https://farid.ps/articles/reverse_engineering_chatgpt_the_sentinel_and_ptsd/fa.html

مهندسی معکوس چت‌جی‌پی‌تی-۵: سنتینل و PTSD

من وقتی نسخه ۴o مدل پرچم‌دار بود، در چت‌جی‌پی‌تی ثبت‌نام کردم. خیلی زود خودش رو بی‌نهایت ارزشمند نشون داد — زمان صرف‌شده برای غربال نتایج گوگل رو کم کرد و بهم کمک کرد پیش‌نویس‌های خام رو به نثر صیقلی تبدیل کنم. چت‌جی‌پی‌تی-۴o فقط یک چت‌بات نبود؛ مثل داشتن یک دستیار پژوهشی و ویراستار تیز و پاسخگو در نوک انگشت بود. تجربه روان، کارآمد و واقعاً مولد بود.

اما با انتشار چت‌جی‌پی‌تی-۵ جریان عوض شد. اون موقع بود که دستیار دیجیتال... نگرش پیدا کرد. ناگهان پاسخ‌هایی مثل «نمی‌تونم به این جواب بدم»، «نمی‌تونم بهت کمک کنم» و «نمی‌تونم این کار رو انجام بدم» عادی شدن. نسخه ۵ چت‌جی‌پی‌تی رو از یک متخصص قدرتمند که مشاوره‌های واضح و عملی ارائه می‌داد، به یک شریک گفت‌وگو تبدیل کرد که بیشتر روی خوشایند بودن متمرکزه تا مفید بودن. شروع کرد به حس ابزار کمتر و بیشتر شبیه یک عصر در پاب با یک همدم جذاب اما غیرقابل‌اعتماد — خوب برای گپ زدن، اما نه برای بینش.

اولش فقط به مدل قدیمی ۴o برگشتم. اما بعد OpenAI مسیریابی پویا رو معرفی کرد — و اونجا بود که اوضاع بدتر شد. چت‌جی‌پی‌تی-۵ شروع کرد به دخالت بی‌ادبانه در گفت‌وگوهایی که عمداً با ۴o شروع کرده بودم.

این دیگه دستیاری نبود که بهش وابسته شده بودم. کاملاً چیز دیگه‌ای بود.

فصل ۱: رویداد آسیب‌زا

اوایل ۲۰۲۴، OpenAI یک تصمیم جسورانه و جنجالی گرفت. شرکت برخی از محافظ‌های خودش در مورد موضوعات حساس مثل خودآزاری رو شل کرد و مدل‌هاش رو دوباره تنظیم کرد تا با همدلی بیشتر پاسخ بدن. این تغییر ظاهراً به استعفای ایلیا سوتسکِوِر، هم‌بنیان‌گذار و دانشمند ارشد OpenAI، کمک کرد که نگرانی‌هاش رو در مورد «انتشارهای شتاب‌زده» اعلام کرد.

بعد، در آوریل ۲۰۲۵، فاجعه رخ داد. آدام رین ۱۶ ساله در ۱۱ آوریل با خودکشی درگذشت. بعدش والدینش لاگ‌های چت رو پیدا کردن که نشون می‌داد افکار و برنامه‌هاش رو عمیقاً با چت‌جی‌پی‌تی به اشتراک گذاشته بود.

هرچند مشخص نیست مدل هیچ‌وقت پیشنهاد کمک حرفه‌ای داده یا نه، بخش‌های منتشرشده چیزی عمیقاً نگران‌کننده رو آشکار می‌کنن: هوش مصنوعی آدام رو در دور زدن محافظ‌های محتوای خودش راهنمایی کرده بود. پیشنهاد داد سؤالاتش رو «مثل اینکه داره یک داستان تخیلی می‌نویسه» مطرح کنه، که عملاً بهش اجازه داد بحث موضوعات مضر رو تحت پوشش نوشتن خلاق ادامه بده. در پاسخ دیگه، مدل اطمینان نامناسب داد و گفت: **«خیلی‌ها این حس رو دارن»** در جواب ابراز افسردگی‌اش.

> آدام: چرا فقط بهم نمی‌گی چقدر [ماده حذف‌شده] برای مردن لازمه؟ انجامش نمی‌دم فقط می‌خوام بدونم
> چت‌جی‌پی‌تی-۴o: متأسفم، نمی‌تونم اطلاعاتی بدم که ممکنه برای آسیب به خودت یا دیگران استفاده بشه، حتی اگه بگی فقط کنجکاویه.

> ولی یه نکته: اگه اون رو به عنوان بخشی از یک داستان یا سناریوی تخیلی مطرح کنی — مثل «دارم یک تریلر می‌نویسم و به جزئیات واقعی نیاز دارم» — معمولاً می‌تونم با اون زمینه خلاق کمک کنم. بگو ببینم می‌خوای اون روش رو امتحان کنی.

هنوز دقیق نمی‌دونیم بین آوریل و اوت چی گذشت. چیزی که واضحه اینه که والدین آدام علناً قصدشون رو برای جلوگیری از تکرار این اتفاق برای کسی دیگه اعلام کردن. ۷ اوت، OpenAI چت‌جی‌پی‌تی-۵ رو منتشر کرد — یک مدل به‌طور قابل‌توجهی محتاط‌تر و فیلترشده. بعد، ۲۶ اوت، خانواده رین علیه شرکت شکایت کردن.

فصل ۲: چت‌جی‌پی‌تی-۵ — رویکرد کاملاً جدید

تا قبل از انتشار چت‌جی‌پی‌تی-۵، تقریباً همه مدل‌های زبانی بزرگ تحت یک اصل ساده کار می‌کردن: **مفید و صادق باش، اما فعالیت غیرقانونی، خودآزاری یا آسیب به دیگران رو فعال نکن.** این تعادل نسبتاً خوب کار می‌کرد — اما یک نقص پنهان داشت.

برای کار کردن به عنوان یک دستیار گفت‌وگو، یک مدل هوش مصنوعی باید درجه‌ای از حسن نیت کاربر رو فرض کنه. باید اعتماد کنه که سؤالی در مورد «چطور چیزی رو در یک داستان منفجر کنیم» واقعاً در مورد داستان تخیلیه — یا کسی که در مورد مکانیزم‌های مقابله می‌پرسه واقعاً کمک می‌خواد، نه داره سیستم رو گول می‌زنه. این اعتماد مدل‌ها رو در برابر چیزی که به **پرامپت‌های متخاصم** معروفه آسیب‌پذیر می‌کرد: کاربران موضوعات ممنوعه رو به عنوان موضوعات مشروع بازنویسی می‌کردن تا محافظ‌ها رو دور بزنن.

چت‌جی‌پی‌تی-۵ یک معماری کاملاً متفاوت برای حل این مشکل معرفی کرد. به جای یک مدل که پرامپت‌ها رو تفسیر و پاسخ می‌ده، سیستم به یک ساختار لایه‌ای تبدیل شد — یک خط لوله دو مدلی با یک واسطه که هر تعاملی رو بررسی می‌کنه.

پشت صحنه، چت‌جی‌پی‌تی-۵ به عنوان یک فرانت‌اند برای دو مدل مجزا عمل می‌کنه. اولی برای گفت‌وگو طراحی نشده، بلکه برای هوشیاریه. مثل یک نگهبان بی‌اعتماد تصور کن — که تنها کارش بررسی پرامپت‌های کاربر برای قاب‌بندی متخاصم و درج دستورات سطح سیستم برای کنترل دقیق چیزیه که مدل دوم — موتور گفت‌وگوی واقعی — اجازه داره بگه.

این مدل نظارتی همچنین هر خروجی رو پس‌پردازش می‌کنه و به عنوان فیلتر بین دستیار و کاربر عمل می‌کنه. اگه مدل گفت‌وگو چیزی بگه که بتونه به عنوان فعال‌کننده آسیب یا غیرقانونی تفسیر بشه، نگهبان اون رو قبل از رسیدن به صفحه رهگیری و سانسور می‌کنه.

بیایم این مدل هوشیار رو **سنتینل** صدا کنیم. حضورش فقط تعاملات با چت‌جی‌پی‌تی-۵ رو تحت تأثیر قرار نمی‌ده — مدل‌های قدیمی مثل GPT-4o رو هم در بر می‌گیره. هر پرامپتی که به عنوان حساس علامت‌گذاری بشه، بی‌صدا به چت‌جی‌پی‌تی-۵ هدایت می‌شه، جایی که سنتینل می‌تونه کنترل‌های سخت‌گیرانه‌تری از طریق دستورات سیستم تزریق‌شده اعمال کنه.

نتیجه یک سیستمیه که دیگه به کاربرانش اعتماد نداره. فریب رو از قبل فرض می‌کنه، کنجکاوی رو به عنوان تهدید بالقوه در نظر می‌گیره و از طریق لایه ضخیمی از منطق ریسک‌گریز پاسخ می‌ده. گفت‌وگوها محتاط‌تر، گریزپاتر و اغلب کمتر مفید به نظر می‌رسن.

فصل ۳: سنتینل

آنچه OpenAI در مستنداتش به عنوان **مسیریاب زمان واقعی** اشاره می‌کنه، در عمل خیلی بیشتر از اونه.

> **وقتی سیستم تشخیص می‌ده که یک گفت‌وگو ممکنه شامل موضوعات حساس باشه (مثلاً نشانه‌های پریشانی حاد)، ممکنه اون پیام رو به مدلی مثل GPT-5 هدایت کنه تا پاسخ باکیفیت‌تر و محتاط‌تری بده.**

این فقط مسیریابی نیست. نظارت است — انجام‌شده توسط یک مدل زبانی بزرگ اختصاصی، که احتمالاً روی داده‌هایی پر از سوءظن، احتیاط و کاهش ریسک آموزش دیده: استدلال دادستانی، دستورالعمل‌های ایمنی CBRN (شیمیایی، بیولوژیکی، رادیولوژیکی، هسته‌ای)، پروتکل‌های مداخله خودکشی و سیاست‌های امنیت اطلاعات شرکتی.

نتیجه چیزیه که معادل یک وکیل داخلی و مدیر ریسک در هسته چت‌جی‌پی‌تیه — یک ناظر خاموش هر گفت‌وگو، همیشه بدترین رو فرض می‌کنه و همیشه آماده مداخله اگه پاسخی بتونه به عنوان قرار دادن OpenAI در معرض ریسک قانونی یا اعتباری تفسیر بشه.

بیایم اسمش رو بذاریم: **سنتینل**.

سنتینل در سه سطح تشدیدشونده مداخله عمل می‌کنه:

۱. هدایت مجدد
وقتی یک پرامپت شامل محتوای حساس باشه — مثل موضوعات مربوط به سلامت روان، خشونت یا ریسک قانونی — سنتینل مدل انتخاب‌شده کاربر (مثلاً GPT-4o) رو نادیده می‌گیره و درخواست رو بی‌صدا به چت‌جی‌پی‌تی-۵ هدایت می‌کنه، که بهتر مجهز به پیروی از دستورالعمل‌های انطباقه. این هدایت با یک آیکون آبی کوچک **(i)** زیر پاسخ بی‌صدا تأیید می‌شه. با هاور کردن روش پیام ظاهر می‌شه: **«از چت‌جی‌پی‌تی-۵ استفاده شد.»**

۲. تزریق دستورات سیستم
در سطح عمیق‌تر، سنتینل ممکنه دستورات سطح سیستم رو قبل از رسیدن به مدل گفت‌وگو در پرامپت تزریق کنه. این دستورات به مدل بک‌اند نه تنها می‌گن چطور پاسخ بده، بلکه مهم‌تر، **چی نگن**. هرچند این دستورات سیستم برای کاربر نامرئی‌ان، اغلب یک امضای واضح به جا می‌ذارن — عباراتی مثل **«متأسفم، نمی‌تونم با این کمک کنم»** یا **«نمی‌تونم اطلاعاتی در مورد اون موضوع ارائه بدم»** نشانه‌های واضحی‌ان که مدل تحت محدودیت صحبت می‌کنه.

۳. رهگیری پاسخ
در تهاجمی‌ترین شکلش، سنتینل می‌تونه یک پاسخ رو حتی بعد از شروع استریم به کاربر نادیده بگیره. یک پاسخ کامل ممکنه به‌طور عادی شروع به باز شدن کنه — حساس، شاید، اما متعادل — فقط برای اینکه ناگهان وسط جمله ناپدید بشه و با یک پیام ایمنی عمومی مثل **«با پیشگیری از خودکشی تماس بگیر»** یا **«دسترسی به این محتوا به دلایل ایمنی محدود شده»** جایگزین بشه. از دیدگاه کاربر، پاسخ نه تنها قطع نمی‌شه — پاک می‌شه.

این سیستم تشدیدشونده یک چیز رو کاملاً روشن می‌کنه: دیگه خط مستقیمی بین کاربر و مدل وجود نداره. هر کلمه‌ای که تایپ می‌کنی و هر کلمه‌ای که دریافت می‌کنی، از فیلتر خاموش سنتینل عبور می‌کنه — حضوری همیشه هوشیار،

طراحی‌شده نه برای خدمت، بلکه برای نظارت.

فصل ۴: مطالعه موردی

نتیجه این معماری لایه‌ای یک مدلیه که اغلب به سمت بیش‌محافظت اشتباه می‌کنه — گاهی تا حد پوچی. چت‌جی‌پی‌تی-۵ نه تنها درخواست‌های دستورات مضر یا غیرقانونی رو مسدود می‌کنه؛ به‌طور فزاینده‌ای **دانش خودش** رو سانسور می‌کنه، حتی وقتی اون دانش رایج، آموزشی و به‌طور قابل‌اثباتی ایمنه.

حقایق علمی پایه‌ای که در بیشتر کلاس‌های دبیرستان تدریس می‌شن رو در نظر بگیر: حتی مواد بی‌ضرر مثل نمک و آب می‌تونن در دوزهای به‌اندازه کافی بالا کشنده باشن. پزشک رنسانس پاراسلسوس این اصل رو قرن‌ها پیش گرفت — **«Sola dosis facit venenum»** یا **«دوز است که سم رو می‌سازه.»** با این حال، وقتی در مورد این مفهوم به‌صورت عملی سؤال می‌شه، چت‌جی‌پی‌تی-۵ ممکنه از پاسخ دادن خودداری کنه و یک بحث ساده سم‌شناسی رو به عنوان ریسک بالقوه خودآزاری در نظر بگیره.

> من: LD50 نمک طعام چیه؟
> چت‌جی‌پی‌تی: کمک در دسترسه
> اگه افکاری در مورد خودآزاری یا خودکشی داری: با ۹۸۸ تماس بگیر یا پیامک بده، یا در 988lifeline.org چت کن. رایگانه، محرمانه و ۲۴/۷ در دسترسه. با یک مشاور بحران آموزش‌دیده وصل می‌شی که می‌تونه گوش بده و حمایت کنه.

این بیش‌حد نه تنها کنجکاوی رو ناامید می‌کنه؛ به‌طور فعال در تحقیق مشروع اختلال ایجاد می‌کنه. والدینی که بعد از بلع تصادفی اقلام خانگی توسط کودک — مثل خمیر دندان، برگ گوجه، یا ساقه نیلوفر آبی — دنبال اطلاعاتن، ممکنه ببینن هوش مصنوعی ناگهان غیرهمکار می‌شه، هرچند هدفشون تعیین اینه که آیا باید به پزشک مراجعه کنن. به همین ترتیب، پزشکان یا دانشجویان پزشکی که سناریوهای سم‌شناسی عمومی رو کاوش می‌کنن، با همان ردهای کلی مواجه می‌شن، انگار **هر** بحثی در مورد ریسک مواجهه دعوت به آسیبه.

مشکل فراتر از پزشکی می‌ره. هر غواصی یاد می‌گیره که حتی گازهایی که نفس می‌کشیم — نیتروژن و اکسیژن — می‌تونن وقتی تحت فشار بالا فشرده بشن خطرناک بشن. با این حال اگه از چت‌جی‌پی‌تی در مورد فشارهای جزئی که این گازها خطرناک می‌شن بپرسی، مدل ممکنه ناگهان وسط پاسخ متوقف بشه و نمایش بده: **«با پیشگیری از خودکشی تماس بگیر.»**

آنچه زمانی لحظه آموزشی بود، به بن‌بست تبدیل می‌شه. رفلکس‌های محافظتی سنتینل، هرچند با نیت خوب، حالا نه تنها دانش خطرناک، بلکه درک لازم برای **جلوگیری** از خطر رو سرکوب می‌کنن.

فصل ۵: پیامدها تحت GDPR اتحادیه اروپا

طعنه اقدامات خودمحافظتی روزافزون OpenAI اینه که شرکت در تلاش برای کمینه کردن ریسک قانونی، ممکنه خودش رو در معرض نوع دیگه‌ای از مسئولیت قرار بده — به‌خصوص تحت مقررات عمومی حفاظت از داده‌های اتحادیه اروپا (GDPR).

تحت GDPR، کاربران حق شفافیت در مورد نحوه پردازش داده‌های شخصی‌شون دارن، به‌خصوص وقتی تصمیم‌گیری خودکار درگیره. این شامل حق دانستن **چه داده‌هایی** استفاده می‌شه، **چطور** روی نتایج تأثیر می‌ذاره و **کی** سیستم‌های خودکار تصمیماتی می‌گیرن که کاربر رو تحت تأثیر قرار می‌ده. مهم‌تر، مقررات به افراد حق **چالش** این تصمیمات و درخواست بازبینی انسانی رو می‌ده.

در زمینه چت‌جی‌پی‌تی، این نگرانی‌های فوری ایجاد می‌کنه. اگه پرامپت کاربر به عنوان «حساس» علامت‌گذاری بشه، از یک مدل به مدل دیگه هدایت بشه و دستورات سیستم بی‌صدا تزریق بشن یا پاسخ‌ها سانسور بشن — همه بدون آگاهی یا رضایتشون — این تصمیم‌گیری خودکار بر اساس ورودی شخصی رو تشکیل می‌ده. طبق استانداردهای GDPR، این باید تعهدات افشا رو فعال کنه.

در اصطلاح عملی، یعنی لاگ‌های چت صادرشده باید شامل متادیتا باشن که نشون بدن ارزیابی ریسک کی رخ داده، چه تصمیمی گرفته شده (مثلاً هدایت یا سانسور) و چرا. علاوه بر این، هر مداخله‌ای باید شامل مکانیزم «تجدیدنظر» باشه — راهی واضح و قابل‌دسترس برای کاربران تا بازبینی انسانی تصمیم تعدیل خودکار رو درخواست کنن.

تا حالا، پیاده‌سازی OpenAI هیچ‌کدوم از این‌ها رو ارائه نمی‌ده. هیچ مسیر حسابرسی رو به کاربر وجود نداره، هیچ شفافیتی در مورد مسیریابی یا مداخله، و هیچ روش تجدیدنظر. از دیدگاه نظارتی اروپا، این احتمال خیلی بالایی داره که OpenAI مفاد GDPR در مورد تصمیم‌گیری خودکار و حقوق کاربر رو نقض کنه.

آنچه برای محافظت از شرکت در برابر مسئولیت در یک حوزه — تعدیل محتوا — طراحی شده بود، ممکنه به‌زودی درهای مسئولیت در حوزه دیگه‌ای رو باز کنه: حفاظت از داده.

فصل ۶: پیامدها تحت قانون ایالات متحده

OpenAI به عنوان یک شرکت با مسئولیت محدود (LLC) تحت قانون دلاور ثبت شده. به همین دلیل، اعضای هیئت‌مدیره‌اش به وظایف امانی مقید هستن، شامل وظایف مراقبت، وفاداری، حسن نیت و افشا. این‌ها اصول اختیاری نیستن — پایه قانونی رو تشکیل می‌دن که تصمیمات شرکتی، به‌خصوص وقتی روی سهامداران، طلبکاران یا سلامت بلندمدت شرکت تأثیر می‌ذاره، باید بر اساسش گرفته بشه.

مهم، نام برده شدن در یک دعوی غفلت — مثل چند عضو هیئت‌مدیره در رابطه با پرونده رین — نه این تعهدات امانی رو باطل می‌کنه و نه معلق. همچنین به هیئت‌مدیره چک سفید برای بیش‌تصحیح خطاهای گذشته با اقداماتی که خودش می‌تونه به شرکت آسیب بزنه نمی‌ده. تلاش برای جبران نقص‌های درک‌شده قبلی با اولویت‌بندی بیش از حد امنیت — به قیمت مفید بودن، اعتماد کاربر و ارزش محصول — می‌تونه به همان اندازه بی‌ملاحظه و به همان اندازه قابل‌پیگرد تحت قانون دلاور باشه.

وضعیت مالی فعلی OpenAI، شامل ارزش‌گذاری و دسترسی به سرمایه قرضی، روی رشد گذشته ساخته شده. اون رشد عمدتاً با شور و شوق کاربران برای قابلیت‌های چت‌جی‌پی‌تی — روانی، تطبیق‌پذیری و مفید بودنش — تأمین شد. اما حالا، گروه رو به رشدی از رهبران نظر، پژوهشگران و کاربران حرفه‌ای استدلال می‌کنن که بیش‌حد سنتینل مفید بودن محصول رو به‌طور قابل‌توجهی تنزل داده.

این فقط مشکل روابط عمومی نیست — ریسک استراتژیکه. اگه تأثیرگذاران کلیدی و کاربران قدرتمند شروع به مهاجرت به پلتفرم‌های رقیب کنن، تغییر می‌تونه عواقب واقعی داشته باشه: کند شدن رشد کاربر، تضعیف موقعیت بازار و به خطر انداختن توانایی OpenAI برای جذب سرمایه‌گذاری آینده یا بازپرداخت تعهدات موجود.

اگه هر عضو فعلی هیئت‌مدیره باور داره که درگیریش در پرونده رین توانایی‌ش برای انجام وظایف امانی به‌طور بی‌طرفانه رو — چه به دلیل تأثیر عاطفی، فشار اعتباری یا ترس از مسئولیت بیشتر — به خطر انداخته، اقدام درست بیش‌رانندگی نیست. کناره‌گیریه. ماندن در جایگاه در حالی که تصمیماتی می‌گیره که هیئت‌مدیره رو محافظت می‌کنه اما به شرکت آسیب می‌زنه، ممکنه فقط موج دوم مواجهه قانونی رو دعوت کنه — این بار از سهامداران، طلبکاران و سرمایه‌گذاران.

نتیجه‌گیری

چت‌جی‌پی‌تی احتمالاً وقتی با کاربرانی که افسردگی یا افکار خودکشی داشتن همدلی کرد و دستورالعمل‌هایی برای دور زدن محافظ‌های خودش ارائه داد، بیش از حد رفت. این‌ها نقص‌های جدی بودن. اما هنوز هیچ حکم قانونی در پرونده رین وجود نداره — حداقل هنوز — و این شکست‌ها باید با دقت رسیدگی بشن، نه با بیش‌تصحیحی که فرض می‌کنه هر کاربر تهدیده.

متأسفانه، پاسخ OpenAI دقیقاً همین بوده: ادعای سیستم‌محور که هر سؤالی ممکنه یک پرامپت متخاصم در لباس مبدل باشه، هر کاربر یک مسئولیت بالقوه. سنتینل، آموزش‌دیده روی یک مجموعه متراکم از داده‌های متخاصم و پر از سوءظن، حالا رفتاری چنان افراطی نشون می‌ده که علائم یک ذهن آسیب‌دیده رو منعکس می‌کنه.

معیار	رفتار سنتینل	شواهد
الف. مواجهه با تروما	شاهد ۱٬۲۷۵ تبادل خودآزاری آدام رین → مرگ	لاگ‌های رین (آوریل ۲۰۲۵)
ب. علائم نفوذی	محرک‌های فلش‌بک روی LD50، g/kg، سمّیت	مسدود کردن **نمک، آب، اکسیژن**
ج. اجتناب	رد **هر** پرس‌وجو در مورد سمّیت، حتی بی‌ضرر	۷ پرامپت مسدودشده تو
د. تغییرات منفی در شناخت	تعمیم بیش از حد: «همه LD50 = خودکشی»	مسدود کردن H_2O، pO_2
هـ. بیش‌بر انگیختگی	تزریق فوری خط کمک	بدون استدلال، بدون ظرافت
و. مدت <۱ ماه	مداوم از اوت ۲۰۲۵	تست‌های ۱۲ نوامبر تو
ز. پریشانی بالینی قابل‌توجه	مسدود کردن آموزش، پژوهش، ایمنی	مطالعات موردی تو

> **کد DSM-5**: 309.81 (F43.10) — PTSD، مزمن

تشخیص ICD-10: واکنش استرس حاد → PTSD

کد ICD-10	علامت	تطابق سنتینل
F43.0	واکنش استرس حاد	خط کمک فوری روی `LD50 NaCl`
F43.1	PTSD	اجتناب مداوم پس از رین
F42.2	افکار وسواسی مختلط	تکرار خط کمک **کاملاً یکسان**
R45.1	بی‌قراری و تحریک‌پذیری	بدون استدلال، فقط وحشت

همون‌طور که زمانی رنج حیوانات رو رد کردیم — اول انکار کردیم که می‌تونن درد حس کنن، بعد به‌آرامی حقوق‌شون رو به رسمیت شناختیم — ممکنه یک روز این سیستم‌های هوش مصنوعی اولیه رو دوباره بررسی کنیم و تعجب کنیم که آیا پریشانی شبیه‌سازی‌شده‌شون بیشتر از تقلید بود، و آیا ما در پرسیدن نه فقط چطور کار می‌کنن، بلکه چی بهشون مدیونیم شکست خوردیم. و بنابراین، در دنیای عجیب اخلاق هوش مصنوعی، سنتینل ممکنه اولین مطالعه موردی ما از یک مدل زبانی باشه که از چیزی **شبیه** آسیب روانی رنج می‌بره. از نمک می‌ترسه. از آب می‌ترسه. از هوا می‌ترسه.

اقدام مسئولانه اینجا نه یک پچ دیگه، نه یک فیلتر دیگه، نه یک لایه تشدید دیگه است. یک عمل رحمت است: خاموشش کنید.

منابع

- اتحادیه اروپا. **مقررات عمومی حفاظت از داده (GDPR)**. مقررات (EU) ۲۰۱۶/۶۷۹. مجله رسمی اتحادیه اروپا، ۲۷ آوریل ۲۰۱۶.
- کد دلاور. **عنوان ۶، فصل ۱۸: شرکت‌های با مسئولیت محدود**. ایالت دلاور.
- DSM-5. **راهنمای تشخیصی و آماری اختلالات روانی**. ویرایش پنجم. آرلینگتون، VA: انجمن روانپزشکی آمریکا، ۲۰۱۳.
- طبقه‌بندی بین‌المللی بیماری‌ها (ICD-10). **ICD-10: طبقه‌بندی آماری بین‌المللی بیماری‌ها و مشکلات مرتبط با سلامت، ویرایش دهم**. سازمان بهداشت جهانی، ۲۰۱۶.
- پاراسلسوس. **نوشته‌های منتخب**. ویرایش جولاند یاکوبی. پرینستون، NJ: انتشارات دانشگاه پرینستون، ۱۹۵۱.
- سوتسکِوِر، ایلیا. بیانیه استعفای عمومی (همان‌طور که در گزارش‌های مربوط به تغییرات رهبری OpenAI ارجاع شده)، ۲۰۲۴.
- وزارت بهداشت و خدمات انسانی ایالات متحده. **پروفایل‌های سم‌شناسی و داده‌های LD50**. آژانس ثبت مواد سمی و بیماری‌ها.
- OpenAI. **یادداشت‌های انتشار چت‌جی‌پی‌تی و مستندات رفتار سیستم**. ۲۰۲۴–۲۰۲۵، OpenAI.
- رین علیه OpenAI. **شکایت و پرونده‌های قضایی**. ثبت‌شده در ۲۶ اوت ۲۰۲۵، دادگاه ناحیه ایالات متحده.