبيعته، والثانية أن هناك إختلافا جوهريا بين الأداء والخلق: فالروبوت قادر على أن يتغلب على أفضل لاعب شطرنج في العالم، ولكنه غير قادر على اختراع قاعدة أخلاقية.

سابعًا– يجب أن يتفق معظم علماء أخلاقيات الآلة على أن التفكير الأخلاقي لا يضمن التوافق الأخلاقي الكامل، وبناء عليه يعتمد الدافع وراء متابعة أخلاقيات الآلة على الادعاء الأكثر تواضعًا بأنه يمكن تحقيق تقدم تدريجي كافٍ لأخلاقيات الآلة لتقديم إسهام إيجابي في المواءمة الأخلاقية لصنع القرار الآلي.

ثامنًا– ما يثير القلق هو أن الذكاء الاصطناعي القوي، والذي يتميز بكونه برنامجًا متطورًا لديه القدرة على حل المشكلات التي يواجهها من تلقاء نفسه دون الرجوع إلى البشر. لذلك فهي تتمتع باستقلال تام عنها.

تاسعًا– إن التزام الكيانات التي تنشر أنظمة خوارزمية بأن تكون منفتحة بشأن كيفية عملها، وأن تكون مسؤولة عن قراراتها القائمة على الخوارزمية، هو ضمانة مهمة ضد إساءة الاستخدام. تمامًا كما يحق للمواطنين الديمقراطيين التدقيق في ممارسة السلطة السياسية ومحاسبتها، كذلك يحق لمكونات الخوارزميات التدقيق، ومساءلة ممارسة سلطة الخوارزميات. ولكن

عندما تنشأ الخلافات، بين الكيانات التي تتخذ القرارات الحسابية وتنفذها، وأولئك الذين يخضعون لها، كيف ينبغي حل هذه الخلافات؟ من ناحية أخرى لا يمكننا أن نتوقع من المواطنين التمسك بمخرجات الخوارزميات التي قد لا يتفقون معها في المعايير المعرفية والمعيارية. ومع ذلك يجب علينا أيضًا تقديم بعض المعايير الإيجابية التي يمكن للكيان من خلالها أن ينجح في توفير حساب مُرضٍ لنظامه الحسابي. الإجابة التي يقترحها العقل العام هي أن الكيان الذي يرغب في تنفيذ خوارزميته يجب أن يكون قادرًا على حساب نظامه بمصطلحات معيارية، ومعرفية يمكن لجميع الأفراد المعقولين في المجتمع قبولها.

عاشرًا- يعدُّ بناء الآلات مع الضمير مهمة كبيرة، ويتطلب جهودًا منسقة من الفلاسفة وعلماء الكمبيوتر والمشرعين والمحامين؛ لأنه " عندما تصبح الآلات أسرع وأكثر ذكاءً وقوة، تصبح الحاجة إلى منحها حسًا أخلاقيًا أكثر إلحاحًا.

الحادي عشر- تم تحديد أربعة موضوعات مهمة سيحتاج البحث المستقبلي العمل على معالجتها وهي:

١- الظروف التي من المحتمل أن يعزز فيها نظام التفكير الأخلاقي التوافق الأخلاقي للآلات، والأهم من ذلك، تحت أي ظروف من المحتمل أن تفشل مثل هذه الأنظمة؟ حتى لو ثبت أن القدرة على التفكير الأخلاقي لديها القدرة على تعزيز التوافق الأخلاقي للآلة بشكل كبير، فيجب موازنة ذلك مقابل مخاطر الإخفاقات المنهجية أو غير المتوقعة. بالإضافة إلى ذلك، كيف يمكننا منع التفكير الأخلاقي الآلي من أن يكون قابلاً للفساد، أو يتم توظيفه لغايات خبيثة، أو خادعة، أو احتيالية؟

٢- كيف نضمن أن مثل هذه الآلات قادرة على التعامل بشكل مناسب مع تعددية القيمة والخلافات العميقة؟ من ناحية أخرى تعد القدرة على التوفيق بين مثل هذه الخلافات أحد الفوائد المحتملة لنظام التفكير الأخلاقي الآلي. ومع ذلك، كما وضحنا لا نريد أن تُستَبعد الآلة التعددية الأخلاقية الحميدة افتراضيًا، على سبيل المثال بافتراض وجود إجابة واحدة ومحددة لجميع المشكلات الأخلاقية.

٣- تحت أي ظروف نعتقد أنه يجب علينا منح حقوق معنوية للآلات؟ هل تفي المتطلبات الأساسية للفاعلية الأخلاقية أيضًا بشروط المبالغة الأخلاقية؟ ما العواقب التي قد تترتب على الاعتراف بالسلوك الأخلاقي للآلات (المتقدمة بشكل مناسب)؟

٤- كيف يمكننا تجنب التفكير الأخلاقي الآلي الذي يقوض مسؤوليتنا الأخلاقية؟ على وجه التحديد، ما التأثير المحتمل للاعتماد على الآلات الأخلاقية على حكمنا الأخلاقي في مختلف القطاعات والبيئات؟ كيف يمكننا الحفاظ على الاستقلالية الأخلاقية والرقابة حيث تصدر الآلات أحكامًا أخلاقية في سيناريوهات التعقيد المتزايد؟

الهوامش

1 - **H. Lacey, Is Science Value-Free?** London, U.K.: Routledge, 1999,p 55. And see P. Kitcher, Science in a Democratic Society. New York, NY, USA: Prometheus Books, 2011, p 23

2 - **S. L. Anderson,** "Machine metaethics," in Machine Ethics, M. Anderson and S. Anderson, Eds. New York, NY, USA: Cambridge Univ. Press, 2011 pp. 21–27.

3 - **Bewaji, John A. I. "**Ethics and Morality in Yoruba Culture," in Kwasi Wiredu (ed.), A Companion to African Philosophy, Oxford: Blackwell Publishing, 2004, pp. 396.

4 - J. Sullins, "When is a robot a moral agent?" Int. Rev. Inf. Ethics, vol. 6, pp. 23–30, Dec. 2006.

5 - R. Tonkens, "A challenge formachine ethics," Minds Mach., vol. 19, no. 3, pp. 421–438, 2009.

6 - G. W. F. Hegel, Elements of the Philosophy of Right, A. W. Wood, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1991,p 76

7 - J. Annas, "Ancient ethics and modern morality," Philos. Perspectives, vol. 6, pp. **119–**136, Jan. 1992. And see, M. Anderson and S. L. Anderson, "Guest editors' introduction: Machine ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 10–11, Jul. 2006.

8 - G. D. Crnkovic and B. Çürüklü, "Robots: Ethical by design," Ethics Inf. Technol., vol. 14, no. 1, pp. 61–71, 2012.

9 - J. H. Moor, "The nature, importance, and difficulty of machine ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 18–21, Jul. 2006**.**

10 - IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being With Autonomous and Intelligent Systems, Version 2. 2017. http://standards.ieee.org/develop/ indconn/ec/autonomous_systems.html

11 - J. Rawls, Political Liberalism. New York, NY, USA: Columbia Univ. Press, 1993,p45 And see R. Binns, "Algorithmic accountability and

public reason," Philos. Technol., 2017. https://doi.org/10.1007/s13347-017-0263-5 تاريخ الدخول على الموقع ٢٠٢٢/٤/١٠

12- **Gary Marcus,** Moral Machines, https://www.newyorker.com/news/news-desk/moral-machines
تاريخ الدخول على الموقع ٢٠٢٢ /٤/١٠

13 - **H. Dreyfuss, "**Why heideggerian AI failed and how fixing it would require making it more heideggerian," Philos. Psychol., vol. 20, no. 2, pp. 247–268, 2007.

14 - **Wiener, N. (1960).** Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358.

15 - ibid, P,132

16 - **M. Anderson, S. L. Anderson, and C. Armen, "**An approach to computing ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 56–63, Jul. 2006.

17 - **ibid**, P,60

18 - **C. Allen, I. Smit, and W. Wallach,** "Artificial morality: Top-down, bottom-up, and hybrid approaches," Ethics Inf. Technol., vol. 7, no. 3, pp. 149–155, 2005.

19 - **M. Anderson, S. L. Anderson, and C. Armen, "**Towards machine ethics," in Proc. IAAA Workshop Agent Org. Theory Pract., San Jose, CA, USA, Jul. 2004, pp. 1–7.

20 - **D. Davidson, Essays on Actions and Events.** Oxford, U.K.: Clarendon Press,P66, 1980.

21 - **J. Searle, "Minds, brains, and programs**," Behav. Brain Sci., vol. 3, no. 3, pp. 417–424, 1980.

22 - **D. C. Dennett, Brainchildren:** Essays on Designing Minds. Cambridge, MA, USA: MIT Press,P44, 1998.

23 - **D. C. Dennett,** The Intentional Stance. Cambridge, MA, USA: MIT Press,P66, 1987.

24 - T. Crane, Intentionality as the Mark of the Mental, Contemporary Issues in the Philosophy of Mind, A. O'Hear, Ed., 1998.

25 - J. H. Moor, "The nature, importance, and difficulty of machine ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 18–21, Jul. 2006.

26 - C. Allen and W. Wallach, Moral Machines: Teaching Robots Right from Wrong. London, U.K.: Oxford Univ. Press,P20, 2009.

27 - M. Dawkins, Why Animals Matter. London, U.K.: Oxford Univ. Press,P33, 2012.

28 - C. B. Jaeger and D. T. Levin, "If Asimo thinks, does Roomba feel the legal implications of attributing agency to technology," J. Hum.-Robot Interact., vol. 5, no. 3, pp. 3–25, 2016.

29 - **STEPHEN CAVE, RUNE NYRUP, KARINA VOLD , AND ADRIAN WELLER,** Motivations and Risks of Machine Ethics, Vol. 107, No. 3, March 2019, p 565

٣٠ - **كزافييه غوشيه ،** هل الآلة وكيل أخلاقي؟ في الشروط المعرفية للحديث عن فلسفة أخلاقية للآلات، ترجمة، خديجة حلفاوي، مؤسسة مؤمن بلا حدود للدراسات والأبحاث. https://www.mominoun.com/articles

31 - **Rawls, J. (1997).** The idea of public reason revisited. *The University of Chicago Law Review. University of Chicago. Law School, 64*(3), 765–807.

32 - ibid, P,65

33 - **Wiener, N. (1960)**. Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358.

34 - **Halevy, A., Norvig, P., & Pereira, F. (2009**). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12.

35 - ibid, P,11

36 - **STEPHEN CAVE, RUNE NYRUP, KARINA VOLD , AND ADRIAN WELLER, Motivations** and Risks of Machine Ethics, Vol. 107, No. 3, March 2019, p 566

37 - ibid, P,66

38 **C. Allen, W. Wallach, and I. Smit,** "Why machine ethics?" IEEE Intell. Syst., vol. 21, no. 4, pp. 12–17, Jul. 2006.

39 - ibid, P,13

40 - M. Brundage, "Limitations and risks of machine ethics," J. Exp. Theor. Artif. Intell., vol. 26, no. 3, pp. 355–372, 2014.

41 - T. M. Powers, "Incremental machine ethics," IEEE Robot. Automat. Mag., vol. 18, no. 1, pp. 51–58, Mar. 2011.

42 - J. H. Moor, "The nature, importance, and difficulty of machine ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 18–21, Jul. 2006.

43 - C. Allen, W. Wallach, and I. Smit, "Why machine ethics?" IEEE Intell. Syst., vol. 21, no. 4, pp. 12–17, Jul. 2006.

44 - **STEPHEN CAVE,** RUNE NYRUP, KARINA VOLD , AND ADRIAN WELLER, Motivations and Risks of Machine Ethics, Vol. 107, No. 3, March 2019, p 567

45 -**C. Allen, W. Wallach, and I. Smit, "**Why machine ethics?" IEEE Intell. Syst., vol. 21, no. 4, p. 16

46 - J. Bryson and A. Winfield, "Standardizing ethical design for artificial intelligence and autonomous systems," Computer, vol. 50, no. 5, pp. 116–119, May 2017. And see K. Baum, M. E. Köhl, and Schmidt, "Two challenges for CI trustworthiness and how to address them," in Proc. 1st Workshop Explainable Comput. Intell., Santiago de Compostela, Spain, Sep. 2017, pp. 1–5.

47 - **K. Baum, M. E. Köhl, and Schmidt, "**Two challenges for CI trustworthiness and how to address them," in Proc. 1st Workshop Explainable Comput. Intell., Santiago de Compostela, Spain, Sep. 2017, p. 4.

48 - C. List and P. Pettit, Group Agency: The Possibility, Design, and Status of Corporate Persons. London, U.K.: Oxford Univ. Press, 2011.

49 - **STEPHEN CAVE, RUNE NYRUP, KARINA VOLD , AND ADRIAN WELLER,** Motivations and Risks of Machine Ethics, Vol. 107, No. 3, March 2019, p 568

50 - G. Marcus. (Nov. 24, 2012). Moral Machines. The New Yorker. https://www. newyorker.com/news/news-desk/moral-machines

تاريخ الدخول على الموقع ٢٠٢٢/٤/١٠

51 - S. L. Anderson, "How machines might help us achieve breakthroughs in ethical theory and inspire us to behave better," in Machine Ethics,M. Anderson and S. Anderson, Eds. New York, NY, USA: Cambridge Univ. Press, 2011, pp. 524–530.

52 - M. Anderson, S. L. Anderson, and C. Armen, "An approach to computing ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 56–63, Jul. 2006.

53 -G. Marcus. (Nov. 24, 2012). Moral Machines. The New Yorker. https://www. newyorker.com/news/news-desk/moral-machines

تاريخ الدخول على الموقع ٢٠٢٢/٤/١٠

54 - M. Brundage, "Limitations and risks of machine ethics," J. Exp. Theor. Artif. Intell., vol. 26, no. 3, pp. 355–372, 2014.

55 - A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok (2017). "Synthesizing robust adversarial examples." https://arxiv.org/abs/1707.07397 تاريخ الدخول على الموقع ٢٠٢٢/٤/١

56 - V. Charsi (2017). "Towards moral autonomous systems." https://arxiv.org/abs/1703.04741 تاريخ الدخول على الموقع ٢٠٢٢/٣/١٠

57 - D. Vanderelst and A. Winfield (2016). "The dark side of ethical robots." https://arxiv.org/abs/1606.02583

58 - ibid

59 - https://manshoor.com/society/artificial-intelligence-morality/

تاريخ الدخول على الموقع ٢٠٢٢/٤/١١

60 - **G. Marcus. (Nov. 24, 2012).** Moral Machines. The New Yorker. https://www. newyorker.com/news/news-desk/moral-machines

تاريخ الدخول على الموقع ٢٠٢٢/٤/١٠

61 - https://manshoor.com/society/artificial-intelligence-morality/

تاريخ الدخول على الموقع ٢٠٢٢/٤/١١

62 - E. Mason, "Value pluralism," in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed. 2015. https://plato.stanford.edu/ archives/sum2015/entries/value-pluralism/

تاريخ الدخول على الموقع ٢٠٢٢/٤/٥

63 - I. Kant, Groundwork of the Metaphysic of Morals.1785.

64 - D. Wiggins, "Weakness of will, commensurability, and the objects of deliberation and desire," in Essays on Aristotle's Ethics, A. O. Rorty, Ed. Berkeley, CA, USA: Univ. California Press,P66, 1980.

65 - M. Stocker, "Abstract and concrete value: Plurality, conflict and maximization," in Incommensurability, Incomparability and Practical Reason, R. Chang, Ed. Cambridge, MA, USA: Harvard Univ. Press,P122, 1997.

66 - in Proc. Int. Conf. Auton. Agents Multiagent Syst. (AAMAS). http://celweb.vuse.vanderbilt.edu/aamas18/

تاريخ الدخول على الموقع ٢٠٢٢/٢/١٠

67 - A. Jaworska and J. Tannenbaum, "The grounds of moral status," in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed. 2017. https://plato.stanford.edu/archives/ fall2017/entries/grounds-moral-status/ تاريخ الدخول على الموقع ٢٠٢٢/١/١٠

68 - W. Quinn, "Abortion: Identity and loss," Philos.Public Affairs, vol. 13, no. 1, pp. 24–54, 1984.

69 - J. McMahan, The Ethics of Killing: Problems at the Margins of Life. London, U.K.: Oxford Univ. Press,P78, 2002**.**

70 - S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it" Science, vol. 358, no. 6362, pp. 486–492, 2017.

71 - A. Cleeremans, "Connecting conscious and unconscious processing," Cogn. Sci., vol. 38, no. 6, pp. 1286–1315, 2014.

72 - J. J. Bryson, "Robots should be slaves," in Close Engagements With Artificial Companions: Key Social, Psychological, Ethical and Design Issue, Y.Wilks and

J. Benjamins, Eds. 2010, pp. 63–74.

73 - T. Harford. Crash: How Computers are Setting us up for Disaster. The Guardian. https://www.theguardian.com/technology/2016/oct/11/crash-howcomputers- are-setting-us-up-disaster تاريخ الدخول على الموقع ٢٠٢٢/٣/٣

[and see moor) T. Harford, Messy: How to Be Creative and Resilient in a Tidy-Minded World. New York, NY, USA: Riverhead,P88, 2016.

٧٤ - **الرينس هالبيباس،** فلسفة تصميم الأخلاق في خوارزميات الآلة للأنظمة المستقلة، مقال منشور

https://www.unsystemsarabia.com9

تاريخ الدخول على الموقع ٢٠٢٢ /٤/١٣

75 - J. Danaher, "The rise of the robots and the crisis of moral patiency," AI Soc., pp. 1–8, 2017. https://doi.org/10.1007/ s00146-017-0773-9

تاريخ الدخول على الموقع ٢٠٢٢/٤/١٠

76 - P. J. G. Lisboa, "Interpretability inmachine learning -Principles and practice," in Fuzzy Logic and Applications. WILF (Lecture Notes in Computer Science), vol. 8256, F. Masulli, G. Pasi, and R. Yager, Eds. Cham, Switzerland: Springer, 2013, pp. 15–21.

74 - **Deary, I. J., Penke, L., & Johnson, W. (2010).** The neuroscience of human intelligence differences. Nature reviews neuroscience, 11(3), 201.

قائمة المصادر والمراجع

أولًا- المراجع العربية والموسوعات.

١- الرينس هالبيباس، فلسفة تصميم الأخلاق في خوارزميات الآلة للأنظمة المستقلة، مقال منشور https://www.unsystemsarabia.com9 تاريخ الدخول على الموقع (٢٠٢٢/٤/١٣)

٢- كزافييه غوشيه، هل الآلة وكيل أخلاقي؟ في الشروط المعرفية للحديث عن فلسفة أخلاقية للآلات، ترجمة، خديجة حلفاوي، مؤسسة مؤمن بلا حدود للدراسات والأبحاث. https://www.mominoun.com/articles

٣- موسوعة ستانفورد للفلسفة، الميتا أخلاق، ترجمة حسان عبيدات، ص ١

ثانيًا- المصادر والمراجع الأجنبية.

1- **A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok (2017).** "Synthesizing robust adversarial examples." https://arxiv.org/abs/1707.07397

2- A. Cleeremans, "Connecting conscious and unconscious processing," Cogn. Sci., vol. 38, no. 6, pp. 1286–1315, 2014.

3- **A. Jaworska and J. Tannenbaum**, "The grounds of moral status," in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed. 2017.

[Online).Available: https://plato.stanford.edu/archives/ fall2017/entries/grounds-moral-status/

4- **Bewaji, John A. I.** "Ethics and Morality in Yoruba Culture," in Kwasi Wiredu (ed.), A Companion to African Philosophy, Oxford: Blackwell Publishing, 2004

5- **C. Allen and W. Wallach**, Moral Machines: Teaching Robots Right from Wrong. London, U.K.: Oxford Univ. Press, 2009.

6- **C. Allen, I. Smit, and W. Wallach**, "Artificial morality: Top-down, bottom-up, and hybrid approaches," Ethics Inf. Technol., vol. 7, no. 3, pp. 149–155, 2005.

7- **C. Allen, W. Wallach, and I. Smit**, "Why machine ethics?" IEEE Intell. Syst., vol. 21, no. 4, pp. 12–17, Jul. 2006.

8- **C. B. Jaeger and D. T. Levin**, "If Asimo thinks, does Roomba feel the legal implications of attributing agency to technology," J. Hum.-Robot Interact., vol. 5, no. 3, pp. 3–25, 2016.

9- **C. List and P. Pettit, Group Agency**: The Possibility, Design, and Status of Corporate Persons. London, U.K.: Oxford Univ. Press, 2011.

10 - **D. C. Dennett, Brainchildren**: Essays on Designing Minds. Cambridge, MA, USA: MIT Press, 1998.

11- **D. C. Dennett, The Intentional Stance**. Cambridge, MA, USA: MIT Press, 1987.

12- **D. Davidson,** Essays on Actions and Events. Oxford, U.K.: Clarendon Press, 1980.

13- Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. Nature reviews neuroscience, 11(3), 201.

14- **D. Vanderelst and A. Winfield (2016)**. "The dark side of ethical robots." https://arxiv.org/abs/1606.02583

15- **D. Wiggins**, "Weakness of will, commensurability, and the objects of deliberation and desire," in Essays on Aristotle's Ethics, A. O. Rorty, Ed. Berkeley, CA, USA: Univ. California Press, 1980.and see moor.

16- E. Mason, "Value pluralism," in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed. 2015. https://plato.stanford.edu/ archives/sum2015/entries/value-pluralism/

17- **G. D. Crnkovic and B. Çürüklü, "Robots**: Ethical by design," Ethics Inf. Technol., vol. 14, no. 1, pp. 61–71, 2012.

18- **G. Marcus.** (Nov. 24, 2012). Moral Machines. The New Yorker. https://www. newyorker.com/news/news-desk/moral-machines

19- G. Marcus. (Nov. 24, 2012). Moral Machines. The New Yorker. https://www. newyorker.com/news/news-desk/moral-machines

20- **G. W. F. Hegel**, Elements of the Philosophy of Right, A. W. Wood, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1991.

21- Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12.

22- **H. Dreyfuss**, "Why heideggerian AI failed and how fixing it would require making it more heideggerian," Philos. Psychol., vol. 20, no. 2, pp. 247–268, 2007.

23- **H. Lacey**, Is Science Value-Free? London, U.K.: Routledge, 1999. And see P. Kitcher, Science in a Democratic Society. New York, NY, USA: Prometheus Books, 2011.

24- I. Kant, Groundwork of the Metaphysic of Morals.1785.

25- **IEEE Global Initiative** on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being With Autonomous and Intelligent Systems, Version 2. 2017. http://standards.ieee.org/develop/ indconn/ec/autonomous_systems.html

26- **in Proc. Int.** Conf. Auton. Agents Multiagent Syst. (AAMAS).

http://celweb.vuse.vanderbilt.edu/aamas18/

27- J. Annas, "Ancient ethics and modern morality," Philos. Perspectives, vol. 6, pp. 119–136, Jan. 1992. And see moor,

28- **J. Bryson and A. Winfield**, "Standardizing ethical design for artificial intelligence and autonomous systems," Computer, vol. 50, no. 5, pp. 116–119, May 2017.

29- **J. Danaher**, "The rise of the robots and the crisis of moral patiency," AI Soc., pp.–8, 2017. https://doi.org/10.1007/ s00146-017-0773-9

30- **J. H. Moor**, "The nature, importance, and difficulty of machine ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 18–21, Jul. 2006.

31- **J. J. Bryson**, "Robots should be slaves," in Close Engagements With Artificial Companions: Key Social, Psychological, Ethical and Design Issue, Y.Wilks and

32- **J. McMahan**, The Ethics of Killing: Problems at the Margins of Life. London, U.K.: Oxford Univ. Press, 2002.

33- **J. Rawls**, Political Liberalism. New York, NY, USA: Columbia Univ. Press, 1993. And see R. Binns, "Algorithmic accountability and public reason," Philos. Technol., 2017. https://doi.org/10.1007/s13347-017-0263-5.

34- **J. Searle**, "Minds, brains, and programs," Behav. Brain Sci., vol. 3, no. 3, pp. 417–424, 1980.

35- **J. Sullins,** "When is a robot a moral agent?" Int. Rev. Inf. Ethics, vol. 6, pp. 23–30, Dec. 2006.

36- K. Baum, M. E. Köhl, and Schmidt, "Two challenges for CI trustworthiness and how to address them," in Proc. 1st Workshop Explainable Comput. Intell., Santiago de Compostela, Spain, Sep. 2017, pp. 1–5.

37- M. Anderson, S. L. Anderson, and C. Armen, "An approach to computing ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 56–63, Jul. 2006.

38- M. Anderson and S. L. Anderson, "Guest editors' introduction: Machine ethics," IEEE Intell. Syst., vol. 21, no. 4, pp. 10–11, Jul. 2006.

39- **M. Anderson, S. L. Anderson, and C. Armen,** "Towards machine ethics," in Proc. IAAA Workshop Agent Org. Theory Pract., San Jose, CA, USA, Jul. 2004, pp. 1–7.

40- **M. Brundage**, "Limitations and risks of machine ethics," J. Exp. Theor. Artif. Intell., vol. 26, no. 3, pp. 355–372, 2014.

41- **M. Dawkins,** Why Animals Matter. London, U.K.: Oxford Univ. Press, 2012.

42- M. Stocker, "Abstract and concrete value: Plurality, conflict and maximization," in Incommensurability, Incomparability and Practical Reason, R. Chang, Ed. Cambridge, MA, USA: Harvard Univ. Press, 1997.

43- **P. J. G. Lisboa**, "Interpretability inmachine learning-Principles and practice," in Fuzzy Logic and Applications. WILF (Lecture Notes in Computer Science), vol. 8256, F. Masulli, G. Pasi, and R. Yager, Eds. Cham, Switzerland: Springer, 2013, pp. 15–21.

44- Rawls, J. (1997). The idea of public reason revisited. *The University of Chicago Law Review. University of Chicago. Law School, 64*(3), 765–807.

45- **R. Tonkens**, "A challenge formachine ethics," Minds Mach., vol. 19, no. 3, pp. 421–438, 2009.

46- **S. Dehaene, H. Lau, and S. Kouider**, "What is consciousness, and could machines have it" Science, vol. 358, no. 6362, pp. 486–492, 2017.

47- **S. L. Anderson**, "How machines might help us achieve breakthroughs in ethical theory and inspire us to behave better," in Machine Ethics, M.

Anderson and S. Anderson, Eds. New York, NY, USA: Cambridge Univ. Press, 2011, pp. 524–530.

48- **S. L. Anderson**, "Machine metaethics," in Machine Ethics, M. Anderson and S. Anderson, Eds. New York, NY, USA: Cambridge Univ. Press, 2011 pp. 21–27.

49- **STEPHEN CAVE, RUNE NYRUP, KARINA VOLD, AND ADRIAN WELLER,** Motivations and Risks of Machine Ethics, Vol. 107, No. 3, March 2019.

50- **T. Crane,** Intentionality as the Mark of the Mental, Contemporary Issues in the Philosophy of Mind, A. O'Hear, Ed., 1998.

51- **T. Harford**. (Oct. 11, 2016). Crash: How Computers are Setting us up for Disaster. The Guardian.

https://www.theguardian.com/technology/2016/oct/11/crashhowcomputers- are-setting-us-up-disaster

52-T. Harford, Messy: How to Be Creative and Resilient in a Tidy-Minded World. New York, NY, USA: Riverhead, 2016.

53- **T. M. Powers**, "Incremental machine ethics," IEEE Robot. Automat. Mag., vol. 18, no. 1, pp. 51–58, Mar. 2011.

54- **V. Charsi** (2017). "Towards moral autonomous systems."

https://arxiv.org/abs/1703.04741

55- **Wiener, N. (1960).** Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358.

56- **W. Quinn**, "Abortion: Identity and loss," Philos.Public Affairs, vol. 13, no. 1, pp. 24–54, 1984. J. Benjamins, Eds. 2010, pp. 63–74.

ثالثا- المواقع الإلكترونية.

1- https://hekmah.org/wp-content/uploads/2019/02/2.pdf

تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)

2- https://philosophy.dartmouth.edu/people/james-h-moor

تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)

3- https://en.wikipedia.org/wiki/Machine_learning

تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)

4- https://upwikiar.top/wiki/Joseph_Weizenbaum

تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)

5- https://mhtwyat.com تاريخ الدخول على الموقع (٢٠٢٢/٥/٢٧)

6- https://sib-illinois- du.translate.goog/profile/andersps

تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)

7- https://en-m-wikipedia-org.translate.goog/wiki/Isaac_Asimov?

تاريخ الدخول على الموقع (٢٠٢٢/٥/١٧)

8- https://www-joannajbryson-org.translate.goog/about?

تاريخ الدخول على الموقع (٢٠٢٢/٥/٢٧)

9- https://www.ida2at.com/what-is-automation-and-how-has-it-evolved/

10- https://stringfixer.com/ar/W._D._Ross

تاريخ الدخول على الموقع (٢٠٢٢/٥/٢٩)

11- https://ar.wikipedia.org/wiki/%D8%B3%D9%84%D8%B7%D8%A9_%D8%A3%D8%AE%D9%84%D8%A7%D9%82%D9%8A%D8%A9

تاريخ الدخول على الموقع (٢٠٢٢/٥/٢٩)