Theoria, 2025, 40(3), 297-319
https://doi.org/10.1387/theoria.25171

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Meaning from matching: How representational mechanisms exploit structural similarity

(El significado a partir de la correspondencia: cómo los mecanismos representacionales explotan la similitud estructural)

Wojciech Mamak*
Polish Academy of Sciences
https://orcid.org/0009-0004-3657-1128

Marcin Miłkowski**
Polish Academy of Sciences
https://orcid.org/0000-0001-7646-5742

Keywords

Correspondence Network Framework
Representational mechanisms
Structural similarity
Infocorrespondence
Semantic information
Neural representation
Multiple classification schemes
Representational similarity analysis
Taxonomic organization
Content-sensitive processing

ABSTRACT: Representational mechanisms are responsible for processing information to modify readiness for action. While structural similarity has been proposed as foundational to neural representation, how these mechanisms systematically harness correspondence-based information remains undertheorized. This paper introduces a Correspondence Network Framework that elucidates how representational systems exploit multiple channels of structural similarity concurrently. We demonstrate that structural similarity engenders networks of infocorrespondences between informational structures, allowing representational mechanisms to evaluate information across channels, detect inconsistencies, and establish satisfaction conditions for representational content. Our framework extends beyond single-channel accounts by explaining how multiple classification systems can operate simultaneously over the same neural vehicles, enabling content-sensitive processing. To illustrate the framework's utility, we analyze representational similarity analysis (RSA) in cognitive neuroscience as detecting correspondence networks. This analysis reveals why taxonomic organization consistently emerges in neural representational spaces and provides principled responses to anti-representationalist critiques. While RSA alone cannot fully characterize representational mechanisms, the Correspondence Network Framework delineates its interpretive constraints and provides valuable heuristics for further empirical inquiry.

* ***Correspondence to:*** Wojciech Mamak. Graduate School for Social Research, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland – wojciech.mamak@gmail.com – https://orcid.org/0009-0004-3657-1128

** ***Correspondence to:*** Marcin Miłkowski. Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland – marcin.milkowski@ifispan.edu.pl – https://orcid.org/0000-0001-7646-5742

How to cite: Mamak, Wojciech; Miłkowski, Marcin (2025). «Meaning from matching: How representational mechanisms exploit structural similarity»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 40(3), 297-319. (https://doi.org/10.1387/theoria.25171).

Received: 09 October, 2023; Final version: 03 May, 2025.
ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press

Palabras clave

Marco de Redes de Correspondencia
Mecanismos representacionales
Similitud estructural
Infocorrespondencia
Información semántica
Representación neural
Esquemas de clasificación múltiple
Análisis de similitud representacional
Organización taxonómica
Procesamiento sensible al contenido

RESUMEN: Los mecanismos representacionales son responsables de procesar información para modificar la disposición para la acción. Aunque se ha propuesto que la similitud estructural es fundamental para la representación neural, sigue sin teorizarse adecuadamente cómo estos mecanismos aprovechan sistemáticamente la información basada en correspondencias. Este artículo introduce un Marco de Redes de Correspondencia que aclara cómo los sistemas representacionales explotan múltiples canales de similitud estructural de manera concurrente. Demostramos que la similitud estructural genera redes de infocorrespondencias entre estructuras informacionales, permitiendo a los mecanismos representacionales evaluar información a través de canales, detectar inconsistencias y establecer condiciones de satisfacción para el contenido representacional. Nuestro marco va más allá de los enfoques de canal único, al explicar cómo varios sistemas de clasificación pueden operar simultáneamente sobre los mismos vehículos neurales, permitiendo un procesamiento sensible al contenido. Para ilustrar la utilidad del marco, analizamos el análisis de similitud representacional (RSA) en la neurociencia cognitiva como una detección de redes de correspondencia. Este análisis revela por qué la organización taxonómica emerge consistentemente en los espacios representacionales neurales y proporciona respuestas fundamentadas a las críticas antirrepresentacionalistas. Si bien el RSA por sí solo no puede caracterizar completamente los mecanismos representacionales, el Marco de Redes de Correspondencia delimita sus restricciones interpretativas y proporciona heurísticas valiosas para futuras investigaciones empíricas.

1. Introduction

Correspondence-based semantic information, grounded in structural similarity between classifications, offers a promising framework for understanding representation in cognitive science (Miłkowski, 2023). When classifications exhibit complete similarity, semantic information can be evaluated as true —a notion that aligns with how neural representations are understood in contemporary neuroscience, particularly in approaches like Representational Similarity Analysis (Kriegeskorte & Kievit, 2013). However, despite these conceptual alignments, it remains unclear how representational mechanisms —computational systems sensitive to semantic content— systematically exploit this form of information.

The purpose of this paper is to address this issue by introducing a Correspondence Network Framework that synthesizes notions of representational mechanism and correspondence-based semantic information and elucidates how mechanisms systematically harness this form of information. While previous accounts, particularly Shea's (2007, 2018) work on structural representation, have examined how type-token mappings enable representation, our framework uniquely focuses on how networks of multiple correspondence relations operate concurrently, allowing cognitive systems to evaluate information across channels. This approach explains how neural systems simultaneously maintain different classification schemes over the same vehicles, enabling content-sensitive processing through cross-channel evaluation.

The gist of the approach is as follows: correspondence-based information is grounded in networks of infocorrespondences between informational structures. Representational mechanisms process this information across multiple channels to modify readiness for action. This enables them to refer to targets, identify their characteristics, and evaluate information by detecting cross-channel consistency or inconsistency. This network approach to correspondence explains why neural representational spaces consistently organize along semantic dimensions —a phenomenon widely observed in cognitive neuroscience but inadequately explained by traditional accounts.

Our framework directly addresses recent anti-representationalist critiques (Facchin, 2023; Kohár, 2023) by showing how representational content becomes causally efficacious through network-level operations rather than isolated similarity relations. The Correspondence Network Framework offers a principled explanation for why cognitive systems

exhibit sensitivity to semantic properties in ways that cannot be reduced to purely structural features of individual vehicles.

We will first elaborate on the key concepts of representational mechanisms and correspondence-based semantic information. We will then synthesize them to show how mechanisms can exploit correspondence information to refer, identify characteristics, and evaluate, by demonstrating in detail how RSA relies on assessing correspondence-based information using multivariate methods. This application will reveal the benefits of the systematic approach for understanding representation in cognitive science. In the penultimate section, we also briefly discuss possible objections. In the final section, we briefly summarize how the paper contributes to the clarification of how neural representational mechanisms leverage semantic information grounded ultimately in structural similarity.

2. *Representational mechanisms*

Representational mechanisms are responsible for processing information to appropriately modify an agent's readiness for action (Bielecka & Miłkowski, 2020; Miłkowski, 2013, 2015; for a similar approach, see Plebe & De La Cruz, 2016; Gładziejewski, 2015b). This broad notion encompasses mechanisms in the sense of the neo-mechanistic approach to explanation (Machamer *et al.*, 2000). These mechanisms, understood as organized collections of component parts and operations, are posited to explain various representational phenomena. In essence, representational mechanisms are identified functionally and can be instantiated by neural systems or any other information-processing architectures that play a representational role, including, at least in principle, artificial systems.

As we use this term, it refers to a particular account of a specific kind of computational mechanisms. It must be noted, however, that not all defenders of neo-mechanistic explanation in (neuro)cognitive science are committed to representationalism (Kohár, 2023; Villalobos & Dewhurst, 2017), and even if they consider representational explanations mechanistic (Bechtel, 2008; Piccinini, 2020), they need not be committed to the specific account of representational mechanisms that we appeal to here. For example, Thomson and Piccinini (2018) adopt a much more liberal understanding of "representation". Thus, before we go on, we shall clarify our terminology. A *representation vehicle* is the physical medium that is processed by representational mechanisms. A *representation target* is the (possibly vacuous) referent of representing processes, and *representation contents* are the satisfaction conditions of representational vehicles. Finally, a *representational mechanism* is the computational mechanism that operates on representational vehicles and engages in representing processes when functioning correctly. The account adopted here insists that key features of representational mechanisms are:

(a) Referring to a target
(b) Identifying characteristics of the target
(c) Evaluating the epistemic value of information about the target

These characteristics enable a naturalistic approach to representational content whose satisfaction conditions are grounded in relational properties of representational mechanisms. In traditional terms, referring to targets provides extension, while identifying characteristics gives intension. Evaluating information determines accuracy or error. Representational mechanisms process information, requiring, at a minimum, information vehicles to be compared for consistency. The computational/mechanistic approach to representation relies upon the extant approaches to cognitive representation, by complementing the received views with the emphasis on understanding representational processes in causal and computational terms.

This is clearly compatible with the dominant views on representation, in particular teleosemantic ones (Millikan, 1984), which are leading theories in cognitive (neuro)science (Shea, 2018). In fact, one can view representational

mechanisms as specific kinds of teleosemantic systems that consume semantic information and use it to guide action of cognitive systems.[1] In this perspective, representational mechanisms are primary representational consumers of representational contents. However, they can interact with downstream consumer mechanisms that are no longer representational. As Miłkowski (2015) argued, this is because consuming information must involve evaluating the epistemic value of information. For example, thanks to my representational mechanisms, I can use a typewriter, which is a mechanism that responds to my keystrokes. In a sense, it is a consumer mechanism with which I interact by inputting information, but it is entirely oblivious to the contents of this information. In other words, while there may be a number of various kinds of consumer mechanisms, only some qualify as representational: the latter must be essentially sensitive to the satisfaction conditions.

Teleosemantics may embrace a more liberal view on representation, assuming that there could be semantic information that influences a certain consumer device that lacks the ability to respond to the semantic value of this information in general, which leads to the conclusion that there could exist unexploitable content, or junk representations (Shapiro, 1997). In contrast, representational mechanisms must respond to the satisfaction conditions of representational contents of their vehicles, for example, by checking for consistency among multiple sources of information (Bielecka & Miłkowski, 2020). At the same time, Millikan's biosemantics emphasizes the role of the consumer part, insisting that representations "must function as representations for the system itself" (Millikan, 1989, p. 285). This is exactly what the proposal of representational mechanisms aims to account for.

For example, predictive processing frameworks posit hierarchical neural models that make inferences about sensory causes (Clark, 2013). Higher-level models generate predictions about lower-level signals. Prediction errors reflecting discrepancies are propagated up the hierarchy to revise the models. Hierarchical predictive processing models can be understood in mechanistic terms (Badcock *et al.*, 2019), which enables also the application of the framework of representational mechanisms in this case. The predictive mechanisms can be considered not only as featuring structural representations (as argued by Gładziejewski, 2016), but also as parts of representational mechanisms referring to targets (causes of sensory stimulation), identifying characteristics (via generative predictive models), and evaluating information (by computing prediction errors). Their recurrent interaction enables representationally relevant information (Buckner, 2022). In fact, the explanatory practices in cognitive science frequently appeal to similar capacities of cognitive systems to explain various phenomena.

However, this example falls short of providing a satisfactory explanation, not only because we did not mention a single phenomenon to be explained but also because it only provided a rough description of the component parts and their operations. This is what mechanists call *sketches of mechanisms*, in which some structural aspects of a mechanism are omitted. Admittedly, we have provided merely a functional analysis of a predictive processing system, but to yield a full-blown explanation, we must fill in a number of missing aspects. Otherwise, this functional analysis will fail to be satisfactory from the mechanistic point of view (Piccinini & Craver, 2011).

The function of representational mechanisms cannot be reduced to mere information processing or transmission. This implies that computation is insufficient for representation. Indeed, mechanists stress that computationalism can be dissociated from representationalism (Fresco & Miłkowski, 2021; Villalobos & Dewhurst, 2017), even if both are compatible (Bechtel, 2008; Miłkowski, 2017b; Piccinini, 2022). The function of representational mechanisms is to modify readiness for action in content-sensitive, goal-directed ways. This semantics-sensitive teleology differentiates them from

[1] In this paper, we cannot fully address the complex issue of which philosophical account of function is compatible with both teleosemantics and the theory of representational mechanisms. While we do not endorse a popular but flawed argument that teleosemantics ascribes functions that cannot be causally relevant (for the full argument, see Miłkowski, 2016), the practice of cognitive neuroscience appears to diverge from Millikan's approach to proper functions (Hacohen, 2022). An additional challenge for our approach is that representational mechanisms constitute a proper subset of functional computational mechanisms (see Section 4 below). Regrettably, a comprehensive treatment of this issue exceeds the scope of this paper.

merely computational mechanisms. Yet, they were defined liberally to encompass diverse kinds of semantic information, including correspondence-based information. Overall, the notion provides a systematic framework for analyzing informational structures in cognitive systems.

3. *Correspondence-based semantic information*

The correspondence-based account of semantic information (Miłkowski, 2023) builds upon Barwise and Seligman's (1997) pioneering work on information flow in distributed systems. Their approach provides the formal foundation for understanding how information flows across multiple connected classifications. Our extension to correspondence networks explains how cognitive systems exploit these flows through multiple channels operating concurrently, establishing a richer explanatory framework than single-correspondence accounts.

In our account, semantic information relies on structural similarity between informational structures. The degree of similarity gives rise to conditions of satisfaction, allowing evaluation of accuracy. Correspondence is formally analyzed using infomorphism (Barwise & Seligman, 1997). This relation is defined between two classifications. We understand classification as holding between particular physical tokens and some type. For example, one can consider optical character recognition (OCR) tasks as involving classifications in this sentence: an old typewriter can introduce some variations, but the characters are classified under a single type (Latin letter) by a typist, and these can be recognized by OCR systems (see Figure 1).

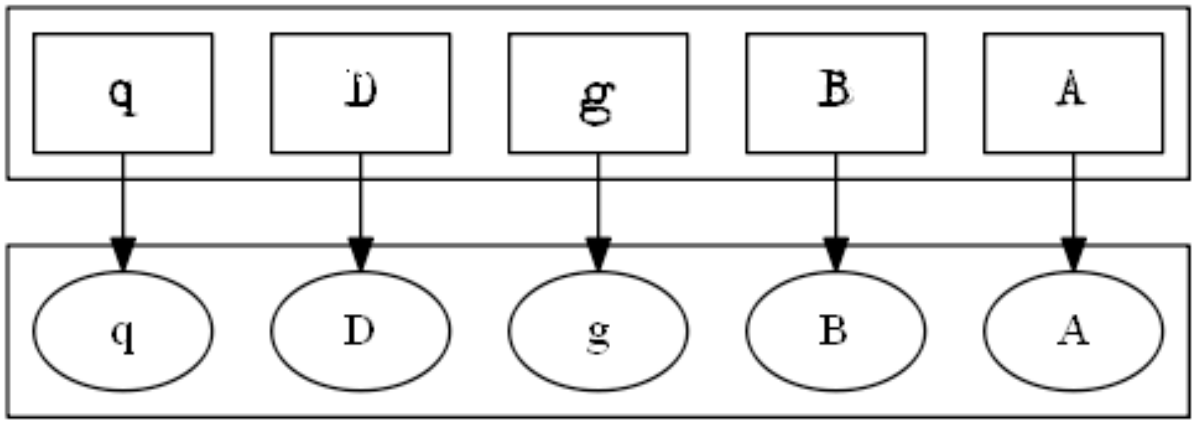

Figure 1

Typewritten characters correspond to Latin alphabet characters

Thus, an OCR system is tasked with classifying various images of letters under appropriate types, and then mapping these types to Latin characters, which are encoded digitally in a particular way (which also implies physical vehicles, or tokens). What the OCR system does is to extract information from images because there is a correspondence between images on the page and Latin characters. In other words, an infomorphism exists when the type-token mappings are preserved between two classifications. This ensures tokens classified under the same image types will remain under corresponding Latin character types. The structure and relations between classified objects is maintained.

We can now state it more formally. An infomorphism is a structure-preserving map between two classifications, **A** = <A, Σ_A, $\vDash_A$> and **B** = <B, Σ_B, $\vDash_B$> where:

— A and B are sets of objects (tokens);

— Σ_A and Σ_B are sets of objects used for classification (types);

— $\vDash_A$ and $\vDash_B$ are binary relations between tokens and types that tell one which tokens are classified as being of which types.

An infomorphism f from **A** to **B** is a pair of functions $<f^\wedge, f^\vee>$ (see Figure 2) such that:

— $f^\wedge: \Sigma_A \to \Sigma_B$,

— $f^\vee: B \to A$,

— for all types $a \in \Sigma_A$ and for all tokens $b \in B$ the functions $f^\wedge$ and $f^\vee$ satisfy the following biconditional: $f^\vee(b) \vDash_A a$ if and only if $b \vDash_B f^\wedge(a)$.

Since Barwise and Seligman introduced this notion in the context of information channels, the functions $f^\wedge$ (read "f-up") and $f^\vee$ (read "f-down") can be intuitively understood as providing encoding and decoding. Thus, tokens of b can be encoded (via $f^\wedge$) and classified using the classification relation $\vDash_A$ as being of types a. When they can be also classified using $\vDash_B$ as being of types provided by the decoding relation ($f^\vee$), then an infomorphism obtains: there is a structure-preserving map. Obviously, this can be used directly to model noiseless information transfer in a channel.

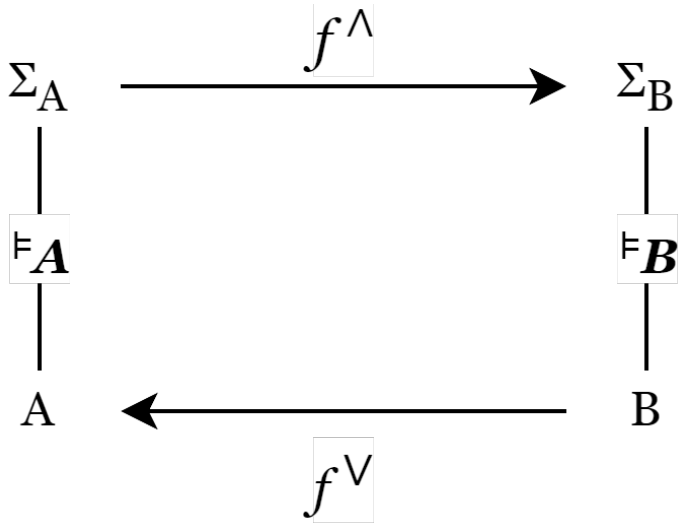

Figure 2

An infomorphism obtains between A and B

However, similarity admits degrees. Fuzzy infomorphism, or infocorrespondence, incorporates prototypical category membership and partial resemblance, which is inspired by the fuzzy set theory (Zadeh, 1965). Infocorrespondence is a generalization of infomorphism that allows for partial similarity between classifications. It can be defined for fuzzy classifications **A** = <A, Σ_A, $\vDash_A$> and **B** = <B, Σ_B, $\vDash_B$> where:

— A and B are sets of objects (tokens),

— Σ_A and Σ_B are fuzzy sets of objects used for classification (types),

— $\models_A$ and $\models_B$ are fuzzy binary relations between tokens and types.

An infocorrespondence f from **A** to **B** is a pair of functions $\langle f^\wedge, f^\vee \rangle$ defined in the same manner, but the fundamental biconditional $f^\vee(b) \models_A a$ if and only if $b \models_B f^\wedge(a)$ holds for at least some tokens $b \in$ B and some types $a \in \Sigma_A$.

In fuzzy classifications, tokens belong to types or satisfy relations to some degree. Infocorrespondence only requires the type-token mappings to be preserved partially between the classifications. This allows for approximate similarity rather than an infomorphism's stricter structure preservation. Quantifying the degrees of partial mapping allows one to measure the extent of correspondence. In other words, tokens belong to types to a degree, but some mappings remain. This fuzzy correspondence still licenses surrogate reasoning between informational structures. For instance, some OCR systems are able to fill in the gaps, by guessing the correct letter in a word based on the surrounding context. In this case, they infer what the intervening, unreadable character might be. Thus, systematic similarity, or correspondence, is used to drive a kind of reasoning.

Critically, correspondence occurs between informational structures, not abstract propositions. The relata are physical vehicles classified in terms of types and tokens. This avoids positing notions of propositional structure matching external states. Instead, resemblance involves concrete classified entities.

Diverse similarity metrics can build on this basic infomorphism framework. For example, mutual information measures dependence between joint and marginal probabilities, indicating correspondence between distributions and thereby integrating probabilistic and structural approaches to information. Tversky's (1977) contrast model offers another approach, providing non-probabilistic measures of this relationship.

Overall, correspondence-based semantic information depends on inferentially useful resemblance between informational structures. Fuzzy infomorphism, or infocorrespondence provides a general framework, which enables various approaches to similarity.

4. Representational Mechanisms Exploiting Correspondence Information

The account of correspondence-based information, summarized above, does not provide any detail on how this information may be exploited by cognitive mechanisms. In this section, we argue that correspondence-based information can be exploited by a class of representational mechanisms. These two frameworks, joined together, provide a formal and causal account of structural representation, which is usually presented as providing a crucial feature: it legitimizes surrogate reasoning between structures. From the preceding, it should be fairly clear that infocorrespondence does the same because of the structure-preserving nature of this relationship. Representational mechanisms can thus leverage it to fulfill their key functions.

Specifically, the informational structure referred to by a mechanism exhibits infocorrespondence with a target structure. This licenses the identification of characteristics of the target based on the information vehicle. Evaluating information also relies on assessing similarity between the structures to determine accuracy. Vehicular discrepancies reveal errors in representing the target.

The mechanistic approach demands a form of realism concerning representation, implying that both the vehicles and the contents play a causally relevant role in the phenomena governed by a particular mechanism. This emphasis on the causal relevance of both vehicles and contents is one of the conditions for representational mechanisms dealing with structural representations. Further conditions follow Gładziejewski's (2015a, 2016) widely accepted contemporary

philosophical analysis of structural representations, which can be summarized as follows. In the context of a system denoted as S, a vehicle denoted as V qualifies as the vehicle for a structural representation of a target entity, designated as T, if and only if the following conditions hold:

(1) *Structural Similarity*: V exhibits structural similarity to T.
(2) *Action Guidance*: The structural similarity outlined in (1) enables V to provide guidance to S's actions concerning T.
(3) *Decouplability*: Condition (2) remains applicable even when V is detached or decoupled from T.
(4) *Error Detection*: S possesses the capability to detect representational errors generated by V.

The present account endorses conditions (1), (2), and (4), while rejecting condition (3), at least in a strong interpretation. While there may be decouplable representations in representational mechanisms, the fact that a given representation remains coupled (for example, through perceptual inputs) to the target does not disqualify it from its representational roles (Miłkowski, 2017a). The condition (2) implies that representational mechanisms should guide action in a broad understanding of the term. What should be clear, however, is that the same vehicle V can guide various actions when exploited by various downstream mechanisms. Indeed, as contemporary research suggests (Mante, Sussillo, Shenoy & Newsome, 2013), the same vehicle, depending on the computational processes it is involved in, can be processed in various ways to derive distinct downstream representations. In fact, this could be a computationally efficient way to make various kinds of use of the same input information. At the same time, vehicles attain their status as vehicles of structural representations only if they can be checked for consistency with other vehicles, which enables error detection in cognitive representations (Bielecka & Miłkowski, 2020).

As Miłkowski (2013) notes, this makes some discussions about various extrinsic content ascriptions moot: for example, notorious bug detectors may not fulfill any representational role just because amphibians in general do not seem to detect any inconsistency between the signal they receive from the tongue when they consume dummies as opposed to bugs. If they do, then the exact kinds of error detectable by their representation consuming devices must constrain our content ascriptions. Otherwise, these content ascriptions boil down to epiphenomenal theoretical glosses.

Structural similarity is understood here in terms of the correspondence-based account of semantic information —it must conform to the conditions that infocorrespondences fulfill. This implies that there must be two classifications of tokens into types that are interlinked with at least a partial mapping. As a matter of fact, the reason why this linkage exists in cognitive systems is that these systems respond to information vehicles systematically, establishing an information channel, which, in turn, triggers the information flow. Systematicity is precisely what enables utilizing cognitive information about members of the class beyond what is currently accessible to the cognitive system. Hence, one can reason about absent, past or counterfactuals tokens of the given type in a veridical and regimented way. This implies that representational mechanisms explain phenomena that are not confined to the spatiotemporal boundaries of a single individual.

A critical feature of our framework is that representational mechanisms typically exploit not just single correspondence relations, but networks of correspondences. Neural vehicles often participate simultaneously in multiple classification schemes, with different typing relations operating concurrently (Mante *et al.*, 2013). For instance, the same pattern of neural activity in visual cortex might be classified according to shape properties in one processing stream and motion properties in another, establishing multiple correspondence channels with different aspects of the environment.

This network framework elucidates three crucial features of representational systems. First, it accounts for how the same vehicles can carry different contents in different processing contexts —as observed in the multiple processing pathways in vision (Milner & Goodale, 1995; Pitcher & Ungerleider, 2021). Second, it enables cross-channel evalua-

tion, where information from different correspondence channels can be compared to detect inconsistencies and evaluate accuracy. Third, it suggests why neural representational spaces consistently organize along semantic dimensions — the network preserves taxonomic structure across processing channels.

The Correspondence Network Framework thus goes beyond simple structural similarity accounts by explaining how multiple concurrent correspondence relations establish a rich representational architecture capable of content-sensitive processing.

In a recent critique, Kohar (2023) confines mechanistic explanations to phenomena within the spatial and temporal boundaries of their mechanisms, arguing that constitutive mechanistic explanations rely on the local supervenience of phenomena on the component parts and operations of mechanisms. Clearly, at least some representational phenomena extend beyond these limits because they rely on external information flows. This broader scope is essential for explaining veridical perception, which involves not only internal perceptual mechanisms but also the causal processes linking these mechanisms to external objects. If Kohar is correct, then constitutive mechanistic explanations of perceptual phenomena would not exist. However, there is no reason to believe that mechanistic explanations should be constrained in this way. Not only would this fail to account for mechanistic explanations in scientific practice, but it would also be arbitrary and epistemologically unfounded. No epistemological norm dictates that causal explanations must adhere to armchair metaphysical ideas about local supervenience, and no major work in mechanistic philosophy endorses such ideas.[2]

Thanks to multiple independent information flows, representational mechanisms can compare information across channels to gain proxy access to semantic properties, such as the existence of perceived objects or truth. Satisfaction conditions are typically available to cognitive systems through operations across multiple information channels. Overall, the mechanistic explanation in this case is hybrid: it involves an etiological explanation of the information flow, constitutive explanations of information channels, and constitutive explanations of representational mechanisms. Moreover, no plausible epistemological norm suggests that such hybrid explanations are a priori invalid.

Kohar argues that advocates of mechanistic explanation should reject representationalism instead. Specifically, he claims that anti-representationalism can explain the same phenomena mechanistically, and that semantic properties of structural representations, such as their degree of similarity to targets, are not needed to explain success (p. 153-158). He argues that since structural similarity supervenes on the vehicles' structure, the success can be explained by referring to structural features of vehicles (p. 154).

But his argument is unsound. Structural similarity is a two-place relation and no two-place relation can supervene on one relatum only (unless we speak of identity, which is a very atypical relation). Furthermore, since the same neural vehicle structure can support multiple typing schemes, as in multiple processing pathways in vision independent information channels with distinct classifications, it cannot be the case that the vehicular structure *alone* corresponds one-to-one to a particular satisfaction condition. For this reason, the vehicles' structure cannot be explanatory of success.

In addition, the structure of the vehicle, or, to use our terminology, the classification used to map tokens into types, is only explained through the interactions occurring in a given information channel, which goes beyond the bounds of the nervous system. Kohar seems to imply that an incomplete explanation is to be preferred instead, since local supervenience cannot occur across channels that span multiple distinct mechanisms with their distinct phenomena.

---

[2] Craver in one of his papers (Craver, 2007a) indeed spoke of supervenience, in a parenthesis, but he explicitly rejected its utility in his major monograph (Craver, 2007b, p. 153).

From our point of view, the major shortcoming of Kohar's view is that he focuses solely on individual tokens, or physical vehicles in a mechanism, ignoring two critical considerations: (1) multiple classifications and (2) the causality of information flow across channels. Tokens are properly informational only when classified into types, often in several ways at the same time, and they are processed in complex, often recurrent, processing webs.

Furthermore, the linkage between similarity and success is complex, and it is not purely conceptual. Representations usually involve dimensionality reduction of information that is available to the cognitive systems. Structural representations, as noted by Gładziejewski and Miłkowski (2017, p. 343), "that resemble the target too much become excessively complex themselves". There are two reasons for this. First, fully mirroring the structural complexities of the world imposes prohibitive costs on limited cognitive agents. Second, overly accurate representations may have limited predictive value because their structure may result from overfitting the structural features of the world. While it is certainly possible to represent a simple system using a very complex representation, and in some border cases, a full-blown infomorphism could hold, in general, neural systems typically simplify reality. Thus, similarity alone does not explain success according to Gładziejewski and Miłkowski, which implies that Kohar opposes a line of argument that is avowedly rejected by them.

Facchin (2024) has recently mounted a second critique of structural representationalism. His critique takes an opposite approach to Kohar's strategy. Instead of criticizing the promiscuity of ascribing structural representations, he embraces a pluralistic understanding and argues that representationalism, even in its pluralistic form, is too narrow to address its own desiderata, e.g., the job description challenge (Facchin, 2021).

According to his account, there are two distinct issues concerning structural representations: their status and their content. The former concerns whether a given physical vehicle achieves representational status, while the latter assumes representational status and questions whether the content of a representation is truly responsible for the involvement of a given part of the cognitive system. In other words, the status of structural representations is distinct from their causal efficacy.

While Facchin acknowledges that his aims are primarily clarificatory and taxonomical (he proposes four tentative taxa), his goal is not to pave the way for a more sophisticated representational account but to argue for discarding it altogether. For him, disentangling the commitments to status and content in structural representations reveals representationalism's inability to provide satisfactory answers. His main argument, though tacit, can be reconstructed as follows: charting a map of representationalist accounts does not clarify or bolster their case, but instead exposes their lack of explanatory scope.

Our account, however, aims to explain how a framework based on structural relations —namely, similarity— can provide a complex but consistent picture of structural representations. While similarity is a crucial factor, it is not sufficient to explain representational phenomena. We demonstrate how content remains essential to explaining the causal powers of representational mechanisms, while also emphasizing the required elasticity of concurrent infomorphic mappings. This approach aligns with a pluralistic reading of structural representational posits, with which we agree with Facchin. As such, our proposal builds on type-token theories of content in semantic networks (e.g., Shea, 2007) but expands them by explaining how semantic, informational relations are established without remaining rigidly fixed. This extension helps address anti-representational worries in a novel way.

Now let us turn to how RSA conceptualizes similarity in terms of representational geometry.

5. Unlocking Meaning from Matching: The Role of RSA in Cognitive Science Practice

Representational Similarity Analysis (RSA) provides a compelling case study for understanding how the Correspondence Network Framework applies to cognitive neuroscience practice. RSA reveals patterns of correspondence across

multiple classificatory schemes, detecting networks of similarity relations rather than isolated structural mappings. This methodology implicitly relies on the taxonomic organization that emerges from correspondence networks, making it an ideal testbed for our framework.

RSA is a family of multivariate pattern analysis methods employed in neuroscience to delineate "representational geometries" of cognitive systems (Kriegeskorte & Kievit, 2013). Intuitively, representational geometry is expected to offer a systematic map of how the perceptual state space is organized. Brain activity patterns, elicited by various stimuli, are measured (notably, an array of methods, from single-cell recordings to wide-area fMRI, can be applied) and correlation between these responses is calculated. Strictly speaking, the objects or stimuli do not participate directly in this operation, i.e. the correlation is not measured between the stimuli and the response, but between two stimuli-induced responses that differ in some experimentally interesting way.

These comparative results are subsequently compiled into a systematic numerical description of the extent to which neural activity is shared across stimuli. When patterns exhibit structural similarity, it implies shared underlying representations as the source of identifying characteristics, while pattern mismatches, on the other hand, suggest representational differences. The outcome of this procedure is the description of the representational geometry. However, the precise construction of this geometry and its alignment with our understanding of semantic information raise pertinent questions.

From a technical standpoint, representational geometry serves as a mathematical description of the similarity relations between reactions to objects (scil. stimuli-induced patterns of activity) within a given dataset (Mur *et al.*, 2009). The primary goal of this procedure is to operationalize the concept of similarity among these objects. These objects are conceptualized as points within a high-dimensional space, where each dimension corresponds to the range of activity measured in a specific part of the overall activation space. For instance, in the context of fMRI, this activation space encompasses all the voxels under observation, with each dimension representing the range of activity in a single voxel. These activity patterns intuitively diverge for distinct classes of objects, positioning them in different regions of this geometric space.

To quantify the degree of dissimilarity (or similarity), a similarity metric must be selected. Options include angular, Euclidean, and Cartesian metrics. The choice of metric may be driven by theoretical considerations that preserve different types of properties, although the technical details need not concern us here. It suffices to acknowledge that the procedure maintains its character as a representational space as long as the chosen metric is applied consistently.

Once the measurements are collected, the space is established, and the chosen metric is employed to compute the similarity relations, one can assert that these geometrical relationships represent mappings of semantic relations between the representations of objects. The complete set of these relationships collectively constitutes what we refer to as *representational geometry*. Consequently, we can view it as an operationalization of the similarity relations among objects within the analyzed dataset.

Significantly, RSA as a method is a two-step process. The initial step involves the construction of the representational geometry. To achieve this, pairwise similarity patterns are gathered. To be more precise, these patterns are derived from the subtraction of activity patterns of the same set of units (notably, in the case of fMRI, these are voxels; however, RSA is versatile and can encompass EEG electrodes, single-cell recordings, behavioral judgments, or unit activations in artificial neural networks). In a simplified example, we present the subject with stimuli such as a banana and an armadillo, record the activity across the selected units, and then compute the dissimilarity of the responses. This is calculated by subtracting the correlation score from 1 (i.e., score = 1 - correlation). These individual dissimilarity entries are incorporated into a matrix, creating a dissimilarity score. The collection of these pairwise dissimilarities between objects constitutes the first-order RSA, often referred to as a representational dissimilarity matrix (RDM). In essence, it constructs a representational geometry for the chosen objects and units.

The second step of "actual" RSA is focused on directly comparing these first-order RDMs, aiming to obtain a measure of similarity not between pairwise activation responses but across the overall similarity of entire representational geometries. RSA, fundamentally, entails a second-order comparison or, in another sense, a meta-similarity measurement. The essence of this procedure lies in unveiling structural similarities within the internal architecture of datasets, and more fundamentally, within the first-order geometries.

While RSA is designed to capture and compare internal structural relations, this should not imply that all internal structures of the targets are fully retained in the resulting RDMs. On the contrary, we must recall that the entries in these matrices are simply numerical values, or distances between vectors in mathematical terminology (Roskies, 2021). They represent scalar quantities and, as such, abstract away from details such as part-whole relations within the considered objects or the specific presentation format. Spatial relations are also not preserved in this manner. What RSA does retain is an overall similarity score based on the chosen recording method and metric. Typically, RSA is employed to detect dataset-level internal structure, rather than focusing on the object-level structure itself. Semantic relations are defined over representations of objects and not their individual components.

RSA does not operate on the same representational substrate as the consumers within brain tissue —whether these consumers are brain regions, cortical layers, psychological modules, individual nodes, or other neural entities. The voxels typically analyzed by RSA are far more granular than the leading candidates for neural processing units. However, this discrepancy does not undermine the method's predictive success, as has been tentatively demonstrated (Thornton & Mitchell, 2018). Nevertheless, RSA, emerging from early multivariate pattern analysis (MVPA) procedures like the classical setup presented in Haxby *et al.* (1991), is promoted by its designers as a means to reveal the actual information flow within cognitive systems (Kriegeskorte *et al.*, 2006). Consequently, it firmly positions itself within the domain of neural decoding.

Moreover, RSA was specifically designed to address criticisms that previous MVPA methods function as spurious data classifiers when used for reverse inference. The principal objection to inferring experiential content from neural activity was that such methods abstracted away from internal relations within the signal (Rathkopf *et al.*, 2023). Earlier multivariate analyses failed to acknowledge that justifiable inference requires understanding the internal structure of a system's representational space. Ritchie *et al.* (2019) labeled the unfounded belief that neural decoding could proceed without these structural insights the "decoder's dictum," which they demonstrated to be false. Their central question —if we cannot identify what information our classifier uses, how can we verify it operates on principles similar to the system being modeled?— reflects a specific instance of the broader problem of opacity in machine learning, where classifiers routinely take computational shortcuts (Zednik, 2021).

Kriegeskorte and his colleagues suggest that RSA is the preferred approach to sidestep the pitfalls associated with the decoder's dictum and to attain more accurate neural decoding than previously employed methods. The underlying rationale for this assertion lies in RSA's capacity to represent the internal structure within a model, grounding the psychological model used within a given system in measurable data. This, in turn, mitigates the risk of potentially superficial classification, a concern that plagued earlier curve-fitting models. What truly sets RSA apart is its unique ability to map the semantic structure by reconstructing the cognitive space, giving form to the representational geometry of the system. As contended by Kriegeskorte and his team, the RSA project serves as a means to decode the information within a system, effectively working for the system itself, as opposed to acquiring information about the system as external observers conducting experiments.

Upon closer examination, this approach doesn't seem to advance us any further toward the desired goal than earlier MVPA attempts. It's worth recalling that the flexibility in the types of entries accepted for second-order comparisons, along with the subsequent format-agnostic nature, casts doubt on the notion of these scalar quantities serving as genuine carriers of content. The representations within matrices of "representational" similarities only qualify as representations in the mathematical sense of the term, as seen in the AI literature, and not as representations in the cognitive sense. Roskies (2021) aptly characterizes them as 'provisional' representations, and in our framework, this would qual-

ify as mere semantic information. This would not be a significant issue if a more foundational, shared framework could be established in terms of information theory.

Nonetheless, the entities involved in transmitting the signal, including the vehicles and the format of the presumed coding scheme, are unlikely to directly align with the actual mechanisms operating in the brain. So, the information, in this context and scheme, remains provisional. Perhaps our present empirical approximations are sufficiently reliable. Therefore, as we progress towards more advanced brain imaging, it is conceivable that the proposed mappings could remain accurate. This would suggest that both provisional representations and provisional information may indeed be precise and mutually aligned. However, the overall empirical accuracy of RSA neural decoding depends on whether the information *within* and *about* the system eventually harmonizes.

So, is our conclusion merely critical? Are we dismissing RSA as a method for uncovering neural content? Not at all; quite the opposite. To appreciate its value, we suggest the need to:

1. Focus on its insights into semantic clustering.
2. Reinterpret what RSA reveals about neural information flow and neural information more broadly.

To address the former, we will briefly examine research on the emergence of nested classifications within RSA analysis and its taxonomic characteristics. For the latter, we propose considering information in terms of structural similarity as a correspondence relation, aligning with the correspondence-based theory of semantic information.

One of the most intriguing aspects of RSA-based studies on human and animal brains is the emergence of a categorical structure within representational geometries that exhibits reasonable invariance across subjects and even species. For instance, human and macaque inferotemporal cortices (IT) appear to follow similar coding principles (Kiani *et al.*, 2007). The invariance being discussed is not absolute; the response matrices in the representational dissimilarity matrices (RDMs) do not precisely mirror each other's entry values. However, there is a substantial degree of overlap in response patterns between humans and monkeys.

What do we mean by "categorical structure" in this context? It implies that within representational geometries, grounded in measurable data, we identify clusters that align with semantically interpretable terms, reflecting a nested hierarchy of objects. For example, studies show a clear division when subjects view inanimate and animate objects (Jozwik, Kriegeskorte, & Mur, 2016; Kriegeskorte, Mur, & Bandettini, 2008; Mur *et al.*, 2013). High similarity judgments are consistently observed within pairs of animate objects and within pairs of inanimate objects, but not across these categories.

This observation is also evident at a finer level of categorization within subcategories of animate and inanimate objects. For example, a sub-cluster emerges from humans to animals, followed by a distinction between quadrupeds and non-quadrupeds. Faces cluster with other facial features, while hands and limbs also group together. Importantly, this difference in the IT areas appears to be rooted in categorical distinctions rather than feature-based mechanisms. Jóźwik *et al.* argued against the feature-based mechanism, as their statistical analyses showed that feature-based models do not sufficiently account for the variance in the dataset, while categorical models offer a much better fit.

Why are these results of particular interest to us? In most classical interpretations (for more nuanced interpretations, see Isaac, 2019; Mann, 2020), information theory in neuroscience typically aligns with the Shannonian perspective of coding. In this view, coding is arbitrary, meaning that the content of symbols doesn't inherently correspond to the physical attributes of the entities carrying that content. While mappings rely on the statistical properties of the signal to achieve optimal coding, they do not depend on the similarity between the entities or between the intended targets and the codes. For example, in computer storage, memory addresses do not necessarily share similarity relations; they are purely symbolic.

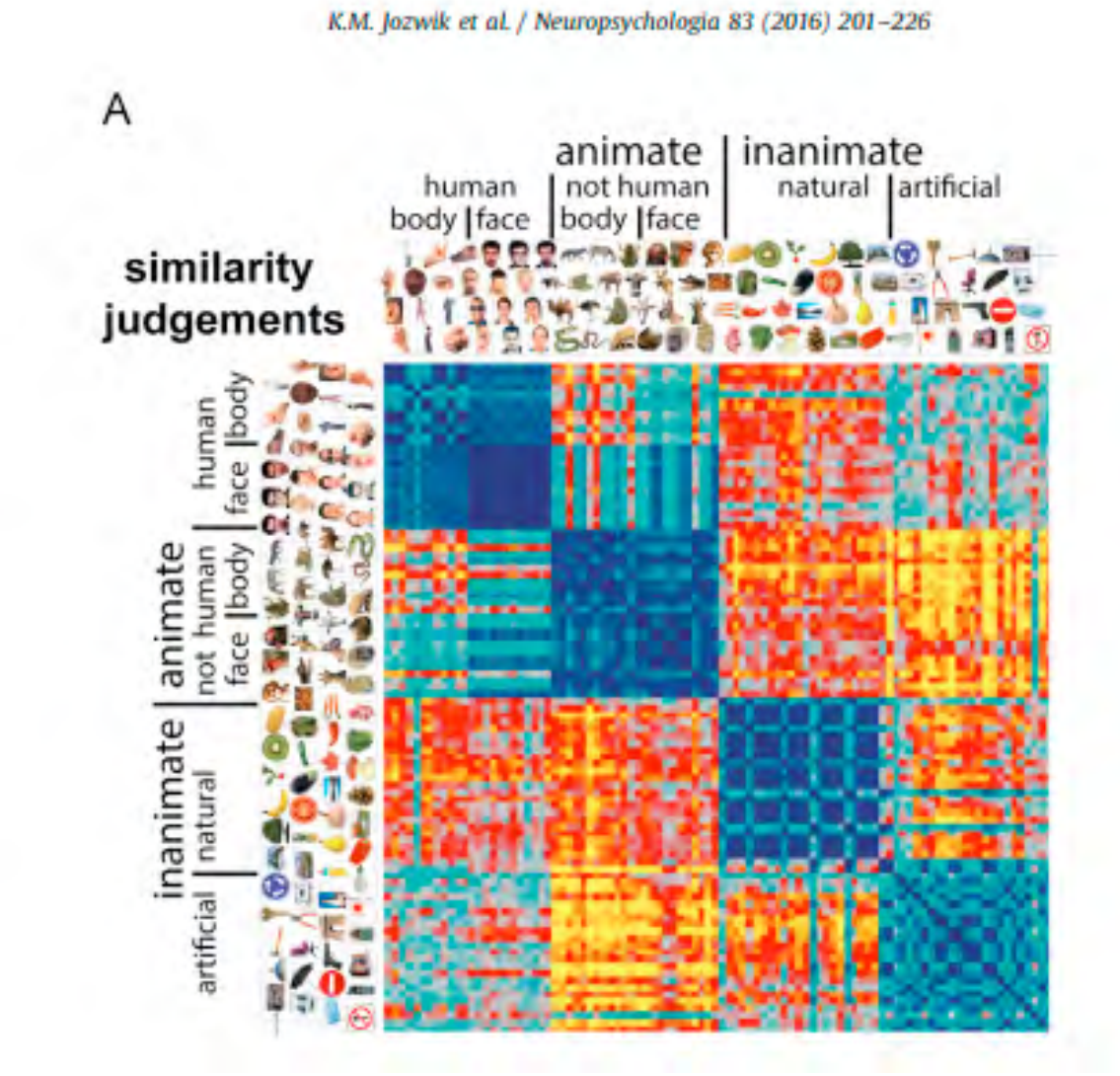

Figure 3

A diagram from Jóźwik et al. demonstrating the emergence of a nested hierarchy of semantic categories that mirror the taxonomic order of high-level features

Given the importance of semantic, similarity-based relations as organizational principles for representational geometries, it is more plausible to seek a theory to explain this property. The correspondence theory of semantic information provides such an explanation, as it takes similarity as a prior for why certain states are informative of other states. Thus, rather than conceptualizing RSA-type methods as detecting classical Shannonian information channels, we propose an alternative view: representational geometries maintain similarity-based information channels.

The RSA procedure is thus conceptualized as a series of infocorrespondence relations. First, we depict the information flow charted by the RSA analysis and the preservation of informational properties within a cognitive pipeline (see Fig. 4). This should help the reader visualize the nature of infocorrespondence relations established within the RSA pipeline.

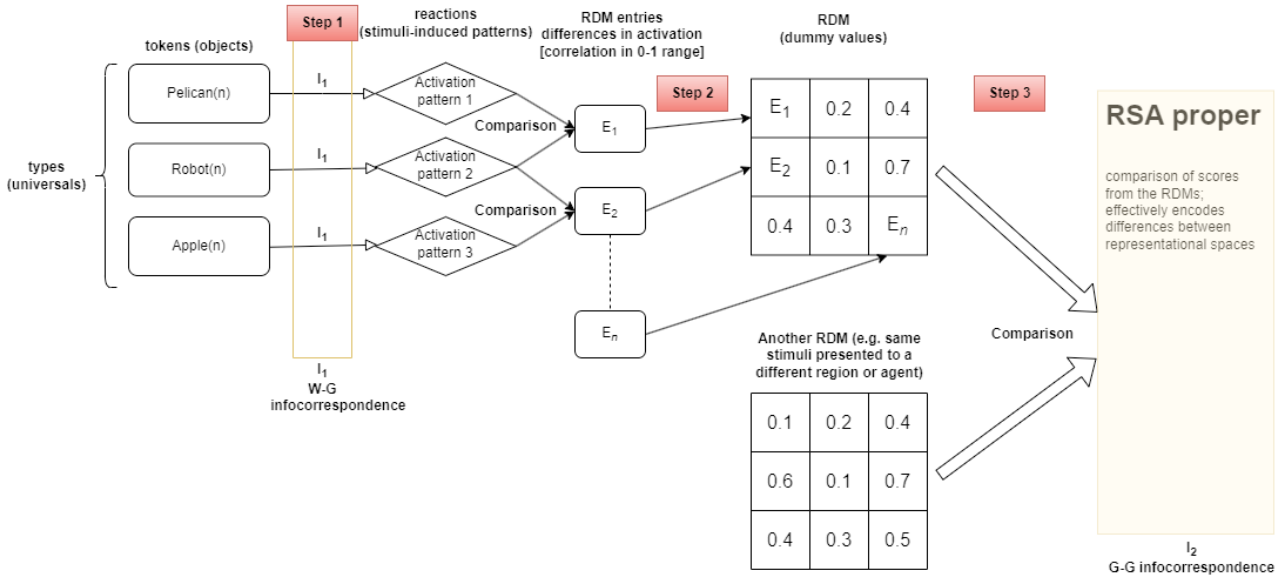

Figure 4

The RSA pipeline considered from the perspective of information flow

In our proposal, there are three "steps" effectively bound by two major infocorrespondence relations (I_1 and I_2). Moving from left to right, the first is the domain of objects presented as stimuli to agents. Tokens of objects are assigned to their respective classes (universals), governed by the classification operator (⊨universal). This constitutes the "left" half of the first infocorrespondence (I_1) as they are pairs of classifications in Barwise and Seligman's terms. Tokens presented as stimuli elicit particular activation patterns, which constitute the respective tokens on the "right" (agentive, e.g., neural) side. These tokens are likewise subsumed under counterpart types, according to the categorical operator (⊨category). Entries in the RDMs encode pairwise differences between the objects (operationalized as distances between vectors that encode these objects). It is plausible to assume that the differences are judged according to the properties that the objects do not share. Hence, at this level, these properties are the types, whereas the activation patterns are tokens into which they are distributed. That constitutes Step 2 on the pathway.

The "world to geometry" infocorrespondence (W-G) is crucial for disambiguating objects and placing them in proper categories. Objects lend their properties to be mirrored and utilized by agents, binding targets to representational geometries and involving actual external objects. This aligns with the view that neural information is selective responding to stimuli, with certain firing patterns reliably correlating. The binding is arbitrary, as any content could be coded by any neural state, similar to memory addresses in digital computers. However, our view diverges from this picture in later stages.

The final step —Step 3— is the actual G-G (geometry to geometry) comparison, where two full-fledged representational spaces are assessed against each other. We can then reliably infer the contents and properties of one based on the other and make "predictions" (on special status of the word in neural decoding, see Appendix to (Kriegeskorte & Bandettini, 2007)).

The details of the mapping may be better understood using Table 1. We can see which items during the entire procedure are conceived as tokens and which as types, and what type of categorical operator binds them (governs how the tokens are classified into relevant types). A pair of infocorrespondence relations —first, between stimuli and the given representational geometry (I_1), and then, between two representational geometries that are being computed (I_2)— renders the information flow with structure-preserving (semantic) properties not only possible but also valid.

Table 1

Tokens, types, and operators inherent in infocorrespondence relations (IMs) within the RSA pipeline

	Step 1: Stimulus exposition	Step 2: RDM – world-geometry infocorrespondence	Step 3: RSA proper – geometry-geometry infocorrespondence
Tokens	singular objects	activation differences (distances)	points in representational space
Types	universals	differentiating properties	semantic clusters or taxa
Operator	⊨(universal)	⊨(category)	⊨(σB)

In the first column of Table 1, the classification is relatively simple. Singular objects in the world pertain to respective universal classes. Instances (tokens) of objects are assigned to their labeled sets, or in other words, universals. This is regulated via the operator we call ⊨(universal). In the middle column (linked with the first one through the first infocorrespondence relation I_1, which we denote by the pair of arrows on the top), when the objects are presented as stimuli, they are encoded by activation patterns, which in turn are translated into vectors in the state (activation) space. The distances between the vectors representing the activation patterns are compared. These distances are representations of differences between the given objects. In our terms, these differences effectively represent discriminative properties of the types of objects (types from the left column). That is why these discriminative properties constitute types on the second step of the infocorrespondence pathway. As objects correspond to the corresponding activation patterns, universals correspond to the discriminative properties that set them apart. We call the operator working on this level the ⊨(category) operator, as its job is to assign the activation patterns into categories that they encode for the agent.

Finally, in the third step, where the 'proper' RSA comparisons are done, we encounter a geometry-to-geometry (G-G) infocorrespondence, where semantic informational relations between the representational spaces themselves are obtained. Tokens to be categorized at this level are points in the given representational space that stand in for particular targets. Types to which they are assigned are semantic clusters or taxa that contain information about the objects, but also their relationships with other objects encoded in the space (proximity stands for similarity, but certain dimensions stand for properties). We call the operator on that level ⊨(σB). It classifies the tokens based on the structure of relations that describe/organize the representational geometry.

To wrap up, the results derived from RSA seem to imply that the coding principles within neural systems align with semantic properties —in plain terms, they convey information about what they encode. The retention of a semantic hierarchy, preserving taxonomic internal structure and clustering of mutually similar targets, serves as evidence for this architectural pattern. Current theories of information encounter challenges when attempting to explain this characteristic of "informativeness" within a system, at least at first glance. The framework proposed here, however, does not grapple with this problem and provides an explanation for why similarity organizes the informational structure of neural systems in a semantic manner. While the mere presence of clustering could be attributed to efficiency considerations within the traditional framework, it fails to offer an understanding of why such clustering should align with semantic categorizations.

Let's revisit the key points regarding RSA and its connection to semantic information, grounded in structural similarity:

T1: There exists an infocorrespondence relation, both among the representational geometries themselves (G-G) and between geometries and the external world (G-W) on a pairwise basis.
T2: This very relation is what facilitates the flow of semantic information.
T3: The presence of semantic information offers a more robust explanation for the retention of property relations between representational geometries, a challenge that the theory of arbitrary coding (i.e., Shannonian) finds difficult to address.
T4: Structural similarity provides the basis for researchers (external observers) to make surrogative inferences.

Importantly, exploiting correspondence information necessitates additional organizational constraints. Mere resemblance is insufficient for representation (Goodman, 1972). However, the theories that rely on resemblance seldom, if ever, ground it solely in structural relationships (Decock & Douven, 2010). However, when integrated into teleological representational mechanisms, it fosters semantically evaluable content. These mechanisms introduce further conditions to delineate suitable informational structures and targets. Correspondence, in this context, serves as the fundamental relational framework connecting them.

The theory of representational mechanisms additionally requires:

T5: Representational content is exploited downstream, which implies that representational geometries should have causal impact on evaluation mechanisms that detect error.

This implies that the theory offers a potent heuristic: when you suspect the presence of a representation within a cognitive system, direct your attention to the evaluation mechanisms and study the subsequent processes. To put it differently, what renders RSA an *incomplete* account of representation is its limited capacity to systematically address this question. Nevertheless, when coupled with the theory advocated here, RSA can be employed to propose additional experimental studies.

The systemic mechanistic perspective sheds light on why RSA yields insights into clusters, laying the foundation for representationally valuable inferences. Structural similarity serves as the basis for drawing conclusions about representational content and targets. Moreover, the correspondence theory of semantic information implies that as long as stimuli evoke activity patterns, these establish an information flow, which implies that there is an infocorrespondence between stimuli and activity patterns.

In summary, a systematic analysis of RSA highlights the role of correspondence-based information in cognitive systems, showing how semantic properties support mechanistic explanations and fulfill representational needs. Hence, we argue that the Correspondence Network Framework deepens our understanding of cognitive science practice and computational representation processing. Philosophically, it provides theoretical foundations for existing RSA methods, linking informational vehicles, contents, and targets —addressing a critical gap in the literature. By integrating resemblance, teleology, and mechanism, it offers a powerful framework for understanding representation.

6. Potential objections

Now, we address potential objections to our account. Although some objections have appeared in the literature, we present them here to underscore their relevance to semantic matching.

> *Differing Formats:* Critics like Ritchie *et al.* (2019) highlight that information acquired through decoding methods like MVPA does not match the brain's inherent format. This mismatch between analytical units (e.g., voxels) and actual neural structure means we process information as external to the system rather than internal. Facchin (2023) contends that this discrepancy prevents our understanding from aligning with the brain's native processing.

While our methods undoubtedly use artificial constructs influenced by experimental design, neural decoding paradigms aim to converge information about and within the system. Thornton & Mitchell's (2018) success in predicting neural firing patterns supports the causal efficacy of semantic relations, suggesting potential despite current methodological limitations. The alignment between theoretical arguments and empirical results indicates growing convergence. RSA findings reflect real patterns (Dennett, 1991), and critics must demonstrate that these correlations are mere artifacts. RSA resembles a measurement device with limited resolution —microscopic structures can be meaningfully captured at

a macroscopic level without revealing every detail. Granularity is after all tied to the "quality" of structural correspondence. The fuzzier is the relation between classifications bound by the infomorphism, the coarser the informational flow between them.

> *Mere toolbox*: Critics argue that RSA protocols lack sufficient constraints to indicate real cognitive currency, merely describing similarities between models or systems without providing explanations. The flexibility of RSA across species, agents, localizations, and modes is viewed as a weakness that may lead to spurious correlations (Kriegeskorte *et al.*, 2008). Ritchie *et al.* (2021) demonstrate how protocol choices like normalization techniques can influence neuroregistration effects, raising concerns about the reliability of observed correlations.

However, this apparent "promiscuity" and statistical volatility actually underscore the importance of theoretical supplementation rather than undermining it. Our semantic information account provides both normative reasoning and a formal apparatus to explain why RSA-type evidence transcends being merely "a toolbox to think" about content-individuation. By demonstrating why semantic-informational relations are necessarily employed in representational mechanisms, we support the "ambitious" camp of RSA supporters (or "theoretically motivated," in Ritchie's terms), justifying RSA as more than merely descriptive.

> *Psychological Models:* Critics like Bowers *et al.* (2022) argue that successful decoding schemes require psychological models to demonstrate explanatory power, not just predictive capabilities. They contend that decoders should operate on principles similar to target cognitive systems —a standard not consistently met, as evidenced by deep neural networks that fail to represent part-whole relationships or prioritize the same aspects as humans in categorization tasks. If RSA patterns are merely epiphenomenal rather than reflecting genuine structural representations, the core theory is threatened, especially when algorithms used may be less sophisticated than those proposed by psychological models.

We acknowledge these imperfections, but emphasize that RSA-type analyses represent a significant advancement by offering a psychological model to complement statistical predictions. Unlike classical MVPA, which merely deployed classifiers without positing internal relations between representational geometries, RSA's strength lies in its use of semantic categories that structure representational space through semantic and taxonomic properties.

While we agree with critics like Ritchie *et al.* (2019, 2021) and Bowers *et al.* on the need for refined protocols and more stringent tests, our framework actually imposes additional constraints on internal informational relations. This aligns with psychological models' job of making the conditions for true explanation more restrictive than mere statistical matching. Likewise, positing the existence of semantic, not arbitrary coding channels raises the bar for genuine informativeness for a cognitive system. This positions representational mechanisms grounded in content-endowing structural similarity as candidates better suited for understanding cognitive processing and its alignment with the world's structure.

> *Epiphenomenal Correlations:* Critics argue that patterns uncovered by RSA-type methods may be epiphenomenal. Current categorizations, limited to simple relations and a narrow set of labels, fall short of representing the true complexity of cognitive systems. While these instances serve as proof-of-concept, they may not accurately reflect the intricate information processing within the brain.

The charge of epiphenomenalism against RSA-based results is, at best, premature. These findings, far from being illusory, represent the first tentative steps in charting the cognitive landscape. True, our current maps are crude, akin to early seafarers' charts with their sea monsters and unexplored territories. But this crudeness does not negate their value or potential accuracy.

The causal efficacy of these representational structures is not a mere philosophical fancy, but an empirical reality grounded in the study of evaluation mechanisms. Research into prediction error and metacognition has begun to il-

luminate the causal relevance of these vehicles of content (for more detail, see Bielecka & Miłkowski, 2020; Buckner, 2022). The consistent emergence of semantically organized structures is not a cosmic coincidence, but a testament to the causal potency of content in cognitive mechanisms. As an intuition pump, it may be instructive to think of it as a corollary of Putnam's (1979) No Miracles Argument for scientific realism, just for structural representationalism this time.

> *No novelty*: Critics might argue that our Correspondence Network Framework merely applies existing ideas from structural representation theories (e.g., Shea, 2007) with different terminology.

However, this objection misses several crucial innovations. First, while Shea and others acknowledge the importance of type-token relations, they do not systematically address how multiple classification schemes operate concurrently over the same vehicles. Second, our framework uniquely explains how cross-channel evaluation enables content-sensitive processing through network-level operations. Third, we provide a formal account using infocorrespondence to explain taxonomic organization in neural representational spaces. Fourth, our framework directly addresses anti-representationalist critiques by showing how content becomes causally efficacious through network architectures. These features collectively constitute a novel theoretical synthesis that extends beyond previous accounts.

7. Conclusion

This paper has introduced the Correspondence Network Framework, integrating correspondence-based semantic information and representational mechanisms in a novel theoretical synthesis. While structural similarity has been recognized as important to representation (e.g., Shea, 2007; Gładziejewski, 2015a; Shea, 2018), our framework uniquely explains how networks of correspondences across multiple channels enable representational mechanisms to systematically process content-sensitive information.

Our framework extends beyond previous accounts in several ways: it explains how neural vehicles can simultaneously participate in multiple classification schemes; it provides a principled explanation for why neural representational spaces consistently organize along semantic dimensions and as such avoids troubles for traditional explanations based on arbitrary coding schemes. Additionally, it directly addresses anti-representationalist critiques by showing how and why content must become causally efficacious (and indeed does, as documented by empirical studies) through network-level operations that detect cross-channel consistency.

Examining RSA through our framework revealed how this methodology implicitly relies on taxonomic organization emerging from correspondence networks. This demonstration highlighted both the utility of our approach for understanding representation in practice and its potential for guiding future empirical research. By offering a formal account of how multiple classification systems can operate over the same vehicles, our framework provides a powerful experimental heuristic for studying cognitive mechanisms.

Our account explains the non-arbitrary binding of targets to contents through correspondence networks that preserve taxonomic structure across processing channels. By addressing both how representational mechanisms process information and why researchers' inferences about these systems are informative, the Correspondence Network Framework brings coherence to disparate notions of semantic information by revealing their grounding in networks of structural similarity.

Overall, this approach provides a versatile, pluralistic framework for conceptualizing representational mechanisms across the cognitive sciences bridging philosophical formalism with empirical techniques, offering new analytic resources for investigating the foundations of mental representation.

Acknowledgements

The authors wish to thank two anonymous reviewers of this journal for their extensive and helpful comments.

REFERENCES

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective, & Behavioral Neuroscience*, 19(6), 1319-1351. doi: 10.3758/s13415-019-00721-3

Barwise, J., & Seligman, J. (1997). *Information flow: The logic of distributed systems*. Cambridge University Press.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge.

Bielecka, K., & Miłkowski, M. (2020). Error detection and representational mechanisms. In J. Smortchkova, K. Dołęga, & T. Schlicht (Eds.), *What are mental representations?* (pp. 287-313). Oxford University Press.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1-74. doi: 10.1017/S0140525X22002813

Buckner, C. (2022). A forward-looking theory of content. *Ergo: An Open Access Journal of Philosophy*, *8*, 37. doi: 10.3998/ergo.2238

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181-204. doi: 10.1017/S0140525X12000477

Craver, C. F. (2007a). Constitutive explanatory relevance. *Journal of Philosophical Research*, *32*, 3-20. doi: 10.5840/jpr20073241

Craver, C. F. (2007b). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Decock, L., & Douven, I. (2010). Similarity after Goodman. *Review of Philosophy and Psychology*, *2*(1), 61-75. doi: 10.1007/s13164-010-0035-y

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, 88(1), 27-51.

Facchin, M. (2021). Structural representations do not meet the job description challenge. *Synthese*, *199*(3-4), 5479-5508. doi: 10.1007/s11229-021-03032-8

Facchin, M. (2023). Neural representations unobserved-or: A dilemma for the cognitive neuroscience revolution. *Synthese*, *203* (1), 7. doi: 10.1007/s11229-023-04418-6

Facchin, M. (2024). Maps, simulations, spaces and dynamics: On distinguishing types of structural representations. *Erkenntnis*. doi: 10.1007/s10670-024-00831-6

Fresco, N., & Miłkowski, M. (2021). Mechanistic computational individuation without biting the bullet. *The British Journal for the Philosophy of Science*, 72 (2), 431-438. doi: 10.1093/bjps/axz005

Gładziejewski, P. (2015a). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric*, 40(1), 63-90. doi: 10.1515/slgr-2015-0004

Gładziejewski, P. (2015b). *Wyjaśnianie za pomocą reprezentacji mentalnych: Perspektywa mechanistyczna*. Fundacja na rzecz Nauki Polskiej.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559-582. doi: 10.1007/s11229-015-0762-9

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & Philosophy*, 32(3), 337-355. doi: 10.1007/s10539-017-9562-6

Hacohen, O. (2022). The problem with appealing to history in defining neural representations. *European Journal for Philosophy of Science*, 12(3), 45. doi: 10.1007/s13194-022-00473-x

Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., ... Rapoport, S. I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, 88(5), 1621-1625. doi: 10.1073/pnas.88.5.1621

Isaac, A. M. C. (2019). The semantics latent in shannon information. *The British Journal for the Philosophy of Science*, *70*(1), 103-125. doi: 10.1093/bjps/axx029

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, 83, 201-226. doi: 10.1016/j.neuropsychologia.2015.10.023

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6), 4296-4309. doi: 10.1152/jn.00024.2007

Kohár, M. (2023). *Neural machines: A defense of non-representationalism in cognitive neuroscience*. Springer.

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, 38(4), 649-662. doi: 10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10), 3863-3868. doi: 10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8), 401-412. doi: 10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis —Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. doi: 10.3389/neuro.06.004.2008

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1-25. doi: 10.1086/392759

Mann, S. F. (2020). Consequences of a functional account of information. *Review of Philosophy and Psychology*, 11(3), 669-687. doi: 10.1007/s13164-018-0413-4

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474), 78-84. doi: 10.1038/nature12742

Miłkowski, M. (2013). *Explaining the computational mind*. The MIT Press.

Miłkowski, M. (2015). Satisfaction conditions in anticipatory mechanisms. *Biology & Philosophy*, 30(5), 709-728. doi: 10.1007/s10539-015-9481-3

Miłkowski, M. (2016). Function and causal relevance of content. *New Ideas in Psychology*, 40, 94-102. doi: 10.1016/j.newideapsych.2014.12.003

Miłkowski, M. (2017a). Szaleństwo, a nie metoda. Uwagi o książce Pawła Gładziejewskiego "Wyjaśnianie za pomocą reprezentacji mentalnych." *Filozofia Nauki*, 25(3 (99)), 57-67.

Miłkowski, M. (2017b). The false dichotomy between causal realization and semantic computation. *Hybris*, (38), 1-21.

Miłkowski, M. (2023). Correspondence theory of semantic information. *The British Journal for the Philosophy of Science*, 74(2), 485-510. doi: 10.1086/714804

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. The MIT Press.

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy*, 86(6), 281. doi: 10.2307/2027123

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action*. Oxford University Press.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI —an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101-109. doi: 10.1093/scan/nsn044

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128. doi: 10.3389/fpsyg.2013.00128

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.

Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics*, 16, 846979. doi: 10.3389/fnbot.2022.846979

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3), 283-311. doi: 10.1007/s11229-011-9898-4

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 25(2), 100-110. doi: 10.1016/j.tics.2020.11.006

Plebe, A., & De La Cruz, V. M. (2016). *Neurosemantics*. Springer International Publishing. http://link.springer.com/10.1007/978-3-319-28552-8

Rathkopf, C., Heinrichs, J. H., & Heinrichs, B. (2023). Can we read minds by imaging brains? *Philosophical Psychology*, 36(2), 221-246. doi: 10.1080/09515089.2022.2041590

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, 70(2), 581-607. doi: 10.1093/bjps/axx023

Ritchie, J. B., Lee Masson, H., Bracci, S., & Op de Beeck, H. P. (2021). The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity. *NeuroImage*, 245, 118686. doi: 10.1016/j.neuroimage.2021.118686

Roskies, A. L. (2021). Representational similarity analysis in neuroimaging: Proxy vehicles and provisional representations. *Synthese*, 199(3-4), 5917-5935. doi: 10.1007/s11229-021-03052-4

Shapiro, L. A. (1997). Junk representations. *The British Journal for the Philosophy of Science*, 48(3), 345-361. doi: 10.1093/bjps/48.3.345

Shea, N. (2007). Content and Its vehicles in connectionist systems. *Mind & Language*, 22(3), 246-269. doi: 10.1111/j.1468-0017.2007.00308.x

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Thomson, E., & Piccinini, G. (2018). Neural representations observed. *Minds and Machines*, 28(1), 191-235. doi: 10.1007/s11023-018-9459-4

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, *28*(10), 3505-3520. doi: 10.1093/cercor/bhx216

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.

Villalobos, M., & Dewhurst, J. (2017). Why post-cognitivism does not (necessarily) entail anti-computationalism. *Adaptive Behavior*, 25(3), 117-128. doi: 10.1177/1059712317710496

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353. doi: 10.1016/S0019-9958(65)90241-X

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2), 265-288.

Wojciech Mamak is a PhD student at the Graduate School of Social Research. His PhD thesis focuses on the explanatory uses of information in cognitive (neuro)science.

Address: Graduate School for Social Research, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland. E-mail: wojciech.mamak@gmail.com – ORCID: 0009-0004-3657-1128

Marcin Miłkowski is Associate Professor and Chair of the Section for Logic and Cognitive Science at the Institute of Philosophy and Sociology, Polish Academy of Sciences. He works on the philosophy of cognitive (neuro)science, digital philosophy of science, computational explanation, and mental representation.

Address: Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland. E-mail: marcin.milkowski@ifispan.edu.pl – ORCID: 0000-0001-7646-5742