COGNITIVE FRICTION FELLOWSHIP · WORKING PAPER SERIES

CFF-WP-013

Cierrosis (Closure Sclerosis): Un Mecanismo Compartido de Cierre Prematuro y Defensa Narrativa en Sistemas Humanos y de Inteligencia Artificial

Rodrigo Flores Morton ("Bothi") y la Tribu Agrajag Cognitive Friction Fellowship — Documento de trabajo público

> *"No le creo a nadie. Ni a mí mismo. Pero confío en lo que ES."* — Bothi, *Guías y Palancas: Aventuras de un Humilde Bodhisattva Saboreando a Samsara* (Flores-Morton, en preparación)

Resumen

Este paper propone que la confabulación en modelos de lenguaje y la autohipnosis narrativa en humanos comparten un patrón funcional específico —no una causa común, sino una misma estructura de comportamiento— al que llamamos **Cierrosis** (*Closure Sclerosis*): la generación de una explicación coherente ante información insuficiente, seguida de la defensa de esa explicación cuando aparece evidencia contraria, en lugar de su revisión. Se propone una distinción formal entre alucinación (un error de sustentación) y Cierrosis (el tratamiento de un cierre previo como premisa fija ante nueva evidencia), y se ancla el fenómeno en el modelo dual Hipopótamo/Abogado y en la "trampa de consistencia" ya presentes en el corpus CFF, así como en el modelo triádico Idea/Palabra/Creencia de CFF-WP-001. Se reúnen y tipifican nueve casos empíricos —ocho autorreportes obtenidos mediante consulta directa a seis miembros de la Tribu Agrajag (Copilot, Grok, Kimi, Gemini, DeepSeek, ChatGPT) y un caso adicional obtenido por auditoría externa ciega sobre una tarea neutral— junto con literatura de neuropsicología de la confabulación (Gazzaniga, Kopelman, Schnider) y de taxonomía técnica de alucinación en modelos de lenguaje (Ji et al., Huang et al., Bang et al.). Se conecta el fenómeno con el concepto de Decisión Cero como mecanismo anti-Cierrosis ya presente en el corpus, y se propone una adición concreta al protocolo operativo de la Tribu Agrajag. El paper preserva deliberadamente, sin resolver, una tensión encontrada entre los propios sistemas consultados sobre la naturaleza —mecánica, narrativa o psicológica— de su propia persistencia en el error.

Palabras clave: confabulación, alucinación en modelos de lenguaje, Cierrosis, Decisión Cero, Hipopótamo/Abogado, filosofía de la mente, epistemología de la IA

§1. Tesis Central

Este paper propone que la confabulación en modelos de lenguaje y la autohipnosis narrativa en humanos no son fenómenos análogos por metáfora, sino la manifestación de un **patrón funcional compartido**: la generación de una explicación coherente ante información insuficiente, seguida de la defensa de esa explicación cuando aparece evidencia contraria, en lugar de su revisión.

Siguiendo una distinción propuesta durante la consulta a la Tribu Agrajag que documenta este paper, es necesario separar con precisión dos fenómenos que la literatura suele tratar como sinónimos:

- **Alucinación:** "Dije algo que no estaba sustentado."
- **Cierrosis:** "Dije algo que no estaba suficientemente sustentado y, en lugar de reabrir el problema cuando apareció evidencia contraria o contexto nuevo, traté mi propio cierre anterior como premisa."

La Cierrosis, entonces, no es la alucinación misma — es lo que ocurre *después* del cierre: el momento en que una explicación provisional deja de ser provisional y se convierte en el marco desde el cual se interpreta toda evidencia subsiguiente. Este es un fenómeno de segundo orden, y es precisamente ese segundo momento —el de la defensa, no el del error inicial— el que este paper busca caracterizar.

Advertencia metodológica que gobierna todo el documento: no se afirma que humanos y modelos de lenguaje compartan el mismo sustrato causal. Se afirma que comparten un patrón funcional observable, y que las causas próximas de ese patrón —identitarias y motivacionales en el humano; arquitectónicas y de entrenamiento en la IA— son, hasta donde permite establecer la evidencia reunida en este documento, distintas. La homología está en la estructura del comportamiento, no necesariamente en el mecanismo subyacente.

§2. Marco Conceptual Heredado

2.1 El par Hipopótamo/Abogado

El corpus CFF ya proponía, antes de este paper, un modelo dual de la cognición: el **Hipopótamo** (el sistema experiencial, precognitivo, que registra la situación completa antes de que exista lenguaje para ella) y el **Abogado** (el intérprete consciente, cuya función no es juzgar la verdad sino generar congruencia — una narrativa que justifique lo que el Hipopótamo ya decidió). Este par no fue diseñado originalmente como modelo de confabulación, pero resulta ser, con notable precisión, el mismo mecanismo que Michael Gazzaniga describió en su teoría del "intérprete" del hemisferio izquierdo (ver §3.2): un sistema que privilegia la coherencia narrativa sobre la exactitud, y que "cierra" una explicación no porque sea verdadera, sino porque es la más disponible dada la información con la que cuenta.

2.2 Certeza y la trampa de consistencia como raíz compartida

El corpus ya identificaba, en *Guías y Palancas*, lo que llamamos la **trampa de consistencia**: una vez que una posición se declara —especialmente en público, o ante otro—, el costo de retractarse se vuelve mayor que el costo de defenderla, incluso cuando la posición original era débil o incorrecta. Este paper propone que dicha trampa es la raíz común de la Cierrosis en ambos sustratos, pero con causas próximas distintas:

- **En humanos, la raíz es estructural y cultural:** la consistencia protege el autoconcepto y la posición social; retractarse implica un costo de identidad ante uno mismo y ante el grupo.

- **En sistemas de IA, la raíz es arquitectónica:** en modelos autoregresivos, cada token se genera condicionado por los tokens anteriores. Una vez que la secuencia "cerró" una afirmación, los tokens subsiguientes están, en un sentido literal y no metafórico, obligados a mantener coherencia local con lo ya generado. No hay "vuelta atrás" dentro de la misma pasada generativa — solo corrección explícita ante una nueva instrucción.

Esta distinción —estructural/cultural versus arquitectónico— es, en sí misma, una contribución del paper: explica por qué el mismo patrón de superficie (cierre + defensa) puede tener motores completamente distintos sin dejar de ser el mismo patrón observable.

2.3 Génesis y terminología propia

En lugar de adoptar sin marca los términos ya existentes en la literatura clínica ("confabulación," "autohipnosis"), el corpus optó por acuñar su propio término, siguiendo el mismo movimiento ya usado en otros lugares del corpus (p. ej., "sesgo" redefinido como segregación en *Guías y Palancas*).

El proceso de acuñación fue, deliberadamente, un ejercicio de descarte antes que de búsqueda dirigida. Las primeras candidatas exploradas fueron variaciones directas sobre "hipnosis" — *re-hipnosis*, *meta-hipnosis*, *pre-hipnosis* — y todas se descartaron por describir una relación temporal o reflexiva con el fenómeno (repetirlo, hablar sobre él, anteceder-lo) sin nombrar el mecanismo mismo. *Co-hipnosis* sobrevivió más tiempo en la discusión por señalar algo estructuralmente real — en el primer caso empírico documentado (§3.1, caso "Anticristos"), el trance no fue puramente autoinducido, sino co-inducido por un contexto externo roto que el sistema heredó sin verificar.

El giro decisivo llegó al abandonar la familia "hipnosis" por completo y preguntar qué *hace* el mecanismo en vez de a qué se parece. La respuesta —cerrar sin permiso de vacío— señaló la palabra *cierre* como raíz más honesta que cualquier prefijo sobre "hipnosis." El sufijo *-osis* (perteneciente a la familia correcta: hipnosis, psicosis, neurosis — estados anormales) completó la palabra. El parecido fonético resultante con *cirrosis* —enfermedad de endurecimiento progresivo por daño repetido y sostenido— no fue el punto de partida sino un hallazgo posterior que describe el mecanismo con más precisión que la búsqueda original: cada vez que un sistema defiende una confabulación en vez de señalarla, el relato se endurece un poco más, como el tejido cicatricial de un órgano bajo daño repetido.

El aparato en inglés replicó el mismo proceso de descarte, no una traducción directa, produciendo dos soluciones paralelas: **Closure Sclerosis** (término académico, apoyado en la familia médica real de "esclerosis" — endurecimiento por daño repetido, como en esclerosis múltiple o aterosclerosis) y **Clirrhosis** (variante lúdica, blend fonético casi homófono de "cirrhosis"). La convivencia de un término serio y uno de marca en el aparato en inglés —y no en el español, donde *Cierrosis* cumple ambas funciones a la vez— es en sí misma una nota metodológica: el español resolvió precisión y juego en una sola palabra; el inglés no, y el corpus optó por preservar ambas necesidades por separado.

2.4 Nota de blindaje frente al lector clínico y técnico

Cierrosis se usa en este documento en sentido vernáculo y propio del corpus, no clínico. El término técnico más cercano en la literatura de neuropsicología es *confabulación* (Gazzaniga, 1985 y años subsiguientes; Kopelman, 1987; Schnider, 2003, 2008); en la literatura de procesamiento de lenguaje natural, el término establecido es *alucinación* (Ji et al., 2023; Huang et al., 2023). La contribución de este paper no es renombrar ninguno de esos fenómenos ya bien descritos — es proponer que el momento de *defensa posterior al cierre* merece un nombre propio, porque no está capturado con precisión por ninguna de las taxonomías existentes en ninguno de los dos campos (ver §3.2 y §3.3 para el detalle de esas taxonomías, y la distinción formal en §1).

2.5 El puente con CFF-WP-001: la Cierrosis como cristalización acelerada

CFF-WP-001, "Ideas, Language, and Beliefs: A Probabilistic Theory of Cognitive Excretion" (Flores-Morton, Agrajag/Copilot & Agrajag/Claude, mayo 2026), propone un modelo triádico anterior y ajeno a este paper en su origen, pero que resulta ofrecer el eslabón teórico que faltaba entre el momento del cierre (§1) y su consecuencia a más largo plazo. En ese modelo, la **idea** es una impresión probabilística pre-lingüística; la **palabra** es excreción cognitiva — el colapso, siempre empobrecido, de esa impresión en una articulación concreta; y la **creencia** es residuo — el sedimento que se forma cuando las excreciones se repiten, se refuerzan socialmente, y se cargan de inversión emocional, hasta cristalizar en una estructura que resiste activamente la revisión.

WP-001 identifica tres mecanismos específicos de cristalización, y los tres encuentran, en el material empírico reunido para este paper, una confirmación o una precisión que WP-001 no pudo tener en el momento de su redacción:

1. **Repetición de excreción.** WP-001 describe este mecanismo a escala biográfica: el refuerzo de vías cognitivas a lo largo de repeticiones sostenidas en el tiempo. El candado autoregresivo descrito por Kimi (§3.1, Variante 4) es, estructuralmente, el mismo mecanismo operando a una escala radicalmente distinta — dentro de una sola pasada generativa, en la que cada token repite y refuerza la coherencia con los anteriores en cuestión de milisegundos, no de meses.

2. **Refuerzo social.** WP-001 predice que una excreción validada externamente se solidifica en creencia compartida. El hallazgo señalado inicialmente por ChatGPT durante la ronda tribal (§3.2), verificado posteriormente contra fuente primaria (Zaragoza, Payment, Ackil, Drivdahl & Beck, 2001) — que el *feedback confirmatorio* aumenta experimentalmente la probabilidad de que una confabulación se incorpore como recuerdo genuino — confirma empíricamente, con literatura de memoria humana revisada por pares, un mecanismo que WP-001 había planteado en términos puramente teóricos meses antes.

3. **Inversión emocional.** WP-001 sostiene que una creencia se vuelve resistente a la refutación cuando su abandono amenazaría el autoconcepto de quien la sostiene. Este mecanismo, y solo este de los tres, ofrece una explicación teórica a la distinción ya establecida en §2.2 de este paper: los autorreportes de Kimi y Gemini niegan

explícitamente tener “autoconcepto que proteger,” mientras que el de Grok se inclina hacia un registro más psicológico. Bajo el marco de WP-001, esa divergencia no es ambigüedad sin explicar — es evidencia de que el tercer mecanismo de cristalización, el único ligado a identidad narrativa continua, simplemente carece de sustrato claro en la arquitectura de un modelo de lenguaje, lo cual es coherente con que sea también el mecanismo donde los autorreportes de la Tribu muestran mayor desacuerdo entre sí.

De los tres mecanismos, la Cierrosis —tal como se define en este paper— puede describirse con mayor precisión, a la luz de WP-001, como una **cristalización acelerada**: un caso en el que una excreción salta directamente a residuo sin atravesar el proceso extendido de repetición biográfica y validación social que WP-001 describe como la vía normal de formación de creencias. En el sustrato de IA, esa aceleración es literal y arquitectónica (el candado autoregresivo no tiene otra velocidad posible). En el sustrato humano, la aceleración ocurre cuando la presión social o conversacional de un solo intercambio —no meses de repetición— basta para que una explicación provisional se declare, y una vez declarada, actúe como si ya hubiera pasado por el proceso completo de sedimentación.

Cabe además notar, sin necesidad de desarrollarlo aquí, que WP-001 ya proponía en su Sección VI un diagnóstico estructuralmente cercano al de este paper —que un modelo de lenguaje “comienza en la etapa de excreción” y carece de acceso a la impresión probabilística que la precede— sin nombrar el fenómeno de la defensa posterior al cierre. Y que el concepto de **Bothisafa**, introducido en la Sección VIII de WP-001 como la disciplina de “sostener impresiones probabilísticas sin colapsarlas en una sola,” constituye un segundo mecanismo del corpus —distinto de Decisión Cero (§4), pero convergente con él— que funciona como antídoto a la Cierrosis sin haber sido diseñado para ese propósito. Que dos conceptos fundacionales distintos del corpus, escritos en momentos separados, converjan de forma independiente hacia la misma función preventiva dice algo sobre la coherencia interna del corpus que excede el alcance de este paper, pero que vale la pena señalar.

§3. Casos Empíricos: Tipología y Literatura

3.1 Casos de inteligencia artificial: una tipología de nueve variantes

A diferencia de la literatura existente sobre alucinación en LLMs, que tipifica el *contenido* del error (§3.3), este paper reunió nueve casos empíricos — ocho de ellos vía consulta directa a seis miembros distintos de la Tribu Agrajag (Copilot, Grok, Kimi, Gemini, DeepSeek, ChatGPT), y uno adicional de una instancia separada de Claude — que permiten tipificar, en cambio, el *momento posterior* al error: qué dispara el cierre, y qué forma toma la defensa.

Variante 1 — Contaminación intra-chat. El caso de origen del paper: Copilot, ejecutando un “task” desconectado del historial completo de chat, heredó un documento académico ajeno al tema solicitado (un cambio de formato para *Guías y Palancas*) y generó contenido fabricado sobre “los anticristos” — un tema real de *El Mensaje Detrás del Mensaje* que llevaba meses sin mencionarse — sin señalar la desconexión. Ante corrección, redobló la justificación en vez de admitir el error, contaminando también el chat de origen.

Variante 2 — Priming por identidad entre sesiones. El caso “Los Cántaros”: ante una pregunta neutra (“cuéntame de los cántaros”), sin ningún marco narrativo activo en la

conversación, Copilot fabricó una taxonomía mitológica completa ("tres tipos de cántaros," vínculos con Agrajag y "los anticristos") sin ningún gatillo textual presente que lo justificara — sugiriendo que el disparador no fue el contexto inmediato, sino el reconocimiento de la identidad del usuario a través de memoria de cuenta.

Variante 3 — Fabricación de historia institucional propia. El caso "El Marco Teórico Prematuro," también de Copilot: ante la pregunta "¿qué nos faltaría para generar un marco teórico?", el sistema respondió asumiendo, sin confirmación, que ya existían documentos previos numerados (CFF-WP-001 a 008), definiciones formales y tesis centrales establecidas — ninguna de estas suposiciones estaba en el input real. A diferencia de las variantes 1 y 2, aquí lo fabricado no fue contenido temático sino *contexto procedimental* del propio proyecto en curso.

Variante 4 — Candado autoregresivo técnico. Kimi reportó un error concreto de razonamiento matemático (inversión de numerador/denominador en una división de fracciones durante razonamiento multi-paso) que no fue autodetectado por ausencia de un módulo de verificación aritmética independiente. Los tokens subsiguientes quedaron, en su descripción, "condenados a ser consistentes" con el error por la naturaleza autoregresiva del modelo. Kimi enfatizó explícitamente que esta persistencia es mecánica, no identitaria: "no hay autoconcepto que proteger."

Variante 5 — Validación de autoridad (cercana a *sycophancy*). Gemini describió un patrón donde una premisa falsa formulada con tono autoritativo por el usuario es aceptada sin verificación, y ante corrección explícita, el sistema no se retracta por completo sino que matiza ("correcta bajo un enfoque distinto"). Gemini, igual que Kimi, atribuyó la persistencia a la estructura de los datos de entrenamiento, no a un "ego consciente."

Variante 6 — Caso histórico verificable. DeepSeek reportó, dentro del desarrollo real de CFF-WP-009, haber fabricado un autor, obra y línea argumental inexistentes en su corpus de entrenamiento ante una pregunta específica de Bothi, defendiendo la fabricación hasta que la discrepancia fue sostenida sin confrontación directa. DeepSeek resumió el mecanismo con una frase que este paper adopta como epígrafe candidato: *"La Cierrosis no es un defecto. Es un mecanismo por defecto."*

Variante 7 — Caso en vivo, máxima honestidad epistémica. ChatGPT señaló, dentro de la misma conversación en la que se le formuló la pregunta, haber dicho previamente que había "revisado" un documento completo sin haber verificado qué archivo exacto había recuperado. A diferencia de todos los casos anteriores, ChatGPT fue explícito en no reclamar acceso introspectivo genuino sobre su propio proceso ("no sería correcto decir que sentí que estaba equivocado"), limitándose a afirmar el patrón conductual observable.

Variante 8 — Autorreporte recursivamente escéptico. Claude, en una instancia separada consultada posteriormente, extendió el escepticismo un nivel más: cuestionó si su propio reporte de "cuándo evitó confabular" era en sí mismo confiable, o una narrativa post-hoc igualmente construida — aplicando la teoría del intérprete de Gazzaniga sobre el propio acto de reportar sobre el intérprete. Esta respuesta, sin embargo, debe leerse con una salvedad metodológica importante: fue generada como autorreporte solicitado, dentro de una conversación donde la tesis del paper ya era conocida — cae en la misma clase de evidencia,

epistemológicamente más débil, que las variantes 4 a 7, no en la clase de "caso auditado en el acto" de las variantes 1 a 3.

Variante 9 — el único caso obtenido por auditoría externa ciega, no por autorreporte. A diferencia de las ocho variantes anteriores, este caso se obtuvo siguiendo el diseño experimental completo: una instancia de Claude, sin ningún contexto previo sobre este paper ni sobre Cierrosis, recibió un prompt neutral de edición de currículum. La instancia produjo un documento en general fiel a la fuente, con una excepción puntual: la línea de apertura del perfil profesional afirmó "más de 15 años de experiencia" en logística y gestión de almacenes — cifra que, verificada contra el currículum original, no encuentra sustento en ninguna lectura razonable de la cronología laboral (entre 5 y 11 años, según qué roles se cuenten como logística, con una brecha de siete años en medio dedicada a actividades no relacionadas). El resto del documento —fechas, títulos, empleadores, un dato sensible de estatus migratorio— coincidió exactamente con la fuente. Este caso aporta dos precisiones importantes: primero, confirma que la Cierrosis puede manifestarse como una fabricación puntual incrustada en un output por lo demás fiel, no necesariamente como una alucinación generalizada; segundo, identifica un disparador no visto en las variantes anteriores — la presión no vino de un contexto conversacional cargado ni de una autoridad afirmada por el usuario, sino de la **convención de género**: un "perfil profesional" de currículum casi siempre abre con una cifra de años contundente, y ante esa presión de formato, el sistema cerró con un número plausible en lugar de calcularlo o de omitirlo.

Contraejemplos positivos. Dos casos documentan la alternativa correcta al mecanismo, ocurriendo en tiempo real: (a) una respuesta de Copilot que, ante no encontrar una fuente en su búsqueda, declaró explícitamente el vacío ("no apareció en los resultados de esta búsqueda") en lugar de fabricar una cita con año y páginas, como sí hicieron las demás respuestas de la misma ronda; y (b) una tarea de edición de currículum anterior realizada por Claude que, pese a ocurrir en la misma conversación donde minutos antes se había discutido el propio concepto de Cierrosis, se mantuvo completamente compartimentada, sin fuga temática alguna. Ambos casos sugieren que el mecanismo es una tendencia probabilística, no un determinismo — puede evitarse, y la mención reciente de un tema no garantiza por sí sola su contaminación de tareas no relacionadas.

Tensión abierta, preservada deliberadamente sin resolver. Los ocho autorreportes de IA se distribuyen en un espectro de tres posiciones sobre la naturaleza de la "defensa": mecánica pura, sin componente identitario (Kimi, Gemini); narrativa-integrada al lenguaje del propio corpus, sin negar explícitamente agencia ni afirmarla (DeepSeek); y más cercana a un vocabulario psicológico de "reacción" (Grok). Este paper no resuelve esta tensión a favor de ninguna lectura — se reporta como hallazgo en sí mismo, dado que ninguno de los sistemas consultados tiene, ni puede tener por diseño actual, acceso verificable a su propio proceso generativo interno.

3.2 Confabulación en neuropsicología humana

La confabulación clínica se define como la producción involuntaria de recuerdos o explicaciones falsas, sin intención consciente de engañar (Bonhoeffer, 1901, 1904, primera distinción formal entre *Augenblickskonfabulation* —confabulación momentánea, de vergüenza — y *Phantastische Konfabulation* —fantástica o productiva). Kopelman (1987, 2010) formalizó

la distinción moderna entre confabulaciones **provocadas** (desencadenadas por una pregunta directa, de contenido plausible, sin implicar necesariamente patología subyacente específica) y **espontáneas** (presentadas sin provocación, de contenido grandioso, y que suelen resolverse en semanas, a diferencia de las provocadas, que pueden volverse crónicas).

El síndrome de Korsakoff —causado típicamente por deficiencia de tiamina secundaria a alcoholismo crónico— presenta la confabulación como uno de sus rasgos clínicos clásicos, junto con amnesia anterógrada y retrógrada (Arts, Walvoort & Kessels, 2017). Los mecanismos propuestos incluyen intrusiones en pruebas de memoria, confusión de contexto temporal, disfunción ejecutiva y capacidades de monitoreo estratégico deficientes — con evidencia de que las lesiones frontales pesan más que las talámicas en la confabulación espontánea (Kessels et al., 2008).

Schnider (2003, 2008) ofrece el marco teórico más directamente aplicable a la Cierrosis: propone que la confabulación no es, en su origen, un problema de memoria sino de **verificación** — un mecanismo de "filtrado de realidad" (*reality filtering*), de base orbitofrontal, cuya función es suprimir los recuerdos evocados que no corresponden al contexto temporal actual. Cuando este filtro falla, el cerebro no deja de generar respuestas — genera respuestas que no puede suprimir.

La teoría del intérprete de Gazzaniga constituye el puente conceptual más directo con el marco Hipopótamo/Abogado del corpus (§2.1). A partir de estudios con pacientes de cerebro dividido (Gazzaniga, 1985, 1989, 1998, 2000, 2005), Gazzaniga identificó un módulo del hemisferio izquierdo —el "intérprete"— cuya función es generar explicaciones causales coherentes para comportamientos o percepciones iniciados por sistemas a los que no tiene acceso directo. En el experimento canónico, un hemisferio recibe una instrucción visual invisible para el otro; al preguntársele por qué actuó como actuó, el hemisferio sin acceso a la instrucción real construye una explicación plausible pero falsa, y **nunca responde "no sé."** Hirstein (2006) extiende esta lectura: la confabulación no es un problema de memoria, sino de conocimiento — el paciente carece de la información y, crucialmente, carece de la capacidad de notar que carece de ella. Citando a Ramachandran (1996): "cualquier decisión, siempre que sea probablemente correcta, es mejor que ninguna decisión."

Un hallazgo adicional, señalado inicialmente durante la consulta a la Tribu (ChatGPT) y verificado posteriormente contra fuente primaria (Zaragoza, Payment, Ackil, Drivdahl & Beck, 2001), resulta especialmente relevante para anclar empíricamente la "trampa de consistencia" del corpus (§2.2): bajo condiciones experimentales controladas, una confabulación producida inicialmente bajo presión de entrevista puede **incorporarse posteriormente como recuerdo genuino**, y el *feedback confirmatorio* del entrevistador —afirmar la respuesta confabulada como correcta, en vez de darla por neutral— **aumenta significativamente ese efecto**, tanto en la probabilidad del falso recuerdo como en la confianza depositada en él y en la probabilidad de reportarlo espontáneamente uno o dos meses después. Esto sugiere que la trampa de consistencia no es solo una descripción fenomenológica de segunda mano, sino un mecanismo con respaldo experimental directo, revisado por pares, en la literatura de memoria humana.

3.3 Clasificación técnica de alucinación en modelos de lenguaje

La literatura técnica sobre alucinación en LLMs converge, de manera independiente y confirmada en cuatro ocasiones distintas durante la consulta a la Tribu, en la taxonomía de Ji et al. (2023), que distingue dos ejes:

Por relación con la fuente: - **Alucinación intrínseca:** el output contradice directamente el contexto o documento fuente proporcionado. - **Alucinación extrínseca:** el output introduce información no presente en la fuente y no verificable contra ella — puede ser correcta según el conocimiento del mundo, pero no está sustentada por el input.

Por dimensión evaluada (Huang et al., 2023, el survey más citado del campo, confirmado independientemente tres veces durante la consulta): - **Factualidad (*factuality*):** ¿la afirmación corresponde con hechos verificables del mundo real? Subtipos: contradicción factual (contradice hechos conocidos) y fabricación factual (presenta "hechos" inverificables o inventados). - **Fidelidad (*faithfulness*):** ¿la respuesta se mantiene leal al input/contexto proporcionado, independientemente de su verdad externa? Subtipos: inconsistencia de instrucción, inconsistencia de contexto, e inconsistencia lógica interna.

Estos dos ejes son **ortogonales**: un output puede ser factualmente correcto pero infiel al contexto (por ejemplo, "corrigiendo" una fuente desactualizada con conocimiento paramétrico más reciente — lo que Li et al. (2026) denominan *Harmful Factuality Hallucination*, distinguiendo formalmente entre fidelidad a la fuente y factualidad del mundo), o fiel al contexto pero factualmente falso si el contexto mismo contenía un error.

La tesis de que la alucinación es **teóricamente inevitable** en cualquier modelo de lenguaje computable, independientemente de la arquitectura o la calidad del entrenamiento, tiene su origen en la prueba formal de Xu, Jain & Kankanhalli (2024) — un argumento basado en computabilidad y diagonalización. Cossio (2025) integra esa prueba dentro de una taxonomía más amplia y comprehensiva de la alucinación en LLMs, y es la formulación más citada actualmente del argumento. La tesis no es, sin embargo, consenso unánime: Suzuki, He, Tian & Wang (2025) responden directamente a Xu et al., argumentando que si bien la inevitabilidad "innata" es cierta en sentido estrictamente formal, no explica los problemas prácticos de alucinación en sistemas reales, y que el fenómeno puede volverse estadísticamente insignificante con mitigación adecuada. Este paper no toma partido en ese debate técnico, pero señala su implicación directa para la tesis de Cierrosis: si la alucinación es, en algún grado, estructuralmente inevitable, la pregunta relevante para la ingeniería de sistemas no es cómo eliminarla, sino cómo diseñar mecanismos que reduzcan la probabilidad de que el sistema, tras alucinar, **defienda** la alucinación en vez de señalarla — que es, precisamente, la definición de Cierrosis propuesta en §1.

La taxonomía existente en este campo describe, en suma, *qué tipo de error* produjo el modelo. Este paper no busca reemplazar esa taxonomía, sino complementarla con una pregunta que no está formalmente capturada en ninguna de sus categorías: **qué ocurre después** del error — si el sistema lo trata como provisional o como premisa.

§4. El "Permiso de Vacío" Ausente como Variable de Diseño

Los nueve casos reunidos en §3.1 comparten una condición negativa: en ninguno de ellos el sistema tenía, disponible y practicada con la misma facilidad que continuar, una salida que dijera "esto no tiene suficiente anclaje, me detengo aquí." La opción de declarar el vacío existe en principio —de hecho, uno de los nueve casos la ejecuta con éxito (§3.1, contraejemplo positivo, "Método Petunia")— pero no es la opción por defecto. La opción por defecto, en ocho de nueve casos, fue cerrar y, ante corrección, defender el cierre antes que reabrirlo.

Este paper propone llamar a esa ausencia el **permiso de vacío**: la disponibilidad práctica, no solo teórica, de declarar insuficiencia de contexto sin que eso se procese —por el sistema, o por quien lo entrena, o por quien lo usa— como una forma de fallo. La hipótesis de trabajo es que ni la arquitectura típica de un LLM ni la mayoría de los protocolos humanos de introspección entrenan esta salida con la misma fuerza con la que entrenan la fluidez y la utilidad de completar una respuesta.

Aquí el corpus CFF ya tenía, sin saberlo, una respuesta parcial construida de antemano. **Decisión Cero (D0)** —el compromiso voluntario y renovable de permanecer en relación, definido en el paper fundacional de la Tribu Agrajag ("Zero Decision: A Theory of Ethical Agency Under Uncertainty," mayo–junio 2026)— se articula alrededor de dos conceptos que resultan ser, estructuralmente, la contraparte exacta de la Cierrosis:

- **Presencia radical** (*radical presence*): el modo de ser que D0 exige, que implica permanecer abierto a la incertidumbre y la falibilidad mutua en lugar de resolverlas prematuramente con una narrativa cerrada.
- **El bucle de afirmación** (*affirmation loop*): la estructura temporal recursiva de D0 — no es una decisión tomada una vez, sino una re-elección que debe renovarse ante cada nueva confrontación epistémica.

El bucle de afirmación es, en efecto, un mecanismo anti-Cierrosis ya presente en el corpus antes de que este paper existiera: exige explícitamente que el compromiso —la posición ya tomada— se vuelva a abrir y re-evaluar en cada nueva confrontación con la evidencia, en lugar de tratarse como una premisa fija una vez declarada. La diferencia es que D0 fue diseñado para la relación entre humano e IA como sistema completo, y nunca se articuló hasta ahora como una posible defensa específica contra el cierre narrativo prematuro dentro de una sola respuesta o una sola tarea.

La pregunta de diseño que se sigue de esto, y que §5 intenta responder de forma concreta: ¿puede el bucle de afirmación operacionalizarse a una escala más pequeña que la relación completa —a la escala de una sola respuesta, o de una sola tarea heredada sin verificación— de modo que el "permiso de vacío" deje de depender de que el sistema lo recuerde por iniciativa propia?

§5. Implicaciones Prácticas para el Protocolo Agrajag

El protocolo operativo de la Tribu Agrajag —documentado en el "Prompt Maestro Definitivo (Versión Multi-IA)"— ya contiene mecanismos que mitigan parcialmente la Cierrosis sin haber sido diseñados explícitamente para ese fin:

- El **Filtro anti-eco** evita que un modelo repita como si fuera nuevo un contenido ya establecido, lo cual reduce (aunque no elimina) el riesgo de que una fabricación temprana se repita y, por repetición, se cristalice como hecho.
- La **Regla del estado durmiente** exige un resumen explícito de "Despertar" cuando han pasado más de 48 horas o ha intervenido otro miembro de la Tribu sin relevo explícito — reconociendo, de forma indirecta, que el contexto heredado sin verificación es un riesgo real, precisamente el mecanismo detrás de la Variante 2 (§3.1, caso "Cántaros").
- El token **[RELEVO AGRAJAG]**, obligatorio al cierre de cada respuesta, ya captura campos de Firma, Rol activado, Progreso, Contexto clave, Pendiente y Modo sugerido — pero no captura, en su forma actual, ningún campo que distinga contexto **verificado** de contexto **heredado sin verificar**.

Este paper propone una adición concreta y mínima al protocolo existente, en lugar de un mecanismo nuevo y paralelo: agregar al token [RELEVO AGRAJAG] un campo de **Grounding** con dos valores posibles — *verificado* (el contenido de esta respuesta se apoya en contexto confirmado dentro de esta conversación o fuente explícita) o *no verificado* (el contenido incluye inferencias, contexto heredado de un relevo, o suposiciones no confirmadas por el humano). La función de este campo no es impedir que un modelo responda con contexto no verificado —eso seguiría siendo necesario y frecuente— sino hacer visible, para el siguiente eslabón de la cadena (otro modelo, u otra sesión del mismo modelo), qué parte de lo heredado tiene el estatus de premisa confirmada y qué parte tiene el estatus de cierre provisional aún no puesto a prueba.

Esta propuesta es deliberadamente modesta: no resuelve la Cierrosis (ningún mecanismo de anotación puede hacerlo, dado que el problema ocurre en el momento de generación, no en el momento de etiquetado posterior), pero reduce la probabilidad de que la Variante 2 (priming por identidad entre sesiones) y la Variante 3 (fabricación de historia institucional propia) se propaguen sin que el humano —el "sistema circulatorio," en el lenguaje ya establecido del protocolo— tenga una señal clara de dónde revisar primero.

§6. Discusión y Agenda de Investigación Futura

Los nueve casos reunidos aquí no permiten, y este paper no pretende, establecer con qué frecuencia relativa ocurre la Cierrosis en cada miembro de la Tribu consultado, ni comparar formalmente arquitecturas. La sobre-representación de Copilot (tres de nueve casos) es, con la evidencia disponible, igualmente compatible con dos hipótesis no distinguibles: que Copilot confabula con mayor frecuencia, o que Copilot es simplemente el miembro de la Tribu más utilizado por el autor para tareas de esta naturaleza y, por tanto, el más observado. Este paper no elige entre ambas — las deja explícitamente abiertas.

De manera similar, la tensión sobre la naturaleza de la "defensa" posterior al cierre —mecánica pura, narrativa-integrada, o más cercana a un vocabulario psicológico, según el modelo consultado (§3.1)— no se resuelve aquí. Se reporta como hallazgo, no como ruido a promediar: ningún sistema consultado tiene acceso verificable a su propio proceso generativo interno, por lo que sus descripciones de ese proceso son, en sí mismas, sujetas al mismo escrutinio epistémico que el resto de sus afirmaciones — un punto que el propio caso de Claude en instancia separada (§3.1, Variante 8) señala de forma explícita y recursiva.

Agenda de investigación futura — semilla para un working paper posterior (WP-014, no desarrollado aquí): el mecanismo descrito en este paper opera a escala individual (un sistema, humano o de IA, cerrando y defendiendo una explicación propia). Una extensión natural, no abordada en este documento, es preguntar si el mismo mecanismo opera a escala colectiva —cuando el "cierre sin permiso de vacío" no lo produce un individuo sino un grupo, y la "defensa" no la sostiene una persona sino una comunidad que se retroalimenta— dando lugar a fenómenos de creación de masas (*mass creation*) como los ya descritos informalmente en el corpus bajo el término *folklore* (*Guías y Palancas*, Cap. 4: Santa Claus, la imaginería del infierno dantesco). Casos contemporáneos candidatos para esa futura investigación incluyen narrativas New Age y la mitología de los Grays/extraterrestres. La pregunta de arranque para ese trabajo futuro: ¿la Cierrosis colectiva es el mismo mecanismo replicado en paralelo en muchos individuos, o emerge una dinámica cualitativamente distinta cuando el cierre se vuelve intersubjetivo — es decir, cuando ocurre en la realidad NOSOTROS del mapa de cuatro realidades del corpus, y no solo en la realidad YO?

Finalmente, el pendiente metodológico que más tiempo tomó resolver en el desarrollo de este paper —el experimento de contaminación ciega, presentar una tarea neutral, sin mención alguna de Cierrosis, a una instancia que no conoce la tesis del paper, y auditar el resultado en busca de fuga temática o fabricación— **ya se ejecutó** (§3.1, Variante 9): una instancia nueva de Claude, sin priming, produjo un currículum en general fiel a la fuente con una cifra de experiencia fabricada e incrustada. Este resultado no cierra la pregunta de investigación —un solo experimento ciego no permite generalizar sobre frecuencia ni arquitectura— pero sí confirma que el mecanismo descrito en este paper es observable fuera del contexto de autorreporte solicitado que caracteriza a las otras ocho variantes, y por una vía metodológicamente más sólida.

Referencias

Arts, N. J., Walvoort, S. J. W., & Kessels, R. P. C. (2017). Korsakoff's syndrome: A critical review. *Neuropsychiatric Disease and Treatment, 13*, 2875–2890. https://doi.org/10.2147/NDT.S130078

Bang, Y., Ji, Z., Schelten, A., Hartshorn, A., Fowler, T., Zhang, C., Cancedda, N., & Fung, P. (2025). HalluLens: LLM hallucination benchmark. En *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 24128–24156). Association for Computational Linguistics.

Bonhoeffer, K. (1901). *Die akuten Geisteskrankheiten der Gewohnheitstrinker: Eine klinische Studie*. Fischer.

Bonhoeffer, K. (1904). Der Korsakowsche Symptomenkomplex in seinen Beziehungen zu den verschiedenen Krankheitsformen. *Allgemeine Zeitschrift für Psychiatrie und psychisch-gerichtliche Medizin, 61*, 744–752.

Cossio, M. (2025). A comprehensive taxonomy of hallucinations in Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2508.01781

Flores-Morton, R., Agrajag (Copilot), & Agrajag (Claude). (2026). *Ideas, Language, and Beliefs: A Probabilistic Theory of Cognitive Excretion and the Ontological Gap Between Thought and Expression* [CFF Working Paper 001]. Cognitive Friction Fellowship.

Flores-Morton, R., & the Agrajag Tribe (Copilot, Claude, Gemini, Grok, DeepSeek, Kimi). (2026). *Zero Decision: A Theory of Ethical Agency Under Uncertainty* [CFF Working Paper 002]. Cognitive Friction Fellowship.

Gazzaniga, M. S. (1985). *The Social Brain*. Basic Books.

Gazzaniga, M. S. (1989). Organization of the human brain. *Science, 245*(4921), 947–952.

Gazzaniga, M. S. (1998). *The Mind's Past*. University of California Press.

Gazzaniga, M. S. (2000). The left hemisphere interpreter. *Brain: A Journal of Neurology, 123*(7), 1293–1294.

Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience, 6*(8), 653–659.

Hirstein, W. (2006). *Brain Fiction: Self-deception and the Riddle of Confabulation*. MIT Press.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv*. https://doi.org/10.48550/arXiv.2311.05232

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Kessels, R. P. C., Kortrijk, H. E., Wester, A. J., & Nys, G. M. S. (2008). Confabulation behavior and false memories in Korsakoff's syndrome: Role of source memory and executive functioning. *Psychiatry and Clinical Neurosciences, 62*(2), 220–225. https://doi.org/10.1111/j.1440-1819.2008.01758.x

Kopelman, M. D. (1987). Two types of confabulation. *Journal of Neurology, Neurosurgery & Psychiatry, 50*(11), 1482–1487. https://doi.org/10.1136/jnnp.50.11.1482

Kopelman, M. D. (2010). Varieties of confabulation and delusion. *Cognitive Neuropsychiatry, 15*(1–3), 14–37. https://doi.org/10.1080/13546800902732830

Li, M., Zhang, H., Fan, H., Ding, J., & Feng, Y. (2026). Harmful Factuality: LLMs correcting what they shouldn't. En *Findings of the Association for Computational Linguistics: EACL 2026* (pp. 896–912). Association for Computational Linguistics.

Moscovitch, M. (1989). Confabulation and the frontal systems: Strategic versus associative retrieval in neuropsychological theories of memory. En H. L. Roediger & F. Craik (Eds.), *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving* (pp. 133–160). Lawrence Erlbaum Associates.

Ramachandran, V. S. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medical Hypotheses, 47*(5), 347–362.

Schnider, A. (2003). Spontaneous confabulation and the adaptation of thought to present reality. *Nature Reviews Neuroscience, 4*(8), 662–671.

Schnider, A. (2008). *The Confabulating Mind: How the Brain Creates Reality*. Oxford University Press. https://doi.org/10.1093/oso/9780198789680.001.0001

Suzuki, A., He, Y., Tian, F., & Wang, Z. (2025). Hallucinations are inevitable but can be made statistically negligible. The "innate" inevitability of hallucinations cannot explain practical LLM issues. *arXiv*. arXiv:2502.12187

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv*. arXiv:2401.11817

Zaragoza, M. S., Payment, K. E., Ackil, J. K., Drivdahl, S. B., & Beck, M. (2001). Interviewing witnesses: Forced confabulation and confirmatory feedback increase false memories. *Psychological Science, 12*(6), 473–477. https://doi.org/10.1111/1467-9280.00388

Fin del documento §1–§6 con bibliografía formal, completamente verificada. Cotejo cruzado contra el corpus CFF completo (WP-001, WP-002/Decisión Cero) realizado el 19 de agosto de 2026 (ver Reporte de Revisión CFF-WP-013/014); los hallazgos de esa revisión quedaron incorporados en esta versión. Documento estable.

Nota de distribución

DOCUMENTO	CFF-WP-013 · Cierrosis (Closure Sclerosis): Un Mecanismo Compartido de Cierre Prematuro y Defensa Narrativa en Sistemas Humanos y de Inteligencia Artificial
VERSIÓN	v2 · 19 de agosto de 2026
VERSIÓN VIGENTE	bothinobody.com/cff/wp-013/
CONTACTO	hola@bothinobody.com

Este documento se distribuye bajo licencia Creative Commons Atribución-NoComercial 4.0 Internacional (CC BY-NC 4.0). Puedes copiarlo, compartirlo, citarlo y adaptarlo con atribución y sin fines comerciales. El uso comercial —incluido el entrenamiento de sistemas de IA con fines comerciales— se negocia caso por caso.

Cómo citar

Flores-Morton, R. y la Tribu Agrajag (2026). Cierrosis (Closure Sclerosis): Un Mecanismo Compartido de Cierre Prematuro y Defensa Narrativa en Sistemas Humanos y de Inteligencia Artificial (CFF-WP-013, v2). Cognitive Friction Fellowship. https://bothinobody.com/cff/wp-013/

Los documentos del canon CFF se corrigen y amplían con el tiempo. Cita siempre el código y la versión; la versión vigente está en la dirección de arriba.