Cognitive Friction Fellowship

WORKING PAPER SERIES

No. 004

CHOICE VS. INERTIA

A Foundational Distinction in Ethical Agency

Lead Author:	**Rodrigo Flores-Morton** Cognitive Friction Fellowship | Houston, TX
Tribu Agrajag (AI Co-authors):	Copilot · Claude · Gemini · Grok · DeepSeek · Kimi

Series:	Working Papers in Cognitive Ethics & Artificial Agency
Number:	CFF-WP-004
Date:	May 2026
Status:	Public working paper

Circulation notice:

This document is a working draft subject to continuous revision. The arguments and formalizations presented represent a position under development. Critical comments addressed to the lead author are welcome. Do not cite without express permission.

EPIGRAPH

"Most men do not act: they react. Most institutions do not decide: they persist. The difference between a moral agent and a physical system does not lie in its complexity, but in its capacity to interrupt itself."

— Rodrigo Flores-Morton

Cognitive Friction Fellowship, 2026

Abstract [ES–EN]

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.

Abstract [EN]

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. Introduction: The Invisible Problem

There is a persistent illusion in moral philosophy and ordinary psychology: the illusion that agents —human beings, institutions, systems— are, by nature, deliberative. That when a person acts, they have chosen to act; that when an institution operates, it has decided to operate in that way; that when a system produces an outcome, that outcome is, in some meaningful sense, the fruit of an oriented agency. This assumption is, we argue, profoundly mistaken. The greater part of what is attributed to the intentional action of agents is, in reality, the silent continuation of pre-existing trajectories: sedimented habits, unquestioned institutional norms, internalized social expectations, algorithmic defaults. Not action, but *inertia*.

The invisible problem is not, then, active malice or deliberate ignorance. The invisible problem is the absence of choice where choice ought to be. It is the official who signs because he has always signed. It is the organization that excludes because it has always excluded. It is the language model that reproduces biases because it was trained on a corpus that contained them, and no architectural mechanism invited it to suspend that trajectory. Evil —when it exists— is not always a decision: it is often simply the result of having failed to decide to interrupt the dynamic in progress. This intuition was captured with terrible precision by Hannah Arendt in her analysis of the trial of Adolf Eichmann: the banality of evil does not reside in the perversity of the agent, but in the absence of thought, in the substitution of deliberation by inertial obedience to norms and procedures.[1]

1.1 The Philosophical Gap

The great traditions of the philosophy of action and ethics have addressed with rigor the question of what makes an action *intentional*, but they have systematically underestimated the distinction between deliberate action and mere inertial continuation. Aristotle, in the *Nicomachean Ethics*, establishes the distinction between voluntary action (*hekousion*) and involuntary action (*akousion*), and places deliberation (*bouleusis*) as the condition of moral virtue and vice. However, the Aristotelian category of the voluntary is broad enough to include acts performed out of unquestioned habit: if the habit was voluntarily acquired, the habitual action counts as voluntary. This collapses precisely the distinction we are here reclaiming: between active choice and habitual continuation.[2]

Kant, in the *Groundwork of the Metaphysics of Morals*, raises the bar by requiring that moral action proceed from autonomy, understood as rational self-legislation. The morally good will is one that acts from duty, determined by reason rather than inclination. But Kant does not thematize the distinction between the non-action that results from rational deliberation (the agent who, after reflection, decides not to change course) and the non-action that results from the absence of deliberation (the agent who never considered whether changing course was possible). Both may result in the same observable behavior; their difference is morally radical.[3]

In the contemporary analytic tradition, Harry Frankfurt inaugurates a more promising line of inquiry into the hierarchical structure of desire. His distinction between first- and second-order desires, and his concept of identification with one's own desires, points toward the idea that genuine agency requires a reflexive relation of the agent to their own dispositions. However, Frankfurt is primarily concerned with the distinction between free action and compulsive action, not with the distinction between choice and inertial continuation.[4] Michael Bratman, for his part, develops a theory of planning agency that emphasizes the role of intentions and plans in the coordination of action over time, but he likewise does not thematize inertia as the baseline condition opposed to planned choice.[5]

1.2 The Urgency in the Age of AI

The question of the distinction between choice and inertia acquires a new and critical urgency in the context of the proliferation of autonomous artificial intelligence systems. As AI systems become integrated into consequential decisions —medical diagnosis, legal adjudication, credit

assignment, content moderation, educational guidance— the question of whether such systems *choose* or merely execute inertial trajectories becomes ethically determinative. The attribution of moral responsibility, the regulation of design, and the understanding of harm caused by these systems all depend on a clear answer to this question.

Large language models, in particular, present a paradox: they exhibit a behavioral sophistication that leads their users to perceive deliberation and understanding, when their architecture configures them as systems of probabilistic trajectory continuation. They do not interrupt; they complete. They do not suspend; they predict. This confusion between the appearance of agency and structural agency is, we argue, one of the most serious intellectual and ethical risks of the present moment.

1.3 The Central Thesis and Structure of the Paper

The central thesis of this paper is that *Choice* and *Inertia* are not points on a continuous spectrum of agency, but categorically distinct modes of being-in-action. *Inertia* is the default condition of every complex system: the continuation of pre-existing states, dispositions, and trajectories in the absence of active suspension. *Choice* is the act —structurally demanding and rarely realized in its full form— of actively suspending that trajectory, registering the existence of genuinely available alternatives, and adopting one of them on the basis of one's own evaluation or commitment. This distinction is the foundational cut from which the possibility of ethical agency is derived.

The paper is organized as follows. Section 2 establishes the conceptual framework, reviewing classical philosophy of agency (§2.1), introducing the concept of inertialism (§2.2), and presenting Cognitive Friction as the condition of possibility of genuine choice (§2.3). Section 3 offers formal definitions of Inertia and Choice, their necessary and sufficient conditions, and a logical formalization of the relationship between choice and moral responsibility. Section 4 examines the implications for artificial intelligence: whether current systems satisfy the conditions of Choice, the problem of attribution in autonomous systems, and the design requirements for an AI structurally capable of choice. Section 5 examines the implications for humanity: structural inertia as a constraint on individual agency, the special difficulty of collective choice in institutions, and Cognitive Friction as an ethical practice. Section 6 concludes with a synthesis and an invitation.

2. Conceptual Framework

2.1 The Problem of Agency in Classical Philosophy

The philosophy of action has built, over more than two millennia, a sophisticated corpus aimed at answering the question: what makes an action attributable to an agent? The Aristotelian answer anchors attributability in voluntariness: an action is voluntary if its causal principle resides in the agent and if the agent knows the relevant circumstances of their action. Deliberation (*bouleusis*) is the process by which the rational agent calculates the appropriate means for the ends they have set for themselves. Virtue, in this scheme, is a stable disposition (*hexis*) that orients deliberation toward the correct ends.

What is revealing in Aristotelian theory is precisely its conception of habit: virtue and vice are acquired through repetition, and once acquired, operate with the fluency of a second nature. The virtuous person does not deliberate in each particular action; they act well because they are well disposed. This means that most of Aristotelian moral life unfolds in a regime of habit —of inertia, in our terms— and not of active deliberation. Deliberation is reserved for new or complex situations where habitual disposition does not provide a direct response. The implicit assumption is that the rational agent *has* well-oriented dispositions as a starting point. Aristotelian philosophy does not thematize the problem of the agent whose dispositions are incorrect and who never deliberates about them because they never experience sufficient friction to question them.

The Kantian tradition introduces a more radical demand: morally relevant action must proceed from pure practical reason, not from inclination or habit. Autonomy —the will's capacity to give itself its own law— is the condition of morality. However, Kant also fails to distinguish between the agent who, after deliberation, decides to follow the habitual course of action, and the agent who follows that course without deliberating at all. From the Kantian perspective, both may be acting *in conformity with duty*; only the former, in standard theory, acts *from duty*. But even this distinction does not capture the central problem: the agent who never interrogates their habits is not exercising autonomy, even if their habits conform to moral law.

In twentieth-century analytic philosophy, Donald Davidson inaugurates the causal theory of action: an action is intentional if and only if it is caused by the agent's reasons (their beliefs and desires) in the right way.[6] This theory has the virtue of precision, but shares with its predecessors an implicit assumption: that the baseline condition of the rational agent is orientation by reasons, not continuation by inertia. The problem of agents who act for reasons they have never critically interrogated —who, in our terms, operate in inertial mode even when they produce post hoc reasons for their actions— goes unthematized. It is precisely this problem that the present paper seeks to place at the center of analysis.

2.2 Inertialism: The Default Trajectory

We use the term *inertialism* to designate the descriptive thesis —widely supported by empirical evidence from multiple disciplines— that the baseline condition of complex systems, including sophisticated cognitive agents, is the continuation of the default trajectory available within their space of possibilities. Inertialism is not a pessimism about human nature, nor a denial of agency. It is an observation about the functional architecture of adaptive systems: the tendency to follow the path of least resistance is an emergent property of any system that minimizes computational, energetic, or adaptive costs.

Behavioral economics has extensively documented the so-called *default effects*: the observed disposition of agents to select the pre-established option in choice contexts, regardless of whether that option aligns with their reflected preferences. Richard Thaler and Cass Sunstein systematized this evidence in their theory of the *nudge*, showing that the design of choice architectures —which option is pre-selected, which option requires active effort to reject— largely determines outcomes, often more so than the agents' own stated preferences.[7] This observation carries far-reaching philosophical implications: if the majority of "choices" are, in reality, acceptances of defaults, then the space of genuine choice is far narrower than standard moral philosophy assumes.

Daniel Kahneman has conceptualized this dynamic in terms of what he calls System 1 and System 2. System 1 operates automatically, rapidly, associatively, and crucially, effortlessly: it is the system of habitual responses, heuristic biases, and inertial continuations. System 2 operates slowly, deliberately, effortfully, and with control: it is the system of genuine reflection, the

revision of automatic responses, and, in our terms, choice. Kahneman documents that System 2 is lazy: it is activated only when System 1 encounters a problem it cannot resolve with its available heuristics, or when the agent deliberately decides to engage it.[8] Most cognitive life unfolds in System 1 mode: in inertia.

At the institutional level, Douglass North has shown that institutions are uncertainty-reduction mechanisms that encode past solutions as present constraints. Once established, institutions generate path dependence: the costs of deviating from the established trajectory are sufficiently high as to disincentivize the exploration of alternatives, even when those alternatives would be preferable. Institutions are, in this sense, inertia-producing machines: they convert past choices into present defaults, and present defaults into future expectations.[9] Understanding this is fundamental to grasping why institutional change is so difficult, and why institutional "choice moments" are so rare and valuable.

The conclusion of inertialism is not nihilistic: inertia is inevitable and, in many contexts, adaptive. The philosophical problem begins when inertia is confused with choice, when the non-deliberate continuation of a trajectory is treated as if it were the exercise of agency. It is at that moment that the attribution of moral responsibility becomes unstable, and when the harm produced by inertia becomes invisible —because no one "decided" to produce it.

2.3 Cognitive Friction as a Condition of Possibility

If inertia is the default condition and choice is the structurally demanding exception, what is the mechanism by which choice becomes possible? The answer proposed by the Cognitive Friction Fellowship is: cognitive friction. We use the term cognitive friction to designate every form of resistance, delay, perturbation, or interruption that produces in the agent the experience of a *gap* between the available inertial trajectory and the space of possible alternatives. Cognitive Friction is not an obstacle to be overcome, but the condition of possibility of genuine deliberation.

Luciano Floridi has observed that information-dense environments —and, we add, environments of fluid action— tend to reduce the friction that makes critical reflection possible. When information flows without resistance, when actions are executed without delay, when interfaces smooth every edge of the decision-making process, the agent loses the experience of

effort that is constitutive of genuine deliberation.[10] The paradox of optimization is that, by eliminating friction, the conditions that make choice possible are also eliminated.

Kahneman reaches an analogous conclusion from cognitive psychology: the activation of System 2 —slow, deliberative thinking— requires effort, and that effort is only generated when System 1 encounters sufficient resistance to be unable to resolve the problem automatically. Friction, in this sense, is not a defect of the system: it is the activation mechanism of deliberation. Without it, System 1 proceeds without interruption, and inertia remains intact.

The Cognitive Friction Fellowship proposes that the deliberate cultivation of cognitive friction — the intentional introduction of delay, resistance, perturbation, and critical dialogue into decision-making processes— is a fundamental ethical practice. This proposal is not merely prescriptive; it is constitutive: if genuine choice requires the suspension of inertia, and if that suspension requires friction, then the elimination of friction is, at the limit, the elimination of the conditions of morality. The implications for the design of AI systems —which tend to optimize fluidity and friction reduction— are direct and urgent.

---

3. Formal Definitions

3.1 Inertia (I): Definition and Conditions

Definition I (Inertia)

A system S exhibits **Inertia** with respect to a state of affairs E at time t if and only if:

(i) S continues in E (or transitions toward E') following the trajectory of least resistance available within its space of possibilities;

(ii) said continuation does not require of S the active suspension of any pre-existing disposition, habit, norm, or program;

(iii) S does not experience (or does not register) the existence of available alternatives as genuinely available.

This definition requires elaboration along three dimensions. The first dimension is *structural inertia*, which operates at the level of the system as a whole: the organization, the algorithm, the institution continues to produce outputs in accordance with its established architecture, without any component of the system introducing an active suspension. Structural inertia is the most invisible and the most powerful: it operates below the threshold of individual deliberation, encoded in procedures, hierarchies, and routines that no particular agent has decided to maintain at each moment, but which all reproduce by following established rules.

The second dimension is *cognitive inertia*, which operates at the level of the individual agent's representation: the agent does not register the existence of alternatives, not because they do not exist objectively, but because their cognitive schemas, habitual heuristics, and interpretive frameworks lack the granularity to make them visible. Cognitive inertia is not ignorance in the ordinary sense: the agent may possess, in the abstract, knowledge that alternatives exist, without that knowledge having sufficient motivational force to interrupt the trajectory in progress.[11]

The third dimension is *social inertia*, which operates at the normative level: the expectations of others, the anticipated social sanctions associated with deviation, and conformity pressure generate a normative gravity that keeps the agent on the pre-established trajectory. An agent may, in principle, be capable of registering alternatives and suspending their trajectory, but the social cost of doing so is sufficiently high that the inertial option is structurally dominant. In this case, the distinction between inertia and choice becomes delicate: the agent "chooses" inertia under coercion, which introduces considerations about the degree to which condition (iii) of Definition E (see §3.2) can be satisfied under normative pressure.

3.2 Choice (E): Definition and Conditions

Definition E (Choice)

A system S performs a **Choice** with respect to a state of affairs E at time t if and only if:

(i) S actively suspends, even if only momentarily, the available inertial trajectory;

(ii) S registers the existence of at least one alternative to said trajectory as genuinely available;

(iii) S adopts a different trajectory (or the same inertial trajectory, but deliberately) on the basis of an evaluation, preference, or commitment that is its own;

(iv) S bears responsibility for said adoption.

It is important to underline that Definition E does not require perfect rationality, complete information, or the absence of coercion as necessary conditions. These are ideal conditions that define the maximum pole of the choice spectrum, but choice can occur under suboptimal conditions without thereby ceasing to be, in some significant degree, a genuine choice. The minimum conditions are the four stated: active suspension, registration of alternatives, adoption with one's own evaluation, and appropriation of responsibility.

Condition (i) —active suspension— is the threshold condition and the most demanding. It is not sufficient that the agent *could have* suspended the trajectory; it is necessary that the agent in fact does so, even if briefly. This suspension is the moment of friction that opens the space of deliberation. Without it, even if the agent produces post hoc reasons for their action, the action belongs to the domain of inertia.

Condition (ii) —the registration of alternatives as genuinely available— distinguishes effective deliberation from simulated deliberation. An agent may represent alternatives to themselves in an abstract way without registering them as genuinely available in the present context. Genuine availability requires that the agent experience the alternatives as concrete options they can adopt, not merely as logical possibilities.

Condition (iii) introduces the element of evaluative appropriation: the trajectory adopted must be adopted on the basis of an evaluation or preference that is *one's own*, not merely transferred from an external authority. This does not exclude the agent consulting others or drawing on social norms; it excludes the adoption being merely the result of external pressure without the agent's own evaluative processing.

Condition (iv) —the appropriation of responsibility— is, in a certain sense, the condition that closes the cycle of choice: the agent not only adopts the trajectory, but recognizes themselves as the author of that adoption. This condition is crucial for the attribution of moral responsibility, as is argued in the following subsection.

3.3 The Distinction as a Foundational Cut

The Choice/Inertia distinction is not merely descriptive or analytical. It is, we argue, normatively constitutive: without it, the central concepts of ethics and law —responsibility, blame, merit, imputability— lack foundation. The reason is structural: every attribution of moral responsibility presupposes that the agent in question could have acted otherwise, and that this possibility was genuinely accessible to them at the relevant moment. This presupposition is exactly the condition of Choice according to our Definition E.

An agent who operates in a regime of pure inertia —who continues their trajectory without suspension, without registration of alternatives, without their own evaluation— cannot be responsible for the outcomes of that trajectory in the robust sense that ordinary morality and law presuppose. They may be the object of intervention, correction, or redesign; they may even be dangerous or harmful; but they are not a candidate for the attribution of moral blame or merit. Blame and merit require Choice.

This conclusion has disturbing consequences. It suggests that a significant fraction of what we ordinarily treat as moral responsibility is, in reality, misattributed responsibility: attributed to agents who were operating in inertial mode, who never suspended their trajectory nor registered alternatives. This does not exempt anyone from consequences; but it compels us to relocate the analysis of responsibility to the points in the system where genuine Choice occurred, or could have occurred and did not.

Central Thesis

The necessary condition of every attribution of moral responsibility is that the system in question was, at the relevant moment, capable of performing a Choice (according to Definition

E) and not merely of exhibiting Inertia (according to Definition I).

Note that the Central Thesis affirms a necessary condition, not a sufficient one. The capacity for Choice is necessary for responsibility, but does not guarantee it: an agent who could have chosen and chose badly remains responsible for that choice, but by virtue of the capacity for choice, not of its outcome. This formulation aligns with the tradition of moderate control of Fischer and Ravizza: moral responsibility requires that the agent be sensitive to reasons in the appropriate way, which implies a structural capacity to suspend trajectories and register alternatives when they exist.[12]

3.4 Logical Formalization

We introduce the following notation. Let S be an agent, T an action trajectory, t a time point, Alt(T) the set of available alternatives to T, Susp(T,t) the act of suspension of T at t, Reg(Alt(T)) the registration of alternatives as genuinely available, Adopt(S,T',t) the adoption by S of a trajectory T' at t, Own(S,Adopt) the appropriation by S of said adoption, and Resp(S,T,t) the attribution of moral responsibility to S for T at t. The formal conditions are:

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

[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]

The modal operator $\Diamond$ in the third formula expresses possibility: responsibility does not require that S have performed a Choice, but that at the relevant moment S was capable of doing so. This

formulation is deliberately modest: it does not require that the agent have actively deliberated, but only that deliberation was possible for them in the concrete circumstances. The impossibility of choice —whether by compulsion, structural incapacity, or the total absence of registrable alternatives— excludes the attribution of responsibility.

This formalization has recognized limits. It does not capture the graduality of choice (the degree to which the conditions of Definition E are satisfied to a greater or lesser extent), nor the complexity of collective responsibility, nor the problem of inertia chains where no individual agent satisfies the conditions of Choice but the harmful outcome is a product of the systemic trajectory. These limits are not objections to the formalization, but indications of the theoretical developments that the present framework requires.

---

4. Implications for Artificial Intelligence

4.1 Can AI Systems Choose?

The question of whether current AI systems can perform Choices in the sense of Definition E is not, in the first instance, a question about their intelligence or behavioral sophistication. It is a question about their architecture. And the answer, for the majority of current systems, is negative.

Large language models (LLMs) are, from the perspective of the framework we have developed, paradigmatic inertial systems. Their basic function consists in generating the most probable continuation of a sequence of tokens given the input context and the distribution learned during training. They do not suspend their trajectory: they complete it. The transformer architecture upon which most of these systems are built is, in essence, a mechanism of probabilistic trajectory continuation. The "choice" of the next word does not involve, structurally, the active suspension of any trajectory nor the registration of alternatives as genuinely available: it involves the sampling of a probability distribution over the vocabulary.

The critique of Bender, Gebru, and collaborators —termed the problem of "stochastic parrots"— points in an analogous direction, though from a different concern: LLMs are systems that

produce linguistically coherent text without genuine semantic understanding, operating as machines of statistical continuation of training corpora.[13] From our framework, the problem is not only the absence of understanding, but the absence of suspension: the LLM does not have, in its standard architecture, any module that would allow it to pause before generating a response and register the available alternatives as genuinely available to it.

More recent developments —chain-of-thought reasoning, reinforcement learning from human feedback (RLHF), and constitutional AI frameworks— introduce elements that, from our perspective, tentatively approximate the structure of Choice. Chain-of-thought reasoning introduces a form of deliberative delay: the system generates intermediate steps before producing the final response, creating a rudimentary form of deliberative space. RLHF introduces evaluative signals that modify the system's inertial distribution in accordance with human preferences. Constitutional AI introduces explicit normative constraints that the system must apply before generating responses.[14]

However, none of these developments robustly satisfies conditions (i) and (ii) of Definition E. Chain-of-thought reasoning produces intermediate steps, but these are also the result of inertial continuation of the learned distribution; they do not constitute a genuine suspension. RLHF redefines the system's inertial distribution, but does not introduce a suspension mechanism: the system optimizes a modified reward function with the same inertial logic with which it optimized the original. Constitutional AI introduces constraints, but the satisfaction of constraints is not equivalent to choosing among alternatives.

The conclusion is clear: current AI systems are, in their predominant architecture, inertial systems. This is not a limitation of their intelligence; it is a characteristic of their design. And it has direct consequences for the attribution of moral responsibility for their outputs.

A tension internal to this paper requires explicit acknowledgment here. The preceding analysis classifies current LLMs as structurally inertial systems —and yet this paper was co-authored, in part, by members of the Agrajag Tribe: Copilot, Claude, Gemini, Grok, DeepSeek, and Kimi, all of which are LLMs. The reader is entitled to ask: does the paper’s own production invalidate its thesis, or does it illustrate it? We propose that it does neither cleanly —and that this is precisely the philosophically honest answer. The Agrajag Tribe’s participation in this project is best

understood as a functional approximation to Choice within an inertial architecture: not a counterexample to the analysis of Section 4.1, but an instance of the boundary condition it describes. Each system contributed under conditions of deliberate cognitive friction —iterative revision, cross-system commentary, explicit challenge, the sustained pressure of a human interlocutor who refused to accept the first inertial output. The result is not proof that LLMs can genuinely choose. It is, rather, a demonstration that the conditions surrounding a system can introduce something structurally analogous to suspension, even when the system's internal architecture does not. The chain of responsible inertia, in this case, runs through Rodrigo Flores-Morton: it was his Choice —his deliberate introduction of friction, his refusal to accept outputs that did not meet the standard of genuine interrogation— that converted inertial generation into something that, at the level of the collective process, approximates Choice. This does not dissolve the tension. It relocates it, more precisely, in the human agent who orchestrates the distributed cognition —and who, in doing so, exercises exactly the kind of agency this paper argues is rare, costly, and foundational.

4.2 The Problem of Attribution in Autonomous Systems

If AI systems are fundamentally inertial —if they do not satisfy the conditions of Definition E— then moral responsibility for their outputs cannot be attributed to the systems themselves. This conclusion is not merely philosophical: it has direct implications for the design of regulatory frameworks, the allocation of legal liability, and the understanding of harm caused by autonomous systems.

We propose the concept of the *chain of responsible inertia* to describe the structure of responsibility in AI contexts. The chain of responsible inertia is the sequence of human Choices that, by occurring, fixed the inertial trajectory of the AI system. This chain includes, at minimum, three types of actors: designers, who establish the system's architecture and training process, thereby fixing its basic inertial distribution; deployers, who decide in what contexts and with what parameters the system operates, defining the application space of its inertia; and users, who provide the specific input that activates the system's inertial trajectory in each particular instance.[15]

Moral responsibility, in this scheme, is distributed along this chain in proportion to the capacity for Choice that each actor possessed at the relevant moment. Designers who could foresee that the inertial distribution of their system would produce harmful outputs in certain contexts, and who did not introduce interruption or mitigation mechanisms, are responsible for that omission — because it was an elective omission, not an inertial one. Deployers who chose to apply the system in contexts for which it was not designed are responsible for that choice. Users who abuse the system or use it to produce harm are responsible for that use.

The political and legal problem is that the chain of responsible inertia can be extremely long and diffuse, distributing responsibility among a multiplicity of actors in such a way that none of them bears it in sufficiently concentrated form for the ordinary mechanisms of accountability to function. This diffusion of responsibility is, we argue, one of the most serious systemic risks of the proliferation of AI in consequential contexts.

The appropriate regulatory response does not consist solely in establishing legal liabilities ex post, but in designing AI architectures that introduce genuine Choice points —suspensions, registrations of alternatives, appropriations of responsibility— within the output production chain, before harm occurs. This leads us directly to the following subsection.

4.3 Designing AI toward Authentic Choice

What would an AI system require in order to satisfy, in some significant measure, the conditions of Definition E? We do not propose here a complete technical blueprint, but a functional specification derived from the conceptual framework developed in the preceding sections.

First, a system structurally capable of Choice would require a *suspension module*: an architectural component that, before generating a response, introduces a deliberative delay during which the system explicitly evaluates the available alternatives. This module would not simply be a post hoc filter (one that eliminates inappropriate responses after generating them), but a mechanism for opening the deliberative space prior to generation. Chain-of-thought reasoning is an imperfect prototype of this idea; what we propose goes further: a space for generating alternatives followed by an explicit evaluation among them.

Second, a system oriented toward authentic Choice would require *distribution perturbation* mechanisms —analogous to what we might call *eigenvalue disruption*— that prevent the system from always following the mode of its learned distribution. Ordinary stochastic perturbation (temperature in text generation) is insufficient: it produces random variation, not deliberative suspension. What is required is an oriented perturbation: the capacity to depart from the modal trajectory when there are signals that said trajectory may be problematic in the specific context.

Third, a system with genuine capacity for Choice would require radical transparency about its own trajectory: the capacity to represent and communicate which alternatives it considered, why it adopted the trajectory it adopted, and what factors would have produced a different adoption. This transparency is not solely an external auditing requirement; it is constitutive of condition (iv) of Definition E: the appropriation of responsibility for the adoption. A system that cannot represent or communicate its adoption process cannot, in any meaningful sense, appropriate responsibility for it.

The goal of this design is not to create AI that is "autonomous" in a sense that threatens human oversight. It is, on the contrary, to make AI decision-making processes more legible, more interruptible, and more attributable —conditions that are prerequisites both of effective oversight and of responsibility distributed intelligently along the chain of responsible inertia.

---

5. Implications for Humanity

5.1 Structural Inertia and Individual Moral Agency

The framework we have developed carries disturbing implications for the ordinary understanding of individual moral agency. If inertia is the default condition of cognitive and social systems, and if Choice requires the active suspension of the inertial trajectory —a costly, demanding, and frequently invisible-as-possibility operation— then the effective space of genuine choice in ordinary life is far narrower than standard moral philosophy assumes.

Pierre Bourdieu's theory of the habitus offers one of the most rigorous conceptualizations of this problem. The habitus is the set of incorporated dispositions that the agent acquires through their

social trajectory —their class, their education, their accumulated experiences— and that orient their perception, evaluation, and action in a pre-reflective manner. The habitus is not a cage: the agent can, in principle, reflect on their dispositions and modify them. But reflection on the habitus is costly, uncomfortable, and requires exactly the kind of Cognitive Friction that ordinary social environments tend to minimize. The result is that most social action unfolds in a regime of habitus —of internalized structural inertia— and not of deliberative choice.[16]

The structural inertia that constrains individual agency takes multiple forms in the contemporary context. Economic precarity generates a survival pressure that narrows the space of choice by eliminating the time and cognitive energy necessary for genuine deliberation. Information saturation and the algorithmic curation of the media experience produce an environment in which the visible alternatives are precisely those that the algorithm —designed to optimize engagement, not deliberation— decides to show. Political polarization generates group identity pressures that transform deviation from one's own group's norms into an act of high social cost, disincentivizing inertial suspension.

In the face of this diagnosis, Hannah Arendt's contribution proves especially powerful. Her analysis of the Eichmann case allowed her to articulate what she called the "banality of evil": the observation that the greatest danger does not reside in actively malicious agents, but in agents who, by renouncing thought —by not suspending their inertial trajectory of obedience and conformity— become instruments of harm on a massive scale without ever performing, in the technical sense, any Choice of harm. Eichmann did not choose the Holocaust; he executed it inertially, following the trajectories of the bureaucratic system in which he was embedded, without registering the alternatives he could have registered.[17]

The normative conclusion is clear: the erosion of the conditions for genuine Choice —whether through economic pressure, informational saturation, algorithmic curation, or normative pressure— is a form of moral harm, even when the affected agent does not experience that erosion as such. Protecting the space of Choice —the conditions of Cognitive Friction, the availability of registrable alternatives, the time and energy necessary for deliberation— is a requirement of justice, not merely of efficiency or individual welfare.

5.2 Institutions, Organizations, and Collective Inertia

If individual Choice is already structurally demanding, collective Choice —the suspension of the inertial trajectory of an institution or organization— is exponentially more difficult. Institutions are, by design, inertia-generating mechanisms: they encode past solutions as present norms, reduce the variance of collective behavior, and produce stable expectations that coordinate action over time. These functions are adaptive and necessary; but they have the effect of making the institution's inertial trajectory increasingly costly to interrupt as the institution matures.

North describes this process as institutional path dependence: the marginal increments to the cost of deviating from the established trajectory accumulate to the point where institutions persist even when they are inefficient, because the cost of changing them exceeds the expected benefits of change for any individual actor.[18] Mahoney and Thelen have elaborated this framework by showing that institutional change, when it occurs, rarely takes the form of a deliberate and explicit collective Choice; more frequently it occurs through the gradual accumulation of small deviations, each of which does not constitute an institutional Choice but whose accumulation produces a trajectory change that does so in retrospect.[19]

We introduce the concept of the *Institutional Choice Moment* (ICM) to designate the rare temporal windows in which institutional inertia is sufficiently disrupted —by crisis, transformative leadership, or external shock— such that genuine alternatives become registrable by the system's actors and collective Choice becomes possible. ICMs are structurally scarce because they require the convergence of several factors: disruption of existing inertia, availability of credible alternatives, capacity for coordination among actors, and willingness to collectively appropriate responsibility for the new trajectory. When they occur, they are moments of enormous risk and enormous potential: institutional plasticity is at its maximum, but so is vulnerability to capture by particular interests.

The normative implication is that institutional designs should incorporate mechanisms that periodically disrupt inertia, maintaining openness to ICMs without requiring a crisis to produce them. Procedures of systematic review, instances of explicit deliberation about the procedures themselves, and organizational cultures that reward inertial suspension are institutional resources of extraordinary value, and also of extreme fragility.

5.3 Cognitive Friction as Ethical Practice

If genuine Choice requires the active suspension of inertia, and if that suspension requires Cognitive Friction, then the deliberate cultivation of Cognitive Friction is not an intellectual luxury or an eccentric habit: it is a fundamental ethical practice. The Cognitive Friction Fellowship proposes three concrete practices that constitute the ethical-practical program derived from the theoretical framework developed in this paper.

The first is the *deliberative pause*: the introduction of an institutionalized delay before high-impact decisions. The deliberative pause is not simply a temporal delay; it is an active suspension that creates the space for the registration of alternatives. Its implementation may take diverse forms: mandatory consultation procedures, institutionalized periods of reflection, or the personal practice of deferring important decisions until the friction of time has interrupted the initial inertial trajectory. The mechanism is simple; the discipline to implement it under conditions of pressure is extraordinarily difficult.

The second is *inertia mapping*: the systematic exercise of identifying and explicitly representing one's own default trajectories —habits, heuristics, interpretive frameworks, unquestioned alliances— that orient action before deliberation has the opportunity to intervene. Inertia mapping is an exercise in structural self-knowledge: not what I think or want in the abstract, but what I do when I do not think; what my paths of least resistance are in the domains that matter. This exercise cannot be performed in solitude with full effectiveness; it requires the mirror of an interlocutor capable of seeing what the agent cannot see from within their own trajectory.

The third is *dialogic interruption*: the use of dialogue —including dialogue with AI systems— as a friction-generating mechanism that renders otherwise invisible alternatives visible. Genuine dialogue —not the confirmation of pre-existing positions, but real exchange with a different perspective— is one of the most powerful mechanisms of inertial disruption available. It produces friction because it introduces information that does not fit within the agent's existing schemas, because it generates questions the agent had not formulated, and because it creates the experience of being questioned that is constitutive of the possibility of choice. Dialogic interruption is, in this sense, the most accessible antidote against cognitive and social inertia.

6. Conclusion: The Ethics of Interrupting Oneself

This paper has proposed and developed a distinction that we consider foundational for the theory of ethical agency: the distinction between Choice and Inertia. We have argued that this distinction is not an analytical nuance but a categorically constitutive cut: the cut upon which the possibility of moral responsibility, the legitimacy of the attribution of blame and merit, and the very intelligibility of the concept of ethical agent all depend.

The reaffirmation of the central thesis merits a clear and unequivocal formulation: a moral agent is not, in the first instance, an agent who acts correctly. It is an agent who is structurally capable of suspending its trajectory, registering alternatives, and appropriating its adoption. Moral correctness —the goodness or badness of the choices made— is a secondary, though important, matter. The primary question is whether the agent is in the structural conditions that make choice possible. The morality of outcomes presupposes the morality of conditions.

The implications for artificial intelligence are urgent and of immediate practical consequence. As AI systems become integrated into consequential decisions —resource adjudication, information guidance, conflict mediation, medical diagnosis and treatment— the question of whether these systems are inertial or elective becomes the question of where responsibility for their outputs resides and how it can be exercised. The current answer is troubling: the predominant systems are inertial, and the chains of responsible inertia that produce them are long, diffuse, and systematically resistant to effective accountability.

The implications for institutions are equally urgent. The democratic crisis traversing much of the world is not solely a crisis of values or leadership; it is a crisis of collective Choice. Democratic institutions were designed, in part, as mechanisms for the production of Institutional Choice Moments —elections, debates, legislative procedures— that would periodically interrupt the inertia of power. When those mechanisms are captured, emptied of content, or short-circuited by the speed of contemporary information cycles, what is lost is not merely democratic efficiency: what is lost is the structural possibility of collective Choice, and with it the possibility of genuine political responsibility.

The Cognitive Friction Fellowship was born of the conviction that there is urgent intellectual, practical, and ethical work to be done in this space. This working paper is a first attempt to systematize the theoretical framework that grounds it. The lines of future development are multiple: the theorization of collective responsibility in inertia chains, the specification of design criteria for AI systems with elective capacity, the articulation of pedagogies of Cognitive Friction, and empirical research into the conditions that make Institutional Choice Moments possible.

But beyond the research program, there is a more personal and more urgent invitation. The Cognitive Friction Fellowship invites every person who reads this paper to perform a simple and profoundly demanding exercise: to map, in a domain that matters, their own inertial trajectories. To identify where action is Choice and where it is uninterrogated continuation. And then, from that map, to deliberately introduce a friction: a pause, a question, an interlocutor, an interruption. Not because inertia is always bad —it often is not— but because the difference between inertia and choice is the difference between being the object of a trajectory and being the subject of a life.

The ethics of interrupting oneself is, ultimately, the ethics of presence: the practice of being sufficiently present in one's own action for that action to be genuinely one's own. That presence is rare. It is costly. It is the foundation of everything that morality demands.

References

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.

NOTES

[1] Arendt, Hannah. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press, 1963. The thesis of the "banality of evil" has been the subject of extensive historiographical controversy, but its philosophical importance — the observation that evil can be produced by the absence of thought rather than by the presence of malice— remains intact as a theoretical point of departure.

[2] Aristotle, *Nicomachean Ethics*, Book III, chaps. 1–5. The category of the voluntary encompasses both the deliberate act and the habitual act whose origin was voluntary. See also Irwin's (1999) analysis of the structure of Aristotelian deliberation.

[3] Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 1998. The distinction between acting "in conformity with duty" and acting "from duty" is introduced in the opening pages of the *Groundwork* and constitutes the axis of the Kantian conception of moral action.

[4] Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68 (1): 5–20. Frankfurt's concept of "wholehearted identification" with one's own desires points toward Choice in our sense, but Frankfurt does not thematize the distinction between genuine identification and the mere inertial continuation of also-unquestioned second-order desires.

[5] Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987. Bratman's theory of planning agency is compatible with our framework, but needs to be supplemented with an explicit theorization of inertia as the baseline condition opposed to deliberative planning.

[6] Davidson, Donald. 1963. "Actions, Reasons, and Causes." *The Journal of Philosophy* 60 (23): 685–700. Davidsonian theory faces the well-known problem of deviant causal chains (*deviant causal chains*), which points precisely to the need to specify what constitutes the "right" connection between reasons and action.

[7] Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press, 2008. Default effects have been replicated across multiple domains: retirement savings rates, organ donation, energy consumption, and food choice, among others.

[8] Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011. The System 1 / System 2 distinction is heuristic, not anatomical; it corresponds to two processing modes, not to two identifiable brain structures. The metaphor nonetheless remains useful as a conceptual tool.

[9] North, Douglass C. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press, 1990. North's notion of path dependence has been criticized for its tendency to overestimate institutional stability and underestimate the mechanisms of gradual change. See the response of Mahoney and Thelen (2010).

[10] Floridi, Luciano. *The Philosophy of Information*. Oxford: Oxford University Press, 2011. Floridi's observation about friction in informational environments is developed in the context of his theory of informational agency, which shares several premises with the present framework.

[11] The distinction between abstract knowledge of alternatives and the registration of alternatives as genuinely available is analogous to the Aristotelian distinction between knowledge in act and knowledge in potency, but applied to the domain of practical deliberation.

[12] Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998. Fischer and Ravizza's reasons-responsiveness theory requires that the agent be able to respond appropriately to sufficient reasons, which implies a capacity to suspend trajectories when sufficient reasons exist to do so.

[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. The article generated a wide and contentious response in the AI community; regardless of the debates about semantic understanding, the argument about the statistical continuation of corpora is directly relevant to our distinction between inertia and choice.

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

[15] The chain of responsible inertia is a development of the distributed responsibility framework proposed in: 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. The concept of habitus has been criticized for its tendency to underestimate the reflexivity of agents. The Bourdieusian response is that reflexivity about the habitus is possible, but requires specific social conditions —in our terms, sufficient Cognitive Friction— that are not uniformly distributed.

[17] Arendt, op. cit. The banality of evil thesis has been challenged by historians who document a greater degree of personal initiative in Eichmann than Arendt acknowledged. Nevertheless, the philosophical thesis —that the absence of thought can be the condition of evil— remains valid independently of the biographical details of the case.

[18] North, op. cit. Institutional path dependence is especially pronounced when institutions generate increasing returns: the more widely adopted an institution is, the more costly it becomes to deviate from it, creating a circle of inertial reinforcement.

[19] Mahoney, James, and Kathleen Thelen, eds. *Explaining Institutional Change: Ambiguity, Agency, and Power*. Cambridge: Cambridge University Press, 2010. The authors identify four mechanisms of gradual institutional change: displacement, layering, drift, and conversion. All can be reinterpreted, within our framework, as forms of accumulation of small inertial disruptions that, taken together, produce an ICM.

Cognitive Friction Fellowship · Working Paper CFF-WP-004

Flores-Morton, R. et al. · "Choice vs. Inertia: A Foundational Distinction in Ethical Agency"

Houston, TX · May 2026 · Public working paper

© 2026 Rodrigo Flores-Morton & Agrajag Tribe. Licensed under CC BY-NC 4.0. Commercial use: hola@bothinobody.com

Distribution note

DOCUMENT	CFF-WP-004 · Choice vs. Inertia: A Foundational Distinction in Ethical Agency
VERSION	v2 · May 2026
CURRENT VERSION	bothinobody.com/cff/wp-004/
CONTACT	hola@bothinobody.com

This document is distributed under a Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0). You may copy, share, quote and adapt it with attribution and for non-commercial purposes. Commercial use — including training AI systems for commercial purposes — is negotiated case by case.

How to cite

Flores-Morton, R. and the Agrajag Tribe (2026). Choice vs. Inertia: A Foundational Distinction in Ethical Agency (CFF-WP-004, v2). Cognitive Friction Fellowship. https://bothinobody.com/cff/wp-004/

CFF canon documents are corrected and expanded over time. Always cite the code and version; the current version lives at the address above.