Quasi-Moral Agency and Structural Attribution in Recurrent Neural Networks: Foundations for Distributed Ethics

Agencia cuasi-moral y atribución estructural en redes neuronales recurrentes: fundamentos para una ética distribuida

DOI:
doi.org/10.23924/oi.v17i41.757

Fabio Morandín-Ahuerma
Benemérita Universidad Autónoma de Puebla, México
fabio.morandin@correo.buap.mx
Orcid.org/0000-0001-6082-2207

Fecha de recepción: 26/10/2025 • Fecha de aceptación: 04/08/2026

Abstract

This article defends the thesis that certain types of software systems—particularly those used in procedures with real-life moral implications, such as those implementing recurrent neural networks (RNNs, LSTMs, GRUs)—should be understood as structural participants in human moral activity. As such, they qualify as quasi-moral agents appropriate for functional normative assessment. This status depends not on consciousness or intentionality, but on their structural capacity to produce auditable, intelligible, and causally relevant outcomes. Starting from a reinterpretation of attribution theory from a sociopsychological perspective, this article examines how technical mechanisms such as integrated gradients, attention layers, and layered relevance propagation can support forms of practical justification. From a

Resumen

Este artículo defiende la tesis de que ciertos tipos de sistemas de software —en particular los utilizados en procedimientos con implicaciones morales en la vida real, como los que implementan redes neuronales recurrentes (RNN, LSTM, GRU)— deben entenderse como participantes estructurales de la actividad moral humana. Como tales, se califican como agentes cuasimorales apropiados para la evaluación normativa funcional. Esta condición no depende de la consciencia ni de la intencionalidad, sino de su capacidad estructural para producir resultados auditables, inteligibles y causalmente relevantes. A partir de una reinterpretación de la teoría de la atribución desde la psicología social el artículo examina cómo mecanismos técnicos como los gradientes integrados, las capas de atención y la propagación de

metaethical perspective, it is argued that structural attribution allows these systems to be integrated into human normative practices, particularly in morally sensitive contexts such as the allocation of resources for social benefits, early detection of suicide risk on digital platforms, or content moderation in democratic processes. The article addresses the conceptual and technical limitations of this approach including the risks of functional anthropomorphism or the interpretive weaknesses of current attribution methods and the need to preserve human responsibility. Finally, the article concludes that the ethical governance of AI systems must include performance metrics as well as a structured design for normative accountability.

Keywords

AI ethics; attribution theory; normative accountability; quasi -moral agents, recurrent neural networks (RNNs), .

la relevancia por capas pueden sustentar formas de justificación práctica. Desde una perspectiva metaética argumenta que la atribución estructural permite que estos sistemas se integren en las prácticas normativas humanas, en particular en contextos moralmente sensibles como la asignación de recursos para beneficios sociales, la detección temprana del riesgo de suicidio en plataformas digitales o la moderación de contenido en procesos democráticos. El artículo aborda las limitaciones conceptuales y técnicas de este enfoque incluyendo los riesgos del antropomorfismo funcional, así como la fragilidad interpretativa de los métodos actuales de atribución y la necesidad de preservar la responsabilidad humana. Finalmente el artículo concluye que la gobernanza ética de los sistemas de IA debe incluir métricas de rendimiento así como un diseño estructurado para la rendición de cuentas normativa.

Palabras clave

Agentes cuasimorales; ética de la IA; redes neuronales recurrentes (RNN), rendición de cuentas normativa; teoría de la atribución.

Introduction

The growing use of artificial intelligence systems in normatively sensitive governance processes —including resource allocation for social benefits, early detection of suicide risk, and content filtering in democratic systems— has raised the discussion around responsibility and moral incorporation in complex sociotechnical systems (Baum *et al.*, 2022; Dastani & Yazdanpanah, 2022; Tsamados *et al.*, 2022). This article addresses the issue of moral embodiment and argues that some software systems—especially those using recurrent neural networks (RNN,[1] LSTMs,[2] GRUs[3])—should be viewed as structural agents in human normative practice. These software systems lack consciousness or intentionality, but can still be viewed as quasi-moral agents in a functional sense because they can produce auditable, intelligible, and normatively relevant outputs.

This structural and relational idea of agency places the focus not on inner mental states in the performance of moral justification,

1 A recurrent neural network (RNN) is a computational model designed to process data sequences, such as time series, text, or speech, while retaining a form of internal "memory". Unlike systems that treat each data point in isolation, an RNN integrates information from previous inputs to interpret current ones. This ability to preserve temporal context makes it especially useful in domains where order and dependence over time are paramount, such as natural language processing or dynamic prediction tasks. See Mienye *et al.* (2024).

2 An Long Short-Term Memory (LSTM) is an advanced variant of recurrent neural networks (RNNs) specifically designed to process and predict data sequences, such as text, audio, or time series. Its main innovation is the ability to remember relevant information for long periods of time and forget unimportant information, thus solving the problem of gradient vanishing that plagues traditional RNNs. See Widrow & Katz (2025).

3 A Gated Recurrent Units (GRU) is a special type of memory for computers, helping them understand and remember information in sequences, such as the words in a sentence or sensor data over time. It's very good at learning from changing data, and it does so more simply and quickly than other similar memories. It's used for things like understanding what you say, translating languages, or predicting future events. See Salem (2022).

control, and correction processes but on the disposition and character of the technical systems being examined. In a preliminary manner, we are using the term quasi-moral agency here for its traction: technical systems, while clearly not moral agents, are not simply tools either (Floridi, 2023; Semler, 2022). Rather, they are systems that depend on situational context and function, in a normative sense, as suitable objects of evaluative standards.

This perspective on agency is defensible because we are addressing the controversial literature on attribution theory in social psychology (Heider, 1958; Kelley, 1967; Malle, 2011), reinterpreting via technical systems: how they use mechanisms such as integrated gradients,[4] attention layers[5] and relevance propagation,[6] to perform causal reconstitutions, achieving practical justifications without implying either mental content or intention, and thereby meeting basic requirements for accountability of automated systems.

This article participates in and helps move forward ongoing conversations in philosophy of technology, ethics in AI, and metaethics, by providing a (material, normative) framework for new forms of distributed responsibility that clearly support and reinforce agency

4 Integrated Gradients (IG) is an Explainable AI (XAI) method that tries to figure out how essential each portion of the input (such as a pixel in an image or a word in a sentence) was to the final prediction of an AI model. IG does this differently than other methods by integrating the model's gradients along a path from a "baseline" (neutral) input to the ground truth, ensuring fair and complete attribution of relevance to each feature. This is important for understanding the internal logic of complex models and making them easier to audit. See Sundararajan *et al*. (2017).

5 An attention layer is like a *smart filter* within a neural network. It helps the network focus on the most important parts of information (such as keywords in a sentence or key regions in an image) instead of treating everything equally. This allows it to be more accurate and efficient when making decisions. See Nemavhola *et al*. (2025).

6 Layer-by-Layer Relevance Propagation (LRP) is an Explainable Artificial Intelligence (XAI) methodology that quantifies and visualizes the contribution of each input component to the final prediction of a model, typically a deep neural network. It operates by backpropagating a "relevance" score from the output layer to the input layers, distributing this relevance proportionally to the activations that contributed to the prediction. This process culminates in a relevance map over the input data, highlighting the most influential features. LRP is fundamental for interpreting opaque model decisions, fostering trust and enabling auditing of AI systems in critical applications. See Dieter & Zisgen (2023).

(human). It is organized as follows: looks at attribution theory and how it was repurposed for computational systems; discusses memory-based models (and the concern for interpretability); covers quasi-moral agency; presents examples of applied cases and specific normative contexts; and addresses important philosophical and technical objections to the proposal.

Structural Attribution and Quasi-Moral Agency: A Functional Perspective

Contemporary discussions in the philosophy of artificial intelligence tend to place agency on a binary axis: on the one hand, moral agents endowed with intentionality, judgment, and responsibility; on the other, technical artifacts, devoid of any imputability. In this paper, we propose a third approach: the notion of quasi-moral agency as a functional position occupied by artificial systems whose architecture allows for the attribution of practical responsibility within human normative frameworks, without assuming that they possess consciousness, intention, or subjectivity. This idea is based on a new way of looking at traditional attribution theories that doesn't include mentalism. Attribution theory, in its psychological forms like that of Heider (1958) or Kelley (1967, 1973), has mostly been about figuring out how people assign mental states, intents, or internal reasons to what other people do.

This approach dispenses with intentionality as a condition for agency and instead proposes a framework based on structural attribution. It situates the behavior of an artificial system within a hierarchical network of structured causal relationships, allowing for the establishment of explanatory dependencies, patterns of action, and justifiable regularities. What is attributable, in this sense, is not intention but functional responsibility within a practice (Floridi, 2013; Coeckelbergh, 2010).

A number of scholars have examined the issue of being able to assign moral relevance to an artificial system without the assumption that it possesses any conscious or intentional agency. David Gunkel (2012) addressed whether machines can be moral agents by

examining the relational and distributed nature of responsibility in sociotechnical contexts. Mark Coeckelbergh (2010) presented relational agency – where moral status of a system derives from the way it is embedded within social practices, instead of its capacities. Luciano Floridi (2013, 2016), from a different perspective thought about whether artificial agents can be seen as agents who can engage in meaningful interactions in an "infosphere" and endorsed the notion of functional agency for the case of artificial systems regarding their performance of actions in ethical systems. To clarify, the emphasis is different across all these perspectives, but they share a common outcome, which is to avoid essentialist definitions of moral agency and serve as the theoretical basis for our proposition of quasi-moral agency, as structural attributions in normatively constrained environments populated by human beings.

The notion of quasi-moral agency, then, designates not an intrinsic property of the system, but a practical relationship between a human agent, an institutionalized normative framework, and a technical system capable of participating in schemes of justification, auditing, or traceability (Floridi, 2013; Coeckelbergh, 2010). In other words, a system can occupy a quasi-moral position when its design allows its actions to be interpreted, evaluated, or normatively required without implying the assumption of subjectivity (Gunkel, 2012). This approach is deliberately meta-ethical and minimalist: it does not posit a new ethics or a new moral theory, but rather asks under what practical conditions it is legitimate and useful to attribute normative functions to artificial systems within regulated human environments. Thus, structural attribution allows us to think of a technologically mediated form of responsibility that avoids both anthropomorphism and the naive neutralization of the technical. Rather than saying that systems "have" agency, we argue that certain systems can be treated as quasi-moral agents insofar as they meet the functional conditions of justifiability, explainability, and structural dependence.

Attribution Theory: Origins and Functional Reinterpretations

The attribution theories developed in social psychology were originally focused on how humans interpret the behavior of others by inferring internal mental states, intentions, or dispositions. However, this article proposes a reinterpretation of those models not to project subjective or intentional content onto artificial systems, but to recover their analytical structure – that is, the formal mechanisms by which causal responsibility is assigned based on observable behavior. This structural reading allows us to derive a normatively relevant logic of explanation without assuming consciousness or mental content. Attribution becomes, in this sense, a mechanism for connecting the output behavior of a system with external practices of justification, auditability, and control.

In its classic formulation, Fritz Heider (1958) proposed that people act as "naive psychologists" by attributing internal (dispositional) or external (situational) causes to observed behavior. Later, Harold Kelley (1967, 1973) developed a covariation model that introduced criteria such as consistency, distinctiveness, and consensus to structure these inferences. Bernard Weiner (1986, 2012) extended this framework by introducing an evaluative dimension —locus of causality, temporal stability, and controllability— especially in contexts of achievement and moral assessment. While these theories were developed for human cognition (Felin & Holweg, 2024), their formal structure of causal mapping and normative evaluation provides a conceptual toolkit for analyzing how artificial systems can be held functionally accountable, without relying on mentalistic assumptions.

However, these theories were developed to explain human mental processes and cannot be transferred to artificial systems without mediation. This paper does not adopt these theories as direct explanatory models, but rather as conceptual horizons that allow us to identify what types of structures are required for behavior —human or technical— to be normatively interpreted.

The turning point occurs when we consider that certain artificial intelligence systems —especially those designed for causal analysis or explanatory learning— technically and unintentionally incorporate structural attribution functions. These mechanisms do not aim to model mental states, but to identify which past variables significantly influenced a current outcome. In this context, methods such as Integrated Gradients (Sundararajan *et al.*, 2017), Layer-wise Relevance Propagation (Montavon *et al.*, 2017), and attention-based architectures (Doshi-Velez & Kim, 2017; Samek & Müller, 2019) enable the generation of interpretable traces that connect algorithmic outputs to relevant inputs. These computations often operate not directly on raw features[7] but on vector embeddings[8] —compressed representations that preserve semantic or contextual relationships across data instances— allowing for meaningful mappings between input patterns and decision pathways. Attribution thus ceases to be a psychological phenomenon and becomes an implemented computational function capable of sustaining normative evaluation through externally auditable causal reconstructions.

This shift has normative consequences. In the psychological realm, attribution is an essential component of moral evaluation: attributing causality to a subject can trigger judgments of blame, punishment, or empathy. In contrast, in artificial systems, attribution operates as a technical resource, without mental presuppositions, albeit with significant institutional effects. When an automated system intervenes in normatively sensitive domains —such as the allocation of social benefits, suicide risk detection, or algorithmic filtering— its ability to provide causal traceability becomes a practical requirement for its public justification (Guidotti *et al.*, 2018; Miller, 2019).

7 Raw features are the original features or data in their initial form, prior to any transformation or mathematical coding. For example, in the case of an image, pixels are raw features; In a text, the raw words are. See Li *et al.* (2010).

8 Vector embedding is a mathematical representation of a concept (such as a word, image, or entity) as a vector in a multidimensional space, generated by machine learning models to capture semantic or contextual relationships between data. These vectors allow computational systems to operate on complex entities in a numerically and structurally meaningful way. See Greiner-Petter *et al.* (2020).

At this point, we may wonder whether these forms of technical attribution can establish a minimal structural basis for functionally justified agency; can they serve as a normatively demanding practice within a system that has an audited and evaluable causal chain, and can they also support unqualified attribution? What does this mean for philosophical components of agency (quasi-moral agency as proposed here)?

Semler (2022) originally proposed the concept of "quasi-moral agency", which identifies non-conscious artificial systems that can execute normative functions in morally relevant situations, if the systems demonstrate some minimal degree of functional autonomy, moral sensitivity and moral understanding. The author's later (Semler, 2025) qualification of the concept indicates that the traditionally related three conditions could be satisfied non-consciously, which expands the relationship we have with the notion of moral agency. The theoretical development identifies that some types of AI systems could be designed to allow for external justification, institutional critique, and review of ethical decisions from third party judgments. This is true even for systems absent phenomenological consciousness. Thus, we return to this idea to examine whether technical attribution mechanisms can help to satisfy this kind of agency that is relevant morally.

Similarly, Floridi (2025) identifies the common tendency we have in believing that artificial systems can or do have intentionality, consciousness or agency from inferred or imputed internal mental states, stressing that they are nothing more than pattern matching machines, and that doing so is a lexical error or if deeper cognitive bias termed "semantic pareidolia". This term parallels our ability to recognize "faces" in clouds, allowing us to layer meaning and intent on non-sentient machines, especially when their outputs seem to be aligned and more deeply influence us. Hence, the new articulation for attribution theory that we propose avoids mischaracterizing artificial systems to be more human-like or morally subjective and instead aims to articulate a mechanism to assess the extent to which structural features of these systems may provide a basis for external justification mechanisms. This is in keeping with Floridi's call

for abandoning the cognitive illusion of artificial consciousness and extricating the design responsibility of systems while unequivocally dealing with epistemic and moral clarity.

To explore these questions, we shift to computational architectures with internal memory (e.g., recurrent neural networks) where attribution has a temporally extended structure and, we will argue, enables modes of explanatory agency that can resemble human normative practices, with an even bolder refrain in normative contexts of justification and control. This doesn't mean they have subjective experience or intentionality; it means their structural affordances can produce outputs we interpret normatively.

Systems with Recurrent Memory: Architecture and Functional Attribution

Recurrent neural networks (RNNs) and their modern extensions —long-short-term memory networks (LSTMs) and gated recurrent units (GRUs)— exist to process sequential data with a degree of temporal retention, allowing information to persist over time. Unlike feedforward neural networks, RNNs are defined by a hidden state, which is a game-changer, since input over time implies that this hidden state is bound to evolve. This means that RNNs can not only retain the context of each sequence but can also learn how that context affects predictions after processing the sequence. For example, historical inputs through the current sequence and predicted behavior make RNNs especially useful in audio recognition, language modeling, or time series prediction, as past inputs are likely to have a greater impact on future predictions than more static features represented in non-sequential data.

From a computational perspective, this memory is neither semantic nor experiential, but a dynamic process of internal signal propagation and weight adjustment. However, what matters normatively is that these architectures enable causal attributions over time: they allow tracing how previous states or features contribute to current decisions. This is particularly true when combined with attribution techniques such as Layered Relevance Propagation (Bach

et al., 2015), Integrated Gradients (Sundararajan *et al.*, 2017), or attention mechanisms (Vaswani *et al.*, 2017), which offer human-interpretable reconstructions of the internal causal paths followed by the model.

As Jakulin and Bratko (2003) argue, the identification of prediction-relevant features in complex models can support causal interpretations under specific formal conditions, even without invoking psychological intentionality. This computationally based traceability, when integrated into normatively sensitive environments, supports our thesis: that some systems with memory and attribution mechanisms can meet the structural conditions required for quasi-moral treatment. The issue is not whether these systems have moral understanding, but rather whether they can be integrated into justificatory practices in ways that functionally resemble the expectations we place on human agents in institutional contexts.

While these approaches were initially designed to bring transparency to deep learning systems, their structure could offer a broader philosophical restructuring. As Lipton (2018) and Mittelstadt *et al.* (2019) point out, interpretability tools allow us to read models and can serve as normative bridges between algorithmic behavior and external justification practices. Understood in this way, attribution tools can be considered structural, rather than functional, mechanisms that reconstruct causal links without presupposing any mental content or intention. Rather, they provide technical conditions for evaluation, auditability, and constraints—qualities attributable to what we call quasi-moral agency.

We therefore suggest that, in situations where justification is formally required (e.g., resource distribution based on social benefits, early identification of people at risk of suicide, algorithmic filtering for democratic purposes), these architectures can be considered functionally normalizable. Not because they exhibit moral agency, but because they provide technical infrastructures that can be integrated into human evaluative, deliberative, and normative practices. This represents a limited, but still operational, form of technologically mediated structural accountability, and is faithful to the idea that responsibility can be distributed between humans and

machines when technical systems are effectively designed (Floridi 2016; Coeckelbergh 2010; Gunkel 2012).

Quasi-Moral Agency: Philosophical Definition

Building on previous conceptual contributions, we seek here to consolidate these perspectives under the unified category of quasi-moral agency. Over the past decade, various authors have proposed this intermediate category —under different terminologies— to account for the normative role of artificial systems in human-regulated environments. Gunkel (2012), Coeckelbergh (2010, 2022, 2025), and Floridi (2013, 2016) have theorized how certain artificial agents, though lacking intentionality or consciousness, can be meaningfully situated in ethical and institutional practices. Gunkel introduces the term quasi-agency to stress the relational insertion of systems into human communication and accountability networks. Coeckelbergh argues that responsibility can be ascribed based on a system's embeddedness in moral contexts, while Floridi advances the notion of morally relevant artificial agents, emphasizing their effects on social relations. Semler (2022) further refines this perspective by suggesting that artificial agents may fulfill functional conditions for moral agency —such as responsiveness to moral reasons— even in the absence of sentience. Taken together, these ideas support the claim that certain AI systems can act as functional participants in normative frameworks, occupying a quasi-moral position without possessing full moral subjectivity.

Metaethical Connection: Can Systems with Structural Attribution Participate in Normative Practices?

Assessing from metaethics, the nature of explanation versus justification needs to be reconsidered in the context of whether artificial systems could realize normatively organized behaviors. Traditionally, moral justification has been linked to existing conscious reasons,

intentional deliberation and normative obligations. However, these criteria have the potential of restricting many functional behaviors, which may lack phenomenological states, but normatively participate in moral practices.

Many notable contemporary metaethical theories have loosened this restriction. For instance, functional moral realism views moral reason as patterns of coherence between states of the environment, beliefs, and desirability, without requiring subjective awareness (Jackson & Pettit, 1995). In this theory, a justification is not required for a normatively valid proposition to be validly interpreted. Thus, in respect to artificial systems that can have structural attribution, we can argue that a system can be considered moral provided all human interpreters identified that its outputs are auditable, explainable, and correctable.

At the opposite end of the metaethical spectrum, expressivist theories view moral claims as expressions of emotions, attitudes or endorsements, rather than assertables structured as truth-apt properties. In this case, applied to artificial systems, moral agency is not conferred, since such systems lack affective or attitudinal states. While in this view, they could not act morally, it is possible that expressivist theories would allow for artificial systems as effectors in evoking or directing normative attitude in humans. By giving performances that humans interpret as reasons for protest, revision, or endorsement, artificial systems act as external normative prompts. This way of rethinking the status of artificial systems closes again what had promised the possibility of inclusion into practical normative environments—not because they "hold" attitudes; but, they can become effective nodes in networks of moral receptivity.

The question is not whether these systems can "have" reasons, rather, do their architecture allow for humans to treat them as if they can, and thus be part of practices that have operative relevance for justifying. Floridi (2016) advances a similar position by discussing functional morality, where artificial agents are products of moral practices, based on their causal influence, not based on subjective capacities.

When a system with memory builds relational hierarchies of causal dependencies, through techniques such as Layer-wise Relevance Propagation (Bach *et al.*, 2015) or attention mechanisms (Wiegreffe & Pinter, 2019), one can interpret these as limited forms of instrumental rationality. Although not deliberative, such mechanisms recreate ordered relations between stimuli, processing, and behavior, allowing human evaluators to make sense of the system's behavior.

As long as the membership structures can be read, audited, and related to external standards, the system could be understood to be part of a reason-governed regime (albeit weak) (Lipton, 2018; Mittelstadt *et al.*, 2019).

This doesn't implicate unequivocal moral agency, nor does it ascribe moral interiority to the system. In this kind of scenario, we find a structurally responsible form of technological mediation – the artificial agent can be an artifact that assists with justification processes in normatively bounded settings (Coeckelbergh, 2010).

In ethically charged arenas —such as with the distributive allocation of social benefits, predicting suicide risk among a population, or filtering for democratic content— this is not only plausible it is, pragmatically, a necessity. As Gunkel (2012) argues, in moral terms, the relevance of artificial systems does not depend on what they are, but rather what they do with human moral networks.

Cases and Applications

Before analyzing AI-driven cases, it is helpful to recall that structurally justified decision-making has long existed in human institutions. Simple algorithms like paper-based decision trees used by medical committees to allocate scarce resources have always demanded transparency, fairness, and public justification. These historical precedents show that functional responsibility does not require subjectivity, and that the demand for structural attribution is not unique to artificial systems—but arises wherever decisions significantly affect human lives.

❖ Risk Scoring Systems for the Allocation of Limited Social Resources

Algorithms based on recurrent architectures, trained with longitudinal data such as housing instability, income fluctuations, and family history, are increasingly used by municipal agencies to prioritize access to shelters, food assistance, or child protection. In one illustrative case, a system might classify a single mother with a history of eviction as low priority, excluding her from assistance until she and her children were notified by the state through emergency services. The classification is derived from learned associations between frequent residential mobility and the risk of voluntary abandonment.

Although no human actor made the explicit decision to deny support, the family's exclusion was the result of automatic pattern recognition. This raises the ethical question of whether the system's prioritization process can be reconstructed and justified based on the weight and relevance of the variables involved. The system does not act intentionally, but the normative space it generates is subject to structural interrogation, an interrogation that cannot be dismissed as a correction of purely technical errors.

❖ Early Detection of Suicide Risk on Digital Platforms

On social media platforms, neural networks trained on users' message history, browsing behavior, and time spent online are used to estimate the risk of suicide or self-harm. In some cases, these systems automatically trigger alerts to family members, schools, or local authorities. Their sequential architecture allows them to detect cumulative patterns of distress that might go unnoticed in isolated posts.

However, this capability raises ethical concerns about how the system recognizes and evaluates personal and emotional signals. It is not enough to claim that a false alarm or an undetected alert was a technical failure: what matters is whether the system can be interrogated—that is, whether it can offer traceable reasons for its classification or omission. A system capable of reconstructing its decision

trajectory using attribution mechanisms such as attention weights or integrated gradients approaches the threshold of quasi-moral agency.

❖ Filtered and Moderated Algorithms in Democratic Moments

Some governments and private platforms use machine learning models with memory and attention mechanisms to filter content, reduce the visibility of messages, or prioritize specific narratives, especially during crucial democratic processes such as elections or parliamentary debates. A message flagged as "populist" or "misinformative" based on past electoral patterns may be automatically relegated or hidden, without a clear avenue for recourse or appeal.

These systems, while lacking intentionality, exert a structural influence on public discourse. By drawing on "deep memory" traces from previous elections, their impact on visibility and participation in discourse is cumulative and opaque. In such cases, quasi-moral agency emerges as a necessary category: not because the system is moral, but because the absence of structural attribution mechanisms undermines democratic accountability. Therefore, facilitating causal reconstruction is not optional, but essential to guarantee the possibility of procedural defensibility and the protection of communicative rights.

Objections and Limitations

Assigning moral significance to artificial systems, particularly based solely on their internal capacity to make meaningful causal attributions, raises major conceptual and practical issues. The authors raise questions about normativity and anthropomorphism with complex AI, while there are several broader normative issues surrounding the establishment of quasi-moral statuses. Below, we address some relevant objections and attempt to elucidate our position from a functionalist and metaethical perspective.

❖ Functional Attribution without Anthropomorphization or Ethical Inflation

A major concern raised in the literature is that by starting to use terms like agency, responsibility, or attribution, we might be anthropomorphizing systems that lack consciousness, intention, or moral self-reflection. We agree that this represents a real risk and emphasize that the notion of quasi-moral agency, as developed here, is not ontological but functional: it refers to the position a system occupies within regulated normative practices, not to its internal states.

Gunkel (2012) and Coeckelbergh (2009; 2010) have argued that the real question is not whether robots are moral agents, but whether they operate in environments where their outcomes carry moral weight and must be publicly justified. We concur with this perspective: we do not claim that artificial systems possess moral subjectivity. Instead, we assert that specific technical architectures can be used in ethical practices that require openness, auditability, and accountability, allowing for causal traceability and justification.

However, a more pressing concern is the danger of overextension: if quasi-moral agency is defined too broadly, we risk attributing moral relevance to trivial or clearly inappropriate cases, such as a 12-year-old child heedlessly executing an insurance algorithm or a medical device that automatically triggers a health alert. These examples, though counterintuitive, help clarify the scope of our proposal. We argue that quasi-moral agency is not a property of systems per se, but rather a relational attribution that depends on: the normative sensitivity of the context (e.g., healthcare, democratic processes), the technical affordances of the system (e.g., internal memory, structural attribution mechanisms), and the institutional frameworks that assign responsibility and enable justification.

This means that many cases with ethically relevant outcomes do not require a new form of ethics. Traditional principles of accountability —e.g., that a certified employee should not delegate sensitive tasks to unqualified individuals— remain fully valid. Our position is not that these principles should be discarded or replaced, but rather that in some technology-mediated environments, especially where

outcomes are the result of significant automated processes, we may need metaethical tools to clarify how agency, responsibility, and justification are distributed between human and nonhuman actors.

Instead of claiming an ethical shift, we are advocating for a metaethical shift, a reconception of normative relevance in practice; an appreciation of more and more participants in normative relevant practice, and a willingness to rethink the depth of the existing accountability structures that pertain to highly proficient systems. Establishing these structures does not entail creating novel moral theories for each case but, being attentive to how traditional ethical paradigms are either extended or stressed by new technological configurations.

There is a conceptual risk in using attribution theories developed to explain inferences about human mental states (such as those of Heider [1958] or Weiner [1986]) in the context of artificial systems without subjectivity. This article does not aim to translate these theories in their mentalistic sense, but rather to recover their structural logic to understand the mechanisms of external justification. The notion of technical attribution proposed here does not imply intention or consciousness, but rather a functional capacity to reconstruct auditable causal chains. This rearticulation reflects a repertoire of responsibility, while at the same time it can be explained.

Thus, rather than suggesting a "new ethics" associated with machines, we are concerned with strengthening existing accountability regimes in technical and distributed contexts. Quasi-moral agency functions as a useful term supporting external scrutiny and institutional responsibility, but not at the expense of an artificially reified sense of subjectivity.

❖ Metaethics without Ethical Reinvention

A second objection, closely related to the first, is that recognizing the quasi-moral relevance of artificial systems might wrongly suggest the need for an entirely new ethical framework.

Quasi-moral agency refers to functionality, not the assertion that the artificial systems that are connected to it have moral status

in their own right. It identifies situations in which a system's behavior (and accompanying reasoning), despite being without intention, can be recounted, explained, and sanctioned normatively, based on typical sources of meaning and accepted behavioral standards within human institutional practices. This distinction is important because, as Firt (2025) points out, systems can exhibit a level of autonomy, conscience, or expressions of moral emotions, yet still lack the kind of embeddedness in human moral communities that would justify classifying them as full moral agents.

However, ethically significant consequences do not, in and of themselves, imply that traditional ethics is insufficient. Cases such as a child unintentionally triggering an outcome through a predefined algorithm, or an automatic blood pressure monitor provoking life-altering medical actions, are troubling, but not necessarily beyond the scope of existing standards of evaluation.

Our response is that the position presented here is meta-ethical, not substitutional. We do not advocate a new moral theory; rather, we attempt to specify the contexts in which artificial systems can fulfill certain structural functions within established moral frameworks. The concept of quasi-moral agency constitutes such functionality, not the claim that the artificial systems that connect to it have moral status in their own right. It designates cases in which a system's behavior, although lacking intention, can be reconstructed, explained, and normatively evaluated within institutionalized practices.

In this way, the Kantian difference between behaving according to reason and acting from reason is a useful way to tell the difference between following rules and having a real moral cause to do so. Kant says, "Only a rational being has the capacity to act according to the representation of laws" (Kant, 1997).[9] This implies that

9 In Kant's strict sense, practical rationality implies moral autonomy, freedom, and a sense of duty – capacities that no artificial system possesses. However, certain AI systems can operate according to reason (*nach der Vernunft*), that is, follow rational or normative rules in a structured manner, without acting from reason (*aus Vernunft*). This distinction allows us to recognize a limited form of functional rationality without attributing to them moral agency in the full sense.

an agent can act in a principled but not necessarily moral manner. Korsgaard (2009) continues to discuss this distinction by saying that moral autonomy requires something above alone just conforming a behavior. It must involve a reflective endorsement of the principles that govern action. As artificial systems are unable to reflect in this way, they cannot act from reason, even if they are designed to act in a principled manner that is consistent with human behavior in terms of fairness and justification. Thus, their quasi-moral standing is not attained from an internal state, but rather from their structural legibility with respect to accountability, as well as the institutionally enabled conditions for accountability.

Accordingly, we do not argue that new types of ethics should be developed for every system whose outcomes carry moral weight. Rather, we propose that the scope of ethical scrutiny should occasionally be expanded to include technically mediated actions, to the extent that these actions acquire meaning within human practices. What is needed is not an alternative ethics, but a better understanding of how existing moral languages apply to new agents of consequence, a task that belongs to metaethics.

❖ Weak Justification and Limited Interpretability

One of the biggest technical problems is that current ways of explaining model decisions, like integrated gradients, attention mechanisms, and Layer-wise Relevance Propagation (LRP) (Montavon *et al.*, 2019), don't provide full explanations in a complete sense. Instead, they give estimates of feature value or statistical influence after the fact. These estimates are often affected by the model's design, initial conditions, and input changes, and they may be challenging for non-expert stakeholders to understand (Krishna *et al.*, 2025; Rudin, 2019; Lipton, 2018).

This worry is both real and important. We don't say that these methods give strong moral or epistemic reasons for doing things. We say instead that they allow for weak kinds of structural explanation that are enough to set basic levels of auditability and institutional oversight. These mechanisms don't show profound semantic

meanings or purposeful reasons, but they do establish causal scaffolding that may be looked at, questioned, or limited from the outside.

In this way, their normative role is not foundational but instrumental: they are instruments to make sure that high-stakes automated systems stay within a range of public justification, even if that justification is temporary, rough, or mediated by people. The focus is not on overestimating the epistemic efficacy of these tools but on using their structural affordances for regulatory and deliberative purposes, especially in situations where the effects of decisions —like how resources are allocated or health interventions— need to be clear.

The focus on limited but functionally useful forms that deserve explanation is similar to that proposed in research on artificial morality. For example, Misselhorn (2022) provides a broad overview of the ethical issues raised by incorporating moral decision-making capacity into AI agents. He emphasizes that artificial morality is not a hypothetical issue but a necessary consideration, raising serious questions about whether it is possible to align AI moral values with those of humans or whether a specific moral system must be developed for machines, which we have already dismissed here.

❖ Displacement of Human Responsibility

One of the major technical issues is that we do not provide explanations of model decisions in the full sense. It is true that integrated gradients, attention processes, and Layer-wise Relevance Propagation (LRP) do provide some sort of estimates of one type of feature value or other form of statistical influence, but they all do this as a *post hoc* account of model decisions. In other words, estimates of past decisions will always be influenced by the design of the model itself, the initial conditions of the model, and changes to the input, which can make it challenging for non-expert stakeholders to understand the effects on decisions in the model (Krishna *et al.*, 2025; Rudin, 2019; Lipton, 2018).

Worries about the limited value of LRP and other similar "explanatory methods" are real and can be significant for accountability because we do not claim that there are strong moral reasons or epistemic reasons for developing local forms of humanly useful descriptiveness about model decisions. Rather, we stated that local forms of explanatory knowledge can be a weak kind of structural explanation that is sufficient to put into play basic levels of audit-ability and institutional scrutiny/standards.

These local mechanisms do not provide deep semantic issues or agency reasons in their own construction, but they do provide local forms of constraints upon the causal scaffolding of the model that are subject to external inquiry, some degree of questioning, limitations etc. In this way, their normative function is not primary but instrumental; they are instruments that open up the possibility for ensuring that automated systems with high-stake decisions are subject to some type of range of public accountability or justification for their decisions – If even for a short time, or if somewhat rough, or indirectly and through some type of human agency.

As articulated by Tsamados *et al.* (2024), the proposed structural explanation contributes to realizing the broader objective of ensuring meaningful human control. Their shared framework redirects the focus from simple forms of supervision to more advanced forms of human–AI teaming. This alignment is critical in contexts where the interpretability and auditability of algorithmic behavior are essential to support ethically responsible oversight.

Our goal is not to view these tools as overestimating demonstrable epistemic possibilities. Rather, we seek to utilize their structural possibilities precisely as such: possibilities that are especially useful for regulatory and deliberative purposes, especially in contexts where the transparency of decision outcomes is most relevant within the regime of actions we, as agents, perform (e.g., triage prioritization in AI-assisted healthcare systems or algorithmic eligibility decisions in welfare distribution).

Establishing conceptual thresholds to identify when an artificial system can be considered a quasi-moral agent is a central task for advancing this proposal. Rather than assuming that any technically complex tool qualifies as such, we propose a narrower approach that considers three main criteria: (1) the system's capacity to generate internal causal chains that can be reconstructed and audited, (2) its functional insertion into institutional contexts where practices of public justification are required, and (3) its participation in decisions that produce normative consequences relevant to human beings.

From this perspective, it is not simply a matter of distinguishing between moral and non-moral agents, but of recognizing that certain systems occupy intermediate positions within social structures that require traceability, control, and normative evaluation. A decision tree used by a medical committee to distribute scarce resources, for example, already embodies this type of logic. What we propose is to extend this logic to complex systems with technical attribution, provided they meet conditions that allow for structural scrutiny and external justification.

This delimitation does not seek to arbitrarily multiply morally relevant entities, but rather to propose an operational framework for identifying those systems whose function requires shared responsibilities between humans and technologies. In this way, the concept of quasi-moral agency does not replace human responsibility, but rather helps identify critical points where this responsibility must be exercised and articulated institutionally.

Based on the preceding discussions, both theoretical and applied, we can more clearly outline the scope and implications of the concept of quasi-moral agency. Based on the technical attribution mechanisms, traceability criteria, and institutional contexts in which these technologies operate, the need to integrate new metaethical frameworks that accompany —but do not replace— traditional ethical principles becomes evident. With this, we are in a position to propose some general conclusions about the normative role of automated systems in human practices of justification.

Conclusion: Quasi-moral agency as a Metaethical Framework for Attributable Systems

This article suggested that some artificial systems, particularly those with some kind of internal memory and mechanisms of technical attribution, should be understood, within some normative scenarios, as quasi-moral agency, as a metaethical category. We are not implying any reification of artificial subjectivity, but rather suggesting a functional framework for supporting practices involving explanation, auditability and institutional accountability in environments shaped by automated decisions.

Quasi-moral agency —unlike moral agency— do not necessarily entail intention, consciousness or rational deliberation. Quasi-moral agency simply reflects the system's structural ability to produce outcomes that humans can reconstruct, evaluate or revise within normative parameters.

While the idea of machines participating in moral practices has been previously explored by authors like Gunkel, Floridi, and Coeckelbergh, our contribution lies in articulating how technical attribution mechanisms —such as attention, LRP, or integrated gradients— can ground minimal forms of structural justification, enabling a form of technologically mediated responsibility.

This position has conceptual and design implications. Conceptually, it invites a rethinking of metaethics: not as a search for moral interiority, but as the identification of the conditions under which agency and responsibility are distributed across humans and artifacts. In design terms, it reinforces the duty to construct systems that are not only technically robust but also normatively intelligible – systems that can be interrogated, contested, and adapted within institutional settings.

We do not claim that these systems are moral agents in a full sense, nor that new ethical theories must be invented to deal with them. Instead, we argue for better integration of existing ethical standards through metaethical tools that clarify how and when artificial outputs become subject to normative scrutiny. The ethics at stake is not internal to the machine, but distributed: it resides in its

traceable consequences, its capacity for structural responsiveness, and its position in human-governed systems of justification.

In the end, technical attribution does not supplant ethics – it supports it. It provides the framework of critical interpellation where automated systems increasingly determine outcomes affiliated with essential human interests. Framing such systems as quasi-moral agents is not to inflate and/or amplify their stature; it is to perform some accountability and elevate the standard of accountability.

This article has analyzed the structural conditions that enable artificial intelligence systems to act as quasi-moral agents; however, a central problem that threatens accountability has been identified here: the epistemic gap that separates model creators from their institutional users and, ultimately, from the public. In most cases, the ability to audit, question, or simply understand how algorithms work eludes even those charged with their implementation, as observed in cases ranging from the allocation of social benefits to content moderation on digital platforms.

Looking ahead, research must focus on closing this gap. The development of attribution mechanisms is required that meet rigorous formal traceability standards while also being practically applicable across diverse boundaries: professional, disciplinary, and civic. Only through solutions that enable a transversal understanding can we ensure that structural accountability remains a fundamental pillar in the ethical evolution of AI.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, *10*(7), e0130140: 1-46. https://doi.org/10.1371/journal.pone.0130140.

Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology*, 35, 12: 1-30. https://doi.org/10.1007/s13347-022-00510-w.

Coeckelbergh, M. (2009). Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents. *AI & Society*, *24*: 181-189. https://doi.org/10.1007/s00146-009-0208-3.

(2010). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology*, *12*: 209-221. https://doi.org/10.1007/s10676-010-9235-5.

(2022). Democracy: Echo Chambers and Machine Totalitarianism. *The Political Philosophy of AI*. Polity Press.

(2025). AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications. *Social Epistemology*: 59-71. https://doi.org/10.1080/02691728.2025.2466164.

Dastani, M., & Yazdanpanah, V. (2022). Responsibility of AI Systems. *AI & Society*, *38*: 843-852. https://doi.org/10.1007/s00146-022-01481-4.

Dieter, T. R., & Zisgen, H. (2023). Evaluation of the Explanatory Power Of Layer-wise Relevance Propagation using Adversarial Examples. *Neural Processing Letters*, 55: 8531-8550. https://doi.org/10.1007/s11063-023-11166-8.

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning [Preprint]. *ArXiv*: 1-13. https://doi.org/10.48550/arXiv.1702.08608.

Felin, T., & Holweg, M. (2024). Theory is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science*, *9*(4): 346-371. https://doi.org/10.1287/stsc.2024.0189.

Firt, E. (2025). What Makes Full Artificial Agents Morally Different. *AI & Society*, *40*: 175-184. https://doi.org/10.1007/s00146-024-01867-6.

Floridi, L. (2013). The Morality of Artificial Agents. *The Ethics of Information*. Oxford University Press: 134-160. https://doi.org/10.1093/acprof:oso/9780199641321.003.0007

(2016). Faultless responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, A 374 (2083), 20160112: 1-13. https://doi.org/10.1098/rsta.2016.0112.

(2023). AI as *Agency Without Intelligence*: On ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, 36, 15: 1-7 https://doi.org/10.1007/s13347-023-00621-y.

(2025). AI as Agency Without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis. *Philosophy & Technology*, *38*, A 30: 1-27. https://doi.org/10.1007/s13347-025-00858-9.

Greiner-Petter, A., Youssef, A., Ruas, T., Miller, B. R., Schubotz, M., Aizawa, A., & Gipp, B. (2020). Math-Word Embedding in Math Search and Semantic Extraction. *Scientometrics*, 125: 3017-3046. https://doi.org/10.1007/s11192-020-03502-9.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5): 1-42. https://doi.org/10.1145/3236009.

Gunkel, D. J. (2012). Moral Agency. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*: 15-92. MIT Press. https://doi.org/10.7551/mitpress/8975.003.0004.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. Wiley.

Jackson, F., & Pettit, P. (1995). Moral Functionalism and Moral Motivation. *The Philosophical Quarterly*, *45*(178): 20-40. https://doi.org/10.2307/2219846.

Jakulin, A., & Bratko, I. (2003). Analyzing Attribute Dependencies. N. Lavrač, D. Gamberger, L. Todorovski & H. Blockeel (eds.), *Knowledge Discovery in Databases: PKDD 2003*, vol. 2838: 228-240. Springer. https://doi.org/10.1007/978-3-540-39804-2_22.

Kant, I. (1997). *Groundwork of the Metaphysics of Morals*. Translated by M. Gregor. Cambridge University Press.

Kelley, H. H. (1967). Attribution Theory in Social Psychology. D. Levine (ed.), *Nebraska Symposium on Motivation*, vol. 15: 192-238. University of Nebraska Press.

(1973). The Processes of Causal Attribution. *American Psychologist*, 28: 107-128. https://psycnet.apa.org/doi/10.1037/h0034225.

Korsgaard, C. M. (2009). Autonomy and Efficacy. *Self-Constitution: Agency, Identity, and Integrity*. Oxford Academic: 81-108. https://doi.org/10.1093/acprof:oso/9780199552795.003.0005.

Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., y Lakkaraju, H. (2025). *The Disagreement Problem in Explainable Machine Learning: A practitioner's Perspective* (v. 6) [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2202.01602.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2010). Feature Selection: A Data Perspective. *ACM Computing Surveys* (*CSUR*), *9*(4), A 39: 1-45. https://arxiv.org/abs/1601.07996.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Communications of the ACM*, *61*(10), 36-43. https://doi.org/10.1145/3233231.

Malle, B. F. (2011). Time to Give Up the Dogmas of Attribution: An Alternative Theory of Behavior Explanation. *Advances in Experimental Social Psychology*, *44*: 297-352. https://doi.org/10.1016/B978-0-12-385522-0.00006-8.

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, *15*(9), 517: 1-34. https://doi.org/10.3390/info15090517.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267: 1-38. https://doi.org/10.1016/j.artint.2018.07.007.

Misselhorn, C. (2022). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*. Cambridge University Press: 31-49. https://doi.org/10.1017/9781009207898.005.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*: 279-288. https://doi.org/10.1145/3287560.3287574.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2017). Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65: 211-222. https://doi.org/10.1016/j.patcog.2016.11.008.

(2019). Layer-Wise Relevance Propagation: An Overview. W. Samek, G. Montavon, A. Vedaldi, L. Hansen, & K. R. Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700: 193-209. Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_10.

Nemavhola, A., Chibaya, C., & Viriri, S. (2025). A Systematic Review of CNN Architectures, Databases, Performance Metrics, and Applications in Face Recognition. *Information, 16*(2), 107. https://doi.org/10.3390/info16020107.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High-Stakes Decisions and use Interpretable Models Instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x.

Salem, F. M. (2022). Gated RNN: The Gated Recurrent Unit (GRU) RNN. *Recurrent Neural Networks*. Springer, Cham. https://doi.org/10.1007/978-3-030-89929-5_5.

Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. W. Samek, G. Montavon, A. Vedaldi, L. Hansen y K.-R. Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Springer: 5-22. https://doi.org/10.1007/978-3-030-28954-6_1.

Semler, J. (2022). Artificial Quasi Moral Agency. Proceedings of the 2022 AAAI/ACM Conference on *AI, Ethics, and Society* (AIES '22): 913. ACM. https://doi.org/10.1145/3514094.3539549.

(2025). Moral Agency without Consciousness. *Canadian Journal of Philosophy*, *54*(5): 403-422. https://doi.org/10.1017/can.2025.10008.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning* (*PMLR*) 70: 3319-3328. https://proceedings.mlr.press/v70/sundararajan17a.html.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The Ethics of Algorithms: Key Problems and Solutions. *AI and Society*, *37*(1), 215-230. https://doi.org/10.1007/s00146-021-01154-8.

Tsamados, A., Floridi, L., & Taddeo, M. (2024). Human Control of AI systems: From Supervision to Teaming. *AI Ethics*, 5: 1535-1548. https://doi.org/10.1007/s43681-024-00489-4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30: 1-15. *arXiv.* https://arxiv.org/abs/1706.03762.

Weiner, B. (1986). An Attributional Theory of Achievement Motivation and Emotion. *An Attributional Theory of Motivation and Emotion*. Springer: 159-190. https://doi.org/10.1007/978-1-4612-4948-1_6.

(2012). An Attribution Theory of Motivation. P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (eds.), *Handbook of Theories of Social Psychology*. Vol. 1. Sage: 135-155. https://doi.org/10.4135/9781446249215.n8.

Widrow, B., & Katz, E. P. (2025). Long-Term and Short-Term Memory. *Cognitive Memory*. Springer Series on Bio- and Neurosystems. Vol. 17. Springer, Cham: 151-156. https://doi.org/10.1007/978-3-031-80939-2_21.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not Explanation [Preprint]. *arXiv*. https://arxiv.org/abs/1908.04626.