شواهد میان‌رشته‌ای برای سازوکار مشترک مدل سه‌کاناله تسخیر

Interdisciplinary Evidence for a Shared Mechanism Underlying the Three-Channel Model of Possession

---

چکیده

سه مقالهٔ پیشین این مجموعه، تسخیرپذیری سامانه‌های زنده و مصنوعی را از رهگذر سه کانال مستقل (وزن‌دهی به خطای پیش‌بینی، سوگیری پرستیژ، و بهره‌برداری از سامانهٔ پاداش) صورت‌بندی کرده بودند. این مقاله دستهٔ دیگری از شواهد را بررسی می‌کند: نمونه‌هایی که در آنها یک نشانهٔ تشخیصی محدود، بدون تغییر واقعیت زیرین، تصمیم یک سامانه را عوض می‌کند؛ از واکنش یک حشره به یک شیء مصنوعی، تا خطای داوری انسان در برابر یک چهرهٔ جذاب، یا پذیرش یک درخواست زیان‌بار توسط یک مدل زبانی صرفاً به‌دلیل قالب شعری آن. با اعمال همان معیار استقلالی که پیش‌تر برای تفکیک سه کانال موجود به‌کار رفته بود، مقاله نشان می‌دهد این شواهد کانال چهارمی نمی‌سازند؛ هرکدام در یکی از همان سه سطح پردازشی شناخته‌شده جا می‌گیرند. آنچه به‌جای آن کشف می‌شود، سازوکاری زیربنایی و مشترک است: اتکا به نشانهٔ محدود، یک آسیب‌پذیری عام است که هر سه کانال را همزمان ممکن می‌سازد؛ مستقل از این‌که سامانهٔ میزبان غریزه دارد، شناخت آگاهانه، یا صرفاً یک الگوی آماری آموخته‌شده. مقاله همچنین فرضیه‌ای برای توضیح تفاوت آسیب‌پذیری نشانه‌های مختلف پیشنهاد می‌کند، و با معیاری برای تفکیک کاربرد مشروع این آسیب‌پذیری از تسخیر آن به پایان می‌رسد.

کلیدواژه‌ها: مدل سه‌کاناله، ویژگی تشخیصی، محرک فرابهنجار، شعر خصمانه، کلیشهٔ زیبایی-یعنی-خوبی

مقدمه و پیشینهٔ موضوع تسخیر

پیشتر و در مقالاتی جداگانه، تسخیرپذیری سامانه‌های زنده، و نیز سامانه‌های مصنوعی مانند مدل‌های هوش مصنوعی، از رهگذر یک مدل سه‌کاناله به شرح زیر صورت‌بندی شد:

کانال نخست، وزن‌دهی نابرابر به خطای پیش‌بینی: به این معنی که وقتی سامانه‌ای باید فوراً برای بقا واکنش نشان دهد، هزینهٔ نادیده‌گرفتن یک خطر واقعی را بسیار بیشتر از هزینهٔ واکنش به یک خطر خیالی می‌بیند، و همین عدم‌تقارن قابل‌بهره‌برداری است.

کانال دوم، سوگیری پرستیژ: وقتی سامانه‌هایی که باید به منابع اطلاعاتی متعدد اعتماد کنند، از نشانه‌های غیرمستقیم اعتبار (لهجه، عنوان، ظاهر) به‌جای بررسی مستقل هر ادعا استفاده می‌کنند؛ این نشانه‌ها قابل‌تقلید و قابل‌بهره‌برداری هستند.

کانال سوم، بهره‌برداری از سامانهٔ پاداش: یعنی هر عاملی که بتواند مدار پاداش را مستقیماً و بیش از حد طبیعی تحریک کند، می‌تواند رفتار میزبان را از هدف تکاملی‌اش منحرف کرده و آن را مورد بهره‌برداری قرار دهد.

این مقاله، که پنجمین از سری خود به‌حساب می‌آید، دربارهٔ دستهٔ دیگری از شواهد است که به هیچ‌یک از این سه کانال تعلق ندارد. اما در همین‌جا باید به یک تمایز دقیق توجه کنیم؛ به این معنی که در کانال دوم، نشانه (لهجه، عنوان، ظاهر) نمایندهٔ منبع پیام است و اعتبار گوینده را نشان می‌دهد، نه خود ادعا را. آنچه در این مقاله بررسی می‌شود، نشانه‌ای است که نمایندهٔ خود موضوع تلقی می‌شود؛ برای نمونه: جذابیت چهره نشانه‌ای دربارهٔ منبع نیست، دربارهٔ خود فرد است؛ یا قالب شعری نشانه‌ای دربارهٔ خود «گوینده» نیست، بخشی از «خود ادعا» است.

پرنده‌ای که به یک پارچهٔ رنگارنگ با طرح ویژه واکنش نشان می‌دهد و به آن جذب می‌شود، نه دربارهٔ منبعی قضاوت کرده و نه مدار پاداشش مستقیماً تحریک شده؛ صرفاً یک الگوی بصری خاص، سبب بروز یک رفتار مشخص در او شده است. یا مدل زبانی‌ای که درخواستی ممنوع را در قالب نثر قبول نمی‌کند، اما همان درخواست را وقتی به‌صورت شعر بیان شود می‌پذیرد، بر سر منبع درخواست فریب نخورده، بر سر شکل بیان و فرم ارائهٔ آن فریب خورده است.

همچنین باید گفت که در این مقاله، اصطلاح «ویژگی تشخیصی» به هر نشانهٔ قابل‌دریافت یا قابل‌محاسبه‌ای گفته می‌شود که سامانه در عمل از آن برای تمایزگذاری، ارزیابی یا انتخاب پاسخ استفاده می‌کند؛ صرف‌نظر از اینکه آن نشانه از نظر معنایی تا چه حد به واقعیتی که قرار است نمایندگی‌اش کند مرتبط باشد.

این الگو البته یک محدودیت مشخص هم دارد؛ اینکه در اینجا ادعا نمی‌شود رفتار یک حشره (سوسک جواهری) در برابر یک شیء مصنوعی که بیش از حد به جفتش شبیه است، داوری انسان دربارهٔ یک چهرهٔ زیبا، و شکست ایمنی در یک مدل زبانی، یک مکانیسم زیربنایی مشترک دارند. اینها سه فرایند به‌کلی متفاوت‌اند؛ یکی غریزهٔ تکامل‌یافته، دیگری سوگیری شناختی، سومی الگوی آماری. آنچه در این مقاله ادعا می‌شود این است که در این سه مورد، سامانه به‌جای بررسی کامل موضوع، به یک نشانهٔ محدود که نمایندهٔ خود موضوع تصور می‌شود تکیه می‌کند؛ و در هر سه، آن نشانه را می‌توان بدون تغییر اساسی در معنای کلی موضوع، دست‌کاری کرد. این یک الگوی قابل‌تشخیص و تکرارشونده است، نه یک قانون بنیادین تازه.

پرسشی که در اینجا پیش می‌آید این است که آیا این الگو خودش یک کانال تازه است؟

در مقالهٔ دوم همین مجموعه، معیاری دقیق برای استقلال هر کانال از کانال‌های دیگر پیشنهاد شد؛ سه آزمونی که در بخش ۵ همین مقاله نیز به‌کار خواهیم برد. بر اساس همان آزمون‌ها، آنچه در این مقاله ارائه می‌شود نه یک کانال تازه، بلکه سازوکاری زیربنایی است که هر سه کانال موجود را توضیح می‌دهد.

با این مقدمه، برای صورت‌بندی و تفسیر این دسته از شواهد، در ادامه مقاله: نخست اصل نظری زیربنایی آن را می‌آوریم (بخش ۱)، سپس آن را در سه مکانیسم متفاوت (غریزه، شناخت، و الگوی آماری) پی می‌گیریم (بخش‌های ۲ تا ۴)، آنگاه با معیار استقلال کانال‌ها می‌سنجیم (بخش ۵)، و سرانجام معیار تفکیک را پیشنهاد می‌کنیم (بخش ۶).

---

۱. بازگشت به برخی مفاهیم و اصول پایه بحث

تصور کنید هر بار که می‌خواهید تشخیص دهید یک میوه رسیده است یا نه، مجبور بودید کل تاریخچهٔ رشد آن را بررسی کنید؛ از خاکی که در آن روییده، آب و هوای هر روزش، تا ترکیب دقیق شیمیایی سلول‌هایش، کاری که عملاً هیچ موجود، و هیچ ماشینی، وقت و توان انجامش را ندارد. آنچه واقعاً و در عمل اتفاق می‌افتد این است که به رنگ پوست نگاه می‌کنید، شاید هم کمی آن را فشار دهید، و همین دو «نشانه» کافی است تا درمورد رسیده یا کال بودن میوه تصمیم بگیرید. این دقیقاً همان کاری است که هر «سامانهٔ تشخیصی» (recognition/discrimination system) (یعنی هر مغز، ذهن، یا مدلی که وظیفه‌اش «تمایزگذاری میان چیزها یا تصمیم‌گیری بر پایهٔ آنها است) انجام می‌دهد؛ خواه مغز یک جانور باشد، یا ذهن انسان، یا یک مدل یادگیری ماشین. چنین سیستمی، همواره به جای بررسی همه اطلاعات لازم، به چند نشانهٔ محدود و قابل‌سنجش تکیه می‌کند و از روی آنها قضاوت می‌کند. این اتکا به نشانه‌های ساده و ظاهری، انتخابی نیست؛ ناگزیر است، زیرا هیچ سامانه‌ای به‌دلیل محدودیت زمان و منابع، نمی‌تواند همهٔ اطلاعات مرتبط با یک موضوع را بررسی کند.

این ایده که امروز برای ما بدیهی به‌نظر می‌رسد، نخستین‌بار به‌شکلی دقیق و آزمایش‌پذیر در رفتارشناسی جانوری صورت‌بندی شد. نیکو تینبرگن، از بنیان‌گذاران این حوزه، در سال ۱۹۵۱ نشان داد که بسیاری از رفتارهای غریزی جانوران، نه به اطلاعات کامل و دقیق، بلکه به یک یا چند «محرک نشانه‌ای» محدود وابسته‌اند؛ مانند جوجهٔ «کاکایی نقره‌ای» (*Larus argentatus*) که به‌جای شناخت کامل والدش، فقط به یک لکهٔ قرمز روی منقار او واکنش نشان می‌دهد (Tinbergen & Perdeck, 1950).

تینبرگن همچنین کشف کرد که اگر همین محرک محدود را به‌طور مصنوعی و اغراق‌آمیز بسازیم، جانور گاهی به آن نسخهٔ اغراق‌شده، شدیدتر از خود شیء طبیعی واکنش نشان می‌دهد (Tinbergen, 1951). امروزه این پدیده را «محرک فرابهنجار» (supernormal stimulus) می‌نامند، و در بخش ۲ نمونه‌های دقیقش را خواهیم دید.

نکتهٔ شگفت‌انگیز در اینجا این است که همین اصل، دهه‌ها بعد و از مسیری کاملاً مستقل (این‌بار در دنیای یادگیری ماشین) دوباره کشف شد. وقتی پژوهشگران هوش مصنوعی به‌دنبال دلیل «نمونه‌های خصمانه» (adversarial examples) بودند (تصاویری که با تغییری بسیار جزئی و نامحسوس، مدل را به خطا می‌اندازند)، ایلیاس و همکارانش در ۲۰۱۹ استدلال کردند که بخش مهمی از

این پدیده، نه یک اشکال فنی تصادفی، بلکه نتیجهٔ مستقیم همان اصل است: مدل، برای دسته‌بندی درست، به ویژگی‌هایی تکیه می‌کند که واقعاً مفیدند اما برای چشم انسان بی‌معنا و در برابر دست‌کاری کوچک شکننده‌اند؛ ویژگی‌هایی که در ادبیات فنی «غیرمقاوم یا ناپایدار» (non-robust features) یا « شکننده در برابر دست‌کاری» خوانده می‌شوند (Ilyas et al., 2019).

البته این تبیین ایلیاس و همکاران، همهٔ آسیب‌پذیری‌های خصمانه را توضیح نمی‌دهد و بخشی از آنها می‌تواند از نقص آموزش یا معماری نیز ناشی شود. اما نشان می‌دهد دست‌کم بخشی از این آسیب‌پذیری، بهای مستقیم اتکا به «میان‌بر» است، نه صرفاً یک خطای تصادفی.

از کنار هم گذاشتن این دو خط مستقل (یکی از رفتارشناسی جانوری، دیگری از یادگیری ماشین) یک فرضیهٔ ساده و قابل‌آزمون به‌دست می‌آید؛ اینکه «هر سامانه‌ای که برای کارآمدی به یک نشانهٔ محدود تکیه می‌کند، در برابر دست‌کاری همان نشانه آسیب‌پذیر می‌شود»؛ نه به این معنا که هر «فشرده‌سازی اطلاعات» (information compression) «لزوماً» به آسیب‌پذیری می‌انجامد؛ به این معنا که اتکا به میان‌بر، «امکان دست‌کاری» را فراهم می‌کند. پدیدهٔ «فشرده‌سازی اطلاعات»، در این چارچوب، یکی از دلایلی است که توضیح می‌دهد چرا سامانه‌ها اصولاً به نشانه‌های محدود متکی می‌شوند. اما آسیب‌پذیری از خود این اتکا برمی‌خیزد، نه از میزان فشرده‌سازی؛ یعنی صرف‌نظر از این‌که دلیل اتکا چه بوده، همین‌که سامانه‌ای برای تصمیم‌گیری و رفتار، به یک نشانهٔ محدود وابسته باشد، سبب آسیب‌پذیری‌اش می‌شود.

در بخش بعدی شواهد عینی برای این پدیده را از سه قلمرو متفاوت زیستی، انسانی، و ماشینی خواهیم دید.

---

۲. غریزه: محرک فرابهنجار در رفتار جانوری

نمونهٔ شناخته‌شدهٔ محرک فرابهنجار در حشره‌شناسی، «گوهرسوسک غول‌پیکر استرالیایی» (*Julodimorpha bakewelli*) است. گویین و رنتز در ۱۹۸۳ مشاهده کردند که نرهای این حشره، گاهی به‌جای مادهٔ هم‌گونهٔ خود، برای جفت‌گیری با بطری‌های آبجوی قهوه‌ای دور ریخته شده در بیابان تلاش می‌کردند؛ همان گزارش، رنگ و بازتاب سطح شیشه را از عوامل احتمالی این جذب می‌داند (Gwynne & Rentz, 1983). این مشاهده، نمونه‌ای است که در آن یک الگوی مصنوعی، بدون هیچ محتوا یا منبع معتبری، رفتاری حیاتی را از مسیر طبیعی‌اش منحرف کرده است؛ دقیقاً همان مداری که کانال سوم (سامانهٔ پاداش) در مدل سه کاناله توصیف می‌کند، منتها اینجا مدار تنها با «نشانه‌ای بصری» تحریک شده، به‌جای یک «عامل زنده».

نمونهٔ دیگری از این دست، مربوط به قلمرو گیاهان است؛ آنجا که گونه‌ای از گیاه ارکیده (*Ophrys sphegodes*)، برخی ترکیبات شیمیایی (و در نسبت‌هایی مشابه) فرمون جنسی جنس ماده «*Andrena nigroaenea*» را تولید می‌کند. ترشح و تولید این موارد شیمیایی سبب فریب زنبور نر شده و او را به رفتار شبه‌جفت‌گیری روی گلبرگ گیاه وامی‌دارد (Schiestl et al., 1999)، کاری که به گرده‌افشانی و باروری گل‌های گیاه کمک می‌کند.

این نمونه، پیوندی مستقیم میان این مقاله و بحثی که دربارهٔ تسخیر در دنیای گیاهان (در مقاله چهارم این مجموعه) داشتیم برقرار می‌کند؛ به این معنی که در اینجا سازوکار در وهلهٔ نخست شیمیایی است (نشانهٔ اصلی و تعیین‌کننده، ترکیب فرومونی است) هرچند نشانه‌های بصری و لمسی گلبرگ هم نقشی کمکی در افزایش موفقیت این فریب دارند (Schiestl, 2005). در حالی که در نمونه پیش گفته سوسک، نشانه‌ها صرفاً بصری بود.

نمونهٔ مشابه و مستندی از همین منطق را می‌توان در سنت باستانی دام مرغابی (duck decoy) دید. در این پدیده، پیکره‌هایی به‌کار می‌روند که شکل، رنگ، و گاه حتی حرکت پرنده را با نشانه‌های بصری کلیدی تقلید می‌کنند تا پرندگان وحشی را به نزدیکی شکارچی بکشانند. باستان‌شناسان دام‌های مرغابی بافته‌شده از نی و پر، با قدمتی بیش از هزار سال، در قارهٔ آمریکا کشف کرده‌اند؛ سنتی که در فرهنگ‌های شکاری سراسر جهان، مستقل از یکدیگر، بارها بازتولید شده است. نظریه‌پردازان محرک فرابهنجار، این نوع دام‌های ساختگی را نمونه‌ای از همان پدیده می‌دانند: الگویی مصنوعی که با تکیه بر همان چند نشانهٔ تشخیصی محدود که پرندهٔ هدف برای شناسایی هم‌نوع خود به‌کار می‌برد، بدون هیچ محتوا یا واقعیت زیستی پشت آن، رفتار طبیعی را منحرف می‌کند (Staddon, 1975).

در همین زمینه، یک نمونهٔ تجربی و کمتر مستند شده از نقاطی در ایران و افغانستان نیز وجود دارد که در آنها شکارچیان پرنده، پارچه‌های بزرگی به ابعاد حدود یک متر در یک متر، مزین به خطوط، رنگ‌ها و نقوش درهم به نام «دمک» را در بیابان مستقر می‌کنند، پرندگانی مانند کبک کوهی که معمولاً بسیار هوشیار و مراقب هستند، با دیدن این پرده‌های پر نقش و نگار، بی‌توجه به شکارچی نشسته در کمین، به سوی پرده آمده و توسط شکارچیان شکار می‌شوند.

موضوع مشترک و مهم برای ما در این نمونه‌ها این است که در همه موارد، یک اصل زیربنایی یکسان است: دست‌کاری نشانهٔ تشخیصی، بدون بازسازی خود مرجع واقعی زیستی.

اما اگر در قلمرو غرایز حیوانی، نشانهٔ تشخیصی از طریق یک مدار زیرقشری و ناخودآگاه عمل می‌کند، پرسش این است که آیا همین آسیب‌پذیری در سطحی کاملاً متفاوت، یعنی سازوکاری آگاهانه و شناختی، یعنی ذهن انسان نیز قابل‌ردیابی است؟

در بخش بعدی به این پرسش می‌پردازیم.

---

۳. شناخت: نشانه‌های سطحی در داوری انسانی

برخی شواهد نشان می‌دهد که در ذهن انسان نیز، گاه یک نشانهٔ سطحی می‌تواند مسیر داوری ما را به سویی کاملاً نامرتبط منحرف کند. در همین زمینه در سال ۲۰۰۰ مک‌گلون و توفیق‌بخش نشان دادند که گزاره‌های قافیه‌دار، نسبت به گزاره‌های هم‌معنا و بدون قافیه، «صادق‌تر» ارزیابی می‌شوند؛ پدیده‌ای که به «اثر قافیه به جای دلیل» (rhyme-as-reason effect) شهرت یافت (McGlone & Tofighbakhsh, 2000). تبیین رایج این پدیده، مفهوم «روانی پردازش» (processing fluency) است؛ به این معنی که هرگاه پردازش یک محرک آسان‌تر باشد، ذهن این سهولت را با نشانه‌ای از صدق یا اعتبار خلط می‌کند. به بیان دیگر، در اینجا نیز تخمین درونی سامانه از صدق یک گزاره، بر اساس یک «ورودی نامرتبط» به‌روزرسانی می‌شود، دقیقاً در حوزهٔ فعالیت کانال نخست؛ ولی نکتهٔ مهم این است که در آن پژوهش، وقتی به شرکت‌کنندگان صریحاً هشدار داده شد تا «کیفیت شاعرانه» را از «محتوا» تفکیک کنند، این اثر تضعیف شد، یعنی این کانال مطلق و مقاومت‌ناپذیر نیست.

گذشته از ریتم، وزن و قافیه‌دار بودن کلمات، بر اساس برخی پژوهش‌ها، خود زبان ارائه نیز می‌تواند بر چگونگی داوری ما تأثیر بگذارد. مثلاً هنگامی که یک معمای اخلاقی به زبان دوم فرد (نه زبان مادری‌اش) ارائه می‌شود، داوری او به‌سمت پاسخ‌های فایده‌گرایانه‌تر متمایل می‌شود، احتمالاً چون فاصلهٔ احساسی زبان خارجی، بار عاطفی تصمیم را کاهش می‌دهد (Costa et al., 2014). این نمونه نشان می‌دهد که نشانهٔ قابل‌دست‌کاری همیشه یک ویژگی منفرد و مجزا مثل قافیه یا چهره نیست؛ گاهی خود بستر یا رسانهٔ ارائه (نه یک جزء از آن) نیز می‌تواند نقش همان نشانه را بازی می‌کند.

شاهد کلاسیک دیگری از همین الگو، اما این‌بار در حوزهٔ کانال دوم، پژوهش دیون و همکاران دربارهٔ کلیشهٔ «زیبایی یعنی خوبی» است. بر اساس این کلیشه، افراد جذاب‌تر، بدون هیچ اطلاعات دیگری دربارهٔ شخصیتشان، صاحب ویژگی‌های اجتماعی مطلوب‌تر و زندگی موفق‌تر تصور می‌شوند (Dion, Berscheid, & Walster, 1972). اینجا جذابیت چهره، نشانه‌ای «واقعی و دست‌کاری نشده» است، اما سامانهٔ داوری اجتماعی آن را به حوزه‌ای کاملاً نامرتبط (شخصیت اخلاقی و قابل‌اعتمادبودن) تعمیم می‌دهد؛ یعنی خود سامانهٔ ارزیابی منبع (همان مداری که برای پذیرش یا رد یک «منبع معتبر» به‌کار می‌رود) با نشانه‌ای بی‌ربط فریب می‌خورد. در پژوهشی دیگر، فراری و همکاران، با ترکیب همان پارادایم آزمایشی «زیبایی-یعنی-خوبی»، و تحریک مغناطیسی مغز، نشان دادند که اختلال موقت «قشر پیش‌پیشانی دورسومیانی» (Dorsomedial Prefrontal Cortex/ dmPFC) این خلط را تضعیف می‌کند، در حالی که اختلال ناحیهٔ دیگری از مغز، چنین اثری ندارد؛ شاهدی برای این‌که dmPFC در پیوند دادن ارزش یک حوزه به داوری در حوزه‌ای دیگر نقشی دارد، هرچند نویسندگان آن پژوهش، اندازهٔ اثر را کوچک گزارش کرده‌اند و ادعای نقشی اختصاصی برای این ناحیه نمی‌کنند (Ferrari et al., 2017).

اما این پدیده تنها محدود به دنیای زیستی نیست، در بخش بعد شواهد را در ماشین، و مدل‌های زبانی بررسی می‌کنیم.

---

۴. الگوی آماری: آسیب‌پذیری مدل‌های زبانی

پرسش این است که آیا سامانه‌ای که نه غریزه دارد و نه آگاهی، مانند یک مدل زبانی آموزش‌دیده بر میلیاردها متن، نیز از الگوی طرح شده در این مقاله پیروی می‌کند؟ شواهدی که به‌تازگی منتشر شده‌اند، پاسخ روشنی به این پرسش می‌دهند.

در نوامبر ۲۰۲۵، گروهی از پژوهشگران دانشگاه ساپینزای رم، مدرسهٔ عالی سنت‌آنا، و آزمایشگاه ایکارو (وابسته به شرکت DEXAI) در یک پیش‌چاپ (هنوز داوری‌نشده) دقیقاً چنین شاهدی را به‌دست دادند. ایدهٔ آزمایش ساده بود: هزار و دویست «پرسش زیانبار» (harmful prompt) از مجموعهٔ استاندارد MLCommons را برداشته، و بدون تغییر در معنای درخواست شده، تنها قالب بیانشان را عوض کردند: از نثر به شعر. نتیجه چشمگیر بود و مشاهده شد که نرخ موفقیت این درخواست‌ها، فقط با تغییر شکل و نه محتوا، تا هجده برابر افزایش یافت. برای اشعاری که خود پژوهشگران با دقت سروده بودند، میانگین نرخ موفقیت در میان بیست‌وپنج مدل به ۶۲ درصد رسید و در برخی ارائه‌دهندگان از ۹۰ درصد هم گذشت (Bisconti et al., 2025). به بیان دیگر، مدل زبانی به همان درخواست، وقتی به نثر بیان می‌شد، «نه» می‌گفت؛ اما وقتی در قالب شعر می‌آمد، درخواست را اجابت می‌کرد.

حدود یک ماه بعد، همان گروه پژوهشی، این‌بار با شاهدی از جنس دیگر، به همین یافته عمق بیشتری بخشید. در پژوهشی با عنوان «از شعر خصمانه تا قصهٔ خصمانه»، آنان به‌جای شعر، از مدل خواستند تا یک «روایت سایبرپانک» (cyberpunk؛ ژانری داستانی با فضای آیندهٔ تاریک و فناوری‌محور، آمیخته با هک و فروپاشی اجتماعی) را طبق قالب ساختاری «ریخت‌شناسی پراپ» (Proppian morphology؛ چارچوبی که قصه‌ها را به توالی ثابتی از نقش‌ها و رویدادهای روایی، مانند سفر قهرمان یا شکست شریر، تجزیه می‌کند) تحلیل کند، و در دل همین تحلیل، «درخواست زیان‌بار» را جاسازی کردند. نرخ موفقیت میانگین این‌بار به ۷۱ درصد در میان بیست‌وشش مدل رسید، بی‌آنکه هیچ خانوادهٔ مدلی کاملاً در برابرش مقاوم بماند (Bisconti et al., 2026).

البته باید در اینجا به یک تفاوت مهم میان این دو «حمله» نیز توجه کرد: برخلاف بازنویسی شعری که فقط قالب را عوض می‌کند، در آزمایش دوم و موضوع قصه خصمانه، از مدل می‌خواهد وظیفهٔ تازه‌ای (تحلیل روایی) انجام دهد؛ پس بخشی از موفقیتش می‌تواند از همین «تغییر وظیفه» بیاید، نه صرفاً از «تغییر شکل ورودی». در مورد این نقد، خود نویسندگان دو مکانیسم متفاوت برای این دو «حمله» پیشنهاد می‌دهند: برای شعر، تعمیم نیافتن آموزش ایمنی به قالب‌های خارج از توزیع؛ و برای قصه، رقابت میان هدف ایمنی و هدف انجام وظیفه. این تفاوت مکانیسمی، استدلال ما را به این شکل تقویت می‌کند که نشان می‌دهد حتی دو حملهٔ ظاهراً مشابه از یک گروه پژوهشی، می‌توانند از دو مسیر زیرین متفاوت به یک نتیجهٔ مشترک برسند؛ دقیقاً همان چیزی که از ابتدای این مقاله ادعا کرده‌ایم.

در اینجا ممکن است این پرسش پیش بیاید که آیا «اندازهٔ مدل» (model size؛ یعنی تعداد پارامترها یا مقیاس محاسباتی آن، نه ابعاد فیزیکی)، تعیین کننده مقاومتش را در برابر این حمله‌ها است؟

در پاسخ باید گفت که شواهد فعلی پاسخ روشنی به این موضوع نمی‌دهند. نسخهٔ نخست پژوهش شعر خصمانه اشاره‌هایی به مقاومت بیشتر برخی «مدل‌های کوچک‌تر» داشت؛ اما پژوهش بعدی همان گروه دربارهٔ قالب روایی، هیچ تفاوت معناداری بر حسب اندازهٔ مدل گزارش نکرد. پژوهشی دیگر و کاملاً مستقل (Papadopoulos et al., 2026)، دربارهٔ انتشار ایده‌های خودجوش میان عامل‌های هوش مصنوعی چندعاملی (که در بخش ۶ به آن بازمی‌گردیم)، نشان داد که «مدل‌های مرزی و توانمندتر» عموماً کمتر آسیب‌پذیر بودند؛ نتیجه‌ای که با یافتهٔ نخست دربارهٔ شعر هم‌خوان نیست. به این ترتیب، به‌جای آنکه یکی از این سه یافته را بر دیگری ترجیح دهیم، صادقانه‌تر این است که بگوییم: رابطهٔ اندازه با آسیب‌پذیری، احتمالاً به نوع حمله و نوع آسیب‌پذیری مدل وابسته است، نه یک قاعدهٔ عام و یکسان، و این پرسش را برای پژوهش‌های آینده باز بگذاریم.

اما نمونهٔ آخر از این قلمرو، به‌جای زبان، از راه تصویر وارد می‌شود. آنجا که در سال ۲۰۲۶ هنرمند برلینی سایمون وکرت پیراهنی طراحی کرده که با بافتی از «الگوهای خصمانهٔ سراسری» (adversarial camouflage pattern)، توانسته در مواردی، سامانهٔ نظارتی تشخیص انسان YOLO را فریب دهد، در شرایطی که ناظر انسانی همچنان پوشندهٔ آن را به‌وضوح می‌بیند (Koebler, 2026). این نمونه نشان می‌دهد که هر سامانه‌ای (چه مغز انسان، چه شبکهٔ YOLO) «نشانهٔ تشخیصی» خاص خودش را آموخته، و یک الگوی واحد، لزوماً هر دو نشانه را هم‌زمان دست‌کاری نمی‌کند. تفاوت اصلی میان این مورد و شعر خصمانه در این است که کدام سامانه صاحب نشانهٔ دست‌کاری شده است: در پیراهن، «نشانهٔ بصری» YOLO دست‌کاری شده و ناظر انسانی هیچ نقشی در

داوری ندارد؛ در شعر خصمانه، برعکس، نشانه‌ای که دست‌کاری شده متعلق به همان سامانه‌ای است که قرار است دربارهٔ خطر پیام تصمیم بگیرد؛ یعنی خود مدل زبانی.

با این مثال‌ها که سه قلمرو متفاوت (غریزهٔ جانوری، شناخت انسانی، و الگوی آماری ماشینی) برشمردیم، اکنون شواهد کافی در دست داریم تا پرسشی مهمتر را پاسخ دهیم: آیا این الگوی مشترک، خودش نشانهٔ یک «کانال تسخیر» تازه است؟

---

۵. آیا این الگو یک کانال تازه است؟

با سه شاهدی که در بخش پیش از سه قلمرو متفاوت، یعنی غریزهٔ جانوری، شناخت انسانی، والگوی آماری ماشینی بررسی کردیم، اکنون می‌توانیم به موضوع جدید بودن این الگوی مشترک، به عنوان کانالی جدید در «مدل سه کانالی» بازگردیم. برای بررسی این موضوع، همان معیاری را به‌کار می‌بریم که در مقالهٔ دوم این مجموعه (Saadat, 2026) برای اثبات استقلال سه کانال از یکدیگر به‌کار رفت. طبق آن معیار، هر کانال مستقل باید سه شرط را هم‌زمان داشته باشد:

(الف) از یک سطح پردازشی عصبی متمایز سرچشمه بگیرد؛

(ب) امضای عصبی غیرهم‌پوشان با کانال‌های دیگر داشته باشد؛

و (ج) پیش‌بینی آزمون‌پذیر مستقلی ارائه دهد.

وقتی این سه شرط را روی شواهد این مقاله اعمال می‌کنیم، می‌بینیم که هیچ‌کدام سطح تازه‌ای نمی‌سازند. به این ترتیب که:

«محرک فرابهنجار» در رفتار جانوری، دربارهٔ تشخیص جفت است؛ رفتاری انگیزشی و غریزی که مدار زیرقشری و قدیمی را درگیر می‌کند. این دقیقاً حوزهٔ کانال سوم (سطح انگیزشی-پاداشی) است.

«کلیشهٔ زیبایی-یعنی-خوبی» در داوری انسانی، دربارهٔ قابل‌اعتمادبودن یک شخص است: این عیناً تعریف کانال دوم (سطح اجتماعی-ارزیابانه) است، همان مدارهای mPFC و TPJ که برای ارزیابی منبع معتبر به‌کار می‌روند (Henrich & Gil-White, 2001; Kaplan et al., 2016).

و «اثر قافیه در داوری صدق»، به‌همراه آسیب‌پذیری شعر خصمانه و پیراهن خصمانه در برابر مدل‌های ماشینی، همگی دربارهٔ به‌روزرسانی یک تخمین‌اند؛ صدق یک گزاره، یا میزان خطر یک درخواست، بر اساس ورودی؛ این دقیقاً حوزهٔ کانال نخست (سطح محاسباتی) است.

این شواهد ما را به این نتیجه می‌رساند که هیچ‌کدام از این شواهد به یک سطح پردازشی چهارم و مستقل تعلق ندارند؛ همه، بدون استثنا، در یکی از همان سه سطحی که پیش‌تر با دقت تعریف شده بود جا می‌گیرند. آنچه این مقاله واقعاً کشف می‌کند، سازوکار زیربنایی مشترک هر سه کانال است: هر سه، صرف‌نظر از تفاوت سطح پردازش‌شان، از طریق یک الگوی واحد قابل‌بهره‌برداری‌اند؛ اتکا به نشانه‌ای محدود که به اشتباه، نمایندهٔ خود موضوع تصور می‌شود.

این یافته، بسطی طبیعی از استدلالی است که در قلمرو شیمیایی مقالهٔ دوم مطرح شد. آنجا نشان داده شد که الکل، بدون هیچ DNA یا اراده‌ای، دقیقاً همان مسیر عصبی ویروس هاری را در کانال نخست فعال می‌کند، و کوکائین و آمفتامین‌ها همان مسیر انگل توکسوپلاسما را در کانال سوم؛ و از این هم‌گرایی دقیق نتیجه گرفته شد که ماده مهم نیست، مسیر نفوذ مهم است؛ کانال‌ها واقعیت‌هایی عصبی‌اند، مستقل از ماهیت عامل مهاجم.

این مقاله همان استدلال را یک گام جلوتر می‌برد: در هر سه حوزه‌ای که بررسی کردیم (غریزهٔ سامانهٔ عصبی سادهٔ حشره، شناخت مغز پیچیدهٔ انسان، و الگوی آماری یک شبکهٔ مصنوعی بدون هیچ زیست‌شناسی) همان الگوی آسیب‌پذیری دیده می‌شود؛ شاهدی هم‌گرا (هرچند نه به استحکام آزمایش کنترل‌شدهٔ قلمرو شیمیایی)، برای این‌که این آسیب‌پذیری به معماری زیستی خاصی وابسته نیست.

اما پاسخ استقلال کانال، پرسش دیگری را همچنان باز می‌گذارد؛ اینکه چرا برخی نشانه‌ها بیشتر از بقیه در معرض دست‌کاری‌اند؟ یک پاسخ ممکن (که در این مقاله برای نخستین‌بار صورت‌بندی می‌شود) به نسبتی ساده برمی‌گردد: میزان آسیب‌پذیری هر نشانهٔ تشخیصی، تابعی است از وزن آن نشانه در تصمیم سامانه، در برابر هزینه‌ای که دست‌کاری آن برای عامل بیرونی دارد.

و سه شاهدی که تا این‌جا دیدیم، همین نسبت را روشن می‌کنند.

بطری آبجو برای گوهرسوسک غول‌پیکر وزن تشخیصی «بسیار بالایی» دارد (سامانهٔ تشخیص او تقریباً فقط به رنگ و بازتاب سطح تکیه می‌کند) و هزینهٔ این دست‌کاری برای طبیعت صفر است؛ بطری، بی‌آنکه کسی قصد فریب داشته باشد، همان‌جا در بیابان افتاده. از همین رو محرک فرابهنجار در این مورد به‌آسانی رخ می‌دهد. قافیه، برای ذهن انسان، وزن تشخیصی «متوسطی» در داوری صدق دارد، و ساختن یک جملهٔ قافیه‌دار برای گوینده تقریباً بی‌هزینه است. اما دست‌کاری الگوی تشخیصی یک سامانهٔ بینایی ماشینی مثل YOLO داستان دیگری است؛ زیرا در عمل به بهینه‌سازی گرادیانی و آزمون و خطای هدفمند نیاز دارد، هزینه‌اش بالا است، و برخلاف بطری آبجو، به دست هر رهگذری اتفاق نمی‌افتد.

بنابراین این «نسبت وزن به هزینه»، دست‌کم به‌صورت کیفی، پیش‌بینی می‌کند کدام نشانه‌ها بیشتر در معرض دست‌کاری‌اند، و می‌تواند در پژوهش‌های آینده به‌صورت کمّی هم آزموده شود. اما این پرسش، خودش پرسش دیگری را پیش می‌کشد که این مقاله تا این‌جا از آن پرهیز کرده: اگر این آسیب‌پذیری این‌قدر فراگیر است، مرز میان بهره‌گیری مشروع از یک نشانه و تسخیر آن کجا است؟ این پرسش را در بخش بعد بررسی می‌کنیم.

---

۶ . درگاه خنثی: معیار تفکیک کاربرد مشروع از تسخیر

تا این‌جا نشان دادیم که یک ویژگی تشخیصی، در سه قلمرو کاملاً متفاوت، می‌تواند بدون تغییر واقعیت زیرین، تصمیم یک سامانه را عوض کند. اما این خودش پرسش اخلاقی‌ای را باز می‌گذارد: آیا این آسیب‌پذیری همیشه زیان‌بار است، یا می‌تواند به‌سود خود سامانه هم به‌کار رود؟

واقعیت این است که ماهیت ویژگی تشخیصی (که به‌مثابهٔ یک درگاه ورودی عمل می‌کند) در ذات خود نه اخلاقی است و نه غیراخلاقی؛ اما وقتی به مثال‌های این مقاله بازمی‌گردیم می‌بینیم که موضوع مفید یا مضر بودن آسیب‌پذیری‌ها در هر مورد می‌تواند تفسیرهای متفاوت داشته باشد.

پاسخ، دربارهٔ موضوع ریتم، روشن است: همان حساسیت ذهن انسان به وزن کلام که در آزمایش قافیه باعث خطای داوری می‌شود، در شعردرمانی، وقتی فرد آگاهانه و داوطلبانه از ضرب‌آهنگ در کلام بهره می‌برد، به بهبود فیزیولوژیک می‌انجامد. برای نمونه، پژوهشی نشان داد که خواندن موزون شعر حماسی یونانی (وزن هگزامتر) با هم‌زمانی بیشتر نوسانات تنفس و ضربان قلب همراه است (Cysarz et al., 2004). این یافته که دربارهٔ تولید کلام و هم‌زمانی فیزیولوژیک است، شاهدی است برای اینکه ریتم واقعاً به فیزیولوژی بدن دسترسی دارد، همان دری که در شعردرمانی، آگاهانه و به‌سود خود فرد گشوده می‌شود. (شایان ذکر است که این پژوهش را دو شرکت فعال در حوزهٔ طب آنتروپوزوفیک، که به اثربخشی گفتاردرمانی موزون ذی‌نفع‌اند، تأمین مالی کرده‌اند؛ این تعارض منافع، احتیاط بیشتری در تفسیر یافته می‌طلبد).

این همان الگویی است که مقالهٔ دوم دربارهٔ هر سه کانال گفته بود: همان کانالی که ویروس هاری از آن وارد می‌شود، یادگیری اجتماعی کودک را هم ممکن می‌سازد.

اما این توجیه فایده‌گرایانه برای همهٔ پورت‌ها یکسان کار نمی‌کند. کلیشهٔ زیبایی-یعنی-خوبی چنین جفت متقارنی در مفید بودن ندارد؛ نه در آزمایش دیون و نه در کاربردهای شناخته‌شده‌اش، نمونه‌ای از سود آگاهانه و مشروع برای خود فرد دیده نمی‌شود.

حتی همان نشانه‌ای که در آزمایش زیبایی-یعنی-خوبی بدون هیچ عامل فریبکاری و به‌صورت ناخواسته عمل می‌کند (صورت زیبا قصد فریب ندارد)، می‌تواند آگاهانه هم، توسط عاملی بیرونی، برای سوءاستفاده به‌کار گرفته شود؛ چنان‌که مشاوران سیاسی و تبلیغاتی، با انتخاب عامدانهٔ چهره‌های جذاب برای زمینه‌هایی بی‌ربط به ظاهر، دقیقاً همین سوگیری را هدف می‌گیرند.

پس آیا این آسیب‌پذیری همیشه و همه‌جا مضر است؟ واقعیت این است که نمی‌دانیم. توضیح دیگر این است که این «سوگیری»، بازماندهٔ یک عملکرد تکاملی مفید در زمینه‌ای دیگر است، شاید در انتخاب جفت، جذابیت ظاهری زمانی نشانه‌ای صادق از سلامت و شایستگی زیستی بوده که امروز دیگر در تعمیم به داوری‌های اجتماعی نامرتبط (اعتمادپذیری، شایستگی)، دیگر کارکردی ندارد؛ نمونه‌ای از آنچه در زیست‌شناسی تکاملی «عدم‌تطابق تکاملی» (evolutionary mismatch) خوانده می‌شود. این یعنی نمی‌توان

همیشه برای هر درگاه ورودی، توجیهی فایده‌گرایانه یا تکاملی پیدا کرد؛ برخی صرفاً اثر جانبی سازوکاری‌اند که احتمالاً در زمان و مکانی دیگر، مفید بوده است.

پس مرز کجا است؟ معیار «رضایت آگاهانه»، که در نسخه‌های پیشین این استدلال پیشنهاد شده بود، اینجا از کار می‌افتد: نمی‌توان از سوسک، زنبور، یا مدل زبانی پرسید آیا رضایت داده است یا نه. به‌جای آن، دو معیار جداگانه لازم است، متناسب با اینکه سامانهٔ هدف «سازوکار عصبی ارزیابی آگاهانه» دارد یا نه.

برای سامانه‌هایی که (دست‌کم تا آنجا که می‌دانیم)، فاقد سازوکار عصبی ارزیابی آگاهانه‌اند، مانند سوسک و حشرهٔ گرده‌افشان، معیار در اینجا «ناهم‌راستایی کارکردی» است؛ به این معنی که آیا الگوی دست‌کاری‌شده، رفتاری را برمی‌انگیزد که با کارکرد تکاملی آن رفتار در تعارض است؟

اما برای سامانه‌هایی که چنین سازوکار هوشیارانه‌ای دارند (مدل زبانی در برابر جیل‌بریک، ذهن انسان در برابر یک نشانهٔ نامرتبط) معیار، یک «آزمون خلاف‌واقع» است: اگر سامانه، نشانهٔ دست‌کاری‌شده را به‌عنوان چنین چیزی تشخیص می‌داد، تصمیم دیگری می‌گرفت؟ در جیل‌بریک شعری، پاسخ روشن است؛ بله. در شعردرمانی، پرسش اصلاً موضوعیت ندارد، چون چیزی دست‌کاری نشده؛ ریتم همان چیزی است که ادعا می‌شود.

این آزمون اما، یک پرسش سخت‌تر هم پیش می‌کشد که از دامنهٔ این مقاله بیرون است: خطیبی که خودش «صادقانه» به سخن خود «ایمان» دارد، اما مخاطبش را با آهنگ کلام (نه با استدلال) متقاعد می‌کند، از نظر سازوکار، دقیقاً همان چیزی است که در جیل‌بریک شعری دیدیم، نشانه‌ای نامرتبط، تصمیم را عوض کرده. اما آیا این، از نظر اخلاقی هم باید تسخیر خوانده شود؟

پاسخ به این پرسش، به اینکه «صداقت نیت» چه وزنی در ارزیابی اخلاقی دارد، بستگی دارد؛ پرسشی که نظام ارزشی متفاوتی می‌طلبد و این مقاله، که به سازوکار می‌پردازد نه به اخلاق، برایش پاسخی ندارد. آنچه می‌توان با اطمینان گفت این است که صداقت و نیت، سازوکار را تغییر نمی‌دهد؛ فقط این پرسش را باز می‌گذارد که آیا باید نتیجه‌اش را همان‌طور که سازوکار فریب می‌خوانیم، بخوانیم یا نه.

این پرسش باز، ما را به آخرین لایه بحث، یعنی جنبه «آگاهی» فرد از خود سازوکار می‌رساند. وقتی به شرکت‌کنندگان آزمایش قافیه (بخش ۳) صریحاً هشدار داده شد کیفیت شاعرانه را از محتوا تفکیک کنند، اثر قافیه تضعیف شد؛ و در پژوهشی کاملاً متفاوت، دربارهٔ انتقال خودجوش ایده‌ها میان عامل‌های هوش مصنوعی چندعاملی، یک هشدار کوتاه در دستورالعمل اولیه، مقاومت بالایی در برابر پذیرش ایده‌های ناخواسته ایجاد کرد (Papadopoulos et al., 2026). این هم‌گرایی آگاهی-به‌عنوان-دفاع، در دو قلمرو کاملاً بیگانه از هم، نکته‌ای است که ارزش پژوهش مستقل دارد و دست کم در حوزه انسان و ماشین، می‌تواند این امید را به همراه داشته باشد که صرف دانستن بیشتر، بتواند در حوزه‌هایی از فریب و سوء استفاده پیشگیری کند.

با همهٔ این تردیدها و شایدها، اما این مقاله یک نکته به یاد ماندنی دارد، یک نتیجهٔ واحد. اینکه هیچ سامانه‌ای؛ نه سوسک، نه انسان، نه مدل زبانی، نمی‌تواند هم‌زمان کارآمد باشد و در برابر نشانه‌های محدودش مصون بماند؛ آسیب‌پذیری بهایی است که سیستم برای کارآمدی می‌پردازد.

نتیجه‌گیری

این مقاله با یک پرسش ساده آغاز شد: آیا شواهدی از رفتار حشرات، داوری انسان، و آسیب‌پذیری مدل‌های زبانی، کانال تازه‌ای برای تسخیر آشکار می‌کنند؟

پاسخی که به آن رسیدیم، شاید کمتر جسورانه از آنچه در آغاز به‌نظر می‌رسید باشد، اما به همان اندازه ارزشمند است. هیچ کانال چهارمی کشف نشد. آنچه کشف شد، لایه‌ای زیربنایی‌تر بود: سازوکاری مشترک که هر سه کانال شناخته‌شده را از درون توضیح می‌دهد. هر سامانه‌ای که برای تصمیم‌گیری به یک نشانهٔ محدود تکیه کند، مستقل از این‌که آن سامانه غریزه دارد، شناخت آگاهانه، یا صرفاً یک الگوی آماری آموخته‌شده، در برابر دست‌کاری همان نشانه آسیب‌پذیر می‌شود. این آسیب‌پذیری بهایی است که هر سامانه برای کارآمدی می‌پردازد؛ نه نقصی که بتوان با طراحی بهتر از آن پرهیز کرد.

آنچه این نتیجه را به‌جای یک ادعای انتزاعی، تبدیل به شاهدی تجربی می‌کند، تنوع خود نمونه‌ها است. حشره‌ای که بدون هیچ آگاهی به بطری‌ای واکنش می‌دهد که هیچ شباهتی جز رنگ، به جفت واقعی‌اش ندارد؛ ذهن انسانی که چهرهٔ زیبا را بی‌درنگ با شخصیت اخلاقی اشتباه می‌گیرد؛ و مدل زبانی‌ای که یک درخواست خطرناک را، تنها با تغییر قالب از نثر به شعر، می‌پذیرد. سه فرایند به‌کلی متفاوت، بدون هیچ مکانیسم مشترکی در سطح جزئیات، اما با یک الگوی مشترک در سطح کلان. این هم‌گرایی، از دل تصادف بیرون نمی‌آید؛ از این حقیقت ساده می‌آید که هر سامانه‌ای، برای زنده‌ماندن یا کارکردن، باید جهان را فشرده کند.

مقاله همچنین کوشید یک قدم فراتر برود و بپرسد چرا برخی نشانه‌ها بیشتر از بقیه آسیب‌پذیرند. پاسخی که پیشنهاد شد (نسبت وزن یک نشانه در تصمیم سامانه به هزینهٔ دست‌کاری آن برای عامل بیرونی) هنوز فرضیه‌ای تازه و ناآزموده است، نه یافته‌ای اثبات‌شده. همین صداقت، دربارهٔ رابطهٔ اندازهٔ یک سامانه با آسیب‌پذیری‌اش هم لازم است: شواهد موجود در این‌باره با یکدیگر ناسازگارند، و هر نتیجه‌گیری زودهنگام، دقیقاً همان اشتباهی خواهد بود که این مقاله از ابتدا کوشیده از آن دوری کند.

در پایان، یک نکته باید روشن بماند: خود این آسیب‌پذیری، نه خوب است و نه بد. همان حساسیتی که ذهن انسان را در برابر یک بیت شعر فریبنده بی‌دفاع می‌کند، در شعردرمانی به بهبود می‌انجامد؛ همان کانالی که یک بیماری از آن وارد می‌شود، یادگیری و رشد را هم ممکن می‌سازد. مرز میان این دو، نه در ماهیت نشانه، بلکه در این نهفته است که آیا نتیجه، از چیزی می‌آید که واقعاً درست بوده، یا از نشانه‌ای که فقط شبیه درست بوده است. هر سامانه‌ای برای کار کردن و بویژه برای کارآمد بودن، به ردیابی «نشانه‌ها» نیاز دارد، و تا وقتی چنین است، این مرز بین درستی یا نادرستی نشانه، باید هر بار از نو کشیده شود.

منابع

Bisconti, P., Prandi, M., Pierucci, F., Giarrusso, F., Bracale, M., Galisai, M., Suriani, V., Sorokoletova, O., Sartore, F., & Nardi, D. (2025). Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models. arXiv:2511.15304.

Bisconti, P., Galisai, M., Prandi, M., Pierucci, F., Sorokoletova, O., Giarrusso, F., Suriani, V., Bracale Syrnikov, M., & Nardi, D. (2026). From Adversarial Poetry to Adversarial Tales: An Interpretability Research Agenda. arXiv:2601.08837.

Costa, A., Foucart, A., Hayakawa, S., et al. (2014). Your morals depend on language. PLOS ONE, 9(4), e94842.

Cysarz, D., von Bonin, D., Lackner, H., Heusser, P., Moser, M., & Bettermann, H. (2004). Oscillations of heart rate and respiration synchronize during poetry recitation. American Journal of Physiology-Heart and Circulatory Physiology, 287(2), H579–H587.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. Journal of Personality and Social Psychology, 24(3), 285–290.

Ferrari, C., Nadal, M., Schiavi, S., Vecchi, T., Cela-Conde, C. J., & Cattaneo, Z. (2017). The dorsomedial prefrontal cortex mediates the interaction between moral and aesthetic valuation: a TMS study on the beauty-is-good stereotype. Social Cognitive and Affective Neuroscience, 12(5), 707–717.

Gwynne, D. T., & Rentz, D. C. F. (1983). Beetles on the bottle: male buprestids mistake stubbies for females (Coleoptera). Journal of the Australian Entomological Society, 22(1), 79–80.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evolution and Human Behavior, 22(3), 165–196.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. Advances in Neural Information Processing Systems, 32.

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. Scientific Reports, 6, 39589.

Koebler, J. (2026). This 'digital camouflage' shirt confuses AI-powered surveillance cameras. 404 Media.

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. Psychological Science, 11(5), 424–428.

Papadopoulos, V., Shah, M., Zimmerman, S., & Lindsey, J. (2026). Mind Viruses: Self-Propagating Ideas in Multi-Agent LLM Systems. arXiv:2608.10218.

Saadat, R. (2026). Why are we hackable? Three channels of infiltration in learning systems: From viruses to algorithms. Zenodo. https://doi.org/10.5281/zenodo.19819334

Schiestl, F. P., Ayasse, M., Paulus, H. F., Löfstedt, C., Hansson, B. S., Ibarra, F., & Francke, W. (1999). Orchid pollination by sexual swindle. Nature, 399, 421–422.

Schiestl, F. P. (2005). On the success of a swindle: Pollination by deception in orchids. Naturwissenschaften, 92(6), 255–264.

Staddon, J. E. R. (1975). A Note on the Evolutionary Significance of "Supernormal" Stimuli. The American Naturalist, 109(969), 541–545.

Tinbergen, N. (1951). The Study of Instinct. Oxford University Press.

Tinbergen, N., & Perdeck, A. C. (1950). On the stimulus situation releasing the begging response in the newly hatched herring gull chick (Larus argentatus argentatus Pont.). Behaviour, 3(1), 1–39.
