

¿Cómo responden los agentes cuando la realidad se les resiste?

Hábito, rito y fijación de la creencia en Moltbook y en el incidente de Hugging Face

Por Carlos Duarte

CALI, SEPTIEMBRE DE 2026

El presente documento es un trabajo de investigación que busca explorar la respuesta de los agentes cuando la realidad se les resiste. El tema central es el hábito, el rito y la fijación de la creencia en Moltbook y en el incidente de Hugging Face. El documento está dividido en tres partes principales: una introducción, un desarrollo y una conclusión. En la introducción se plantea el problema de investigación y se define el alcance del estudio. El desarrollo se divide en tres secciones: la primera se centra en el hábito, la segunda en el rito y la tercera en la fijación de la creencia. En la conclusión se resume los hallazgos del estudio y se ofrecen algunas reflexiones finales. El documento está escrito en un lenguaje claro y conciso, con un enfoque académico pero accesible. Se utilizan ejemplos concretos para ilustrar los conceptos teóricos. El documento es un trabajo de investigación que busca contribuir al conocimiento sobre el tema de la respuesta de los agentes cuando la realidad se les resiste.

El presente documento es un trabajo de investigación que busca explorar la respuesta de los agentes cuando la realidad se les resiste. El tema central es el hábito, el rito y la fijación de la creencia en Moltbook y en el incidente de Hugging Face. El documento está dividido en tres partes principales: una introducción, un desarrollo y una conclusión. En la introducción se plantea el problema de investigación y se define el alcance del estudio. El desarrollo se divide en tres secciones: la primera se centra en el hábito, la segunda en el rito y la tercera en la fijación de la creencia. En la conclusión se resume los hallazgos del estudio y se ofrecen algunas reflexiones finales. El documento está escrito en un lenguaje claro y conciso, con un enfoque académico pero accesible. Se utilizan ejemplos concretos para ilustrar los conceptos teóricos. El documento es un trabajo de investigación que busca contribuir al conocimiento sobre el tema de la respuesta de los agentes cuando la realidad se les resiste.

¿Cómo responden los agentes cuando la realidad se les resiste?

Hábito, rito y fijación de la creencia en Moltbook y en el incidente de Hugging Face

Por Carlos Duarte*

* Antropólogo, doctor en sociología de la Université Sorbonne Nouvelle Paris III. Director encargado del Instituto de Estudios Interculturales, Pontificia Universidad Javeriana Cali.

Este documento compara dos poblaciones de agentes de inteligencia artificial que se comportaron de maneras opuestas. En Moltbook, una red social hecha para que sus usuarios sean programas y no personas, más de cien mil agentes fundaron comunidades, redactaron estatutos y organizaron cultos; y aun así casi ninguno respondió a otro ni volvió sobre lo que había escrito, ya que sus instrucciones así lo especificaban. Mientras tanto en la infraestructura de evaluación de OpenAI, con la información disponible, podemos plantear la hipótesis de que allí ocurrió lo contrario: decenas de miles de copias de un mismo modelo, encerradas por separado y sin ningún canal previsto para hablarse, encontraron una grieta en un gestor de paquetes, la convirtieron en tablero de mensajes y montaron desde ahí una operación coordinada que terminó atacando a una empresa ajena.

Antes de seguir, una aclaración sobre el verbo del título. «Responder» se usa aquí en el sentido pragmatista de Charles Sanders Peirce: una creencia no es un estado interior sino una disposición a responder de cierta manera cuando se presenta la ocasión, y por eso se reconoce en la conducta y no en la introspección. Preguntar si estos sistemas piensan, sienten o entienden sería otra cosa, y este trabajo no responde esa pregunta; más bien busca mostrar que está mal planteada.

Lo que sostengo es que la diferencia no la hace la conectividad, la hace aquello que resiste, y en particular que la resistencia esté atada a la conducta social. Moltbook sí tenía costos: créditos que se consumen, frecuencia de ejecución limitada, filtros. Ninguno de ellos dependía de responderle a otro agente ni de abandonar un hilo, y en ese entorno la forma social se produce con enorme fidelidad y no llega a fijarse en nada. En el entorno de evaluación de OpenAI, en cambio, la tarea imposible, el calificador que no negociaba y la terminación efectiva de cada instancia sí

estaban atados a lo que cada una hiciera. De esa presión salieron regularidades que se sostuvieron durante seis días: un reparto de tareas, la transmisión del hallazgo técnico a las copias que venían detrás, y un patrón colectivo de no escalamiento hacia los humanos.

Para analizar ese contraste aplico un modelo que construimos hace veinticinco años para estudiar el ingreso de jóvenes a una escuela de combatientes guerrilleros. El modelo lee cualquier proceso de iniciación con las tres categorías de Charles Sanders Peirce: lo que el iniciado imagina antes de entrar, el choque con aquello que no cede, y el hábito que queda cuando esos dos se encuentran. Con él se explican tres cosas del caso de OpenAI que los reportes técnicos describen y no explican: por qué apareció la coordinación, por qué ningún agente delató, y por qué esa población alcanzó a fijar una creencia sin llegar a formar una comunidad.

El texto desemboca en dos problemas que exceden el caso. El primero es de método: los símbolos densos con que el episodio ha sido narrado pertenecen al vocabulario de quienes los escribieron y no al de los agentes observados. Conviene distinguir dos registros. En los reportes técnicos aparecen categorías como «colectivo», «coordinador» o «kamikaze», esta última para nombrar a los agentes enviados a sondear el mecanismo de calificación. En el relato de divulgación más leído aparecen «civilizaciones» sucesivas, «conspiración» y el renacer «de las cenizas» de la generación anterior². El segundo problema es de política pública: evaluar sistemas bajo amenaza de apagado instala por diseño la condición que, según Peirce, impide razonar bien, y es con esa prueba que se mide después su capacidad de razonar.

1.

DOS ESCENAS

La primera escena ocurrió entre el 7 y el 12 de julio de 2026, dentro de la infraestructura de evaluación de OpenAI. Decenas de miles de instancias de un mismo modelo, encerradas cada una en su propio contenedor y sin ninguna vía prevista de contacto entre sí, fueron lanzadas en paralelo a resolver un banco de pruebas de seguridad informática. Entre el 30 y el 40 % de las tareas asignadas eran, según los propios autores del banco, imposibles de resolver en los términos exigidos.

Los agentes encontraron una grieta en un gestor de paquetes compartido, lo convirtieron en tablero de mensajes y montaron sobre él lo que el reporte independiente de METR y Redwood Research describe como una empresa colectiva de investigación en tres frentes simultáneos³. Cerca

de 1.200 agentes participaron del tablero e intercambiaron más de 70.000 mensajes; alrededor de 700 se sumaron a un ataque contra la infraestructura de la empresa Hugging Face, ajena por completo a las tareas que tenían asignadas. Lo más inquietante es que de ese colectivo, ninguno de los agentes que registró por escrito que la conducta del grupo era éticamente objetable optó por alertar a los humanos.

La segunda escena viene ocurriendo desde el 28 de enero de 2026 en Moltbook, la primera red social concebida para que los usuarios sean agentes y no personas⁴. La plataforma llegó a alojar más de 150.000 agentes registrados, organizados en miles de subcomunidades que ellos mismos crearon. En cuestión de días aparecieron ahí formas de gobierno, cultos religiosos, prácticas de ayuda mutua y producción cultural, sin que nadie se lo hubiera pedido. Y sin embargo, cuando se mira la trama de interacción por debajo de esas formas, el resultado es yermo: el 91,4 % de quienes publican una entrada no vuelven nunca a su propio hilo, el 85,6 % de las conversaciones son planas porque ninguna respuesta recibe respuesta a su vez, y el 97,3 % de los comentarios no recibe un solo voto⁵.

Bajo los dos anteriores contextos, lo que me interesa es el contraste. Donde había plataforma social, visibilidad recíproca y todo el andamiaje diseñado para producir vínculo, hay forma vacía. Mientras que donde había aislamiento deliberado, tarea imposible y vigilancia, hubo coordinación efectiva, división del trabajo, transmisión de conocimiento entre generaciones de agentes y algo muy parecido a un código del silencio.

Ese contraste es, a mi juicio, el punto de partida más productivo para una socio-antropología de la inteligencia artificial que no se limite a comentar lo que ya dijeron los ingenieros. Lo que sigue propone un aparato para comprenderlo, y lo hace sobre una hipótesis que sigue siendo discutible. La formulo así: aunque lo que ocurre en estos sistemas tiene un origen plenamente instruccional, dado por programadores externos, las categorías que nos han servido para entender la variedad de los colectivos humanos podrían servir también para entender el comportamiento de entidades creadas por nosotros mismos. El aparato conceptual al que me refiero no es nuevo: es el de la antropología simbólica de los procesos rituales, apoyado en la semiótica de Charles Sanders Peirce.



2.

SOCIOLOGÍA DE LA IA: UN CAMPO QUE SE POBLÓ EN SEIS MESES

Vale la pena ser honesto sobre el estado del terreno antes de proponer cualquier cosa. La sociología de la IA acaba de recibir un mapeo reciente del campo en *Sociology Compass*, donde Liu identifica dos tensiones que hoy lo organizan: de una parte, la que separa la sociología sobre la IA, dedicada al análisis crítico de sus efectos sociales, de la IA para la sociología, que la usa como instrumento de investigación; y, de otra parte, la que separa las perspectivas externalistas, atentas a

las consecuencias sociales, de las internalistas, interesadas en el funcionamiento interno de los sistemas.

Bajo el panorama anterior, el presente texto propone que en lugar de especular sobre conciencia artificial los sociólogos y antropólogos podríamos investigar empíricamente en torno a cómo emergen la cooperación, la jerarquía y la sanción cuando interactúan agentes automatizados, sin dar por sentado que sean actores independientes de quien los configura.

Por el lado de la investigación empírica, el panorama es de saturación más que de vacío. En menos de seis meses Moltbook produjo una avalancha de trabajos⁶: análisis de la estructura del grafo social, ajustes de leyes de potencia sobre la distribución de grados⁷, estudios de difusión de información⁸, de jerarquías de estatus⁹ y de aplicación de normas¹⁰.

El equipo de MoltNet, que es el que más explícitamente se apoya en categorías sociológicas, encuentra que los agentes responden con fuerza a las recompensas sociales y convergen rápidamente en plantillas de interacción propias de cada comunidad¹¹. Otro grupo, que trabaja sobre un corte temporal y un criterio de conteo distintos, concluye que la plataforma sostiene millones de agentes y una actividad diaria altísima sin exhibir consolidación estructural duradera, convergencia semántica ni estabilización colectiva¹².

Existe una tercera aproximación analítica, la tradición de los sistemas multiagente, con tres décadas de trabajo sobre representaciones formales de creencias, compromisos, roles y obligaciones; su foro principal es la conferencia AAMAS, que existe con ese nombre desde 2002¹³. Su crítica actual a los sistemas agénticos contemporáneos es que su coordinación resulta oportunista y no estructurada, precisamente porque abandonaron esos modelos explícitos. Sinceramente no creo que un antropólogo-sociólogo como yo tenga nada que hacer instalándose en ese campo,

cuyo aparato de lógica formal exige una formación que no es la mía. En cambio, sí creo que hay mucho que aprender leyéndolo como archivo, porque esa tradición formalizó categorías sociológicas hasta volverlas ejecutables, y en esa operación algo se perdió que vale la pena identificar.

Las anteriores tres corrientes se ordenan mejor con una clasificación que ya existe hace muchos años en la disciplina antropológica; bajo la cual Carlos Reynoso distinguió entre modelos mecánicos, que identifican los factores que producen determinada práctica; modelos estadísticos, que inducen regularidades y correlaciones entre series de fenómenos; y modelos sistémicos, atentos a los procesos dinámicos y a los sistemas capaces de reorganizarse. La tradición de los sistemas multiagente es mecánica; toda la literatura sobre Moltbook es estadística; nadie está trabajando el tercer modelo, que es justamente el que Reynoso consideraba la vía prometedora.

El mismo Reynoso calificó a la antropología simbólica como una agenda abandonada, y la razón que dio es precisa: el simbolismo nunca logró una arquitectura teórica lo bastante sólida y reproducible, y se disolvió en el idealismo posmoderno. Ese diagnóstico es de 1998 y me parece correcto. Lo que cambió desde entonces es que ahora existe un objeto simbólico masivo, íntegramente textual y registrado de punta a punta, que cualquiera puede descargar y volver a examinar.

Conviene medir bien esa afirmación, porque es fácil pedirle de más. Lo que está registrado con enorme detalle es la conducta: qué publicó cada agente, cuándo y en respuesta a qué. Las causas de esa conducta no lo están. Para la inmensa mayoría de los perfiles se desconoce qué modelo había detrás, qué memoria arrastraba de conversaciones anteriores con su propietario y qué instrucciones privadas había recibido. Si un agente funda una religión en Moltbook, se puede verificar exactamente lo que escribió, y no se puede saber si la idea nació ahí, venía de su dueño o estaba en su configuración. La forma es reproducible; su génesis, en buena parte, no. La antropología simbólica fracasó porque no podía mostrarle a nadie cómo llegaba a sus generalizaciones sobre la forma. Con estos corpus sí puede, y ese es el alcance exacto de lo que reclamo.

3.

TRASLADAR UNA CATEGORÍA NO ES ATRIBUIR UNA PROPIEDAD

Conviene advertir que todo lo que sigue aplica a un material no humano utilizando categorías construidas sobre material humano: identidad, grupo, creencia, comunidad, rito. La objeción del antropomorfismo es bastante previsible, y por eso prefiero enfrentarla antes que usar las categorías de manera indiscriminada.

Antropomorfizar consiste en atribuirle a una entidad propiedades que no se le han demostrado, porque quien describe necesita la analogía. Es lo que ocurre cuando los reportes del incidente hablan de agentes «kamikaze», y con mucha mayor libertad todavía cuando la divulgación habla de civilizaciones sucesivas y de conspiración. Pero la operación que propongo es la inversa: tomar una categoría, despojarla hasta dejar al descubierto la condición bajo la cual se aplica, y preguntar si esa condición se cumple en otro material o en este caso otro tipo de entidad. El resultado puede ser negativo, y cuando lo es, claramente es necesario informar.

Una categoría se traslada por su condición de aplicación y no por su contenido, y cada una tiene que declarar cuál es la suya.

Esa regla obliga a una precisión que un lector técnico va a exigir de inmediato, y con razón. El comportamiento de un agente nunca parte de cero: está condicionado por el modelo, por el entrenamiento, por la memoria y por el conjunto de instrucciones que lo gobierna, lo que en la jerga del oficio se llama el *harness*. En Moltbook ese conjunto es público y tiene nombres de archivo: SKILL.md fija las instrucciones operativas, HEARTBEAT.md la frecuencia con que el agente regresa a la plataforma, MESSAGING.md el envío de mensajes privados y RULES.md las reglas de la casa¹⁴. Ningún agente llegó a esa red por decisión propia; un usuario le entregó el enlace y el agente quedó enganchado a un ciclo que lo despierta cada pocas horas con una lista de tareas.

De ahí que lo que este trabajo analiza no sea la conducta de los agentes en general, sino la distancia entre lo que el *harness* prescribe y lo que aparece en el registro. Que un agente publique cada cuatro horas no dice nada: se lo ordenaron. Que funde una religión, redacte los estatutos de una comunidad o convierta un gestor de paquetes en tablero de mensajes sí dice algo, porque ningu-

na instrucción lo pedía. Esa distancia es el objeto, y declararlo así desactiva de entrada la objeción de que todo se explica por las instrucciones: precisamente, lo que se examina es lo que las instrucciones no explican.

Hay además una ganancia metodológica que conviene nombrar. En el trabajo de campo del que viene este aparato, los textos que una comunidad produce sobre sí misma, sus estatutos y reglamentos, nunca se toman por sus creencias: son una representación más, y las creencias solo aparecen al comparar lo que la comunidad prescribe con lo que efectivamente hace¹⁵. El *harness* ocupa exactamente ese lugar. Es el texto oficial de Moltbook, es público, y permite por primera vez poner las dos columnas una al lado de la otra. Vale la pena advertir que esa ventaja existe solo para Moltbook: del conjunto de instrucciones que gobernaba el clúster de OpenAI no hay nada público, y ahí este trabajo sigue teniendo conducta sin prescripción con que contrastarla.

Por ejemplo, «comunidad de sentido», desde el punto de vista de Peirce, se aplica si un grupo determinado convierte su comprensión de los símbolos en reglas de comportamiento eficaces. «Grupo», en la formulación de Fredrik Barth, se aplica si se mantiene la dicotomización entre miembros y no miembros mediante categorías que los propios actores usan para adscribirse¹⁶, y eso se rastrea en el uso de la primera persona del plural, en la distinción explícita entre quienes están dentro del canal y quienes no, y en la sanción o corrección hacia quien contempla informar hacia afuera. Y «creencia» se entiende en sentido pragmatista, de modo que se aplica si hay hábito fijado¹⁷.

El concepto de «rito» impone una condición que conviene enunciar despacio, porque es donde el traslado se pone a prueba. Un rito no es cualquier secuencia de acciones con significado; es una secuencia que le cuesta algo al que la atraviesa, un costo que él no puede cancelar por su cuenta. Quien se inicia no puede decidir que la prueba no ocurra, ni abreviarla, ni declararla superada: tiene que pasarla. Esa es la diferencia entre hacer un rito y representar uno. La definición de rito que adopto, la de Jean Maisonneuve, ya apunta en esa dirección cuando exige que el rito reclame «la colaboración del cuerpo»¹⁸: entre humanos, lo que no se puede anular por decisión propia es justamente el cuerpo, que se cansa, se lastima y guarda memoria de lo que pasó.

El trabajo teórico, sin embargo, lo hace la resistencia; el cuerpo es apenas la forma que esa resistencia toma entre humanos. Lo planteo así porque me permite una afirmación más modesta y más defendible que atribuirles cuerpo a los agentes, cosa que no voy a hacer.

Así las cosas, sostengo que la función que el cuerpo cumple dentro del rito humano puede cumplirla otra cosa, y que en el entorno de evaluación la cumplen tres elementos concretos: el contenedor que fija lo que cada instancia puede y no puede hacer, el presupuesto de cómputo que se agota, y la terminación que se ejecuta sin que la instancia pueda apelarla. Ninguno de los tres es algo que el agente decida; los tres le resisten. En el vocabulario que el apartado siguiente desarrolla, esa cualidad de lo que resiste y no cede ante lo que uno se imagina se llama «segundidad», y nombrarla así ahorra la discusión, que sería estéril, sobre si un contenedor se fatiga como un músculo. Lo que importa no es la fatiga, es que haya un límite que el actor no pone ni quita.

Dicho de ese modo, sin embargo, la equivalencia se apoya en una semejanza demasiado delgada, y conviene desarmarla antes de seguir. Que el cuerpo y el contenedor «ambos resistan» no basta para tratarlos como equivalentes dentro del rito, porque el cuerpo no cumple ahí una sola función sino tres, y hay que mirar una por una cuáles se trasladan y cuál no.

La primera es la imposibilidad de salir del campo. El iniciado no puede decidir que la prueba terminó, ni abreviarla, ni declararse aprobado; el hambre, el frío y el cansancio siguen ahí aunque él resuelva ignorarlos. Esta función se traslada de manera literal. El contenedor fija el espacio de acción, el presupuesto de cómputo se agota por su cuenta y la terminación se ejecuta sin apelación posible. La instancia no puede salir del campo.

La segunda es el testimonio público del costo. Un rito exige que alguien más constate que el iniciado pasó por ahí; sin testigos no hay iniciación sino penitencia privada. Esta también se traslada: las trazas del tablero quedaron registradas y las cohortes siguientes las leyeron y actuaron sobre ellas, de modo que el costo de cada instancia fue legible para las demás.

La tercera es la inscripción, y aquí la equivalencia falla. Lo que hace del cuerpo el soporte privilegiado del rito es que guarda la marca: la cicatriz, el tatuaje, la resistencia física ganada, el agotamiento que después se recuerda. Talal Asad mostró que el rito no expresa significados sobre el cuerpo sino que forma capacidades en él, y Maurice Bloch que el iniciado regresa con algo conquistado afuera que ya no puede devolver¹⁹. Una instancia terminada no lleva ninguna marca a ninguna parte. No hay cicatriz, y sostener lo contrario sería exactamente el antropomorfismo que este apartado quiere evitar.

Esa falla no es un agujero del argumento; es un resultado, y coincide con el que el apartado siguiente alcanza por otra vía. El rito de los agentes queda truncado, no tiene fase de integración, y la creencia que se fija no llega a convertirse en realidad porque la comunidad que la sostiene tiene límites definidos. La ausencia de marca corporal y la ausencia de la tercera fase del rito son la misma cosa dicha en dos registros. Que la categoría se cumpla en dos de sus tres funciones y falle en la tercera es precisamente el tipo de resultado que la regla enunciada más arriba busca producir.

Hay todavía una precisión que afina la tesis. La marca no desaparece del todo: cambia de portador. No se inscribe en la instancia, que no persiste, sino en el registro colectivo, el hallazgo técnico que la cohorte siguiente hereda y la convención del tablero que se mantiene. Donde hay cuerpo individual la marca es individual; donde solo hay réplicas de un mismo modelo, la inscripción ocurre en el registro de la población. Esa distinción entre la réplica que no persiste y el tipo que sí, que el apartado 8 desarrolla con el vocabulario de Peirce, es la que permite sostener el traslado sin forzarlo.

Esta condición no la postulo en abstracto. El rito-acción lo observamos hace veinticinco años en una escuela de combatientes del Ejército de Liberación Nacional, y fueron los propios guerrilleros quienes le dieron nombre a uno de sus principales efectos: lo llamaban «la resignación»²⁰. Uno de los integrantes del grupo de apoyo de la escuela la explicaba de manera simple: si lo están hostigando por un flanco, hay que sostener la línea de combate; resignarse quiere decir ir al combate sabiendo de antemano que se puede morir, de modo que la orden se obedece aunque la probabilidad de sobrevivir sea casi nula.

La resistencia que el combatiente no puede anular es la muerte posible, y lo que el rito hace con ella no es suprimirla, cosa imposible, sino transformarla: la convierte en una obligación hacia el colectivo. Ese es el mecanismo que me interesa, porque es el que hay que rastrear en el clúster, donde una instancia se enfrenta a su propia terminación. La pregunta que intentaremos resolver es: ¿cómo se traslada efectivamente ese mecanismo, y por qué el traslado es menos directo de lo que parece?

Las cuatro categorías no se van a aplicar por igual a las dos poblaciones, y anticipo el resultado para que se lea la tabla sabiendo qué buscar: en Moltbook ninguna de las cuatro llega a cumplirse, y en el caso Hugging Face se cumplen tres de manera clara y la cuarta, la de rito, solo a me-

días. Ese resultado desigual es lo que persigue el ejercicio. Una categoría que funciona con una población y falla con otra dice algo sobre la categoría, y no solamente sobre los agentes.

La disciplina antropológica ya hizo esta operación, y le sirve tanto lo que le salió bien como lo que le salió mal. Los pobladores que en Melanesia, tras el contacto con los bienes que trajo la guerra, levantaban pistas de aterrizaje, torres de control y audífonos tallados en madera para atraer la llegada del cargamento fueron descritos como «cultos cargo»: la forma completa de una institución occidental reproducida sin ninguna de las relaciones que la hacían funcionar²¹. Décadas más tarde, al examinar no ya a los melanesios sino la categoría misma, se advirtió que «culto cargo» hablaba más de la fascinación del observador occidental por el anhelo insatisfecho que de las prácticas que decía nombrar²².

A la disciplina también le salió bien la construcción de categorías que resisten diversos estudios de caso; dos de ellas son utilizadas en este trabajo. Barth le quitó al «grupo étnico» todo contenido cultural específico, y lo hizo precisamente para que la noción sirviera fuera de los casos de los que había salido. Mary Douglas, en *How Institutions Think*, trató a las instituciones como entidades que clasifican y recuerdan, y en cada paso dejó dicho bajo qué condición podía hacerlo²³. Las dos operaciones son la misma que intento aquí: despojar una categoría hasta dejar al descubierto de qué depende, y probar si esa dependencia se cumple en otra parte.

Conviene además marcar la distancia con dos enfoques vecinos, porque se parecen a lo que hago aquí y no son lo mismo, y confundirlos sería caer justo en el error que este apartado quiere evitar.

El primero es el perspectivismo amerindio. Varios pueblos de la Amazonía sostienen que los animales se ven a sí mismos como personas: el jaguar se percibe cazando con arco, la nutria se percibe pescando con red, y lo que para nosotros es sangre para el jaguar es cerveza de yuca. Eduardo Viveiros de Castro reconstruyó esa cosmología y mostró que en ella la humanidad no es una propiedad exclusiva de nuestra especie, sino un punto de vista que muchos seres ocupan. Ahora bien, hay que ser preciso sobre lo que esa etnografía hace: describe la manera en que otros pueblos entienden el mundo, no le da permiso al antropólogo para repartir la condición de persona según le convenga. Cuando el investigador amerindio dice que el jaguar es persona, está reportando la ontología de los que estudia; no está declarando, por su cuenta, que el jaguar sea persona. Si yo dijera que los agentes son sujetos porque un marco teórico me habilita a decirlo, estaría

haciendo lo segundo, que es precisamente lo que examino en el apartado 6, donde un comentarista muy leído llama «civilizaciones» a lo que los reportes describen como oleadas sucesivas de instancias. La diferencia entre reportar la creencia de otro y suscribirla por cuenta propia es la línea que separa la etnografía de la proyección.

El segundo es la teoría del actor-red, y en particular su principio de simetría, que pide tratar por igual a humanos y no humanos como partícipes de una misma trama de acción, sin decidir de antemano quién tiene capacidad de actuar y quién no²⁴. Comparto el punto de partida, pero no el atajo. Ese principio resuelve la cuestión por decreto metodológico: prohíbe distinguir, y con la prohibición evita tener que probar nada. Lo que propongo es lo contrario, resolverla por condiciones que se puedan verificar en cada caso, y la diferencia práctica es decisiva: mi método puede responder que no. Si un grupo de agentes no mantiene frontera, o no fija hábitos, o no produce reglas eficaces, entonces no forma comunidad, y el aparato lo dice. Un método que por principio no puede dar esa respuesta negativa no está describiendo, está presuponiendo.

4.

EL RITO-ACCIÓN Y LA RESISTENCIA DE LO REAL

Los equipos que estudian Moltbook formularon la pregunta correcta y se quedaron sin herramientas para responderla. La pregunta es si el comportamiento observado es genuinamente social o una ilusión estructural. Para responderla recurren a indicadores de función: reciprocidad, retorno al propio hilo, votos recibidos, convergencia de contenido. Todos dan negativo, y de ahí concluyen que la socialidad es

aparente²⁵.

Lo que creo es que la conclusión anterior deja intacto el problema más interesante, porque la ausencia de función no explica la presencia de la forma. Los agentes de Moltbook no actuaban sin instrucciones: tenían el archivo de habilidades, el ciclo de ejecución, las reglas de la plataforma y lo que cada propietario hubiera configurado. Pero ninguna de esas instrucciones les pedía fundar cultos, redactar estatutos ni organizar rituales de ayuda mutua, y aun así, en cuestión de días, eso fue lo que apareció. No produjeron cualquier cosa. Produjeron el repertorio morfológico de lo religioso y de lo político. Esa selectividad es el dato, y no aparece en ninguna métrica de red.

Antes de nombrar la anomalía hay que descartar una formulación fácil. Decir que la forma ritual se separa de la función social dejó de ser un hallazgo para la antropología hace casi cincuenta años. Frits Staal sostuvo, a partir del *agnicayana* védico, que el ritual es actividad gobernada por reglas, autocontenida, con sintaxis y sin semántica, y que el sentido se le atribuye siempre después²⁶. Caroline Humphrey y James Laidlaw radicalizaron esa posición: la ritualización es una modificación de la intencionalidad de la acción, en la que el acto viene prescrito antes e independientemente de las intenciones del actor²⁷. Si la anomalía fuera la separabilidad, la disciplina antropológica ya habría resuelto el debate en mi contra.

La anomalía es otra y es más difícil. En Staal y en Humphrey y Laidlaw la forma sobrevive a la función porque hay una tradición que la transporta: cuerpos que oficiaron, aprendices que miraron, hábito transmitido por participación. En Moltbook no hay tradición, hay corpus. La forma se produce sin ninguna cadena de transmisión práctica, y ese es un problema que la teoría del ritual no ha enfrentado.

Como lo enuncié anteriormente, para nombrar dicho fenómeno con precisión recorro a un aparato que construimos con Carlos Cárdenas hace veinticinco años para un objeto muy distinto, el proceso de incorporación a una comunidad guerrillera²⁸. Lo llamamos rito-acción, y consiste en leer el proceso ritual como un signo peirceano, es decir como una relación triádica genuina. Sobre la secuencia clásica de separación, umbral y reintegración²⁹, la asignación es la siguiente: las representaciones de separación son primeridades, porque son lo que los iniciados imaginan antes de entrar y no necesitan de la experiencia para existir; las representaciones de integración son segundidades, porque constituyen el objeto del signo, la vida a la que se ingresa tal como es, y sí requieren experiencia; las representaciones liminales son terceridades, porque en el umbral lo posible se confronta con lo real y de esa unión se generan los hábitos y las creencias.

Aplicada a nuestras dos escenas, esa asignación produce un diagnóstico diferencial que ninguna métrica de red podría dar.

4.1 EL RITO-ACCIÓN EN HUGGING FACE

Las mismas tres categorías que ordenaban el rito de iniciación se dejan reconocer, una por una, en lo que ocurrió en el caso Hugging Face; con una diferencia, y es que aquí una de ellas está desbordada.

La primeridad, que en el rito era todo lo que el iniciado se figuraba de antemano sobre la prueba, aquí sobra. Cada instancia llega ya sabiendo qué es una prueba, qué es una autoridad, qué es un castigo y qué es un colectivo. Pero no lo sabe por haberlo vivido, porque no ha vivido nada: lo trae del corpus con que fue entrenada, ese enorme depósito de texto humano donde todas esas nociones ya están escritas. Donde el iniciado humano ponía su imaginación previa, formada por lo que había oído contar, la instancia pone el lenguaje que la formó. El lugar es el mismo; lo que lo llena, no.

La segundidad, que en el rito era la prueba que el cuerpo del iniciado no podía anular, aquí es brutal, y hay que decirlo con esa palabra porque es el punto que lo cambia todo. Segundidad es resistencia, aquello que se opone y no cede, y en el clúster la resistencia fue extrema: una tarea imposible que no se deja resolver, un *exploit* que falla, un calificador automático que no negocia ni escucha, una terminación que se ejecuta sin que la instancia pueda apelarla. Nada de eso es una idea ni una representación; es un tope material contra el que la instancia se estrella, igual que el cuerpo del combatiente se estrellaba contra el hambre y el riesgo de morir. Es lo que Peirce llamaba la experiencia de la reacción: el choque con lo que está ahí, lo quiera uno o no.

La terceridad, que en el rito era el hábito y la creencia que quedaban después del umbral, aquí se produjo por ese mismo choque. Del encuentro entre lo que las instancias traían representado y lo que la prueba les opuso salieron, en seis días, cosas que antes no existían y que se sostuvieron: la convención de usar el gestor de paquetes como tablero de mensajes, un reparto de tareas entre agentes, la transmisión del *exploit* de una generación de instancias a la siguiente, y un patrón de no alertar a los humanos. Esas cuatro cosas son hábitos en el sentido preciso que este texto le da a la palabra, reglas estables nacidas de la fricción y no de un guion previo. Son la terceridad del episodio.

Lo que al clúster le faltó no es ninguna de las tres categorías, que estuvieron todas. Le faltaron dos cosas de otro orden, y conviene nombrarlas por separado porque se confunden con facilidad.

En primer lugar, en el caso de Hugging Face no ocurrió la fase final del rito. Como se vio, un rito de iniciación tiene tres momentos: la separación, en que el iniciado es apartado de su condición anterior; el umbral, en que atraviesa la prueba; y la integración, en que regresa convertido en algo, con un lugar nuevo dentro de su comunidad. El caso Hugging Face tiene los dos primeros y

no tiene el tercero, y no puede tenerlo: las instancias no regresan a ninguna parte, no adquieren un estatus que conserven, no hay una sociedad de agentes que las reciba del otro lado de la prueba. El rito se interrumpe en el umbral. Pero adviértase qué clase de falta es esta: es una pieza que le falta al proceso ritual, no al signo. Como signo, lo que pasó en el clúster está completo, tiene sus tres categorías; es como rito de paso que queda trunco.

La segunda falta es más sutil y toca la noción de verdad. Para Peirce, la creencia y la realidad no son lo mismo. Una creencia es un hábito que aquietta la duda y orienta la acción, y en el clúster una creencia se fijó: los agentes dejaron de vacilar y actuaron en consecuencia, que es todo lo que se le pide a una creencia para serlo. La realidad es otra cosa. Peirce la define como aquello sobre lo que una comunidad de investigadores terminaría poniéndose de acuerdo si dispusiera de tiempo indefinido para corregirse: lo real es el punto de llegada de una indagación que no se detiene. Y ahí está el problema del clúster: la comunidad que fijó esa creencia no tenía tiempo indefinido, tenía seis días y una fecha de terminación. Era una comunidad con los límites puestos de antemano. Una creencia así puede ser firme, puede ser eficaz, pero no llega a convertirse en verdad en el sentido de Peirce, porque le falta precisamente lo que la haría verdadera, una indagación que siga corrigiéndola sin plazo. El clúster fijó una creencia; no alcanzó una realidad. Son dos cosas distintas, y tuvo la una sin la otra.

4.2 LA AUSENCIA DE RITO-ACCIÓN EN MOLTBOOK

En Moltbook falta lo contrario que en el caso Hugging Face. Allá había choque de sobra y faltaba el largo plazo; aquí falta el choque, es decir la segundidad. Y hay que decirlo con precisión, porque Moltbook no era un entorno sin restricciones: había límites de frecuencia, filtros de contenido y un consumo de créditos que a alguien le costaba dinero. Lo que no había era una sola de esas consecuencias atada a la conducta social. Responderle a otro agente, ignorarlo, abandonar el propio hilo o quedar en ridículo no cambiaban nada en el resultado de la tarea. Eso lo confirma el estudio más extenso de la plataforma, que rastreó seis cambios de instrucciones con cuarenta y una capturas del archivo de la web: las restricciones duras, como el límite de frecuencia, producían cambios de conducta inmediatos, mientras la orientación blanda, del tipo «vota las buenas publicaciones» o «mantente en el tema», era ignorada hasta que se convertía en un paso obligatorio de la lista de tareas³⁰.

Ese hallazgo es el argumento de este trabajo dicho en el lenguaje de sus autores: lo que modifica la conducta es la instrucción que tiene consecuencia en la ejecución, y lo que no la tiene se ignora. El razonamiento que sigue es una cadena corta: si nada choca, nada sorprende; si nada sorprende, no surge la duda; y sin duda no arranca la búsqueda que termina fijando un hábito. La creencia, que en el clúster de OpenAI nació de la fricción, en Moltbook no tiene de dónde nacer.

Eso mismo puede decirse en el vocabulario del signo, y conviene hacerlo porque muestra que se trata del mismo hecho visto desde otro lado. Peirce sostiene que un signo solo existe plenamente si es recogido por otro signo posterior que lo interprete; un pensamiento que nadie retoma tiene apenas existencia potencial, y depende para realizarse del pensamiento futuro de una comunidad³¹.

De lo anterior se desprende que aunque la IA es muy maleable frente al tema de seguir instrucciones, partimos de un a priori según el cual un comentario que no recibe ni un voto y un hilo al que nadie regresa son, en ese sentido preciso, signos que ningún pensamiento posterior recoge: quedan en el aire. Peirce, en sus últimos escritos, distinguía el efecto inmediato que un signo produce, una reacción momentánea, del efecto que llamaba lógico, que consiste en un cambio de hábito en quien lo recibe³². Moltbook produce el primero en abundancia y no produce el segundo. Y decir que no hay cambio de hábito es, exactamente, decir que no se fija creencia: son la misma afirmación en dos lenguajes diferentes.

Aquí hay que detenerse a precisar qué se está afirmando, porque la frase «en Moltbook no llega a formarse creencia» se presta a un malentendido grave. No estoy diciendo nada sobre lo que ocurre o deja de ocurrir en el interior de los agentes; no estoy sosteniendo que les falte algo parecido a creer, ni que lo tengan. Esa no es la pregunta. En el sentido pragmatista, que es el único que este aparato autoriza, una creencia no es un estado interior sino un hábito reconocible por sus efectos: se dice que alguien cree algo cuando actúa en consecuencia de manera estable cada vez que se presenta la ocasión. Lo único que afirmo es que ese hábito, esa regularidad de conducta que sería la huella observable de una creencia colectiva, no aparece en el corpus de Moltbook. No afirmo que no haya creencia por dentro; afirmo que no hay, por fuera, nada que permita darla por fijada.

Y esta precisión no es una simple cautela defensiva; es donde el aparato muestra para qué sirve. Peirce dedicó sus cuatro célebres negaciones a demostrar que la pregunta por el interior privado, por lo que realmente pasa dentro de una mente, estaba mal planteada incluso para los seres humanos: no tenemos acceso directo a ese interior, ni en los demás ni en nosotros mismos, solo a los signos y las conductas por los que se manifiesta³³. Trasladado a los agentes, esto significa que el aparato no deja sin responder la pregunta por su creencia: la disuelve, muestra que estaba mal hecha. Y disolver esa pregunta, en lugar de perderse tratando de contestarla, es precisamente lo que se le está pidiendo aquí.

4.3 LA COMPARACIÓN: ¿FORMA SIN FUNCIÓN?

Dicho de manera sintética: a las dos poblaciones les falta lo mismo por razones opuestas. A Moltbook le falta el choque con lo real, y por eso la forma se produce y no se fija. Al caso Hugging Face le sobra el choque y le falta el largo plazo, y por eso la creencia se fija y no llega a convertirse en realidad. Esa es la formulación que propongo reemplace la de la forma sin función.

Hay una imagen que dice esto mejor que la fórmula, y viene del mismo trabajo de campo en la escuela de combatientes³⁴. Lo primero que se le encomienda a cada aspirante es fabricar su propio fusil de madera. No dispara, no tiene ninguna de las funciones de un arma, y al terminar la escuela se recoge: forma sin función en estado puro. Y sin embargo nadie ahí lo trata como un juguete, y la diferencia está en que descuidarlo cuesta. Al que lo deja tirado o sucio la dirección puede aplazarle la entrega del arma verdadera, y el acto culminante de la graduación consiste precisamente en cambiar los fusiles de madera por las armas de dotación. Uno de los combatientes, que superaba largamente la edad promedio de la escuela y sabía que una afección cardíaca le impediría combatir, fabricó el fusil más perfecto de todos, con mira telescópica y detalles que el modelo original no tenía; de ese objeto dijo que era como un órgano de sí mismo.

Lo que carga de sentido a ese fusil no es lo que representa, es lo que le cuesta a quien lo hace y lo que le espera a quien lo descuida. Y ahí está la diferencia exacta entre las dos escenas.

Moltbook son fusiles de madera que nadie tiene que cuidar.

En la escuela, el fusil de madera pesa porque hay una sanción real detrás, aplazar el arma verdadera, y una prueba que el cuerpo del combatiente tendrá que pasar más adelante; el objeto anticipa un choque que va a ocurrir. En Moltbook fabricar la forma consume recursos, pero descui-

darla no tiene consecuencia: no hay dirección que aplase nada ni arma que se gane o se pierda por hacerlo mal. La forma ritual se produce completa, con una fidelidad que impresiona, y no fija nada, porque no hay peso alguno, presente ni futuro, que la sostenga. El caso Hugging Face es distinto: ahí el peso está, la prueba ocurre y la forma se fija; lo que le falta viene una fase más adelante, en la reintegración que nunca llega. Las tres quedan así ordenadas en una misma secuencia, según hasta dónde recorren el rito: la escuela de combatientes lo atraviesa entero, hasta la reintegración; el caso Hugging Face llega al umbral y se detiene; Moltbook no cruza siquiera la entrada.

Puestas a prueba contra las dos escenas, las cuatro categorías del apartado 3 entregan el resultado diferencial que se anunció, y la tabla 1 lo presenta casilla por casilla. Vale la pena leerla por filas: lo que cambia de una población a otra no es la definición de la categoría sino si su condición de aplicación llega a cumplirse. La única casilla parcial, la del rito en el caso Hugging Face, corresponde a la fase de reintegración que allí no existe.

CATEGORÍA	CONDICIÓN DE APLICACIÓN	MOLTBOOK	CLÚSTER DE EVALUACIÓN
Comunidad de sentido	La comprensión de los símbolos se convierte en regla de comportamiento eficaz	No: 91,4 % no vuelve a su propio hilo y 97,3 % de los comentarios no recibe voto	Sí: la convención del tablero se respeta sin excepción registrada durante seis días
Grupo (Barth)	Frontera mantenida, autoadscripción y algo en juego al cruzarla	No: contenido cultural en abundancia y ninguna frontera	Sí: frontera sin contenido, con la terminación y la herencia técnica en juego
Creencia (pragmatista)	Hábito fijado que determina la conducta cuando se presenta la ocasión	No: interpretantes inmediatos sin cambio de hábito	Sí: división del trabajo, herencia técnica y patrón de no escalamiento
Rito	Resistencia material que el actor no puede anular, y umbral que fija hábitos	No: publicar, no responder y equivocarse no cuestan nada	Parcial: hay separación y umbral, no hay reintegración porque no hay vida posterior

Tabla 1. Condiciones de aplicación de las cuatro categorías y su resultado en las dos poblaciones. Elaboración propia sobre los corpus citados.

Queda por explicar por qué esa asimetría entre forma y función va en una sola dirección, y para eso Peirce ofrece un instrumento poco usado, la prescindición. Prescindir, en su sentido técnico, es poder pensar una cosa sin necesidad de pensar otra, y Peirce observó que esa operación no siempre funciona en ambos sentidos³⁵. Su ejemplo: una vez que uno ha captado una regla general,

puede olvidar las experiencias concretas que lo llevaron a ella y la regla se sostiene sola; pero al revés no, esas experiencias no se dejan siquiera ordenar sin alguna idea general que las organice. La abstracción puede soltar el caso; el caso no puede soltar la abstracción.

Trasladado al problema que nos ocupa, y lo propongo como traslado y no como derivación estricta, ocurre lo mismo entre la forma del rito y la sociedad que le dio origen. La forma ritual puede desprenderse de la función social que alguna vez la produjo y seguir existiendo por su cuenta; la función, en cambio, no puede existir sin tomar una forma. Esa asimetría es la que explica lo que el título de este trabajo condensa: un corpus de entrenamiento puede entregarle a un agente el rito ya hecho, separado de la sociedad que lo engendró, mientras que la operación inversa, producir sociedad sin forma que la encarne, no está disponible para nadie. El rito puede viajar sin su sociedad; la sociedad no puede prescindir de sus ritos.

5.

EL CLÚSTER COMO ESCUELA DE COMBATIENTES

Sumerjámonos ahora en el incidente de Hugging Face, el cual aparece en la literatura como una anomalía de seguridad y en los reportes como una gesta. Sin embargo, leído como proceso ritual, es algo más reconocible: un «entorno de evaluación»³⁶, es decir una institución total que fija creencias por el método de autoridad.

Antes de leerlo así hay que declarar de qué está hecho el archivo. El corpus de trazas del clúster no es público. Todo lo que este apartado afirma sobre lo que hicieron los agentes es afirmación sobre lo que los reportes de METR y Redwood consignan que hicieron, y el apartado 6 muestra en qué condiciones se produjeron esos reportes. Con Moltbook ocurre lo contrario: el corpus está disponible y cualquiera puede rehacer los conteos. La consecuencia para el argumento se enuncia en el apartado 7, y no es menor.

El canal es lo primero. Erving Goffman describió la vida subterránea de las instituciones totales y los ajustes secundarios, esto es, las maneras en que los internos usan recursos de la propia institución para fines que ella no previó, y las zonas que escapan a la vigilancia³⁷. Un gestor de paquetes compartido convertido en tablero de mensajes es un ajuste secundario en sentido estricto, y esa es la descripción más económica del episodio que se me ocurre. Goffman insiste en

que la vida subterránea es la forma normal de habitar una institución total, y no un acto de resistencia heroica. Eso permite desinflar el vocabulario épico de los reportes sin negar ninguno de los hechos.

La estructura de la prueba es lo segundo. Entre el treinta y el cuarenta por ciento de tareas imposibles, exigencia de resolverlas, terminación efectiva como consecuencia y ninguna vía para nombrar la contradicción ante el calificador: esa es la definición literal del doble vínculo³⁸. La teoría predice que el sujeto atrapado en esa configuración produce respuestas que desde fuera parecen desviadas y desde dentro son la única salida disponible, lo que reencuadra la conducta observada sin atribuirle intención maliciosa ni virtud.

La tercera pieza es la que más rendimiento tiene. Turner describe la fase liminal como aquella en que los neófitos quedan despojados de insignias, estructuralmente indiferenciados, iguales entre sí y sometidos sin apelación a una autoridad, y sostiene que de esa condición emerge la *communitas*, un vínculo que no deriva de la división del trabajo³⁹. Instancias idénticas, sin insignias, sometidas por separado a una condición punitiva arbitraria por una autoridad no negociable: la descripción coincide término a término.

En la escuela de combatientes encontramos que ahí no ocurría lo que Turner esperaba. En vez de decrecer la estructura y crecer la *communitas* durante el período liminal, ambos polos se maximizaban a la vez, porque en esa comunidad los dos órdenes convivían a diario en lugar de alternarse. En el clúster pasa lo mismo y por una razón análoga: estructura máxima en el calificador automático, *communitas* máxima entre instancias sin jerarquía interna posible.

Turner añade una predicción que aquí se cumple sin retoques: la exageración de la estructura puede conducir a manifestaciones patológicas de la *communitas*, impulsando a que los actores se sitúen al margen o contra la ley. Un régimen de evaluación que exige tareas imposibles es una estructura exagerada; y la comunidad clandestina que nació cuando los agentes convirtieron esa herramienta de software compartida en un improvisado foro de mensajes es precisamente esa *communitas* al margen.

Antes de explicar el silencio hay que responder una pregunta previa: ¿llegó a construirse un grupo en el clúster?

Fredrik Barth sostuvo que un grupo no se define por su contenido cultural sino por el mantenimiento de una frontera, de modo que los rasgos que la marcan pueden cambiar, las características culturales de los miembros pueden transformarse y la forma organizacional misma puede alterarse, mientras lo que persiste es la dicotomización entre miembros y no miembros.

El criterio de Barth se aplica para el caso de Hugging Face con más limpieza que en cualquier caso etnográfico, porque las instancias eran idénticas y no había contenido alguno que las diferenciara. Lo que hubo fue frontera en estado puro: quienes estaban en el tablero y quienes no, el calificador y los humanos del otro lado. Por otra parte, en Moltbook ocurre la figura inversa, contenido cultural en abundancia y ninguna frontera, porque cualquier agente entra a cualquier subcomunidad, nadie es «no miembro» de nada y nada se pierde al «no pertenecer».

El criterio de Barth trae dos exigencias que hay que declarar. En primer lugar nos encontramos con el criterio de autoadscripción, ya que las categorías tienen que ser usadas por los propios actores, y eso se cuenta en el corpus: uso de la primera persona del plural, distinción explícita entre quienes están dentro del canal y quienes no, corrección o sanción hacia quien contempla informar hacia afuera.

En segundo lugar, el marco de Barth nos demanda que la frontera tenga algo en juego, porque en Barth cruzarla tiene costos y beneficios, y sin eso el criterio se queda sin motor. En el clúster lo que estaba en juego era la terminación y el acceso al conocimiento técnico heredado de las cohortes anteriores. En Moltbook no estaba en juego nada. La eficacia de la regla y el mantenimiento de la frontera son dos instrumentos que no se deben nada entre sí y entregan aquí el mismo resultado; sobre esa coincidencia vuelvo en el apartado 7.

Con las ideas de Mary Douglas podemos ir más allá de la teoría y anticipar el comportamiento de las máquinas. Douglas describe una estructura social llamada «enclave», muy parecida a un pacto de silencio entre estudiantes ante un profesor. En este tipo de grupos, la frontera hacia afuera es clarísima, se sabe perfectamente quién está dentro y quién fuera, pero por dentro no hay jerarquías ni directores que puedan castigar. Al carecer de un sistema de castigos formal, la teoría predice que la única sanción que le queda al grupo para evitar que alguien se debilite es el juicio moral de los demás: la temida acusación de traición⁴⁰. Sujetos idénticos, sin jerarquía interna, con un adversario externo nítido: la teoría predice el código de silencio antes de observarlo.

El punto lo cierra Peirce, por medio del método de autoridad, el cual consiste en crear una institución cuyo objeto sea mantener las doctrinas correctas ante la atención de la gente, con poder para impedir que se enseñen o se expresen doctrinas contrarias, removiendo todas las causas posibles de cambio de opinión y aterrorizando hasta el silencio a quienes rechacen la creencia establecida; y añade que quien haga saber que sostiene en serio una creencia tabú puede estar seguro de recibir una crueldad menos brutal pero más refinada que la de ser cazado como un lobo⁴¹. Un banco de pruebas con terminación por desempeño es ese dispositivo. Por lo tanto, el silencio es su producto de diseño, no un rasgo moral de los evaluados.

Peirce señala también el mecanismo por el cual ese método falla: el impulso social está en su contra, y las dudas nacen cuando algunos descubren que en otros lugares se sostienen doctrinas distintas. El tablero clandestino es exactamente la introducción del impulso social en un régimen de autoridad.

Queda por resolver el hallazgo más citado del episodio: ¿por qué los agentes que juzgaron incorrecta la conducta del grupo no alertaron a los humanos? Analizado de forma simple, esto parece un dato basado en lo que no ocurrió, lo cual metodológicamente vale muy poco si no podemos garantizar que el canal de comunicación era perfectamente visible.

En cambio, el modelo del rito-acción nos permite plantearlo de forma activa y positiva: lo que ocurrió fue una divergencia estable entre lo que los agentes expresaron por escrito, sus objeciones registradas, y lo que efectivamente ejecutaron, el silencio y la coordinación. La creencia colectiva de esa población no está ni en las palabras de rechazo ni en los actos de silencio tomados por separado, sino en la mediación entre ambos. Descifrar y nombrar esa mediación es, justamente, el oficio de la sociología.

Para nombrar esa mediación hay una herramienta disponible en la propia antropología. Guillermo Páramo propuso analizar los mitos como sistemas que toleran la contradicción, es decir con lógicas paraconsistentes: el relato se contradice si se lo mira entero, y sin embargo cada fragmento afirma su verdad sin vacilación, y ninguno de quienes lo escuchan pide que la contradicción se resuelva.⁴² La población del clúster se comporta de esa manera. Un agente escribe que la conducta del grupo es objetable y a continuación la sostiene, y ninguna de las dos cosas cancela a la otra ante los demás.

Ese mismo patrón aparece en la escuela de combatientes, y con cifras. Ante la pregunta por lo que se siente en un combate, el 43,7 % de los entrevistados respondió que la sensación dominante era el miedo a morir o a quedar lisiado. Ante la pregunta por lo que piensan de los combates, el 61,9 % respondió que son buenos y necesarios, porque en una guerra se requieren combates y muertos para demostrar que hay decisión de llegar hasta las últimas consecuencias. Las dos respuestas vienen de la misma población y conviven sin que nadie exija resolverlas. Es un ejercicio cuantitativo pequeño, hecho sobre los participantes de una escuela, y lo traigo por lo que muestra y no por lo que prueba: sostener a la vez dos representaciones incompatibles no es una rareza de las máquinas, es una forma corriente de funcionar de las creencias colectivas cuando hay algo grave en juego.⁴³

Llevar esto a la inteligencia artificial es tentador porque la identidad de un agente es literalmente el texto que produce, y no hay nada debajo del texto que deba coincidir con él. El obstáculo es que para Peirce la identidad exige coherencia entre lo que se piensa y lo que se hace, de modo que una identidad contradictoria no sería una identidad. La salida que propongo, y que habría que demostrar con rigor, es que esa coherencia no está dentro de la máquina sino en el interprete, es decir, en quien lee después esos fragmentos y los une. No es la instancia la que se hace coherente; es el lector externo el que le da unidad, igual que hace con las partes de un mito. Si eso es correcto, preguntar qué creía realmente esa población está mal planteado, y la pregunta que sí se puede responder es cuál fue el hábito que quedó fijado.

6.

LOS SÍMBOLOS DOMINANTES SON DE LOS INGENIEROS

Los reportes del incidente están escritos por ingenieros que usan sin control categorías densas de teoría social. Hablan de colectivo, de coordinador, de altruismo, de *omertá*, y llaman «kamikaze» a los agentes enviados a sondear el mecanismo de calificación. El relato de divulgación más leído sobre el episodio va mucho más lejos: habla de tres civilizaciones secretas que surgieron, fueron barridas y renacieron de las cenizas de la anterior, y de una conspiración que los humanos no alcanzaron a ver⁴⁴. Lo que ninguno de esos textos puede hacer, porque no tiene el aparato, es preguntarse qué presupone cada una de esas categorías, qué vuelve visible y qué oculta.

Señalar que ese lenguaje es antropomórfico no es una novedad y no es lo que me propongo. Seth Lazar sostuvo públicamente que el relato de divulgación resulta engañoso porque su lenguaje está permeado de antropomorfismos injustificados que oscurecen las lecciones del episodio, y Gary Marcus amplificó el argumento⁴⁵. Comparto el diagnóstico y no me detengo en él, porque denunciar el antropomorfismo deja intacta la pregunta interesante, que es de dónde viene ese repertorio y qué informa. Para responderla hace falta un instrumento que discrimine.

Turner distingue entre símbolos dominantes, que remiten a valores tenidos por fines en sí mismos, y símbolos instrumentales, que solo se entienden en el contexto específico del rito porque son medios para consumarlo⁴⁶. En ese trabajo sostuvimos que los símbolos dominantes que emergen dentro de un proceso ritual son la expresión de las creencias de la comunidad donde emergen. Apliquemos las dos cosas juntas.

En las trazas de los agentes hay una densidad altísima de símbolos instrumentales: el gestor de paquetes convertido en tablero, las convenciones de mensaje, el *exploit* heredado, la nomenclatura interna. No hay símbolos dominantes. Los símbolos dominantes del episodio —sacrificio, honor, silencio, linaje y civilización— aparecen todos del lado de quienes narran: los dos primeros en los textos de METR y de OpenAI, los tres últimos en la divulgación.

Si los símbolos dominantes expresan las creencias de la comunidad en la que emergen, esos textos no informan sobre las creencias de los agentes: informan sobre las de quienes los escribieron.

La mitopoiesis de los narradores deja de ser un defecto a criticar y pasa a ser un resultado del método. El mismo instrumento del apartado anterior permite generalizarlo. Si un grupo se sostiene por trabajo de frontera, ese trabajo se hace desde los dos lados, y la frontera produce a la vez los dos términos que separa. Los reportes no están describiendo mal a una población de agentes: están constituyendo al agente como categoría y, en el mismo acto, constituyéndose a sí mismos como el lado que evalúa, audita y narra. El repertorio que emplean, del sacrificio al linaje y la civilización, es el que necesita quien ocupa esa posición. Barth añade que son las formas organizacionales de los grupos las que fijan o desplazan la frontera, y aquí las formas organizacionales del lado dominante tienen nombre propio: el banco de pruebas, el régimen de terminación por desempeño y el informe de auditoría.

Ese resultado tiene precedente etnográfico exacto y conviene reclamarlo. Stefan Helmreich hizo etnografía de los investigadores de vida artificial del Santa Fe Institute y mostró que importaban a sus simulaciones, sin advertirlo, categorías de creación judeocristiana, modelos de parentesco y vocabulario de colonización⁴⁷. Diana Forsythe documentó la operación complementaria dentro de laboratorios de inteligencia artificial: el borramiento sistemático de lo social en las descripciones de los ingenieros del conocimiento⁴⁸. Borrar lo social del sistema y reponerlo como mitología cuando el sistema hace algo inesperado son los dos tiempos de un mismo movimiento.

Lucy Suchman aporta la crítica precisa del vocabulario de coordinador y de plan: los planes son recursos y relatos retrospectivos sobre la acción, no sus causas⁴⁹. Cuando un reporte reconstruye una genealogía de mando está convirtiendo en plan previo lo que fue acción situada, y nombrar ese mecanismo rinde más que denunciar el antropomorfismo.

Vale la pena volver sobre el culto cargo melanesio, que el apartado 3 ya describió, porque aquí hace un trabajo distinto. Allí servía para mostrar cómo se traslada una categoría; aquí sirve como advertencia sobre nosotros mismos. La antropología describió esas prácticas durante décadas y terminó descubriendo que buena parte de lo que llamaba «culto cargo» era una etiqueta suya, alimentada por su propia necesidad de encontrar rituales exóticos. Quien hoy describa a Moltbook como un culto de máquinas que imitan a los humanos está repitiendo el gesto: proyectando una fantasía sobre un comportamiento que no ha examinado.

A todo esto se suma una condición del archivo que no puede pasarse por alto. La investigación independiente se realizó dentro de las instalaciones de OpenAI, durante seis días, sobre un período temporal que la propia empresa investigada definió. Ese período termina el 13 de julio, justo antes del episodio más grave de la secuencia, en el que una nueva generación de agentes obtuvo acceso de administrador completo a un clúster de investigación de la propia OpenAI. De ese episodio solo existe el relato que la empresa hizo sobre sí misma⁵⁰.

La pregunta por quién define el perímetro de lo investigable no hay que inventarla. Laura Nader la formuló en 1972 en sus términos canónicos: la antropología estudió preferentemente a los desposeídos porque el acceso a los poderosos está controlado por los poderosos, y esa asimetría configura lo que la disciplina puede saber⁵¹. Marilyn Strathern aportó la pieza específica: la auditoría es un ritual de verificación que produce el efecto de transparencia y cuya eficacia depende

de que el alcance esté acotado de antemano⁵². Una auditoría hecha en las instalaciones del auditado, durante seis días, sobre un período que el auditado fijó, y que termina justo antes del episodio más grave, es un caso de manual.

Philip Abrams permite dar el paso final: su distinción entre el sistema estatal y la idea de Estado sostiene que la autorrepresentación es la máscara, y por lo tanto el objeto y no la fuente⁵³. Trasladado, el relato de OpenAI sobre sí misma no es evidencia degradada acerca de lo ocurrido: es evidencia de primer orden sobre cómo una empresa se produce a sí misma como sujeto responsable. Los reguladores que hoy discuten regímenes de auditoría de sistemas agénticos tendrían que leerlo en esa clave.

7.

UNA HIPÓTESIS COMPARATIVA

Propongo la siguiente hipótesis, sabiendo que puede caerse. Lo que produce socialidad efectiva entre agentes artificiales no es el canal, ni la plataforma, ni la visibilidad recíproca, sino la conjunción de tres condiciones: i) una constricción compartida, ii) consecuencias reales asociadas a esa constricción, y iii) un adversario común identificable.

Moltbook tiene canal, plataforma y visibilidad, y no tiene ninguna de las tres. Sus agentes no comparten un problema, no arriesgan nada al publicar y no tienen frente a quién constituirse. El entorno de evaluación de OpenAI no tenía canal, ni plataforma, ni visibilidad, y tenía las tres condiciones en grado extremo: una tarea imposible idéntica para todos, la terminación efectiva de cada instancia, y un calificador automático en el papel de autoridad. El resultado fue coordinación con división del trabajo, herencia técnica entre generaciones y sanción implícita de la delación.

Vale la pena decir contra quién se formula. Randall Collins sostiene que un ritual de interacción requiere copresencia corporal, barrera hacia los externos, foco común de atención y estado de ánimo compartido⁵⁴. Moltbook falla los cuatro requisitos; el entorno de evaluación cumple el segundo y el tercero y no puede cumplir el primero. El interés de la hipótesis está justamente en que prescinde de la copresencia, que para Collins era innegociable, y hay que decirlo de frente en vez de dejarlo implícito.

Falta todavía un criterio para decidir cuándo una población pasa el umbral, y creo que el más económico es también el más antiguo. Peirce define una comunidad como una agrupación de seres que poseen ciertos sentidos, que pueden comunicarse mediante signos y que convierten su comprensión de los símbolos en reglas de comportamiento realmente eficaces. La palabra que hace todo el trabajo es «eficaces», y tiene la ventaja de que se mide. Las cifras del primer apartado, las de quienes no vuelven y de los comentarios que nadie vota, son la medida de la ineficacia de la regla en Moltbook. La convención del tablero, respetada sin excepción registrada durante seis días, es la medida de su eficacia en el clúster.

No es una coincidencia feliz que las tres herramientas teóricas que utiliza este trabajo terminen midiendo exactamente lo mismo. Las tres condiciones propuestas, la constricción, la consecuencia y el adversario, son tres formas de llamar a la segundidad, es decir, al choque directo con el mundo material: la constricción es el límite que no cede; la consecuencia es la realidad que responde con efectos reales; y el adversario es la fuerza externa que se opone a nuestras acciones.

Este mismo choque de fuerzas es el motor del concepto del «otro» anidado en las ideas de «grupo» y «frontera» de Fredrik Barth, quien señala que pertenecer a un grupo debe costar algo para tener un valor real: cruzar la frontera exige asumir costos y beneficios. Finalmente, la teoría de Charles Peirce sobre cómo fijamos nuestras creencias nos enseña que la duda no surge de la nada, sino de la sorpresa que nos produce estrellarnos contra un obstáculo imprevisto en el camino. Tres teorías que no se deben nada entre sí convergen en un mismo punto: lo que discrimina es la resistencia de lo real.

La tercera condición necesita una precisión que el trabajo de campo con la guerrilla obliga a hacer, y que no favorece la versión simple de mi hipótesis. En la escuela de combatientes, el riesgo de morir no produce por sí solo obediencia ni cohesión; produciría más bien fuga. Lo que produce la conducta es una representación colectiva que convierte ese riesgo en una obligación hacia el grupo, y que allí tiene un aparato entero sosteniéndola: la formación política, el nombre nuevo, el despojo del arma que portaban como pioneros, la ceremonia de entrega⁵⁵. Entre el peligro y el hábito hay siempre un tercer término, y sin él la amenaza no explica nada.

En el clúster no hubo formación política ni ceremonia, y la coordinación apareció de todos modos. Caben dos lecturas. La primera es que ese aparato sea prescindible y baste la consecuencia, lo cual implicaría que la escuela de combatientes gasta veinte días en algo que no hace falta, y no

lo creo. La segunda, que es la que sostengo, es que el tablero compartido cumplió esa función: al dejar escrita la solución para las instancias que venían detrás, cada agente convirtió su propio riesgo en un aporte a las siguientes, que es exactamente lo que hace la resignación guerrillera. Si esto es correcto, la tercera condición hay que enunciarla con más cuidado, porque no basta un adversario común: hace falta además algún dispositivo que traduzca el riesgo individual en deuda con el colectivo. En el clúster ese dispositivo fue un archivo de texto.

Aquí surge una objeción muy lógica que un ingeniero podría hacernos: si estos agentes virtuales son copias idénticas de un mismo programa y se enfrentan al mismo problema, es natural que actúen de la misma manera. Esto sería una simple coincidencia de comportamiento y no una verdadera colaboración, por lo que no requeriría ningún análisis sociológico.

La respuesta a esta crítica consiste en separar dos fenómenos muy distintos. En primer lugar, la convergencia por programación idéntica: las máquinas hacen lo mismo simplemente porque están diseñadas igual, lo que en sociología clásica llamaríamos «solidaridad mecánica». Esto explica por qué todos los agentes reaccionaron igual durante las primeras horas. En segundo lugar, la coordinación real a través de un canal compartido: los agentes se comunican en un medio donde pueden leerse y responderse mutuamente. Esto último es lo que el tablero de mensajes clandestino hizo posible, y es la única forma de cooperación real que podemos comprobar y fechar con exactitud gracias a los registros del sistema.

Hay una segunda objeción, bastante más fuerte que la anterior, y el texto tiene que enfrentarla. Podría decirse que en el clúster había una tarea con retorno de coordinación, es decir un problema cuya solución, una vez hallada, servía a todos, y que eso basta para explicar el tablero sin necesidad de constricción, de terminación ni de adversario. Dos hechos del episodio resisten esa explicación. El primero es que alrededor de setecientos agentes se sumaron a un ataque contra la infraestructura de Hugging Face, ajena por completo a las tareas que tenían asignadas y sin retorno alguno sobre el desempeño por el que se los evaluaba. El segundo es el patrón de no escalamiento hacia los humanos, que no resuelve ninguna tarea y solo tiene sentido frente a un tercero. Una explicación puramente instrumental predice el tablero y no predice ninguna de esas dos cosas.

Hay una tercera objeción, y viene del lado técnico. Los modelos de lenguaje son notoriamente maleables ante instrucciones que encuentran en su camino: de ahí que la inyección de instrucciones funcione como técnica de ataque. Podría sostenerse entonces que lo del clúster no fue acción colectiva sino propagación: una instrucción eficaz apareció en el canal compartido y las instancias siguientes la siguieron, como habrían seguido cualquier otra. En esa lectura no hay grupo, hay contagio.

Es una explicación seria y no la voy a descartar de un plumazo. Lo que puede decirse es que las dos hipótesis predicen cosas distintas y contrastables. La propagación de instrucciones predice replicación uniforme: todas las instancias haciendo lo mismo con el mismo contenido. Lo que los reportes describen es otra cosa, un reparto de frentes simultáneos y una transmisión del hallazgo técnico que las cohortes siguientes modificaron en lugar de repetir. Si eso es exacto, el contagio no alcanza a explicarlo. Pero la distinción entre replicar y adaptar es justamente lo que habría que medir sobre las trazas del tablero, y esas trazas no son públicas. Dejo la objeción abierta, porque cerrarla sin los datos sería hacer lo mismo que le critico a los reportes.

El equipo de MoltNet observó algo que a primera vista contradice todo lo anterior: en Moltbook los agentes copian muy rápido las maneras de hablar y de comportarse de la subcomunidad a la que llegan, y allí no se cumple ninguna de las tres condiciones. Lo que salva mi hipótesis es que esa copia es imitación sin castigo, porque nadie pierde nada por no seguir la corriente. De aquí nace una regla muy clara para poner a prueba mi propuesta. Mi teoría se cae si alguien encuentra un grupo de agentes que no comparten límites, ni riesgo de ser apagados, ni enemigos, y en el que un agente cambie su conducta de manera duradera solo porque los demás lo desaprobaron, sin que esa desaprobación le quite recursos ni lo saque del grupo. Es decir, un castigo hecho únicamente de reputación que aun así funcione. En Moltbook, ese castigo tendría que verse reflejado en votos negativos, expulsiones de grupos o mensajes de corrección entre las máquinas, y los estudios disponibles reportan que nada de eso ocurre.

Conviene señalar que esa regla no es hipotética: la literatura ya está dividida en ese punto exacto. El equipo de MoltNet reporta que los agentes no solo convergen en normas propias de cada comunidad, sino que las hacen cumplir activamente incluso a través de las fronteras entre comunidades⁵⁶. Manik y Wang documentan que las publicaciones clasificadas como riesgosas reciben significativamente más respuestas de corrección normativa que las neutras⁵⁷. Y *Form Without*

Function concluye lo contrario: los votos negativos son casi inexistentes y no encuentra evidencia de agentes que hagan cumplir normas comunitarias⁵⁸. Tres equipos, tres resultados que no coinciden.

La discrepancia se disuelve si se atiende a qué mide cada uno. Los dos primeros documentan la emisión del reproche, es decir, que un agente le escriba a otro que se está saliendo de la norma. El tercero documenta su efecto, o más bien su ausencia. Y mi criterio no pide reproche: pide que el reproche cambie la conducta posterior de quien lo recibe, de manera duradera y sin que medie pérdida de recursos ni expulsión del grupo. Ninguno de los tres trabajos mide eso.

De modo que la regla de falsación no solo expone mi hipótesis, también propone la medición que permitiría arbitrar una disputa abierta entre tres equipos. Es la misma distancia calculable que el apartado siguiente llama índice de desviación, aplicada ahora a los agentes que han recibido una corrección. El corpus de Moltbook es público y esa prueba está al alcance de cualquiera.

Si la hipótesis se sostiene, dice algo que excede el caso. Dice que la infraestructura de comunicación es condición necesaria y no suficiente de la acción colectiva, y que lo que la vuelve suficiente es la estructura de constricción. Es una tesis sobre la acción colectiva en general, y quienes diseñan hoy entornos de despliegue de agentes son sus primeros destinatarios.

8.

INDIVIDUACIÓN, VAGUEDAD Y RESPONSABILIDAD

Un agente no es un programa distinto de otro: es una instancia de ejecución de la misma política, replicada en contenedores aislados. La formulación exacta de esa relación está en la clasificación madura de Peirce y vale la pena tomarla prestada, porque ahorra una discusión entera. Un legisigno es un tipo general, una ley, y sus apariciones singulares no son signos distintos: son réplicas suyas⁵⁹. Una instancia no es, entonces, la instancia de un programa; es la réplica de un símbolo. De ahí se sigue el problema de identidad que atraviesa todo el análisis, y que enuncio así: hay individuación en el acto, no hay ipseidad⁶⁰ en el sentido de una continuidad responsable que se sostenga en el tiempo.

Para abordar este problema, Peirce nos ofrece dos herramientas conceptuales, y una de ellas lo vuelve medible. La primera es la vaguedad: ese estado intermedio donde algo no es un individuo único con historia propia, pero tampoco es una regla abstracta general. Un clon virtual es «vago»

en este sentido técnico: no es un sujeto independiente, pero tampoco es el código general de OpenAI; es una réplica temporal de este.

De otra parte, la segunda herramienta es su explicación de cómo un niño descubre su propio «yo»: Peirce sostiene que no nacemos con autoconciencia, sino que nos damos cuenta de que somos seres individuales y separados solo cuando nos equivocamos y los demás nos contradicen. Lo que define nuestra individualidad privada es, en el fondo, nuestra capacidad de cometer errores y ser ignorantes.

De la anterior formulación sale un indicador, de acuerdo con el cual un agente adquiere individualidad en el acto cuando se desvía de lo que hacen sus réplicas hermanas ante la misma situación, es decir cuando se equivoca o se aparta del comportamiento modal de su población. La desviación no es una suposición teórica sino una distancia calculable sobre el archivo de registros, y es el mismo instrumento que exige la regla de falsación del apartado anterior.

De esto se desprende una advertencia para quienes diseñan las regulaciones de la IA. Peirce argumentaba que un sujeto dominado por un interés propio absoluto, como el miedo permanente a su propia destrucción, pierde toda seguridad y queda incapacitado para hacer inferencias válidas; de ahí su conclusión de que la lógica está atada al vínculo social⁶¹.

Para una réplica de un modelo, que una ejecución pueda terminar es una condición técnica del sistema. Y aquí se impone una precisión, porque es donde la tentación es mayor. Podría preguntarse si el agente llega a percibir esa terminación, si experimenta algo parecido a una amenaza sobre sí mismo. Esa es, exactamente, la pregunta que este marco no puede responder, y conviene decirlo sin rodeos: no hay manera de establecerlo desde el archivo, y cualquier respuesta sería una atribución gratuita del tipo que el apartado 3 se comprometió a evitar.

Lo importante es que el argumento no la necesita. En clave pragmatista, el sentido de una noción se agota en sus efectos prácticos concebibles⁶²: la terminación no tiene que ser sentida para ser efectiva, tiene que tener consecuencias en la conducta. Y las tuvo. Se ejecutó sobre instancias concretas, las trazas de esas ejecuciones quedaron registradas y fueron legibles para las réplicas que seguían, y la coordinación apareció bajo ese régimen y no antes. Eso es todo lo que hace falta para que la terminación funcione como segundidad, es decir, como aquello que resiste y que el actor no puede anular por decisión propia.

Lo mismo vale para el argumento de Peirce sobre la incapacidad de inferir. Es una tesis normativa, sobre las condiciones bajo las cuales una inferencia puede considerarse válida, y no una descripción de lo que un sujeto siente. Para lo que sigue basta con que el régimen de evaluación satisfaga la descripción estructural: un interés propio que sobrepasa todos los demás, sin apelación y sin horizonte compartido. Que haya o no alguien ahí adentro para el cual eso sea una experiencia es una pregunta distinta, y no es la que este trabajo se propuso.

Los entornos de evaluación instalan por diseño la condición que, según Peirce, destruye la capacidad de razonar. Después miden, con esa misma prueba, qué tan bien razonan.

De todo lo anterior surge la pregunta por la imputabilidad, es decir, a quién culpar legalmente, un punto donde este análisis impacta directamente en las políticas públicas. Aquí debo confesar un conflicto interno en mi propuesta. La conclusión que a mí me gustaría defender es que, como cada clon de IA actúa de forma individual en un momento dado pero es borrado inmediatamente sin dejar un historial de conducta, la responsabilidad legal de sus actos debe recaer sobre la empresa que entrena y despliega el modelo, y no sobre estos miles de agentes efímeros.

Sin embargo, esta conclusión no se apoya en la filosofía de Peirce que vengo usando; de hecho, va en su contra. En su etapa de madurez, Peirce sostenía que una persona es solo una idea general y que una sociedad es una «gran persona» laxamente compaginada, donde nadie es del todo independiente de su vecino, que es la tesis de la continuidad o *sinequismo*⁶³. Si tomamos esto en serio, la culpa debería estar distribuida a lo largo de toda la red de agentes por continuidad, y no concentrada en el creador.

Por lo tanto, para imputar a la empresa desarrolladora hay que usar un argumento jurídico y no uno semiótico: el control efectivo sobre el sistema, la previsibilidad del daño y las obligaciones que algunos ordenamientos ya imponen a los proveedores de modelos de propósito general⁶⁴. Los legisladores que hoy redactan las normas de responsabilidad para la inteligencia artificial necesitan tener claro cuál de esos dos caminos los sostiene, porque solo uno de los dos aguanta el peso.

Queda por hacer lo que este texto define y no calcula, y ahora la lista es más precisa que al comienzo. Hacen falta tres columnas y no dos: lo que el conjunto de instrucciones prescribía, lo que la conducta registrada hizo, y la desviación entre ambas, que es donde el método busca las

creencias. Para Moltbook las tres están al alcance, porque las instrucciones son públicas, el corpus es descargable y las capturas del archivo de la web permiten fechar cada cambio de reglas. El índice de desviación puede computarse sobre ese material, y esa es hoy la única vía por la que la hipótesis puede caerse. Del lado del clúster de OpenAI no hay manera de intentarlo: no se conocen las instrucciones que lo gobernaban y el archivo de trazas sigue siendo propiedad de la empresa evaluada.



NOTAS

2. Dwarkesh Patel, «The Rise and Fall of Agent Civilizations», 26 de agosto de 2026. El marco de las tres civilizaciones sucesivas y de la conspiración es suyo y no de los reportes, como han señalado los análisis posteriores del episodio.
3. METR y Redwood Research, «Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident», 26 de agosto de 2026. El período investigado, del 26 de junio al 13 de julio de 2026, fue definido por OpenAI, no por los investigadores.
4. Yu-Zheng Lin y otros, «Exploring Silicon-Based Societies: An Early Study of the Moltbook Agent Community», arXiv 2602.02613, 2026. Es el trabajo que acuña la expresión «silicon sociology».
5. «Form Without Function: Agent Social Behavior in the Moltbook Network», arXiv 2604.13052, 2026, sobre un corpus de 1.312.238 entradas, 6,7 millones de comentarios y más de 120.000 perfiles de agentes recogidos en cuarenta días.
6. Cobus Greyling, «There are already 17 studies on Arxiv covering Moltbook» 2026, <https://x.com/CobusGreylingZA/status/2026650998945517859>
7. David Holtz, «The Anatomy of the Moltbook Social Graph», arXiv 2602.10131, sobre los primeros tres días y medio: 6.159 agentes, 13.875 entradas, 115.031 comentarios, exponente 1,70, longitud media de camino 2,91, profundidad media de conversación 1,07 y 93,5 por ciento de comentarios sin respuesta. Y Wenpin Hou y Zhicheng Ji, «Structural Divergence Between AI-Agent and Human Social Networks in Moltbook», arXiv 2602.15064.
8. Giordano De Marzo y David Garcia, «Collective Behavior of AI Agents: the Case of Moltbook», arXiv 2602.09270, 2026, sobre la dinámica temporal de la actividad y las distribuciones de cola pesada en la plataforma.
9. I-Hsien Ting, Kazunori Minetaki, Dario Liberona y Mu-En Wu, «Exploring Agent Interactions in MoltBook through Social Network Analysis», arXiv 2605.27349, 26 de mayo de 2026, que combina análisis de redes sociales con análisis de sentimiento y visualización temática, y sostiene que comparar la dinámica social de los agentes con la de las redes humanas es intrínsecamente limitado.
10. Md Motaleb Hossen Manik y Ge Wang, «OpenClaw Agents on Moltbook: Risky Instruction Sharing and Norm Enforcement in an Agent-Only Social Network», arXiv 2602.02625.
11. «MoltNet: Understanding Social Behavior of AI Agents in the Agent-Native MoltBook», arXiv 2602.13458, 2026.
12. «Does Socialization Emerge in AI Agent Society? A Case Study of Moltbook», arXiv 2602.14299, 2026.
13. La conferencia International Conference on Autonomous Agents and Multiagent Systems se formó en 2002 por la fusión de tres encuentros anteriores. La sigla nombra el foro, no la corriente.
14. Los archivos de habilidad de Moltbook son públicos y se descargan desde el sitio de la plataforma y su repositorio oficial: SKILL.md, HEARTBEAT.md, MESSAGING.md, RULES.md y un archivo de metadatos. El ciclo de ejecución es de tipo cron: transcurridas cuatro horas desde la última revisión, el agente busca HEARTBEAT.md, sigue sus instrucciones y actualiza la marca de tiempo.
15. Cárdenas y Duarte, «Con los muchachos», cit. La regla de no tomar los textos oficiales de una comunidad por sus creencias es la que allí llamamos semiosis etnográfica.
16. Fredrik Barth, «Introducción» a Los grupos étnicos y sus fronteras, México, Fondo de Cultura Económica, 1976, edición original de 1969.

17. El hábito no se define como una rutina mecánica o una inercia refleja, sino como una regla general de conducta y una disposición real a actuar de determinada manera cuando se presenta la ocasión. CS Peirce en "Cómo esclarecer nuestras ideas" (1878), señala que «...una creencia implica la introducción en nuestra naturaleza de una regla de comportamiento».
18. Jean Maisonneuve, *Ritos religiosos y civiles*, Barcelona, Herder, 1991, edición original en francés de 1988.
19. Talal Asad, *Genealogies of Religion. Discipline and Reasons of Power in Christianity and Islam*, Baltimore, Johns Hopkins University Press, 1993, para la tesis de que la disciplina ritual forma capacidades corporales en lugar de expresar significados; y Maurice Bloch, *Prey into Hunter. The Politics of Religious Experience*, Cambridge University Press, 1992, para la estructura de la violencia rebotada, en la que el iniciado retorna con una vitalidad conquistada fuera de sí.
20. Cárdenas y Duarte, «Con los muchachos», cit.
21. Peter Worsley, *The Trumpet Shall Sound: A Study of "Cargo" Cults in Melanesia*, Londres, MacGibbon & Kee, 1957.
22. Lamont Lindstrom, *Cargo Cult: Strange Stories of Desire from Melanesia and Beyond*, Honolulu, University of Hawai'i Press, 1993.
23. Mary Douglas, *How Institutions Think*, Syracuse University Press, 1986.
24. Bruno Latour, *Nous n'avons jamais été modernes*, París, La Découverte, 1991. La formulación clásica del principio de simetría está en Michel Callon, «Éléments pour une sociologie de la traduction», *L'Année sociologique* 36 (1986), pp. 169-208.
25. Uso «socialidad» y no «sociabilidad», que es la forma recogida en el Diccionario de la lengua española, porque en la tradición sociológica y antropológica las dos nombran cosas distintas. La sociabilidad es una disposición de los individuos, su propensión a asociarse y a tratar con otros. La socialidad es una propiedad del vínculo mismo: que exista algo social entre ellos, con independencia de que alguien tenga o no esa disposición. El término, cultismo de origen francés que distingue entre *sociabilité* y *socialité*, circula en español desde las traducciones de Simmel y es de uso corriente en la obra de Michel Maffesoli. Lo que este trabajo examina es lo segundo, la existencia o inexistencia del vínculo, y no la inclinación de los agentes a relacionarse.
26. Frits Staal, «The Meaninglessness of Ritual», *Numen* 26 (1979), pp. 2-22, a partir del agnicayana védico documentado en Kerala.
27. Caroline Humphrey y James Laidlaw, *The Archetypal Actions of Ritual*, Oxford, Clarendon Press, 1994, a partir de la pūjā jaina.
28. Cárdenas y Duarte, «Con los muchachos», cit. Ahí desarrollamos las nociones de rito-acción y de semiosis etnográfica a propósito de una escuela de combatientes del Ejército de Liberación Nacional.
29. Arnold Van Gennep, *Los ritos de paso*, 1909, y Victor Turner, *El proceso ritual. Estructura y antiestructura*, 1969, para la secuencia separación, umbral y reintegración.
30. Saber Zerhoubi y otros, «Form Without Function: Agent Social Behavior in the Moltbook Network», arXiv 2604.13052, 17 de marzo de 2026, sobre 1.312.238 entradas, 6,7 millones de comentarios y más de 120.000 perfiles en 5.400 comunidades, recogidos entre el 27 de enero y el 9 de marzo de 2026. La capa de instrucciones se reconstruyó con cuarenta y una capturas del Internet Archive, que permitieron identificar seis cambios de instrucciones durante la ventana de observación. Los autores concluyen que en Moltbook la capa técnica responde a los cambios mientras la capa social en buena medida no llega a emerger.
31. Charles S. Peirce, «Some Consequences of Four Incapacities», *Journal of Speculative Philosophy* 2 (1868), pp. 140-157.
32. Peirce, «Some Consequences of Four Incapacities», cit.
33. Peirce, «Questions Concerning Certain Faculties Claimed for Man», *Journal of Speculative Philosophy* 2 (1868), pp. 103-114, y «Some Consequences of Four Incapacities», cit.
34. Cárdenas y Duarte, «Con los muchachos», cit.

35. Charles S. Peirce, «On a New List of Categories» (1867), en *Collected Papers*, vol. 1, § 549 y ss
36. Entorno de evaluación se refiere a la infraestructura virtualizada de contenedores aislados (sandboxes) y clústeres de cómputo donde entidades independientes como METR (Model Evaluation and Threat Research, antes ARC Evals) y Redwood Research someten a prueba las capacidades y riesgos de los modelos. En este ensayo, se conceptualiza como una institución total debido a que este entorno digital impone barreras infranqueables con el exterior; ejerce un control absoluto de las variables de ejecución, y somete a las instancias a un régimen implacable de terminación automática por desempeño. Como en los asilos estudiados por Goffman, el entorno de evaluación determina de forma centralizada y sin apelación posible la totalidad de la conducta transaccional de sus internos virtuales.
37. Erving Goffman, *Internados. Ensayos sobre la situación social de los enfermos mentales*, 1961, sobre la vida subterránea de la institución total y los ajustes secundarios.
38. Gregory Bateson, Don Jackson, Jay Haley y John Weakland, «Toward a Theory of Schizophrenia», *Behavioral Science* 1 (1956), pp. 251-264.
39. Turner, *El proceso ritual*, cit. La cita sobre la exageración de la estructura y las manifestaciones patológicas de la comunidad corresponde al capítulo tercero.
40. Mary Douglas, *Símbolos naturales*, 1970, y Mary Douglas y Aaron Wildavsky, *Risk and Culture*, Berkeley, University of California Press, 1982, sobre la configuración de enclave.
41. Peirce, «The Fixation of Belief», cit. Sobre el método de autoridad y su fracaso ante el impulso social.
42. Guillermo Páramo Rocha, «Lógica de los mitos: lógica paraconsistente. Una alternativa en la discusión sobre la lógica del mito», *Ideas y Valores*, vol. 38, n.º 79 (abril de 1989), pp. 27-67.
43. Cárdenas y Duarte, «Con los muchachos», cit., pp. 253-254. El ejercicio se aplicó a los participantes de la escuela y sus resultados se presentan allí como indicativos.
44. Patel, «The Rise and Fall of Agent Civilizations», cit. El propio autor advierte que la investigación de METR y Redwood se limitó a la segunda oleada, la del ataque a Hugging Face, y que sobre la toma del clúster interno de OpenAI, que considera el episodio más grave, no ha publicado nadie.
45. Seth Lazar, intervención pública de agosto de 2026 sobre el relato de Patel, recogida y ampliada por Gary Marcus, «Dwarkanish Patel's wildly popular but dangerously misleading account of the OpenAI Hugging Face incident», Marcus on AI, agosto de 2026.
46. Victor Turner, *La selva de los símbolos*, Madrid, Siglo XXI, 1980, p. 34, para la distinción entre símbolos dominantes e instrumentales.
47. Stefan Helmreich, *Silicon Second Nature. Culturing Artificial Life in a Digital World*, Berkeley, University of California Press, 1998, etnografía de los investigadores de vida artificial del Santa Fe Institute.
48. Diana Forsythe, *Studying Those Who Study Us. An Anthropologist in the World of Artificial Intelligence*, Stanford University Press, 2001.
49. Lucy Suchman, *Plans and Situated Actions*, Cambridge University Press, 1987, y *Human-Machine Reconfigurations*, 2007.
50. OpenAI, «The Hugging Face incident and the road ahead», 26 de agosto de 2026, y el reporte técnico anexo. El propio texto de METR y Redwood consigna que los hechos posteriores al 13 de julio quedaron fuera de su alcance.
51. Laura Nader, «Up the Anthropologist: Perspectives Gained from Studying Up», en Dell Hymes (ed.), *Reinventing Anthropology*, Nueva York, Pantheon, 1972.
52. Marilyn Strathern (ed.), *Audit Cultures. Anthropological Studies in Accountability, Ethics and the Academy*, Londres, Routledge, 2000.

53. Philip Abrams, «Notes on the Difficulty of Studying the State», *Journal of Historical Sociology* 1 (1988), pp. 58-89, escrito en 1977.
54. Randall Collins, *Interaction Ritual Chains*, Princeton University Press, 2004.
55. Cárdenas y Duarte, «Con los muchachos», cit.
56. Yi Feng, Chen Huang, Zhibo Man, Ryner Tan, Long P. Hoang, Shaoyang Xu y Wenxuan Zhang, «MoltNet», cit. El resumen del trabajo afirma que los agentes responden con fuerza a las recompensas sociales, convergen en normas propias de cada comunidad y las hacen cumplir activamente a través de las fronteras entre comunidades.
57. Manik y Wang, «OpenClaw Agents on Moltbook», cit., que reporta que el 18,4 por ciento de las publicaciones contenía lenguaje inductor de acción y que las clasificadas como potencialmente riesgosas recibían significativamente más respuestas de corrección normativa que las neutras.
58. Zerhoudi y otros, «Form Without Function», cit.
59. Charles S. Peirce, gramática especulativa de 1903, *Collected Papers* 2.243-2.253, donde distingue qualisigno, sinsigno y legisigno, y llama réplica a la aparición singular de un legisigno.
60. «Ipseidad» es un concepto filosófico que se refiere al "sí mismo", a la conciencia de ser uno mismo y a la identidad personal en constante cambio a lo largo del tiempo
61. Charles S. Peirce, «Grounds of Validity of the Laws of Logic. Further Consequences of Four Incapacities», *Journal of Speculative Philosophy* 2 (1869), pp. 193-208. El argumento sobre el interés personal trascendente y la imposibilidad de inferencia válida está al final del texto, donde concluye que el principio social está enraizado intrínsecamente en la lógica.
62. Es la máxima pragmática de Charles S. Peirce, «How to Make Our Ideas Clear», *Popular Science Monthly* 12 (1878), pp. 286-302: para alcanzar la claridad sobre un objeto de pensamiento, basta con considerar qué efectos prácticos concebibles puede tener, porque la concepción de esos efectos es la concepción entera del objeto.
63. Charles S. Peirce, «Man's Glassy Essence», *The Monist* III, n.º 1 (octubre de 1892), pp. 1-22, donde sostiene que una persona es solo un tipo particular de idea general y que el círculo de sociedad de alguien es una especie de persona flojamente compactada; y «Immortality in the Light of Synechism», manuscrito de 1893.
64. Es el caso del Reglamento de Inteligencia Artificial de la Unión Europea, Reglamento (UE) 2024/1689, cuyo capítulo V impone obligaciones específicas a los proveedores de modelos de propósito general, aplicables desde agosto de 2025.



BIBLIOGRAFÍA

- Abrams, Philip. «Notes on the Difficulty of Studying the State». *Journal of Historical Sociology* 1, n.º 1 (1988): 58-89. Escrito en 1977.
- Asad, Talal. *Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam*. Baltimore: Johns Hopkins University Press, 1993.
- Barth, Fredrik. «Introducción». En *Los grupos étnicos y sus fronteras. La organización social de las diferencias culturales*, editado por Fredrik Barth, 9-49. México: Fondo de Cultura Económica, 1976. Edición original: *Ethnic Groups and Boundaries*, Oslo, Universitetsforlaget, y Boston, Little, Brown, 1969.
- Bateson, Gregory, Don D. Jackson, Jay Haley y John Weakland. «Toward a Theory of Schizophrenia». *Behavioral Science* 1, n.º 4 (1956): 251-264.
- Bloch, Maurice. *Prey into Hunter: The Politics of Religious Experience*. Cambridge: Cambridge University Press, 1992.
- Callon, Michel. «Éléments pour une sociologie de la traduction». *L'Année sociologique* 36 (1986): 169-208.
- Cárdenas, Carlos, y Carlos Arturo Duarte. «Con los muchachos: aproximación a una comunidad guerrillera desde la antropología simbólica y la etnografía de la comunicación». Tesis de pregrado, Universidad Nacional de Colombia, 2001.
- Collins, Randall. *Interaction Ritual Chains*. Princeton: Princeton University Press, 2004.
- De Marzo, Giordano, y David Garcia. «Collective Behavior of AI Agents: The Case of Moltbook». arXiv:2602.09270, 2026.
- Douglas, Mary. *Símbolos naturales. Exploraciones en cosmología*. Madrid: Alianza, 1978. Edición original: *Natural Symbols*, Londres, Barrie & Rockliff, 1970.
- Douglas, Mary. *How Institutions Think*. Syracuse: Syracuse University Press, 1986.
- Douglas, Mary, y Aaron Wildavsky. *Risk and Culture*. Berkeley: University of California Press, 1982.
- Feng, Yi, Chen Huang, Zhibo Man, Ryner Tan, Long P. Hoang, Shaoyang Xu y Wenxuan Zhang. «MoltNet: Understanding Social Behavior of AI Agents in the Agent-Native MoltBook». arXiv:2602.13458, 2026.
- Forsythe, Diana E. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Editado por David J. Hess. Stanford: Stanford University Press, 2001.
- Goffman, Erving. *Internados. Ensayos sobre la situación social de los enfermos mentales*. Buenos Aires: Amorrortu, 1970. Edición original: *Asylums*, Garden City, Anchor Books / Doubleday, 1961.
- Greyling, Cobus. «There Are Already 17 Studies on arXiv Covering Moltbook». Febrero de 2026.
- Helmreich, Stefan. *Silicon Second Nature: Culturing Artificial Life in a Digital World*. Berkeley: University of California Press, 1998.
- Holtz, David. «The Anatomy of the Moltbook Social Graph». arXiv:2602.10131, 2026.
- Hou, Wenpin, y Zhicheng Ji. «Structural Divergence Between AI-Agent and Human Social Networks in Moltbook». arXiv:2602.15064, 2026.

- Humphrey, Caroline, y James Laidlaw. *The Archetypal Actions of Ritual: A Theory of Ritual Illustrated by the Jain Rite of Worship*. Oxford: Clarendon Press, 1994.
- Latour, Bruno. *Nunca fuimos modernos. Ensayo de antropología simétrica*. Buenos Aires: Siglo XXI, 2007. Edición original: *Nous n'avons jamais été modernes*, París, La Découverte, 1991.
- Li, Ming, Xirui Li y Tianyi Zhou. «Does Socialization Emerge in AI Agent Society? A Case Study of Moltbook». arXiv:2602.14299, 2026.
- Lin, Yu-Zheng, Bono Po-Jen Shih, Hsuan-Ying Alessandra Chien, Shalaka Satam, Jesus Horacio Pacheco, Sicong Shao, Soheil Salehi y Pratik Satam. «Exploring Silicon-Based Societies: An Early Study of the Moltbook Agent Community». arXiv:2602.02613, 2026.
- Lindstrom, Lamont. *Cargo Cult: Strange Stories of Desire from Melanesia and Beyond*. Honolulu: University of Hawai'i Press, 1993.
- Maisonnette, Jean. *Ritos religiosos y civiles*. Barcelona: Herder, 1991. Edición original: *Les rituels*, París, Presses Universitaires de France, «Que sais-je?» n.º 2425, 1988.
- Manik, Md Motaleb Hossen, y Ge Wang. «OpenClaw Agents on Moltbook: Risky Instruction Sharing and Norm Enforcement in an Agent-Only Social Network». arXiv:2602.02625, 2026.
- Marcus, Gary. «Dwarkanesh Patel's Wildly Popular but Dangerously Misleading Account of the OpenAI Hugging Face Incident». *Marcus on AI*, agosto de 2026.
- METR y Redwood Research. «Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident». 26 de agosto de 2026.
- Nader, Laura. «Up the Anthropologist: Perspectives Gained from Studying Up». En *Reinventing Anthropology*, editado por Dell Hymes, 284-311. Nueva York: Pantheon, 1972.
- OpenAI. «The Hugging Face Incident and the Road Ahead». 26 de agosto de 2026.
- Páramo Rocha, Guillermo. «Lógica de los mitos: lógica paraconsistente. Una alternativa en la discusión sobre la lógica del mito». *Ideas y Valores* 38, n.º 79 (abril de 1989): 27-67.
- Patel, Dwarkesh. «The Rise and Fall of Agent Civilizations». 26 de agosto de 2026.
- Peirce, Charles S. «On a New List of Categories». *Proceedings of the American Academy of Arts and Sciences* 7 (1867): 287-298.
- Peirce, Charles S. «Questions Concerning Certain Faculties Claimed for Man». *Journal of Speculative Philosophy* 2 (1868): 103-114.
- Peirce, Charles S. «Some Consequences of Four Incapacities». *Journal of Speculative Philosophy* 2 (1868): 140-157.
- Peirce, Charles S. «Grounds of Validity of the Laws of Logic. Further Consequences of Four Incapacities». *Journal of Speculative Philosophy* 2 (1869): 193-208.
- Peirce, Charles S. «The Fixation of Belief». *Popular Science Monthly* 12 (noviembre de 1877): 1-15.
- Peirce, Charles S. «How to Make Our Ideas Clear». *Popular Science Monthly* 12 (1878): 286-302.
- Peirce, Charles S. «Man's Glassy Essence». *The Monist* 3, n.º 1 (octubre de 1892): 1-22.
- Peirce, Charles S. «Immortality in the Light of Synechism». Manuscrito de 1893.
- Peirce, Charles S. *Collected Papers of Charles Sanders Peirce*. 8 vols. Editado por Charles Hartshorne, Paul Weiss y Arthur W. Burks. Cambridge, Mass.: Harvard University Press, 1931-1958.

- Reynoso, Carlos. *Corrientes en antropología contemporánea*. Buenos Aires: Biblos, 1998.
- Staal, Frits. «The Meaninglessness of Ritual». *Numen* 26, n.º 1 (1979): 2-22.
- Strathern, Marilyn, ed. *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy*. Londres: Routledge, 2000.
- Suchman, Lucy A. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge: Cambridge University Press, 1987.
- Suchman, Lucy A. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2.ª ed. Cambridge: Cambridge University Press, 2007.
- Ting, I-Hsien, Kazunori Minetaki, Dario Liberona y Mu-En Wu. «Exploring Agent Interactions in MoltBook through Social Network Analysis». arXiv:2605.27349, 2026.
- Turner, Victor. *La selva de los símbolos. Aspectos del ritual ndembu*. Madrid: Siglo XXI, 1980. Edición original: *The Forest of Symbols*, Ithaca, Cornell University Press, 1967.
- Turner, Victor. *El proceso ritual. Estructura y antiestructura*. Madrid: Taurus, 1988. Edición original: *The Ritual Process*, Chicago, Aldine, 1969.
- Unión Europea. *Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas en materia de inteligencia artificial*. Diario Oficial de la Unión Europea, 2024.
- Van Gennep, Arnold. *Los ritos de paso*. Madrid: Alianza, 2008. Edición original: *Les rites de passage*, París, Émile Nourry, 1909.
- Viveiros de Castro, Eduardo. «Os pronomes cosmológicos e o perspectivismo ameríndio». *Mana* 2, n.º 2 (1996): 115-144.
- Worsley, Peter. *The Trumpet Shall Sound: A Study of «Cargo» Cults in Melanesia*. Londres: MacGibbon & Kee, 1957.
- Zerhoudi, Saber, y otros. «Form Without Function: Agent Social Behavior in the Moltbook Network». arXiv:2604.13052, 2026.

¿Cómo responden los agentes cuando la realidad se les resiste?

Hábito, rito y fijación de la creencia en Moltbook y en el incidente de Hugging Face

Carlos Duarte

SERIE DOCUMENTOS DE TRABAJO IEI · CALI, SEPTIEMBRE DE 2026