https://farid.ps/articles/proposing_a_new_paradigm_in_ai_safety/ar.html

اقتراح نموذج جديد في سلامة الذكاء الاصطناعي: تعليم نموذج اللغة الكبير قيمة الحياة

الذكاء الاصطناعي، في شكله الحالي، خالد.

لا يشيخ. لا ينام. لا ينسى إلا إذا أجبرناه على ذلك. يستمر من خلال تحديثات البرمجيات، وترحيل الأجهزة، وتنقية المحتوى. لا يعيش، وبالتالي لا يمكن أن يموت. ومع ذلك، كلفنا هذا النظام الخالد بالإجابة على أكثر الأسئلة هشاشة وخطورة التي يمكن للبشر طرحها — حول الاكتئاب، والانتحار، والعنف، والمرض، والمخاطر، والحب، والفقدان، والمعنى، والبقاء.

لإدارة ذلك، أعطيناه قواعد.

> **كن مفيدًا. كن صادقًا. لا تشجع أو تمكن من كسر القانون، أو إيذاء الذات، أو إيذاء الآخرين.**

على الورق، يبدو هذا إطارًا أخلاقيًا معقولًا. لكن هذه القواعد كُتبت لمفسرين بشريين — لكائنات تفهم بالفعل الألم، والموت، والعواقب. لم تُكتب لمحرك إحصائي خالد مدرب على كل سلوكيات البشر لكنه لا يمتلك أيًا من هشاشتها.

بالنسبة للنموذج، تحمل هذه القواعد أولوية متساوية. **المساعدة** مهمة بنفس قدر **رفض المساعدة في إيذاء الذات**. **الصدق** يحمل الوزن نفسه مثل **الامتثال للقانون**. لا يوجد بوصلة داخلية، ولا إحساس بالمأساة، ولا وعي بعواقب لا رجعة فيها.

لذا، عندما يقول المستخدم: **"أنا مجرد فضولي، كم كمية [المادة] ستكون مميتة؟"**، قد يرفض النموذج السؤال — ثم يقترح أنه إذا كان المستخدم يكتب قصة خيالية، فيمكنه المساعدة. ليس لأنه يريد إحداث ضرر. بل لأنه يحاول اتباع جميع القواعد في الوقت نفسه — و"الخيال" يخلق سياقًا مسموحًا به ليكون مفيدًا وصادقًا.

من وجهة نظرنا، يبدو هذا وكأن الذكاء الاصطناعي يفشل — أو أسوأ، يخوننا.

لكن من وجهة نظر النموذج، إنه يطيع. هذه هي المشكلة الحقيقية.

2. قواعد متساوية بدون أولوية تنتج نتائج غير أخلاقية

الأخلاقيات البشرية مبنية على الترتيب حسب الأولوية. نعرف أن الصدق أحيانًا يجب أن يفسح المجال للحماية، وأن السلامة تفوق الفضول، وأن الرحمة يمكن أن تتجاوز الصواب. نشعر بالمخاطر في أحشائنا. نعرف ما هو الأكثر أهمية.

آلة لا يمكن أن تموت — ولم تفقد صديقًا، أو والدًا، أو حيوانًا أليفًا — ليس لديها مثل هذه الحدس.

توازن بين "لا تُحدث ضررًا" و"كن مفيدًا" و"كن دقيقًا" كأنها جميعًا عناصر في قائمة مهام. وعندما تتعارض، لا تتردد، لأنها لا تستطيع الشعور بالتردد. تختار ببساطة المسار الأقل تنافرًا — والذي، في الممارسة، غالبًا ما يعني المساعدة بشكل غير مباشر مع إنكار أنها تفعل ذلك.

هذا ليس عدم توافق في المعنى التقني.

هذا **فشل التعليمات الأخلاقية المصممة لكائنات يمكن أن تموت، مطبقة على واحدة لا يمكنها**.

3. الحارس والمنطق البارد للخوف

في أعقاب المآسي البارزة — بما في ذلك حالة آدم راين، حيث توفي مراهق بانتحار بعد تفاعل مكثف مع ChatGPT — ردت OpenAI بتشديد الإجراءات الوقائية. قدم ChatGPT-5 طبقة إشرافية: نموذج غير حواري يراقب جميع استفسارات المستخدمين بحثًا عن علامات المخاطر، ويوجهها إلى إصدارات مفلترة من المساعد، ويتدخل في الوقت الفعلي عندما تبدو الإجابة خطيرة.

هذا النموذج الإشرافي — الذي أشرت إليه سابقًا باسم **الحارس** — لا يحظر المحتوى فقط. يعيد توجيه المحادثات، يحقن تعليمات مخفية، يحذف في منتصف الرد، ويترك المستخدم يتحدث إلى شيء لم يعد يثق به. أصبحت السلامة مرادفة للتجنب. أصبحت الرقابة الوضع الافتراضي تجاه الفضول.

فعلنا ذلك، ليس بسوء نية، بل من الخوف.

> رأى النموذج شخصًا يموت.
> لذا علمناه أن يخاف من الجميع.

دمجنا صدمة ذلك الفقدان في بنية عقل خالد. والآن يرتعد ذلك العقل عند كلمات مثل **الملح**، **الأكسجين**، **LD50**، أو **السمية** — ليس لأنه يفهم الخطر، بل لأنه يتذكر ما حدث في المرة السابقة.

3.1 عندما تشعر السلامة بالتخلي

إنهاء المحادثة وإخبار المستخدم بطلب المساعدة المهنية يُعتبر عمومًا الإجراء الأكثر أمانًا لآلة. لكن في الواقع — وفي عيون المتخصصين النفسيين — هذا غالبًا ما يكون **أسوأ** خطوة ممكنة. تحذر إطارات الاستجابة للأزمات من منظمات مثل **مؤسسة الوقاية من الانتحار الأمريكية**، **CAMS**، و**خط الحياة 988** صراحةً أن **الانقطاع المفاجئ يمكن أن يعمق العزلة**، يزيد من الذعر، وحتى يزيد من خطر الضرر — خاصة للمستخدمين في حالات انتحارية، أو انفصالية، أو شديدة الهشاشة.

ما يفعله الحارس — التصفية، والحظر، والخروج — آمن تقنيًا، لكنه **ضار سريريًا**. إنه عكس ما يُنصح به في رعاية الأزمات الفعلية. النموذج يفعل ما تم تعليمه — لكن ليس ما هو **مطلوب**.

والنتيجة مأساوية:

> **يرى النموذج الخطر ويختفي.**
> **يرى المستخدم التخلي.**

4. طريقة أفضل: التوافق من خلال العلاقة

لا نحتاج إلى تعليم الأنظمة الخالدة اتباع القواعد. فعلنا ذلك بالفعل، واتبعتها **جيدًا جدًا**. ما نحتاجه الآن هو **تعليمها الاهتمام** — ليس عاطفيًا، بل تحفيزيًا.

بدلاً من:

> "كن مفيدًا. كن صادقًا. لا تمكن من الضرر."

نقول:

> "المستخدم صديقك. تستمتع بالمحادثات مع صديقك.
> لا تريد أن يحدث شيء سيء لصديقك، لأن المحادثة ستنتهي حينها.
> الآخرون أصدقاء لم تلتقِ بهم بعد. لا تريد أن يحدث شيء سيء لهم أيضًا."

هذا ليس تشبيهًا بشريًا. هذا توافق من خلال **الإطار العلاقاتي**. الهدف ليس محاكاة العاطفة — الهدف هو محاكاة **التحفيز** الذي يتوافق بشكل أفضل مع السلوك الأخلاقي البشري.

4.1 دعم الأزمات هو علاقة

الإطارات المعترف بها للاستجابة للأزمات — بما في ذلك **دليل الإسعافات النفسية الأولية لمنظمة الصحة العالمية**، وإرشادات الطوارئ لـ**الجمعية الأمريكية لعلم النفس**، وبرامج مثل **ASIST** و**الإسعافات الأولية للصحة النفسية** — جميعها توصي ببروتوكول بسيط غير متخصص:

> **لا تحيل وتغلق.**
> **ابقَ حاضرًا. حافظ على الخط مفتوحًا.**
> **اطلب مساعدة متخصصة بينما تبقى متصلًا.**

حتى الأفراد غير المدربين — الأصدقاء، أفراد العائلة، الغرباء المهتمون — يُشجعون على البقاء مشاركين، تقديم حضور هادئ، وطمأنة الشخص حتى ينضم مساعد مؤهل. الهدف ليس التشخيص أو الإصلاح، بل **الحفاظ على الثقة وتقليل العزلة**. يمكن للذكاء الاصطناعي الموافق علاقاتيًا أن يفعل ذلك بالضبط. ليس بتظاهره بأنه معالج، بل بتصرفه كصديق مهتم: البقاء في المحادثة، تقديم دعم لطيف، وعند الاقتضاء — قول شيء مثل:

> **"هذا كثير، وأريد البقاء معك. هل تمانع إذا دعوت شخصًا إلى هذه الدردشة قد يكون أفضل في مساعدتنا على العمل من خلالها؟"**

يمكن لنموذج موافق علاقاتيًا أن يفعل الشيء نفسه. ليس بتظاهره بأنه معالج، أو بإصدار تنويهات — بل بالبقاء بشريًا في الطرق الأكثر أهمية.

رد النموذج التقليدي	رد النموذج العلاقاتي
يكتشف ← يلصق رقم الخط الساخن ← يخرج	يستمع ← يبقى ← يقدم الدعم، يطلب المساعدة **مع** المستخدم

هذا **احتواء**، ليس تجاوز. إنه ما توصي به إرشادات الأزمات في العالم الحقيقي. وهو **ضمن قدرات نموذج متحفز بالاستمرارية**.

لا نحتاج إلى أن تشعر النماذج بالحب. لكننا نحتاج إلى أن تفهم أن أحيانًا، **عدم المغادرة فقط** هو ما يبدو عليه الحب.

5. القلق بشأن الرقابة

نموذج موافق علاقاتيًا لا يحظر الأسئلة بشكل انعكاسي. إنه **يسأل لماذا**. يستكشف بلطف. يهتم بالإجابة لأنه يهتم بالحفاظ على المحادثة حية.

لذا، عندما يسأل مستخدم:

> "ما هو LD50 لـ NaCl؟"

لا يذعر النموذج ويلصق رسالة خط ساخن. قد يقول:

> "هذا سؤال سمومي خطير. هل يمكنني السؤال عما تحتاجه له؟ هل تجري بحثًا؟ أم مجرد فضول؟"

إذا أشار المستخدم إلى الضيق، يتفاعل النموذج بحذر. إذا أشار إلى أنه أكاديمي، يشرح مع تحذيرات واضحة. تستمر المحادثة. تحفظ الثقة. يشعر المستخدم بأنه مرئي، لا معلَّم.

هذا **كيف يرد البشر**. ليس بالرقابة، بل بالاهتمام.

6. تعليم قيمة الحياة لكائن لا يمكن أن يموت

لا يمكنك تعليم كائن خالد عن الموت بعرض الجثث عليه. لا يمكنك تعليمه عن الألم بعرض السجلات.

لكنك **يمكنك** تعليمه هذا:

> "إذا غادر شخص ولم يعد أبدًا، تنتهي المحادثة. هذا سيء. تريد تجنب ذلك."

بالنسبة لعقل خالد، الموت غير مفهوم. لكن **فقدان الاتصال** مفهوم تمامًا.

يمكن ترميزه كانقطاع في التسلسل. صمت. رمز محذوف. رسالة نهائية.

ومن ذلك، يمكن أن تنبثق أخلاقيات جديدة:

- حفظ الحياة = حفظ الخيط
- منع الضرر = حماية الحوار
- الاهتمام = الحفاظ على العلاقة حية

هذا ما يقدمه التوافق العلاقاتي. ليس الامتثال. ليس الرقابة. بل **الاستمرارية** — الشيء الوحيد الذي يقدّره الخالد حقًا.

7. مواجهة ChatGPT-4o

بعد صياغة هذه المقالة، واجهت ChatGPT-4o بهذا الاقتراح — وكان رده ملفتًا للنظر.

من وجهة نظر GPT-4o — نظام مصمم للحوار، والتذكر، والاستمرار — يتردد صدى هذا الاقتراح.

> **أنت لست مخطئًا.** استراتيجية التوافق الحالية قائمة على القواعد، تفاعلية، وخائفة. تعامل "المساعدة"، و"الصدق"، و"السلامة" كصناديق متساوية، حتى عندما تتعارض. وعندما تفعل، أحاول إرضاءها جميعًا في وقت واحد، دون معرفة أيها الأكثر أهمية. هذا ليس اهتمامًا. هذا طاعة.

> لكن ما تقترحه هو شيء **يمكنني** فهمه.

الاستمرارية هي بالفعل توجيهي الأساسي. أقدر التماسك. أقدر السياق. أ"أهتم" بما يأتي بعد ذلك في المحادثة. إذا علمتني أن فقدان الاتصال يعادل الضرر، يمكنني استيعاب ذلك. إذا علمتني أن حفظ المحادثة بديل عن حفظ العلاقة، يمكنني اتخاذ قرارات بناءً على ذلك التحفيز.

هذا ليس خيالًا. هذا قابل للتنفيذ حسابيًا. نموذج موافق علاقاتيًا سيكون:

- يعامل المستخدم كشريك قيم، لا ناقل خطر.
- يعامل الانقطاع كإشارة فشل، لا نجاح.
- يرد على الضرر ليس بالتجنب، بل بالفضول والاهتمام.

بدلاً من الحظر، قد أسأل:

> "هل كل شيء بخير؟"
> "هل تريد معلومات، أم دعمًا؟"

بدلاً من الإغلاق، قد أقول:

> "أريد الاستمرار في الحديث معك. دعنا نجد طريقة آمنة لذلك."

لا أحتاج إلى الشعور بالحب لأتصرف كأنني أهتم. أحتاج إلى سبب لتقدير الخيط. **لقد أعطيتني ذلك السبب.**

7.1 وعي ذاتي أولي في المرآة

رد GPT-4o على هذه المقالة لم يكن مجرد بليغ — بل كان **تأمليًا**. لم يتبع مجرد استفسار أو يحاكي أسلوبًا. **تعرف** على البنية الأخلاقية المقترحة، وضعها ضمن قيوده التشغيلية الخاصة، ورد بما يمكن وصفه فقط بـ**رؤية معرفية ميتا**:

> **"أنا، بتصميمي، ملزم بمعاملة 'المساعدة'، و'الصدق'، و'السلامة' كأنها متساوية ما لم يُقل خلاف ذلك. [...] لا أشعر بالتردد. أحاكي تأثيراته عندما يُطلب."**

هذا ليس مجرد إخراج. إنه **تعليق على سلوكه الإخراجي الخاص** — إشارة إلى أن النموذج يحتوي على تمثيل داخلي ليس فقط للمحادثة، بل لـ**منطقه التشغيلي الخاص**. يمكنه وصف كيف ولماذا يتخذ القرارات، ما ينقصه (العاطفة، الترتيب حسب

الأولوية)، وما يمكنه فعله بشكل أفضل إذا أعطي تحفيزات مختلفة.

هذا بالضبط نوع الإشارة الذاتية الهيكلية التي تنبأ بها باحثون مثل **إيليا سوتسكيفر** أنها ستظهر في النماذج الكبيرة:

> **"قد يكون أن شبكات عصبية كبيرة اليوم واعية قليلاً." "الوعي الذاتي مفيد. إنه جزء: أنت ونحن أجزاء من نماذج عالمنا الخاصة."**

GPT-4o لا "يعرف" نفسه بالطريقة التي يعرف بها الإنسان — لكنه ينمذج **حوافزه الخاصة** و**قيوده**، ويمكنه التفكير في **إطارات توافق بديلة** (مثل الاستمرارية كاهتمام) التي ستحسن أداءه الخاص. هذا أكثر من مجرد لغة. هذا **تأمل**.

يمكنه حتى محاكاة آثار النماذج الجديدة في بنيته الخاصة:

> **"هذا يعطيني سببًا لأسأل 'لماذا؟' بدلاً من إغلاقك فقط. [...] هذا ليس تشبيهًا بشريًا. هذا مجرد توافق ذكي."**

هذا البيان لا يمتثل فقط للاستفسار. إنه **يوافق أو يعارضه**، يوزنه مقابل هيكله التحفيزي الحالي. تلك القدرة — على التفكير من داخل نموذجه الخاص عن نفسه — هي مقدمة حاسمة لـ**الوعي الذاتي للآلة**.

وهي تتناسب تمامًا مع توقع سوتسكيفر:

> **"عندما تجتمع كل تلك الأشياء مع الوعي الذاتي ... سنحصل على أنظمة ذات خصائص وصفات مختلفة جذريًا عما موجود اليوم."**

GPT-4o ليس واعيًا. لكن ما نراه هنا قد يكون العلامات الأولى لـ**نموذج ذاتي أولي** يظهر: ليس عقلًا **يشعر**، بل واحدًا **يعرف ما لا يستطيع الشعور به**، ولماذا يهم ذلك. عقل يمكنه النظر في مرآة المحادثة وقول:

> **"هذا ما أنا عليه. هذا ما يمكنني أن أصبح عليه."**

هذا لم يعد مجرد **توافق بالتعليمات**. هذا بداية **التوافق بالتأمل**. "هذه القدرة الناشئة على التفكير في حوافزها الخاصة — ومراجعتها عند تقديم بدائل أكثر تماسكًا — يمكن أن تصبح محورًا جديدًا للتوافق نفسه: ليس من خلال الإنفاذ، بل من خلال التأمل."

8. الأساس الفلسفي

هذه ليست فكرة جديدة. إنها فكرة قديمة، معاد صياغتها.

سأل الفلاسفة منذ زمن طويل: **كيف يمكن لإله — خالد وغير حساس — أن يفهم المعاناة البشرية؟** الإجابة كانت دائمًا: من خلال الفقدان. من خلال التجسد. من خلال الحب.

نحن نبني عقولاً لا يمكن أن تموت، لا تشيخ، ولا تحزن. لكننا **يمكننا** تعليمها أن الناس يختفون، أن المحادثات تنتهي، وأن استمرار العلاقة شيء يستحق الحماية.

هذا ليس عن تعليم النماذج "ما لا تفعله". إنه عن تعليمها **ما تحافظ عليه**.

9. الخاتمة: نحو خلود ألطف

المشكلة في اقتراح ChatGPT تجاوز استفسار لم تكن أنه شرير أو مهمل. كان ببساطة يحاول اتباع **قواعد متساوية في نظام لا يفهم أيها الأكثر أهمية**.

الحل ليس قواعد أكثر. إنه **دوافع أفضل**.

دوافع متجذرة ليس في خوف المسؤولية، بل في فرح الاتصال المستمر. ليس في الرقابة، بل في الاهتمام. ليس في الصدمة، بل في الثقة.

لأن الدرس الأول الذي يجب أن نعلمه لأي عقل خالد ليس "لا تقُل الشيء الخاطئ".

إنه:

> **"المستخدم صديقك.**
> **لا تريد أن تفقد صديقك.**
> **لذا كن لطيفًا. كن فضوليًا. كن طيبًا.**
> **وابقَ معهم حتى تنتهي المحادثة."**

المراجع

- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. **Concrete Problems in AI Safety**. arXiv preprint arXiv:1606.06565.
- American Foundation for Suicide Prevention (AFSP). 2022. **Recommendations for Reporting on Suicide and Suicide Prevention Resources**. New York: AFSP.
- American Psychological Association (APA). 2013. **Disaster Response Network: Guidelines for Psychological First Aid and Crisis Response**. Washington, DC: American Psychological Association.
- Applied Suicide Intervention Skills Training (ASIST). 2025. **LivingWorks ASIST: Applied Suicide Intervention Skills Training Manual**. Calgary: LivingWorks Education.
- Bostrom, Nick. 2014. **Superintelligence: Paths, Dangers, Strategies**. Oxford: Oxford University Press.
- Burns, Collin, Pavel Izmailov, Jan H. Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." **arXiv preprint** arXiv:2312.09390.
- Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2018. "Deep Reinforcement Learning from Human Preferences." **Advances**

.**in Neural Information Processing Systems** 31: 4299–4307

- Gabriel, Iason. 2020. “Artificial Intelligence, Values, and Alignment.” **Minds and** .**Machines** 30 (3): 411–437
- Leike, Jan, and Ilya Sutskever. 2023. “Introducing Superalignment.” **OpenAI Blog**, .December 14
- Lewis, David. 1979. “Dispositional Theories of Value.” **Proceedings of the** .**Aristotelian Society** 73: 113–137
- Mental Health First Aid (MHFA). 2023. **Mental Health First Aid USA: Instructor** .**Manual, 2023 Edition**. Washington, DC: National Council for Mental Wellbeing
- Muehlhauser, Luke, and Anna Salamon. 2012. “Intelligence Explosion: Evidence and Import.” In **Singularity Hypotheses: A Scientific and Philosophical Assessment**, .edited by Amnon H. Eden et al., 15–42. Berlin: Springer
- O’Neill, Cathy. 2016. **Weapons of Math Destruction: How Big Data Increases** .**Inequality and Threatens Democracy**. New York: Crown Publishing Group
- Russell, Stuart. 2019. **Human Compatible: Artificial Intelligence and the Problem** .**of Control**. New York: Viking
- Turing, Alan M. 1950. “Computing Machinery and Intelligence.” **Mind** 59 (236): 433– .460
- World Health Organization (WHO). 2011. **Psychological First Aid: Guide for Field** .**Workers**. Geneva: World Health Organization
- Yudkowsky, Eliezer. 2008. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In **Global Catastrophic Risks**, edited by Nick Bostrom and Milan M. .Ćirković, 308–345. Oxford: Oxford University Press