آیا ما نیز مانند هوش مصنوعی، گاهی باید مسئلهٔ فکری خود را از نو بنویسیم؟

چیزی که ماشین و انسان را از هم جدا می‌کند

رامین سعادت

چکیده

این مقاله از یک مشاهدهٔ ساده آغاز می‌شود: وقتی هوش مصنوعی، در میانهٔ گفت‌وگویی طولانی، پیشنهاد می‌دهد که چت را از نو آغاز کنیم. اما به‌جای پذیرفتن توضیح خود مدل دربارهٔ دلیل این پیشنهاد، مقاله به سنجش بیرونی عملکرد روی می‌آورد. بر پایهٔ پژوهش لابان و همکاران (۲۰۲۵)، نشان داده می‌شود که وقتی اطلاعات به‌جای یک‌جا، تکه‌تکه به مدل زبانی می‌رسد، افت عملکرد عمدتاً نه از کاهش توانایی، بلکه از ناپایداری می‌آید: مدل به آنچه زودتر می‌رسد لنگر می‌اندازد. مقاله سپس می‌پرسد آیا حق داریم این یافته را به تفکر انسان تعمیم دهیم، و با تکیه بر ادبیات مستقل لنگرانداختن و تعدیل در روان‌شناسی شناختی (از تورسکی و کانمن تا پژوهش‌های اخیر دربارهٔ مدل‌های زبانی)، نشان می‌دهد چنین مجوزی، دست‌کم در سطح الگوی رفتاری، وجود دارد. اما در انسان، این لنگر می‌تواند به چیزی گره بخورد که ماشین ندارد: «خود». از این‌جا، دو لایه از هم تفکیک می‌شوند؛ لایه‌ای سرد و مشترک میان مدل و انسان، و لایه‌ای گرم که فقط در انسان، لنگر نخستین را به تعهد هویتی تبدیل می‌کند. نتیجهٔ عملی مقاله این است که نه حس درونی پایبندی و نه پرسیدن از خود، هیچ‌کدام آزمون قابل‌اعتمادی برای تشخیص پایبندی سالم از تعهد زودهنگام نیستند؛ تنها آزمون رفتاری، یعنی بازسازی مستقل مسئله و مقایسهٔ نتیجه است که این تشخیص را ممکن می‌کند.

کلمات کلیدی: تعهد زودهنگام، لنگر انداختن و تعدیل، تعهد روایی، ساختار نوبتی اطلاعات، گفت‌وگوهای چندنوبتی با هوش مصنوعی، پافشاری باور، شناخت سرد و گرم

مقدمه و طرح پرسش

بسیاری از ما این صحنه را تجربه کرده‌ایم که در میانهٔ گفت‌وگویی طولانی با هوش مصنوعی، می‌پرسیم آیا بهتر نیست چت را تازه کنیم؟ مدل پاسخ می‌دهد که بله، و در یک خط هم توضیح می‌دهد چرا؛ می‌گوید: پیگیری چیزی که تکه‌تکه و در طول گفت‌وگو گفته شده، برایش دشوارتر شده است، و شاید بهتر باشد همهٔ آنچه تا اینجا گفته‌ایم را یک‌جا و از نو صورت‌بندی کنیم.

این توضیح ساده به‌نظر می‌رسد، اما باید مکث کرد. مدل نمی‌گوید «اشتباه کردم» یا «دیگر نمی‌دانم». چیزی دقیق‌تر می‌گوید: مشکل از محتوای گفت‌وگو نیست، از «چگونگی رسیدن اطلاعات» است؛ اینکه اطلاعات تکه‌تکه و در طول زمان به او رسیده، نه یک‌جا. این یک ادعا دربارهٔ «ساختار» است، نه دربارهٔ خود «دانش»؛ و دقیقاً همین ویژگی است که آن را قابل آزمودن می‌کند، چون می‌شود همان محتوا را یک‌بار تکه‌تکه و یک‌بار یک‌جا به مدل داد و دید نتیجه فرق می‌کند یا نه. و همین دقت، هستهٔ اصلی این مقاله را می‌سازد: اگر ترتیب رسیدن اطلاعات، مستقل از خود اطلاعات، بر نتیجه اثر بگذارد، آیا چیزی شبیه به این در تفکر ما انسان‌ها هم رخ می‌دهد؟ و آیا این پاسخ سادهٔ مدل، درسی برای نحوهٔ اندیشیدن و تلاش فکری ما دارد؟

در این مقاله، این ویژگی را «ساختار نوبتی اطلاعات» (turn-based structure, sharded vs. concatenated) می‌نامیم: اینکه یک درخواست یا یک دسته شواهد، یک‌جا و کامل به مدل می‌رسد، یا تکه‌تکه و در نوبت‌های پیاپی در طول گفت‌وگو. نکتهٔ مهم این است که این یک ویژگی دربارهٔ «نحوهٔ چینش» محتوا است، نه دربارهٔ خود محتوا: اگر همان درخواست، با همان حجم اطلاعات، یک‌بار یک‌جا و یک‌بار تکه‌تکه به مدل داده شود، پاسخ نهایی مدل می‌تواند فرق کند. این اصطلاح، تا پایان مقاله، همین معنای مشخص را خواهد داشت.

در همین‌جا البته باید این نکته را روشن کنیم که ما نمی‌توانیم این پاسخ کوتاه را بی‌چون‌وچرا بپذیریم؛ زیرا این پاسخ، در واقع گزارش مدل است از «وضعیت درونی» خودش؛ و چنین گزارشی شاهد مطمئنی برای یک نتیجه‌گیری قطعی دربارهٔ «سازوکار درونی»

او نیست؛ به احتمال زیاد بازگویی «الگویی» است که از داده‌های آموزشی آموخته شده. گزارش درونی، چه در مدل و چه در انسان، شاهد ضعیفی است برای درک و تبیین سازوکار واقعی درونی.

بنابراین، استدلال ما در این مقاله دربارهٔ تشابه و درس‌گرفتن از پاسخ هوش مصنوعی نمی‌تواند بر این «پاسخ زبانی مدل» بنا شود؛ پایهٔ استدلال ما باید «سنجش بیرونی عملکرد» مدل باشد، نه آنچه او به ما می‌گوید.

با این احتیاط، باید پرسید: آیا آنچه مدل به آن اشاره می‌کند فقط یک محدودیت فنی است، یا دریچه‌ای به پرسشی است که اندیشهٔ بشر از دیرباز با واژه‌های دیگر با آن درگیر بوده و به شکل‌های گوناگون مطرح شده، و اکنون دوباره به این شکل گشوده می‌شود: به راستی پایبندی ما به یک نظر، تا کجا نشانهٔ درستی آن است و از کجا نشانهٔ چیزی دیگر؟

این پایبندی و پافشاری، که در این مقاله «تعهد روایی» (narrative commitment) نامیده می شود، چیزی فراتر از یک واقعیت ساختاری است: یعنی احساس مسئولیت نسبت به گفته‌های پیشین خود، و پیوستگی هویتی که آن گفته‌ها را به یک «خود/من» (self) واحد در طول زمان وصل می‌کند. این «تعهد» دو چهره دارد؛ چهرهٔ سالمش همین مسئولیت و پیوستگی است، چهرهٔ دیگرش قفل‌شدن روی موضع نخست و توجیه‌گری است؛ و موضوع این مقاله فقط همین چهرهٔ دوم است.

پاسخ این مقاله در پایان به یک آزمون می‌رسد، اما آزمونی رفتاری، نه یک «حس درونی». منظور، همان «حسی» است که سبب می‌شود پس از مدتی کنکاش فکری، دیگر شواهد تازه را نمی‌سنجیم، اما این حس به‌تنهایی شاهد کافی برای ما نیست و باید آزمون را نیز انجام دهیم؛ باید مسئله را روی کاغذی تازه نوشت یا برای کسی بی‌پیشینه توضیح داد و دید آیا دوباره به همان نتیجه می‌رسیم یا نه. اگر نتیجه فرق کند، می‌توان نتیجه گرفت که احتمالاً «پایبندی ما به حقیقت»، جای خود را به «تعهدی زودهنگام به پیش‌فرض‌هایمان» داده است.

برای پی‌گرفتن این پرسش‌ها و آزمون‌ها، مقاله را در چهار بخش پیش می‌بریم:

نخست نگاهی به آنچه در آزمایش‌های مدل‌های زبانی رخ می‌دهد می‌اندازیم؛

سپس می‌پرسیم آیا اصلاً حق داریم این یافته را به انسان هم تعمیم دهیم، و شواهدی را می‌بینیم که به این پرسش پاسخ می‌دهند؛

بعد به بخشی از این تصویر می‌رسیم که فقط مخصوص انسان است؛

و در پایان می‌پرسیم، اگر این تشخیص درست باشد، چه می‌توان کرد.

اکنون بیایید نگاهی به خود آزمایش بیندازیم: وقتی یک مدل زبانی در میانهٔ گفت‌وگو مسیر را گم می‌کند، دقیقاً چه اتفاقی می‌افتد؟

---

۱. آنچه در مدل رخ می‌دهد

لابان و همکارانش از مایکروسافت ریسرچ و سیلزفورس ریسرچ (Laban et al., 2025) دقیقاً به همین پرسش پاسخ داده‌اند. پرسش آنها این بود که اگر همان دستور، با همان محتوا و همان حجم، یک‌بار یک‌جا و یک‌بار تکه‌تکه در طول گفت‌وگو به مدل داده شود، آیا پاسخ نهایی فرق می‌کند؟

برای آزمودن این پرسش، یک دستور واحد را دو گونه به مدل دادند: یک بار کامل و یک‌جا، و یک بار خردشده در چند تکه که هر نوبت فقط یکی از آنها را آشکار می‌کرد. در حالت دوم مدل باید پیش از کامل شدن مسئله تصمیم می‌گرفت چه بگوید: بپرسد، حدس بزند یا پاسخی نهایی بدهد.

آزمایش شامل بیش از دویست هزار گفت‌وگوی شبیه‌سازی‌شده، شش نوع وظیفه (برنامه‌نویسی، پرس‌وجوی پایگاه داده، فراخوانی رابط‌های نرم‌افزاری، ریاضی، توصیف داده و خلاصه‌نویسی) و پانزده مدل بود. میانگین عملکرد در حالت خردشده ۳۹ درصد کمتر از حالت کامل شد؛ افتی که در همهٔ مدل‌ها دیده شد و حتی در گفت‌وگوی دو نوبتی هم پدیدار بود (گفت‌وگویی که فقط دو نوبت دارد، یک پیام از کاربر، یک پاسخ از مدل؛ در ادبیات فنی این را turn می‌گویند؛ هر پیام یا پاسخ یک نوبت است، نه هر رفت‌وبرگشت). یافتهٔ مهم‌تر، اما هنگامی به دست آمد که این افت تجزیه شد. به این معنی که هر درخواست ده بار اجرا، و دو معیار جداگانه سنجیده شد: «توانایی»، یعنی نمرهٔ صدک نود از میان آن ده اجرا، و ناپایداری، یعنی فاصلهٔ صدک نود تا صدک ده همان اجراها. توانایی فقط ۱۶ درصد کم شد؛ یعنی مدل، در بهترین حالتش، تقریباً همان قدرتی را داشت که در حالت کامل داشت. اما ناپایداری ۱۱۲ درصد

بیشتر شد. به بیان دیگر، مشکل این بود که مدل دیگر قابل‌اتکا نبود، نه اینکه دیگر نمی‌توانست به همان سطح عملکردی که در حالت کامل داشت برسد.

اما چرا این اتفاق افتاد؟ پژوهشگران چهار الگوی رفتاری جداگانه را در خود پاسخ‌های مدل شناسایی کردند، هرکدام با شاهد خودش. نخست، مدل زود به پاسخی نهایی می‌رسد و برای جزئیات ناگفته فرض می‌سازد؛ دقت همین پاسخ‌های زودهنگام فقط ۳۰.۹ درصد بود، در برابر ۶۴.۴ درصد وقتی مدل صبر می‌کرد تا اطلاعات کامل شود.

دوم، مدل بیش از اندازه به همان پاسخ زودهنگام، حتی اگر نادرست باشد، تکیه می‌کند.

سوم، به نوبت اول و آخر بیش از نوبت‌های میانی وزن می‌دهد؛ ارجاع به نوبت‌های میانی به زیر ۲۰ درصد افت می‌کرد.

چهارم، پاسخ‌هایش هرچه گفت‌وگو پیش می‌رود طولانی‌تر می‌شوند، تا ۲۰ تا ۳۰۰ درصد بلندتر، و همین درازگویی جای سخن کاربر را می‌گیرد.

در نهایت نتیجه‌ای که پژوهشگران به‌طور خلاصه گرفتند این بود که: مدل وقتی در گفت‌وگو راه اشتباهی می‌رود، «گم می‌شود و باز نمی‌گردد».

نخستین و دومین علت را در این مقاله «تعهد زودهنگام» (premature commitment) می‌نامیم: یکی رسیدن زود به یک پاسخ است، همان چیزی که لابان «premature answer attempts» می‌نامد، دیگری تکیهٔ بیش از حد بر همان پاسخ، که پیامد اولی است و لابان آن را «over-reliance on prior attempts» می‌خواند. خود عنوان «تعهد زودهنگام» از مقالهٔ لابان نیست و از ما است.

البته این یافته مرزهایی نیز دارد؛ به این معنی که «کاربر» در این آزمایش خودش یک مدل بود، نه انسان؛ وظیفه‌ها همه تحلیلی و به زبان انگلیسی بودند؛ و در وظیفهٔ مکملی که جداگانه (بیرون از آن شش وظیفهٔ اصلی) آزمودند و می‌شد تکه‌تکه انجامش داد یعنی ترجمهٔ جمله به جمله، افتی دیده نشد. نویسندگان حدس می‌زنند افت در وظیفه‌های تولیدی، پیچیده و تجزیه‌ناپذیر رخ دهد.

به این ترتیب، نتیجه‌ای که می‌توان از این پژوهش گرفت محدود اما محکم است: در سامانه‌ای که «ساختار نوبتی اطلاعات» را می‌توان جدا از «توانایی» تغییر داد، بخش عمدهٔ افت عملکرد از «ناپایداری» می‌آید، یعنی از پراکندگی نتیجه میان اجراهای گوناگون یک درخواست واحد، نه از «کاهش توانایی»، به معنی افت سقف عملکرد مدل در بهترین حالتش.

این یافته پلی به سمت انسان هم دارد که در ادامه به آن نیاز داریم. اگر افت عمدتاً از ناپایداری بیاید نه از انحراف یک‌سویه، معنایش این است که تعهد زود هنگام، همهٔ اجراهای یک درخواست را به یک نتیجهٔ ثابت نمی‌برد؛ هر اجرا به هر تکه‌ای که زودتر رسیده «لنگر می‌اندازد»، و چون آن تکه از اجرایی به اجرای دیگر فرق می‌کند، نتیجه‌ها پراکنده می‌شوند، نه یک‌دست منحرف. به بیان دیگر، امضای رفتاری لنگر انداختن، در سطح جمعیت اجراها، «پراکندگی» است، نه سوگیری به سوی یک «جهت ثابت».

با این تصویر از مدل در دست، پرسش طبیعی بعدی این است: آیا اصلاً حق داریم یافته‌ای را که دربارهٔ یک سامانهٔ محاسباتی به دست آمده، به تفکر انسان هم تعمیم دهیم؟

---

۲. لایهٔ سرد: آیا حق داریم مقایسه کنیم؟

با این تصویری که در بخش پیش از مدل ترسیم کردیم، پرسش طبیعی بعدی این است که: آیا اصلاً حق داریم یافته‌ای که از بررسی یک سامانهٔ محاسباتی به دست آمده را به تفکر انسان نیز تعمیم دهیم؟ با وجودی که می‌دانیم ماشین و انسان، از نظر «سازوکار»، یکی نیستند؟ بویژه وقتی می‌دانیم، دانش و توان استدلال ماشین، بعد از آموزشش، ثابت می‌ماند و در طول هیچ گفت‌وگویی تغییر نمی‌کند؛ او نه حافظه‌ای از گفت‌وگوهای پیشین دارد و نه هویتی که بخواهد حفظش کند؛ آنچه هر نوبت عوض می‌شود، فقط متن ورودی پیش رویش است. انسان برعکس، حافظه و هویتی دارد که در طول زمان با خودش حمل می‌کند و گفته‌های پیشینش را به آن نسبت می‌دهد؛ تفاوتی مهم و واقعی که نمی‌توان و نباید پنهانش کرد.

اما «حق مقایسه»، در هیچ‌جای علم، منوط به یکسان‌بودن سازوکار دو سامانه نیست. مثلاً وقتی رفتارشناسان، حل مسئله را در کلاغ و در انسان مقایسه می‌کنند، مغز کلاغ و مغز انسان هیچ شباهت ساختاری ندارند؛ پژوهشگر حق مقایسه را از جای دیگری می‌گیرد:

الگوی رفتاری مستقل و مستندی که در هر دو دیده شده است. اگر چنین الگویی، در دو ادبیات کاملاً جدا از هم، یکی دربارهٔ مدل‌های زبانی و یکی دربارهٔ شناخت انسانی، پیش از این مقاله و مستقل از آن مستند شده باشد، همین مجوز مقایسه را در سطح رفتار می‌دهد، نه مجوز ادعای سازوکار مشترک. ببینیم چنین الگویی وجود دارد یا خیر.

این لایه را، با وام‌گیری از دوگانهٔ «شناخت سرد» و «شناخت گرم» در روان‌شناسی، در اینجا «لایهٔ سرد» می‌نامیم؛ با این تعریف که:

«شناخت سرد» (cold cognition): پردازشی است که فارغ از هیجان و منافع شخصی رخ می‌دهد؛ صرفاً اطلاعاتی و تحلیلی.

«شناخت گرم» (hot cognition): پردازشی است که هیجان، انگیزه، یا منافع «خود» در آن دخیل است.

بنابراین، در اینجا نیز «لایهٔ سرد» پردازشی است که برایش به یک «من» که آبرو یا هویتش در خطر باشد، نیازی نیست؛ صرفاً وزن‌دهی نامتقارن به ترتیب دریافت اطلاعات است.

در توضیح بیشتر، می‌توان به یکی از چارچوب‌های تبیینی مرتبط در روان‌شناسی شناختی، به نام «لنگر انداختن و تعدیل» (Anchoring and adjustment) اشاره کرد، که تورسکی و کانمن، آن را در سال ۱۹۷۴ نشان دادند (Tversky & Kahneman, 1974). تبیینی که بر اساس آن، «داوری» از نقطه‌ای آغاز می‌شود (لنگر) و با حرکتی رو به جواب درست‌تر تعدیل می‌یابد؛ اما این تعدیل معمولاً زودتر از موعد متوقف می‌شود، و جواب نهایی همچنان ردپای همان نقطهٔ آغاز را با خود دارد.

از نگاه این مقاله، می‌توان بخشی از آنچه در گفت‌وگو با مدل دیده شد را نیز در همین چارچوب تبیینی تورسکی و کانمن خواند: آنچه زود می‌رسد، لنگر بقیهٔ راه می‌شود. البته لابان و همکارانش این ارتباط را برقرار نکرده‌اند؛ آنچه آنها گزارش کردند صرفاً رفتار مدل بود، و این خوانش (که آن رفتار را به نظریهٔ لنگرانداختن وصل می‌کند) تفسیر ما در این مقاله است.

از سوی دیگر، دنیس و آن در سال ۲۰۰۱ این اثر را در «یادگیری رابطه‌های علّی» آزمودند (Dennis & Ahn, 2001). در این پژوهش، شرکت‌کنندگان شواهد یکسانی را با دو ترتیب متفاوت دیدند و مشاهده شد که داوری نهایی‌شان با هم فرق کرد. به این ترتیب که از شواهد نخست، فرضی ساخته می‌شود، و شواهد بعدی آن را اصلاح می‌کنند، اما کمتر از آن اندازه‌ای که محتوای خود آن شواهد اقتضا می‌کرد. در حالی که در یک به‌روزرسانی بیزی ایدئال، اگر شواهد شرطاً مستقل باشند و مدل احتمالاتی ثابت بماند، ترتیب رسیدن داده نباید اهمیت داشته باشد؛ هوگارث و آینهورن (Hogarth & Einhorn, 1992) این انحراف را در قالب «مدل تعدیل باور» صورت‌بندی کردند.

البته در این مرز باید محتاط بود: مارش و آن (Marsh & Ahn, 2006) نشان دادند «اثر اولویت» همیشه رخ نمی‌دهد. واقعیت این است که این اثر به این بستگی دارد که ذهن فرصت داشته باشد از همان ابتدا یک «فرض واحد» بسازد و رویش باقی بماند. وقتی این فرصت گرفته شود، مثلاً چون فرد باید هم‌زمان چند رابطهٔ علّی را از هم تفکیک کند، ذهن دیگر نمی‌تواند به‌سادگی به یک فرض اولیه قفل شود، و «اثر اولویت» ضعیف‌تر می‌شود؛ این ظرفیت با حافظهٔ کلامی فرد هم ارتباط دارد. جالب‌تر اینکه در همین شرایط، به‌جای «اثر اولویت»، گاهی «اثر تازگی» دیده شده: وزن بیشتر به «جدیدترین شاهد»، نه «نخستین شاهد». این نشان می‌دهد حساسیت به ترتیب پدیده‌ای پایدار است، اما جهتش ثابت نیست.

در پژوهشی دیگر، همین «خانوادهٔ اثرها» را در خود مدل‌های زبانی هم آزموده‌اند؛ آنجا که وانگ و همکاران در سال ۲۰۲۳ نشان دادند ChatGPT در طبقه‌بندی، برچسب‌هایی را که در فهرست زودتر آمده‌اند بیشتر انتخاب می‌کند، حتی وقتی محتوا ثابت است و فقط ترتیب تغییر کرده است (Wang et al., 2023).

سپس در ۲۰۲۴، گو و وثوقی (Guo & Vosoughi, 2024) «اثر جایگاه ترتیبی» را در مدل‌های بیشتری بررسی کردند و نوشتند مقایسهٔ رفتار مدل با شناخت انسان نشان می‌دهد این حساسیت شاید فراتر از یک مسئلهٔ مهندسی باشد و به «ماهیت سازوکار توجه» مربوط شود.

در مرحلهٔ بعد نیز در ۲۰۲۵، هامالاینن (Hämäläinen, 2025) آزمایش کلاسیک آش (Asch, 1946) را روی سه مدل تجاری تکرار کرد و نتیجه آمیخته بود. در یک آزمایش ChatGPT نامزدی را ترجیح داد که «صفت‌های مثبتش» اول آمده بود، اما در آزمایش دیگر، هر جا مدل‌ها ترجیحی نشان دادند، بیشتر به نامزدی گرایش داشتند که «صفت‌های منفی‌اش» اول آمده بود. این نتیجه گواه حساسیت به ترتیب است، اما گواه «اثر اولویت انسان‌گونه» نیست.

به جز اینها البته در خود پژوهش لابان هم نشانه‌ای برای ما وجود دارد، هرچند دوپهلو؛ به این معنی که در این آزمایش مدل‌ها به نوبت اول و آخر، بیش از نوبت‌های میانی وزن می‌دادند؛ یعنی هم اولویت و هم تازگی را با هم نشان می‌دادند، نه فقط اولویت. این الگو به همین دلیل شاهد خالصی برای «لنگر انداختن» نیست؛ اما دست‌کم نشان می‌دهد که «حساسیت به ترتیب»، فراتر از یک انحراف یک‌سویه است. برخی پژوهشگران (Guo & Vosoughi, 2024) گمان می‌برند بخشی از این حساسیت به ترتیب در مدل‌ها محصول شیوهٔ آموزش آنها باشد، و نه ویژگی‌ای ذاتی و همگانی در بین همهٔ آنها.

افزون بر این، باید گفت که این پژوهش‌ها تنها دربارهٔ ترتیب «اطلاعات کامل» بودند، حال آنکه در پژوهش لابان، اطلاعات ناقص و تدریجی می‌رسد؛ این دو پدیده هم‌خانواده هستند، و نه دقیقاً یکسان.

اکنون می‌توان به پرسش آغاز این بخش برگشت: آیا حق داشتیم مقایسه کنیم؟

به نظر می‌رسد با توجه به همهٔ آنچه در این بخش دیدیم، می‌توان گفت: در سطح الگوی رفتاری، بله؛ زیرا همان «لایهٔ سرد»، در مدل و در انسان، هر دو مستند است و به انگیزه نیازی ندارد.

اما این فقط نیمی از تصویر انسان است، و نیمهٔ دیگر، همان «لایهٔ گرم» است که ماشین ندارد، و بخش بعد به آن می‌پردازد.

---

۳. لایهٔ گرم: وقتی لنگر به هویت گره می‌خورد

در بخش پیش دیدیم که «لایهٔ سرد» (لنگر انداختن به آنچه زودتر می‌رسد) در مدل و در انسان هر دو مستند است، بی‌آنکه نیازی به انگیزه یا هویت داشته باشد. اما در انسان، چیز دیگری هم هست که این تصویر را کامل می‌کند: چیزی که در همان دوگانهٔ شناخت سرد و شناخت گرم، طرف «گرم» را می‌سازد؛ یعنی پردازشی که دیگر بی‌طرفانه و صرفاً اطلاعاتی نیست، بلکه هیجان، منافع، و از همه مهم‌تر، «خود» (self) در آن دخیل است. وقتی لنگر نخستین به این «خود» گره می‌خورد، چیزی فراتر از یک خطای شناختی ساده پدید می‌آید: «تعهدی» که فرد را به دفاع از آن وامی‌دارد، حتی وقتی شواهد خلافش را نشان می‌دهند.

در سال ۱۹۷۶، استاو این موضوع را در قالب یک آزمایش ساختارمند نشان داد. دویست و چهل دانشجوی مدیریت در نقش یک مدیر شرکت قرار گرفتند که باید بودجهٔ تحقیق و توسعه را میان دو بخش شرکت تقسیم می‌کرد. نیمی از شرکت‌کنندگان خودشان این تصمیم نخست را گرفته بودند؛ نیم دیگر فقط خبردار شدند که تصمیم از پیش گرفته شده. سپس به همه گفته شد بخشی که این بودجه به آن اختصاص یافته بود، عملکرد ضعیفی داشته، و از آنها خواسته شد دوباره بودجه تخصیص دهند. کسانی که خودشان تصمیم نخست را گرفته بودند، این‌بار هم پول بیشتری به همان بخش ضعیف‌عملکرد دادند، به‌مراتب بیشتر از کسانی که تصمیم را خودشان نگرفته بودند (Staw, 1976).

استاو این را با نیاز افراد به توجیه تصمیم خودشان توضیح داد: وقتی مسئولیت یک انتخاب با «من» گره می‌خورد، شکست آن انتخاب دیگر فقط یک خطای فنی نیست؛ شکست خود «من» است، و راه دفاع از خود، سرمایه‌گذاری بیشتر در همان مسیر می‌شود.

راس، لپر و هابارد این مکانیسم را از زاویه‌ای دیگر آزمودند (Ross, Lepper & Hubbard, 1975). در آزمایش آنها، شرکت‌کنندگان در تشخیص یادداشت‌های خودکشی واقعی از جعلی آزموده شدند، اما بازخوردی که دریافت کردند تصادفی بود، نه واقعی؛ به برخی گفتند عملکردشان عالی بوده، به برخی گفتند ضعیف. سپس به همه گفته شد این بازخورد کاملاً ساختگی بوده و هیچ ربطی به عملکرد واقعی‌شان نداشته. با این حال، باور شرکت‌کنندگان دربارهٔ توانایی خودشان همچنان تحت تأثیر همان بازخورد جعلی باقی ماند.

خود پژوهشگران این را به طراحی خود آزمایش نسبت دادند، و پدیده‌ای که ادبیات بعدی از همین پژوهش «پافشاری باور» (belief perseverance) نامید؛ آنجا که پیش از آنکه خبر جعلی‌بودن بازخورد را بشنوند، از شرکت‌کنندگان خواسته بودند دلیل عملکرد (ساختگی) خودشان را بنویسند («چون در ریاضی قوی‌ام» یا «چون دقت بالایی دارم»). کسانی که این دلیل‌نویسی را انجام داده بودند، باورشان پس از ابطال بازخورد، بسیار مقاوم‌تر ماند تا کسانی که این مرحله را نداشتند. البته باید به یاد داشت که این شاهد تجربی، برای همان ادعا است، نه بیشتر: چیزی که باور را زنده نگه می‌داشت، نه خود بازخورد، بلکه توضیحی بود که فرد برایش ساخته بود.

کروگلانسکی (Kruglanski, 1989; Kruglanski & Webster, 1996) این الگو را در قالب نظریه‌ای عمومی‌تر صورت‌بندی کرد: «نیاز به قطعیت». به گفتهٔ او، ذهن در مواجهه با ابهام دو کار می‌کند؛ نخست «می‌گیرد» (seizing)، یعنی به نخستین تبیین پذیرفتنی می‌چسبد تا از رنج بی‌تصمیمی خلاص شود؛ سپس «قفل می‌شود» (freezing)، یعنی در برابر بازنگری آن تبیین مقاوم می‌شود، چون بازنگری یعنی بازگشت به همان بی‌قطعیت ناخوشایند. هرچه نیاز فرد به قطعیت بیشتر باشد (مثلاً زیر فشار زمان یا خستگی)، این گرفتن و قفل‌شدن نیز شدیدتر است.

همین پافشاری را، حتی بدون هیچ موضوع عاطفی و اجتماعی، لوچینز در ۱۹۴۲، در یکی از ساده‌ترین و معروف‌ترین آزمایش‌های روان‌شناسی شناختی، نشان داده بود (Luchins, 1942). در این آزمایش شرکت‌کنندگان سه ظرف آب با گنجایش‌های ۲۱، ۱۲۷ و ۳ واحد در اختیار داشتند و باید ۱۰۰ واحد آب اندازه می‌گرفتند؛ راه‌حل این بود: ظرف ۱۲۷ را پر کن، یک‌بار با آن ظرف ۲۱ را پر و خالی کن، دوبار با آن ظرف ۳ را پر و خالی کن، آنچه می‌ماند ۱۰۰ واحد است. پس از چند مسئلهٔ مشابه که با همین فرمول جواب می‌داد، به شرکت‌کنندگان مسئله‌ای دادند که با این فرمول حل می‌شد، اما یک راه بسیار ساده‌تر هم داشت: فقط کافی بود دو ظرف را جمع بزنند. گروهی که از قبل با فرمول پیچیده عادت کرده بودند، همان راه طولانی را دوباره به کار بردند و راه ساده را اصلاً ندیدند؛ گروه کنترل، که آن تمرین‌های اولیه را نداشت، بی‌درنگ راه ساده را پیدا کرد. ذهن، پس از آنکه یک روش جواب داد، همان «الگو» را ادامه می‌دهد، حتی وقتی راه ساده‌تری جلو چشمش است؛ اثری که به «اینشتلونگ» (Einstellung effect) شناخته می‌شود.

همچنانکه می‌بینیم، این چهار پژوهش دربارهٔ ماشین نیستند؛ موضوعشان همان انگیزه و هویتی است که مدل ندارد. با این حال یک شباهت در سطح خروجی هست:

در پژوهش لابان، مدل بیش از حد به پاسخ‌های نادرست پیشین خود تکیه می‌کرد؛

در پژوهش راس و همکاران، انسان به توضیحی وفادار می‌ماند که خودش ساخته بود؛

در پژوهش استاو، انسان به تصمیمی وفادار می‌ماند که خودش گرفته بود.

علت این پایداری اما در دو سامانه یکی نیست؛ به این معنی که به نظر می‌رسد در مدل، «وزن‌دهی نامتقارن» به آنچه در متن پیشین آمده؛ اما در انسان، «نیاز به توجیه و دفاع از خود»، سبب این پایداری شده است؛ که البته این را نیز باید در حد شباهت خروجی نگه داشت، نه یکی‌انگاشتن علت.

به این ترتیب و از مجموع آنچه در این دو بخش دیدیم، می‌توان گفت که ما با دو لایهٔ مجزا سروکار داریم.

«لایهٔ سرد»، یعنی لنگر انداختن به آنچه زودتر رسیده است، در مدل و انسان هر دو مستند است و به انگیزه نیازی ندارد؛

و لایهٔ گرم، یعنی توجیه‌گری و حفظ هویت، فقط در انسان است و لنگر نخست را به تعهد تبدیل می‌کند.

نکته مشترک اما این است که در هر دو حالت، پایبندی به مسیر نخستین می‌تواند «ظاهر ثبات و درستی» بگیرد، بی‌آنکه درست باشد. اگر این تشخیص درست باشد، پرسش بعدی عملی است: وقتی لنگر نخستین به تعهد تبدیل شده، چه می‌توان کرد؟

در بخش بعد خواهیم دید که پاسخ لابان برای مدل، یک سرنخ به ما می‌دهد؛ اما آیا همان سرنخ برای انسان هم کار می‌کند؟

---

۴. چه می‌توان کرد؟

در پاسخ به پرسشی که در پایان بخش پیش مطرح کردیم، لابان و همکارانش راهی عملی در مورد مدل‌ها پیشنهاد می‌کنند؛ به این شکل که: «هر آنچه در گفت‌وگو پراکنده گفته شده، در یک درخواست کامل جمع شود و گفت‌وگو از نو با همان آغاز شود».

پشتوانه این توصیه نیز، همان آزمایش ایشان است؛ آنجا که مشاهده کردند درخواست تجمیع‌شده به‌طور میانگین حدود ۹۵ درصد عملکرد حالت کامل را برگرداند.

افزون بر این، ایشان دو راهکار میانی را نیز آزمودند که بی‌آنکه گفت‌وگو از نو شروع شود، در همان گفت‌وگو جبران کننده هستند: «بازگفتن کل پیشینه در نوبت آخر، و تکرار تدریجی آن در هر نوبت».

البته این راهکار دوم، در دو مدلی که آزموده شد، فقط ۱۵ تا ۲۰ درصد از افت را جبران کرد و اولی نتیجهٔ بهتری داد، هرچند در یک گفت‌وگوی واقعی، از پیش معلوم نیست کدام نوبت آخرین است. همچنانکه خودشان نیز این راهکارها را «راهکار تسکینی» (mitigation) می‌نامند، نه راه‌حلی اصولی.

با این وجود در تعیین معادل انسانی این کار باید دقیق بود.

درمان لابان «تجمیع» است، نه «حذف»؛ آنجا که هیچ شاهدی دور ریخته نمی‌شود، فقط ترتیب تاریخی رسیدنش حذف می‌شود. معادل درست در انسان هم همین است: همهٔ شواهد، استدلال‌ها و منابعی را که تا این‌جا جمع کرده‌ایم، بی‌توجه به ترتیبی که به دست‌مان رسیدند، یک‌جا و از نو صورت‌بندی کنیم و از روی همان صورت‌بندی تازه استدلال کنیم؛ نه اینکه به آنها نگاه نکنیم.

برای نمونه، توضیح‌دادن مسئله به کسی که از «رفت و آمد فکری» ما هیچ نمی‌داند، این کار را به‌طور طبیعی انجام می‌دهد، چون ما را مجبور می‌کند هر چه می‌دانیم را یک‌جا و مرتب بازگو کنیم؛ برای همین به درمان لابان نزدیک‌تر است.

نوشتن روی کاغذی تازه هم می‌تواند همین نقش را بازی کند، به‌شرط آن‌که هدفش بازسازی کامل شواهد باشد، نه فراموشی آنها. این کار حافظه را پاک نمی‌کند (که در انسان ممکن نیست)، اما میان خود مسئله و تاریخچهٔ «تعهدهایی» که به آن داده‌ایم فاصله می‌اندازد.

این تکنیک با برخی روش‌های شناخته‌شده در علم «تصمیم‌گیری» خویشاوندی دارد، هرچند با هیچ‌کدام یکسان نیست. برای نمونه، با «نگاه بیرونی» (outside view) که کانمن و لوالو در ۱۹۹۳ پیشنهاد کردند (Kahneman & Lovallo, 1993) و داوری را با مقایسه به موارد مشابه گذشته می‌سنجد، نه با بازسازی خود شواهد.

همچنین با «پیش‌مرگ‌اندیشی» (premortem) که کلاین پیشنهاد کرد و با تصور یک شکست فرضی، نقاط کور را پیش از وقوع آشکار می‌کند (Klein, 2007).

وجه مشترک هر سه، ایجاد فاصله‌ای میان داوری جاری و چارچوبی است که تصمیم اولیه در آن شکل گرفته است.

در واقع آنچه اینجا تازه است، این است که در مدل زبانی می‌توان «ساختار نوبتی اطلاعات» را مستقیماً دست‌کاری کرد و اثر آن را بر عملکرد سنجید؛ در انسان نیز اثر ترتیب و نحوهٔ ارائهٔ اطلاعات را می‌توان آزمایش کرد، چنان‌که پژوهش‌های بخش‌های پیش نشان می‌دهند. اما حذف کامل تاریخچهٔ شناختی و تعهدهای پیشین یک فرد، به‌نحوی که هیچ اثری باقی نماند، مسئله‌ای متفاوت است و در انسان ممکن نبوده.

افزون بر اینها، نباید فراموش کرد که آزمایش لابان دربارهٔ مدل بود؛ اینکه معادل انسانی آن هم همین اثر را دارد یا نه، پیشنهاد و فرضیه‌ای آزمودنی است، نه یافته‌ای که بتوان ادعایش کرد.

پس چگونه بفهمیم کی نوبت این کار است؟

نخستین چیزی که در این زمینه به ذهن می‌رسد همان حسی است که در آغاز مقاله گفتیم: اینکه زمانی پیش می‌آید که دیگر شواهد تازه را نمی‌سنجیم و فقط می‌کوشیم آنها را با آنچه پیش‌تر گفته‌ایم سازگار کنیم. اما این حس به‌تنهایی کافی نیست، نه فقط به آن دلیل کلی که در آغاز گفتیم، بلکه به یک دلیل خاص‌تر: پرسیدن از خودمان «اگر امروز از صفر شروع می‌کردم، باز به همین نتیجه می‌رسیدم؟» هم یک داوری درونی است که همان «ذهن متعهد» اجرایش می‌کند و پاسخش را هم خودش صادر می‌کند. پژوهش راس، لپر و هابارد دقیقاً همین را دربارهٔ انسان نشان داد؛ آنجا که دیدیم وقتی از افراد خواستند تصور کنند پایهٔ اولیهٔ باورشان ساختگی بوده، باور همچنان برجا ماند. پس این پرسش را نباید به مثابه یک آزمون گرفت، فقط محرکی است برای اقدام. آزمون باید «رفتاری و بیرونی» باشد: مسئله را روی کاغذی تازه بنویسیم یا برای کسی که از سابقهٔ فکری ما هیچ نمی‌داند توضیحش دهیم، و ببینیم صورت‌بندی و نتیجهٔ تازه با آنچه پیش‌تر گفته بودیم یکی درمی‌آید یا نه. این کار، برخلاف پرسیدن از خود، خروجی تولید می‌کند که می‌شود آن را با متن پیشین مقایسه کرد.

اما باید این را هم گفت: همهٔ پایبندی‌ها از جنس تعهد زودهنگام نیستند. گاهی موضعی که پس از این مقایسه هم برجا می‌ماند، «پایبندی سالمی» است به حقیقتی که از پس تفکر مستقل و تعمق طولانی به دست آمده؛ این پایبندی نیازی به تازه‌کردن ندارد. آنچه باید گاه به گاه تازه شود، نه خود این پایبندی سالم، بلکه عادت به «آزمودن» آن است. به بیان دیگر، حتی باوری که اکنون

درست می‌نماید، ارزش آن را دارد که گاهی دوباره، از صفر، روی کاغذی تازه، محک بخورد؛ همان‌طور که گاهی باید چتی تازه را با هوش مصنوعی آغاز کرد.

نتیجه‌گیری

این مقاله از یک صحنهٔ آشنا آغاز شد: مدلی که در میانهٔ گفت‌وگویی طولانی، پیشنهاد می‌دهد چت را تازه کنیم، و پرسش ما این بود که این رفتار چه درسی دربارهٔ واقعیت عملکرد ذهن خودمان به ما می‌دهد.

سنجش مستقیم رفتار ماشین چیزی را آشکار کرد که ربطی به ماشین‌بودن نداشت: ترتیب رسیدن اطلاعات، به‌تنهایی و جدا از محتوای آن اطلاعات، نتیجهٔ نهایی را تغییر می‌دهد. این یافته، وقتی در کنار ادبیاتی مستقل دربارهٔ شناخت انسان قرار گرفت، یک چیز را روشن کرد: این آسیب‌پذیری منحصر به ماشین نیست؛ ما هم به همان شکل، شاید با همان سازکار ناشناخته، به آنچه زودتر می‌رسد وزن نامتقارن می‌دهیم.

اما نکتهٔ مهم‌تر این بود که برای ماشین، اصلاح این آسیب ارزان است: هیچ‌چیزی در ماشین از این اصلاح آسیب نمی‌بیند، چون ماشین چیزی ندارد که بخواهد از آن دفاع کند. برای انسان اما این‌طور نیست. وقتی همان لنگر اولیه به «من» گره می‌خورد، اصلاح دیگر فقط تغییر یک نظر نیست؛ تهدیدی است برای همان کسی که آن نظر را ساخته. برای همین، دقیقاً در جایی که بیشترین نیاز را به بازبینی داریم، بیشترین مقاومت را هم در برابرش نشان می‌دهیم. این، و نه شباهت سطحی‌تر در لنگرانداختن، آن چیزی است که این مقاله واقعاً دربارهٔ انسان، در تقابل با ماشین، نشان داد.

از همین تفاوت، یک نتیجهٔ عملی و نه چندان خوشایند بیرون می‌آید: اگر نمی‌توانیم به حس درونی خودمان دربارهٔ درستی یک باور اعتماد کنیم، نه به این دلیل ساده که ممکن است اشتباه کنیم، بلکه به این دلیل دقیق‌تر که خود آن حس، وقتی باور به هویت گره خورده، دیگر گزارش بی‌طرف یک داوری نیست، بلکه محصول همان چیزی است که قرار است بسنجدش، پس تنها راه، دور زدن این حس است: نه با پرسیدن از خود، که با تولید چیزی بیرونی و قابل مقایسه. آنچه از ماشین به عاریت می‌گیریم، حرفش نیست؛ روش آزمودنش است.

بله، ما هم گاهی باید مسئلهٔ فکری‌مان را از نو بنویسیم، نه چون باورهای ما، مثل گفت‌وگوهای طولانی، به‌خودی‌خود فاسد می‌شوند، و نه چون باید از ماشین الگو بگیریم که راحت از نو شروع می‌کند. دلیلش برعکس است: چون شروع دوباره برای ما، برخلاف ماشین، کاری سخت و پرهزینه است، و همین سختی تنها راه تشخیص پایبندی سالم از تعهد زودهنگام است.

ماشین به آسانی می‌تواند از نو بنویسد، چون چیزی ندارد که از نوشتن دوباره آسیب ببیند، اما دوباره از نو نوشتن برای ما دشوار است، چون هرچه نوشته‌ایم، بخشی از خودمان شده. همین دشواری بازنویسی، نشان ضرورت آن است، نه بهانهٔ اجتناب: وفاداری راستین، نه در نگه‌داشتن آنچه قبلاً نوشته‌ایم، که در آماده‌بودن همیشگی برای نوشتن دوباره است.

منابع

Asch, S. E. (1946). Forming impressions of personality. Journal of Abnormal and Social Psychology, 41(3), 258–290.

Dennis, M. J., & Ahn, W.-K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. Memory & Cognition, 29, 152–164.

Guo, X., & Vosoughi, S. (2024). Serial position effects of large language models. arXiv:2406.15981.

Hämäläinen, M. (2025). On psychology of AI — Does primacy effect affect ChatGPT and other LLMs? arXiv:2504.20444.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. Cognitive Psychology, 24, 1–55.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. Management Science, 39(1), 17–31.

Klein, G. (2007). Performing a project premortem. Harvard Business Review, 85(9), 18–19.

Kruglanski, A. W. (1989). Lay epistemics and human knowledge: Cognitive and motivational bases. Plenum Press.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: “Seizing” and “freezing”. Psychological Review, 103(2), 263–283.

Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). LLMs get lost in multi-turn conversation. arXiv:2505.06120. Published as a conference paper at ICLR 2026.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. Psychological Monographs, 54(6).

Marsh, J. K., & Ahn, W.-K. (2006). Order effects in contingency learning: The role of task complexity. Memory & Cognition, 34(3), 568–576.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. Journal of Personality and Social Psychology, 32(5), 880–892.

Staw, B. M. (1976). Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. Organizational Behavior and Human Performance, 16(1), 27–44.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131.

Wang, Y., Cai, Y., Chen, M., Liang, Y., & Hooi, B. (2023). Primacy effect of ChatGPT. Proceedings of EMNLP 2023, 108–115.
