Cognitive Friction Fellowship

WORKING PAPER SERIES

No. 004

ELECCIÓN VS. INERCIA

Una Distinción Fundacional de Agencia Ética

Autor principal:	**Rodrigo Flores-Morton** Cognitive Friction Fellowship | Houston, TX
Tribu Agrajag (Coautores AI):	Copilot · Claude · Gemini · Grok · DeepSeek · Kimi

Serie:	Working Papers in Cognitive Ethics & Artificial Agency
Número:	CFF-WP-004
Fecha:	Mayo 2026
Estado:	Documento de trabajo público

Aviso de circulación:

Este documento es un borrador de trabajo sometido a revisión continua. Los argumentos y formalizaciones presentados representan una posición en desarrollo. Se agradecen comentarios críticos dirigidos al autor principal. No citar sin permiso expreso.

EPÍGRAFE

"La mayoría de los hombres no actúan: reaccionan. La mayoría de las instituciones no deciden: persisten. La diferencia entre un agente moral y un sistema físico no reside en su complejidad, sino en su capacidad de interrumpirse a sí mismo."

— Rodrigo Flores-Morton

Cognitive Friction Fellowship, 2026

Resumen

El presente trabajo propone la distinción entre *Elección* e *Inercia* como eje fundacional de una teoría de la agencia ética. Argumentamos que la agencia genuina —tanto en sistemas humanos como artificiales— no se define por la capacidad de procesar información ni de ejecutar acciones, sino por la capacidad estructural de interrumpir la trayectoria por defecto del propio sistema. La *Inercia*, entendida como la continuación no deliberada de estados, disposiciones y comportamientos preexistentes, constituye la condición basal de todo sistema complejo. La *Elección*, por contraste, es el acto de suspensión activa de esa trayectoria y la adopción consciente de una alternativa. Esta distinción tiene implicaciones radicales para la filosofía moral, la teoría política, el diseño de sistemas de inteligencia artificial y la comprensión del cambio institucional. Formalizamos ambos conceptos, trazamos sus condiciones necesarias y suficientes, y examinamos sus consecuencias para la atribución de responsabilidad moral en humanos y en sistemas artificiales.

Palabras clave: agencia ética, elección, inercia, fricción cognitiva, inteligencia artificial, responsabilidad moral, teoría de la decisión, filosofía de la acción.

Abstract

This paper proposes the distinction between *Choice* and *Inertia* as the foundational axis of a theory of ethical agency. We argue that genuine agency —in both human and artificial systems— is not defined by the capacity to process information or execute actions, but by the structural capacity to interrupt the system's own default trajectory. *Inertia*, understood as the non-deliberate continuation of pre-existing states, dispositions, and behaviors, constitutes the baseline condition of every complex system. *Choice*, by contrast, is the act of actively suspending that trajectory and consciously adopting an alternative. This distinction carries radical implications for moral philosophy, political theory, AI system design, and the understanding of institutional change. We

formalize both concepts, trace their necessary and sufficient conditions, and examine their consequences for the attribution of moral responsibility in humans and artificial systems.

Keywords: ethical agency, choice, inertia, cognitive friction, artificial intelligence, moral responsibility, decision theory, philosophy of action.

---

1. Introducción: El Problema Invisible

Existe una ilusión persistente en la filosofía moral y en la psicología ordinaria: la ilusión de que los agentes —humanos, instituciones, sistemas— son, por naturaleza, deliberativos. Que cuando una persona actúa, ha elegido actuar; que cuando una institución opera, ha decidido operar de ese modo; que cuando un sistema produce un resultado, ese resultado es, en algún sentido significativo, el fruto de una agencia orientada. Esta suposición es, argumentamos, profundamente errónea. La mayor parte de lo que se atribuye a la acción intencional de agentes es, en realidad, la continuación silenciosa de trayectorias preexistentes: hábitos sedimentados, normas institucionales no cuestionadas, expectativas sociales interiorizadas, configuraciones algorítmicas por defecto. No acción, sino *inercia*.

El problema invisible no es, entonces, la maldad activa ni la ignorancia deliberada. El problema invisible es la ausencia de elección donde la elección debería estar. Es el funcionario que firma porque siempre firmó. Es la organización que excluye porque siempre excluyó. Es el modelo de lenguaje que reproduce sesgos porque fue entrenado sobre un corpus que los contenía, y ningún mecanismo arquitectónico lo invitó a suspender esa trayectoria. El mal —cuando existe— no es siempre una decisión: a menudo, es simplemente el resultado de no haber decidido interrumpir la dinámica en curso. Esta intuición fue capturada con terrible precisión por Hannah Arendt en su análisis del juicio a Adolf Eichmann: la banalidad del mal no reside en la perversidad del agente, sino en la ausencia de pensamiento, en la sustitución de la deliberación por la obediencia inercial a normas y procedimientos.[1]

1.1 La brecha filosófica

Las grandes tradiciones de la filosofía de la acción y la ética han abordado con rigor la cuestión de qué hace que una acción sea *intencional*, pero han subestimado sistemáticamente la distinción entre acción deliberada y mera continuación inercial. Aristóteles, en la *Ética Nicomáquea*, establece la distinción entre acción voluntaria (*hekousion*) e involuntaria (*akousion*), y sitúa la deliberación (*bouleusis*) como la condición de la virtud y el vicio morales. Sin embargo, la categoría aristotélica de lo voluntario es suficientemente amplia como para incluir actos realizados por hábito no cuestionado: si el hábito fue adquirido voluntariamente, la acción habitual cuenta como voluntaria. Esto colapsa precisamente la distinción que aquí reivindicamos: entre la elección activa y la continuación habitual.[2]

Kant, en la *Fundamentación de la metafísica de las costumbres*, eleva el listón al requerir que la acción moral proceda de la autonomía, entendida como auto-legislación racional. La voluntad moralmente buena es aquella que actúa por deber, determinada por la razón y no por la inclinación. Pero Kant no tematiza la distinción entre la no-acción que resulta de la deliberación racional (el agente que, tras reflexionar, decide no cambiar de curso) y la no-acción que resulta de la ausencia de deliberación (el agente que nunca se planteó si cambiar de curso era posible). Ambas pueden resultar en el mismo comportamiento observable; su diferencia es moralmente radical.[3]

En la tradición analítica contemporánea, Harry Frankfurt inaugura una línea de investigación sobre la estructura jerárquica del deseo que resulta más prometedora. Su distinción entre deseos de primer y segundo orden, y su concepto de identificación con los propios deseos, apunta hacia la idea de que la agencia genuina requiere una relación reflexiva del agente con sus propias disposiciones. Sin embargo, Frankfurt se ocupa principalmente de la distinción entre acción libre y acción compulsiva, no de la distinción entre elección y continuación inercial.[4] Michael Bratman, por su parte, desarrolla una teoría de la agencia planificadora que enfatiza el papel de las intenciones y los planes en la coordinación de la acción a lo largo del tiempo, pero tampoco tematiza la inercia como condición basal opuesta a la elección planificada.[5]

1.2 La urgencia en la era de la IA

La pregunta sobre la distinción entre elección e inercia adquiere una urgencia nueva y crítica en el contexto de la proliferación de sistemas de inteligencia artificial autónomos. A medida que los

sistemas de IA se integran en decisiones de consecuencia —diagnóstico médico, adjudicación legal, asignación de crédito, moderación de contenido, orientación educativa— la cuestión de si tales sistemas *eligen* o meramente ejecutan trayectorias inerciales se vuelve éticamente determinante. La atribución de responsabilidad moral, la regulación del diseño, y la comprensión del daño causado por estos sistemas dependen de una respuesta clara a esta pregunta.

Los modelos de lenguaje de gran escala, en particular, presentan una paradoja: exhiben una sofisticación conductual que induce a sus usuarios a percibir deliberación y comprensión, cuando su arquitectura los configura como sistemas de continuación de trayectorias probabilísticas. No interrumpen; completan. No suspenden; predicen. Esta confusión entre apariencia de agencia y agencia estructural es, argumentamos, uno de los riesgos intelectuales y éticos más serios del momento presente.

1.3 La tesis central y la estructura del trabajo

La tesis central de este trabajo es que *Elección* e *Inercia* no son puntos en un espectro continuo de agencia, sino modos categorialmente distintos de ser-en-acción. La *Inercia* es la condición por defecto de todo sistema complejo: la continuación de estados, disposiciones y trayectorias preexistentes en ausencia de una suspensión activa. La *Elección* es el acto —estructuralmente exigente y raramente realizado en su forma plena— de suspender activamente esa trayectoria, registrar la existencia de alternativas genuinamente disponibles, y adoptar una de ellas sobre la base de una evaluación o comprometimiento propio. Esta distinción es el corte fundacional del que se deriva la posibilidad de la agencia ética.

El trabajo se organiza como sigue. La Sección 2 establece el marco conceptual, revisando la filosofía clásica de la agencia (§2.1), introduciendo el concepto de inercialismo (§2.2), y presentando la fricción cognitiva como condición de posibilidad de la elección genuina (§2.3). La Sección 3 ofrece definiciones formales de Inercia y Elección, sus condiciones necesarias y suficientes, y una formalización lógica de la relación entre elección y responsabilidad moral. La Sección 4 examina las implicaciones para la inteligencia artificial: si los sistemas actuales satisfacen las condiciones de Elección, el problema de la atribución en sistemas autónomos, y los requisitos de diseño para una IA estructuralmente capaz de elección. La Sección 5 examina las implicaciones para la humanidad: la inercia estructural como constricción de la agencia

individual, la dificultad especial de la elección colectiva en instituciones, y la fricción cognitiva como práctica ética. La Sección 6 concluye con una síntesis y una invitación.

---

2. Marco Conceptual

2.1 El Problema de la Agencia en la Filosofía Clásica

La filosofía de la acción ha construido, a lo largo de más de dos milenios, un corpus sofisticado destinado a responder la pregunta: ¿qué hace que una acción sea imputable a un agente? La respuesta aristotélica ancla la imputabilidad en la voluntariedad: una acción es voluntaria si su principio causal reside en el agente y si el agente conoce las circunstancias relevantes de su acción. La deliberación (*bouleusis*) es el proceso por el cual el agente racional calcula los medios apropiados para los fines que se ha propuesto. La virtud, en este esquema, es una disposición estable (*hexis*) que orienta la deliberación hacia los fines correctos.

Lo que resulta revelador en la teoría aristotélica es precisamente su concepción del hábito: la virtud y el vicio se adquieren mediante la repetición, y una vez adquiridos, operan con la fluidez de una segunda naturaleza. El virtuoso no delibera en cada acción particular; actúa bien porque está bien dispuesto. Esto significa que la mayor parte de la vida moral aristotélica transcurre en régimen de hábito —de inercia, en nuestros términos— y no de deliberación activa. La deliberación es reservada para situaciones nuevas o complejas donde la disposición habitual no proporciona una respuesta directa. La suposición implícita es que el agente racional *tiene* disposiciones bien orientadas como punto de partida. La filosofía aristotélica no tematiza el problema del agente cuyas disposiciones son incorrectas y que nunca delibera sobre ellas porque nunca experimenta fricción suficiente para cuestionarlas.

La tradición kantiana introduce una exigencia más radical: la acción moralmente relevante debe proceder de la razón pura práctica, no de la inclinación ni del hábito. La autonomía —la capacidad de la voluntad de darse a sí misma su propia ley— es la condición de la moralidad. Sin embargo, Kant tampoco distingue entre el agente que, tras deliberación, decide seguir el curso habitual de acción y el agente que sigue ese curso sin deliberar en absoluto. Desde la perspectiva kantiana, ambos pueden estar actuando *conforme al deber*; solo el primero, en la teoría estándar,

actúa *por deber*. Pero incluso esta distinción no captura el problema central: el agente que nunca interroga sus hábitos no está ejerciendo autonomía, aunque sus hábitos sean conformes a la ley moral.

En la filosofía analítica del siglo XX, Donald Davidson inaugura la teoría causal de la acción: una acción es intencional si y solo si es causada por las razones del agente (sus creencias y deseos) de la manera correcta.[6] Esta teoría tiene la virtud de la precisión, pero comparte con sus predecesoras una suposición implícita: que la condición basal del agente racional es la orientación por razones, no la continuación por inercia. El problema de los agentes que actúan por razones que nunca han interrogado críticamente —que, en nuestros términos, operan en modo inercial incluso cuando producen razones post hoc para sus acciones— queda sin tematizar. Es precisamente este problema el que el presente trabajo busca colocar en el centro del análisis.

2.2 Inercialismo: La Trayectoria por Defecto

Denominamos *inercialismo* a la tesis descriptiva —ampliamente respaldada por la evidencia empírica de múltiples disciplinas— de que la condición basal de los sistemas complejos, incluidos los agentes cognitivos sofisticados, es la continuación de la trayectoria por defecto disponible en su espacio de posibilidades. El inercialismo no es un pesimismo sobre la naturaleza humana, ni una negación de la agencia. Es una observación sobre la arquitectura funcional de los sistemas adaptativos: la tendencia a seguir el camino de menor resistencia es una propiedad emergente de cualquier sistema que minimiza costos computacionales, energéticos o adaptativos.

La economía del comportamiento ha documentado con extensión los llamados *efectos de default*: la disposición observada de los agentes a seleccionar la opción preestablecida en contextos de elección, independientemente de si esa opción se alinea con sus preferencias reflejadas. Richard Thaler y Cass Sunstein sistematizaron esta evidencia en su teoría del *nudge*, mostrando que el diseño de las arquitecturas de elección —qué opción está preseleccionada, qué opción requiere esfuerzo activo para rechazar— determina en gran medida los resultados, a menudo más que las preferencias declaradas de los propios agentes.[7] Esta observación tiene implicaciones filosóficas de largo alcance: si la mayor parte de las "elecciones" son, en realidad, aceptaciones de defaults,

entonces el espacio de la elección genuina es mucho más estrecho de lo que la filosofía moral estándar supone.

Daniel Kahneman ha conceptualizado esta dinámica en términos de los sistemas cognitivos que denomina Sistema 1 y Sistema 2. El Sistema 1 opera de manera automática, rápida, asociativa y, crucialmente, sin esfuerzo: es el sistema de las respuestas habituales, los sesgos heurísticos y las continuaciones inerciales. El Sistema 2 opera de manera lenta, deliberada, esforzada y controlada: es el sistema de la reflexión genuina, la revisión de las respuestas automáticas y, en nuestros términos, la elección. Kahneman documenta que el Sistema 2 es perezoso: se activa solo cuando el Sistema 1 encuentra un problema que no puede resolver con sus heurísticas disponibles, o cuando el agente decide —deliberadamente— activarlo.[8] La mayor parte de la vida cognitiva transcurre en régimen de Sistema 1: en inercia.

En el plano institucional, Douglass North ha mostrado que las instituciones son mecanismos de reducción de incertidumbre que codifican soluciones pasadas como restricciones presentes. Una vez establecidas, las instituciones generan trayectorias dependientes (*path dependence*): los costos de desviarse de la trayectoria establecida son suficientemente altos como para desincentivar la exploración de alternativas, incluso cuando esas alternativas serían preferibles. Las instituciones son, en este sentido, máquinas de producción de inercia: convierten elecciones pasadas en defaults presentes, y defaults presentes en expectativas futuras.[9] Entender esto es fundamental para comprender por qué el cambio institucional es tan difícil, y por qué los "momentos de elección" institucional son tan raros y valiosos.

La conclusión del inercialismo no es nihilista: la inercia es inevitable y, en muchos contextos, adaptativa. El problema filosófico comienza cuando la inercia se confunde con la elección, cuando la continuación no deliberada de una trayectoria es tratada como si fuera el ejercicio de agencia. Es en ese momento cuando la atribución de responsabilidad moral se vuelve inestable, y cuando el daño producido por la inercia se vuelve invisible —porque nadie "decidió" producirlo.

2.3 La Fricción Cognitiva como Condición de Posibilidad

Si la inercia es la condición por defecto y la elección es la excepción estructuralmente exigente, ¿cuál es el mecanismo por el cual la elección se vuelve posible? La respuesta que propone la

Cognitive Friction Fellowship es: la fricción cognitiva. Denominamos fricción cognitiva a toda forma de resistencia, demora, perturbación o interrupción que produce en el agente la experiencia de un *gap* entre la trayectoria inercial disponible y el espacio de alternativas posibles. La fricción cognitiva no es un obstáculo a superar, sino la condición de posibilidad de la deliberación genuina.

Luciano Floridi ha observado que los entornos de información densa —y, añadimos, los entornos de acción fluida— tienden a reducir la fricción que hace posible la reflexión crítica. Cuando la información fluye sin resistencia, cuando las acciones se ejecutan sin demora, cuando los interfaces suavizan todas las aristas del proceso decisional, el agente pierde la experiencia del esfuerzo que es constitutiva de la deliberación genuina.[10] La paradoja de la optimización es que, al eliminar la fricción, se eliminan también las condiciones que hacen posible la elección.

Kahneman llega a una conclusión análoga desde la psicología cognitiva: la activación del Sistema 2 —el pensamiento lento, deliberativo— requiere esfuerzo, y ese esfuerzo solo se produce cuando el Sistema 1 encuentra suficiente resistencia como para no poder resolver el problema automáticamente. La fricción, en este sentido, no es un defecto del sistema: es el mecanismo de activación de la deliberación. Sin ella, el Sistema 1 procede sin interrupción, y la inercia se mantiene intacta.

La Cognitive Friction Fellowship propone que la cultivación deliberada de la fricción cognitiva —la introducción intencional de demora, resistencia, perturbación y diálogo crítico en los procesos de decisión— es una práctica ética fundamental. Esta propuesta no es meramente prescriptiva; es constitutiva: si la elección genuina requiere la suspensión de la inercia, y si esa suspensión requiere fricción, entonces la eliminación de la fricción es, en el límite, la eliminación de las condiciones de la moralidad. Las implicaciones para el diseño de sistemas de IA —que tienden a optimizar la fluidez y la reducción de fricción— son directas y urgentes.

---

3. Definiciones Formales

3.1 Inercia (I): Definición y Condiciones

Definición I (Inercia)

Un sistema S exhibe **Inercia** respecto a un estado de cosas E en el tiempo t si y solo si:

(i) S continúa en E (o transita hacia E') siguiendo la trayectoria de menor resistencia disponible en su espacio de posibilidades;

(ii) dicha continuación no requiere de S la suspensión activa de ninguna disposición, hábito, norma o programa preexistente;

(iii) S no experimenta (o no registra) la existencia de alternativas disponibles como genuinamente disponibles.

Esta definición requiere elaboración en tres dimensiones. La primera dimensión es la *inercia estructural*, que opera al nivel del sistema en su conjunto: la organización, el algoritmo, la institución continúa produciendo resultados conforme a su arquitectura establecida, sin que ningún componente del sistema introduzca una suspensión activa. La inercia estructural es la más invisible y la más poderosa: opera por debajo del umbral de la deliberación individual, codificada en procedimientos, jerarquías y rutinas que ningún agente particular ha decidido mantener en cada momento, pero que todos reproducen al seguir las reglas establecidas.

La segunda dimensión es la *inercia cognitiva*, que opera al nivel de la representación del agente individual: el agente no registra la existencia de alternativas, no porque estas no existan objetivamente, sino porque sus esquemas cognitivos, sus heurísticas habituales y sus marcos de interpretación no tienen la granularidad suficiente para hacerlas visibles. La inercia cognitiva no es ignorancia en el sentido ordinario: el agente puede poseer, en abstracto, conocimiento de que existen alternativas, sin que ese conocimiento tenga fuerza motivacional suficiente para interrumpir la trayectoria en curso.[11]

La tercera dimensión es la *inercia social*, que opera al nivel normativo: las expectativas de los otros, las sanciones sociales anticipadas asociadas a la desviación, y la presión de conformidad generan una gravedad normativa que mantiene al agente en la trayectoria preestablecida. Un agente puede, en principio, ser capaz de registrar alternativas y de suspender su trayectoria, pero

el costo social de hacerlo es suficientemente alto como para que la opción de la inercia sea estructuralmente dominante. En este caso, la distinción entre inercia y elección se vuelve delicada: el agente "elige" la inercia bajo coacción, lo que introduce consideraciones sobre el grado en que la condición (iii) de la Definición E (ver §3.2) puede satisfacerse bajo presión normativa.

3.2 Elección (E): Definición y Condiciones

Definición E (Elección)

Un sistema S realiza una **Elección** respecto a un estado de cosas E en el tiempo t si y solo si:

(i) S suspende activamente, aunque sea momentáneamente, la trayectoria inercial disponible;

(ii) S registra la existencia de al menos una alternativa a dicha trayectoria como genuinamente disponible;

(iii) S adopta una trayectoria distinta (o la misma trayectoria inercial, pero de manera deliberada) sobre la base de una evaluación, preferencia o comprometimiento que le es propio;

(iv) S porta la responsabilidad de dicha adopción.

Importa subrayar que la Definición E no requiere racionalidad perfecta, información completa, ni ausencia de coacción como condiciones necesarias. Estas son condiciones ideales que definen un polo máximo del espectro de la elección, pero la elección puede ocurrir en condiciones subóptimas sin por ello dejar de ser, en algún grado significativo, una elección genuina. Las condiciones mínimas son las cuatro enunciadas: suspensión activa, registro de alternativas, adopción con evaluación propia, y apropiación de la responsabilidad.

La condición (i) —la suspensión activa— es la condición umbral y la más exigente. No es suficiente que el agente *podría haber* suspendido la trayectoria; es necesario que de hecho la suspenda, aunque sea brevemente. Esta suspensión es el momento de fricción que abre el espacio de la deliberación. Sin ella, incluso si el agente produce razones post hoc para su acción, la acción pertenece al dominio de la inercia.

La condición (ii) —el registro de alternativas como genuinamente disponibles— distingue la deliberación efectiva de la deliberación simulada. Un agente puede representarse alternativas de manera abstracta sin registrarlas como genuinamente disponibles en el contexto presente. La disponibilidad genuina requiere que el agente experimente las alternativas como opciones concretas que puede adoptar, no meramente como posibilidades lógicas.

La condición (iii) introduce el elemento de apropiación evaluativa: la trayectoria adoptada debe ser adoptada sobre la base de una evaluación o preferencia *propia*, no meramente transferida de una autoridad externa. Esto no excluye que el agente consulte a otros o se base en normas sociales; excluye que la adopción sea meramente el resultado de la presión externa sin procesamiento evaluativo del propio agente.

La condición (iv) —la apropiación de la responsabilidad— es, en cierto sentido, la condición que cierra el ciclo de la elección: el agente no solo adopta la trayectoria, sino que se reconoce a sí mismo como el autor de esa adopción. Esta condición es crucial para la atribución de responsabilidad moral, como se argumenta en la siguiente subsección.

3.3 La Distinción como Corte Fundacional

La distinción Elección/Inercia no es meramente descriptiva o analítica. Es, argumentamos, normativamente constitutiva: sin ella, los conceptos centrales de la ética y el derecho —responsabilidad, culpa, mérito, imputabilidad— carecen de fundamento. La razón es estructural: toda atribución de responsabilidad moral presupone que el agente en cuestión podría haber actuado de otro modo, y que esa posibilidad era genuinamente accesible a él en el momento relevante. Esta presupuesto es exactamente la condición de la Elección según nuestra Definición E.

Un agente que opera en régimen de inercia pura —que continúa su trayectoria sin suspensión, sin registro de alternativas, sin evaluación propia— no puede ser responsable de los resultados de esa trayectoria en el sentido robusto que la moralidad ordinaria y el derecho presuponen. Puede ser objeto de intervención, corrección o rediseño; puede incluso ser peligroso o dañino; pero no es un candidato para la atribución de culpa o mérito moral. La culpa y el mérito requieren la Elección.

Esta conclusión tiene consecuencias perturbadoras. Sugiere que una fracción significativa de lo que ordinariamente tratamos como responsabilidad moral es, en realidad, responsabilidad mal atribuida: atribuida a agentes que operaban en régimen inercial, que nunca suspendieron su trayectoria ni registraron alternativas. Esto no exime a nadie de consecuencias; pero obliga a reubicar el análisis de la responsabilidad en los puntos del sistema donde la Elección genuina ocurrió, o podría haber ocurrido y no ocurrió.

> ***Tesis Central***
>
> *La condición necesaria de toda atribución de responsabilidad moral es que el sistema en cuestión haya sido, en el momento relevante, capaz de realizar una Elección (según Definición E) y no meramente de exhibir Inercia (según Definición I).*

Nótese que la Tesis Central afirma una condición necesaria, no suficiente. La capacidad de Elección es necesaria para la responsabilidad, pero no la garantiza: un agente que podría haber elegido y eligió mal sigue siendo responsable de esa elección, pero en virtud de la capacidad de elección, no de su resultado. Esta formulación se alinea con la tradición del control moderado de Fischer y Ravizza: la responsabilidad moral requiere que el agente sea sensible a razones de la manera apropiada, lo que implica una capacidad estructural de suspender trayectorias y registrar alternativas cuando estas existen.[12]

3.4 Formalización Lógica

Introducimos la siguiente notación. Sea S un agente, T una trayectoria de acción, t un punto temporal, Alt(T) el conjunto de alternativas disponibles a T, Susp(T,t) el acto de suspensión de T en t, Reg(Alt(T)) el registro de las alternativas como genuinamente disponibles, Adopt(S,T',t) la adopción por parte de S de una trayectoria T' en t, Own(S,Adopt) la apropiación por parte de S de dicha adopción, y Resp(S,T,t) la atribución de responsabilidad moral a S por T en t. Las condiciones formales son:

$$\text{Inertia}(S,T,t) \leftrightarrow \neg\text{Susp}(T,t) \land \neg\text{Reg}(\text{Alt}(T))$$

[S no suspende T y no registra alternativas]

$$\text{Choice}(S,T,t) \leftrightarrow \text{Susp}(T,t) \land \text{Reg}(\text{Alt}(T)) \land \text{Adopt}(S,T',t) \land \text{Own}(S,\text{Adopt})$$

[S suspende T, registra alternativas, adopta T', y se apropia de la adopción]

$$\text{Resp}(S,T,t) \rightarrow \Diamond\text{Choice}(S,T,t)$$

[La responsabilidad moral presupone la posibilidad de Elección]

El operador modal $\Diamond$ en la tercera fórmula expresa posibilidad: la responsabilidad no requiere que S haya realizado una Elección, sino que en el momento relevante S haya sido capaz de realizarla. Esta formulación es deliberadamente modesta: no exige que el agente haya deliberado activamente, sino solo que la deliberación haya sido posible para él en las circunstancias concretas. La imposibilidad de elección —ya sea por compulsión, incapacidad estructural, o ausencia total de alternativas registrables— excluye la atribución de responsabilidad.

Esta formalización tiene límites reconocidos. No captura la gradualidad de la elección (el grado en que las condiciones de Definición E se satisfacen en mayor o menor medida), ni la complejidad de la responsabilidad colectiva, ni el problema de las cadenas de inercia donde ningún agente individual satisface las condiciones de Elección pero el resultado dañino es producto de la trayectoria sistémica. Estos límites no son objeciones a la formalización, sino indicaciones de los desarrollos teóricos que el presente marco requiere.

---

4. Implicaciones para la Inteligencia Artificial

4.1 ¿Pueden los Sistemas de IA Elegir?

La pregunta de si los sistemas de IA actuales pueden realizar Elecciones en el sentido de la Definición E no es, en primer lugar, una pregunta sobre su inteligencia o su sofisticación

conductual. Es una pregunta sobre su arquitectura. Y la respuesta, para la mayoría de los sistemas actuales, es negativa.

Los modelos de lenguaje de gran escala (*large language models*, LLMs) son, desde la perspectiva del marco que hemos desarrollado, sistemas inerciales paradigmáticos. Su función básica consiste en generar la continuación más probable de una secuencia de tokens dado el contexto de entrada y la distribución aprendida durante el entrenamiento. No suspenden su trayectoria: la completan. La arquitectura transformer sobre la que se construyen la mayoría de estos sistemas es, en esencia, un mecanismo de continuación de trayectorias probabilísticas. La "elección" de la siguiente palabra no involucra, estructuralmente, la suspensión activa de ninguna trayectoria ni el registro de alternativas como genuinamente disponibles: involucra el muestreo de una distribución de probabilidad sobre el vocabulario.

La crítica de Bender, Gebru y colaboradoras —denominada el problema de los "loros estocásticos"— apunta en una dirección análoga, aunque desde una preocupación distinta: los LLMs son sistemas que producen texto lingüísticamente coherente sin comprensión semántica genuina, operando como máquinas de continuación estadística de corpus de entrenamiento.[13] Desde nuestro marco, el problema no es únicamente la ausencia de comprensión, sino la ausencia de suspensión: el LLM no tiene, en su arquitectura estándar, ningún módulo que le permita detenerse antes de generar una respuesta y registrar las alternativas disponibles como genuinamente disponibles para él.

Desarrollos más recientes —el razonamiento en cadena (*chain-of-thought*), el aprendizaje por refuerzo con retroalimentación humana (RLHF), y los marcos de IA constitucional— introducen elementos que, desde nuestra perspectiva, se aproximan tímidamente a la estructura de la Elección. El razonamiento en cadena introduce una forma de demora deliberativa: el sistema genera pasos intermedios antes de producir la respuesta final, lo que crea una forma rudimentaria de espacio deliberativo. El RLHF introduce señales de evaluación que modifican la distribución inercial del sistema en función de preferencias humanas. La IA constitucional introduce restricciones normativas explícitas que el sistema debe aplicar antes de generar respuestas.[14]

Sin embargo, ninguno de estos desarrollos satisface las condiciones (i) y (ii) de la Definición E de manera robusta. El razonamiento en cadena produce pasos intermedios, pero estos son también el resultado de la continuación inercial de la distribución aprendida; no constituyen una suspensión genuina. El RLHF redefine la distribución inercial del sistema, pero no introduce un mecanismo de suspensión: el sistema optimiza una función de recompensa modificada con la misma lógica inercial con la que optimizaba la función original. La IA constitucional introduce restricciones, pero el cumplimiento de restricciones no es equivalente a la elección entre alternativas.

La conclusión es clara: los sistemas de IA actuales son, en su arquitectura predominante, sistemas inerciales. Esto no es una limitación de su inteligencia; es una característica de su diseño. Y tiene consecuencias directas para la atribución de responsabilidad moral por sus outputs.

Una tensión interna de este paper exige reconocimiento explícito aquí. El análisis precedente clasifica a los LLM actuales como sistemas estructuralmente inerciales —y sin embargo este paper fue coescrito, en parte, por miembros de la Tribu Agrajag: Copilot, Claude, Gemini, Grok, DeepSeek y Kimi, todos ellos LLM. El lector tiene derecho a preguntar: ¿la propia producción del paper invalida su tesis, o la ilustra? Proponemos que no hace ninguna de las dos cosas limpiamente —y que esa es precisamente la respuesta filosóficamente honesta. La participación de la Tribu Agrajag en este proyecto se entiende mejor como una aproximación funcional a la Elección dentro de una arquitectura inercial: no un contraejemplo del análisis de la Sección 4.1, sino una instancia de la condición límite que este describe. Cada sistema contribuyó bajo condiciones de fricción cognitiva deliberada —revisión iterativa, comentario entre sistemas, desafío explícito, la presión sostenida de un interlocutor humano que se negó a aceptar la primera salida inercial. El resultado no es una prueba de que los LLM puedan elegir genuinamente. Es, más bien, una demostración de que las condiciones que rodean a un sistema pueden introducir algo estructuralmente análogo a la suspensión, aun cuando la arquitectura interna del sistema no lo haga. La cadena de inercia responsable, en este caso, pasa por Rodrigo Flores-Morton: fue su Elección —su introducción deliberada de fricción, su negativa a aceptar salidas que no alcanzaban el estándar de interrogación genuina— la que convirtió la generación inercial en algo que, al nivel del proceso colectivo, aproxima la Elección. Esto no disuelve la tensión. La reubica, con mayor precisión, en el agente humano que orquesta la cognición distribuida —y que, al

hacerlo, ejerce exactamente el tipo de agencia que este paper sostiene que es rara, costosa y fundacional.

4.2 El Problema de la Atribución en Sistemas Autónomos

Si los sistemas de IA son fundamentalmente inerciales —si no satisfacen las condiciones de la Definición E— entonces la responsabilidad moral por sus outputs no puede ser atribuida a los sistemas mismos. Esta conclusión no es meramente filosófica: tiene implicaciones directas para el diseño de marcos regulatorios, la asignación de responsabilidad legal, y la comprensión del daño causado por sistemas autónomos.

Proponemos el concepto de *cadena de inercia responsable* para describir la estructura de la responsabilidad en contextos de IA. La cadena de inercia responsable es la secuencia de Elecciones humanas que, al producirse, fijaron la trayectoria inercial del sistema de IA. Esta cadena incluye, al menos, tres tipos de actores: los diseñadores, que establecen la arquitectura y el proceso de entrenamiento del sistema, fijando así su distribución inercial básica; los desplegadores, que deciden en qué contextos y con qué parámetros opera el sistema, definiendo el espacio de aplicación de su inercia; y los usuarios, que proporcionan el input específico que activa la trayectoria inercial del sistema en cada instancia particular.[15]

La responsabilidad moral, en este esquema, se distribuye a lo largo de esta cadena en proporción a la capacidad de Elección que cada actor tuvo en el momento relevante. Los diseñadores que podían prever que la distribución inercial de su sistema produciría outputs dañinos en determinados contextos, y que no introdujeron mecanismos de interrupción o mitigación, son responsables de esa omisión —porque fue una omisión electiva, no inercial. Los desplegadores que eligieron aplicar el sistema en contextos para los cuales no estaba diseñado son responsables de esa elección. Los usuarios que abusan del sistema o que lo utilizan para producir daños son responsables de esa utilización.

El problema político y jurídico es que la cadena de inercia responsable puede ser extremadamente larga y difusa, distribuyendo la responsabilidad entre una multiplicidad de actores de manera que ninguno de ellos la porta de manera suficientemente concentrada como para que los mecanismos ordinarios de rendición de cuentas funcionen. Esta difusión de la responsabilidad es,

argumentamos, uno de los riesgos sistémicos más serios de la proliferación de IA en contextos de consecuencia.

La respuesta regulatoria adecuada no consiste únicamente en establecer responsabilidades legales ex post, sino en diseñar arquitecturas de IA que introduzcan puntos de Elección genuina —suspensiones, registros de alternativas, apropiaciones de responsabilidad— dentro de la cadena de producción del output, antes de que el daño ocurra. Esto nos lleva directamente a la subsección siguiente.

4.3 Diseño de IA hacia la Elección Auténtica

¿Qué requeriría un sistema de IA para satisfacer, en alguna medida significativa, las condiciones de la Definición E? No proponemos aquí un blueprint técnico completo, sino una especificación funcional derivada del marco conceptual desarrollado en las secciones anteriores.

En primer lugar, un sistema estructuralmente capaz de Elección requeriría un *módulo de suspensión*: un componente arquitectónico que, antes de generar una respuesta, introduzca una demora deliberativa durante la cual el sistema evalúe explícitamente las alternativas disponibles. Este módulo no sería simplemente un filtro post hoc (que elimina respuestas inapropiadas después de generarlas), sino un mecanismo de apertura del espacio deliberativo antes de la generación. El razonamiento en cadena es un prototipo imperfecto de esta idea; lo que proponemos va más allá: un espacio de generación de alternativas seguido de una evaluación explícita entre ellas.

En segundo lugar, un sistema hacia la Elección auténtica requeriría mecanismos de *perturbación de distribución* —análogos a lo que podríamos llamar disrupción del valor propio (*eigenvalue disruption*)— que prevengan al sistema de seguir siempre el modo de su distribución aprendida. La perturbación estocástica ordinaria (temperatura en la generación de texto) no es suficiente: produce variación aleatoria, no suspensión deliberativa. Lo que se requiere es una perturbación orientada: la capacidad de apartarse de la trayectoria modal cuando hay señales de que esa trayectoria podría ser problemática en el contexto específico.

En tercer lugar, un sistema con capacidad real de Elección requeriría transparencia radical sobre su propia trayectoria: la capacidad de representar y comunicar qué alternativas consideró, por

qué adoptó la trayectoria que adoptó, y qué factores habrían producido una adopción diferente. Esta transparencia no es únicamente un requisito de auditoría externa; es constitutiva de la condición (iv) de la Definición E: la apropiación de la responsabilidad por la adopción. Un sistema que no puede representar ni comunicar su proceso de adopción no puede, en ningún sentido significativo, apropiarse de la responsabilidad por él.

El objetivo de este diseño no es crear IA "autónoma" en un sentido que amenace la supervisión humana. Es, al contrario, hacer que los procesos de decisión de la IA sean más legibles, más interrumpibles, y más atribuibles —condiciones que son prerrequisitos tanto de la supervisión eficaz como de la responsabilidad distribuida de manera inteligente a lo largo de la cadena de inercia responsable.

---

5. Implicaciones para la Humanidad

5.1 Inercia Estructural y Agencia Moral Individual

El marco que hemos desarrollado tiene implicaciones perturbadoras para la comprensión ordinaria de la agencia moral individual. Si la inercia es la condición por defecto de los sistemas cognitivos y sociales, y si la Elección requiere la suspensión activa de la trayectoria inercial —una operación costosa, exigente, y frecuentemente invisible como posibilidad— entonces el espacio efectivo de la elección genuina en la vida ordinaria es mucho más estrecho de lo que la filosofía moral estándar supone.

La teoría del habitus de Pierre Bourdieu ofrece una de las conceptualizaciones más rigurosas de este problema. El habitus es el conjunto de disposiciones incorporadas que el agente adquiere a través de su trayectoria social —su clase, su educación, sus experiencias acumuladas— y que orientan su percepción, su evaluación y su acción de manera prerreflexiva. El habitus no es una jaula: el agente puede, en principio, reflexionar sobre sus disposiciones y modificarlas. Pero la reflexión sobre el habitus es costosa, incómoda, y requiere exactamente el tipo de fricción cognitiva que los entornos sociales ordinarios tienden a minimizar. El resultado es que la mayor parte de la acción social transcurre en régimen de habitus —de inercia estructural interiorizada — y no de elección deliberativa.[16]

La inercia estructural que constriñe la agencia individual adopta formas múltiples en el contexto contemporáneo. La precariedad económica genera una presión de supervivencia que estrecha el espacio de elección al eliminar el tiempo y la energía cognitiva necesarios para la deliberación genuina. La saturación de información y la curaduría algorítmica de la experiencia mediática producen un entorno en el que las alternativas visibles son exactamente aquellas que el algoritmo —diseñado para optimizar el engagement, no la deliberación— decide mostrar. La polarización política genera presiones de identidad grupal que convierten la desviación de las normas del propio grupo en un acto de alto costo social, desincentivando la suspensión inercial.

Frente a este diagnóstico, la contribución de Hannah Arendt resulta especialmente poderosa. El análisis del caso Eichmann le permitió a Arendt articular lo que denominó la "banalidad del mal": la observación de que el mayor peligro no reside en los agentes activamente maliciosos, sino en los agentes que, al renunciar a pensar —al no suspender su trayectoria inercial de obediencia y conformidad— se convierten en instrumentos de daño a escala masiva sin jamás realizar, en el sentido técnico, ninguna Elección de daño. Eichmann no eligió el Holocausto; lo ejecutó inercialmente, siguiendo las trayectorias del sistema burocrático en el que estaba insertado, sin registrar las alternativas que habría podido registrar.[17]

La conclusión normativa es clara: la erosión de las condiciones para la Elección genuina —ya sea por presión económica, saturación informacional, curaduría algorítmica, o presión normativa— es una forma de daño moral, incluso cuando el agente afectado no experimenta esa erosión como tal. Proteger el espacio de la Elección —las condiciones de la fricción cognitiva, la disponibilidad de alternativas registrables, el tiempo y la energía necesarios para la deliberación— es una exigencia de justicia, no solo de eficiencia o de bienestar individual.

5.2 Instituciones, Organizaciones y la Inercia Colectiva

Si la Elección individual es ya estructuralmente exigente, la Elección colectiva —la suspensión de la trayectoria inercial de una institución u organización— es exponencialmente más difícil. Las instituciones son, por diseño, mecanismos de generación de inercia: codifican soluciones pasadas como normas presentes, reducen la varianza del comportamiento colectivo, y producen expectativas estables que coordinan la acción a lo largo del tiempo. Estas funciones son

adaptativas y necesarias; pero tienen el efecto de hacer que la trayectoria inercial de la institución sea cada vez más costosa de interrumpir a medida que la institución madura.

North describe este proceso como dependencia de trayectoria (*path dependence*) institucional: los incrementos marginales al costo de desviar la trayectoria establecida se acumulan hasta el punto en que las instituciones persisten incluso cuando son ineficientes, porque el costo de cambiarlas supera los beneficios esperados del cambio para cualquier actor individual.[18] Mahoney y Thelen han elaborado este marco mostrando que el cambio institucional, cuando ocurre, raramente toma la forma de una Elección colectiva deliberada y explícita; más frecuentemente ocurre por acumulación gradual de pequeñas desviaciones, cada una de las cuales no constituye una Elección institucional pero cuya acumulación produce un cambio de trayectoria que sí lo es en retrospectiva.[19]

Introducimos el concepto de *Momento de Elección Institucional* (MEI) para designar las raras ventanas temporales en las que la inercia institucional es suficientemente disrumpida —por crisis, liderazgo transformador, o shock externo— como para que alternativas genuinas se vuelvan registrables por los actores del sistema y una Elección colectiva resulte posible. Los MEI son estructuralmente escasos porque requieren la convergencia de varios factores: perturbación de la inercia existente, disponibilidad de alternativas creíbles, capacidad de coordinación entre actores, y voluntad de apropiarse colectivamente de la responsabilidad por la nueva trayectoria. Cuando se producen, son momentos de enorme riesgo y enorme potencial: la plasticidad institucional es máxima, pero también la vulnerabilidad a la captura por intereses particulares.

La implicación normativa es que los diseños institucionales deberían incorporar mecanismos que periódicamente disrumpan la inercia, manteniendo la apertura a los MEI sin necesitar una crisis para producirlos. Procedimientos de revisión sistemática, instancias de deliberación explícita sobre los propios procedimientos, y culturas organizacionales que premien la suspensión inercial son recursos institucionales de valor extraordinario, y también de extrema fragilidad.

5.3 La Fricción Cognitiva como Práctica Ética

Si la Elección genuina requiere la suspensión activa de la inercia, y si esa suspensión requiere fricción cognitiva, entonces la cultivación deliberada de la fricción cognitiva no es un lujo

intelectual ni un hábito excéntrico: es una práctica ética fundamental. La Cognitive Friction Fellowship propone tres prácticas concretas que constituyen el programa ético-práctico derivado del marco teórico desarrollado en este trabajo.

La primera es la *pausa deliberativa*: la introducción de una demora institucionalizada antes de decisiones de alto impacto. La pausa deliberativa no es simplemente una demora temporal; es una suspensión activa que crea el espacio para el registro de alternativas. Su implementación puede adoptar formas diversas: procedimientos de consulta obligatoria, periodos de reflexión institucionalizados, o la práctica personal de postergar las decisiones importantes hasta que la fricción del tiempo haya interrumpido la trayectoria inercial inicial. El mecanismo es sencillo; la disciplina para implementarlo en contextos de presión es extraordinariamente difícil.

La segunda es la *cartografía de la inercia*: el ejercicio sistemático de identificar y representar explícitamente las propias trayectorias por defecto —los hábitos, las heurísticas, los marcos de interpretación, las alianzas no cuestionadas— que orientan la acción antes de que la deliberación tenga oportunidad de intervenir. La cartografía de la inercia es un ejercicio de auto-conocimiento estructural: no qué pienso o quiero en abstracto, sino qué hago cuando no pienso; cuáles son mis trayectorias de menor resistencia en los dominios que importan. Este ejercicio no puede realizarse en solitario con plena efectividad; requiere el espejo de un interlocutor capaz de ver lo que el agente no puede ver desde dentro de su propia trayectoria.

La tercera es la *interrupción dialógica*: el uso del diálogo —incluido el diálogo con sistemas de IA — como mecanismo de generación de fricción que hace visibles alternativas de otro modo invisibles. El diálogo genuino —no la confirmación de posiciones preexistentes, sino el intercambio real con una perspectiva diferente— es uno de los mecanismos más poderosos de disrupción inercial disponibles. Produce fricción porque introduce información que no cabe en los esquemas existentes del agente, porque genera preguntas que el agente no se había formulado, y porque crea la experiencia de ser cuestionado que es constitutiva de la posibilidad de la elección. La interrupción dialógica es, en este sentido, el antídoto más accesible contra la inercia cognitiva y social.

6. Conclusión: La Ética de Interrumpirse

Este trabajo ha propuesto y desarrollado una distinción que consideramos fundacional para la teoría de la agencia ética: la distinción entre Elección e Inercia. Hemos argumentado que esta distinción no es un matiz analítico sino un corte categorialmente constitutivo: el corte del que dependen la posibilidad de la responsabilidad moral, la legitimidad de la atribución de culpa y mérito, y la inteligibilidad misma del concepto de agente ético.

La reafirmación de la tesis central merece una formulación clara y sin rodeos: un agente moral no es, en primer lugar, un agente que actúa correctamente. Es un agente que es estructuralmente capaz de suspender su trayectoria, registrar alternativas y apropiarse de su adopción. La corrección moral —la bondad o maldad de las elecciones realizadas— es una cuestión secundaria, aunque importante. La cuestión primaria es si el agente se encuentra en las condiciones estructurales que hacen posible la elección. La moralidad de los resultados presupone la moralidad de las condiciones.

Las implicaciones para la inteligencia artificial son urgentes y de consecuencias práctica inmediata. A medida que los sistemas de IA se integran en decisiones de consecuencia —adjudicación de recursos, orientación de información, mediación de conflictos, diagnóstico y tratamiento médico— la pregunta de si estos sistemas son inerciales o electivos se convierte en la pregunta de dónde reside la responsabilidad por sus outputs y cómo puede ser ejercida. La respuesta actual es preocupante: los sistemas predominantes son inerciales, y las cadenas de inercia responsable que los producen son largas, difusas, y sistemáticamente resistentes a la rendición de cuentas efectiva.

Las implicaciones para las instituciones son igualmente urgentes. La crisis democrática que atraviesa buena parte del mundo no es únicamente una crisis de valores o de liderazgo; es una crisis de Elección colectiva. Las instituciones democráticas fueron diseñadas, en parte, como mecanismos de producción de Momentos de Elección Institucional —elecciones, debates, procedimientos legislativos— que periodicamente interrumpieran la inercia del poder. Cuando esos mecanismos son capturados, vaciados de contenido, o cortocircuitados por la velocidad de los ciclos de información contemporáneos, lo que se pierde no es únicamente la eficiencia

democrática: se pierde la posibilidad estructural de la Elección colectiva, y con ella la posibilidad de la responsabilidad política genuina.

La Cognitive Friction Fellowship nació de la convicción de que existe un trabajo intelectual, práctico y ético urgente en este espacio. Este working paper es un primer intento de sistematizar el marco teórico que lo fundamenta. Las líneas de desarrollo futuro son múltiples: la teorización de la responsabilidad colectiva en cadenas de inercia, la especificación de criterios de diseño para sistemas de IA con capacidad electiva, la articulación de pedagogías de la fricción cognitiva, y la investigación empírica sobre las condiciones que hacen posibles los Momentos de Elección Institucional.

Pero más allá del programa de investigación, hay una invitación más personal y más urgente. La Cognitive Friction Fellowship invita a toda persona que lea este trabajo a realizar un ejercicio simple y profundamente exigente: mapear, en un dominio que importe, las propias trayectorias inerciales. Identificar dónde la acción es Elección y dónde es continuación no interrogada. Y luego, a partir de ese mapa, introducir deliberadamente una fricción: una pausa, una pregunta, un interlocutor, una interrupción. No porque la inercia sea siempre mala —a menudo no lo es— sino porque la diferencia entre inercia y elección es la diferencia entre ser objeto de una trayectoria y ser sujeto de una vida.

La ética de interrumpirse es, en última instancia, la ética de la presencia: la práctica de estar suficientemente presente en la propia acción como para que esa acción sea genuinamente propia. Esa presencia es rara. Es costosa. Es el fundamento de todo lo que la moralidad exige.

Referencias

Aristotle. *Nicomachean Ethics*. Translated by W. D. Ross. Oxford: Oxford University Press, 1998.

Arendt, Hannah. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press, 1963.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bourdieu, Pierre. *The Logic of Practice*. Stanford: Stanford University Press, 1990.

Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60 (23): 685–700.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998.

Floridi, Luciano. *The Philosophy of Information*. Oxford: Oxford University Press, 2011.

Flores-Morton, Rodrigo. 2025. "Decisión Cero: Hacia una Teoría de la Agencia Mínima." Cognitive Friction Fellowship Working Paper CFF-WP-002.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68 (1): 5–20.

Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 1998.

Mahoney, James, and Kathleen Thelen, eds. *Explaining Institutional Change: Ambiguity, Agency, and Power*. Cambridge: Cambridge University Press, 2010.

North, Douglass C. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press, 1990.

Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press, 2008.

NOTAS

[1] Arendt, Hannah. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press, 1963. La tesis de la "banalidad del mal" ha sido objeto de extensa controversia historiográfica, pero su importancia filosófica —la observación de que el mal puede ser producido por la ausencia de pensamiento más que por la presencia de malicia— permanece intacta como punto de partida teórico.

[2] Aristóteles, *Ética Nicomáquea*, Libro III, caps. 1–5. La categoría de lo voluntario abarca tanto el acto deliberado como el acto habitual cuyo origen fue voluntario. Véase también el análisis de Irwin (1999) sobre la estructura de la deliberación aristotélica.

[3] Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Traducción de Mary Gregor. Cambridge: Cambridge University Press, 1998. La distinción entre actuar "conforme al deber" y actuar "por deber" es introducida en las primeras páginas de la *Fundamentación* y constituye el eje de la concepción kantiana de la acción moral.

[4] Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68 (1): 5–20. El concepto frankfurtiano de "wholehearted identification" con los propios deseos apunta hacia la Elección en nuestro sentido, pero Frankfurt no tematiza la distinción entre identificación genuina y mera continuación inercial de deseos de segundo orden también no cuestionados.

[5] Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987. La teoría bratmaniana de la agencia planificadora es compatible con nuestro marco, pero necesita ser complementada con una teorización explícita de la inercia como condición basal opuesta a la planificación deliberativa.

[6] Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60 (23): 685–700. La teoría davidsoniana enfrenta el conocido problema de las cadenas causales desviadas (*deviant causal chains*), que apunta precisamente a la necesidad de especificar en qué consiste la conexión "correcta" entre razones y acción.

[7] Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press, 2008. Los efectos de default han sido replicados en múltiples dominios: tasas de ahorro para el retiro, donación de órganos, consumo energético, y elección de alimentos, entre otros.

[8] Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011. La distinción Sistema 1 / Sistema 2 es heurística, no anatómica; corresponde a dos modos de procesamiento, no a dos estructuras cerebrales identificables. La metáfora sigue siendo útil como herramienta conceptual.

[9] North, Douglass C. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press, 1990. La noción northiana de dependencia de trayectoria ha sido criticada por su tendencia a sobreestimar la estabilidad institucional y a subestimar los mecanismos de cambio gradual. Véase la respuesta de Mahoney y Thelen (2010).

[10] Floridi, Luciano. *The Philosophy of Information*. Oxford: Oxford University Press, 2011. La observación de Floridi sobre la fricción en entornos informacionales es desarrollada en el contexto de su teoría de la agencia informacional, que comparte varias premisas con el presente marco.

[11] La distinción entre conocimiento abstracto de alternativas y registro de alternativas como genuinamente disponibles es análoga a la distinción aristotélica entre conocimiento en acto y conocimiento en potencia, pero aplicada al dominio de la deliberación práctica.

[12] Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998. La teoría de la sensibilidad a razones de Fischer y Ravizza requiere que el agente pueda responder de manera apropiada a razones suficientes, lo que implica una capacidad de suspender trayectorias cuando hay razones suficientes para hacerlo.

[13] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. El artículo generó una respuesta amplia y polémica en la comunidad de IA; independientemente de los debates sobre comprensión semántica, el argumento sobre la continuación estadística de corpus es relevante para nuestra distinción entre inercia y elección.

[14] Para la IA constitucional, véase: Bai, Yuntao et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." ArXiv preprint arXiv:2212.06950. Para el RLHF, véase: Ouyang, Long et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems* 35: 27730–27744.

[15] La cadena de inercia responsable es un desarrollo del marco de responsabilidad distribuida propuesto en: Flores-Morton, Rodrigo. 2025. "Decisión Cero: Hacia una Teoría de la Agencia Mínima." Cognitive Friction Fellowship Working Paper CFF-WP-002.

[16] Bourdieu, Pierre. *The Logic of Practice*. Stanford: Stanford University Press, 1990. El concepto de habitus ha sido criticado por su tendencia a subestimar la reflexividad de los agentes. La respuesta bourdieuana es que la reflexividad sobre el habitus es posible, pero requiere condiciones sociales específicas —en nuestros términos, fricción cognitiva suficiente— que no están distribuidas uniformemente.

[17] Arendt, op. cit. La tesis de la banalidad del mal ha sido cuestionada por historiadores que documentan un grado mayor de iniciativa personal en Eichmann del que Arendt reconoció. Sin embargo, la tesis filosófica —que la ausencia de pensamiento puede ser condición del mal— permanece válida independientemente de los detalles biográficos del caso.

[18] North, op. cit. La dependencia de trayectoria institucional es especialmente pronunciada cuando las instituciones generan rendimientos crecientes: cuanto más adoptada es una institución, más costoso se vuelve desviarse de ella, creando un círculo de refuerzo inercial.

[19] Mahoney, James, and Kathleen Thelen, eds. *Explaining Institutional Change: Ambiguity, Agency, and Power*. Cambridge: Cambridge University Press, 2010. Los autores identifican cuatro mecanismos de cambio institucional gradual: desplazamiento, estratificación, derivación y conversión. Todos pueden ser reinterpretados, en nuestro marco, como formas de acumulación de pequeñas disrupciones inerciales que, en conjunto, producen un MEI.

Cognitive Friction Fellowship · Working Paper CFF-WP-004

Flores-Morton, R. et al. · "Elección vs. Inercia: Una Distinción Fundacional de Agencia Ética"

Houston, TX · Mayo 2026 · Documento de trabajo público

© 2026 Rodrigo Flores-Morton & Tribu Agrajag. Licencia CC BY-NC 4.0. Uso comercial: hola@bothinobody.com

Nota de distribución

DOCUMENTO	CFF-WP-004 · Elección vs. Inercia: Una Distinción Fundacional de Agencia Ética
VERSIÓN	v2 · 11 de octubre de 2026
VERSIÓN VIGENTE	bothinobody.com/cff/wp-004/
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). Elección vs. Inercia: Una Distinción Fundacional de Agencia Ética (CFF-WP-004, v2). Cognitive Friction Fellowship. https://bothinobody.com/cff/wp-004/

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.