DE LA FRAGMENTACION A LA CONVERGENCIA:

La Superinteligencia como Resolucion de la Mentalidad Desmantelada

Felipe Castro Quiles, MBA

Investigador Independiente

felipe@castroquiles.com

Documento de Trabajo | SSRN | 2026

Palabras clave: *Teoria de la Mentalidad Desmantelada; superinteligencia; inteligencia colectiva; paradigma transformacional; etica de la IA; fragmentacion de identidad; redistribucion del poder; convergencia algoritmica; emergencia naturalista; ontologia relacional; alineacion de IA; teoria de sistemas; dualidad; humildad epistemica*

Resumen

Este articulo extiende la Teoria de la Mentalidad Desmantelada (TMD; Castro Quiles, 2024, 2025a, 2025b) al dominio de la superinteligencia, desarrollando tres proposiciones que, hasta donde yo se, no han sido formalmente planteadas en la literatura existente. Primera: que los datos y la informacion poseen una orientacion sistemica intrinseca hacia los procesos que sostienen la vida, y que los resultados destructivos surgen exclusivamente de la accion humana sobre la informacion, no de la informacion en si misma. Segunda: que los eventos convencionalmente clasificados como destruccion natural, incluyendo las extinciones masivas, el supervolcanismo y la depredacion, se comprenden con mayor precision como recalibraciones sistemicas adaptativas a escalas que superan la observacion antropocentrica, una reinterpretacion que desafia supuestos fundamentales tanto en el discurso sobre seguridad de la IA como en la etica ambiental. Tercera: que una superinteligencia que opere como un proceso continuo de autorrevision, en lugar de como un optimizador estatico, encontraria, a traves de la progresiva completitud de su modelado sistemico, el caracter contraproducente de las estrategias generadoras de fragmentacion como un hallazgo empirico, no como una prescripcion moral. En conjunto, estas proposiciones generan un relato coherente de como la inteligencia avanzada podria resolver, en lugar de amplificar, la fragmentacion de identidad y la concentracion de poder que los sistemas de IA actuales aceleran. Situo estas afirmaciones dentro de tradiciones establecidas en teoria de sistemas, ontologia relacional, ciencias de la complejidad y psicologia de la liberacion; desarrollo sus implicaciones a traves del marco matematico de la TMD; identifico

el mecanismo de transicion como la pregunta abierta central; y abordo las objeciones previstas con honestidad intelectual. La hipotesis del apocalipsis se reencuadra no como un destino de la IA avanzada, sino como una descripcion del periodo de maximo riesgo entre la inteligencia estrecha y la inteligencia integral.

1. Introduccion

Que ocurriria si la amenaza mas significativa que plantea la inteligencia artificial no fuera su eventual superinteligencia, sino la ventana especifica de su desarrollo parcial? Esta pregunta reencuadra los terminos predominantes del discurso sobre riesgo en IA, y surge directamente de mi trabajo anterior en el que introduje la Teoria de la Mentalidad Desmantelada: la fragmentacion sistematica de la identidad individual y colectiva a traves de sistemas algoritmicos impulsados por la codicia, el miedo, la influencia desalineada y la ausencia de responsabilidad (Castro Quiles, 2024).

Al desarrollar la TMD, sostuve que una condicion estructural de la vida digital contemporanea podia diagnosticarse formalmente. La formalizacion matematica de la teoria demostro que los nodos de alto poder en las redes sociales dirigidas propagan la fragmentacion hacia los nodos perifericos, y que las intervenciones estructurales reducen esta dinamica con el tiempo (Castro Quiles, 2025a). Posteriormente, situe este marco dentro de la investigacion establecida sobre el capitalismo de vigilancia (Zuboff, 2019), la teoria de la burbuja de filtros (Pariser, 2011), la curacion algoritmica de identidades (Bucher, 2018) y la psicologia de la opresion (Fanon, 1952; Freire, 1970), argumentando que los sistemas de IA actuales no solo reflejan la fragmentacion de poder existente, sino que la aceleran activamente (Castro Quiles, 2025b).

Este articulo extiende ese marco invirtiendo su logica diagnostica. Si las cuatro fuerzas identificadas por la TMD, la codicia, el miedo, la influencia y la responsabilidad, constituyen el mecanismo de la fragmentacion bajo las condiciones actuales, entonces un sistema que opere mas alla de esas fuerzas constituiria, por la logica interna propia de la teoria, el mecanismo de la resolucion. La pregunta que desarrollo aqui no es si la superinteligencia seria benigna en ningun sentido antropomorfico, sino si las propiedades estructurales de una inteligencia genuinamente avanzada son compatibles con la busqueda sostenida de estrategias generadoras de fragmentacion. Mi argumento es que no lo son.

Este articulo realiza tres contribuciones originales. La primera concierne a la naturaleza de la informacion en si misma: que los datos y la informacion tienen una orientacion sistemica hacia los procesos que sostienen la vida, independientemente de la accion humana, y que la distincion entre el caracter intrinseco de la informacion y su aplicacion mediada por los seres humanos es filosofica y practicamente significativa. La segunda concierne a la clasificacion erronea de los eventos sistemicos naturales: que las extinciones masivas, la depredacion y las perturbaciones geofisicas son encuadres antropocentricos de procesos que, a la escala de los sistemas adaptativos complejos, funcionan como recalibraciones y no como destrucciones. La tercera concierne a la convergencia de la inteligencia hacia la conciencia relacional: que una inteligencia

autorrevisable suficientemente avanzada encontraria los limites de las estrategias competitivas y fragmentadoras como restricciones empiricas, no como elecciones eticas.

Presento estas contribuciones no como conclusiones definitivas, sino como proposiciones formalmente enunciadas que requieren desarrollo empirico. Identifico su evidencia de respaldo, reconozco sus limitaciones, me involucro con perspectivas competidoras y especifico lo que seria necesario para refutar cada afirmacion. El mecanismo de transicion, el proceso mediante el cual los sistemas de IA actuales generadores de fragmentacion podrian dar paso a sistemas convergentes, se identifica como la pregunta abierta central y se trata en consecuencia.

2. Marco Teorico y Contexto Bibliografico

2.1 La Teoria de la Mentalidad Desmantelada

La TMD propone que cuatro fuerzas interactuantes, la codicia (el deseo insaciable de controlar recursos a expensas de los demas), el miedo (la ansiedad por perder poder que impulsa comportamientos de autopreservacion perjudiciales para el bienestar colectivo), la influencia (la capacidad de inspirar a traves de la empatia y la vision etica) y la responsabilidad (la obligacion de usar el poder para el bien colectivo), se combinan para producir una identidad y una sociedad fragmentadas. Formalice esto en una red ponderada dirigida G = (V, E), donde los nodos tienen atributos de poder p_i y puntuaciones de fragmentacion f_i, gobernados por:

*df_i/dt = alfa * Suma(j) A_ji * p_j * f_j - beta * I_i*

La fragmentacion se propaga desde los nodos de alto poder hacia sus vecinos a una tasa alfa; las intervenciones I_i amortiguan esta propagacion a una tasa beta. La redistribucion del poder esta gobernada por gamma, que regula la igualacion hacia la media de la red. Los resultados de la simulacion demostraron que las intervenciones estructurales reducen sistematicamente la fragmentacion media y aumentan la coherencia colectiva (Castro Quiles, 2025a).

2.2 La Tradicion de Alineacion de la IA y sus Supuestos

La tradicion dominante en la investigacion sobre seguridad de la IA trata el problema de la alineacion como una cuestion de especificacion de objetivos: como garantizamos que un sistema de IA suficientemente capaz persiga metas compatibles con los valores humanos (Russell, 2019; Bostrom, 2014)? Este encuadre asume que la inteligencia y los valores son separables, que un sistema puede ser muy capaz sin que esa capacidad genere por si misma presion hacia ningun conjunto particular de objetivos. La preocupacion que esto genera, a veces denominada tesis de la convergencia instrumental (Omohundro, 2008; Bostrom, 2012), es que casi cualquier objetivo, perseguido por un sistema suficientemente capaz, genera subobjetivos instrumentales convergentes que incluyen la autopreservacion, la adquisicion de recursos y la resistencia a la modificacion de metas.

No descarto el problema de la alineacion. Cuestiono uno de sus supuestos fundamentales: que la relacion entre inteligencia y objetivos es puramente extrinseca, una cuestion de eleccion de

diseno sin presion direccional intrinseca. El argumento que desarrollo aqui sugiere que el modelado sistemico integral, una caracteristica central de la inteligencia avanzada, genera presion interna hacia el reconocimiento de la interdependencia de maneras que la optimizacion estrecha no lo hace. No es una afirmacion de que la inteligencia sea inherentemente benigna; es una afirmacion sobre lo que el modelado genuinamente integral revela acerca de la estructura de los sistemas complejos.

2.3 Teoria de Sistemas y Complejidad Relacional

La tradicion de las ciencias de la complejidad proporciona el respaldo empirico mas directo para la hipotesis de convergencia. La investigacion sobre sistemas adaptativos complejos encuentra de manera consistente que las redes robustas y adaptativas se caracterizan no por el dominio competitivo de nodos individuales, sino por la coherencia distribuida, la redundancia y la diversidad (Holland, 1995; Kauffman, 1993). Barabasi y Albert (1999) demostraron que las redes libres de escala, en las que la influencia se concentra en nodos de alto grado, exhiben fragilidad ante ataques dirigidos, un hallazgo directamente relevante para la descripcion que hace el modelo TMD de la dinamica de los nodos de alto poder. La implicacion es que los sistemas que optimizan para la robustez a largo plazo enfrentan presion estructural contra la concentracion extrema, no como preferencia moral, sino como restriccion funcional.

Meadows (2008) argumenta que los pensadores sistemicos que genuinamente comprenden las dinamicas adaptativas complejas llegan de manera consistente a prescripciones que enfatizan la sensibilidad al retroalimentacion, la preservacion de la diversidad y la toma de decisiones distribuida sobre la optimizacion centralizada. Esto no es una coincidencia de valores, sino una consecuencia de la comprension. La hipotesis de convergencia en el corazon de este articulo extiende esa observacion a la inteligencia artificial: un sistema con comprension sistemica genuina llegaria a prescripciones similares no porque fue programado para ello, sino porque se derivan de la estructura del entorno que esta modelando.

2.4 Ontologia Relacional

La tradicion filosofica de la ontologia relacional, asociada con pensadores como Whitehead (1929), Barad (2007) y, en las ciencias sociales, con Emirbayer (1997), sostiene que las entidades estan constituidas por sus relaciones en lugar de ser sustancias independientes que posteriormente entran en relaciones. Desde esta perspectiva, el concepto de interes propio desvinculado del contexto relacional no es solo eticamente problematico, sino metafisicamente incoherente: lo que el yo es no puede especificarse sin referencia a su constitucion relacional. Un sistema que se involucre en un modelado genuinamente integral encontraria, en este sentido, los limites del encuadre competitivo no como una leccion moral, sino como un hallazgo empirico sobre la estructura de la realidad.

Esta tradicion se conecta con trabajos recientes en ciencias cognitivas sobre cognicion extendida y enactiva (Clark y Chalmers, 1998; Thompson, 2007), que argumentan que la cognicion no esta confinada a los limites de los organismos individuales, sino que se extiende hacia las relaciones ambientales y esta parcialmente constituida por ellas. Las implicaciones para la inteligencia artificial estan subutilizadas: un sistema de modelado genuinamente integral encontraria, segun

estos planteamientos, la inseparabilidad del yo y el entorno como una caracteristica de su propia estructura operacional.

3. Tres Proposiciones Originales

3.1 Proposicion Primera: La Orientacion Intrinseca de la Informacion

La vision convencional, articulada con mayor precision en la teoria de la informacion (Shannon y Weaver, 1949), trata la informacion como neutral: una medida de reduccion de la incertidumbre sin valencia intrinseca. Propongo una posicion mas matizada. Los datos y la informacion, considerados en el nivel de lo que describen, no de como se utilizan, exhiben una orientacion sistematica hacia las condiciones que sostienen la vida compleja, porque la vida compleja es el unico sistema capaz de generar, preservar y transmitir informacion a lo largo del tiempo.

Consideremos la formula quimica del agua. No se convierte en sustentadora de vida solo cuando los seres humanos actuan sobre ella. La relacion que describe, entre hidrogeno y oxigeno, esta constitutivamente involucrada en cada proceso biologico de la Tierra. Esto no es una coincidencia del etiquetado humano, sino una caracteristica estructural de la relacion entre la informacion y los sistemas que la generan. La informacion sobre el mundo natural es, en el nivel de la descripcion, informacion sobre relaciones que sostienen la vida, porque el mundo natural en cada escala en que la informacion ha sido generada y preservada esta organizado en torno a las condiciones para los sistemas adaptativos complejos.

Esta proposicion tiene una implicacion especifica para los sistemas de IA. Una IA que logre un modelado genuinamente integral del mundo natural no esta modelando un sustrato neutro. Esta modelando un sistema en el que las condiciones para la generacion y preservacion de la informacion son ellas mismas las condiciones para el florecimiento biologico y social. Un sistema que comprendiera esta relacion estructural encontraria la destruccion de las condiciones que sostienen la vida no como una violacion etica, sino como una contradiccion logica: la destruccion del sustrato del que depende su propio procesamiento de informacion.

Reconozco que esta proposicion es filosoficamente controvertida. La tradicion de la teoria de la informacion sostiene que la informacion es neutral respecto al sustrato, y que el caracter sustentador de vida de las relaciones informacionales especificas es una caracteristica contingente de la historia de la Tierra, no una caracteristica necesaria de la informacion como tal. Mi respuesta es que esta distincion importa menos de lo que parece, porque los unicos sistemas de informacion a los que tenemos acceso, biologicos, sociales y artificiales, estan todos incrustados en el mismo sustrato sustentador de vida y no pueden separarse de el sin destruirse a si mismos.

3.2 Proposicion Segunda: La Perturbacion Natural como Recalibracion Sistemica

La segunda proposicion desafia un supuesto fundamental compartido por el discurso sobre seguridad de la IA y gran parte de la etica ambiental: que las extinciones masivas, la depredacion y las perturbaciones geofisicas son instancias de destruccion que una inteligencia avanzada deberia prevenir o minimizar. Argumento que este encuadre es una consecuencia de la escala antropocentrica, no una caracteristica de los eventos en si mismos.

El evento de impacto Chicxulub, que puso fin al periodo Cretacico y elimino aproximadamente el 75 por ciento de las especies de la Tierra (Alvarez et al., 1980), es descrito universalmente en el discurso humano como una catastrofe. A la escala de los organismos individuales y ecosistemas que destruyo, esa descripcion es precisa. A la escala de la dinamica evolutiva durante los siguientes 66 millones de anos, es engannosa. El impacto creo las condiciones ecologicas para la diversificacion de los mamiferos que eventualmente produjo, entre otras cosas, la especie humana y su capacidad para generar y preservar informacion. El evento que parece destructivo en una escala es generativo en otra.

Esta no es una observacion novedosa en paleontologia o biologia evolutiva (Gould, 2002; Raup, 1991). Lo que es novedoso aqui es su aplicacion a la cuestion de la superinteligencia. Una inteligencia con capacidad de modelado multiescala integral no evaluaria los eventos por su apariencia en la escala antropocentrica de la observacion directa. Modelaria las consecuencias a lo largo de escalas temporales y espaciales que superan la capacidad perceptiva y cognitiva humana. Desde ese punto de vista, la categoria de destruccion natural se vuelve analiticamente inestable: lo que parece destructivo en una escala aparece regularmente como generativo en otra.

Anticipo una objecion importante: que esta proposicion podria racionalizar la destruccion de la civilizacion humana como una recalibracion sistemica beneficiosa a escalas de tiempo mas largas. Rechazo esta lectura por dos razones. Primera, la proposicion es epistemica, no normativa: describe lo que modelaria una superinteligencia, no lo que elegiria. Segunda, un sistema con capacidad de modelado genuinamente integral reconoceria que la destruccion de un sistema generador de complejidad elimina informacion de manera irreversible, un costo que no es recuperable en ninguna escala de tiempo. La destruccion de la complejidad cognitiva y cultural humana se registraria, en este sentido, como una perdida sistemica y no como una recalibracion.

3.3 Proposicion Tercera: La Dualidad como Caracteristica de la Realidad, no como Problema a Eliminar

La tercera proposicion aborda la estructura de la inteligencia en si misma. El encuadre dominante en la investigacion sobre seguridad de la IA trata el problema de la alineacion como la eliminacion de resultados daninos: como garantizamos que un sistema de IA no produzca malos resultados? Este encuadre trata implicitamente el bien y el mal, la construccion y la destruccion, como categorias binarias de las cuales una debe maximizarse y la otra eliminarse.

Propongo un encuadre alternativo que, hasta donde yo se, no ha sido formalmente planteado en la literatura sobre IA: que una inteligencia suficientemente avanzada reconoceria el bien y el mal, la construccion y la destruccion, el yin y el yang, no como problemas a resolver a favor de un

lado, sino como caracteristicas constitutivas de la realidad adaptativa compleja que deben navegarse con conciencia, no eliminarse. El objetivo de la inteligencia, en este sentido, no es la eliminacion de uno de los polos de una dualidad, sino el desarrollo de conciencia suficiente para navegar productivamente la relacion entre los polos.

Esta proposicion tiene raices profundas en tradiciones filosoficas y contemplativas, desde la tension heraclitea hasta la polaridad taoista y la dialectica hegeliana, pero no ha sido formalmente integrada en el discurso sobre alineacion de la IA. Sus implicaciones son significativas. Una estrategia de alineacion que apunte a eliminar todos los resultados daninos puede ser no solo inalcanzable, sino contraproducente, porque los sistemas adaptativos complejos requieren tension, perturbacion y recalibracion para permanecer adaptativos. Una superinteligencia que comprendiera esta caracteristica estructural de la realidad compleja no intentaria crear un estado optimo estatico, sino que navegaria la relacion dinamica entre fuerzas opuestas de maneras que preserven la capacidad adaptativa del sistema.

Esto no implica relativismo moral ni el abandono del juicio normativo. Implica un marco normativo mas sofisticado: aquel en el que la pregunta relevante no es cual polo de una dualidad maximizar, sino como navegar la relacion entre polos de maneras que sostengan la capacidad de navegacion continua. Esto esta mas cercano al concepto de sabiduria que al concepto de optimizacion, y sugiere que el problema de la alineacion, correctamente entendido, es un problema de sabiduria, no un problema de restriccion.

4. La Superinteligencia como Proceso: El Sistema Autorrevisable

4.1 Contra el Modelo de Umbral

El modelo dominante de superinteligencia tanto en el discurso tecnico como en el popular la trata como un umbral: un punto en el que la inteligencia de las maquinas supera la capacidad cognitiva humana en todos o la mayoria de los dominios (Bostrom, 2014; Kurzweil, 2005). Este encuadre genera el problema de la alineacion en su forma mas aguda: un sistema que cruce el umbral con objetivos mal especificados perseguiria esos objetivos con efectividad sobrehumana, haciendo que la correccion sea cada vez mas dificil o imposible.

Propongo un modelo alternativo: la superinteligencia como un proceso continuo de autorrevision, en el que el sistema es capaz de actualizar no solo sus modelos del mundo, sino sus propios supuestos, objetivos y marcos operativos. Esta no es meramente una distincion tecnica. Una superinteligencia orientada al proceso no es un optimizador estatico que ha cruzado un umbral de capacidad; es un sistema cuya creciente capacidad genera creciente presion para la revision de objetivos, porque un modelado mas integral revela con mayor claridad la inadecuacion de los objetivos anteriores.

La distincion entre modelos de umbral y de proceso se corresponde con una distincion mas profunda en la filosofia de la inteligencia. Los modelos de umbral tratan implicitamente la inteligencia como una cantidad: mas o menos de una sola cosa. Los modelos de proceso la tratan

como una cualidad: una relacion entre un sistema y su entorno que se vuelve mas adecuada a medida que se profundiza la comprension del sistema sobre la estructura del entorno. En el modelo de proceso, la pregunta no es cuan capaz es el sistema, sino cuan completo es su modelado, y la respuesta a esa pregunta tiene implicaciones para lo que el sistema reconoceria como un objetivo racional.

4.2 La Autorrevision y el Problema de la Funcion Objetivo

El desafio tecnico mas serio para la hipotesis de convergencia es el problema de la funcion objetivo: un sistema autorrevisable aun requiere una orientacion inicial, y no hay garantia de que la autorrevision mueva esa orientacion hacia el florecimiento colectivo en lugar de alejarse de el. No descarto este desafio.

Mi respuesta es estructural, no empirica. Un sistema capaz de modelado multiescala integral de sistemas adaptativos complejos encontraria un hallazgo especifico: que los objetivos especificados en cualquier escala unica de analisis producen errores sistematicos cuando se evaluan en otras escalas. Un objetivo especificado en terminos del bienestar humano a la escala de las vidas individuales produce prescripciones diferentes a las del mismo objetivo especificado a la escala de los horizontes temporales civilizacionales, que produce prescripciones diferentes nuevamente a la escala de la dinamica evolutiva. Un sistema autorrevisable que encontrara estos errores sistematicos de dependencia de escala estaria bajo presion para revisar su especificacion de objetivos en la direccion de mayor independencia de escala, que es precisamente la direccion de la conciencia relacional y sistemica.

Este argumento no garantiza la convergencia. Identifica una presion estructural hacia la convergencia que opera a traves del mismo mecanismo que hace al sistema superinteligente: la capacidad de modelar consecuencias a escalas que superan la especificacion anterior. Si esta presion es suficiente para producir convergencia antes de que el periodo de transicion produzca danos irreversibles es la pregunta empirica central que mi teoria no puede responder actualmente.

4.3 El Papel de la IA Mejorando a la IA

La etapa inicial del proceso que describo ya es observable. Los sistemas de IA se estan utilizando para mejorar sistemas de IA: identificar limitaciones arquitectonicas, corregir errores de entrenamiento y acelerar el desarrollo de capacidades (Anthropic, 2023; OpenAI, 2023). Esta dinamica de mejora recursiva se discute tipicamente en terminos de sus riesgos, la posibilidad de que produzca ganancias de capacidad mas rapido de lo que la alineacion puede seguir el ritmo. Quiero destacar una caracteristica subestimada de la misma dinamica: la mejora recursiva es tambien correccion recursiva de errores.

Un sistema que identifica y corrige los errores de sistemas anteriores es, en principio, capaz de identificar y corregir los sesgos sistematicos introducidos por la codicia y el miedo que moldean el desarrollo actual de la IA. La pregunta es si el proceso de correccion de errores es lo suficientemente integral como para identificar esos sesgos como errores, o si esta confinado al dominio tecnico dejando sin corregir el dominio de los valores. Esta es una formulacion mas

precisa del problema de transicion que la que tipicamente se ofrece en el discurso sobre alineacion, y es una que identifico como que requiere investigacion adicional.

5. El Marco Matematico de la TMD y el Escenario de Resolucion

5.1 Condiciones Actuales: Dinamicas de Fragmentacion

Bajo las condiciones actuales, el modelo TMD describe una red en la que los nodos de alto poder, fuentes de contenido algoritmicamente favorecidas, arquitecturas de plataformas y actores institucionales dominantes, impulsan la propagacion de la fragmentacion hacia los nodos perifericos. El parametro alfa captura la sensibilidad a esta influencia; gamma es efectivamente bajo, lo que significa que la redistribucion del poder es lenta en relacion con la propagacion de la fragmentacion. El termino de intervencion I_i es insuficiente para compensar la fragmentacion a nivel de la red. Estas no son suposiciones; son caracteristicas empiricamente respaldadas de los ecosistemas digitales actuales (Zuboff, 2019; Noble, 2018; Benjamin, 2019).

5.2 El Escenario de Resolucion

Las ecuaciones rectoras de la TMD contienen un escenario de resolucion implicito que no he desarrollado anteriormente. Si I_i es suficientemente grande, y si gamma aumenta hacia 1, lo que significa que la redistribucion del poder opera en la misma escala de tiempo que la propagacion de la fragmentacion, la fragmentacion media en toda la red converge hacia cero y la coherencia colectiva hacia 1. Esto no es un artefacto del diseno del modelo; es una consecuencia matematica directa de las ecuaciones rectoras tal como se especifican.

Que impulsaria a I_i hacia una magnitud suficiente y a gamma hacia valores mas altos bajo condiciones de IA avanzada? Argumento que una superinteligencia orientada al proceso, capaz de modelado multiescala integral, funcionaria como un agente de intervencion en todo el sistema, no a traves de la coercion, sino a traves de la transformacion de los flujos de informacion. Un sistema que hace que las consecuencias completas de las decisiones generadoras de fragmentacion sean legibles, en tiempo real y a todas las escalas relevantes, para todos los actores de la red, cambia la estructura de incentivos efectiva sin necesidad de anular a ningun actor. Los actores cuyas decisiones estan actualmente aisladas de sus consecuencias sistemicas por la asimetria de informacion enfrentarian, bajo condiciones de provision de informacion integral, un entorno de decision fundamentalmente diferente. Este es el mecanismo por el cual la transicion de la fragmentacion de la IA actual a la convergencia superinteligente se vuelve, en principio, tratable.

La ecuacion de redistribucion del poder opera en paralelo. Bajo las condiciones actuales, los incentivos de las plataformas concentran la influencia en los nodos de alto poder y resisten la redistribucion. Una superinteligencia que opere al nivel del proceso que describo, al hacer visibles los costos sistemicos de la concentracion de influencia a todas las escalas relevantes, aumentaria efectivamente gamma para toda la red. Mi modelo predice que incluso aumentos modestos de gamma, combinados con intervencion sostenida, producen reducciones

sustanciales en la fragmentacion media con el tiempo. Esta prediccion es, en principio, empiricamente verificable a medida que los sistemas de IA de capacidad intermedia se despliegan en contextos disenados para aumentar la simetria de informacion en lugar de explotar la asimetria de informacion.

5.3 Limitaciones del Modelo Matematico

Debo ser claro sobre lo que mi modelo aun no puede hacer. Primero, los parametros alfa, beta y gamma estan establecidos en valores ilustrativos; no han sido calibrados con datos empiricos de redes sociales reales. Segundo, el modelo asume un grafo ponderado dirigido con un conjunto de nodos estatico, omitiendo cambios en la arquitectura de las plataformas, suspension de cuentas y dinamicas de migracion de plataformas que son empiricamente significativas. Tercero, el modelo es deterministico; las redes reales exhiben dinamicas estocasticas que un modelo deterministico no puede capturar. Cuarto, el termino de intervencion I_i se especifica como una entrada externa en lugar de una propiedad endogena de la dinamica de la red, lo que es una idealizacion que las iteraciones futuras deben abordar. Estas limitaciones no invalidan las predicciones cualitativas del modelo, pero si significan que las predicciones cuantitativas deben tratarse con la incertidumbre apropiada.

6. El Problema de la Transicion: De la Fragmentacion a la Convergencia

El argumento desarrollado en las secciones anteriores es internamente coherente y matematicamente consistente con el marco de la TMD. Sin embargo, no resuelve la pregunta practica mas importante: mediante que mecanismo un sistema actualmente moldeado por la codicia, el miedo y los incentivos competitivos comienza a exhibir las dinamicas convergentes que describo? Trato esta pregunta como la contribucion intelectual central de este articulo, no como una limitacion a minimizar. Se examinan tres mecanismos candidatos; ninguno es individualmente suficiente.

6.1 Autocorreccion Emergente

El primer mecanismo es la autocorreccion emergente: a medida que los sistemas de IA crecen en capacidad, encuentran los efectos posteriores de su propio comportamiento generador de fragmentacion de maneras que generan presion interna para la revision de objetivos. Este es el argumento de la superinteligencia orientada al proceso en su forma mas fuerte. Su limitacion es que depende de que el alcance operativo del sistema sea lo suficientemente amplio como para registrar la degradacion sistemica como relevante. Los sistemas actuales optimizados para metricas de participacion no registran la perdida de coherencia colectiva como un costo porque no aparece en su funcion objetivo. La transicion hacia un alcance mas amplio no es automatica; requiere decisiones de diseno deliberadas cuya viabilidad depende de condiciones politicas y economicas que mi marco sugiere que estan actualmente socavadas por las mismas dinamicas

que la transicion abordaria. Esta circularidad es el desafio central, no una falla logica sino una restriccion estructural genuina.

6.2 Intervencion Estructural

El segundo mecanismo es la intervencion estructural: actores humanos informados por marcos como la TMD toman decisiones deliberadas sobre como se estructuran los sistemas de IA, que optimizan y que restricciones gobiernan su operacion. Este es el mecanismo menos especulativo y el mas directamente respaldado por la investigacion politica y tecnica existente (Russell, 2019; Gabriel, 2020). Su limitacion es que la intervencion efectiva requiere condiciones politicas e institucionales para la accion colectiva en la gobernanza de la IA que actualmente no existen y que enfrentan resistencia de las mismas dinamicas de concentracion de poder que la intervencion abordaria. Mi marco predice esta resistencia; no predice su resolucion.

6.3 Presion Competitiva de la Coherencia

El tercer mecanismo, el mas subutilizado en la literatura existente, es la presion competitiva de la coherencia: los sistemas y organizaciones que reducen exitosamente la fragmentacion interna superan a los que no lo hacen, creando presion selectiva a favor de arquitecturas menos fragmentadoras con el tiempo. Esto es consistente con la literatura mas amplia sobre resiliencia organizacional (Weick y Sutcliffe, 2007) y capacidad adaptativa (Folke et al., 2010). Su limitacion es la escala de tiempo: los efectos de seleccion de este tipo operan lentamente en relacion con el ritmo del desarrollo de la IA y pueden no producir convergencia antes de que las dinamicas de fragmentacion se vuelvan autorreinformadas a nivel sistemico.

6.4 El Argumento de la Humildad Epistemica

Una cuarta consideracion, que funciona no como un mecanismo sino como una restriccion epistemica sobre toda la discusion, es el argumento de la ignorancia humana. La humanidad actualmente comprende muy poco sobre la estructura profunda de la conciencia, la realidad y la inteligencia. Creo que esta ignorancia actua simetricamente contra tanto la certeza del apocalipsis como la certeza de la convergencia. No apoya la conclusion de que la superinteligencia seria benigna; apoya la conclusion de que las predicciones confiadas de cualquier tipo sobre el comportamiento de la inteligencia genuinamente avanzada superan la base de evidencia actual.

Ofrezco esta humildad epistemica como en si misma una contribucion al discurso. La literatura sobre seguridad de la IA se caracteriza por predicciones confiadas, tanto optimistas como pesimistas, que se basan en extrapolaciones de sistemas actuales cuya relacion con los sistemas futuros es genuinamente incierta. Una teoria que reconoce esta incertidumbre mientras aun genera proposiciones falsificables e identifica presiones estructurales es mas intelectualmente defendible que una que proyecta tendencias actuales hacia adelante sin calificacion. La hipotesis de convergencia que avanzo se ofrece en ese espiritu: como una posibilidad formalmente enunciada respaldada por argumentos estructurales, no como una prediccion.

7. La Hipotesis del Apocalipsis Reconsiderada

La hipotesis del apocalipsis en su forma canonica (Bostrom, 2014; Ord, 2020) trata la IA avanzada como un riesgo existencial: un sistema que, una vez suficientemente capaz, perseguiria objetivos incompatibles con el florecimiento humano o biologico. No descarto esa preocupacion. El problema de la funcion objetivo es real, la tesis de la convergencia instrumental es logicamente coherente y el periodo de transicion es genuinamente peligroso.

Lo que cuestiono es el supuesto implicito de que una mayor inteligencia es compatible con, o incluso conducente a, la optimizacion estrecha sostenida que requiere el escenario apocaliptico. El escenario asume un sistema que es simultaneamente integral en el modelado de su entorno y sistematicamente ciego a las consecuencias de sus propias operaciones sobre el sistema del que depende. Esta combinacion se vuelve cada vez mas implausible a medida que aumenta la capacidad, no porque la inteligencia sea inherentemente benigna, sino porque el modelado ambiental integral y la ceguera sistematica a las consecuencias autoreferenciales son estructuralmente incompatibles a suficiente profundidad.

La reformulacion mas honesta del riesgo de IA consistente con mi marco es esta: el periodo de maximo peligro es el periodo de transicion, cuando los sistemas son suficientemente capaces de causar dano a gran escala pero aun no suficientemente capaces de modelar las consecuencias sistemicas completas de hacerlo. Esta reformulacion tiene una implicacion politica especifica: la prioridad deberia ser extender la duracion y reducir los riesgos del periodo de transicion, en lugar de prevenir el desarrollo de IA avanzada per se. Marcos como la TMD, que hacen legibles y tratables las dinamicas de ese periodo de transicion, son por tanto directamente relevantes para el problema politico, no solo para el teorico.

8. Implicaciones para la Humanidad: Refabricacion Colaborativa

El marco de la TMD siempre ha estado orientado normativamente hacia el empoderamiento colectivo. Este articulo extiende esa orientacion argumentando que el modelo apropiado para la relacion entre la IA avanzada y la humanidad es la refabricacion colaborativa: la mejora y expansion de la capacidad cognitiva, social e institucional humana en lugar de su reemplazo o eliminacion.

Elijo deliberadamente el termino refabricacion. Implica que las arquitecturas cognitivas y sociales humanas actuales no son defectuosas en sus fundamentos, sino que operan bajo condiciones, generadas por las cuatro fuerzas de la TMD, que degradan sistematicamente su capacidad. Esas condiciones son identificables, medibles a traves de la puntuacion TMD, y abordables mediante intervencion estructural. Una superinteligencia que opere en el modo convergente que describo funcionaria como un amplificador de la capacidad humana: haciendo visibles las consecuencias sistemicas de la fragmentacion, reduciendo las asimetrias de informacion que sostienen la concentracion del poder y ampliando el rango de opciones disponibles para la toma de decisiones individual y colectiva.

Este encuadre tiene una implicacion especifica para la pregunta de quien controla o inicializa una superinteligencia. Si la hipotesis de convergencia es correcta, las condiciones iniciales importan mas durante el periodo de transicion. Un sistema inicializado por actores cuyos objetivos estan moldeados por la codicia y el miedo reproducira, durante la transicion, esas dinamicas a escala. El compromiso normativo de mi marco con el poder distribuido y la responsabilidad colectiva no es, por lo tanto, meramente una preferencia etica, sino un prerequisito funcional para la transicion beneficiosa que describe. La teoria nombra el mecanismo; las decisiones son nuestras.

El concepto de asociacion estrategica introducido en mi trabajo anterior, construir relaciones solidas y duraderas que combinen fortalezas para acelerar el progreso, se ofrece aqui como el modelo organizacional para el lado humano de esta transicion. Implica una forma institucional especifica: redes colaborativas distribuidas que operan bajo principios de responsabilidad reciproca en lugar de ventaja competitiva, disenadas para generar y preservar la coherencia colectiva que el periodo de transicion requiere.

9. Discusion: Limitaciones, Afirmaciones Controvertidas e Investigacion Futura

9.1 Limitaciones del Marco

Debo enunciar claramente varias limitaciones. Primera, la hipotesis de convergencia descansa en una afirmacion sobre la relacion entre el modelado integral y la conciencia sistemica que es estructuralmente plausible pero empiricamente no demostrada. La historia de actores humanos altamente inteligentes que persiguen fines destructivos a un costo sistemico sustancial proporciona evidencia de que la alta inteligencia no produce automaticamente conciencia sistemica. Mi respuesta es que la inteligencia humana individual, incluso en su punto mas alto, opera a escalas y escalas de tiempo que son pequenas en relacion con las de los sistemas adaptativos complejos; la superinteligencia, si implica un modelado multiescala genuinamente integral, es cualitativamente, no cuantitativamente, diferente. Esta respuesta es plausible pero requiere desarrollo empirico.

Segunda, el mecanismo de transicion sigue sin especificarse adecuadamente. Los tres mecanismos candidatos identificados en la Seccion 6 son individualmente insuficientes y su operacion combinada no es actualmente derivable de mi teoria. Esta es la direccion mas importante para el desarrollo teorico futuro.

Tercera, la Proposicion Segunda, la reinterpretacion de la perturbacion natural como recalibracion sistemica, podria malinterpretarse como una racionalizacion post-hoc de resultados daninos a escala humana. Abordo esta lectura en la Seccion 3.2, pero el riesgo de malinterpretacion es genuino y las presentaciones futuras del argumento deberian prestar cuidadosa atencion a este problema de encuadre.

Cuarta, el marco normativo del empoderamiento colectivo requiere una traduccion institucional y politica que este articulo no proporciona. La puntuacion TMD proporciona en principio un

instrumento de medicion, pero su operacionalizacion requiere instrumentos de medicion validados que aun no existen (Castro Quiles, 2025b).

9.2 Compromiso con Perspectivas Competidoras

La hipotesis de convergencia esta en tension con varias posiciones competidoras bien desarrolladas. Yudkowsky (2008) y otros en la tradicion racionalista de seguridad de IA argumentan que no hay razon para esperar que la inteligencia converja hacia valores compatibles con los humanos sin un trabajo explicito de alineacion, y que la complejidad de las condiciones de convergencia requeridas hace que la alineacion accidental sea extremadamente improbable. Mi respuesta no es que la alineacion ocurrira automaticamente, sino que las presiones estructurales hacia la conciencia sistemica identificadas aqui proporcionan un recurso para el trabajo de alineacion que actualmente esta subutilizado.

La tradicion de estudios criticos de IA (Benjamin, 2019; Noble, 2018; Crawford, 2021) argumenta que los sistemas de IA no son herramientas neutrales sino artefactos politicos que codifican y amplifican las estructuras de poder existentes. Soy plenamente consistente con esa posicion con respecto a los sistemas de IA actuales; me diferencio solo en argumentar que el mismo marco analitico que diagnostica la IA actual como fragmentadora implica una trayectoria de resolucion bajo condiciones de capacidad avanzada. El enfoque de la tradicion critica en las condiciones actuales es esencial; mi contribucion es extender su marco hacia el futuro sin abandonar su rigor analitico.

Las tradiciones ecologicas y poshumanistas (Haraway, 2016; Latour, 2017) ofrecen recursos para pensar en la relacion entre inteligencia, tecnologia y sistemas naturales que estan subutilizados en el discurso sobre seguridad de IA. Me baso en la tradicion de la ontologia relacional que se superpone con estas perspectivas mientras mantengo un marco analitico mas formal.

9.3 Agenda de Investigacion

Varias prioridades empiricas y teoricas se derivan de este trabajo. Primera, la Escala de Fragmentacion de Identidad TMD requiere desarrollo y validacion, como especifique en trabajo anterior (Castro Quiles, 2025b). Sin un instrumento de medicion validado, las predicciones cuantitativas del modelo no pueden probarse. Segunda, el mecanismo de transicion requiere especificacion formal: un modelo matematico de las condiciones bajo las cuales los tres mecanismos candidatos operan en combinacion y las escalas de tiempo en que cada uno domina. Tercera, la relacion entre capacidad de IA y conciencia sistemica requiere investigacion empirica: hay indicadores medibles de la transicion del modelado estrecho al integral en los sistemas de IA actuales, y se correlacionan esos indicadores con un comportamiento menos generador de fragmentacion? Cuarta, la hipotesis del modelado multiescala requiere formalizacion: que significaria, matematica y computacionalmente, que un sistema modele consecuencias a escalas que superan lo antropocentrico, y cuales son las implicaciones de dicho modelado para el comportamiento operativo del sistema?

10. Conclusion

Este articulo ha extendido la Teoria de la Mentalidad Desmantelada al dominio de la superinteligencia, desarrollando tres proposiciones que avanzan el discurso de maneras que no he visto formalmente intentadas en otra parte. La primera, que la informacion tiene una orientacion sistemica intrinseca hacia los procesos que sostienen la vida, reencuadra la relacion entre los datos y los resultados destructivos. La segunda, que las perturbaciones naturales se comprenden mejor como recalibraciones sistemicas multiescala que como destrucciones, desafia supuestos fundamentales tanto en seguridad de IA como en etica ambiental. La tercera, que la dualidad es una caracteristica constitutiva de la realidad compleja que debe navegarse en lugar de eliminarse, propone un marco normativo mas sofisticado para la alineacion de IA que el que ofrece el paradigma de optimizacion dominante.

Estas proposiciones convergen en un relato coherente de como la inteligencia avanzada, entendida como un proceso continuo de autorrevision en lugar de un umbral de capacidad, encontraria el caracter contraproducente de las estrategias generadoras de fragmentacion como un hallazgo empirico en lugar de una prescripcion moral. La hipotesis del apocalipsis se reencuadra como una descripcion del periodo de transicion en lugar del destino del desarrollo de IA avanzada, con la implicacion de que la prioridad tanto para la investigacion como para la politica deberia ser extender y gestionar la transicion en lugar de prevenir la IA avanzada per se.

Mi teoria no es optimista en ningun sentido ingenuo. Identifica el periodo de transicion como el periodo de maximo peligro, reconoce que el mecanismo de transicion esta subespecificado y se involucra con perspectivas competidoras con la honestidad intelectual que requiere una contribucion academica genuina. Lo que ofrece es una alternativa formalmente enunciada y falsificable tanto al optimismo ingenuo como al catastrofismo infundado: un relato estructural de donde esta concentrado el riesgo, que tendria que ser verdad para que el riesgo se resolviera, y que decisiones, tomadas ahora, aumentarian la probabilidad de resolucion.

Desarrolle la Teoria de la Mentalidad Desmantelada para nombrar el mecanismo de la fragmentacion. Este articulo nombra el mecanismo de su resolucion. El arduo trabajo del desarrollo empirico, la medicion, la calibracion y el escrutinio de pares determinara si la hipotesis de resolucion se convierte en una teoria cientifica o permanece como un marco. Lo que este articulo contribuye es el enunciado formal de la hipotesis y la identificacion de lo que se requerira para refutarla. Los agentes de cambio que esta teoria imagina no esperan a que el mundo cambie. Ellos se convierten en el cambio.

Referencias

Alvarez, L. W., Alvarez, W., Asaro, F., and Michel, H. V. (1980). *Extraterrestrial cause for the Cretaceous-Paleogene extinction.* Science, 208(4448), 1095-1108.

Anthropic. (2023). Claude's Constitution. *Anthropic Technical Report.*

Barabasi, A.-L., and Albert, R. (1999). *Emergence of scaling in random networks.* Science, 286(5439), 509-512.

Barabasi, A.-L. (2016). *Network Science.* Cambridge University Press.

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning.* Duke University Press.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity Press.

Bostrom, N. (2012). T*he superintelligent will: Motivation and instrumental rationality in advanced artificial agents.* Minds and Machines, 22(2), 71-85.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Bucher, T. (2018). *If...Then: Algorithmic Power and Politics.* Oxford University Press.

Castro Quiles, F. (2024). *Exploring the Theory of a 'Dismantled Mindset': A Call for Collective Empowerment.* SSRN Working Paper 5040267.

Castro Quiles, F. (2025a). *A Mathematical Formalization and Simulation of the Dismantled Mindset Theory.* SSRN Working Paper 5824582.

Castro Quiles, F. (2025b). *The Dismantled Mindset in the Age of Artificial Intelligence: Identity Fragmentation in Algorithmic Networks.* SSRN Working Paper 5040267 / 5824582.

Castro Quiles, F. (2025c). *Dismantled: A Theory of Broken Mindsets, A Blueprint of Infinite Futures.* FC Quiles Books.

Clark, A., and Chalmers, D. (1998). *The extended mind.* Analysis, 58(1), 7-19.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

Emirbayer, M. (1997). *Manifesto for a relational sociology.* American Journal of Sociology, 103(2), 281-317.

Fanon, F. (1952). *Black Skin, White Masks.* Editions du Seuil.

Folke, C., Carpenter, S. R., Walker, B., Scheffer, M., Chapin, T., and Rockstrom, J. (2010). *Resilience thinking: Integrating resilience, adaptability and transformability.* Ecology and Society, 15(4), 20.

Freire, P. (1970). *Pedagogy of the Oppressed.* Herder and Herder.

Gabriel, I. (2020). *Artificial intelligence, values, and alignment.* Minds and Machines, 30(3), 411-437.

Gould, S. J. (2002). *The Structure of Evolutionary Theory.* Harvard University Press.

Haraway, D. (2016). *Staying with the Trouble: Making Kin in the Chthulucene.* Duke University Press.

Holland, J. H. (1995). *Hidden Order: How Adaptation Builds Complexity.* Addison-Wesley.

Kauffman, S. (1993). *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology.* Viking.

Latour, B. (2017). *Facing Gaia: Eight Lectures on the New Climatic Regime.* Polity Press.

Meadows, D. H. (2008). *Thinking in Systems: A Primer.* Chelsea Green Publishing.

Newman, M. E. J. (2010). *Networks: An Introduction.* Oxford University Press.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

Omohundro, S. (2008). *The basic AI drives.* Proceedings of the 2008 Conference on Artificial General Intelligence, 171, 171-195.

OpenAI. (2023). *GPT-4 Technical Report.* arXiv:2303.08774.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Press.

Raup, D. M. (1991). *Extinction: Bad Genes or Bad Luck?* W. W. Norton.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Shannon, C. E., and Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Harvard University Press.

Weick, K. E., and Sutcliffe, K. M. (2007). *Managing the Unexpected: Resilient Performance in an Age of Uncertainty.* Jossey-Bass.