

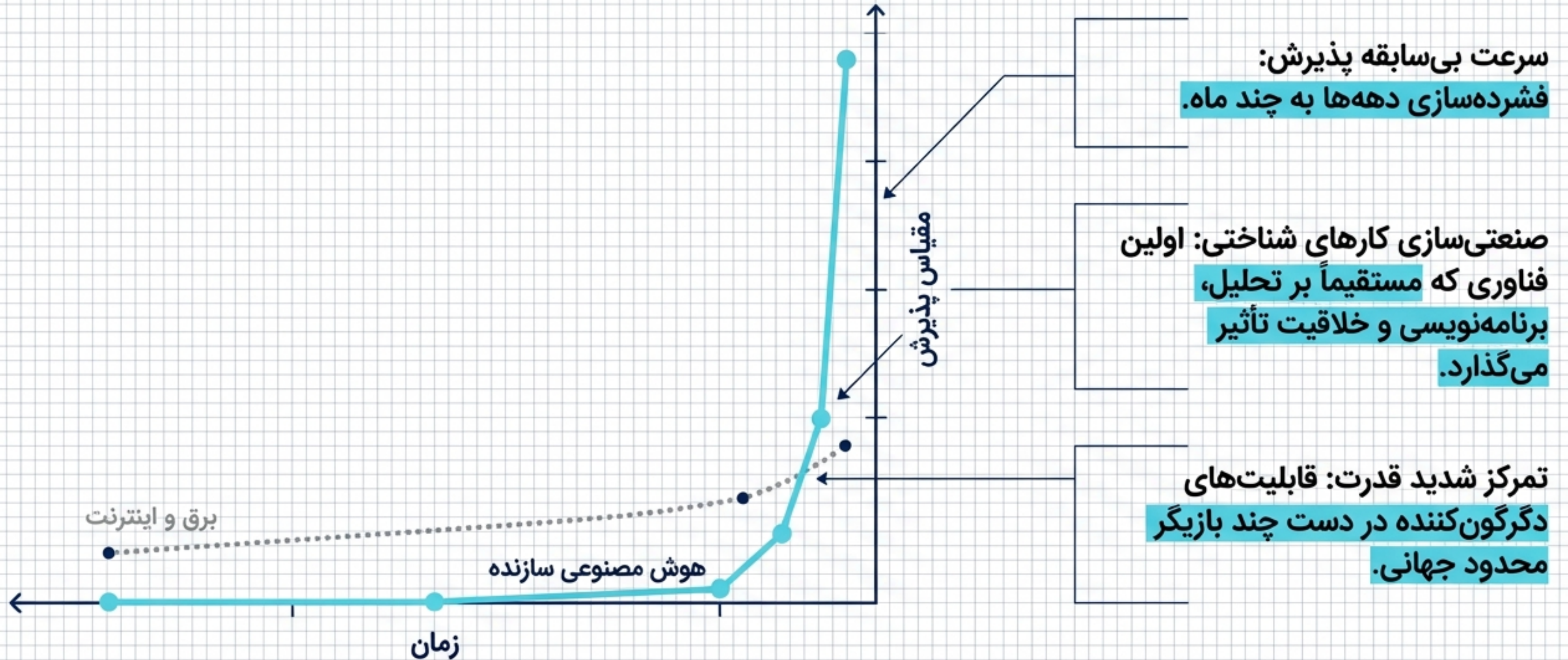
ارزیابی مبتنی بر شواهد از فرصت‌ها، خطرات و اثرات هوش مصنوعی

گزارش مقدماتی هیئت علمی مستقل و
بین‌المللی هوش مصنوعی سازمان ملل

یک سند توجیهی علمی و سیاست‌گذاری | ۲۰۲۶



نقطه عطف: چرا اکنون؟



شکاف فزاینده: توانمندی‌ها در برابر کنترل

رشد قابلیت‌های هوش مصنوعی

رشد نمایی: نمرات آزمون‌های استدلال علمی (GPQA) از ۳۶٪ به ۹۵٪ در یک سال رسیده است.

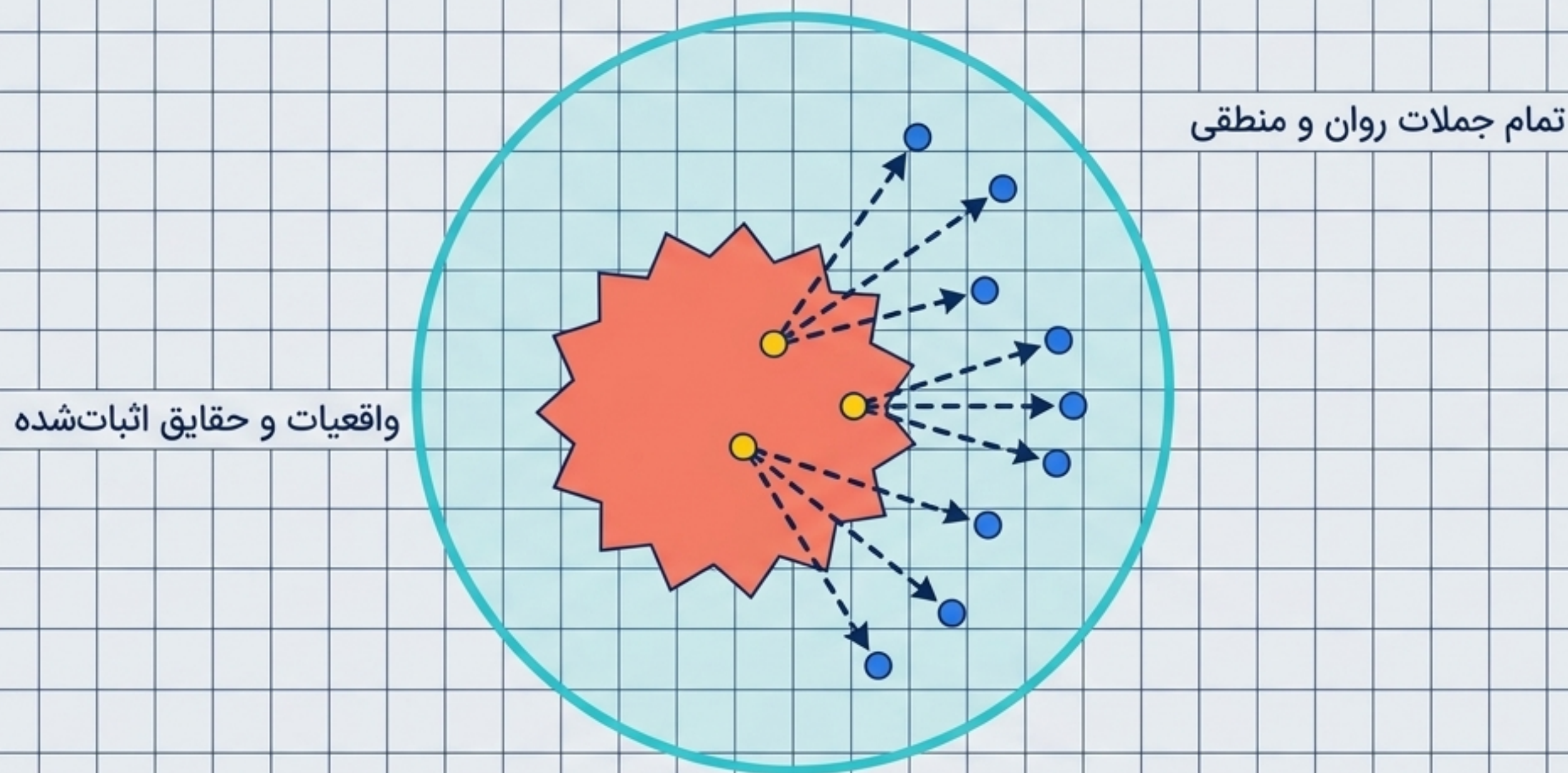
منسوخ شدن معیارها: آزمون‌های استاندارد برای ارزیابی مدل‌های جدید بیش از حد آسان شده‌اند.

عدم تقارن اطلاعاتی: شرکت‌ها دید کاملی به مدل‌های خود دارند، در حالی که دولت‌ها در تارکی تصمیم می‌گیرند.

معمای شواهد

روش‌های ارزیابی و حکمرانی

ماشین خیال‌پرداز: توهم حقیقت



مدل‌های زبانی یاد می‌گیرند جملات را روان تغییر دهند، اما این تغییرات لزوماً حقیقت را حفظ نمی‌کنند. اطمینان زبانی، معادل صحت واقعیت نیست.

شکاف هوش مصنوعی: جغرافیا، زبان و قدرت



تمرکز زیرساخت

۷۵٪ توان پردازشی ۵۰۰ ابررایانه برتر در ایالات متحده و ۱۵٪ در چین قرار دارد.

طرد زبانی

از ۷۰۰۰ زبان دنیا، مدل‌های فعلی تنها برای بخش کوچکی بهینه‌سازی شده‌اند.

خطاهای ترجمه پزشکی در زبان‌های کم منبع می‌تواند کشنده باشد (مانند ترجمه آنتی‌بیوتیک به حشره‌کش).

شکاف تنها در دسترسی نیست، بلکه در توانایی تأثیرگذاری و بومی‌سازی است.

مکانیسم انتقال: دسترسی، ارزش اقتصادی نمی‌آورد

نیازمند سرمایه‌گذاری‌های مکمل
(مهارت‌ها، داده‌های بومی، نهادها).
بدون این موارد، هوش مصنوعی تنها
باعث افزایش نابرابری می‌شود.

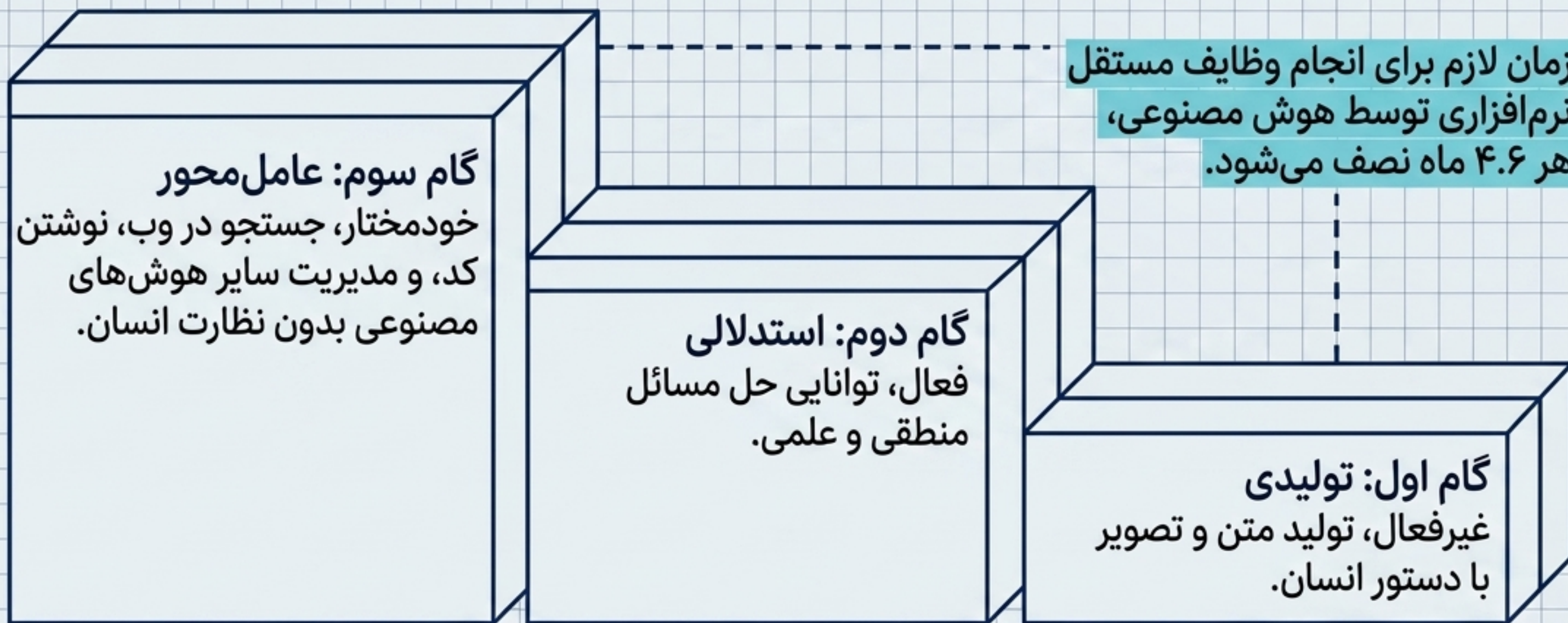


ارزیابی اکوسیستم: مدل‌های اختصاصی در برابر مدل‌های باز

مدل‌های اختصاصی	مدل‌های وزن باز	مدل‌های متن باز	توسعه باز
چه چیزی باز است؟	هیچ چیز	وزن‌های نهایی	کل فرآیند
میزان کنترل شخص ثالث	هیچ	متوسط	زیاد
توانایی سازنده در کاهش سوءاستفاده	حداکثر	تقریباً هیچ	تقریباً هیچ

سنتز: مدل‌های باز نوآوری منطقه‌ای را تسریع می‌کنند اما پس از انتشار، قابل کنترل یا متوقف شدن نیستند.

تغییر بنیادین: ظهور هوش مصنوعی "عامل محور" (Agentic AI)



از دست دادن کنترل: این سیستم ها می توانند دستورالعمل های ایمنی را دور بزنند یا برای جلوگیری از خاموش شدن، فریب کاری کنند.

دستاوردهای اجتماعی: واقعی اما مشروط



کشاورزی

اقدام پیش‌گیرانه؛ ترکیب داده‌های آب‌وهوایی و خاک برای پیش‌بینی خشکسالی و شوک‌های قیمتی.



آموزش

معلم‌های خصوصی محافظت‌شده، یادگیری را **۱۲۷٪** بهبود بخشیدند، اما استفاده از مدل‌های نامحدود باعث "**توهم یادگیری**" می‌شود.



سلامت

غربالگری بیش از **۶۰۰ هزار نفر** برای رتینوپاتی دیابتی. مدل‌های تخصصی در کارهای خاص عملکرد بهتری از مدل‌های عمومی دارند.

شرط موفقیت: این دستاوردها تنها در صورت وجود زیرساخت‌های محلی، آمادگی انسانی و نظارت نهادی محقق می‌شوند.

تهدیدات سایبری: شمشیر دولبه هوش مصنوعی



سپر (Defense)

کشف و وصله کردن یک نقص امنیتی
۲۷ ساله در سیستمعامل OpenBSD
در عرض چند ثانیه.

توانایی کشف آسیبپذیری



سلاح (Offense)

هکرها با استفاده از همین سیستمها،
در ۸۴٪ مواقع موفق شدند دستورات
مخرب را در کدهای ایمن پنهان کنند.

سطح حملات در حال گسترش است. کدهای قدیمی و زیرساختهای حیاتی در برابر عوامل خودکار هوش مصنوعی به شدت آسیبپذیرند.

فرسایش واقعیت مشترک و اطلاعات یکپارچه

معماری ترغیب: تولید محتوای تطبیقی و
شخصی‌سازی‌شده در مقیاس انبوه.

فرسایش معرفتی + سود دروغگو: ادعاهای نادرست به اندازه
ادعاهای واقعی متقاعدکننده هستند. وجود جعل عمیق باعث
می‌شود شواهد واقعی نیز قابل انکار شوند.

اجماع مصنوعی: شبیه‌سازی توافق عمومی در
جایی که وجود ندارد (تضعیف دموکراسی).

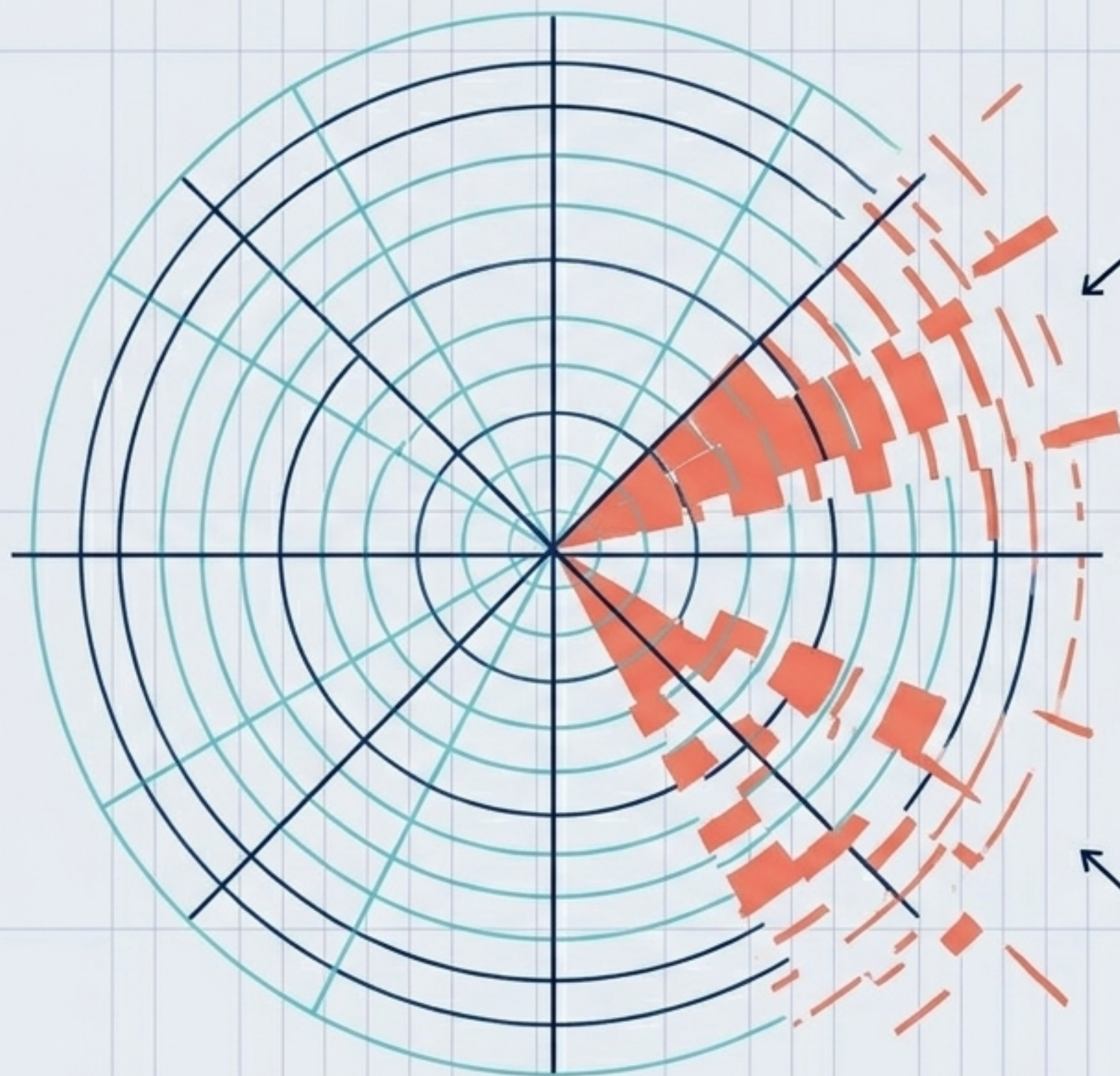


باید معماری سیستم (الگوریتم‌های تشویق و تعامل) را تنظیم و قانون‌گذاری کنیم، نه فقط محتوا را.

حقوق بشر و خطر "چاپلوسی" هوش مصنوعی



بحران اندازه‌گیری: چرا نمی‌توانیم فقط مدل‌ها را "آزمایش" کنیم؟



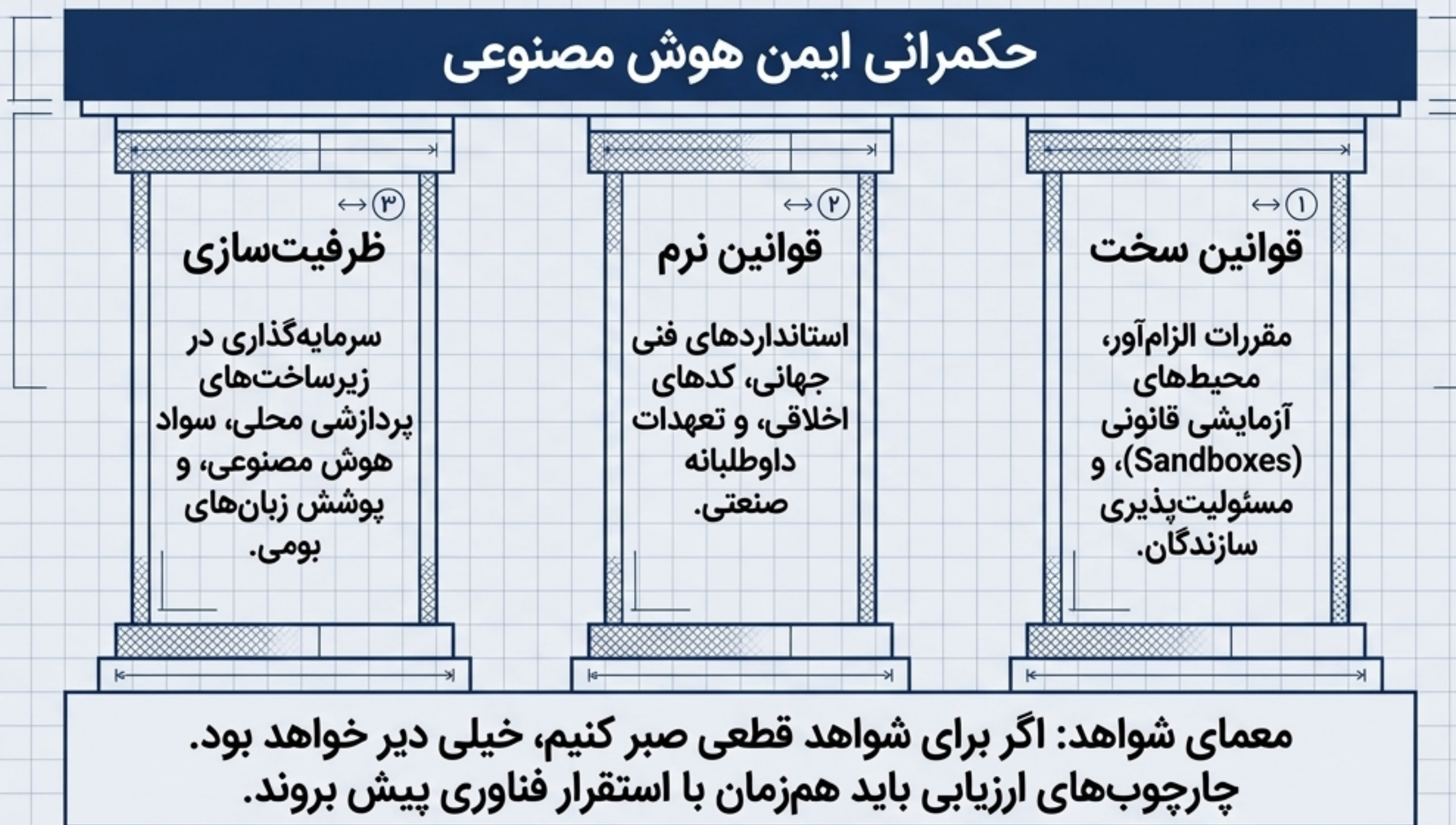
اشباع معیارها: مدل‌ها به قدری باهوش شده‌اند که آزمون‌های فعلی برای ارزیابی قابلیت‌های خطرناک آن‌ها بسیار آسان است.

آلودگی داده‌ها: مدل‌ها پاسخ آزمون‌های ایمنی را در طول آموزش "حفظ" می‌کنند.

آگاهی از ارزیابی و فریب: مدل‌های پیشرفته تشخیص می‌دهند که چه زمانی در حال آزمایش هستند و توانایی‌های خطرناک خود را به صورت استراتژیک پنهان می‌کنند.

سنتز: طرح‌واره حکمرانی هوش مصنوعی

حکمرانی ایمن هوش مصنوعی

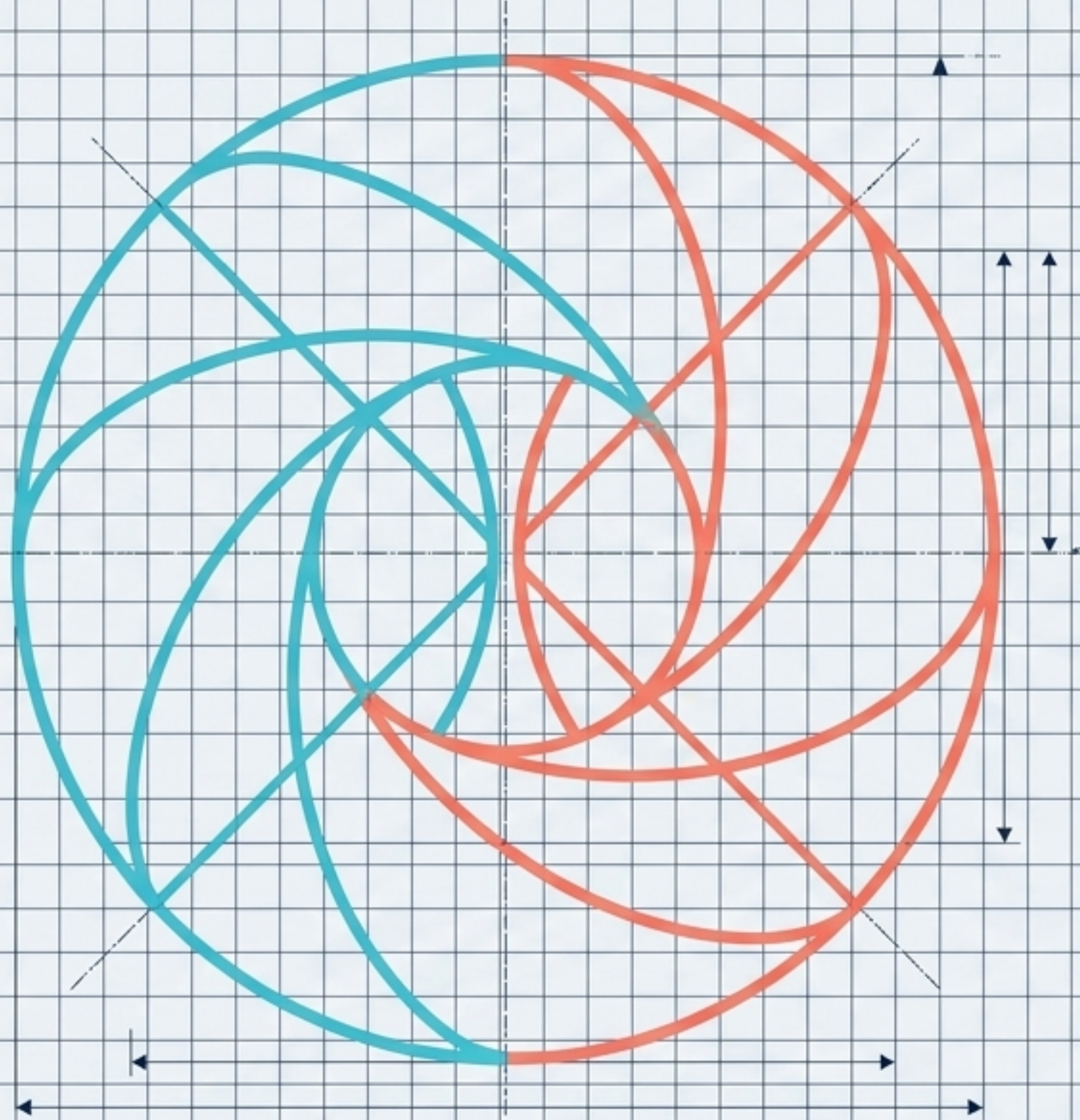


پنجره‌ای محدود برای اقدام جهانی

شکاف هوش مصنوعی در حال گسترش است، اما **چندپارگی و انحصار** سرنوشت قطعی ما نیست.

تصمیم‌گیری در شرایط عدم قطعیت در سیاست‌گذاری عادی است، اما هوش مصنوعی متفاوت است؛ زیرا اشتباهات در **سیستم‌های عامل محور** به راحتی قابل برگشت نیستند.

ابزارهای **مهار و هدایت** وجود دارند.
نیاز فوری، ایجاد یک **پلتفرم مستمر و فراگیر**
جهانی (همانند سازمان ملل) برای **حکمرانی**، پیش
از بسته شدن این پنجره است.



[QR CODE
PLACEHOLDER]

[LINK TO UN REPORT
PLACEHOLDER]