Teorema, 2025, 44(1), pp. 1-27
https://doi.org/10.30827/trif.33614

Neural representations: A Mechanistic Challenge to Suárez's Inferential Conception

Aníbal M. Astobiza

Universidad de Granada

Abstract: This article examines Mauricio Suárez's inferentialist account of scientific representation in light of recent advances in neuroscience and artificial intelligence (NeuroAI). While it offers valuable pragmatic insights, I argue it is insufficient to capture the dynamic, computational, and biological nature of neural representations. Drawing on the mechanistic, functionalist, and representationalist (MFR) approach and empirical findings, I maintain they are not mere abstract entities but are embodied in the physical and functional properties of neural systems. I challenge arguments against the necessity and sufficiency of similarity and isomorphism, highlighting computational transformations, functional roles, and the directionality of processing that shape content. The Hodgkin–Huxley model and neurocomputationalism employing artificial neural networks support the MFR and challenge Suárez's inferentialism. I conclude that a mechanistic and computational understanding provides a more comprehensive and empirically grounded framework for modelling and representation in the mind sciences.

Keywords: *neural representation, mechanistic explanation, computational models, inferential account, cognitive neuroscience, artificial intelligence, Mauricio Suárez.*

Resumen: Este artículo examina la explicación inferencialista de la representación científica de Mauricio Suárez a la luz de avances recientes en neurociencia e inteligencia artificial (NeuroAI). Aunque aporta valiosas perspectivas pragmáticas, sostengo que es insuficiente para captar la naturaleza dinámica, computacional y biológica de las representaciones neuronales. Apoyado en el enfoque mecanicista, funcionalista y representacionalista (MFR) y en hallazgos empíricos, sostengo que no son meras entidades abstractas, sino que están encarnadas en propiedades físicas y funcionales de los sistemas neuronales. Cuestiono argumentos contra la necesidad y suficiencia de similitud e isomorfismo, destacando transformaciones computacionales, roles funcionales y direccionalidad del procesamiento que configuran contenido. El modelo Hodgkin-Huxley y el neurocomputacionalismo con redes neuronales artificiales apoyan el MFR y desafían el inferencialismo de Suárez. Concluyo que una comprensión mecanicista y computacional ofrece un marco más completo y empírico para la modelización y la representación en las ciencias de la mente.

Palabras clave: *representación neuronal, explicación mecanicista, modelos computacionales, enfoque inferencial, neurociencia cognitiva, inteligencia artificial, Mauricio Suárez.*

Received: 24 April 2025 **Accepted**: 16 July 2025 **Published**: 23 July 2025

Teorema. Revista Internacional de Filosofía
ISSN/ISSN-e: 1888-1254

* This article has been funded by Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía (Grant EMEC_2023_00442).

1. Introduction

For many years, the goal of cognitive science has been to unravel the complex mechanisms underlying intelligent behavior. Recently, artificial intelligence (AI), particularly artificial neural networks (ANNs), has ushered in a new era of computational modeling in neuroscience to achieve this goal. This development has led to the rise of a field known as NeuroAI (Zador et al., 2023), which utilizes ANNs to simulate and understand brain computations underlying cognitive abilities. However, the philosophical foundations of modeling and representation in NeuroAI remain debated, despite strong empirical evidence supporting certain claims.

As we will see in sections III and V, one of those claims is that neural representations[1] are regularly observed, manipulated, and measured by experimental neuroscientists and computer scientists working in AI (Tootell et al., 1988; Lewis and Kristan, 1998; Matsumato and Komatsu, 2005; Liu et al., 2012; Spillman, 2014; Shi et al., 2023; Johnston and Fusi 2023; Courellis, Minxha, Cardenas, *et al.* 2024). Additionally, there are philosophical reasons to argue that these representations are not mere theoretical postulates (Artiga, 2023; Piccinini, 2020; Ramsey, 2007; Poldrack, 2021; Shea 2018).

In his recent book, *Inference and Representation: A Study in Modeling Science* (2024), hereafter *IR*, Mauricio Suárez proposes an inferentialist account of scientific representation, emphasizing the pragmatic use of representations in scientific practice and their role in enabling inferences. Suárez provides valuable insights into the nature of representation across various natural and social sciences, exemplified through modeling practices in economics (Phillips-Newlyn machine), ecology (Lotka-Volterra model), astrophysics (models of stellar structure), and the kinetic theory of gases (billiard models). However, the applicability of this inferential account to neural representations in NeuroAI warrants further scrutiny. Suárez is explicit about the scope of his project, stating, "I will assume from the start that the aim of any account of scientific representation is to understand modeling practice... The inferential conception... is, I submit, the best account of representation for the purposes of understanding the practice of modeling in science" (Suárez 2024, p. 4). His focus is therefore on what I termed *imposed representations*: public, socially constructed artifacts like diagrams, equations, and scale models created by scientists for specific inferential purposes (This point is elaborated further at the end of this introduction).

Although *IR* never claims to explain neural representation, the book does offer a *general* theory of scientific representation that is explicitly presented as an alternative to similarity- and isomorphism-based accounts. In a charitable spirit, I therefore treat Suárez's view as a candidate for *extension* to the neural domain and ask a conditional question: *were the inferential conception applied to neural systems, would it illuminate the*

[1] To be more precise, a neural representation, for the purposes of this article, is defined as a pattern of activity in a neural system (the *vehicle*) that systematically co-varies with a feature of the environment, the organism's body, or an abstract problem space (the *content*). Crucially, this pattern of activity must be used by other parts of the organism's cognitive architecture to guide behavior or drive further computations. The central philosophical debate, which this article addresses, concerns the nature of the relationship between the vehicle and its content, and what properties are necessary for it to successfully perform its functional role. For clarity, this article will adhere to the following terminology: "intrinsic neural representation" refers to the biological patterns of activity within a nervous system that carry content. The "MFR account" or "mechanistic models" refer to the scientific and philosophical theories developed by researchers to explain these intrinsic phenomena.

representational capacities that neuroscientists actually measure and manipulate? My answer is negative, and it is negative precisely because *IR* express a theory that denies that structural similarity or isomorphism plays any necessary explanatory role. In other words, my objective could be seen as exploring the possibility of extending Suárez's analysis beyond its intended applications.

Empirical work shows that many canonical neural codes—retinotopic, tonotopic and somatotopic maps, hippocampal place cells, grid cells, head-direction cells, etc.—traffic in stable, topologically preserved correspondences between neural activity and stimulus space (Poldrack 2021). These are not optional conveniences but *mechanistic pre-conditions* for fast, computation-ready inference by the organism itself, as stressed by mechanistic–functionalist and representationalist accounts of cognition (MFR).

Extending Suárez's inferentialist programme to the neural domain is not a *category mistake* but a standard test of theoretical scope. If a general account of scientific representation cannot accommodate the most intensively studied and mechanistically characterised representational system we know—the mammalian brain—its claim to general practice for modeling science is weakened. Hence I examine, *conditional on that extension*, whether the denial of structural similarity and isomorphism could still sustain explanatory power.

Recent empirical work indicates that it cannot. Large-scale Neuropixels recordings show that accurate navigation collapses when the stable, hexagonal phase relations among grid-cell modules are transiently disrupted; animals behave as if *spatial content has vanished*, despite all downstream decision circuitry remaining intact (Vollan et al., 2025). Similarly, one-shot formation of entorhinal maps in novel environments depends on fixed correspondences between visual landmarks and grid-cell firing phases, establishing a veridical mapping after a *single* exposure (Wen et al., 2024). These findings reveal that what carries content for the organism is the maintained topological homology between neural activity and external metric space—a relation the inferentialist view declares explanatorily irrelevant.

A complementary body of work on hippocampal replay deepens the point. Hippocampal networks "compose" new spatial policies offline by binding grid-derived vectors into fresh conjunctive codes; on re-entry to the environment, rodents act optimally *without any additional learning* (Bakermans et al., 2025). Computational analyses and multi-area recordings converge on the interpretation that zero-shot generalisation is possible only because upstream codes preserve structural isomorphism, allowing replay to recombine positions as manipulable vectors (Johnston and Fusi, 2023; Courellis et al., 2024). If those structural correspondences are erased or decorrelated, predictive sweeps and behavioural transfer both fail.

Taken together, these data vindicate the core mechanistic–functionalist–representational (MFR) thesis: neural vehicles *must* instantiate similarity or isomorphism relations—often dynamically enforced—to be usable by the system that generates them (Piccinini, 2020). Because Suárez's account treats such relations as neither necessary nor sufficient, its explanatory resources evaporate precisely where structural mapping is demonstrably indispensable. The MFR framework, by contrast, recognises similarity and isomorphism as contingent but frequently *obligatory* constraints within a full causal-mechanistic explanation; it therefore succeeds where the inferential conception stalls.

Therefore, this article aims to critically examine Suárez´s inferential account of scientific representation in light of recent advancements in neuroscience and AI, particularly within the framework of the mechanistic, functionalist, and representationalist (MFR) account of cognition. I argue that while Suárez´s account offers valuable insights into the pragmatic use of representations in scientific models, it falls short in fully capturing the dynamic, computational, and biological nature of *neural representations*.

Empirical research has identified symmetry (isomorphism) and universality (similarity) as fundamental principles in understanding neural representations (See, section III and IV). Symmetry denotes the invariance and stability of these representations when subjected to certain transformations, ensuring consistent processing across different contexts. Universality implies that common computational principles might underlie cognitive processes across diverse systems—both natural and artificial—and species. This suggests that neural representations are governed by shared structural and functional properties, facilitating a unified understanding of cognition across different domains (Sanborn, Shewmake, Olshausen, and Hillar, 2023; Courellis, Minxha, Cardenas, *et al.* 2024).

Drawing upon the MFR account and empirical findings from recent studies, I propose that Suárez´s arguments against the necessity and sufficiency of similarity and isomorphism in our representational practices do not extend to the domain of neural representation.

Neural representations are not merely abstract entities used for inference, but are embodied in the physical and functional properties of neural systems (and even artificial systems). I highlight the role of computational transformations, functional roles, and the inherent directionality of neural processing in shaping representational content.

Furthermore, I examine how the Hodgkin-Huxley model and the field of neurocomputationalism, which utilize artificial neural networks (ANNs) to model brain computations, provide empirical support for the MFR account and challenge the potential extension Suárez´s inferentialist account to the domain of neural representations (neuroscience).

Through a detailed analysis of the relevant literature and a critical evaluation of Suárez´s arguments, this article aims to contribute to the ongoing dialogue between philosophy and neuroscience regarding the nature of representation and its role in understanding cognition (Baker, Lansdell, and Kording 2022). I ultimately argue that a mechanistic, computational understanding of neural representation offers a more comprehensive and empirically grounded framework for understanding modeling and scientific representation than Suárez´s inferential account, particularly in the context of NeuroAI. Suárez´s inferentialist account focuses on how representations enable practical inferences in the context of science inquiry. According to Suárez, representations do not need to maintain a similarity or isomorphism relationship with what they represent; their value lies in their utility for generating valid inferences. Suárez´s own formulation makes the point explicit: once a source possesses *representational force* toward a target and affords the "specific inferential capacities" that let competent agents draw surrogate conclusions, it already *counts* as a representation, because "neither representational force nor inferential capacity is committed to any means of representation".

Structural likeness can certainly *enhance* a model's epistemic virtues—Suárez grants that isomorphism or similarity "in form are the means of the representations with which we are working" and may underwrite accuracy—but they remain optional embellishments, not part of what *makes* A represent B. Hence my claim stands: on Suárez's inferential conception a representation need not preserve similarity or isomorphism; its value, *qua* representation, lies in enabling the right inferences. This stance contrasts with the MFR perspective, which holds that neural representations are intrinsically linked to the physical and functional properties of target objects. The critical point is that the brain´s method of representation differs fundamentally from that of other cognitive agents, such as scientists. Therefore, it is essential to seek models that most accurately capture how brains build models. In my view, only the MFR account can achieve this.

Computational and mechanistic models in neuroscience, such as the Hodgkin-Huxley model, have demonstrated how the MFR account is the most viable strategy or solution to understanding neural components, neural representations, computational transformations and their specific functional roles. These models not only replicate the observable behavior of neurons, but also provide a mechanistic understanding of the underlying processes. I argue that Suárez´s inferentialist account is insufficient to fully explain neural representations due to its emphasis on pragmatic use over intrinsic properties. Neural representations are not mere abstractions for inference; they are embodied in the physical and functional properties of neural systems.

Central to this article is the distinction between *intrinsic representations*, arising naturally within biological systems (autogenic), and *imposed representations*, intentionally constructed by human agents for scientific purposes. This distinction critically informs my argument that neural representations require fundamentally different explanatory frameworks, such as MFR, distinct from inferential conceptions emphasizing pragmatic inferential utility. Imposed representations are public, artefactual representational tools intentionally constructed by cognitive agents—most notably, scientists—for the purposes of reasoning, prediction, and communication. This category includes maps, diagrams, mathematical equations, and scale models, such as the Phillips-Newlyn machine or the blueprints for the Forth Rail Bridge. Their representational force is established through pragmatic and social conventions, and their value lies in their utility for enabling competent users to draw surrogate inferences about a target system. Suárez's inferential conception, as detailed in *IR*, provides a powerful and nuanced account of this type of representation, focusing rightly on the "practice of modeling in science".

In contrast, *intrinsic representations* are autogenic (self-generating). They are not constructed by an external agent but arise from the inherent causal and functional architecture of a system, such as a nervous system or a sophisticated artificial network. Neural representations—like retinotopic maps or place cell activity—are paradigm examples. They do not represent something *for the organism* in the way a map does for a user; rather, they represent information *for other downstream neural mechanisms* within the system. Their representational content and directionality are not determined by social convention but by their specific causal-mechanistic role in guiding the system's adaptive behavior. It is this class of representation that the Mechanistic, Functionalist, and Representationalist (MFR) account seeks to explain. Therefore, this article argues that these two types of representation operate under fundamentally different constraints. Suárez's arguments against the necessity of similarity and isomorphism are compelling for the flexible, pragmatic domain of imposed scientific models. However, I contend that for intrinsic neural representations, structural correspondence is often not an optional “means” of representation but a mechanistic necessity—a prerequisite for the system to perform computations and generate inferences efficiently and reliably. Consequently, extending the inferential account beyond its intended domain reveals its limitations, and highlights why the MFR framework offers a more complete and empirically grounded explanation for the nature of representation in the sciences of the mind.

This distinction allows me to precisely frame my argument. I do not claim that Suárez's account fails on its own terms. In fact, Suárez himself appears to set aside the domain of intrinsic representation. As he notes, “Neither should we require that a theory of scientific representation be able to explain how humans have evolved the capacity to generate representations, or mental images of the world; although this is an independently interesting issue” (Suárez 2003, p. 226). This article takes up precisely that “independently interesting issue”. I argue that the intrinsic representations that constitute our evolved cognitive capacities operate under a different set of constraints—mechanistic, computational, and biological—that are not fully captured by a purely inferentialist framework. Therefore, our critique is not that Suárez's theory is wrong, but that it is necessarily incomplete. A full account of representation in the sciences of the mind requires a complementary framework, like the MFR account, that can explain the foundational, intrinsic representations upon which the imposed models of scientific practice are ultimately built. Before analyzing Suárez's specific arguments against substantive accounts, it is crucial to first situate his entire project within a broader theory of representation. An anonymous reviewer helpfully directed my attention to the work of Callender and Cohen (2006), whose framework reveals a fundamental limitation in Suárez's approach. Their central claim is that “there is no special problem about scientific representation”; rather, scientific representation is a species of representation in general, that is, mental representations. Following a “General Gricean” strategy, they argue that all non-fundamental forms of representation—such as diagrams, artworks, and scientific models—are derivative. Their ability to represent is inherited from a more fundamental class of representations, which are typically identified with the mental states (intentions, beliefs) of their users . As they put it, the philosophical action lies not in explaining the derivative cases, but in providing a metaphysical account of the fundamental bearers of content. The representational power of any non-mental object, from a stop sign to a scientific model, is constituted by “a stipulation, together with an underlying theory of representation for mental states” (Callender and Cohen 2006, p. 78).

Viewed through this lens, Suárez's inferential conception is a detailed and valuable pragmatic analysis of how a community *uses* a specific class of derivative representations (imposed models). However, by explicitly setting aside the question of how humans evolved the capacity for mental representation, Suárez brackets off the very foundation from which the models he studies derive their representational power. His core concept of “representational force”, for example, is left as a primitive. From a General Gricean perspective, this “force” is precisely the intentionality of the user's mental states, which Suárez’s theory declines to explain. This makes his account philosophically incomplete: it describes the pragmatics of using representations but cannot explain what makes them representational in the first place.

This critique provides a powerful justification for my project. The MFR account, in direct contrast to Suárez's approach, tackles the problem of fundamental representation head-on. It aims to provide a naturalistic, mechanistic explanation of the intrinsic neural representations that form the physical basis of the mental states that ground all other forms of representation. Therefore, the MFR framework is not merely a theory for a different, "subpersonal" domain; it is a theory for the foundational domain of cognition that makes the imposed, inferential practices of science possible. Suárez's attempt to isolate the practice of modeling from its cognitive-neural foundations is thus ultimately untenable from a robustly naturalistic perspective. With this foundational critique in place, we now turn to how the specific constraints of intrinsic representations clash with Suárez's arguments against structural correspondence.

For the sake of clarity and precision, my analysis will adhere to a distinction between three different levels of analysis, a distinction helpfully suggested by an anonymous reviewer. It is crucial to separate: (i) Meta-scientific conceptions, which are philosophical frameworks that describe and evaluate scientific practice. Both Suárez's inferentialism and the MFR account operate primarily at this level. (ii) Scientific theories and models, which are the specific representational tools created by scientists to explain and predict phenomena (e.g., the Hodgkin-Huxley model of the action potential). (iii) Real-world mechanisms, functions, and representations, which are the phenomena in the world that scientists study (e.g., the actual flow of ions across a neuron's membrane). This article's central argument operates by analyzing the relationship between these levels. I contend that Suárez's meta-scientific conception (i) provides an inadequate account of the scientific practice of neuroscience because it mischaracterizes the essential relationship that neuroscientists aim to establish between their scientific models (ii) and the real neural mechanisms (iii) they target. Specifically, I argue that in neuroscience, this relationship must be one of mechanistic correspondence, a feature that Suárez's inferentialism deems non-essential. The MFR framework (i), I propose, offers a more faithful meta-scientific account of this specific scientific practice.

2. Suárez's inferentialist account of representation

Mauricio Suárez, in his book *IR*, traces the genealogy of what he terms the "modeling attitude", a stance toward scientific work and discovery that emerged prominently in the nineteenth century. This modeling attitude emphasizes the creation and use of models as central to scientific inquiry, distinguishing it from earlier or even later periods where models were not as systematically utilized or philosophically scrutinized.

The "modeling attitude" has its roots in the scientific practices and philosophical reflections of the late nineteenth century. Suárez identifies two main schools contributing to its development: the British school, led by figures like James Clerk Maxwell and William Thomson (Lord Kelvin), and the German-speaking school, represented by Heinrich Hertz or Ludwig Boltzmann. These schools, though historically related, evolved in different contexts and contributed distinct insights into the role and nature of models in science.

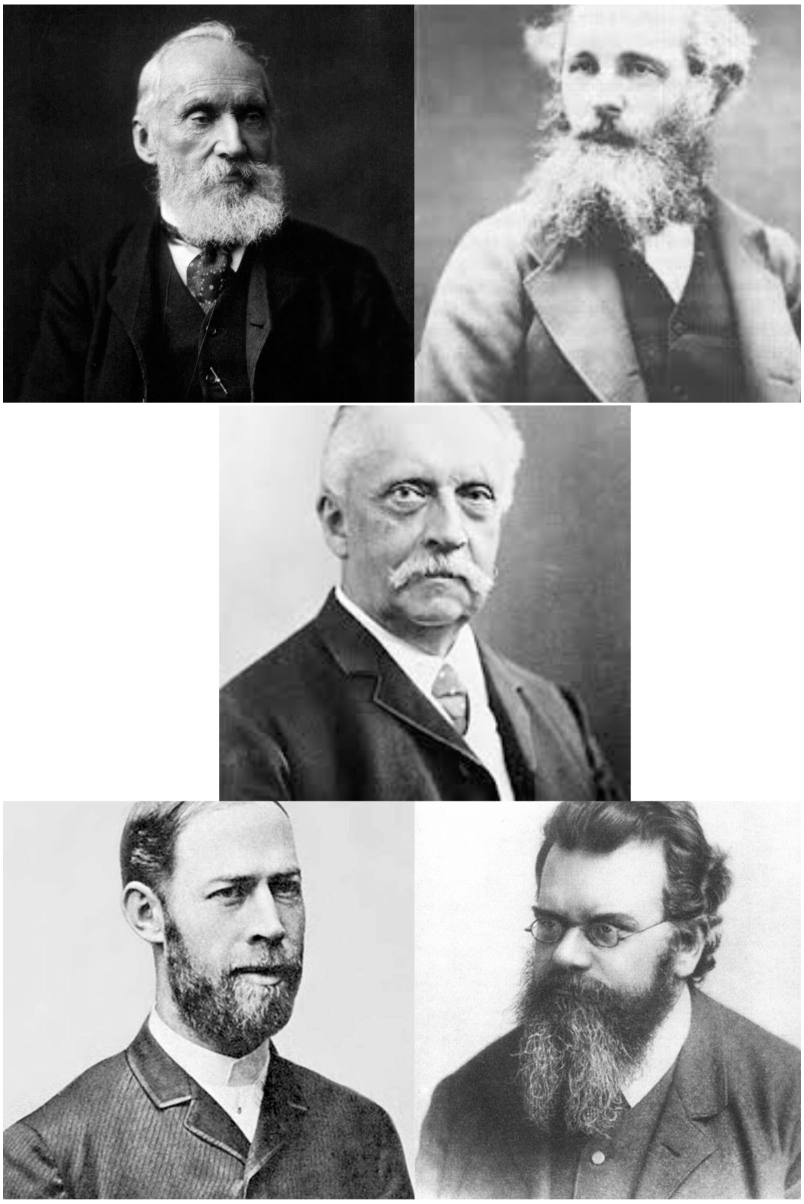

Figure 1: Members of the two schools or traditions in the philosophy of modeling.

Centered around Maxwell and Thomson, the British school emphasized the use of analogical reasoning and concrete physical models to understand phenomena, particularly in the realm of electrodynamics. This school laid the groundwork for a methodological approach that deeply integrated modeling into scientific practice. Hertz and Boltzmann, influenced by Helmholtz, advanced a more abstract and theoretical form of modeling. Their work contributed significantly to the *Bildtheorie* (theory of images), which framed models as formal frameworks with internal structures defined by principles and consequences, emphasizing the plurality of representations and their pragmatic utility in scientific inquiry

Contrary to the standard narrative found in philosophy of science textbooks, the thematic history of the discipline is not a simple linear progression from the received view or syntactic view to the semantic view. Suárez's exploration of the "modeling attitude" (Suárez 2024, p. 20) reveals its enduring influence on scientific methodology. He argues that the principles established by nineteenth-century modelers, such as reasoning by analogy, methodological abstractionism, and the use of formal frameworks, continue to underpin contemporary scientific practices. More importantly, these principles predate the semantic view and the focus on modeling practice that emerged in the 1970s.

The "modeling attitude" enable scientists to create models that are not merely descriptive, but also predictive and explanatory, providing a robust means of understanding complex phenomena. In framing his inferentialist account of representation, Suárez builds upon the historical modeling attitude to propose a deflationary account of scientific representation. This account focuses on the pragmatic use of models to facilitate inferences about target phenomena without focusing into metaphysical questions of reference or denotation. According to Suárez, a true representation is one that allows a competent and informed agent to draw accurate inferences from the model to the target, emphasizing the practical and contextual aspects of scientific modeling. This is what Suarez says explicitly:

> On the inferential conception, a true representation is defined as a representation of a target B by some source A that allows a competent and informed agent to draw surrogate inferences from A to only true conclusions about B. Since the kind and degree of competence as well as the required level of information are essentially dependent not just on the source and the target and their properties but also on the context of inquiry, the expression true model is revealed to be a mere façon de parler: a model (a scientific representation in general), by itself, cannot be true or false. It can be said to be so only in a derivative sense, one that depends on its context of use and application. (Suárez 2024, p. 15).

Mauricio Suárez champions an inferential conception of representation. This conception is characterized by its deflationary and pragmatist stance (Suarez 2024, p. 227), rejecting the notion that representation is solely grounded in substantive relations like similarity or isomorphism. Instead, Suárez emphasizes the pragmatic use of representations in scientific practice and their role in enabling inferences. Suarez uses a terminology that is now standard in scientific representation and modeling studies. It distinguishes between *source* and *target.* The *source* is the object or entity that performs the representational work while the *target* is the represented object or entity. In addition, he distinguishes between the *means* and the *constituent* of scientific representation (Suarez 2024, p. 6). Suárez elaborates this definition in more detail:

> Throughout the book, I shall define the terms source and target as follows. For any pair {X, Y}: if X represents Y, then X is the source and Y the target of the representation. Or, in other words, X is the representational source and Y its representational target. This definition is very minimal. The sources of models can be concrete objects, abstract or fictional entities, mathematical equations or structures, conceptual archetypes, or sets of sentences in some natural or artificial language. The targets can be concrete, abstract, or fictional systems and their states, objects, and/or properties, processes, or experimental data. There is nothing in the properties of these objects or their relations that I suppose is required for them to fulfill their roles; the only requirement on any source-target pair {X, Y} is that "X represents Y" is true. Finally, I make no definitional assumptions as to what makes such a statement true either since its truth makers are likely to be many and varied (Suárez 2024, p. 47).

This explanation provides a minimalist yet flexible framework for understanding representational relationships in the context of scientific practice. However, as will be argued throughout this article, this perspective may be insufficient to fully capture the dynamic, computational, and biological nature of neural representations within the realm of NeuroAI. Suarez in *IR* posits two necessary conditions for representation:

representational force and inferential capacity. Regarding the representational force, the model or source must be directed towards the target, established through its intended use and community practices. This force is a socially established convention maintained by normative practices within a scientific community. Elsewhere, Suarez (2003) argues that neither isomorphism nor similarity can ground the representation relation for models and they have a requirement of directionality: "models are about their targets, but targets are not about models" (Frigg and Nguyen, 2017 p. 55). In other words, *sources* or models are free and independent of any strict relationship with *targets* and are exclusively due to vertical (within models) or horizontal (between models) inferential rules.

Suárez argues that representation is fundamentally about drawing inferences from a source (the model) to gain knowledge about a target (the phenomenon being modeled). He views models as tools for conveying information, not necessarily requiring perfect accuracy or mirroring of reality. Inferential capacity means that the model must enable informative inferences about the target, going beyond mere denotation. Suárez elaborates:

> The source must be the right kind to allow informative inference regarding the target. The condition does not require the inferences to be infallible or to be true conclusions about B, but there is an important clause that requires them to reveal aspects of B that do not follow from the mere existence of a representational relation. There must be other informative inferences about B that can be drawn from A for "A represents B" to be true (2024, p. 9).

This perspective implies that representation is not merely about having a one-to-one correspondence between the model (A) and the target (B). Instead, it is about the capacity of the model to allow the user to draw relevant and novel inferences about the target. These inferences should go beyond what is immediately obvious from the representational relationship itself. This criterion underscores the importance of the model's utility in scientific practice, where the goal is often to uncover new insights and generate understanding that extends beyond the initial representation. Suárez´s view challenges notions of representation that emphasize similarity or isomorphism. By focusing on inferential capacity, he shifts the emphasis to the pragmatic aspects of how models function within scientific inquiry. This approach aligns with a deflationary and pragmatist view of representation, where the success of a model is measured by its ability to produce useful and informative inferences rather than by its adherence to a strict representational accuracy.

Moreover, Suárez's emphasis on the informative nature of inferences brings attention to the epistemic values of models. It highlights how models serve as epistemic tools that aid scientists in exploring and understanding phenomena. This perspective resonates with the broader shift in the philosophy of science towards understanding scientific models as instruments of inquiry rather than mere depictions of reality. Suárez emphasizes the pragmatic dimension of representation, positing that the utility of a representation lies in its ability to function within scientific practice. This pragmatic approach contrasts with traditional views that prioritize resemblance or structural similarity between the model and the target. Instead, Suárez argues that the effectiveness of a representation is determined by how well it facilitates the generation of inferences about the target phenomenon of interest.

Suárez in *IR* introduces the inferential conception as follows: "The inferential conception of representation [inf]: A represents B only if (i) A's representational force is B and (ii) A has specific inferential capacities toward B" [Suárez, (2024), p. 166].

This minimal account is designed to address the limitations of more substantive accounts, such as similarity or isomorphism theories, which impose stronger conditions that often lead to objections (more on this in section IV). By contrast, the inferential conception remains flexible and context-sensitive, allowing it to apply to various representational devices and practices. Suárez emphasizes the minimalist nature of the inferential conception, arguing that it provides only the pragmatic conditions for representation without imposing overly stringent requirements. This makes the framework flexible and applicable across various scientific domains. The contextual nature of inferential capacities means their effectiveness depends on the expertise and knowledge of the agents using the models, thereby grounding representation in the practices of the scientific community.

The inferential conception of representation offers a pragmatic, flexible, and minimalist framework that emphasizes the functional role of models in generating informative inferences. By focusing on representational force and inferential capacity, this conception provides a robust foundation for understanding scientific representation across diverse contexts. However, the research programme in scientific representation and modeling studies carried out by Suarez with his inferential conception of representation is not able to account for neural representations as we will see in the next section.

3. The MFR account

Before proceeding, it is essential to define what constitutes a representation within the MFR framework. This account moves beyond purely logical or pragmatic definitions to offer a set of mechanistic and functional conditions. Drawing on the work of Piccinini (2020), Shea (2018), and others, a neural state functions as an intrinsic structural representation if it meets four key criteria: a) *Structural correspondence*: The vehicle (the neural state) must be isomorphic or similar to its target in a way that preserves relevant structural relations; b) *Causal linkage*: The representation is typically caused by the target it represents, at least during a formative or learning period; c) *Decouplability*: The system must be able to use the representation to guide behavior even in the absence of the target (e.g., in memory or planning); d) *Functional role*: The representation must be used by downstream consumer systems in a way that is specific to its content to guide behavior adaptively. These conditions are not merely philosophical stipulations but form a testable, mechanistic hypothesis about how nervous systems process information to control behavior.

The Mechanistic, Functionalist, and Representationalist (MFR) account offers a robust framework for understanding biological cognition, positing that cognitive capacities are explained by multilevel neurocognitive mechanisms that perform neural computations over neural representations (Piccinini, 2020). To appreciate its challenge to the scope of Suárez's inferentialism, we must first address a foundational concept in the philosophy of mind: the distinction between the personal and subpersonal levels. While Suárez's account clearly operates at the personal level of scientific practice, I argue that this level lacks the explanatory autonomy required to insulate it from a mechanistic analysis. Allow me to be explicit. It is accurate that the terms "neuro" or "neural" do not appear in *IR*, and Suárez does not establish the modeling and representational practices in neuroscience as his intended target. My point is not that Suárez *should* have written a chapter on neurons; it is that the book itself presents the inferential conception as a domain-neutral theory of scientific representation, if not why then present different cases where it is useful (e.g. economics, ecology, astrophysics. . .), so neural codes supply a legitimate stress-test. Suárez insists that the two-factor schema—representational force plus specific inferential capacities—are "*the two most general necessary conditions one can provide for cognitive representation*" (Suárez 2024, p. 160) adding that "*no more general necessary conditions are forthcoming*" (Suárez p. 161). He underscores this breadth again when defining source and target: "*Anything can in principle play the role of the source or the target . . . Our only assumption is that 'X represents Y' is true*"(Suárez 2024, p. 85). Given those explicit aspirations, neural representations—central in contemporary cognitive neuroscience—fall squarely inside the theory's intended scope. My strategy, therefore, is not a category error: I am taking Suárez at his own word and asking whether the two-condition recipe remains explanatorily adequate when confronted with well-characterised neural mapping systems (retinotopy, grid cells, etc.). The evidence suggests it does not, because those systems rely on structural correspondences that the inferential account declares inessential. If Suárez wishes to restrict the theory to *imposed* representations alone (the specific sense in which I employ the term "imposed representations" is clarified below but this issue was discussed in further detail at the end of the Introduction), that is perfectly coherent—but it would amount to relinquishing the general claim quoted above, and it would leave the mechanistic-functionalist-representationalist (MFR) framework as the better guide for intrinsically generated representations.

A potential objection to this strategy is that it conflates two explanatorily autonomous domains, a mistake rooted in ignoring the well-established philosophical distinction between the personal and subpersonal levels of explanation. This article takes a different approach. Rather than accepting this distinction as a given and thereby insulating Suárez's account from this analysis, I will argue that the distinction itself is scientifically and philosophically tenuous. Recent work in the philosophy of cognitive science suggests that a naturalistic approach leads to a "flattened" view of the mind, where personal and subpersonal processes are seen as co-contributing factors in a single, integrated causal architecture (Rupert 2023). Therefore, to fully justify my cross-domain comparison and demonstrate why the MFR account is not merely a theory for a separate domain but a fundamental framework for cognition, it is necessary to first critically assess the distinction that supposedly keeps these domains apart.

As introduced by Dennett (1969), the personal level typically refers to explanations that attribute states like beliefs, desires, and intentions to the person as a whole, often involving consciousness and rationalization. The subpersonal level, in contrast, involves explanations in terms of unconscious, mechanistic processes within the person's cognitive architecture, such as the computational steps in early vision described by Marr (1982). Many philosophers have used this distinction to create an "isolationist" dialectic, where facts about the subpersonal level are deemed irrelevant to questions about the personal level.

However, a consistent naturalism challenges the utility of this distinction. As Rupert (2023) argues, a "flattened view" of the mind, where there is no robust personal level, is more consistent with the practice of contemporary cognitive science. Scientific practice often relies on mixed models where states traditionally considered personal (e.g., explicit attitudes) and subpersonal (e.g., implicit attitudes) are treated as co-contributing causes on a single explanatory plane. For example, Perugini's (2005) structural equation model of behavior shows explicit and implicit attitudes as nodes in a single causal network, with no ontological or explanatory "layering" separating them. In such models, the distinction between levels "plays no role in accounting for the data".

This "flattened" perspective reframes my central argument. The MFR account is not merely a theory for a separate, isolated subpersonal domain. Instead, it describes the fundamental, mechanistic nature of the cognitive architecture that *produces* all representations, including the "imposed" scientific models that are the subject of Suárez's inferentialism. The very capacity of a scientist to use a model for surrogate reasoning depends on an underlying cognitive system whose representational capacities are grounded in mechanisms best explained by the MFR account. Therefore, my critique is not a category error. I argue that the philosophical firewall between the personal and the subpersonal is scientifically untenable. Consequently, an account of scientific representation cannot be fully divorced from the mechanistic principles that govern the cognitive systems doing the representing. The features highlighted by the MFR account, such as structural correspondence and computational transformation, are not confined to a "lower" level but are foundational to the brain's ability to model the world—the very activity that Suárez seeks to characterize.

Central to the MFR account is the idea that neural representations are not merely abstract, inferential or fictional entities, but are grounded in the physical and mechanistic properties of the brain. Piccinini (2020) argues that for something to count as a representation, it must have semantic content and an appropriate functional role, which involves serving as a "stand in" for X to guide behavior with respect to X.

This functional role allows for internal states to guide behavior even when their targets are not immediately present, emphasizing the isomomorphism between internal states and their targets, causal connections from targets to internal states, the possibility of decoupling internal states from their targets, and a role in action control (p. 261). This functional perspective highlights the importance of understanding how neural representations are used by the brain to generate adaptive behaviors. The MFR account recognizes the multi-level nature of neurocognitive mechanisms, where different levels of organization (molecular, cellular, network) contribute to cognitive processes. Coelho Mollo and Vernazzani (2023) argue that this multi-level organization gives rise to a diversity of representational formats, each with its own computational profile determined by the constraints on the transformations that the underlying neural vehicles can undergo. This diversity of formats allows for flexibility and adaptability in neural representation, enabling the brain to process a wide range of information and generate appropriate responses.

A key aspect of the MFR account is the emphasis on computational transformations as the basis for understanding representational content. This view shifts the focus from the static relationship between a representation and its target to the dynamic processes that manipulate and transform representations within neural networks. This view aligns with the broader mechanistic perspective in neuroscience, which seeks to explain cognitive phenomena by identifying the underlying neural mechanisms and processes, neural representations being one of these instances (Johnston and Fusi 2023).

The contrast between the MFR account and Suarez´s inferential conception reveals fundamental differences in how representations are understood within NeuroAI and standard theorizing in philosophy of science. The MFR account, with its emphasis on multilevel mechanisms and computational transformations, provides a detailed framework for explaining how the brain processes information and generates behavior. It highlights the importance of understanding the physical and functional properties of neural representations, offering a comprehensive view of cognitive processes.

On the other hand, Suarez´s inferential conception offers a pragmatic approach to scientific representation, focusing on the utility of representations in generating inferences and advancing knowledge. This approach is particularly valuable in the context of scientific modeling, where the goal is often to develop tools that can predict and explain phenomena, rather than to uncover the exact mechanisms underlying these phenomena. The comparative analysis between the MFR account and Suarez´s inferential conception continues to elucidate key points of divergence in the understanding of representations within NeuroAI. While Suarez´s approach underscores the inferential utility of representations, the MFR account dives deep into the mechanistic underpinnings, offering a more granular view of how representations are instantiated and utilized within neural circuits.

This distinction is critical when considering the practical applications in NeuroAI, where the aim is to model and replicate neural processes. NeuroAI modeling often adopts a mechanistic perspective, striving to emulate the structure and function of biological neural networks. For instance, convolutional neural networks (CNNs) in AI are designed based on hierarchical layers that mimic the visual processing pathways in the primate brain. Yamins and DiCarlo (2016) illustrate that CNNs not only share architectural similarities with the primate visual system, but also exhibit comparable hierarchical and adaptive processing capabilities.

In NeuroAI, it is crucial to distinguish between imposed and intrinsic representational activities. I believe that this distinction of mine is necessary to understand the modeling and representations made by certain cognitive agents, such as human beings engaged in scientific practice, as opposed to the modeling and representation generated by certain natural systems such as the nervous system and in particular the brain. Imposed representational activity includes scientific, and aesthetic representations that are deliberately constructed by humans to serve specific purposes. Intrinsic representational activity refers to natural representations that occur inherently within biological systems, such as neural and biological representations, but also even in artificial systems. These are not externally imposed, but arise from the inherent properties and functions of the neural systems (or even artificial systems) themselves. Neural representations, for instance, involve the brain´s intrinsic ability to encode and process information about the external world. The brain is constantly producing and generating world models that allow the organism to interact efficiently with its environment, otherwise it would not be able to survive.

In other words, entities or biological organisms with intrinsic representational activity, thanks to their nervous systems, have to verifiably represent their environment in order to perform functions and objectives necessary for survival, such as development, growth and interaction with peers. If they did not faithfully represent the disparity of behavioural stimuli of others, objects, space, events... to distinguish potential mates from enemies, friends, etc. an organism with intrinsic representational activity would not survive.

Piccinini (2020) argues that nervous systems perform complex control functions in a computationally tractable way, which necessitates processing structural representations (remember that structural representation includes four elements: a) isomorphism between an internal state and their targets, b) causal connection from at least some targets and their targets, c) the possibility of some internal states to be decoupled from their targets and d) and a role in action control. Although some inferentialist and deflationary authors question this, the dominant mechanistic literature—Piccinini's four-factor schema allied with recent philosophical-neuroscientific defences by Shea 2018, Artiga 2023 and Baker, Lansdell and Kording 2022—still treats isomorphism, causal linkage, decouplability and action-control as the most exact and explanatorily fruitful definition of structural representation). Nervous systems need to integrate information from various sensory modalities, such as shapes, colors, distances, sounds, and chemical signals, to construct internal models that guide the organism´s behavior. These models/representations enable the organism to make fine distinctions between similar stimuli, such as differentiating prey from predators or family members from potential mates, which are critical for survival. Piccinini (2020, p. 264) presents the following argument to introduce how nervous systems needs structural representations:

> The Argument from Complex Control
>
> 1. Nervous systems perform complex control functions in a computationally tractable way.
>
> 2. Performing complex control functions in a computationally tractable way
>
> requires processing structural representations.
>
> ---
>
> Therefore, nervous systems process structural representations.

This argument underlines the necessity for neural systems to develop internal models that can process and represent complex environmental information accurately and efficiently. Organisms with nervous systems that cannot perform these functions in a computationally tractable way would not survive, as they would be unable to respond appropriately to their environment.

But there is another type of argument I would like to apply to defend the MFR account, this one of my own making. Neuroscience is both a basic and a translational science, meaning that its findings serve to increase our curiosity and knowledge about how the brain works but at the same time it seeks to design more precise models to quantitatively test mechanistic hypotheses of the brain and obtain experimentally testable predictions with the ultimate goal of refining our understanding of neural systems in health and disease. In other words, neuroscience has a clinical dimension. If the models it tries to build are not empirically validated in a verifiable way, people suffering from neuropsychiatric disorders, people suffering from strokes, people suffering from head injuries, people suffering from migraines, and so on and so forth… they would have no hope.

In order to effectively understand and explain the nervous system in both healthy and diseased states, any explanatory framework must adhere to the *principle of verisimilitude*. This principle demands that models and theories closely approximate reality, ensuring that representations accurately reflect the complex nature of neural processes. The MFR (Mechanistic, Functionalist, and Representationalist) account satisfies this requirement by emphasizing the intrinsic representational activity within nervous systems, providing a detailed and realistic depiction of how neural mechanisms function.

The MFR account not only accounts for the structural and functional aspects of neural representations but also captures their representational nature. By incorporating verisimilitude, it ensures that models of neural processes are not only theoretically sound but also practically applicable. This approach allows for a more nuanced understanding of both normal and pathological states of the nervous system, making it a powerful framework for advancing research in neuroscience and NeuroAI. The MFR account´s commitment to verisimilitude enhances its capacity to offer insights that are both empirically robust and theoretically coherent.

While the debate between inferentialist and mechanistic accounts of representation can seem abstract, it has profound methodological implications when grounded in the practical goals of a specific scientific field. Neuroscience is unique in its dual role as both a basic science aimed at understanding and a translational science aimed at clinical intervention. This dual role imposes a powerful constraint on its models that I term the clinical imperative: for a model of a neural system to be useful in diagnosing, treating, or curing a pathology, it must accurately represent the underlying causal mechanism responsible for that pathology. Consider the goal of developing treatments for disorders rooted in representational dysfunction, such as spatial neglect following a parietal lobe stroke or disorganized thought in schizophrenia. A purely inferential model, which treats the underlying mechanism as a "black box" and is valued only for its predictive output, offers no clear path for intervention. A clinician cannot prescribe a drug to target an "inferential capacity" or perform surgery on a "representational force". Instead, effective intervention requires a model that correctly identifies the relevant mechanistic components—such as a specific neural circuit, a population of neurons, a receptor type, or a neurotransmitter pathway—and how their interactions produce the cognitive function in question.

This practical requirement for mechanistic accuracy strongly favors the MFR account for explaining intrinsic neural systems. The MFR framework, by its very nature, is committed to achieving high mechanistic verisimilitude—a term I use here to denote the accurate representation of the causal structure of the target mechanism. Its primary goal is to "open the black box" and detail the real causal structure that produces a phenomenon.

The inferential conception, however, is agnostic about underlying mechanisms. It values models for their predictive utility, allowing for "felicitous falsehoods" and instrumentally useful idealizations that may bear no resemblance to the actual causal story. While this is a powerful approach for imposed models in many sciences, it falls short of the demands of a science that must physically interact with its object of study to restore function. A pharmacologist designing a drug needs a model that accurately reflects the targeted molecular pathway, not an "as-if" story that happens to yield good predictions.

This is why the MFR account fulfills what I call the *principle of verisimilitude*:

The argument from the principle of verisimilitude

1. To effectively explain and intervene upon the nervous system in both healthy and diseased states, any explanatory framework must produce models with high mechanistic verisimilitude—that is, models that accurately capture the causal structure of the target neural mechanisms.

2. The MFR account is inherently oriented towards achieving mechanistic verisimilitude by identifying and modeling the real components and interactions within a system. The inferential conception, in contrast, is indifferent to mechanistic accuracy, prioritizing only inferential power.

Therefore, the MFR account provides a framework that satisfies the core explanatory and practical demands of neuroscience, making it a more robust and adequate framework for understanding intrinsic neural representations than the inferential conception

The argument from the principle of versesimilitude underlines a trivial and straightforward idea. If neuroscientists did not investigate neural representations as they emerge naturally, there would be no possibility to clinically treat patients suffering from neuropsychiatric disorders and to satisfy scientific curiosity about the functioning of the brain. But not all authors are happy with the view of representation offered by the MRF account. For example, Carrillo and Knuuttila (2023) present a significant critique of the MFR account, focusing on its application to abstract models in neuroscience, such as the Hodgkin-Huxley model (Hodgkin and Huxley 1939, Hodgkin, Huxley and Katz 1951). They argue that the MFR account struggles with abstract models that lack detailed mechanistic descriptions. The Hodgkin-Huxley model, for instance, does not provide a granular account of ion transport mechanisms but rather uses mathematical abstractions to describe the action potentials of neurons. Therefore, the mechanistic account struggles with abstract models like Hodgkin-Huxley, which lack detailed descriptions of ion transport mechanisms.

The Hodgkin-Huxley model is a mathematical model that describes how action potentials in neurons are initiated and propagated. It was developed in the 1950s by Alan Hodgkin and Andrew Huxley, who conducted experiments on the squid giant axon. The model focuses on the changes in conductance of ion channels (sodium and potassium) in the neuron´s membrane during an action potential. The main components of the model are: a) membrane potential: This is the electrical potential difference across the neuron´s membrane. It changes as ions move in and out of the cell through ion channels. b) ion channels: These are proteins in the membrane that allow specific ions to pass through. The model focuses on sodium (Na+) and potassium (K+) channels, c) conductance: This refers to the ease with which ions can flow through a channel. In the model, conductance is represented by variables that change over time and finally d) gating variables: These variables control the opening and closing of ion channels. They are influenced by the membrane potential and time.

The problem is that Carrillo and Knuuttila (2023) believe that one of the canonical mechanistic models, such as Hodgkin-Huxley model, is not a clear example of MFR account and that because it is an abstract model it does not mechanistically explain the nerve impulse of nerve cells. I disagree. The Hodgkin-Huxley model, a seminal example of mechanistic model in neuroscience, provides a mechanistic explanation of the action potential by describing changes in a neural membrane´s voltage conductivity. Initially, the model omitted lower-level mechanistic details about how changes in membrane permeability arise, both because these details were unknown and to afford the model greater generality. While some have described the model as non-explanatory or non-mechanistic, it is more accurately characterized as a mechanism sketch that evolved into a mechanism schema. This means it explains the phenomenon of the action potential at one mechanistic level (membrane conductivity changes) while abstracting away from lower mechanistic levels (specific ion channel activities). Thus, the Hodgkin-Huxley model exemplifies a mechanistic approach by providing a detailed, albeit abstract, account of the processes underlying action potentials.

According to Aizawa and Headley (2022), abduction—or inference to the best explanation—is a legitimate strategy for justifying compositional claims when direct observation is limited, as was true for Hodgkin and Huxley, who in the early1950s lacked direct molecular knowledge of ion channels and their subcomponents (Hodgkin and Huxley1952). The Hodgkin-Huxley model, despite its abstract nature, successfully uses abduction to infer the roles of ion fluxes in generating action potentials. This approach demonstrates that even abstract models can provide mechanistic insights by focusing on the most relevant causal factors, supporting the validity of the MFR account. Aizawa and Headley (2022) emphasize that the Hodgkin-Huxley model demonstrates the successful use of abduction in mechanistic explanations, highlighting that abstraction focuses on the most relevant causal factors at different levels of organization.

Neurocomputationalism, a paradigm prevalent in the cognitive sciences, posits that the human brain processes information through computation. This concept is embraced by NeuroAI, which acknowledges that the human brain generates representations with informational content through computational processes. The development of high-performing deep convolutional neural networks (DCNNs) has been a breakthrough in this field. These models have demonstrated impressive accuracy in object recognition tasks and have been shown to align well with neural responses in the primate ventral visual stream. This alignment suggests that these models capture some of the essential mechanisms underlying object recognition in the brain (Kar and DiCarlo 2023). We could say that what they capture is how *intrinsic* representations are generated, whether in natural systems (brains) or in artificial systems (artificial neural networks).

4. Demarcating the domains: Why Suárez's arguments highlight the need for the MFR account

In response to Mauricio Suárez´s inferential conception of representation presented in *IR*—which emphasizes the pragmatic and inferential roles of models in scientific practice while rejecting representations as a substantive relation based on similarity or isomorphism—the MFR account advances a comprehensive alternative framework. Let me explain. In *IR* Suárez presents the inferential conception as a universal account of scientific representation, intended to supersede similarity and isomorphism-based views across the board . The book nowhere restricts its remit to imposed (i.e. artefactual, researcher-constructed) models; on the contrary, Suárez invites application to any representational practice. Once that invitation is accepted, neural representation becomes a crucial test-bed. Neurophysiology shows that cognition depends on vehicles whose structure preserves key topological relations to what they represent—

retinotopic, tonotopic, place- and grid-cell codes are canonical cases. The MFR framework systematises this evidence: structural correspondence is *often obligatory* for a representation to play its causal-computational role in the organism (Piccinini 2020). Hence an account that declares similarity/isomorphism explanatorily idle is empirically inadequate for the very paradigm of intrinsic representation. Since *IR* theory makes broad claims, the neural examples discussed in MFR present challenges to the theory´s general application. Alternatively, if we consider a division of theoretical labor—where intrinsic representations follow MFR-type constraints while imposed representations align with inferential-pragmatic norms—the apparent conflict diminishes, though this approach necessarily narrows the inferential conception´s domain. In either case, MFR remains relevant to Suárez´s proposal by highlighting important boundary conditions that any comprehensive account should acknowledge This section will outline the key points of this reponse. Suárez presents five key arguments against the substantive theories of representation, specifically targeting similarity ([sim]) and isomorphism ([iso]) as inadequate for explaining scientific representation. Here is a concise summary of these arguments based on his detailed critique (I refer the reader to *IR* to see these arguments):

1. The Argument from Variety

Similarity and isomorphism do not apply to the full range of scientific representations. Suárez observes that scientific practice involves a wide variety of representational devices, each with different means of representation (e.g. scale, analogue, mathematical, etc.). This empirical fact shows that relying solely on [sim] or [iso] is too restrictive and cannot account for the diversity of representational tools used in science. For example, while physical models like toy bridges might rely on similarity, abstract mathematical models often do not fit neatly into similarity or isomorphism frameworks.

> "The argument from variety: [sim], [iso] do not apply to all representational devices" [Suárez, (2024), p. 104].

2. The Logical Argument

Similarity and isomorphism lack the logical properties necessary for representation. Representation, as a concept, is non-reflexive, non-symmetrical, and non-transitive. However, similarity is reflexive and symmetrical, and isomorphism is reflexive, symmetrical, and transitive. This fundamental mismatch means that neither [sim] nor [iso] can adequately capture the logical nature of representation. The logical argument states that representation is inherently directional (non-symmetrical), whereas similarity and isomorphism are logically symmetric and reflexive. For imposed models, Suárez correctly concludes that an additional condition, which he calls "representational force" is required to supply this directionality.

> "The logical argument: [sim], [iso] do not possess the logical properties of representation" [Suárez, (2024), p. 110].

3. The Argument from Misrepresentation

Similarity and isomorphism cannot account for misrepresentation. Scientific models often misrepresent their targets through idealization, abstraction, or simplification. [Sim] and [iso] theories struggle to explain these inaccuracies because they imply a one-to-one correspondence between model and target. For instance, Newtonian mechanics provides an approximately correct representation of the solar system but is not isomorphic to the actual motions of celestial bodies when relativistic effects are considered.

> "The argument from misrepresentation: [sim], [iso] do not make room for the ubiquitous phenomena of mistargeting and/or inaccuracy" [Suárez, (2024), p. 113].

4. The Nonnecessity Argument

Similarity and isomorphism are not necessary for representation. Suárez argues that representation can occur even when similarity or isomorphism is absent. For example, equations and mathematical models represent physical phenomena despite lacking any physical resemblance to their targets. The relevant properties that define representation do not always include similarity or isomorphism.

> "The nonnecessity argument: [sim], [iso] are not necessary for representation. Representation can obtain even if [sim], [iso] fail" [Suárez, (2024), p. 115].

5. The Nonsufficiency Argument

Similarity and isomorphism are not sufficient for representation. Even if similarity or isomorphism exists, it does not guarantee representation. Representation requires a directional and intentional relationship where the source leads to an understanding of the target. This directionality is not captured by [sim] or [iso], which merely describe structural or relational similarities without explaining how they function as representations.

> "The nonsufficiency argument: [sim], [iso] are not sufficient for representation. Representation may fail to obtain even if [sim], [iso] hold" [Suárez, (2024), p. 117].

However, despite the consistency of these arguments to criticise what Suarez considers to be a reductive naturalism, or in other words to find a sufficient and necessary representational relationship between any source and the target, these arguments do not work for a special class of representations: neural representations. Neural representations are autogenic, i.e. they are not created by *scientists*. They are the result of evolutionary history and selection processes that have resulted in neural systems that faithfully represent the surrounding environment, and those organisms that are equipped with complicated and complex nervous systems can generate a model of the world in order to adapt and survive.

In order to understand neural representations, the scientist has to look for models as close as possible to how these neural representations actually represent. In fact, the MFR account is the most valid strategy for explaining and understanding neural representations. The MFR account counters *The Argument from Variety* by acknowledging that representational diversity is expected due to the multi-level nature of neural mechanisms. For instance, the visual processing pathways in the brain, such as the differences between V1 and IT in vision, exhibit diverse computational profiles that explain varied representational formats (Coelho Mollo and Vernazzani 2023). Representation is a key concept in neuroscience and artificial intelligence (NeuroAI). Representations are internal states or processes within a system (biological or artificial) that are *about* something else. The *content* of a representation is what it is about. The physical state or process that carries the content is the *vehicle*. In addition to content and vehicle, representations can also be characterized by their *format*.

Coelho Mollo and Vernazzani (2023) propose a computational view of representational formats, arguing that formats are individuated by the computational profiles of vehicles. The computational profile refers to the set of constraints that determine the computational transformations a vehicle can undergo. This view emphasizes that the standard of success for a theory of representational formats is not how well the formats fit our pre-theoretic expectations, but rather the epistemic value it has in guiding research. Thus, for intrinsic representations, variety is not an argument against a foundational role for structural correspondence; rather, it is evidence of how different forms of structural correspondence are mechanistically optimized for different cognitive functions.

The MFR account resolves this problem for intrinsic representations by grounding directionality not in an abstract convention but in the physical, causal architecture of the neural system. Information processing in the brain is an inherently directional causal chain, flowing from sensory transducers through successive computational stages to motor effectors. The isomorphism in a retinotopic map, for example, is not a standalone logical relation to which directionality must be added; it is a structural property that is created and maintained *by* this directional causal process. The arrow of representation is an empirical fact of the mechanism's causal flow, not a separable pragmatic posit. Relying on public representations like maps as analogies for this process can indeed be misleading, precisely because it obscures this fusion of structure and causality inherent in neural systems. The MFR account responds to *The Logical Argument* by arguing that the use of public representations, such as maps and pictures, as analogies for internal representations (neural representations) can be misleading. Relying on public representations as analogies can oversimplify the complexity of internal representations. Retinotopic (Benson and Winawer 2018) and tonotopic (Langers and van Dijk 2011) maps are orderly mappings of sensory information in the brain. In a retinotopic map, the spatial arrangement of neurons in the visual cortex corresponds to the spatial arrangement of light on the retina. Similarly, in a tonotopic map, the arrangement of neurons in the auditory cortex corresponds to the frequency of sound waves.

These representations could be seen as mirroring reality in the sense that they preserve spatial or frequency relationships present in the external stimuli. And the same can be said about face cells (Kanwisher et al. 1997) and place cells (Moser, Kropff and Moser 2008), head direction cells (Taube et al. 1990) and concept cells (Quiroga, (2005). These cells and retinotopic or tonotopic maps can be seen as examples of neural representations that exhibit a degree of isomorphism with the external world, enabling efficient processing and invariant representation of sensory information. Lots of experimental evidence within NeuroAI show that neural representations, such as retinotopic and tonotopic maps, often exhibit invariant, similar, and isomorphic characteristics relative to the objects they represent (Johnston and Fusi 2023; Courellis, Minxha, Cardenas, *et al.* 2024).

Numerous studies in NeuroAI (Acosta et al. 2023) demonstrate that neural representations, such as retinotopic and tonotopic maps, consistently exhibit invariant, similar, and isomorphic properties relative to the objects they represent. These representations maintain a structured correspondence with sensory inputs, preserving spatial and frequency relationships, and enabling accurate and efficient information processing. This evidence underscores the significance of structural mappings in neural systems, highlighting their role in ensuring that the brain´s internal representations mirror external stimuli´s spatial and frequency characteristics

In relation to *The Argument from Misrepresentation,* Suárez correctly argues that imposed scientific models are often useful *because* they are idealizations or "felicitous falsehoods" that misrepresent their targets. A strict isomorphism or similarity account struggles to explain how an inaccurate model can still be a representation. For an organism's intrinsic representations, however, significant misrepresentation is not a useful idealization but a catastrophic failure. The brain's models must be veridical enough to guide survival-critical actions . An organism that misrepresents the location of a predator or a food source will not survive long. In the MFR framework, misrepresentation is treated as a computational error or a sign of pathology, which the brain actively works to minimize through error-correction mechanisms like predictive coding. While a scientist can choose to use an inaccurate model for pragmatic reasons, the biological constraints on the nervous system demand a high degree of fidelity. Thus, the argument from misrepresentation, which is powerful for imposed models, highlights a key constraint—the need for veridicality—that governs intrinsic neural representations.

The MFR account would likely argue that Suárez´s *The Nonnecessity Argument* while valid in some contexts, does not undermine the importance of similarity and isomorphism in understanding representational formats within nervous systems which are a instance of intrinsic representational activity. In this view, similarity and isomorphism, while not strictly necessary for representation in a general sense, play a crucial role in determining the computational profile of vehicles and, consequently, the format of representation in neural representations. For example, the spatial arrangement of place cells in the hippocampus, while not a perfect map, exhibits a degree of isomorphism with the spatial environment that facilitates efficient navigation and spatial reasoning (Moser, Kropff and Moser 2008; Courellis, Minxha, Cardenas, *et al.* 2024).

Suárez's argument that similarity and isomorphism are not sufficient for representation rests on the observation that representations can fail even when these relations hold. However, the representations he refers to are part of the category of *imposed* representational activity—that is, representations intentionally created by scientists—and not the special kind of neural representations that are part of the category of *intrinsic* or *autogenic* representational activity, which are not intentionally created but generated by natural systems such as nervous systems. The necessity of isomorphism is not derived from a definition of representation, but from the functional demands on the biological mechanism itself. It is a contingent, empirical necessity for the system to perform its job. Finally, Suárez argues that similarity or isomorphism is not sufficient for representation, as an accidental correspondence does not establish a representational link; directionality and intended use are missing. The MFR account fully agrees that a standalone isomorphism is insufficient. However, within a neural mechanism, the isomorphism is *never* standalone. It is one integral component of a functional system that also fulfills the other necessary conditions for representation: a causal link to the target, the ability to be decoupled from the target for use in memory and planning, and a determinate functional role in controlling action . The sufficiency is provided by the entire integrated mechanism. The structural correspondence is not accidental; it is a feature that the mechanism has evolved to create and maintain precisely because it is essential for the system to fulfill its adaptive function. The sufficiency is therefore mechanistic and causal, not merely logical or conventional The MFR account would respond to *The Nonsufficiency Argument* by saying that similarity and isomorphism are essential for creating mappings that ensure consistency and reliability in representation. The spatial arrangement of neurons in the visual cortex corresponds to the spatial arrangement of light on the retina.

This structural mapping is crucial for maintaining the integrity of visual information as it is processed by the brain. If we assume that, *ceteris paribus*, the computational process of representation functions correctly and faithfully represents the surrounding environment, then the organism can carry out the functions that contribute to its growth, development, and interaction with the environment. In short, it can survive. To put it simply, without isomorphic representation, an organism would not understand its reality. A simple example may help illustrate this point. If an organism with a visual sensory modality that creates representations of stimuli is presented with a rapidly approaching stimulus, but represents this stimulus using an inferentialist account as suggested by Suárez, it could apply a framework of inferences that, given the organism´s learned experiences, would not lead it to represent the stimulus as a predator. If the approaching stimulus is indeed a predator, the organism would not have time to react and escape before being eaten.

Suárez's inferential conception is explicitly a theory of imposed representations—the models scientists build. My analysis, however, evaluates it against the standards of intrinsic representations—the neural processes that constitute cognition. A reviewer has rightly questioned whether this constitutes a "displacement among levels of analysis", given that Suárez's theory was never intended to be a theory of neural mechanisms. I argue that this comparison is not a category error but a necessary philosophical stress-test. All public on-fundamental forms of representation are derivative. Their capacity to represent is inherited from a more fundamental class of representations, which are the mental states of their users. The scientific models Suárez analyzes are therefore derivative representations, enabled by the fundamental, intrinsic representations occurring in the brains of scientists. From this perspective, a theory of derivative representation cannot be philosophically self-sufficient if its core principles contradict the nature of the fundamental representations that ground it. Suárez's theory, by bracketing off the cognitive and neural origins of representation as an "independently interesting issue", isolates itself from its own foundations. This project is to bridge that gap. I challenge the inferential conception not by misapplying it, but by evaluating its coherence in light of the mechanistic realities of the fundamental representational system—the brain.

5. Concluding remarks

In this article, I have critically examined Mauricio Suárez's inferentialist account of scientific representation in light of recent advancements in NeuroAI. While Suárez provides valuable insights into the pragmatic use of representations in scientific practice, his account cannot be extended to capture the dynamic, computational and biological nature of neural representations. Just as Jorge Luis Borges´s story "On Exactitude in Science" illustrates the futility of creating a one-to-one map that becomes indistinguishable from reality itself, the MFR account recognizes that representations are inherently selective, but accurate, abstractions. Unlike the empire in Borges´s tale that collapses under the weight of its perfectly accurate yet impractical map, neural representations encode essential information efficiently without mirroring reality in exhaustive detail, however, the reality they represent is faithfully reflected. Through a detailed analysis, I have argued that neural representations are not merely abstract entities used for inference but are deeply rooted in the physical and functional properties of nervous systems. The MFR account offers a more comprehensive framework for explaining and understanding neural representations. This account emphasizes the importance of computational transformations and functional roles of neural processing. Empirical evidence from neuroscience, such as the Hodgkin-Huxley model, supports the view that neural representations are embodied in the mechanistic and functional architecture of the brain. Furthermore, I have addressed Suárez's arguments against the necessity and sufficiency of similarity and isomorphism for representation. While Suárez argues that these structural relations are neither necessary nor sufficient, I have demonstrated that in the context of neural representations, similarity and isomorphism play a crucial role in ensuring accurate and efficient information processing. The MFR account acknowledges the diversity of representational formats and the adaptive nature of neural representations, providing a robust response to Suárez's critiques. To sum up, a mechanistic and computational understanding of neural representation, as advocated by the MFR account, offers a more empirically grounded and comprehensive framework for modeling and representation in the sciences of the mind than Suárez's inferentialist account. This approach not only aligns with recent empirical findings, but also provides deeper insights into the nature of cognitive processes and the functioning of nervous systems.

Acknowledgements

I am grateful to Professor Mauricio Suárez for inviting me to his book symposium at the XI Conference of the Spanish Society of Logic, Methodology and Philosophy of Science (University of Oviedo, Spain). I am likewise indebted to the panellists for their constructive feedback. I also thank the audience for their attendance. Above all, I wish to acknowledge Gui Sanches de Oliveira, whose sustained and meticulous coordination of the symposium, and of the special issue project that did not come to fruition, critically shaped and supported the present work. As usual, any errors are solely attributable to the author.

References

Acosta, F., Conwell, C., Sanborn, S., Klindt, D., & Miolane, N. (2023). Relating representational geometry to cortical geometry in the visual cortex. NeurIPS 2023 Workshop on Unifying Representations in Neural Models.

Artiga, M. (2023). Understanding structural representations. *The British Journal of Philosophy of Science*. https://doi.org/10.1086/728714

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*, 26, 942–958.

https://doi.org/10.1016/j.tics.2022.08.014

Bakermans, J. J., Warren, J., Whittington, J. C., & Behrens, T. E. (2025). Constructing future behavior in the hippocampal formation through composition and replay. *Nature Neuroscience*, 1-12. https://doi.org/10.1038/s41593-025-01908-3

Benson, N., & Winawer, J. (2018). Bayesian analysis of retinotopic maps. eLife, 7. https://doi.org/10.7554/eLife.40224

Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. T*HEORIA. An International Journal for Theory, History and Foundations of Science*, *21*(1), 67–85. https://doi.org/10.1387/theoria.554

Carrillo, N., & Knuuttila, T. (2023). Mechanisms and the problem of abstract models. European *Journal for Philosophy of Science*, 13(3), 1-19. https://doi.org/10.1007/s13194-023-00515-y

Coelho Mollo, D., & Vernazzani, A. (2023). The formats of cognitive representation: A computational account. Philosophy of Science, 1-20. https://doi.org/10.1017/psa.2023.123

Courellis, H. S., Minxha, J., Cardenas, A. R., Smith, L. M., Holliday, A. M., Johnson, E. L., Wright, M. J., Aum, D. J., Braud, J., Salma, A., Pauli, W. M., Mamelak, A. N., & Rutishauser, U. (2024). Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 1-19. https://doi.org/10.1038/s41586-024-07799-x

Dennett, D. C. 1(969). *Content and Consciousness*. Abingdon, UK: Routledge

Frigg, R., & Nguyen, J. (2017). Models and representation. In L. Magnani & T. Bertolotti (Eds.), *Springer Handbook of Model-Based Science* (pp. 49-102). Springer.

Hodgkin, A. L., & Huxley, A. F. (1939). Action potentials recorded from inside a nerve fibre. *Nature*, 144(3651), 710–711. https://doi.org/10.1038/144710a0

Hodgkin, A. L., Huxley, A. F., & Katz, B. (1952). Measurement of current voltage relations in the membrane of the giant axon of Loligo. *The Journal of Physiology*, 116(4), 424–448. https://doi.org/10.1113/jphysiol.1952.sp004716

Johnston, W. J., & Fusi, S. (2023). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, 14, 1040. https://doi.org/10.1038/s41467-023-36583-0

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. Journal of Neuroscience, 17(11), 4302– 4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kar, K., & DiCarlo, J. J. (2023). The quest for an integrated set of neural mechanisms underlying object recognition in primates. *arXiv.* https://doi.org/10.48550/arXiv.2312.05956

Langers, D. R., & van Dijk, P. (2011). Mapping the tonotopic organization in human auditory cortex with minimally salient acoustic stimulation. *Cerebral Cortex*, 22(9), 2024-2038. https://doi.org/10.1093/cercor/bhr282

Lewis, J. E., & Kristan, W. B., Jr. (1998). Representation of touch location by a population of leech sensory neurons. *Journal of Neurophysiology*, 80(5), 2584–2592. https://doi.org/10.1152/jn.1998.80.5.2584

Liu, X., Ramirez, S., Pang, P. T., Puryear, C. B., Govindarajan, A., Deisseroth, K., & Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. Nature, 484(7394), 381-385. https://doi.org/10.1038/nature11028

Marr, D. 1(982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman and Company.

Matsumoto, M., & Komatsu, H. (2005). Neural responses in the macaque V1 to bar stimuli with various lengths presented on the blind spot. *Journal of Neurophysiology*, 93(5), 2374–2387. https://doi.org/10.1152/jn.00811.2004

Moser, E., Kropff, E., & Moser, M. (2008). Place cells, grid cells, and the brain's spatial representation system. Annual Review of Neuroscience, 31, 69-89. https://doi.org/10.1146/annurev.neuro.31.061307.090723

Perugini, M. (2005). Predictive Models of Implicit and Explicit Attitudes. *British Journal of Social Psychology* 44: 29–45

Piccinini, G. (2020). *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford University Press.

Poldrack, R. A. (2021). The physics of representation. *Synthese*, 199, 1307–1325. https://doi.org/10.1007/s11229-020-02793-y

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045), 1102-1107. https://doi.org/10.1038/nature03687

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press.

Rupert, R. D. (2023). Naturalism Meets the Personal Level: How Mixed Modelling Flattens the Mind. https://philpapers.org/rec/RUPNMT

Sanborn, S., Shewmake, C., Olshausen, B., & Hillar, C. (2023). Bispectral neural networks. ICLR 2023.

Shea, N. (2013). Naturalising representational content. *Philosophy Compass*, 8(5), 496–509. https://doi.org/10.1111/phc3.12033

Shea, N. (2018). *Representation In Cognitive Science*. Oxford University Press.

Shi, Y., Bi, D., Hesse, J. K., Lanfranchi, F. F., Chen, S., & Tsao, D. Y. (2023). Rapid, concerted switching of the neural code in inferotemporal cortex. *bioRxiv*. https://doi.org/10.1101/2023.12.06.570341

Spillmann, L. (2014). Receptive fields of visual neurons: The early years. *Perception*, 43(11), 1145– 1176. https://doi.org/10.1068/p7721

Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science*, 17(3), 225-244. https://doi.org/10.1080/0269859032000169442

Suárez, M. (2024). *Inference and Representation: A study In Modeling Science*. University of Chicago Press.

Taube, J. S., Muller, R. U., & Ranck, J. B., Jr. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. J*ournal of Neuroscience*, 10(2), 420-435. https://doi.org/10.1523/JNEUROSCI.10-02-00420.1990

Tootell, R. B., Switkes, E., Silverman, M. S., & Hamilton, S. L. (1988). Functional anatomy of macaque striate cortex. II. Retinotopic organization. *Journal of Neuroscience*, 8(5), 1531– 1568. https://doi.org/10.1523/JNEUROSCI.08-05-01531.1988

Vollan, A. Z., Gardner, R. J., & Moser, E. I. (2025). Left–right-alternating theta sweeps in entorhinal–hippocampal maps of space. *Nature, 639*, 995–1005 https://doi.org/10.1038/s41586-024-08527-1

Yamins, D., & DiCarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19, 356–365. https://doi.org/10.1038/nn.4244