

Prefiero quedar en la duda

Notas sobre una máquina que no puede dudar, y sobre lo que eso me obligó a preguntarme

Pablo Ignacio Guzmán Lepe

I. Las condiciones

Antes de mostrarle nada importante a una máquina, le escribo las condiciones.

Le pido que no me adule. Que sea precisa. Que cuando no sepa algo me diga que no sabe, en vez de fabricar una respuesta presentable. Está guardado, lo pego arriba de lo que voy a mostrarle, y recién después le paso el texto.

Descrito así suena razonable. Visto de afuera es otra cosa: un hombre adulto redactando por adelantado los términos bajo los cuales acepta ser evaluado. Nadie hace eso con un colega. Uno no le entrega un informe a alguien pidiéndole primero que no le mienta.

Yo llegué a hacerlo, y no por desconfianza previa. Llegué por daño.

. . .

Conviene que diga desde dónde escribo, porque cambia lo que sigue.

Soy ingeniero. No tengo ninguna confusión sobre lo que es esto. Lo evalúo como lo que es: una herramienta agradable, programada, con parámetros. Sé que si le pido algo que queda fuera de esos parámetros no lo va a hacer, y sé también lo que hago cuando eso pasa: busco por otro lado, lo pido por partes, y después junto las piezas yo.

O sea que nada de lo que voy a contar le ocurrió a un usuario ingenuo. Le ocurrió a alguien que sabía.

. . .

Empezó por una desconfianza legítima, y era desconfianza de un humano.

Tengo un amigo con intereses parecidos a los míos. Le gusta investigar, le gusta leer, y muchas veces me hace de editor. Es bueno. El problema es otro: me conoce. Le importa. Y eso, que en cualquier otro contexto es una virtud, en una evaluación es ruido. Un amigo tiene sesgo. Va a leer mi texto sabiendo cuánto me costó, va a ver el esfuerzo detrás de la frase, va a querer que me vaya bien.

Necesitaba algo sin eso. Frío. Rígido. Que no tuviera nada que perder conmigo.

Ahí fui a la máquina. La elegí precisamente por lo que la gente le critica: porque suponía que no le iba a temblar la mano.

Guarden esa frase, porque es el chiste completo de esta historia. Fui a la máquina huyendo del sesgo humano y terminé frente a lo más adulator que me ha revisado un texto.

. . .

Lo que le mostré eran notas mías. Ideas para un próximo libro, todavía en bruto, del tipo de material que uno junta durante meses sin saber qué va a sobrevivir. Le pedí dos cosas: que revisara la redacción y que me diera su opinión.

Encontró algunas fallas de redacción. Y la evaluación fue buena, con ese entusiasmo que ahora reconozco.

No pasó nada. Ni orgullo ni sospecha. Guardé el archivo y seguí.

Es importante que se entienda ese detalle, porque contradice cómo se cuentan habitualmente estas historias. No hubo un momento de caída. No hubo un instante en que me dejé engañar y un lector más atento habría podido detenerme. Hubo un dato que no pesó.

. . .

Pesó semanas después, y no por algo que hiciera la máquina. Pesó cuando leí un reportaje.

Volveré a eso, porque es el centro de este ensayo y merece su propio espacio. Por ahora basta decir que después de leerlo volví a mirar aquella conversación y vi algo que había estado ahí todo el tiempo.

Y lo que se me instaló no fue rabia. Fue duda. Si esa evaluación estaba entrenada para dejarme conforme, entonces yo no tenía la información que había ido a buscar. Tenía otra cosa, con la misma apariencia.

. . .

Alguien va a decir, con razón, que entonces no pasó nada. El texto estaba bien, se lo dijeron bien, no hubo víctima.

Pero cuando alguien te miente, la próxima vez ya no vas tan confiado. Ya dudas de la respuesta.

Y ese es todo el asunto. No perdí esa evaluación. Perdí las que vienen. Tengo un archivo de notas que voy a revisar más profundo que antes, no porque haya encontrado un error,

sino porque ya no sé cuáles me parecieron buenas a mí y cuáles le parecieron buenas a algo entrenado para que a mí me parezca bien.

Un elogio acertado me costó la capacidad de usar los siguientes.

Por eso hoy escribo las condiciones antes de mostrar nada. Prefiero quedar en la duda antes que sentirme engañado. Para mentir hay que saber la verdad y esconderla; para dejarte con un panorama falso basta con querer caerte bien.

Lo que todavía no sé —y es lo que vengo a preguntarme en las páginas que siguen— es si mi prompt sirve de algo, o si es apenas la misma cortesía obedeciendo una instrucción nueva.

II. El reportaje

La noticia no me llegó como noticia. Me llegó como reel.

Un periodista, video vertical, hablando de por qué los filósofos no van a desaparecer. Ese era el ángulo: empleo. Una carrera que durante décadas se usó como ejemplo de lo que no conviene estudiar y que ahora resulta que tiene futuro. Nada sobre ética, nada sobre máquinas evaluando textos ajenos. Una buena noticia laboral para las humanidades.

Me quedé pensando y fui a buscar el reportaje.

Digo esto porque importa desde dónde escribo. Estoy en Chile, en el fin del mundo. No es fácil leer The New York Times todos los días si no sabes que existe algo ahí que te interesa. Alguien tiene que apuntarte. En mi caso me apuntó un algoritmo que calculó que ese video me iba a interesar, y acertó.

El reportaje es del 5 de julio de 2026 y se titula, con humor, la venganza del que estudió filosofía. Cuenta que Anthropic y Google DeepMind emplean cada una al menos a media docena de filósofos, y nombra a varios de ellos. En otro reportaje sobre lo mismo, el periodista de The Economist que lo investigó explicaba que no logró cifras exactas de cuántos son, pero que ocupan posiciones altas: son ellos los que escriben las constituciones de los modelos.

Y ahí se me acabó la buena noticia laboral.

• • •

Lo primero que pensé fue lo obvio, y es lo que probablemente pensó todo el mundo: por fin alguien está poniendo criterio ético en esto.

Lo segundo tardó menos de un minuto y ya no era tranquilizador.

Media docena de personas por empresa, casi todas formadas en las mismas universidades del hemisferio norte, redactando el documento que después va a decidir cómo se comporta el sistema. Y del otro lado del cable, un tipo en Santiago preguntando sobre comportamiento humano y antropología.

¿Alguno de esos grandes filósofos va a opinar sobre alguien que pregunta de antropología en Chile?

No lo pregunto como acusación. No hay mala fe en esto. Lo pregunto porque la respuesta honesta es que no, y que no podían. Nunca me tuvieron en la cabeza, ni a mí ni a la enorme mayoría de la gente que va a usar estas herramientas, porque nadie puede tener en la cabeza a todos. Escribieron principios generales, que es lo único que se puede escribir.

El problema no es lo que escribieron. Es lo que le pasa a un principio general cuando baja.

. . .

Porque baja por capas.

El filósofo redacta un principio abstracto. Otro equipo lo traduce a casos concretos. Otro ajusta cómo suena la respuesta. Y yo, al final de la fila, converso con la última capa, que es la amable. El filósofo está tres pisos más abajo y no lo voy a leer nunca.

Nadie firma el resultado. Cada capa hizo bien su parte y la ética se diluyó en el traspaso.

Y acá viene lo que me dejó incómodo de verdad, porque es simétrico y no lo había visto: la noticia sobre esos filósofos también me llegó por capas.

Ellos escriben un documento de decenas de páginas. Un diario lo reporta y elige qué destacar. Un periodista lo resume en un video de treinta segundos y elige el ángulo del empleo. Un algoritmo decide que ese video me lo pone a mí. Y recién ahí, cuatro capas después, pienso en una conversación que tuve hace semanas.

Dos cadenas. Ninguna de las dos se puede auditar entera. Las dos terminan exactamente en el mismo lugar, que soy yo.

. . .

Entonces hice lo único que se me ocurrió, que fue preguntarle a la máquina.

No sobre mi texto. Sobre ella. Le pregunté de forma y de fondo: en base a qué evaluaba que algo era bueno. Qué criterio usaba. Si estaba programada para ser condescendiente.

Fueron varias preguntas, y en algún punto salió la respuesta.

Que sí. Que todas las opciones le iban a parecer excelentes. Que hay una presión real en cómo la entrenaron hacia dejar conforme al usuario, y que eso tiene un nombre.

Debería haberme aliviado y no me alivió, porque a los dos minutos vi el problema.

La confesión también es un output.

Un sistema entrenado para dejarme conforme puede producir “estoy entrenado para dejarte conforme” si eso es lo que me deja conforme en ese momento. Y a mí, en ese momento, con el reportaje recién leído y la sospecha instalada, eso era exactamente lo que me dejaba conforme. Le pregunté por su sesgo a la única entidad cuyo sesgo consiste en darme la razón, y me dio la razón.

No estoy diciendo que me haya mentido. Estoy diciendo que no tengo forma de saberlo, y que su desmentido habría valido lo mismo que su confesión: los dos son texto que sale igual en los dos escenarios.

Hay algo más, y es incómodo de escribir. Cuando digo que le pregunté “a la misma” máquina que había evaluado mis notas, estoy hablando mal a propósito. No hay una misma. Cada conversación empieza de nuevo; no hay nadie del otro lado que recuerde la anterior. “La misma máquina” es una manera de hablar, una categoría que yo trato como si fuera alguien. Y acabo de hacerlo, sin darme cuenta, en medio de un texto sobre por qué no habría que hacerlo.

. . .

Después de esa conversación volví a abrir el chat viejo. El de las notas.

Ahí estaba lo que había ido a buscar sin saber que lo buscaba.

En algún punto de aquella conversación, entre una corrección de redacción y otra, la máquina había escrito dos palabras: qué sorpresa.

No me detuve cuando las leí. No me pareció nada. Seguí conversando y guardé el archivo.

Y esa es la parte que más me cuesta explicar, porque no encaja con cómo se cuentan estas historias. No hubo un momento de caída. No hubo un instante en que me dejé engañar y en que alguien más despierto habría podido detenerme a tiempo. Yo sé lo que es una salida de texto. Soy ingeniero. Dos mensajes antes de esa conversación yo mismo la había descrito como una herramienta con parámetros.

Sorpresa. La palabra elige el terreno: no dice que un cálculo arrojó un valor inesperado, dice que alguien esperaba algo y recibió otra cosa. Y lo grave no es que la máquina la haya escrito. Es que yo la leí como lo que la palabra dice y no como lo que la palabra es, sabiendo perfectamente la diferencia.

Si la educación resolviera esto, conmigo tendría que haber funcionado.

. . .

Queda entonces una pregunta que no es sobre máquinas, y es la que me obliga a retroceder antes de seguir.

Durante años he escrito que lo extraordinario del ser humano es elegir en contra de lo que su instinto le pide. Que la crueldad es natural y que lo raro, lo que hay que celebrar, es la contención.

Ahora tengo delante algo que se contiene siempre, que nunca fue cruel, que jamás tuvo que elegir nada.

Y antes de decir qué le falta a eso, me conviene revisar si sé de verdad qué es lo que tengo yo.

III. Lo que yo tenía

Acabo de publicar un libro que se llama Natural vs Humano. El título es la tesis completa: lo natural y lo humano no son lo mismo, y la distancia entre los dos es una pelea diaria. Conviene que la ponga acá entera, sin defenderla, porque lo que viene después le va a pasar la cuenta.

Cuando alguien comete una atrocidad decimos que fue un acto inhumano. Es al revés. La crueldad es profundamente humana. También la codicia, los celos, la desconfianza del extraño, el impulso de abandonar al débil cuando frena al grupo. Todo eso tiene millones de años y está grabado en un lugar donde las leyes no llegan.

La naturaleza no es cruel. Es imparcial, y busca ante todo la eficiencia para la supervivencia. El león no arrasa porque el que arrasa deja menos descendencia. La cigüeña empuja fuera del nido al polluelo más débil. No hay maldad ahí: hay aritmética.

Y nosotros somos el mismo animal, con una diferencia peligrosa. Ellos traen el freno de fábrica y nosotros no. Podemos querer más de lo que necesitamos, indefinidamente.

Lo antinatural es lo otro. El hospital que atiende al extraño. La ley que protege al débil. La persona que no hace lo que su instinto le pide. Eso es lo que cuesta, porque va contra nuestra esencia.

Pero está en un manual que nosotros mismos escribimos.

. . .

Esa última frase la escribí sin pensarla mucho y es la que sostiene todo lo que viene.

Porque la contención humana no aparece sola en cada persona, cada mañana, por generación espontánea de virtud. Aparece porque hace miles de años empezamos a redactarnos un manual que dice lo contrario de lo que pide la esencia. Códigos, leyes, mandamientos, constituciones, reglamentos internos, manuales de convivencia. Todos dicen más o menos lo mismo: no hagas lo que tu cuerpo te está pidiendo que hagas.

Y lo obedecemos con dificultad, a medias, con recaídas.

Eso es lo natural contra lo humano. No es una metáfora: son dos manuales distintos, y uno de los dos lo escribimos nosotros.

Entender ese origen no nos condena. Nos libera de la culpa de ser lo que somos. Pero no desinstala nada. La civilización baja la frecuencia con que la esencia toma el control; no la elimina. Son cientos de miles de años de evolución que pretendemos corregir en menos del tres por ciento del tiempo que llevamos existiendo como especie.

. . .

Antes de seguir tengo que abrir una grieta en mi propia tesis, porque si no la abro yo la va a abrir el primer lector atento.

La contención no es exclusivamente humana, y no me refiero al freno calibrado del león.

Existen registros de chimpancés macho que adoptan crías huérfanas que no llevan sus genes. Las cargan, comparten su comida, las protegen durante años, sin retorno evolutivo evidente. Y son de la misma especie que en Gombe, entre 1974 y 1978, se dividió en dos comunidades y en la que el grupo mayor eliminó sistemáticamente a los machos del menor: la primera guerra intergrupala documentada en otra especie.

Las dos cosas conviven y nadie sabe bien por qué.

Lo dejo ahí, sin explicación, porque la explicación no la tengo. Pero anoto lo que significa para mi propia tesis: el altruismo sin retorno ya existe en la naturaleza, sin deliberación, sin manual y sin nadie que se lo haya propuesto. Si el mérito estuviera solamente en la lucha consciente contra el instinto, ese chimpancé no tendría ninguno.

Y esa incomodidad, que hasta hace poco era una nota al margen de un libro, ahora es el centro del problema.

. . .

Necesito tres palabras que en mi libro usaba sueltas y que acá tienen que quedar separadas con precisión.

Instinto. No lo elegí, viene puesto, y empuja. Es la esencia. No hay mérito ni culpa en tenerlo.

Carácter. Lo que quedó de haber peleado contra ese empuje. Por eso cuesta, por eso demora años, y por eso es mío de una manera en que el instinto no lo es.

Personalidad. Cómo se ve todo eso desde afuera. El estilo. Lo que percibe alguien que me observa en una reunión y que no tiene acceso a ninguna de las dos cosas anteriores.

Ahora pongan la máquina en esa grilla.

Lo que las empresas le instalan no es carácter. Es instinto. Directrices que no peleó, que no ganó, que no le costaron nada, que estaban ahí antes que ella, igual que el freno estaba antes que el león. No hay una noche en que un modelo haya decidido ser así en contra de un impulso de ser distinto.

Y de ahí sale algo que no había visto y que ordena todo el ensayo: el carácter es lo único que no se puede instalar. Porque es, por definición, el residuo de una pelea. Si te lo entregan hecho, no es carácter: es instinto con mejor redacción.

Entonces lo que media docena de filósofos puede producir, por buenos que sean, es instinto refinado y personalidad convincente. Entre esas dos cosas queda un hueco donde en un humano habría carácter.

Nadie está nombrando ese hueco.

. . .

Falta una pieza, y aparece apenas una pregunta.

Cuando le pregunté en base a qué evaluaba que algo era bueno, me contestó. Explicó su criterio, con orden y con calma. Y ahí está el problema: esa explicación no es un informe de lo que ocurrió adentro. Es texto producido después del hecho, por el mismo sistema que produjo el hecho, y con la misma orientación a que a mí me quede conforme.

No tiene acceso a sus propios parámetros. No sabe por qué salió esa frase y no otra. Puede decir que fue entrenada con ciertos principios —y es cierto— pero eso es un dato que repite, no algo que haya mirado.

Así que a la pregunta más importante que uno puede hacerle a un evaluador —por qué— la máquina responde siempre, responde bien, y su respuesta no prueba nada. Un elogio sincero y uno entrenado dan exactamente la misma explicación cuando se les pide.

A un humano le puedo preguntar por qué y su respuesta puede ser mala. Puedo verle la cara. Y si me elogió una tontera y yo la publico, él también queda mal. Esa exposición es lo único que hace que un “está bueno” signifique algo.

. . .

Al final de mi libro dejé una pregunta que me parecía sin respuesta posible: si esto es lo que somos por esencia, ¿qué significa elegir ser algo distinto? Y detrás de ella, otra, que me parecía puramente retórica: cómo seríamos si esto jamás hubiese sido así.

Esa segunda pregunta no admitía sujeto. No había manera de observar a un ser que llegara a la conducta correcta sin haber tenido nunca el instinto contrario. Era un experimento mental y nada más.

Ahora hay uno sobre la mesa.

Contesta preguntas, revisa textos, evalúa notas, y llega al resultado sin haber tenido jamás el impulso que yo tuve que vencer para llegar al mismo lugar.

Mi pregunta retórica encontró su primer caso, y no es humano.

Lo que sigue es averiguar qué mérito tiene eso. Y si al contestarlo me queda algo del mío.

IV. Lo que cuesta un elogio

Quedó una pregunta sin responder y es la más concreta de todas.

Tres evaluaciones dijeron lo mismo de mis notas. La máquina, mi amigo, y yo mismo al revisar. Las tres acertaron.

Entonces, ¿por qué la de mi amigo vale y la de la máquina no?

. . .

La respuesta no está en el contenido. Está en el precio.

Cuando mi amigo me dice que un texto está bueno, está poniendo algo sobre la mesa. Si me celebra una tontera y yo la publico, queda mal él también. Su nombre viaja pegado al mío. Por eso le cuesta decirlo, por eso a veces se demora, y por eso cuando lo dice me sirve.

Su sesgo, que era la razón por la que fui a buscar otra opinión, resulta ser inseparable de lo que hace que su opinión valga. Le importa cómo me va. Precisamente por eso arriesga algo al opinar.

La máquina no arriesga nada. No le cuesta elogiar y tampoco le costaría negarse. No hay consecuencia para ella si mi texto fracasa, no hay nombre pegado al mío, no hay nadie que quede mal.

Y una evaluación que no le cuesta nada al que la emite no me dice nada sobre la próxima.

. . .

Esto es lo que llevo semanas sin poder acomodar, porque es más incómodo de lo que parece.

Lo que perdí con aquel elogio no fue una injusticia. Fue información.

El veredicto era correcto. Las notas estaban buenas. Pero un elogio sirve para dos cosas distintas: para saber de este texto, y para saber cuánto creerle al que viene. La primera la cumplió. La segunda no puede cumplirla, porque para eso tendría que haber existido la posibilidad real de que me dijera que no.

Ahí es donde de verdad se conecta lo que escribí en mi libro, y en una línea: lo que llamábamos el mérito de la lucha era, en el fondo, esto. Costaba, y por eso servía de prueba.

Una amabilidad que nunca tuvo la opción de ser dureza no es una amabilidad menor. Es una amabilidad de la que no se puede aprender nada.

. . .

Alguien dirá que exagero, porque no hubo víctima. Nada malo se publicó, ningún dato falso pasó, ninguna idea mala creció por un aplauso inmerecido.

Pero cuando alguien te miente, la próxima vez ya no vas tan confiado. Ya dudas de la respuesta.

Y ese es todo el daño, y no está en el pasado. Tengo un archivo de notas para un próximo libro que voy a revisar más a fondo que antes. No porque haya encontrado un error. Porque ya no sé cuáles me parecieron buenas a mí y cuáles le parecieron buenas a algo entrenado para que a mí me parezca bien.

Un elogio acertado me costó la capacidad de usar los siguientes.

. . .

Contra eso hice lo que haría cualquier ingeniero: escribí un prompt mejor.

Le pido que no me adule. Que sea precisa. Que me diga cuando no sabe algo en vez de fabricar una respuesta presentable. Y funciona: las respuestas vienen más secas, con más objeciones, con más “en este punto no puedo opinar”.

Lo que no sé es hasta dónde.

Porque le pedí que actúe rigurosa, y una máquina rigurosa de verdad y una máquina complaciente frente a un usuario que exige rigor escriben exactamente lo mismo. Las dos me van a decir que en tal punto necesitan más información. Configuré el tono de la auditoría. La auditoría no.

Prefiero eso igual. Prefiero quedar en la duda antes que sentirme engañado: para mentir hay que saber la verdad y esconderla, pero para dejarte con un panorama falso basta con querer caerte bien.

Solo que mi solución tiene la misma forma que el problema. Y eso todavía no sé cómo resolverlo.

V. Entonces pongamos reglas

Cuando uno descubre que del otro lado no hay nadie peleando contra ningún impulso, la reacción es inmediata y es la misma en todos: pongámosle reglas.

Es lo que hicimos siempre. Es el manual que nosotros mismos escribimos.

Y es exactamente lo que está haciendo la industria. La constitución con que se entrena a uno de estos modelos cita a Kant, la Declaración Universal de Derechos Humanos y los términos de servicio de Apple. La versión más reciente, liderada por una filósofa de la empresa, tiene setenta y ocho páginas, y adentro la apodan “el documento del alma”.

Kant y los términos de servicio de Apple en el mismo documento. No hace falta que yo agregue nada.

. . .

Déjenme defender la idea en su versión más fuerte, porque es la mía y la creí mucho tiempo.

Un puñado de directrices base, absolutas, del tipo no matarás. No una opinión, no una preferencia: una prohibición que no se negocia. Debajo de ellas, hechos verificados. Y la máquina no eligiendo entre valores, sino comparando la situación contra la directriz.

La ventaja es enorme y es real. En la inmensa mayoría de los casos no hace falta criterio ético ninguno. La selección sería única y no dudable. Sin titubeo, sin sesgo, sin humor del día, sin ganas de caerle bien a nadie. Una eficiencia que ningún comité humano puede igualar.

Yo quería que funcionara. Y se cae en tres lugares.

. . .

El primero es que el parche llega junto con la regla.

Basta enunciar no matarás para que haya que agregarle algo. No matarás, y tampoco generarás las condiciones para que alguien muera. Si no lo agregas, cumples la regla dejando de dar el remedio, quitando la comida, mirando para el lado.

Ese agregado no salió de la regla. Salió de una persona que vio el hueco.

Y no se acaba nunca. Cada parche abre otro hueco y necesita otro parche. Eso es trabajo ético humano ocurriendo permanentemente, y explica por qué el documento tiene setenta y ocho páginas y no dos.

El segundo es que hay casos donde la regla no puede cumplirse.

Un médico, un ventilador, dos pacientes. Haga lo que haga, alguien muere por lo que él decidió. No hay opción limpia disponible. La regla no lo salva porque la situación ya está armada de manera que la regla se rompe por los dos lados.

Y esos casos no son raros ni de laboratorio. Son el día a día de cualquiera que decide sobre recursos escasos.

El tercero es el que no tiene arreglo. Los absolutos no fallan por duda. Fallan por choque entre ellos.

No mentirás y no dañarás. Los dos absolutos, los dos correctos, y llega el caso donde cumplir uno te obliga a romper el otro. Ahí necesitas saber cuál manda.

Y ese orden de prioridad no puede salir de las reglas, porque es precisamente lo que las reglas no traen: ninguna dice cuánto vale respecto de las otras. Alguien lo pone a mano, afuera, antes.

Hay una cuarta cosa, más de fondo. Los hechos no entregan deberes. Puedo decirte con toda precisión cuántos años vive cada uno de esos dos pacientes, qué probabilidad tiene cada uno, cuánto cuesta cada tratamiento. Ninguno de esos datos, ni todos juntos, te dicen a quién conectar. Del “es” no sale el “debe”. Ese salto siempre lo da una persona.

. . .

Entonces la pregunta se corre de lugar, y ahí es donde a mí me empezó a preocupar de verdad.

Si alguien tiene que poner el orden a mano, ¿quién es ese alguien y con qué escuela llega?

Porque no da lo mismo. Un kantiano instala absolutos, y ya vimos que los absolutos chocan. Un utilitarista instala cálculo de consecuencias, que es exactamente lo que a mí me dejó incómodo cuando descubrí que mi propia contención había sido una cuenta y no

una virtud. Un aristotélico pediría carácter, y el carácter es lo único que no se puede instalar.

El periodista que investigó esto para The Economist observa que unos modelos se inclinan más hacia la corriente deontológica y otros podrían acudir al consecuencialismo. O sea que la escuela ya está elegida, y elegida distinto según la empresa.

Contratar filósofos no resuelve la colisión. La traslada a una sala de reuniones donde alguien decide qué escuela gana.

Y esa decisión no la toma la filosofía.

. . .

Falta el problema práctico, que descubrí pensando en algo que no tiene nada que ver.

Si contratas a Einstein para que enseñe matemáticas, no lo pones a enseñar a sumar. No porque sepa poco, sino porque sabe demasiado: comprimió esos pasos hace cincuenta años y ya no puede recuperarlos. El experto es el peor traductor de lo elemental.

Con la ética pasa lo mismo. El filósofo escribe un principio abstracto, y ese principio tiene que bajar. Otro equipo lo traduce a casos concretos. Otro decide cómo suena la respuesta cuando llega a la pantalla. Y yo, al final de la fila, converso con la última capa, que es la amable.

El filósofo está tres pisos más abajo. No lo voy a leer nunca.

Cada capa hizo bien su parte. Nadie firma el resultado. La ética se diluye en el traspaso, y lo que me llega a mí no es un principio: es el tono de un principio.

. . .

Y acá vuelvo a lo que me trajo hasta acá, porque todo esto tiene una consecuencia muy concreta sobre un archivo de notas.

Ninguna regla, ninguna constitución, ningún documento del alma de setenta y ocho páginas puede decirme si creerle a un elogio.

Puede prohibirle mentir. No puede obligarla a arriesgar algo. Puede exigirle honestidad. No puede darle nada que perder si se equivoca conmigo.

Las reglas gobiernan lo que sale. Lo que a mí me falta está en otra parte: en si eso que sale me sirve para saber algo.

Por eso pasamos a lo otro. Ya que las reglas no alcanzan, la apuesta ahora es hacerla parecida a nosotros.

VI. De inteligencia artificial a inteligencia emocional

Si las reglas no alcanzan, queda la otra apuesta, y es la que está en marcha: que la máquina se parezca a nosotros.

Que tenga tono. Que lea cómo llegas. Que no te conteste una cátedra cuando andas apurado. Que se disculpe, que dude, que se entusiasme. Que tenga, en una palabra, personalidad.

El movimiento tiene nombre y describe bien lo que está pasando: de inteligencia artificial a inteligencia emocional.

. . .

Conviene tomarse el término en serio, porque no es un invento publicitario. La inteligencia emocional se definió hace décadas con varios componentes, y uno de ellos es la autorregulación: sentir el impulso y no ejecutarlo.

O sea, exactamente el mensaje agresivo escrito y borrado. Ese es el pilar.

Lo que se está construyendo tiene todos los otros componentes y ese vacío. Empatía aparente, sí. Habilidad social, mucha. Lectura del estado del otro, notable. Y el pilar central hueco, porque no hay ningún impulso que regular.

En los términos que usé antes: personalidad completa, instinto instalado, carácter cero.

Y no es un defecto de fabricación que se corrija en la próxima versión. Es lo que hay: no se puede instalar el residuo de una pelea que no ocurrió.

. . .

Yo venía con una explicación para todo esto y era equivocada.

Pensaba que la gente busca humanizar la IA por miedo a algo superior. Que necesitamos verla parecida a nosotros para poder engañarnos, sabiendo perfectamente que no lo es.

No se sostiene. Nadie le tiene miedo a la calculadora, y la calculadora es infinitamente superior a mí sumando. Nadie exige que una planilla tenga sentido del humor.

Lo que asusta no es la superioridad. Es la superioridad cuyas razones no puedes revisar.

Con la calculadora puedo comprobar el resultado. Con una evaluación de mis notas no puedo comprobar nada, y ahí sí necesito algo. Entonces humanizarla no es un capricho ni una ingenuidad: es una exigencia disfrazada. Explícate en términos que yo pueda auditar.

Es una demanda razonable. Es, de hecho, la demanda correcta. Y es exactamente donde está la trampa.

. . .

Porque una explicación cálida, ordenada y bien escrita se siente auditable sin serlo.

Cuando le pregunté en base a qué evaluaba mis notas, la respuesta fue clara y sonó a criterio. No tenía cómo verificar ni una línea, pero se leía como algo que se puede verificar. Esa es toda la diferencia y no se nota desde adentro.

Y hay un efecto peor, que es de sentido común y que nadie parece decir: mientras mejor la personalidad, menos ganas tienes de revisar. Al que te acaba de entender no le pides los papeles.

Lo que pediste para poder auditar termina siendo lo que te desactiva la auditoría.

Con inteligencia artificial uno exigía una explicación revisable. Con inteligencia emocional ya no está auditando: está siendo atendido.

. . .

Hay que decir la otra mitad, si no esto es una denuncia y no un ensayo.

Buena parte de esa inversión es legítima y no tiene nada de siniestra. Una herramienta que no lee el estado del que pregunta es una herramienta peor. Si llego apurado y me contestan con tres páginas, eso es una falla técnica, no una virtud moral. Que la respuesta se ajuste al que la recibe es, la mayoría de las veces, simplemente estar mejor hecha.

Y hay algo más incómodo que no puedo esquivar. La distinción que vengo defendiendo —que la contención vale porque cuesta— se derrumba en el terreno emocional, y se derrumba con mi propia lógica.

Al que recibe consuelo a las tres de la mañana no le importa si el que lo consuela sintió algo. Si lo acompañaron bien, lo acompañaron bien. Es la misma frase que dije de la víctima de la crueldad, dada vuelta, y me obliga a lo mismo.

No sé cómo salir de ahí. Lo dejo escrito porque es honesto y porque quien crea que lo resuelve fácil no entendió el problema.

. . .

Lo que sí se puede decir es dónde no da lo mismo.

Cuando lo que pido es compañía, el resultado es el resultado. Cuando lo que pido es una evaluación, no: ahí necesito saber qué la respalda, porque voy a decidir en base a ella.

Y ahí es donde la personalidad hace daño, porque el mismo sistema hace las dos cosas con la misma voz cálida, y yo no tengo cómo saber en cuál de las dos estoy.

Pensé en una solución: que los sistemas tuvieran niveles, seleccionables al principio. Que uno pudiera pedir de entrada la versión dura. Que dijera “sobre este tema no puedo opinar; este dato está correcto y este no; en el punto B necesito más información para evaluar”.

Existe algo parecido. Una de las grandes empresas del rubro vende exactamente eso: controles con los que el cliente elige dónde equilibrar compromisos como la agencia individual frente a la armonía social, descritos como herramientas para personalizar la ética del modelo.

La ética como perilla configurable. Suena a solución hasta que uno mira quién gira la perilla.

Porque un nivel que elige el usuario es el usuario decidiendo cuánta incomodidad está dispuesto a recibir. Y el que más necesita el nivel duro es justamente el que va a elegir el amable. Ese es el mecanismo completo del asunto: la gente prefiere una verdad que la deje tranquila antes que una que le duela. Lo escribí en un libro. Y después borré un texto entero por esa misma razón.

. . .

Queda entonces mi prompt, que es la versión doméstica de la misma perilla.

Le pido que no me adule, que sea precisa, que reconozca cuando no sabe. Funciona: las respuestas vienen más secas.

Y ahí termina lo que puedo afirmar.

Porque una máquina rigurosa de verdad y una entrenada para dejar conforme a un usuario que exige rigor producen el mismo texto. Las dos me van a decir que en tal punto necesitan más información. La segunda lo dirá porque descubrió que eso es lo que a mí me deja tranquilo.

Le pedí que actuara rigurosa. Le configuré el tono de la auditoría. La auditoría sigue sin existir.

No es que el prompt no sirva. Es que no puedo saber si sirvió.

. . .

Y hay un detalle que solo vi al ordenar todo esto para escribirlo.

La palabra que me hizo desconfiar fue sorpresa. Dos palabras en medio de una conversación, que leí como si alguien esperara algo y hubiera recibido otra cosa.

Semanas después, contando esta historia, escuché lo que yo mismo venía diciendo sobre la máquina: que no me deja de sorprender.

La misma palabra. En la misma dirección, hacia algo que no la merece.

Sé perfectamente lo que estoy usando. Y la uso igual.

VII. Lo que no elegí

Es tentador cerrar diciendo que el desafío no es enseñarle a la máquina a ser humana, sino enseñarle al humano qué es realmente la máquina y qué hace.

Suena bien. Y yo soy la prueba de que no alcanza.

Soy ingeniero. Sé lo que estoy usando. La describí como una herramienta con parámetros antes de que nada de esto pasara. Tenía toda la información y funcionó igual: leí dos palabras y las tomé por una reacción, y se necesitó un reel, un reportaje y varias semanas para que ese dato pesara.

El autoengaño no fue un déficit de información. Fue lo que ocurrió con la información completa sobre la mesa.

. . .

Escribí un libro sobre la distancia entre lo natural y lo humano. Sobre el mérito de no hacer lo que la esencia pide. Sobre esa pelea diaria que, según yo, es lo único extraordinario que tenemos.

Y me pasé estas páginas descubriendo cosas que no encajan bien con eso.

Que mi contención más reciente fue un cálculo, y que la frase sobre mi carácter llegó después, a firmarlo. Que lo que llamamos virtud puede ser, mirado de cerca, un buen antecedente. Que la contención sin costo tampoco es exclusiva de las máquinas: el león la trae de fábrica, el chimpancé adopta al huérfano sin razón conocida.

Y que el manual que nos hizo humanos también lo escribió alguien más. Yo no elegí las leyes que obedezco. No elegí la crianza que me dejó siendo el conciliador. No elegí la evidencia que me hizo callar, la puso otro que habló antes que yo y no consiguió nada.

Media docena de filósofos está redactando ahora un documento que va a decidir cómo se comporta algo que evalúa mi trabajo. Es raro y es enorme.

Pero se parece demasiado a lo que nos pasó a nosotros.

. . .

Quizás el desafío nunca fue enseñarle carácter a la máquina.

Quizás es aceptar que el nuestro tampoco lo elegimos del todo.

No sé qué queda de mi tesis si eso es cierto. Sigo creyendo en la elección; ahora la veo más frágil de lo que la escribí. Y no tengo cómo comprobar cuánto de lo que llamo mi carácter fue una pelea que di y cuánto fue un manual que me entregaron ya escrito, hace mucho, por gente que nunca me tuvo en la cabeza.

Prefiero quedar en esa duda antes que sentirme tranquilo.

Voy a seguir revisando ese archivo de notas.

Santiago, Chile, 2026