IMT School for Advanced Studies, Lucca

Lucca, Italy

Nuances of Action:

Understanding action categories and their brain representation through naturalistic fMRI

PhD Program in CCSN

XXXVI Cycle

By

Lorenzo Teresi

2025

The dissertation of Teresi Lorenzo is approved.

PhD Program Coordinator: Prof. Emiliano Ricciardi, IMT School for advanced Studies Lucca

Advisor: Prof. Emiliano Ricciardi

The dissertation of Teresi Lorenzo has been reviewed by:

Prof. Corrado Sinigaglia, University of Milan

Prof. Maria Ida Gobbini, University of Bologna

IMT School for Advanced Studies, Lucca

2025

Contents

Acknowledgements **vii**

Vita **viii**

Publications **ix**

Presentations **x**

Abstract **xi**

Abbreviations **xii**

List of Figures **xiv**

List of Tables **xv**

Introduction **1**

Chapter 1

1. The Search for the Mark of Action **3**

1.1. Is Three the Magic Number? The Basic Principles of Actions and Action Taxonomies **5**

1.1.2 The Hierarchical Structure of Actions **5**

1.2 The Role of Outcome in Action Individuation **6**

1.3 Multiplicity of Goal Representations **7**

1.4 Goal Representation in Action Observation and Control **8**

1.5 Action Taxonomies: An Introduction **9**

1.5.1 Several types of taxonomies **11**

1.6 Conclusion **18**

Chapter 2

2. Introduction: Defining Actions Across Disciplines **21**
2.1 Levels of Representation: A Practical Illustration **22**
2.2 Representational Formats **24**
2.3 Representational Formats and Levels of Description **25**
2.4 On the Interface Problem **26**
2.5 Map of Taxonomies **27**
2.5.1 Sensorimotor Format **29**
2.5.2 Perceptual Format **33**
2.5.3 Semantic Format **34**
2.5.4 Integrating the Levels: The Interface Problem **35**
2.6 Conclusion **37**

Chapter 3

3. Introduction **38**
3.1 Aims and Purposes of a Taxonomy **42**
3.2 Key Ideas of Taxonomic Practice from Biological Systematics **45**
3.3 On the Demarcation of Species **46**
3.4 On the Ontological Basis of Biological Taxonomy **48**
3.4.1 On the Levels at Which Natural Selection Operates

49
3.5 Taxonomic Monism or Pluralism **51**
3.5.1 On the Relationship Between Description and Theory **52**
3.6 The Role of Theory in Taxonomy **53**
3.7 Applying Biological Taxonomic Principles to Actions **54**
3.7.1 On the Demarcation of Actions **55**

3.8 Ontological Basis of an Action Taxonomy **56**
3.9 Taxonomic Monism or Pluralism in Actions **57**
3.10 Biological Taxonomic Principles and the Action Domain **57**
3.11 Organizing Taxonomic Attributes of Action (TAA): A Tentative Proposal **58**
3.12 Conclusion **64**

Chapter 4

4. Heuristic and Organizing Principles of Action Taxonomies **66**
4.1 Heuristic Principles for Action Taxonomies **68**
4.2 Methodological Considerations **73**
4.3 Granularity **75**
4.4 Interactivity **76**
4.5 Observation-Production **77**
4.6 Hierarchy **78**

4.7 Abstraction **79**
4.8 Methodological Perspective on Heuristic Principles **79**
4.9 Conclusion **82**

Chapter 5

5. Introduction **85**
5.1 The Current Study: Investigating Action Recognition during Film Viewing **88**
5.2 Methods **89**
5.2.1 fMRI Datasets **89**
5.2.2 Acquisition and Preprocessing of fMRI Data **90**
5.2.3 Analysis and Annotation of the Films **93**
5.3 Model Description and Construction **94**
5.4 Voxel-wise Encoding Modeling **101**
5.5 Results **104**
5.6 Discussion **111**
5.7 Conclusions **114**

Chapter 6

6. Introduction **116**
6.1 Methods **119**
6.2 Results **150**

6.3 Discussion **166**

6.4 Conclusions **174**

Conclusions **176**
Appendix **180**
Bibliography **186**

Acknowledgements

One of the studies collected in the dissertation are based on co-authored papers in progress.

Chapter 3 is a pre-print paper under submission:

Marras, L.*, Teresi L.*, Simonelli, F.*, Setti, F., Ingenito, A., Handjaras, G., Ricciardi, E. (2025). "Neural representation of action features across sensory modalities: a multimodal fMRI study." (submitted to Special Issue of Neuroimage)

I acknowledge the use of artificial intelligence in refining and restructuring this thesis, while ensuring that no content was generated anew. The AI served as a tool for enhancement, never as a substitute for original thought and scholarly rigor.

First and foremost, I extend my deepest gratitude to my advisor, Emiliano Ricciardi, whose guidance, patience, and insight have been invaluable throughout this journey. His mentorship has been instrumental in shaping my research and academic growth.

To my family—my father, my mother, Luca, and my sister Viola—thank you for your unwavering support, love, and encouragement. Your belief in me has been a constant source of strength.

A heartfelt thanks to the entire IMT community, whose stimulating intellectual environment and collaborative spirit have greatly enriched my experience. Special recognition goes to Federica Ruzzante, a true companion in this PhD journey, and to Roberto, for his friendship and insightful discussions. I am also grateful to Gesine and Na for their camaraderie and encouragement along the way.

This work is the result of not only personal effort but the collective support of a brilliant and inspiring network of individuals. To all who have contributed, in ways big and small, I am profoundly thankful.

Vita

July 9, 1994	Born, Monza (MB)
September 2020	Degree in Philosophy Final mark: 110/110 cum laude Università Statale di Milano
December 2022 - July 2023	Traineeship in Nijmegen, involved in project on contextual linguistic prediction (predictive coding) in auditory deprived and typically developing subjects in the laboratory of Floris de Lange

Publications

Presentations

Noto, Sicily, “For a Systematics of Action Taxonomies”, 13 December 2021, AISC 2021

Trento, Italy, “Investigating social projection: a computational approach”, 15 December 2022, AISC 2022

Montreal, Canada, “Towards a comprehensive neural mapping of action dimensions in naturalistic stimuli”, Virtual poster presentation, 22 July 2023 - 26 July 2023

Abstract

The nature of human action remains a debated issue, with no definitive resolution. This problem can be approached from two perspectives. The first is theoretical: refining our understanding of what constitutes an action and developing a taxonomy that effectively captures its essential features can provide a more structured framework for investigating its neurobiological basis. The second is empirical: to fully grasp the complexity of human action, research must incorporate actions occurring in ecologically valid settings, such as naturalistic stimuli (e.g., movies).

This thesis addresses both dimensions. On the theoretical side, I propose new high-level models for interpreting action, beginning with fundamental definitions and examining the minimal conditions required for an effective taxonomy. I highlight the inherent limitations and unavoidable incompleteness in any attempt to comprehensively map the action phenomenon. Additionally, I review existing empirical approaches in neuroscience that have sought to identify the key features representing actions in the brain.

On the empirical side, I employ functional magnetic resonance imaging (fMRI) with naturalistic stimuli to investigate how actions are encoded in the human brain. Using a theoretically grounded action model, I examine the stability of action representations across different movies, demonstrating that action coding remains consistent despite variations in narrative and sensory complexity. I then conduct a detailed analysis of a single movie (*101 Dalmatians*), exploring how model fit varies across different sensory modalities. The findings reveal that certain action features have a stable impact on brain activity regardless of the modality through which they are perceived. By integrating theoretical modeling with empirical validation in naturalistic settings, this work advances our understanding of action representation in the brain and underscores the importance of studying actions in contexts that closely mirror real-world experiences.

Abbreviations

TAA: Taxonomic Attributes of Action

fMRI: Functional magnetic resonance imaging

IFG: Inferior frontal gyrus

IPL: Intraparietal lobule

STS: Superior temporal sulcus

MT: Middle temporal

vPMC: Ventral premotor cortex

dPMC: Dorsal premotor cortex

TPJ: Temporo-parietal junction

AON: Action observation network

ISC: Inter-subject correlation

BOLD: Blood oxygenation level-dependent

TR: Repetition time

TE: Echo time

FOV: Field-of-view

GIST: Global Image Similarity Transform

PCA: Principal component analysis

LOOCV: Leave-one-fold-out cross-validation

FDR: False discovery rate

FWER: Family-wise error rate

SNR: Signal-to-noise ratio

VPA: Variance Partitioning Analysis

CCA: Canonical Correlation Analysis

CKA: Centered Kernel Alignment

List of Figures

Fig. 1. *Conceptual overview of the basic principles of actions*

Fig. 2. *Representational formats and their relationship to action representation*

Fig. 3. *Biological taxonomic principles guiding construction of an action taxonomy*

Fig. 4. *Features in guiding the construction of an action taxonomy*

Fig. 5. *Trade-offs in action taxonomy construction*

Fig. 6. *Analysis steps of the project*

Fig. 7. *List of high-level action features of the high-level categorical action model*

Fig. 8. *Original feature matrix of the semantic action model*

Fig. 9. *Average correlation matrix for the three movies*

Fig. 10. *Model performance in the three datasets for high-level semantic action model*

Fig. 11. *Model performance in the three datasets for low-level visual model*

Fig. 12. *Model performance in the three datasets for high-level categorical action model*

Fig. 13. *Taxonomic Action Model*

Fig. 14. *Brain encoding of action features in the audiovisual modality*

Fig. 15. *Impact of sensory modality on action representation in the brain*

Fig. 16. *Spatial domain specificity of the different action models*

List of Tables

Table 1: *Characterization of the features and domains of the taxonomic action model*

Table 2: *Spearman Correlations between sensory modalities*

Introduction

The concept of "action" has been studied from multiple perspectives, including philosophy, cognitive neuroscience, psychology, and robotics. Each discipline has shaped the definition of action differently, leading to diverse theories and approaches. However, despite substantial contributions from these fields, there is still no unified framework that integrates these perspectives into a coherent understanding of actions. A key challenge lies in developing a taxonomy of actions that reflects the multi-level complexity and diversity of actions, from basic motor functions to high-level cognitive processes. This thesis aims to address these challenges by exploring the role of action representation and proposing an integrative framework that embraces a pluralistic approach.

The motivation behind this research lies in the lack of consensus in defining and categorizing actions, especially across different levels of description and representational formats. By taking a deflationary pluralism stance, this thesis aims to embrace different theoretical views without enforcing a single, restrictive definition of action. The proposed framework will consider multiple aspects of action representation: sensorimotor, perceptual, and semantic formats. By integrating insights from each of these domains, we can develop a richer and more nuanced understanding of action, contributing to the advancement of theoretical and empirical studies in the field. In this first chapter, we will list several ways of describing actions, discussing various taxonomies that, while not conclusive

individually, together offer different perspectives on the same problem.

This thesis will address key research questions, such as how to evaluate existing action taxonomies, the role of motor and goal representations, and the relationship between theoretical and practical approaches to actions. By building on a foundation of empirical evidence and theoretical analysis, this work aims to propose an adaptable framework for understanding actions, one that can enhance explanatory power and guide future research in cognitive neuroscience and psychology.

Chapter 1

1. The Search for the Mark of Action

The first aspect of the problem of action is understanding whether there is any distinctive “mark” (or collection of traits, or properties) of actions, distinguishing them from mere movements or behavioral motor reactions. It has acquired many meanings through diverse periods and disciplines. Under various standpoints, since ancient times, philosophers have searched for the distinctive property that differentiates bodily actions from unintentional movements or reflexes — what we will call the "mark" of action.

The standard causal theory of action identifies intentionality as the criterion distinguishing an action from something that merely happens to an agent. Other authors (among others, Frankfurt, 1978; Bach, 1978; Butterfill, 2001, 2012; Sinigaglia, 2013; Fridland, 2017; Pacherie, 2018) have argued that propositional attitude and explicit decision-making are not necessarily implied by purposive and skillful action. Evidence from, for example, how the cat's intentional behavior adapts to the environmental context suggests that intention is not necessary to define an action; rather, expertise in the guidance or control of action is a sufficient condition.

More recently, empirical and theoretical work by Jackson and Cross (2011), and Kaufman and Newen (2023) has

contested this view of intentions as necessarily requiring complex language competencies. They argue that non-human animals, such as spiders or rats, possess complex planning abilities that support flexible, intentional behavior. These considerations challenge the view that complex action essentially involves intentions or even linguistic competencies.

These insights notwithstanding, there is still no clear evidence available to conclusively settle the controversy about the "mark" of action. For this reason, we would like to suggest a deflationary pluralism—representing a variety of perspectives along with their respective action taxonomies. By embracing such plurality, we will resist the temptation of prematurely ruling out various approaches and be able to accept relevant contributions from philosophy, cognitive neuroscience, psychology, robotics, and linguistics.

However, embracing pluralism does not mean embracing all indistinguishable action taxonomies. Hence, in the section below, I list the minimum conditions any plausible theory of action must satisfy if it is supposed to provide sufficient and necessary conditions for actions. We will then explain, focusing on different theories consonant with our pluralism, how such theories adopt experimental methodologies and levels of granularity in conceptualizing actions.

1.1 Is Three the Magic Number? The Basic Principles of Actions and Action Taxonomies

When limiting our discussion to bodily actions, we propose that any theory of action must account for several key features of how actions are instantiated. Specifically, every theory should satisfy three minimal explanatory conditions —the "three basic principles of action":

- Actions have hierarchical structures.
- Actions are distinguished by their outcomes (or goals).
- A single action can be described in multiple ways.

1.1.2 The Hierarchical Structure of Actions

An action should consist of a hierarchy of simpler actions, or sub-actions, whereas atomic actions cannot be broken down further and are analyzed only at the kinematic level (Grafton & Hamilton, 2007; Gallese & Sinigaglia, 2011). For instance, the action of "drinking from a cup of tea" involves perceiving and representing the cup (Jeannerod, 1994), reaching for it, grasping it (Rizzolatti et al., 2001), bringing it to the mouth, preparing the facial muscles, and finally ingesting the tea. These sub-actions occur in sequence but exist at the same level of complexity. Additionally, each sub-action is hierarchically superior to simpler ones, such as the coordinated movements of the fingers required to grasp the cup (Fitts & Posner, 1967; Gentilucci et al., 1997).

This structure reflects the means-goal relationship: the coordinated finger movements aim toward grasping the

cup as the goal (Rosenbaum et al., 2007; Newell, 1991). The concept of an action hierarchy also touches on the "granularity of action"—different theories may focus on different levels of granularity depending on their goals and underlying assumptions (Butterfill & Sinigaglia, 2014; Grafton, 2007). For example, a philosophical approach focusing on motor goal representations might adopt a broader-grained analysis, but it could still be quite detailed (Fridland, 2014). Ultimately, each theory's goals and commitments determine the analysis level and the methods used (Gallagher, 2005).

In this context, we aim to exclude theories that fail to meet our criteria for valid action theories. Any theory that does not go beyond the kinematic level of analysis—i.e., that only examines movement—fails to satisfy the first principle of action and is incompatible with our deflationary pluralism. As a result, such theories will be rejected from our pluralistic action taxonomy.

1.2 The Role of Outcome in Action Individuation

The notion of an "outcome" refers to an actual or possible state of affairs, such as "drinking a cup of tea." It distinguishes actions from mere movements or bodily reflexes and functions as the "action individuation principle." However, the concept of outcome is not without its limitations. The mere existence of an outcome or motor/cognitive representation does not automatically constitute an action since intentions may fail to produce corresponding movements or the desired outcome. For

instance, merely intending to "hit a tennis ball like Roger Federer" does not mean the action will succeed.

To effectively individuate actions, each theory of action should include a correctness condition that allows us to determine whether an action has been successfully instantiated. This involves differentiating between outcomes and end-states, as well as between token-outcomes (specific to individual actions) and type-outcomes (applicable to multiple actions). Not all outcomes are end-states, and not all result from actions. Instead, goal representations (i.e., representations of end-states) help identify correctness conditions.

By conceptualizing outcomes in representational terms, mainly focusing on goal representation, we understand their role in organizing and coordinating bodily movements to achieve an intended outcome, enabling us to uniquely individuate actions.

1.3 Multiplicity of Goal Representations

While each action requires a token-outcome representation, this does not mean a single action must have only one representation. For example, "hitting a forehand like Roger Federer" can also be described as "sending the ball to the opposite corner." This demonstrates that one action can have multiple token-outcome representations. Additionally, individuals may execute the same type of action while representing it differently, illustrating that one type-action can be characterized by various token-outcome representations. This multiplicity is consistent with our

explanatory pluralism, which allows each action theory to describe actions differently, assigning appropriate goal representations based on the context.

1.4 Goal Representation in Action Observation and Control

We propose that goal representation is sufficient for action individuation, questioning the necessity of intentionality as an explanatory element. The standard causal theory suggests that intentionality harmonizes mental states like desires and beliefs with bodily movements to achieve a goal. However, we argue that a causal account and second-order representations (e.g., beliefs and desires) may not be required to explain actions. Instead, a model based on temporal stability, supported by appropriate neural representations, could adequately explain both purposive and skillful actions.

In this framework, coordinated bodily movements are selectively guided by motor representations that organize sub-actions and movements toward an intended end-state (Gallese & Sinigaglia, 2011). The motor system coordinates actions through sensory and mnemonic feedback, activating goal-oriented motor plans (Rosenbaum et al., 2001). This approach eliminates the need for meta-representations, such as propositional attitudes (Pacherie, 2008), and emphasizes the role of motor goal representation as a sufficient reason for action (Jeannerod, 1997). Our pluralism, therefore, is essentially descriptive, aiming to provide a comprehensive mechanical and informational

explanation of actions without focusing on their ontological basis (Fridland, 2014).

Beyond its executive function, the motor system also contributes to action observation and understanding. The neural circuits involved in executing actions are similarly activated during action observation, indicating that the motor-neural circuit plays a role in action comprehension. Goal representation, therefore, serves not only as an action individuation principle but also as the essence of what Pacherie (2018) termed "motor intentionality."

In conclusion, our deflationary pluralistic approach to action taxonomies emphasizes the need for specialized taxonomies that accept goal representation as the principle of action individuation, hierarchical action structures as an essential condition, and the idea that actions can be described in multiple ways. This approach excludes mere bodily movements from being considered actions and supports explanatory pluralism by integrating insights from philosophy, cognitive neuroscience, psychology, robotics, and related fields to deepen our understanding of actions.

We will now move on to explore suitable action taxonomies.

1.5 Action Taxonomies: An Introduction

In the present section, we introduce the main approaches to studying different types of actions. Actions are multilevel entities, and taxonomies focused on low-level features,

such as shape and kinematics, differ from those addressing semantic and more abstract features, such as locomotion or communicative actions. Moreover, some features of actions are transversal across different taxonomies. For example, *temporal extension* is essential—researchers may focus on brief actions, such as grasping an object, or extended sequences of actions involving complex processes.

Another guiding feature is *granularity*, which is influenced by the chosen methodology: EEG studies often focus on the microstructure of temporal action, while fMRI studies examine longer time frames, often exploring brain activity localization corresponding to different action types and emphasizing the semantic representation of various actions in the brain. By integrating these approaches, researchers can learn about the temporal structure and localization of action comprehension.

Behavioral-level research reveals trade-offs based on research focus. For example, studying specific action components, such as grasping small versus large objects, requires a narrow focus, whereas studying action variability necessitates broader sampling, often without the same level of detail. Attention and emotional engagement heavily influence how actions are processed, both at kinematic and abstract levels (Alaerts et al., 2011; Castiello et al., 2010; Shahdloo et al., 2022; Wurm & Schubotz, 2012).

For instance, behavioral relevance can be estimated by studying the effects of removing contextual cues on action recognition. Gaze anticipation is a key indicator of motor primitives comprehension, such as grasping (Ambrosini et

al., 2013; Costantini et al., 2010). The following sections will explore how taxonomies are shaped by the chosen level and perspective, with a more detailed review of proposed taxonomies.

1.5.1 Several types of taxonomies

Action Concepts and Dimensions

Current taxonomies often conflate prediction and explanation, although these are distinct concepts. While prediction is practically useful, relying solely on prediction may not lead to a comprehensive understanding of brain functions. Theories must make predictions that are then empirically validated (Breiman, 2001; Yarkoni & Westfall, 2017). Predictive success helps distinguish between competing models, even if they are theoretically equivalent.

However, complex models like deep learning may not offer human-interpretable explanations. In neuroscience, spatial and temporal resolution constraints often limit the predictive value of models like fMRI-based encoding. Despite these limitations, models must still support predictions and offer interpretability, distinguishing between descriptive, mechanistic, and explanatory models (Dayan & Abbott, 2005). Insights from animal studies, such as those involving mirror neurons, remain fundamental to building these models.

Mirror neurons inherently embody a predictive framework: they fire upon observing an action as if the observer is

performing the action. The predictive and the embodied aspects of mirror neuron theory align well with the predictive coding account, conceptualizing action understanding as involving anticipation. Most researchers agree on the concept of motor primitives—the basic units of actions that form complex behaviors (Rizzolatti & Craighero, 2004; Rizzolatti & Fabbri-Destro, 2010). Mirror neurons help explain how we quickly understand others' actions based on our experiences and expertise (Becchio et al., 2012; Cavallo et al., 2016). Studies show that motor experience is superior to observation in predicting subsequent actions (Aglioti et al., 2008). However, research in this field has often focused on simpler motor acts like grasping, limiting the exploration of other action types. Identifying motor primitives and understanding how they combine into complex actions is key to advancing action taxonomy.

Action Concepts and Dimensions

Describing the brain representation of complex actions remains a challenging task. Tarhan and Konkle (2020) used a survey-based approach to model effector involvement and action targets, predicting voxel activation per action clip. Their data-driven approach used clustering to reveal distinct action categories, such as large-scale versus small-scale actions, and actions with social aspects. Wurm and Caramazza (2017), on the other hand, employed a theory-driven approach to demonstrate that brain regions are organized topographically along dimensions of sociality and transitivity, similar to how object categories are

represented in the brain. Their findings showed that sociality and transitivity are computed along a gradient in the lateral occipitotemporal cortex (LOTC), with sociality distinctions in the dorsal LOTC and transitivity distinctions in the ventral LOTC.

These studies illustrate that action dimensions are continuously represented in the brain, supporting both theory-driven and data-driven approaches to action categorization. Although Tarhan and Konkle's clustering approach differed from Wurm and Caramazza's hypothesis-based work, both contribute to our understanding of how the brain represents actions and the underlying semantic information.

The goal of these studies is to extract relevant dimensions from cortical maps representing different action categories, using techniques like multivariate pattern analysis (MVPA) and representational similarity analysis (RSA) (Haxby et al., 2014). These methods demonstrate that representations of both object and action categories are distributed across the cortex. By integrating dimensions from different sources—such as subject ratings and low-level features—researchers can create dissimilarity matrices that reveal correlations between voxel activation and semantic features. Similar methodologies have also been used to explore emotional processing.

Linking Prediction to Action Taxonomies

Recent studies reveal that predicting actions involves more than just movement—it also depends on conceptual

similarity within a high-dimensional semantic space. Thornton and Tamir (2019, 2021) demonstrated that the human brain organizes conceptual action knowledge to enable predictions, showing that similar actions are more likely to follow one another (e.g., walking and running). Their use of an automatic classifier on naturalistic movie stimuli, combined with subjective action similarity ratings, confirmed that conceptual similarity plays a key role in predicting action transitions. This suggests that motor and semantic systems contribute to our understanding of actions, with a trade-off between detailed motor simulation and broader semantic categorization depending on the task context.

Mirror neurons naturally embody a predictive mechanism, firing when observing actions and facilitating rapid understanding through motor simulation (Rizzolatti & Sinigaglia, 2008, 2016). This process is complemented by the semantic system, which provides a broader conceptual framework, particularly during passive observation, such as when watching a movie. The interplay between motor and semantic systems raises questions about how different types of understanding—motoric versus semantic—contribute to action recognition.

Action Prediction

In high-dimensional semantic space, the proximity of actions can predict transition probabilities between them. For instance, walking and running often occur close together temporally, whereas the sequence might matter more for actions like eating after grocery shopping.

Thornton and Tamir (2019) used an automatic classifier to label actions in a naturalistic setting (e.g., *Sherlock Holmes* episodes) and gathered behavioral similarity ratings for linguistic classes of actions. These ratings were then used to create a high-dimensional space where actions were positioned, and their proximity was found to predict their temporal occurrence in naturalistic stimuli. The ability to decode these predictions from fMRI cortical maps further supports the idea that the brain's organization of conceptual action knowledge is tuned to predicting actions in both active scenarios and passive viewing contexts.

Passive vs. Active Action Prediction

The findings raise the question of whether spontaneous action prediction significantly differs from passive observation. Evidence suggests that the organization of action knowledge inherently supports prediction, aligning with the brain's predictive coding model. Conceptual similarity, rather than movement kinematics or effector type, may serve as a stronger predictor of subsequent actions, even though the mirror neuron system (MNS) is thought to use these low-level features for rapid action simulation. This suggests that motor and semantic systems serve distinct roles: motor simulation is vital for distinguishing fine-grained motor acts, while the semantic system helps with broader, context-based categorization.

Mirror neuron theories posit that humans instantly recognize familiar actions, using internal motor simulation (Rizzolatti & Fabbri-Destro, 2010). In contrast, recognizing unfamiliar actions requires more time and effort (Cavallo et

al., 2016; Koul et al., 2019). The semantic system is likely more involved when broader categorization suffices, such as during movie-watching. This reveals a potential trade-off in action taxonomy, similar to the variance-bias trade-off in machine learning: focusing on granularity when describing specific actions may limit generalization, whereas more abstract descriptions lose sensory detail.

Attention

Attention plays a crucial role in modulating action comprehension by influencing the selectivity of semantic maps in the brain based on task relevance and context. Similarly, mirror neuron activity is affected by context, prior experience, and social relevance (Kilner et al., 2006; Campbell & Cunnington, 2017). These findings suggest that the interaction between attention and mirror neurons helps adapt action understanding to the specific situation.

Other findings point in the direction of a focusing of the brain on certain semantic categories, produced by giving semantic cues to the participants (Çukur et al., 2013)

Linguistic Taxonomies

Vendler's classic linguistic taxonomy of verbs identifies four main categories of action—activities, accomplishments, achievements, and states—based on their aspectual properties. While this linguistic categorization provides a useful framework, it may only partially represent how the brain distinguishes different types of actions. However, exposure to language from birth helps

shape conceptual boundaries, which may lead to alignment between language-based categories and cognitive representations (Cangelosi & Stramandinoli, 2018). Embodied cognition theories further emphasize the interplay between language and sensorimotor representations, with abstract concepts grounded in the body's sensorimotor experiences (Jamrozik et al., 2016; Martin, 2016).

Computer Vision and Robotics

In computer vision and robotics, efficient taxonomies for categorizing actions have also been developed, often incorporating insights from cognitive psychology and neuroscience. Actions and objects exhibit a many-to-one relationship, where similar actions can be performed on different objects, and different grasps can be employed depending on the action goal (Butterfill & Sinigaglia, 2014). These taxonomies are designed to quickly categorize observed actions, ensuring context-sensitive representation.

Abstraction in Action Taxonomies

Action taxonomies can encompass both abstract and concrete dimensions. For example, Tarhan and Konkle (2020) demonstrated that features like visibility and involvement correlate with the brain's action representations. Balancing granularity—whether focusing on specific kinematic details or higher-level conceptual features—remains challenging in developing comprehensive action taxonomies and building effective models to explore them through neuroimaging techniques.

1.6 Conclusion

Exploring action taxonomies reveals the complexity and multi-dimensionality of actions as multi-level entities. Integrating different methodologies, such as EEG and fMRI, may help in further bridging the gap between temporal dynamics and localization of brain activity related to actions. Predictive and explanatory models are essential for understanding how actions are organized in the brain, with predictive coding and mirror neuron theories highlighting the interplay between motor simulation and semantic categorization.

As evidenced by the studies discussed, the conceptual similarity between actions suggests that semantic cues play a significant role in predicting actions, complementing the fast motor simulations provided by the mirror neuron system. Attention processes and linguistic taxonomies further influence how actions are understood, indicating that action comprehension is influenced by both low-level sensory details and high-level conceptual understanding.

By adopting a pluralistic approach that incorporates goal representation, hierarchical action structure, and multiple descriptions of actions, we can better understand the intricate ways in which the brain represents and processes actions. This deflationary pluralism accommodates diverse action taxonomies and integrates insights from philosophy, cognitive neuroscience, psychology, robotics, and linguistics, offering a comprehensive framework for

understanding actions at different levels of granularity and abstraction.

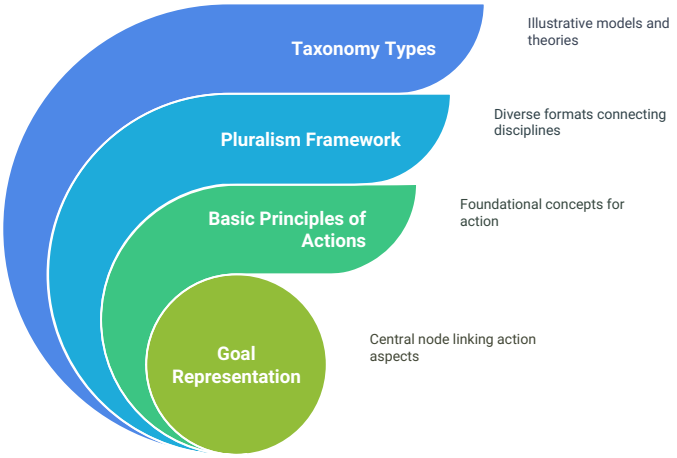

Fig. 1: A conceptual overview of the core components of action understanding, highlighting the three basic principles of actions, the central role of goal representation, and the pluralistic taxonomy framework integrating motor, perceptual, and semantic dimensions across multiple disciplines.

Chapter 2

2. Introduction: Defining Actions Across Disciplines

Multiple disciplines have long been interested in characterizing what constitutes an action. Philosophy of mind (Butterfill, 2014), computer vision (Efros, 2003), robotics (Zech et al., 2018), cognitive neuroscience (Thornton et al., 2019), and (neuro)linguistics (Humpreys et al., 2013) have all contributed to understanding actions from their unique perspectives. Each field's perspective is shaped by its specific needs—ranging from practical applicability to theoretical understanding—often without acknowledging the progress or insights from other disciplines. Consequently, actions remain a sparse and distributed topic in the scientific literature, with little consensus regarding their conceptual and neurobiological organization.

Even within cognitive neuroscience, researchers have examined actions from various perspectives and levels of analysis. While contributions are abundant, several theoretical aspects remain poorly defined, such as the absence of a coherent taxonomy of actions. Integrating recent contributions with earlier insights into a unified schema could prove helpful. Here, we attempt to systematize what is known about actions, organizing this knowledge through conceptual tools mainly derived from the philosophy of mind. Specifically, in the rest of this paper, we will apply the notions of levels of description and representational formats to map out most of the taxonomies of action.

2.1 Levels of Representation: A Practical Illustration

The concept of levels of representation (Land et al., 2013; Grafton et al., 2007) can be illustrated with a simple example—such as the action of "grasping a mug":

- At the basic motor level, grasping involves an effector (e.g., the hand) that closes around an object, using a specific kind of grasp from several possibilities (Giszter, 2015).
- The movements must happen in a specific sequence to achieve the desired outcome: for instance, one cannot grasp before reaching for the mug (Houghton et al., 1996).
- The action involves essential elements, such as an effector and an object.
- There must be an agent with an appropriate effector to perform the action (e.g., a bus cannot grasp mugs), *contextual relevance* and correct *affordances* (Gibson, 1977).
- Grasping has contextual meaning (Wurm et al., 2012) only in certain circumstances (e.g., "I grasped the mug and drank a sip of tea"), not in others (e.g., "I grasped a mug and called the police").
- Knowledge of a specific action can be obtained through direct execution (motor format), observation (perceptual format), or linguistic representation (e.g., "grasping a mug"). Each format conveys distinct information and is inherently partial (Shepherd, 2021)
- "Grasping a mug" encompasses a variety of possible specific actions, yet we know how to generalize from

> similar actions and recognize it when observing another person perform it, despite the infinite variations in its execution (Hu et al., 2023)

These examples demonstrate that action knowledge is stratified. Knowing how to perform "grasping a mug" differs from knowing that a specific instantiation of that action is one of many possible exemplars. Grasping a specific mug at a specific moment represents a token from the action type "grasping a mug." In other words, the particular action observed or performed is a specific instance of a general category.

In the philosophical literature, this distinction rests on what is often called: "knowing how" versus "knowing that" (Ryle, 1945; Ferretti & Caiani, 2023). Knowing how to attach a linguistic description to an action is distinct from knowing how to execute it. Performing an action requires a different granularity than language can provide, and potentially a different representational format, such as a procedural (metric-based) system versus a symbolic one. These foundational distinctions shape our conceptual approach to actions.

The central question then becomes: what do we consider an 'action'? If we extend this notion to include the concept of an action, or the semantics of an action, we encounter the challenge of how different action formats interact—a problem referred to as the "interface problem" (Burnston, 2017; Ferretti & Caiani, 2019)

We intuitively understand actions as means to achieve specific goals, and this informs our sense of levels: we can have a surface representation of an action, composed of concatenated motor acts, or we can access a higher-level representation that highlights the essential elements of action (e.g., agent and outcome). These various forms of action knowledge are represented differently, as explored in the next sections.

2.2 Representational Formats

The concept of 'representational format' is well-established in cognitive science and philosophy (Shepherd, 2018). It refers to the vehicle through which specific content is represented. The same content can be conveyed using different formats—for example, instructions to locate treasure can be described using an allocentric (map-based) or egocentric (personal viewpoint) format.

We distinguish action taxonomies based on their preferred representational format: semantic, perceptual, or motor. These tools help to identify whether two taxonomies describe the same phenomenon, the same set of actions, even if they use different formats or levels of description. Taxonomies at different levels describe actions with varying durations and representational formats. For instance, a taxonomy of grasping will necessarily exist at a different descriptive level than a taxonomy of everyday observed actions—they both describe actions but emphasize different aspects.

Within these levels of description, the representational format concept allows for useful distinctions among different types of actions. We argue that certain levels of description are better suited to specific representational formats, with common action features emerging from considering studies organized within our proposed framework.

2.3 Representational Formats and Levels of Description

To illustrate, consider the action "going to the cinema". This action can be analyzed at various levels of description:

- At a high-level, this action might be represented propositionally, describing the intention and planning involved.
- Breaking it down further, sub-actions are linked to the overarching goal, with sub-actions recursively breaking into smaller components. Ultimately, at the lowest level, actions are analyzed purely in terms of kinematic features (e.g., the actual body movements to get to the cinema). At this level, a motor format is most appropriate.

These representational formats are:

1. **Sensorimotor Format**: Represents precise bodily movements required for action, ensuring accuracy in muscle activation. For instance, grasping a specific mug requires precision in muscle and joint activation (Pouget et al., 2000).

2. **Perceptual Format**: Involves categorizing actions based on stimulus-dependent features (e.g., visual characteristics or movement). For example, effector recognition and movement features are perceptual formats (McDonough et al., 2020).
3. **Semantic Format**: Refers to abstract, modality-independent representations that describe actions in conceptual terms (e.g., transitivity or communicative content, e.g., Wurm et al., 2017).

Before delving into taxonomic proposals, it is important to recognize that streams of information are not entirely segregated by format. At some point, perceptual, motor, and semantic formats must interlock to facilitate comprehensive action comprehension (Burnston, 2017). Our classification emphasizes differences in representational formats, which is useful for organizing taxonomies and research on action dimensions.

2.4 On the Interface Problem

The "interface problem" (Burnston, 2017) arises when attempting to explain how different representational formats interact in the brain. High-level intentions in a semantic format ultimately lead to fine-grained motor actions in a sensorimotor format, but how this happens is still an open question. How do different representational formats communicate? Understanding these mechanisms could benefit both empirical research and theoretical developments in cognitive neuroscience.

The idea of representational formats is just one perspective for approaching action phenomena, serving as a conceptual tool that does not require prior assumptions. While some taxonomies may overlap or fit into multiple categories, we argue that simplicity and clarity outweigh accuracy in this context. Modeling, particularly in brain sciences, often involves drawing artificial boundaries around phenomena that are inherently interconnected. Actions do not exist in isolation; they constantly interact with other cognitive processes, such as emotional responses. However, for the purposes of organizing our understanding of actions, abstraction is useful.

The discussion will conclude by examining how coordinated investigation of empirical and theoretical approaches could further elucidate the neural correlates at each level of action description. We also address the role of representational formats and propose future steps for developing a more comprehensive and coherent taxonomy of actions.

2.5 Map of Taxonomies

To further clarify our framework, we distinguish among sensorimotor, perceptual, and semantic formats in terms of the relevant literature on action dimensions. This review does not explicitly separate action production from observation, as these processes share common cognitive and neural bases. The following three sections will present taxonomies that emphasize specific representational formats, using examples from the literature (Butterfill, 2014; Burnston, 2021; Quandt et al., 2017).

1. **Sensorimotor Format**: The format of the representation that ensures accurate bodily movements. For example, grasping requires motor representations that specify how muscles and joints should move to achieve a specific outcome.
2. **Perceptual Format**: Involves categorizing actions based on stimulus-dependent features. Actions are understood in terms of their sensory characteristics, often involving multiple sensory modalities (McDonough, 2020).
3. **Semantic Format**: Represents the conceptual aspects of actions, often using abstract, propositional content. Examples of semantic dimensions include transitivity, communicative content, and linguistic classifications (e.g., tool use or telicity, see Wurm et al., 2017).

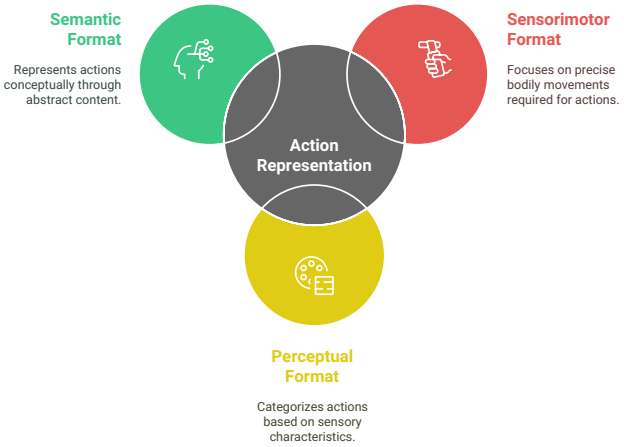

Fig. 2: Representational formats and their relationship to action representation. Action representation can be decomposed into different representational codes that together try to capture (always partially) the process that we call "action understanding". We summarize these codes as pertaining to different ways through which the brain manipulates information. We subdivide them in sensory (based on a code defined on the basis of the sensory channel), sensorimotor (code implied in the transformation needed to execute actions), and semantic (abstract or linguistic, possibly amodal information on action). The three codes are not to be viewed as mutually exclusive in the process of "action understanding".

Ultimately, a complete model of actions would combine models at all levels, incorporating all representational formats. Each level contributes a distinct perspective, helping to build a rich, multidimensional understanding of actions. This integrative perspective requires addressing the interactions between levels and the interdependencies among different formats.

2.5.1 Sensorimotor format

For over 30 years, computer vision and robotics have focused on recognizing and classifying actions (Cutkosky, 1989). Many taxonomies have recently been developed in these fields to label and categorize observed actions in the most efficient manner. These taxonomies draw inspiration from studies on action representation in cognitive neuroscience (Feix et al., 2016).

Robotics tends to focus on the performative aspect of human actions, while computer vision is primarily concerned with how humans recognize actions. Robotics is interested in understanding both recognition and execution (including mechanical effectors and conjoint action), whereas computer vision focuses purely on action recognition (Saif et al., 2018; Beddiar et al., 2020). Mirror neuron theories propose that sensorimotor experience allows humans to understand actions flexibly and adaptively. By contrast, applications like autonomous driving need a classification based solely on surface cues, highlighting the need for purely perceptual analysis.

Actions as Effector-Driven

Another approach models actions based on the effectors involved. Effectors are the body parts used to perform actions, such as the hand, arm, or foot. Studies show that different types of manipulation are limited, but there are numerous objects that can be manipulated (Liu, 2015). This highlights a relationship of many-to-one between objects and types of manipulation, such that the same grasp can be used for different objects, and different grasps can be used for the same object, depending on the aim of the action.

The most recent attempt to organize this complexity systematically comes from Liu and colleagues (2021) and Yang and colleagues (2021), who revised prior efforts and provided a new framework. This framework makes explicit the relationship between tasks, subtasks, and specific movements. It can be extended to other effectors, such as the mouth, which can also perform tasks like manipulation

(e.g., chewing), reaching (e.g., biting), or simple contact (e.g., kissing).

A detailed effector-based account of simple motor acts must account for the relative involvement of multiple body parts. To eat, for instance, involves the coordinated action of hands, mouth, and arms. Thus, an action involving multiple effectors can only be fully understood in terms of their relative contribution.

Actions During Development

Motor skills develop through a combination of environmental interaction and genetic factors. Macro-categories of action that emerge early in development could serve as the ontogenetic predecessors for the wide variety of adult actions. For example, studies suggest that toddlers are influenced by conventional actions associated with objects (e.g., "putting shoes on") before differentiating specific object categories (Hagihara, 2020).

Prenatally, infants engage in general movements (GMs)—movements that involve all body parts and vary in speed, amplitude, and direction (Hadders-Algra, 2004). These movements help infants explore the potential combinations of movement available to them, which serves as a basis for developing motor skills. After birth, primary variability is the process of exploring motor possibilities, while secondary variability refers to the ability to select effective movements in a given context.

The early motor repertoire includes reaching and manipulation abilities, postural control, and locomotor activities. These foundational movements form the basis of adult motor skills, which expand and refine based on tool use, cultural influences, and social interactions.

Actions and Lesions

Liepmann's classic classification of apraxia describes two main types: ideomotor apraxia and ideational apraxia. Ideational apraxia involves difficulty in sequencing motor acts for a goal, while ideomotor apraxia prevents the correct execution of actions even though patients possess the required motor and sensory abilities (Liepmann, 1908; Buxbaum et al., 2018). These patients struggle with imitation and tool use, although they may perform actions spontaneously. This highlights the difficulty of translating action semantics (propositional format) into an executable motor program (motor format).

Anosognosia for Hemiplegia (AHP) provides another example of dissociation between motor awareness and sensory output. In AHP, patients with a paralyzed side of the body believe they can move it and report performing tasks involving both hands (Garbarini et al., 2019). The disorder suggests that an impaired ability to compare predicted sensory consequences with actual sensory feedback could underlie this condition. The premotor cortex (PMC) appears crucial for monitoring whether intended actions are successfully executed.

These cases illustrate the distinction between sensory awareness of action and its implementation at the motor level. However, there is a symbiotic relationship between perceptual and motor formats. Without a mechanism for tracking action correctness, the concept of action loses coherence.

2.5.2 Perceptual Format

Actions as Spatio-Temporal Relations

Ziaeetabar et al. (2021) investigated actions as sequences of simple movements, emphasizing spatio-temporal relations between objects and limbs. They developed a matrix representation to capture the likelihood of specific relationships (e.g., touching or dynamic spatial relations) and classify actions accordingly. Actions were represented by clustering relations based on similarity. Although limited to hand actions and devoid of other cues like context and gaze, this approach uses perceptual features to classify actions at a granular level.

Actions as Dynamic Visual Features

In computer vision, human action recognition focuses on extracting information from visual stimuli, which may differ significantly in low-level features (Zhang, 2017). Handcrafted features, such as point-light displays or space-time interest points (Laptev, 2005), were initially employed to model the relevant features of actions. Later, deep learning approaches took over, applying supervised, unsupervised, and hybrid methods. The introduction of

deep learning marked a revolution in action recognition, using temporal convolution layers to capture action sequences.

2.5.3 Semantic Format

Actions as Conceptual Entities

Recent studies by Wurm and Caramazza (2021) highlighted the role of the ventral "what" stream in recognizing actions. This approach suggests that action recognition involves more than just the Action Observation Network (AON), and that conceptual understanding plays a role in understanding observed actions.

Tarhan and Konkle (2020) took a data-driven approach to identify dimensions relevant to action representation in the brain. Their research involved clustering voxels based on activation patterns when participants observed brief action clips. The resulting clusters represented different types of actions, emphasizing features such as sociality and effector involvement. These features represent a high-level categorization, in contrast to approaches that focus on low-level, perceptual components.

Actions as Predictive Dimensions

Thornton and Tamir (2019) sought to map the conceptual space of actions by predicting the likelihood of transitions between actions. Their model, built from participants' ratings of similarity between action clips, successfully predicted action transitions and aligned with fMRI data,

suggesting that conceptual action knowledge is structured to facilitate prediction. This aligns with the predictive coding account, in which conceptual similarity helps predict subsequent actions more effectively than cues like movement kinematics or effector type.

The interplay between motor and semantic systems raises questions about understanding different types of action. Predicting subtle differences in action execution requires motor simulation, while placing an action within a broader context relies on semantic understanding. This dual approach suggests that the motor and semantic systems serve different purposes, each contributing to action understanding in different ways.

Actions as Linguistic Concepts

Vendler (1957) introduced a taxonomy of verbs based on their aspectual qualities: activities, accomplishments, achievements, and states. These categories help clarify distinctions in verb types and have influenced subsequent linguistic taxonomies (Smith, 1997). However, language may not always align with the brain's representation of actions. Language provides a framework that may partially shape action categories, but it is ultimately just one tool among many (Mazzuca et al., 2018).

2.5.4 Integrating the Levels: The Interface Problem

The "interface problem" refers to the challenge of understanding how a high-level, propositionally structured intention (e.g., "I want to grasp this mug") results in specific

motor actions. This problem deals with how different representational formats communicate. How does an abstract intention propagate to the motor system, directing precise movement sequences? Although empirical evidence shows that such communication exists, we do not fully understand the mechanisms.

Recent studies have investigated how different representational formats interact in the brain. For example, Urgen (2019) described the selectivity of different brain regions toward various representational formats. Baldassano et al. (2017) proposed that different temporal receptive windows might underlie action processing, similar to how language understanding involves varying levels of temporal integration.

One potential approach is to identify regions of the brain whose representational format changes significantly during processing. For example, studies have shown that certain brain regions (e.g., areas near the visual cortex) are sensitive to both visual and language representations of the same concept (Popham et al., 2021). These regions could help us understand the mechanisms of representational format change and interaction.

Further investigation of the structural and functional connectivity of these regions, even combined with targeted interventions (e.g., TMS or tDCS), could provide insights into how different formats integrate during action representation. Studying these interactions could help us develop a more comprehensive understanding of how the

brain achieves the complex task of translating abstract intentions into concrete motor behaviors.

2.6 Conclusion

It is possible to advance the suggestion that neuroanatomically distinct taxonomic (similarity-based) and thematic (event-based) action-related systems exist. This is a diverse way to frame the same thing, the coexistence of two levels in the literature and the brain, action as a concept and action as a concrete exemplar both in production and in observation. A similar semantic system exists for objects and actions in the left posterior temporal lobe (Kalenine et al., 2016).

CHAPTER 3

3. Introduction

Classifications have long been fundamental to scientific inquiry, offering an orderly framework that enhances understanding and exploration of complex concepts and phenomena. Taxonomies, in particular, enable researchers to establish systematic relationships between diverse objects and processes, revealing both existing knowledge and potential gaps that warrant further exploration. The biological sciences have successfully employed taxonomies to categorize life, establish evolutionary relationships, and create a structure that enhances the interpretability of new findings. In a similar vein, a taxonomic approach within the action domain—focusing on human actions, motor skills, and cognitive mechanisms—offers the potential to advance scientific understanding and practical applications across a range of disciplines, including neuroscience, robotics, computer vision, and even social cognition.

The human nervous system, fundamentally structured to generate actions, operates under the imperative of survival and adaptation to natural environments (Catmur & Heyes, 2019). Given this imperative, actions play a pivotal role in the architecture of the nervous system, a principle observable even from childhood as infants learn to link actions with object nouns (Hagihara et al., 2020). However, despite the foundational significance of actions, no general, agreed-upon action taxonomy currently exists, and efforts

toward systematic classification have been fragmented across distinct disciplines. The current approaches to categorizing actions—such as the classifications used in robotics for manipulation tasks (Cutkosky, 1989)—remain domain-specific and are often limited by their targeted applications. A comprehensive taxonomy, incorporating insights from these various fields, could enable a holistic approach to action research by bridging gaps and promoting collaboration across disciplines like cognitive neuroscience, psychology, computer vision, and philosophy.

A coherent action taxonomy has several potential benefits. First and foremost, it provides a shared framework, allowing different research fields to understand and build upon one another's progress by providing a comprehensive picture of the current state of knowledge. Classifications generally accelerate research by highlighting gaps and areas that need exploration, thereby foreshadowing future research directions (Vegas et al., 2009). We believe that an action taxonomy, in particular, is crucial to evaluate and interpret new and existing contributions from a panoramic point of view.

The biological sciences have been a rich source of inspiration for constructing taxonomies, and their methods may be similarly applied to human actions. Biological systematics, which flourished since Linnaeus, offers a well-established model for organizing diversity and creating hierarchical systems that relate similar entities in meaningful ways (Mayr, 1969; Cracraft, 1978; Cracraft & Donoghue, 2004). Drawing parallels between action

research and biological taxonomy could help to solve common challenges and foster novel approaches. For instance, analogies between species and actions could illuminate the relationships between various action types—considering them as analogous to different "species" of action with shared features and notable differences. Moreover, the fundamental issues faced in taxonomy construction in biology, such as the role of granularity, the level of abstraction, and hierarchical relationships, could similarly benefit the construction of an action taxonomy.

In this chapter, we propose to explore how taxonomic approaches inspired by biological classification systems can be adapted to enhance the field of human actions. Specifically, we summarize how a taxonomic approach could contribute to action research in three key ways:

1. Structure: A taxonomy provides a formal and precise language for describing actions and their components within a defined relationship structure.

2. Theory: It offers the basis for a quasi-axiomatic framework, sometimes referred to as "taxonomic theory" (Pavlinov, 2020), which determines the attributes relevant for creating taxonomies. This opens a discussion on which set of attributes—here termed "taxonomic attributes" (TAs)—are of utmost importance.

3. Clarity: Taxonomies, being theory-laden, clarify the theoretical foundation underlying the classification,

helping researchers understand which aspects of action are emphasized.

In this context, we would like to clarify what an action taxonomy is not supposed to be: a) A taxonomy is not a theory in itself: While taxonomies provide a classification, they do not inherently offer an explanation. Theories are required to explain why specific relationships or similarities exist, thus giving meaning to the taxonomy. b) Taxonomic Attributes (TAs) depend on the underlying theory: The attributes used to create a taxonomy are theory-dependent and reflect particular perspectives or theoretical commitments. Classifications are never entirely transparent —they always entail theoretical assumptions. While theory may be embedded in the resulting classification, the classification process can attempt to remain objective.

No taxonomy can substitute for a theory on, for instance, the generation of specific actions at the brain level. However, a robust, well-conceived taxonomy could draw from the extensive data collected in the field of action research and, in doing so, propose new hypotheses or avenues of inquiry. Organizing information in an orderly way is never futile; it opens new combinations and highlights patterns that might otherwise be overlooked.

To further contextualize these ideas, we examine the biological systematic concepts that are translatable to a neurobiologically inspired taxonomy of actions. We also consider whether common challenges in the field of biological systematics have been confronted or solved, and how these lessons could inform action taxonomy.

This exploration will be guided by the following questions:

1. What are the aims and purposes of a taxonomy, and why is it useful in the context of actions?

2. What lessons from biological taxonomies can be translated to a neurobiologically informed taxonomy of actions?

3. Are there similar issues encountered in both action research and biological systematics that have been addressed in the biological domain?

By addressing these questions, this chapter aims to establish a foundation for a systematic taxonomy of human actions, borrowing concepts and methods from biological systematics to advance interdisciplinary understanding. A comprehensive action taxonomy is not only theoretically important but also highly practical, as it holds the potential to unify research, promote collaboration, and enhance applications in fields like robotics, neuroscience, and computer vision (Giese & Rizzolatti, 2015). Thus, through the exploration of biological analogies, we hope to lay the groundwork for an organized and collaborative approach to the science of human actions.

3.1 Aims and Purposes of Taxonomy

Taxonomic practice is grounded in the understanding that classification predicates are inherently theory-laden. All classifications are influenced by the theoretical context

within which it is conceived and ultimately developed (Scott-Ram, 1990).

We can distinguish between two primary attitudes toward taxonomic practice: descriptive and theoretical, as discussed in the philosophy of biology literature (Scott-Ram, 1990). According to this view, classification can be either theoretical or descriptive. However, in both approaches, the goal is to encapsulate the maximum available information in a taxonomy—essentially, "the more characters, the better."

In the descriptive approach, the primary goal is to minimize theoretical bias as much as possible. The idea is to exclude explanatory elements from classifications, resulting in a catalog or database in which new data are structured naturally, without metaphysical assumptions. This approach produces a convenient repository of all actions, organized with their respective features. By doing so, descriptive taxonomies focus on being as objective and theory-independent as possible.

Conversely, in the theoretical approach, the primary goal is to construct a classification system that carries as much theoretical information as possible. Here, the question is: how much theory can be embedded into the taxonomy itself? Ideally, the taxonomy should reflect an established theory, ensuring that newly discovered data are consistent with the framework provided by the theory. This approach emphasizes the interpretative richness that theories can contribute to taxonomic structures.

Importantly, both taxonomic perspectives agree that classifications must be informative—the relationships they entail among entities should provide meaningful insights. Furthermore, taxonomies should be accessible to practitioners. If a taxonomy is overly complex and difficult to use, it loses its utility. An implicit aim of any taxonomy is, therefore, to create "organized knowledge" that is easy to navigate and apply.

These two taxonomic approaches yield different outcomes and serve distinct purposes. Despite their differences, we can still derive some common principles (Vegas et al., 2009) that both approaches share:

1. Providing a Set of Unifying Constructs: Classifications serve as a common language that facilitates communication among researchers. For instance, the Linnaean nomenclature in biology remains widely used because it provides a shared reference system for communication and exchange of knowledge.

2. Identifying Interrelationships: Taxonomies help identify connections between entities. Mendeleyev's periodic table, for example, had a profound impact on our understanding of the atomic structure by illustrating relationships between different elements. Conversely, when little is known about a particular subject, such as certain types of bacteria, their classification becomes challenging.

3. Identifying Knowledge Gaps: Taxonomies can also highlight knowledge gaps. For example, gaps in the periodic table led to the search for new elements, and in

some cases, elements were even predicted before they were discovered.

3.2 Key Ideas of Taxonomic Practice from Biological Systematics

The defining feature of taxonomies lies in the relationships they establish among entities, such as species or action categories (Godfray, 2002). The capacity of a taxonomy to encode and inscribe these relationships—using nomenclature, hierarchical categories, and group formation —defines its effectiveness. The question is: which characteristics provide the most fundamental understanding of the relationships between different entities? For instance, which features are most informative in distinguishing between a cat and a dog?

A core aspect of taxonomies is their relational nature. Taxonomic categories are defined relationally, meaning that they depend on the relationships with all other entities within the classification. This is similar to the idea that "a species is what it is in relation to all other species." Without context, an entity's definition lacks meaning. For example, consider the statement: "Lorenzo is shorter than Luca but heavier than Emiliano." Without any absolute measurements, these descriptions rely on the relationships between the individuals, yet they remain informative and sufficient to distinguish among them.

Taxonomic attributes (TAs) are used to describe these characteristics. TAs that connect different living forms serve to group species and are influenced by underlying

theories. For instance, phenetics groups organisms based purely on their overall similarity (Mayr, 1969). Evolutionary cladistics (Cassis et al., 2010) and phylogenetic systematics (Lieberman, 2011) use homologies to infer ancestral lineages and establish relationships based on evolutionary history.

By understanding the relational and theory-laden nature of taxonomies, we can appreciate their importance in scientific inquiry and the organization of knowledge. A robust taxonomy in the field of human action research, inspired by biological principles, could significantly contribute to our understanding of actions and their neural bases, providing an organized framework that enhances interdisciplinary communication and collaboration.

3.3 On the Demarcation of Species

Since the 1960s, biological taxonomists have recognized that the classical use of categories is inadequate for biological taxonomy, particularly for the demarcation of species (Farris, 1967; Ereshefsky, 2007). According to classical logic (Jansen, 2007), for an object A to be part of a class B, it must possess a precise set of features that all entities in that class have. However, this approach is not suitable for defining biological categories such as Reptilia, Mammalia, and Teleostei (Farris, 1967). There are no definitive sets of diagnostic criteria that can uniquely demarcate a single species. Even at the species level, it is impossible to establish principles that uniquely identify a species, as new instances are continually discovered,

requiring an enlargement or accommodation of the category.

As a result, the concept of a "species" is, in principle, considered a spatio-temporal entity—a real, dynamic category (Richards, 2010; Wilkins, 2014). In practice, taxonomists must rely on examining the totality of specimens belonging to a species. Moreover, the ongoing debate on biological essentialism (Devitt, 2008) shows that while demarcation is sometimes convenient, in some taxa there is less variability between species compared to others.

One way to address this problem is to avoid using Aristotelian logic for species classification. The notion of monotypic and polytypic concepts (Beckner, 1959) is derived from Wittgenstein's idea of "family resemblances" (Wittgenstein, 1953). In this context, members of a class do not need to possess all of the features shared by the group as a whole. A polytypic concept functions like family resemblances: each group member shares most—but not necessarily all—of the taxonomic attributes (TAs) present in the group. Consequently, the boundaries between species are fluid because the concept of species itself is inherently fluid.

Another approach to overcoming the demarcation problem is to apply fuzzy logic to species classification. Fuzzy logic is a recent and influential branch of logic that has proven useful for studying within-species variations and provides a more biologically rooted definition of species boundaries (Xu, 2011; Pappas, 2006). Unlike classical binary logic, fuzzy logic allows for degrees of truth between zero and

one. For instance, consider the statement, "A is an adult." If "A" is an infant, the truth value may be close to 0.1, whereas for an 18-year-old, it could be 0.6, and for a 35-year-old, the value may approach 1. Belonging to a specific class, therefore, becomes a matter of degrees, which is a more biologically plausible conclusion. Fuzzy logic is also useful for defining when a new species emerges, allowing for a graded transition rather than a sudden shift from one species to another.

3.4 On the Ontological Basis of Biological Taxonomy

The ontological question in biological taxonomy concerns the basic unit of the taxonomy: "What constitutes the fundamental unit of analysis?" In biological systematics, the species level is often considered the primary unit of classification. However, this is not as straightforward as it may appear. There is still uncertainty regarding what makes an individual member of a species belong to that group. Devitt (2008) argues that the philosophy of biology often confuses the question of species demarcation with the question of what makes an individual dog a member of the species Canis familiaris. Some features must be considered essential, albeit with a degree of fuzziness, for an individual to belong to a given species.

Moreover, the starting point for constructing a biological taxonomy may not be the species themselves but rather the individual specimens. In biosystematics, geographic races may be more fundamental taxonomic units than species (Pavlinov, 2020; Mayr, 1965). Even geographical races or populations are abstractions, although less so than the

concept of a "Linnaean species," as they are closer to the individuals that comprise a given species.

From this perspective, taxonomy construction can be seen as a bottom-up process, beginning with individual organisms and progressing to the taxonomic units we recognize today. However, categories and relationships are frequently established top-down. Categories and groupings are often postulated to order taxa or are based on intuitive understandings, such as the differentiation between Animalia and Vegetalia—a distinction that remains common in biological discussions, even though the number of biological kingdoms is still debated.

In conclusion, it is crucial to establish a clear ontological basis for taxonomic analysis. This ensures that taxonomies can be evaluated using consistent and comparable parameters. Although this level of decision-making might appear simple, it has profound implications. It cascades through the classification, influencing the relationships within the taxonomy and determining the level of granularity in the final classification. Therefore, determining the ontological level of a taxonomy fundamentally determines the objects that the taxonomy will classify, as well as the extent to which biological "reality" is included in the classification process to achieve the desired objectives.

3.4.1 On the Levels at Which Natural Selection Operates

Evolutionary thought is a prevalent concept in biology, and understanding the scale at which evolutionary mechanisms

function is crucial for systematics. There is ongoing discussion regarding which level of biological organization natural selection primarily acts upon: the micro-scale (genes, molecules, individual organisms) or the macro-scale (populations, species, groups). The evidence suggests that evolution acts at multiple levels, involving both micro and macro processes (Okasha, 2002; Okasha, 2006).

It is, therefore, not sufficient to say that evolution selects only the genotype or phenotype. Rather, both the genotype and phenotype are influenced by selection, alongside population-level selection. These various levels, hierarchically organized from the micro to the macro, may all function as vectors of selection in a coordinated manner. This suggests interactions between lower-level biological units, such as molecules or genes, and higher-level constructs, such as entire populations, each influencing the evolutionary trajectory of the other.

The idea of biological complexity has been discussed in the literature in terms of contextuality and radical openness (Chu, 2011; Fomin et al., 2021). Contextuality involves recognizing that a model of a biological system inevitably leaves out some aspects of the system, as it is impossible to capture all factors simultaneously. Radical openness refers to the fact that there are no closed, natural systems; a modeler must artificially draw boundaries and categories to make a system comprehensible. In this way, biology—particularly evolutionary biology—requires an understanding of the limitations inherent in any classification scheme and the influences that operate at multiple levels.

3.5 Taxonomic Monism or Pluralism

The discussion regarding whether taxonomic research should aim for a single unified classification (monism) or accept the existence of multiple taxonomies (pluralism) has been ongoing for years. This debate arises partly because biological systematics has no singular classification framework, and many taxonomists argue that different taxonomies may reflect distinct perspectives on biological diversity (Pavlinov, 2020). Thus, a macro-taxonomy might be a long-term goal for the field, one that could help bridge the gap between disparate approaches.

A macro-taxonomy can be considered a general theory that encompasses and explains particular taxonomies, elucidating overlaps and differences among them. Such a framework could be instrumental in reducing inconsistencies between taxonomies and organizing systematic work more effectively. It would not only show the relationships between different taxonomies but also provide insights into how these taxonomies complement each other and in what ways they diverge.

Currently, biological systematics is dominated by the phylogenetic approach, which is focused on the evolutionary relationships between species. Alongside phylogenetics, there are other classification methods, including phenetics, numerical taxonomy, typological taxonomy, and biosystematics. These approaches have different emphases and methodologies, and they continue to influence the phylogenetic program in various ways. For example, the notion of a prototypical developmental plan,

which originated in typology, has influenced modern phylogenetics.

3.5.1 On the Relationship Between Description and Theory

With the exception of purely descriptive classification methods that avoid theoretical explanations, most taxonomies incorporate theory into their formation. In evolutionary systematics, data—primarily morphological and genetic—are compared, and the theory (of evolution) provides a set of principles that explain the similarities and differences observed in the comparative analysis. This theoretical grounding allows taxonomists to infer a hierarchical organization of the taxa being studied.

For instance, evolutionary theory may infer that the species shared a common ancestor if two species share the same Bauplan (i.e., the structural body plan of a taxon that distinguishes a particular group of organisms). The concept of Bauplan refers to the specific structural layout of a taxonomic group, such as the distinctive characteristics of feline species, which are shared by all members of that group. Without an articulated theory, establishing boundaries or even defining the components of taxonomies would not be straightforward (Feest, 2010).

Phylogenetic trees and other classifications can thus be viewed as "best guesses" regarding the evolutionary relationships among a group of organisms. These representations are inherently provisional and subject to change with the discovery of new species and data.

Research studies often use different sets of comparative data or alternative methodologies, which can lead to divergent phylogenetic trees or hypotheses about relationships among species (Delsuc et al., 2005; Sanderson et al., 2003a; Sanderson et al., 2003b).

Theories underlying classification also shape how relationships between species are represented. This relationship is visually encoded in structures such as phylogenetic diagrams. For instance, evolutionary cladistics interprets a branching pattern where a new species "buds off" from an old one, yet retains the name of the original species. Although the species itself may not change, the branching implies the emergence of a new lineage. This kind of theoretical framework can also create differences in interpretation: for example, between phenetics, which emphasizes overall similarity, and evolutionary cladistics, which focuses on shared ancestry and homologies.

In summary, the structure and representation of relationships between taxa are strongly influenced by the underlying theory. The choice of a specific taxonomic approach will inevitably impact how relationships among species are visualized and understood, and the integration of both descriptive and theoretical elements is necessary to achieve a meaningful classification.

3.6 The Role of Theory in Taxonomy

While some approaches to taxonomy attempt to minimize the influence of theoretical frameworks, it is ultimately

impossible to fully separate theory from the categorization process. In biology, the phenetics school of taxonomy has faced criticisms for being a mere statistical analysis of comparative data, contributing little beyond accelerating and automating comparative analyses (Blackwelder, 1967). Weighing taxonomic characters and determining their placement in a hierarchy require theoretical understanding and human interpretation. Additionally, mathematical approaches often do not yield unique solutions to optimization problems. D'Arcy Thompson's work demonstrates this: related species' morphological differences can be represented through various mathematical transformations without providing a unique representation of their evolutionary transformations (Thompson, 1917; Arthur, 2006).

Ultimately, a description devoid of theory is not useful. To accurately represent biological systems and construct a "tree of life" that shows evolutionary relationships between species, biologically sound theories with meaningful assumptions are required.

3.7 Applying Biological Taxonomic Principles to Actions

Can actions be categorized in a similar way as biological organisms? Do they exhibit taxonomic attributes that allow researchers to distinguish and organize them? Multiple attributes have been highlighted in the literature to distinguish different types of actions. However, for these distinctions to be meaningful, they must be groupable and hierarchical—otherwise, they would simply result in clusters of actions without any structured relationships.

Widely shared in biological systematics, the concept of taxonomic attributes (TAAs) can be adopted in the action domain. To reach an agreement on a taxonomy of actions, it is essential to establish a common agenda to direct research, such as studying brain topography to understand the representations of various actions. Alternatively, a taxonomy could be constructed by systematically comparing available metrics for action classification, using approaches similar to the phenetic program in biology, which relies heavily on mathematics (Mayr, 1969).

Two critical distinctions need to be emphasized. First, the appropriate level of granularity for tracking TAAs is a subjective choice, highly dependent on the goals of the research discipline involved in classification. This work remains agnostic about the level of detail, as we believe the decision should be driven by the specific research objectives. Second, we view systematic biology as a heuristic tool—a means of advancing the understanding of homologies among actions and testing their relationships across different types of characters.

We argue that, although a comprehensive taxonomy of actions is still some distance away, the attempt to build one is the best way to advance research coherently in this field.

3.7.1 On the Demarcation of Actions

In biological systematics, demarcation is an ongoing challenge (as discussed in first chapter). A similar problem exists in categorizing actions. Actions are often context-dependent and exhibit considerable variability, which

makes defining clear boundaries difficult. Therefore, using flexible approaches—such as fuzzy logic, which allows membership in categories to be treated as a matter of degrees rather than a binary decision—could prove beneficial for categorizing actions.

3.8 Ontological Basis of an Action Taxonomy

In building an action taxonomy, a fundamental decision is identifying the appropriate first level of categorization. In biological taxonomy, this is often the species level. For actions, a similar approach might involve adopting distinctions present in natural language, which provide a starting point for categorizing actions. However, simply accepting linguistic distinctions assumes that language accurately reflects the full range of action diversity. Instead, it might be preferable to use language-based distinctions as a foundation while refining or creating new distinctions where existing terminology falls short.

In the study of brain activation associated with actions, the focus is often on isolated laboratory tasks, which may not fully capture the complexity of actions in the real world. The same neural mechanisms may be involved in performing diverse actions (e.g., gripping a ball vs. running a marathon), but each action has a unique set of challenges related to planning and execution. Questions arise about the nature of action representation in the brain: Is there a unique "fingerprint" for each action? How much overlap exists between different actions? And how can brain measures such as EEG, fMRI, ECOG, or MEG help to elucidate these distinctions? Furthermore, what behavioral

referents can be used to compare these neural responses—such as observed or performed actions (Majdandzic et al., 2007; Niki et al., 2019)?

Just as evolutionary biology considers multiple levels of natural selection, actions can also be understood at multiple levels. These levels might range from basic motor units responsible for simple movements to complex, goal-directed activities involving extensive coordination across body systems. Understanding these levels is important for building a comprehensive taxonomy that integrates both micro-level processes and macro-level coordination.

3.9 Taxonomic Monism or Pluralism in Actions

Since a taxonomy is ultimately a tool for organizing information, why should we aim for a single, unified taxonomy of actions? A more modest, yet potentially useful, proposal is to develop distinct taxonomies for different research areas within the broader field of action studies. For example, a taxonomy specific to action perception might include attributes that encode the identity of actions observed. By pre-structuring hierarchical relationships among these attributes, researchers can generate hypotheses and test them empirically. Continuous refinement of these hypotheses could eventually yield a robust taxonomy of action perception.

3.10 Biological Taxonomic Principles and the Action Domain

Can actions exhibit taxonomic attributes that allow for their meaningful classification? For a taxonomy to be valid, it must be both hierarchical and groupable, establishing relationships that offer more than mere clustering of categories. Attributes such as tool-mediated actions versus non-tool-mediated actions might be nested within broader categories, like transitive actions (i.e., actions involving objects).

The development of an action taxonomy necessitates an agreement on the central tenets—those attributes most influential in understanding actions. To reach such a consensus, researchers might need to establish a shared agenda, such as understanding the neural representation of different types of actions, or adopting more numerically driven approaches similar to those used in biological phenetics (Mayr, 1969).

Although a fully developed taxonomy of actions may be distant, pursuing one is the most coherent way to advance knowledge in this area. By applying biological taxonomy principles to the action domain, researchers can make significant strides toward understanding how actions can be organized, classified, and systematically studied.

3.11 Organizing Taxonomic Attributes of Action (TAA): A Tentative Proposal

In this section, we present a classification and detailed description of the main Taxonomic Attributes of Action (TAA) as defined by the current literature. The purpose of organizing these attributes is to build a systematic and

hierarchical understanding of actions. We will first group TAAs into major action domains, providing a structural framework to observe interrelationships among these attributes. At the end of this chapter, we propose a tentative taxonomy of actions using TAAs, drawing from biological classification frameworks as an inspiration.

The major action domains proposed in this taxonomy represent a mix of different theoretical approaches and have an intuitive coherence analogous to the biological concept of regna. These domains aim to capture the complexity of actions and organize TAAs in a way that highlights their relevance across different levels and perspectives.

Major Action Domains

1. Behavioral Action Domain
   This domain focuses on the mechanical aspects of actions, including low-level features and practical applications, such as robotics and computer vision. Behavioral attributes are concerned with the physical execution and real-life application of actions, encompassing performance and recognition components.
   Robotics and computer vision, for instance, often emphasize the analysis of human movement and mechanics to enhance automated action recognition and performance. The objective is to label and categorize observed actions effectively and efficiently, often for practical purposes such as human-robot interaction and machine learning in

computer vision. A variety of taxonomies have been proposed to classify actions based on their low-level features, including grasp types, manipulation types, and tool-mediated versus non-tool-mediated actions (Nakawala et al., 2018; Zech et al., 2018; Feix et al., 2016; Cutkosky, 1989).
In neuroscience, this domain is related to the neural processes responsible for enabling rapid and automatic responses to observed actions. For example, the mirror neuron system (Caggiano et al., 2012) plays a significant role in action recognition by mapping observed actions onto the same neural circuits that would be activated if the action were performed by the observer. These mechanisms have also been framed within the predictive coding framework (Amoruso & Urgesi, 2018; Cerliani et al., 2022), where the brain's ability to predict and simulate observed actions contributes to more efficient interaction and recognition of joint actions (Della Gatta et al., 2017; Zunino et al., 2020).

2. Linguistic Domain
This domain explores the language-like structure of actions and focuses on understanding both action performance and action observation at a syntactic and semantic level. This includes research into how actions are perceived in terms of their intuitive similarities and dissimilarities and the hierarchical structures they can form.
Linguistic attributes of action encompass syntax and semantics, which are essential for a comprehensive understanding of actions. The syntactic aspect refers

to how basic units of action (similar to linguistic words) are combined and concatenated to produce complex actions (analogous to sentences) (Martins et al., 2019; Ziaeetabar et al., 2021; Worgotter et al., 2020). The semantic aspect of actions involves understanding the meaning and relationships between different actions, as captured in intuitive similarity ratings (Tarhan & Konkle, 2020) and studies on action concatenation to achieve distal goals (Braukmann et al., 2017; Thomas et al., 2018; Majdandzic et al., 2007). These dimensions also apply to how multiple actions are sequenced and coordinated to achieve more complex outcomes, effectively forming an action "grammar."

3. Brain-Based Action Domain
   The brain-based domain aims to understand how actions are represented in neural terms. This involves capturing actions through their neural signatures, as measured by brain imaging methodologies like EEG, fMRI, MEG, and electrophysiology. The main focus is to integrate data from different modalities to build a unified, brain-based taxonomy of actions.
   The neural basis of actions is often investigated in terms of how different brain regions represent similar or distinct actions. The temporal receptive windows (TRWs) theory (Baldassano et al., 2021) suggests that different brain areas are tuned to different timescales, thus creating a hierarchy of action understanding from simple, fine-grained actions to more abstract goals. Research using

methods such as representational similarity analysis has shown that different types of actions are represented in overlapping but distinct neural regions (Thornton & Tamir, 2020; Tarhan et al., 2021; Wurm et al., 2017). Further integration of different neural frameworks - such as predictive coding (Badcock et al., 2019) and the common coding theory (Hommel et al., 2001) - may provide a better understanding of the neural implementation of actions.

Algorithmic and Temporal Considerations in Action Taxonomies

Actions are inherently algorithmic, often involving a sequence of steps to achieve a goal. This reliance on algorithmic thinking implies the presence of a temporal hierarchy, which could serve as a basis for classifying actions. For instance, individuals with dyspraxia often struggle with implementing the sequential steps required for common tasks, which could suggest an underlying difficulty in processing the temporal hierarchy of action (Miller et al., 2014).

These temporal hierarchies correspond to the frequency with which actions are sampled and understood, forming a framework that may be similar to what has been proposed for linguistic comprehension (Fedorenko et al., 2016). This temporal arrangement can be mapped onto specific brain areas with corresponding temporal receptive windows (Baldassano et al., 2021). Thus, actions can be understood at different levels depending on their temporal sampling frequency:

- Low-level TAA: Attributes related to shape, fine kinematics, and mechanical components (often studied in computer vision, robotics, and psychophysics) correspond to areas with fast TRWs, such as primary sensory areas.

- Intermediate TAA: Attributes relating to effector-object relations and specific components of actions correspond to middle-level TRWs, which involve more complex neural processes that combine perceptual information into meaningful action segments.

- High-level TAA: Attributes involving goals, intentions, and the broader meaning of actions are associated with higher-level TRWs, which reflect an abstract understanding of the action's purpose.

Contributions from Computer Vision

The contributions from computer vision to action taxonomy are also significant, particularly in the development of action recognition and labeling algorithms. There are two primary types of features in action recognition: hand-crafted features and deep-learned features. Hand-crafted features, such as improved trajectories and Fisher vectors, have been effective for many years in describing action kinematics and dynamics (Wang et al., 2011). More recently, deep learning approaches, such as convolutional neural networks (CNNs), have allowed for greater autonomy in learning action features (Simonyan & Zisserman, 2014; Shi et al., 2019).

To leverage these features for taxonomic purposes, a promising approach is to use representational similarity analysis to compare the activation of the final layers of CNN architectures across different actions. This could reveal data-driven similarities and differences between actions, which can then be organized into a hierarchical framework through clustering techniques.

3.12 Conclusion

Using the taxonomic attributes identified in the literature, we have proposed a tentative taxonomy of actions that is both hierarchical and domain-specific. The behavioral action domain emphasizes low-level kinematic and mechanical attributes that are relevant to robotics and real-life execution. The linguistic domain captures higher-order relationships between actions, including their syntactic concatenation and semantic properties. Finally, the brain-based domain aims to develop a unified neural representation of actions, integrating multiple levels of observation and methodological approaches.

This classification framework is intended to guide future research in understanding and organizing the vast diversity of human actions. A key aspect of this approach is that it is flexible: different levels of granularity can be employed depending on the research question, allowing for both high-level conceptual understanding and fine-grained mechanistic analysis.

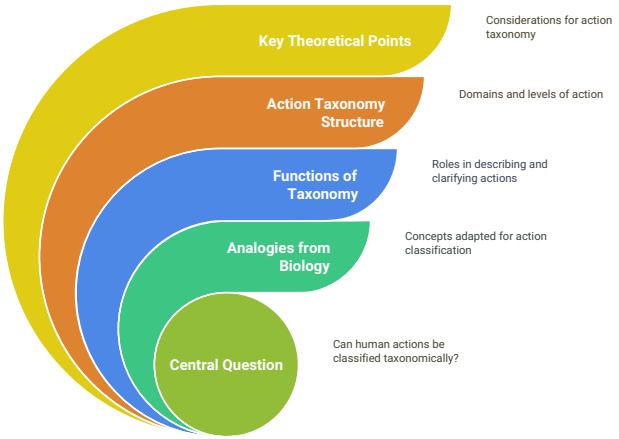

Fig. 3: Multi-layered diagram illustrating how principles from biological taxonomy inform the classification of human actions. Central concepts such as taxonomic attributes, fuzzy boundaries, and the Bauplan analogy are linked to three core functions—structure, theory, and clarity—and organized across behavioral, linguistic, and brain-based domains, each with low-, intermediate-, and high-level attributes.

Chapter 4

4. Heuristic and Organizing Principles of Action Taxonomies

The central question driving the development of action taxonomies is: "What is the best way to segment the action phenomenon?" The goal, undertaken from diverse perspectives, is to identify an appropriate method for discriminating, dividing, and ultimately explaining the action phenomenon in the most efficient, detailed, and non-invasive way. The aim is to approach action as Lord Wenhui's butcher does in the famous Zen tale from the Zhuangzi, cutting effortlessly through natural divisions without exerting unnecessary force:

"I go along with the natural makeup, strike in the big hollows, guide the knife through the big openings, and follow things as they are. So I never touch the smallest ligament or tendon, much less a main joint. There are spaces between the joints, and the blade of the knife has really no thickness. If you insert what has no thickness into such spaces, then there's plenty of room—more than enough for the blade to play."

In this analogy, we aspire to segment the action phenomenon precisely as it is, without imposing artificial or excessive divisions, but instead finding ways to accommodate all empirical evidence as we delineate the profile of our knowledge of actions.

Before delving into heuristic principles and their application to existing action taxonomies, we must address a philosophical question: can we hope to develop a macro-taxonomy that effectively fuses all levels of action understanding, or must we concede that "different perspectives yield different truths" in the study of actions? Is it even possible to develop a taxonomy that accounts for both the kinematic details and the expansive variety of human actions? The answer is definitively yes, though such an endeavor may be impractical given current methodological constraints. Fortunately, more and more datasets containing precise kinematic descriptions of actions are becoming available online (Chung et al., 2020; Smaira et al., 2020; Soomro et al., 2012).

Theoretically, if methodological barriers were removed, we could describe actions with maximal detail. We could record neural activity at the level of single neurons and capture their temporal dynamics at millisecond resolution throughout the brain, while dissecting stimuli into increasingly fine-grained semantic and kinematic models. The effector-specific taxonomies described by Castiello (2005) and Liu (2015) could be extended to all effectors. It appears there are no theoretical limits to our capacity for description, only methodological ones.

However, this raises new questions: How many models or taxonomies could be derived from the same stimuli? Likely, countless models could be proposed. Moreover, how does the brain transform dynamic entities, such as actions, even before we become consciously aware of them? The brain is inherently dynamic and hierarchically

organized (Heer et al., 2017; Raut et al., 2020). It processes information across multiple dimensions at different stages, reflecting the complexity of action perception and production.

To return to the analogy from the Zhuangzi, the natural joints of the action phenomenon are found at the "junctures" within the information processing stream, where dramatic changes occur in the type of encoded information. Therefore, pursuing a useful taxonomy of actions means illuminating our knowledge of actions by emphasizing the areas we do not yet understand—delineating the object by coloring in the background.

The following sections propose several heuristic principles that can help conceptualize and evaluate action taxonomies, recognizing their strengths and limitations. These principles are derived from philosophical considerations and aim to organize and unify action-related research. Given that there are few action taxonomies available in the literature, many references in the following sections may refer repeatedly to the same taxonomies. Additionally, we will often draw from studies that investigated specific dimensions of action but did not attempt to develop a comprehensive taxonomy. The goal is to piece together these fragmented studies to synthesize a broader understanding of actions.

4.1 Heuristic Principles for Action Taxonomies

The principles outlined below serve as criteria for navigating existing action taxonomies and evaluating them

within a common framework. By comparing various taxonomies and their empirical foundations, we aim to identify gaps in the current landscape of action research and guide future questions and experiments.

1. Granularity
   The principle of granularity refers to how finely actions are described. Each description is granular at a particular level. If researchers study the kinematics of action, they may achieve an exceptionally detailed description of actions' mechanical properties. However, if they switch perspectives—say, to brain activity—the level of detail may change substantially. In action production, a highly granular description could include the precise pattern of muscle contractions required to perform an action. On the other hand, granularity may be lower when analyzing broad neural patterns involved in the same movement.
   Importantly, granularity must be balanced with completeness, defined as how much of the actual variety of human actions is captured by the taxonomy. There is often a trade-off between these two elements: achieving high granularity might mean analyzing a smaller subset of actions, whereas capturing the full diversity of human actions might necessitate using broader and less detailed descriptors. Figure 1 below illustrates this trade-off between granularity and completeness. Taxonomies focusing on a broad observational standpoint often sacrifice granularity for completeness and vice versa.

2. Interactivity
   The level of interactivity in a task (e.g., predicting the next action or judging the congruency of an observed action) determines how the brain's action-related networks are engaged. Interactive tasks—those that require active engagement, such as joint action prediction—recruit more motor resources and result in distinct patterns of brain activity (Butterfill & Sebanz, 2014). Passive tasks, such as simple observation, engage different processes. We can also consider interactivity in the context of action production: the complexity of a performed action (e.g., walking vs. learning new dance steps) can vary greatly in terms of required planning and attention, consequently involving different neural resources. The relevance of attention within interactive tasks is well established. Attending to certain features of actions modifies both single-neuron activity (Caggiano et al., 2012; Toschi, 2017) and regional brain activation patterns (Nastase et al., 2017; Shahdloo et al., 2021). This modulation of action perception by attentional focus highlights the interaction between interactivity and hierarchy. For example, attentional focus can warp action representations in the brain, selectively enhancing response patterns relevant to the attended features.

3. Observation-Production
   Many taxonomies distinguish between action observation and action production. Computer vision and robotics, for instance, often consider only one side of the problem. Robotics is primarily interested

in action production—creating robots that move as efficiently as humans—while computer vision focuses on recognizing and understanding observed actions. However, cognitive neuroscience emphasizes the interaction between action production and action perception, with evidence suggesting that the two processes share functional overlap (De Kleijn et al., 2014). The mirror neuron system is an illustrative example: mirror neurons activate both when observing an action and when performing the same action (Casile & Sinigaglia, 2013).

4. Hierarchy
   Hierarchical organization is implicit in all actions. Actions are embedded in a series of nested hierarchies, both conceptually and within the brain. Complex actions require coordination across multiple stages, and long-duration actions necessitate memory to understand how individual components contribute to the overall goal. There are also notable similarities between narrative comprehension and action comprehension, as both require integration over extended timescales (Baldassano et al., 2017; K. Pastra, 2012). The hierarchical structure of actions is evident at multiple levels of processing, from simple motor acts (e.g., grasping) to sequences of actions integrated into a narrative (Ziaeetabar et al., 2021).

5. Abstraction
   Different taxonomies classify actions at different

levels of abstraction, ranging from concrete features (e.g., grasping type) to more abstract concepts (e.g., sociality). Taxonomies modeling low-level kinematic features of actions span a different space than those focusing on higher-level semantic features (e.g., locomotion, communication). This separation is reflected in neural correlates, with a posterior-to-anterior gradient of representation for abstract versus concrete features in the lateral occipitotemporal cortex (LOTC) (Wurm et al., 2017). Considering different levels of abstraction is essential when integrating multiple taxonomies or understanding the overlap between seemingly distinct action categories.

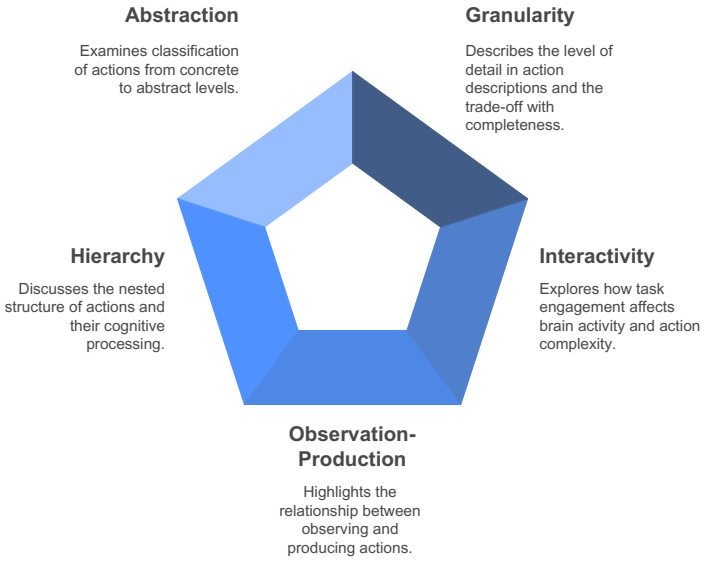

Fig. 4: A graphical representation of the most relevant distinctions and features for an action taxonomy highlighted in the section. First, granularity: how fine and detailed is the description of the stimulus, both from a brain and a stimulus perspective? To this principle connects a methodologically implied one, that we could call completeness. Does the variety of actions considered in the taxonomy reflect the actual variety of actions that we humans do and observe? Are there biases toward certain types of actions in actual experiments? And how does this depend on the granularity of the analysis? Second, abstraction: how much distant and transformed is the action information in the brain as compared to the source (Spunt et al., 2016)? Notice that this feature could be applied to both representations of action in the brain (e.g., the notion of a supramodal representation of actions as in Ricciardi et al., 2013) and to hidden layers of an ANN processing human actions. Third, completeness: Fourth, active-passive: does the taxonomy consider the differences in the underlying mechanisms necessary to passively perceive actions and to actively perceive actions? Actively can mean several attitudes towards the observed actions: it is meant here to indicate tasks in which the subject has to pay attention, with various modulations, to different features of the actions shown. The concept also applies to active or passive production of actions: here, the subjects observe features of its own actions actively, or the actions are only a means for an end, no particular attention in the experimental context is paid to the action itself (but rather to its consequences). Fifth, hierarchy: each action is incapsulated within a hierarchical intentional cascade; simultaneously, the kinematics of action is as well wrapped up in a hierarchy of smaller and smaller acts. Each action in a certain sense is a matryoshka of actions: this hierarchy is probably present also in the brain (Cattaneo et al., 2010), even if coded in ways we don't fully understand for the moment.

4.2 Methodological Considerations

The heuristic principles described above are deeply influenced by the methodologies applied to study action phenomena. Granularity is constrained by methodological limitations. For example, EEG is optimal for analyzing temporal features of action perception, while fMRI is more suitable for investigating the spatial distribution of neural activation, but with lower temporal precision. Integrating these methodologies allows for a comprehensive understanding of both the localization and temporal dynamics of action comprehension (Ortigue et al., 2009).

The principle of interactivity is also crucial from a methodological perspective. Interactivity determines the degree of participant engagement, which in turn influences the level of cognitive and motor processing required. For example, attentional focus on specific action features can significantly modify action perception, as demonstrated in naturalistic action observation (Shahdloo et al., 2021).

Observation-production studies must account for the overlap between these processes and recognize that action understanding often involves implicit prediction. For example, gaze anticipation is an indicator of action comprehension during motor primitive observation (Ambrosini et al., 2013; Costantini et al., 2014). In contrast, for complex or unfamiliar actions, more explicit task requirements are needed to assess comprehension.

A notable challenge encountered in this review is the limited number of comprehensive action taxonomies available in the literature. Most studies tend to focus on a select few dimensions of actions, which is understandable

given the experimental difficulties of analyzing data across multiple dimensions and levels. At times, the discussion will rely on studies that did not explicitly aim to provide a complete account of actions but instead focused on specific dimensions.

The following sections will utilize the set of heuristic principles illustrated before. These principles will provide a framework for evaluating each action taxonomy and model in a consistent manner, allowing us to determine the contributions of each taxonomy to the broader understanding of action. For example, an effective action taxonomy must account for the overlap between action production and perception—such as what is captured by the literature on mirror neurons, which provides insights into the functional interdependence between motor programming and action understanding. As such, this overlap between perception and production must be considered for evaluating action taxonomies, alongside the other principles discussed.

4.3 Granularity

The first principle is granularity—the level of detail at which an action is described. The granularity can vary across different levels of analysis. A researcher focused on the kinematics of actions may seek a highly detailed description of the movement, while a broader description might be appropriate when switching to a different level of analysis, such as understanding brain activity associated with these movements.

For instance, at the production level, granularity can refer to the precise coordination of muscle contractions involved in executing an action, whereas, at the level of brain activity, a shift occurs in the type of detail and focus required. A critical trade-off arises: the completeness of the mapping—how comprehensive the coverage of action types is—often comes at the cost of granularity. This trade-off is illustrated in Figure 1, where the extent to which taxonomies can describe a wide range of human actions depends on the adopted level of granularity.

4.4 Interactivity

The principle of interactivity concerns the degree to which tasks involve active participation, which can significantly influence brain activity related to motor processes. Interactive tasks can take various forms, such as predicting the outcome of an observed action, choosing between congruent and incongruent responses, or engaging in an action. Tasks with lower levels of interactivity can include naturalistic stimulation or orthogonal tasks, where the participant's response is mostly independent of the observed action.

Regarding action production, it may seem redundant to discuss interactivity since executing an action is inherently an active task. However, distinctions can still be made based on gradation and expertise. For example, performing a simple, everyday action like walking involves less cognitive effort compared to learning a complex dance routine. In the same manner, an experimental task involving joint action (e.g., collaborative tasks; Butterfill,

2014) requires more planning and attention, thereby activating different motor-related networks.

The attentional processes involved are crucial for the interactivity principle, as evidenced by multiple studies that demonstrate how attention modulates action observation and execution at both the single neuron level (e.g., Caggiano et al., 2012; Toschi et al., 2017) and across brain regions in neuroimaging studies (e.g., Haxby et al., 2020; Nastase et al., 2017; Bach et al., 2017). A study by Nastase and colleagues (Nastase et al., 2018) found that directing attention towards different dimensions of action —such as observing actions performed by animals compared to attending to taxonomic information—alters the multivariate response patterns across brain regions. Attention also warps semantic maps in the brain, shifting the response selectivity of certain brain areas according to action features or categories (e.g., Shadloo et al., 2021).

In summary, we need to consider whether taxonomies or studies account for the various levels of interaction during action observation, and whether the actions are isolated or occur under specific situational constraints.

4.5 Observation-Production

An important distinction is whether the taxonomy considers both action production and perception. This principle is pivotal for action-related disciplines such as computer vision and robotics. Robotics focuses primarily on producing human-like actions, whereas computer vision

emphasizes the accurate recognition of actions from visual data.

In cognitive neuroscience, the emphasis is not only on action kinematics but also on understanding how action production and perception are functionally linked within the brain, as demonstrated by mirror neuron research (e.g., De Kleijn et al., 2014). Although these two processes are often treated separately, there is a significant overlap, and understanding this overlap is crucial for developing comprehensive action taxonomies.

4.6 Hierarchy

The concept of hierarchy can be approached from two perspectives: the hierarchical nature of actions themselves and the hierarchical processing of these actions in the brain. For an action to be produced or understood, information must pass through multiple cortical levels. Depending on one's perspective of brain functioning—whether representationalist or enactivist (Friston et al., 2018; Gadziejewski et al., 2016)—the interpretation of these hierarchies may vary.

Longer sequences of actions inherently require a form of narrative understanding, akin to language comprehension (Baldassano et al., 2017). Such temporal extension is a fundamental feature across taxonomies, with some researchers focusing on brief, isolated actions (e.g., Ahlheim et al., 2014), while others investigate more complex, chained action sequences (e.g., Ziateebar et al., 2021). Based on the literature on mirror neurons, we can

distinguish between narrative prediction and motor prediction. Narrative prediction is more relevant to understanding extended action sequences, whereas motor prediction is better suited for isolated, immediate actions.

4.7 Abstraction

Abstraction refers to the distinction between higher-level conceptual features of actions and more concrete, observable features. Some taxonomies focus on low-level features such as shape and kinematics, whereas others describe more abstract attributes such as action goals and sociality (e.g., Tarhan & Konkle, 2020). This separation between low-level and high-level features also has a neural correlate. For example, Wurm et al. (2017) found a posterior-to-anterior gradient in the lateral occipitotemporal cortex (LOTC) that transitions from representing concrete action features to abstract ones.

A study by Bach and colleagues (2017) also illustrated how different levels of abstraction influence brain processing: action novelty and ambiguity activate motor-related regions, while familiar actions are primarily associated with semantic object-related associations rather than movements. This distinction highlights the existence of multiple abstraction levels across actions.

4.8 Methodological Perspective on Heuristic Principles

From a methodological viewpoint, these heuristic principles are deeply intertwined with the methods employed in studies of action comprehension. **Granularity**,

for instance, is influenced by the type of methodology used —EEG studies excel at exploring the temporal structure of actions (e.g., Bub et al., 2018, Ortigue et al., 2010, Pomiechowska et al., 2017), while fMRI studies better capture spatial localization of brain activity (e.g., Ortigue et al., 2009). Thus, fMRI is often used to study more abstract levels of action comprehension, while EEG captures fine temporal details.

At the behavioral level, there are similar trade-offs between granularity and completeness. Studying the specific trajectories of an effector during a particular action requires focusing on a limited number of action types (e.g., grasping), whereas investigating the full diversity of human actions necessitates a broader approach, at the cost of reduced granularity.

Interactions between action observation and other cognitive processes, such as attention and emotional engagement, further complicate the picture, as they can alter both kinematic and abstract action processing (e.g., Alaerts et al., 2011; Castiello et al., 2010; Shadloo et al., 2021). Thus, the interactivity principle is highly relevant from a methodological perspective, influencing the abstraction and hierarchy of action processing. Different tasks can result in different processing hierarchies (e.g., Isik et al., 2018).

One effective approach to assessing the importance of specific action dimensions is to investigate the impact of removing certain cues on action recognition (e.g., El-Sourani et al., 2019). For instance, removing contextual cues

reduces action recognition performance, while removing key parts of an action sequence impacts the ability to predict its outcome. Such analyses, ideally conducted across a wide variety of actions and taxonomic dimensions, help determine which aspects of action information are crucial for comprehension.

Distinguishing between action observation and action prediction further refines the evaluation of comprehension. In the case of motor primitives (e.g., grasping), gaze anticipation can serve as a marker of understanding (e.g., Ambrosini et al., 2013; Costantini et al., 2014). However, in passive observation tasks—such as viewing short action clips—understanding is often implied. The **interactivity** principle thus helps account for such differences.

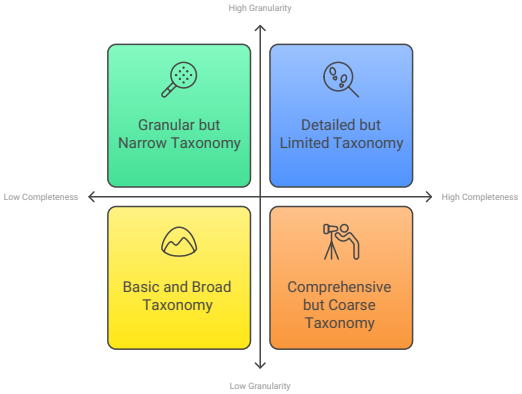

Fig. 5: The figure shows that taxonomies providing a more comprehensive extension (i.e., capturing a wide variety of actions) often sacrifice granularity in describing the constituents of these

actions, while more detailed taxonomies tend to capture fewer types of actions. However, even though obtaining both completeness and granularity across all levels of hierarchy is methodologically challenging, it remains theoretically possible.

In this context, these principles can help organize and critically evaluate proposed taxonomies of action by establishing a structured reference frame, providing deeper insight into how individual models relate to each other and how each one contributes uniquely to our understanding of action. In doing so, we can determine the strengths and weaknesses of existing taxonomies and suggest directions for future research. Such directions could include investigating aspects of action not thoroughly examined, integrating multidisciplinary perspectives, and developing experimental methods that yield greater granularity while still retaining diversity in action representation.

In summary, the principles discussed here represent an invaluable heuristic for guiding research into action taxonomies. These principles allow for a nuanced approach to characterizing action by incorporating diverse dimensions, from abstraction and granularity to interaction levels and hierarchical structure. The journey toward a unifying taxonomy for action phenomena is ongoing, but by employing these principles as both guidelines and evaluation tools, researchers can develop taxonomies that do not merely classify actions but also reveal underlying processes and interconnections.

4.9 Conclusion

In conclusion, a comprehensive and insightful action taxonomy is not a static endeavor but a dynamic interplay between theory, empirical findings, and heuristic inquiry. By focusing on core principles such as granularity, completeness, abstraction, interaction, and hierarchy, we can systematically break down the complex phenomenon of human action and capture its essence across varying dimensions. Each principle provides a distinct lens through which to view and interpret the nature of actions, helping us understand the different ways in which actions can be classified, represented, and understood.

However, more important than these individual principles is their collective application. Much like pieces of a puzzle, these elements are most effective when brought together to generate a complete picture. It is through this holistic approach—one that acknowledges both the diversity and interconnectedness of action representations—that we can make significant strides in action research. This synthesis of theoretical, empirical, and heuristic principles has the potential to foster breakthroughs, uncover hidden dimensions, and inspire new and transformative ways of thinking about human actions in both cognitive neuroscience and related disciplines.

Chapter 5

5. Introduction

A fundamental question in cognitive neuroscience is to what extent different human brains operate similarly. While neuroimaging studies have consistently revealed substantial similarities across individuals, these results often stem from highly controlled experimental conditions, which limit individual variability. Most studies investigating cortical similarity utilize simplified stimuli, such as single, isolated images (e.g., IAPS, Lang, 2005), making it difficult to generalize findings to more complex, ecologically valid settings (Hasson et al., 2004).

Naturalistic stimuli, such as movies, offer a richer framework for studying brain function due to the contextual integration of multiple sensory modalities and their dynamic, multi-object environment. This richness enables a more holistic examination of cognitive processes, unlike the traditional focus on controlled, simple tasks. The use of naturalistic stimuli reveals complex neural activations, suggesting that different levels of perceptual, semantic, and motor dimensions underlie action recognition, and these processes may not fully overlap with traditional simple task-based findings.

Humans recognize thousands of objects and actions, though it is unlikely that each category has a distinct neural region dedicated to it. Instead, studies show that the brain

utilizes a continuous semantic space for categorization based on similarity between categories (Seger & Peterson, 2013; Freedman & Assad, 2016). Categorization in the brain is thought to proceed through a two-stage model: initial perceptual identification followed by a categorical decision process (Freedman et al., 2003; Scholl et al., 2014). Early sensory regions are responsible for perceptual processing, while the prefrontal and parietal cortices are implicated in decision-making processes, such as classifying stimuli (Riesenhuber & Poggio, 2000).

Moreover, symbolic labeling (naming) is essential to categorization in real-world settings. Symbols enhance categorical learning, object recognition, and abstraction (Althaus & Westermann, 2016; Lupyan et al., 2007). However, the neural correlates of symbolic use in categorization remain underexplored. It has been hypothesized that categorizing stimuli through symbols may generate shared neural representations between objects and their symbolic forms (Viganò et al., 2021).

A wealth of research indicates that perceiving and generating actions are tightly linked. Mirror neurons, originally discovered in the premotor cortex (area F5) of macaques, are active both during the execution of an action and while observing the same action performed by others (Gallese et al., 1996; Rizzolatti et al., 1996). In humans, the observation and execution of actions engage the inferior frontal gyrus (IFG), inferior parietal lobule (IPL), and superior temporal sulcus (STS). These areas support the direct mapping of observed actions onto motor

representations, facilitating action understanding (Iacoboni et al., 2005).

The perception of actions in naturalistic settings, such as movies, involves the integration of perceptual, semantic, and motor processes. Naturalistic stimuli elicit consistent brain responses across individuals, as evidenced by fMRI, EEG, MEG, and ECoG studies (Hasson et al., 2004; Bhattasali et al., 2019). These responses are distributed across sensory, associative, and higher-order cortical regions, suggesting that naturalistic paradigms capture richer and more ecologically valid neural representations than traditional controlled stimuli.

Naturalistic stimuli, such as movies, can evoke synchronized brain responses across individuals, not only in sensory areas but also in higher-order cortical regions involved in attention and social processing (Hasson et al., 2004). The use of inter-subject correlation (ISC) measures has proven particularly effective in analyzing such responses, as they do not require a strict parametric model (Pajula et al., 2012). This method allows for better detection of complex, distributed activations related to continuous natural stimuli.

Studies involving prolonged movie viewing have demonstrated hierarchical temporal receptive windows (TRW) within the brain, wherein lower-order sensory regions show rapid responses, while higher-order associative areas require integration over longer timescales (Hasson et al., 2008). This temporal hierarchy suggests that

distinct areas integrate different scales of information to support coherent perception in real-world environments.

5.1 The Current Study: Investigating Action Recognition during Film Viewing

This study aims to explore which features of actions serve as the best predictors of fMRI data acquired during naturalistic stimulation, aiming at creating models as rich as possible, to capture all the underlying complexity of the act of understanding actions in movies. Participants watched three films ("101 Dalmatians," "Forrest Gump," and "The Grand Budapest Hotel"), and brain activity was recorded using fMRI. Two models were compared: a categorical model, which included descriptors of actions (e.g., context, agent type, effector visibility, transitivity), and a semantic model, encompassing both the agent and object of the action.

The results revealed that both models significantly predicted brain activity within networks involved in action observation ($p<0.05$). The semantic model showed stronger representation in early visual areas, lateral occipital, and fusiform cortices, while the categorical model showed greater representation in the superior temporal sulcus, parietal areas, and dorsal extrastriate regions. The observed high correlation between models supports the hypothesis that integrated processing across multiple dimensions underlies the natural comprehension of actions.

5.2 Methods

5.2.1 fMRI Datasets

We analyzed three fMRI datasets from separate studies, each containing naturalistic stimuli.

- *Dataset 1* (Setti et al., 2023): Fifty participants watched a shortened version (~50 minutes) of the film *101 Dalmatians* (1996) in three modalities (audiovisual, visual-only, and audio description). The participants were typically developed individuals, as well as congenitally blind and deaf individuals. For our analysis, we selected 10 typically developed subjects (n=10, age 35 ± 13 years, 8 females) who viewed the audiovisual version during fMRI scanning.
- *Dataset 2* (Studyforrest, Phase II, http://www.studyforrest.org): Fourteen participants (n=14, mean age 29.4 years, range 20-40 years, 6 females) viewed a modified German-dubbed version of *Forrest Gump* (~90 minutes) (Zemeckis, 1994). These participants were previously involved in a study investigating the dynamics of emotional processing under naturalistic stimulation (Lettieri et al., 2019). We utilized the preprocessed data from that study for subsequent analyses.
- *Dataset 3* (Visconti di Oleggio Castello et al., 2020): Twenty-five subjects (n=25, mean age 27.52 ± 2.26 years, 13 females) watched the second part of the movie *Grand Budapest Hotel* (~45 minutes).

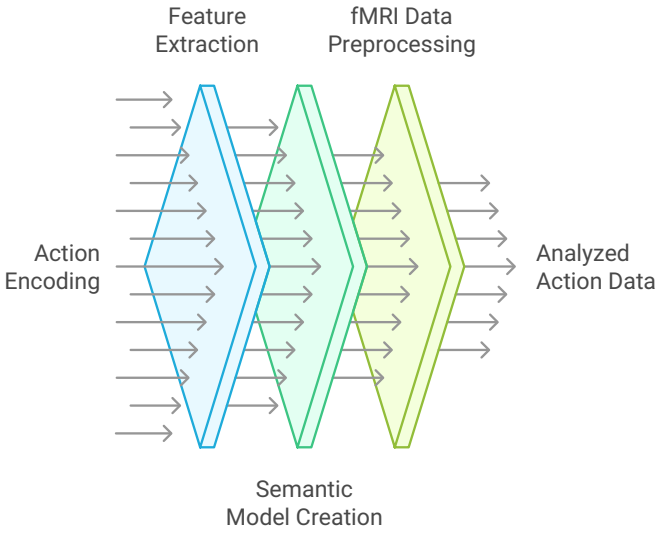

Fig. 6: Overview of the analysis steps of the project. First, a rater (L.T.) encoded each action based on the features extracted from relevant literature on action understanding and naturalistic stimuli. Semantic model was instead created on the basis of a word-to-vec representation of action and object labels assigned to action event. The same approach was applied to the three different movies. fMRI data were preprocessed in a similar fashion in all the three cases.

5.2.2 Acquisition and Preprocessing of fMRI Data

Dataset 1: *101 Dalmatians*

fMRI data were acquired using a Philips 3T Ingenia scanner equipped with a 32-channel head coil. Functional images were collected using the following parameters: GRE-EPI,

TR = 2000 ms, TE = 30 ms, flip angle = 75°, FOV = 240 mm, in-plane acquisition matrix = 80×80, slice thickness = 3 mm, voxel size = 3×3×3 mm, 38 ascending sequential axial slices, with a total of 1614 volumes for six runs of the movie and 256 volumes for a control run.

The preprocessing pipeline was performed using AFNI_17.1.12 (Cox, 1996) and included standard steps. Initially, scanner-related noise was removed by applying spike correction (3dDespike). Then, slice timing correction (3dTshift) was performed, followed by head motion correction using the first run as reference (3dvolreg). Spatial smoothing was applied using a Gaussian kernel (3dBlurToFWHM, 6 mm, full width at half maximum). The normalized runs underwent detrending using Savitzky-Golay filtering (order 3, window size 200 points) in MATLAB R2019b (MathWorks Inc., Natick, MA, USA). Runs were concatenated, and multiple regression analysis was performed (3dDeconvolve) to remove head movement-related signals and motion peak regressors (framewise displacement above 0.3). Volumes were subsequently registered non-linearly to MNI-152 standard space (3dQWarp).

Dataset 2: ***Forrest Gump***

Images were acquired with a Philips Achieva dStream 3T MRI scanner using a 32-channel head coil. Functional images were collected using T2*-weighted gradient-echo echoplanar imaging (TR = 2000 ms, TE = 30 ms, flip angle = 90°, bandwidth = 1943 Hz/Px). The acquisition consisted of 35 ascending axial slices (slice thickness = 3.0 mm) with

80×80 voxels (3.0×3.0 mm) and a field of view (FOV) of 240 mm, providing whole-brain coverage. A total of 3599 volumes were acquired across eight segments of the movie.

The preprocessing was conducted by Lettieri et al. (2019) using ANTs (Avants et al., 2009) and AFNI v.17.2.00 (Cox, 1996). This included brain extraction, non-linear registration to MNI-152 space, slice timing correction, head movement correction (3dvolreg), and spatial smoothing (FWHM = 6 mm). The hemodynamic signal was detrended to remove slow drifts, head movement, and physiological artifacts.

Dataset 3: ***Grand Budapest Hotel***

Functional and structural volumes were acquired using a Siemens Magnetom Prisma 3T MRI scanner with a 32-channel phased-array head coil at the Dartmouth Brain Imaging Center. BOLD functional images were collected using gradient-echo echoplanar imaging (TR/TE = 1000/33 ms, flip angle = 59°, isotropic voxel size = 2.5 mm, matrix size = 96×96, FOV = 240×240 mm, 52 axial slices). Three dummy scans were acquired at the start of each session for signal stabilization. Structural T1-weighted scans were obtained with MPRAGE (TR/TE/TI = 2300/2.32/933 ms, voxel resolution 0.9375×0.9375×0.9 mm).

The preprocessing steps were conducted following Lettieri et al. (2019), using in-house scripts. The functional data were downsampled to match the temporal resolution (TR = 2s) of the other datasets. The initial 10 TRs were removed due to film onset delay, leaving 1477 volumes in total.

5.2.3 Analysis and Annotation of the Films

To categorize the features of the presented stimuli, we annotated the films using a mix of fully automated and human labeling approaches, as described in the previous section. Three independent annotators (L.T., A.I., L.M.) systematically identified the presence or absence of each of the dimensions outlined in Table 1 for each action in the film *101 Dalmatians*. This was used as a proxy for the consistency of the annotations for the other two movies. Each stimulus was analyzed run by run, and the timecodes of each action were extracted with a temporal resolution of hundredths of a second. Some dimensions presented choices beyond 1 or 0 (e.g., the "main effector" dimension, which included 13 levels). For certain dimensions, NaN was also accepted, indicating that none of the given options applied to a specific action (e.g., in the multi-agent or joint-action section when the action involved only a single agent). The timestamps for each action were determined by the first of the two annotators (L.T.).

Inter-rater consistency in annotating the films was estimated by correlating the two resulting design matrices after encoding all columns of the matrix using custom MATLAB functions. As described earlier, the semantic model of action was based on the verb associated with the action and the noun of the object/agent annotated by the two raters. Action verbs were chosen to describe the action in a generalized manner, avoiding rare or unusual verbs. Particular care was also taken in selecting nouns to describe either the associated object (if present) or the agent (if present).

It is believed that the noun provides contextual information to the verb. Generally, when an action was intransitive, the agent was described instead of the associated object. These actions were often intransitive and communicative (e.g., a dog barking or a man talking), whereas actions in which the object was described by a noun were typically transitive and tool-mediated.

5.3 Model Description and Construction

The following models constitute the three different levels or perspectives onto which the actions displayed in the movie have been captured. The low-level visual model captures the perceptual properties of the film stimuli by analyzing motion energy across time. Using a set of Gabor wavelets tuned to different spatial and temporal frequencies, this model quantifies the dynamics of visual information, emphasizing rapidly changing motion patterns that are known to engage early visual and motion-sensitive cortical areas, such as the occipital cortex and middle temporal (MT) regions. In contrast, the high-level categorical action model represents actions based on a structured set of conceptual features, including transitivity, tool mediation, sociality, and effector involvement. By categorizing actions along these dimensions, this model provides a framework for understanding how the brain encodes goal-directed movements and interaction contexts, engaging regions associated with action recognition and social cognition, such as the superior temporal sulcus (STS) and temporoparietal junction (TPJ). Finally, the high-level semantic action model captures the linguistic and conceptual associations of actions by leveraging word

embeddings derived from a word2vec representation. This model encodes the semantic relationships between action verbs and their associated agents or objects, allowing for an exploration of how meaning is structured in the brain. Previous research suggests that such representations recruit lateral temporal and prefrontal regions, facilitating the integration of conceptual knowledge with perceptual experience.

Low-level visual model

The total level of motion energy was computed for each second of the film using a set of 6,555 motion energy descriptors, consisting of a quadrature pair of space-time Gabor filters (Gabor wavelets with three distinct temporal frequencies: 0 - static energy -, 2, and 4 Hz; Nishimoto et al., 2011). The model describes each frame of the film based on a set of specific spatial frequencies, orientations, and temporal frequencies that capture rapidly changing visual information.

High-level categorical action model

To categorize actions during the viewing of films, we constructed a model considering the following 17 dimensions: context, interaction, target, predictability, agent, main effector, effector visibility, transitivity, action description, object description, sociality, dynamics, durability, telicity, iterativity, multi-agent/joint action, and tool mediation. The specific coding for each dimension is described below.

- **Context** describes the scene in which the action is occurring, categorized into three levels (e.g., indoors, urban outdoors, rural outdoors). This dimension is represented predominantly in the visual cortex and fusiform gyrus (Masson & Isik, 2021).
- **Interaction scale** (Tarhan & Konkle, 2020) measures the level of movement required to complete an action (e.g., knitting vs. running).
- **Target** describes the directionality of movement, coded into four levels: towards the agent, another human agent, an animal, or an object. Previous studies showed that interacting with another human primarily involves the occipital face area (OFA), fusiform face area (FFA), and superior temporal sulcus (STS) (Tarhan & Konkle, 2020).
- **Predictability** describes the predictability of actions observed (0 = unpredictable, 1 = predictable). Predictable actions are associated with the Action Observation Execution Network (AOEN), while unpredictable ones activate regions associated with Theory of Mind (ToM) (Thomas et al., 2018).
- **Agent** identifies whether the action is performed by a human or non-human agent, represented in the extrastriate visual cortex and the fusiform gyrus (Haxby et al., 2020).
- **Main effector** is categorized into 13 levels based on the body part involved in the action (e.g., hand, arm, leg). Different parts of the body are represented in distinct regions, such as the intraparietal sulcus (IPS) and temporo-occipital cortex (Tarhan & Konkle, 2020).

- **Effector visibility** describes whether the effector is visible or inferred, involving visual cortices and temporoparietal regions (Tarhan & Konkle, 2020).
- **Transitivity** indicates whether the action involves an object interaction. Transitive actions are represented in the ventral lateral occipitotemporal cortex (LOTC) (Wurm et al., 2017).
- **Action description and object description** are part of the semantic model, represented mainly in the visual cortex and fusiform gyrus (Masson & Isik, 2021).
- **Sociality** describes whether the action involves social interaction, represented in the dorsal LOTC and STS (Wurm et al., 2017).
- **Dynamics** (Aflalo et al., 2020) indicates whether an action involves movement (e.g., running vs. thinking).
- **Durability** (Aflalo et al., 2020) specifies whether the action is instantaneous or sustained over time.
- **Telicity** (Aflalo et al., 2020) describes actions with a clear endpoint.
- **Iterativity** (Aflalo et al., 2020) indicates whether the action requires repetitive movement patterns. All these dimensions are represented within the action observation network, particularly in the posterior parietal cortex.
- **Multi-agent/joint action** distinguishes between actions performed by multiple agents or jointactions. Involves visual, prefrontal, parietal cortices, fusiform gyrus, and STS (Sebanz et al., 2016).

- **Tool mediation** describes whether the action requires the use of a tool, represented in the dorsal parietal cortices, the ventral stream, and the left posterior middle temporal gyrus (Ricciardi et al., 2013).

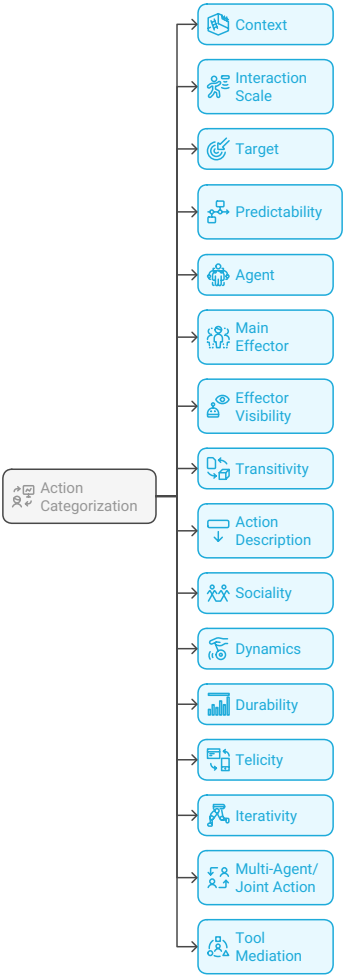

Fig. 7: Comprehensive list of action features used to categorize actions across the three films. Each detected action was systematically annotated with categorical descriptors capturing distinct facets of its representation.

High-level semantic action model

For the high-level semantic action model, we used the *word2vec* algorithm (Mikolov et al., 2013). Word2vec was employed to obtain semantic representations for the verb of the action and its associated agent/object from a corpus of Italian words curated in our lab (Lettieri et al., 2022). The embeddings (128 dimensions, window size 5, CBOW architecture) were averaged across each 2-second interval (matching the fMRI temporal resolution). Principal Component Analysis (PCA) was applied to preserve components explaining at least 75% of the variance, resulting in 49 columns for *101 Dalmatians*, 56 columns for *Forrest Gump*, and 48 columns for *Grand Budapest Hotel*. The resulting columns were convolved with a standard gamma function as a hemodynamic response function.

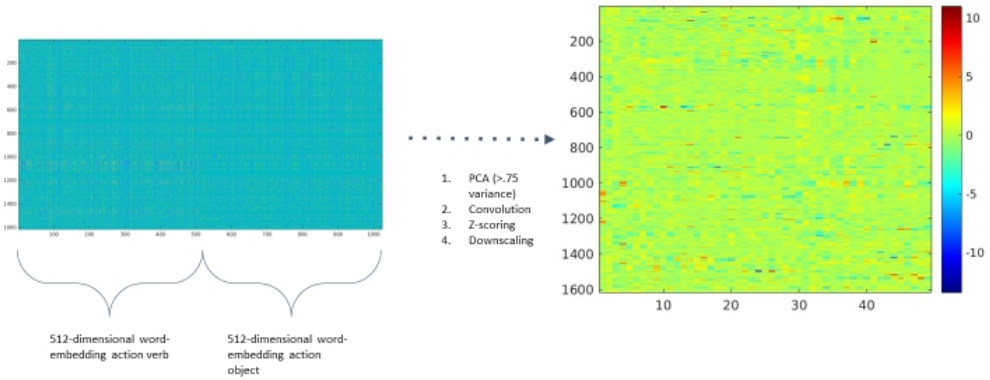

Fig. 8: The original feature matrix, representing the high-level semantic action model (comprising the action verb and object name), underwent a series of preprocessing steps before inclusion in the regression analysis. Dimensionality reduction was first applied to minimize overfitting while preserving key variance in the semantic features. The reduced feature set was then convolved with a hemodynamic response function (HRF), followed by z-score normalization to standardize the features (mean = 0, standard deviation = 1). Finally, the data were downsampled to align with the fMRI temporal resolution (2 seconds).

To account for potential collinearity between high-level action comprehension and motion within the stimulus, we extracted the motion energy from all three films using MATLAB (R2020a, The MathWorks, Natick, MA). After resampling all films to the same size (i.e., 600 x 600 pixels), the film stimuli were analyzed at the highest available temporal resolution (i.e., frame duration, which was 0.04 s at 25 frames per second for "101 Dalmatians," and 0.04 s at 24 frames per second for both "Forrest Gump" and "The

Grand Budapest Hotel"). The low-level motion features were extracted using the GIST model (Oliva & Torralba, 2006).

5.4 Voxel-wise Encoding Modeling

A voxel-wise encoding modeling approach with cross-validation (k=5) was used to predict voxel-specific BOLD responses evoked during film viewing. For each voxel, BOLD responses were modeled as a linear combination of feature spaces created by tagging the films and convolving with a gamma function. Specifically, ordinary least squares (OLS) regression was implemented to estimate the beta weights of the stimulus features.

Using MATLAB's built-in functions (R2022a, The MathWorks, Natick, MA), we applied principal component analysis (PCA) to the predictor matrices and extracted components that explained 99% of the variance. This process was employed to reduce the dimensionality of the feature space (~45 predictors before PCA). Across all three films, the selected components ranged between 20 and 25. Subsequently, in a cross-validation scheme, we estimated for each partition—based solely on the training set's beta parameters—the t-statistic of each voxel, and selected the top 5 features with the best t-statistic for the encoding procedure. The purpose of this step was to further reduce the dimensionality of the feature space, as methodological studies have demonstrated that a voxel cannot represent more than five dimensions. Using this method, we estimated the beta parameters for each voxel, utilizing only

the principal components identified in the feature selection phase.

The data were first divided into five partitions using MATLAB's cvpartition function. Of these, four partitions were used to estimate beta weights. After model estimation, the data excluded from the remaining fifth partition in the loop were used to evaluate model performance, using the r-squared measure. We assessed model performance by computing the relationship between predicted and actual BOLD responses.

For each voxel separately, this process was repeated five times, and model performance was averaged across repetitions.

Specifically, we defined a leave-one-fold-out cross-validation (LOOCV) loop, in which we estimated the t-statistics of each voxel based solely on the training set. For each voxel, a voxel-wise multiple regression was performed to obtain the t-statistics of the beta coefficients for the model descriptors. The t-statistics were computed by averaging the beta coefficients across subjects and dividing them by their standard errors. We decided to limit the maximum number of features predicting a BOLD signal in a voxel to 5, given that the encoding properties of small cortical areas (or voxels) appear to be based on relatively low dimensionality (Diedrichsen et al., 2013; Ahlheim & Love, 2018).

Subsequently, in the final encoding procedure, the selected features were used to predict the test set of a cross-

validation loop (Fig. 10), thus providing an unbiased estimate of model performance. Using the estimated beta weights, the unseen BOLD responses were predicted as follows:

$$Y_{predicted} = X \cdot estimate\, d_{\beta\square} + \square_{\square} \epsilon$$

where Y-predicted is an array of size TR (total number of fMRI volumes from the test set) composed of predicted BOLD signals. The final step involved calculating the amount of variance explained by the model in each voxel. Model performance was evaluated using the R-squared (R^2) statistic.

Two separate encoding models were constructed to assess which voxel could be accurately predicted by the linear combination of categorical or semantic features of the actions.

We calculated the prediction performance map across participants for each model. Group-level analyses were conducted using a non-parametric permutation test to identify voxels that showed prediction performance significantly greater than chance. One thousand null models were generated by randomly shuffling the rows of the design matrix obtained after PCA analysis (Fig. 11). Subsequently, for the null models, we followed the exact same procedure used previously for the original model (feature selection, encoding), ultimately obtaining a measure of the performance (R-squared) of the null models for each voxel (1000 R-squared values per voxel; Fig. 12). From the empirical null distribution of prediction

performance, one-tailed p-values were calculated and adjusted using FDR and FWER corrections. The group-averaged prediction performance maps for each model were thresholded at P-FDR < 0.05 and P-FWER < 0.05 and were projected onto the cortical surface.

5.5 Results

To assess collinearities among the features, we computed a correlation matrix for each film's feature set. Figure X illustrates the correlation matrices for 101 Dalmatians, Forrest Gump, and Grand Budapest Hotel. We observed notable collinearities among certain features, particularly in high-level action dimensions such as sociality, telicity, and iterativity. These correlations suggest shared variance among features that may reflect overlapping cognitive processes during action observation.

To investigate how these collinearities impacted the model's performance, we applied Principal Component Analysis (PCA) to reduce dimensionality and address multicollinearity. The first components captured the majority of the variance, indicating that a reduced set of dimensions sufficiently represented in every movie the

features while minimizing redundancy.

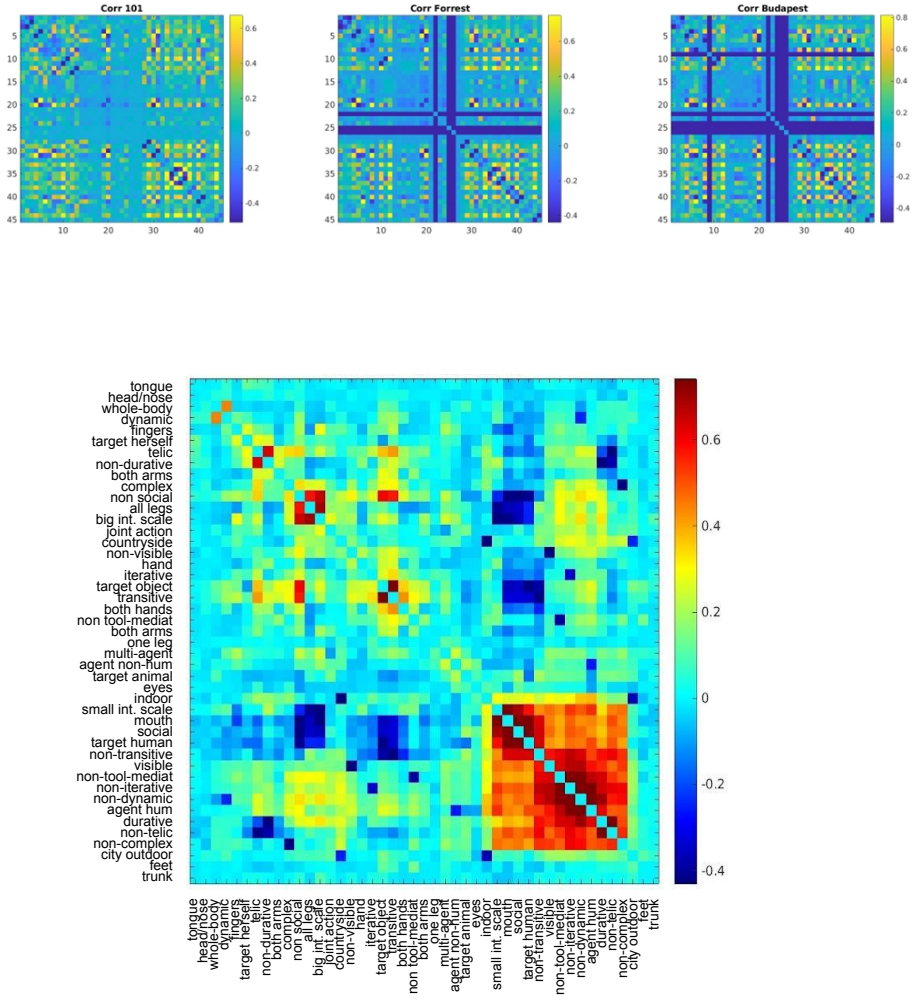

Fig. 9: From the left, three small correlation matrices representing the correlation between the annotated features in each movie, the first

correlation matrix is derived from tagging of the movie "101 Dalmatians", the second from the movie "Forrest Gump" and the third from the movie "Grand Budapest Hotel". The bigger matrix is the average of the three different correlation matrices, underlying once more that regressors correlate with a similar intensity and form clusters of similar shape in the three different movies' ratings.

Our voxel-wise encoding models demonstrated robust prediction performance across all three films. Figure 7 shows the group-averaged R-squared maps for each film, indicating that the models significantly captured the variance in BOLD responses with differences in the performance related to the SNR varying in the different datasets.

For the 101 Dalmatians dataset, the average variance explained was 0.04% (P-FDR < 0.05), predominantly in occipital and posterior temporal regions. Similarly, for Forrest Gump and Grand Budapest Hotel, the models achieved significant R-squared values in occipital and posterior-temporal and temporal cortices for all the three different models. Importantly, despite differences in stimuli and narrative complexity, the models consistently captured variance across these naturalistic settings, supporting their generalizability for further use in an in-depth single movie analysis (chapter 3).

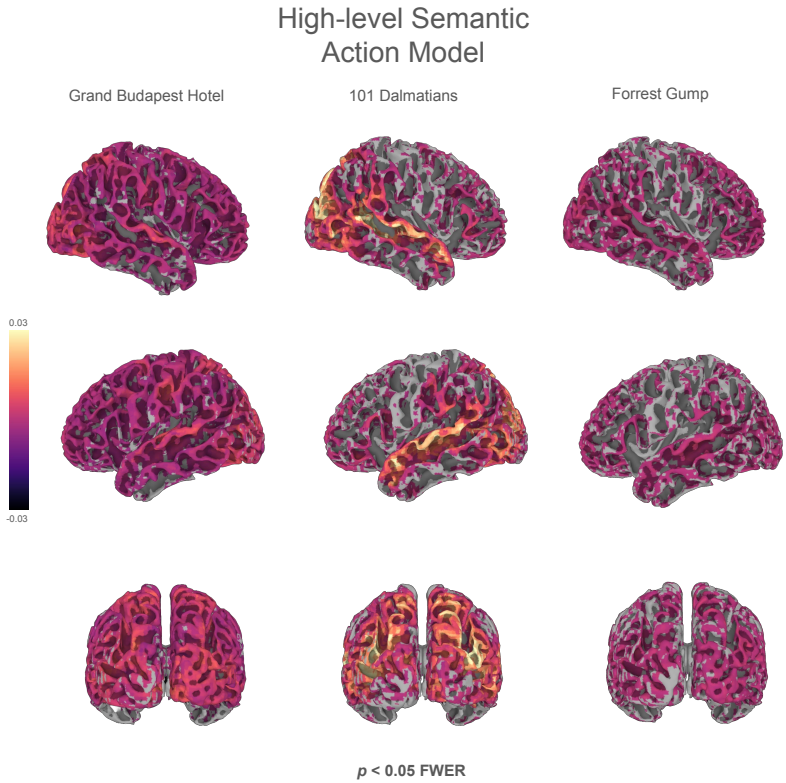

Fig. 10: FWER-corrected R^2 results for the three datasets (corresponding to the three movies) for the high-level semantic action model. Brain surfaces are displayed from three orientations—posterior and lateral views from both hemispheres—illustrating the spatial distribution of model performance across the cortex.

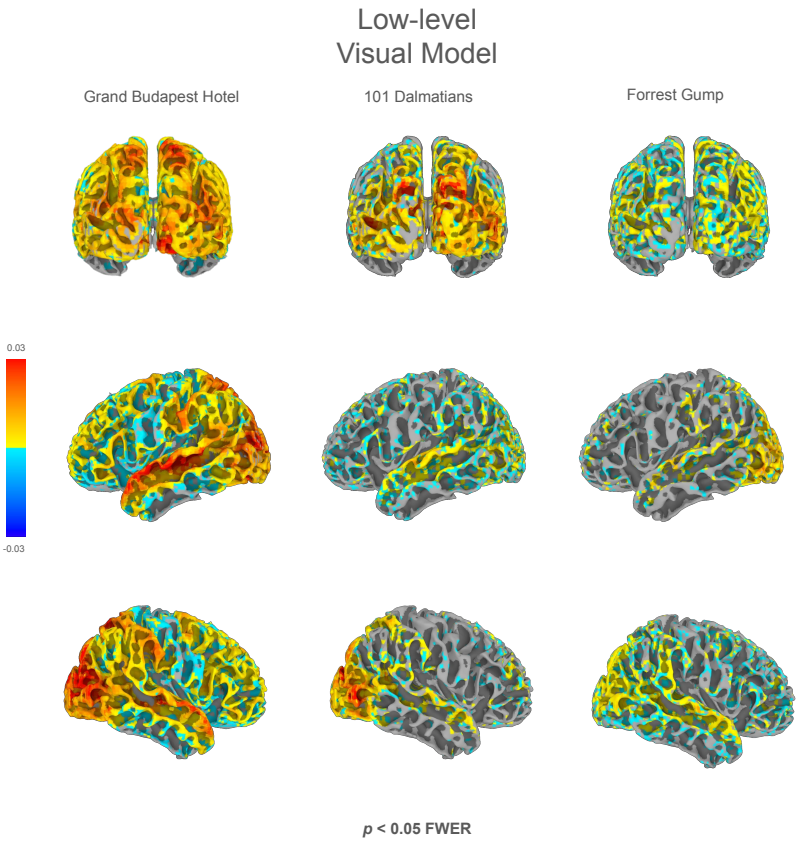

Fig 11: FWER-corrected R^2 results for the three datasets (corresponding to the three movies) for the low-level visual model. Brain surfaces are displayed from three orientations—posterior and lateral views from both hemispheres—illustrating the spatial distribution of model performance across the cortex.

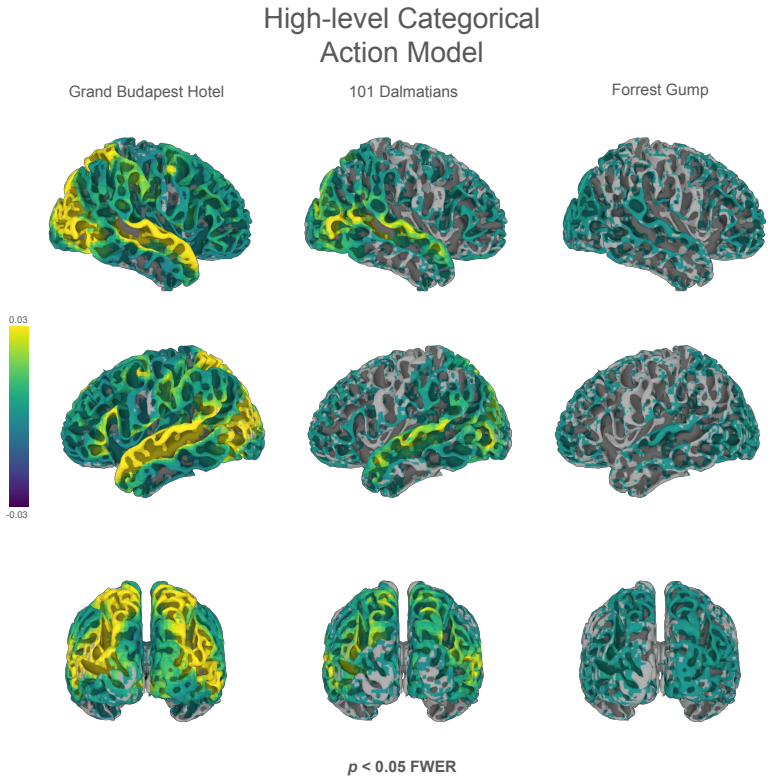

Fig. 12: FWER-corrected R^2 results for the three datasets (corresponding to the three movies) and the high-level categorical action model. Brain surfaces are displayed from three orientations—posterior and lateral views from both hemispheres—illustrating the spatial distribution of model performance across the cortex.

A permutation test was conducted to evaluate the statistical significance of model predictions, with null distributions

generated from shuffled design matrices. Observed R-squared values exceeded null distributions (P-FWER < 0.05), confirming that the models' predictive power was not due to chance.

We identified significant activation within the Action Observation Network (AON), including the superior temporal sulcus (STS), premotor cortex, and inferior parietal lobule, consistent with previous research linking these regions to action understanding and social cognition (Avenanti et al., 2013; Zhou et al., 2023; Mizuguchi et al., 2016). The STS, a core component of the AON, plays a crucial role in processing observed movements and inferring others' intentions (Cortese, 2021). The premotor cortex, particularly in conjunction with the inferior frontal gyrus, is involved in anticipatory simulation of observed actions (Avenanti et al., 2013), while the inferior parietal lobule contributes to mapping observed actions onto motor representations, a key process in action recognition and social interaction (Nelissen et al., 2011).

We also highlight similarities between low-level models (motion energy), semantic models (word embeddings of verbal etiquettes of each action), high-level action models (transitivity, telicity, effectors involved, etc.), that will be subsequently further investigated in chapter 3 of the present thesis.

We observed macro-level similarities between model predictions across the three films. Specifically: the low-level motion energy model predominantly explained variance in early visual areas, including the occipital cortex and motion-sensitive middle temporal (MT) regions, reflecting the model's emphasis on perceptual features.

In contrast, the high-level categorical action model captured variance in superior temporal and TPJ regions, suggesting an involvement of social and cognitive processing in action understanding.

The high-level semantic action model showed activation in lateral temporal cortices and the angular gyrus, indicating its role in processing linguistic and contextual information. These differences highlight the hierarchical nature of action representation, from low-level perceptual encoding to high-level semantic integration.

5.6 Discussion

Our analysis revealed differences in signal-to-noise ratio (SNR) across the three datasets, likely due to variations in scanning parameters and film characteristics. Specifically, the 101 Dalmatians dataset showed lower temporal SNR compared to Forrest Gump and Grand Budapest Hotel, which could be attributed to differences in narrative complexity and auditory-visual integration requirements.

Variability in SNR is known to influence the reliability of BOLD responses, impacting model performance across regions involved in action observation and semantic processing (Dubois & Adolphs, 2016). Moreover, naturalistic stimuli have been shown to elicit variable SNR depending on sensory modalities and scene transitions, which might explain the observed differences (Mandelkow et al., 2017).

These SNR differences underscore the importance of considering stimulus-specific factors when interpreting model performance, as high temporal dynamics and auditory complexity can reduce SNR, particularly in associative and prefrontal cortices (Abreu et al., 2018).

To address multicollinearity and improve model generalizability, we applied Principal Component Analysis (PCA) to reduce dimensionality in the feature space. PCA effectively preserved key variance while mitigating noise, aligning with previous studies showing that dimensionality reduction enhances model interpretability and performance in complex fMRI datasets (Adhikari et al., 2019).

Interestingly, the first few principal components captured the majority of variance in all three films, suggesting a shared low-dimensional structure underlying action representation. This finding supports the hypothesis that action semantics are encoded in compressed neural subspaces, consistent with hierarchical models of action observation (Lv et al., 2015).

Moreover, the application of PCA facilitated cross-dataset comparisons by harmonizing the dimensionality of feature

sets, allowing us to directly compare model performance across films with varying narrative complexity and sensory modalities. This approach enhances reproducibility and robustness, particularly in naturalistic fMRI studies with heterogeneous stimuli (Afshin-Pour et al., 2014).

Our encoding models demonstrated robust performance across all three films, indicating that action representation is consistently captured despite differences in narrative structure, temporal dynamics, and cultural context. This generalizability suggests that the brain encodes action semantics through a hierarchical network involving perceptual, cognitive, and social processing areas, consistent with the Action Observation Network (AON) (Mandelkow et al., 2016).

Notably, model performance was not confined to the 101 Dalmatians dataset, but generalized well to Forrest Gump and Grand Budapest Hotel, demonstrating that our feature set captures invariant aspects of action semantics. This cross-dataset consistency aligns with findings that naturalistic stimuli engage distributed neural networks in a reproducible manner, irrespective of cultural and linguistic differences (Thirion et al., 2014).

The use of multiple films with distinct narrative arcs allowed us to identify common neural coding schemes for action understanding, enhancing the ecological validity of our models. This robustness across stimuli emphasizes the potential of using diverse naturalistic datasets to study complex cognitive functions.

A key strength of this study is the reproducibility of model performance across three independent datasets, underscoring the robustness of our encoding models. Our use of cross-validation and permutation testing ensured that observed prediction accuracy was not due to overfitting, confirming the models' generalizability (Mandelkow et al., 2017).

Reproducibility was particularly high in perceptual regions for low-level motion models, and in the STS and TPJ for high-level categorical and semantic models, consistent with established neural correlates of action observation and social cognition (Lv et al., 2015).

5.7 Conclusions

By demonstrating consistent model performance across films with varying narrative complexity and cultural contexts, we provide compelling evidence for a shared neural architecture supporting action representation. This reproducibility supports the use of naturalistic stimuli for investigating complex cognitive functions, bridging the gap between controlled experiments and real-world cognition (Dubois & Adolphs, 2016).

Chapter 6

Neural representation of action features across sensory modalities: a multimodal fMRI study

6. Introduction

Action processing in neuroscience explores how the brain encodes, processes, and retrieves information related to actions (Grafton & Hamilton, 2007; Kilner, 2011; Giese & Rizzolatti, 2015). In the last decades, particular interest has arisen around the role of the so-called action representation or action observation network (AON) which is believed to play a crucial role in understanding and interpreting observed actions (Decety & Grèzes, 1999; Gallese et al., 2004; Rizzolatti & Craighero, 2004). This network spans a wide extent of the cortical mantle and comprises distant, yet functionally interconnected regions, extending from the inferior ventral and dorsal premotor cortex (vPMC and dPMC) to the bilateral occipitotemporal (LOTC) and the parietal cortices.

Various theoretical frameworks have been proposed in the literature to explain the mechanisms underlying action representation in the human brain, each contributing unique insights into the cognitive and neural mechanisms involved. Evidence suggests that the human brain encodes the specific features contributing to action recognition (e.g., kinematics, object-centered goals, motor acts) through a hierarchical and distributed organization (Grafton & Hamilton, 2007; Kilner, 2011). This existing literature has primarily explored the representation of distinct sets of features - e.g., effector-target interaction (Beurze et al., 2007), target-agent identity (Chambon et al., 2014), social-

emotional valence (De Gelder & Van den Stock, 2011). However, the emergence of high-level conceptual representations of action, and the degree to which feature coding is shared across constituent regions of the AON remain subjects of ongoing debate (Simonelli et al., 2024).
Furthermore, while much of the existing literature has focused on the visuomotor processing of actions, the ability of the AON to encode the properties of actions across different sensory modalities is still of particular interest, as it may reveal fundamental principles of neural representation and cognitive processing (Calvert et al., 2004; Beauchamp, 2005). Recent studies have suggested that the AON exhibits a degree of invariance to the modality of stimulus presentation, indicating that the brain can extract and represent action-related information regardless of whether it is presented visually, auditorily, or through a combination of both (James et al., 2011; Kirsch & Cross, 2015; Copelli et al., 2022). In this regard, evidence exists that congruent auditory and visual stimuli enhance the activation of the AON compared to unimodal presentations (Bischoff et al., 2014) and that the integration of auditory and visual information leads to a more robust representation of actions in the brain (Calvert et al., 2004). Although this invariance is crucial for understanding how individuals interpret and respond to actions in a dynamic environment where sensory inputs can vary widely, the extent to which the encoding of distinct action features is

stable across sensory modalities remains poorly understood.

Here, we assessed *whether* and *to what extent* distinct features of action representation are processed under naturalistic conditions across different sensory modalities. Taking advantage of the rich literature based on behavioral and functional studies (e.g., Van Elk et al., 2014; Kemmerer, 2021; Kabulska & Lingnau, 2023), we purposely defined a comprehensive taxonomic model for action representation that allows a detailed categorization of action-related events even in naturalistic condition, when action features are often interacting and overlapping. Specifically, a set of six domains of action features - *Space, Effector, Agent & Object, Social, Emotion, and Linguistic* - was defined a priori encompassing multiple action features in terms of their context, execution, and underlying dynamics (Fig. 11A and Tab.1). Furthermore, we acquired brain responses using functional Magnetic Resonance Imaging (fMRI) in three groups of participants exposed to the multimodal audiovisual, the visual-only, or the auditory-only versions of the live-action movie *101 Dalmatians* (S. Herek, Great Oaks Entertainment & Walt Disney, 1996). Our aim was twofold: first, we tested the ability of our model to predict fMRI activity by measuring the impact of each action domain through a multi-voxel encoding analysis (Haynes, 2015; Naselaris et al., 2011); second, we

determined whether the representation of action domains was comparable across sensory modalities within the AON. Importantly, here we adopted a variance partitioning approach to disentangle the unique contribution of action features, whereas previous research on action representation has largely focused on individual dimensions in isolation (Kemmerer, 2021), often overlooking potential confounds due to collinearity among action categories.

6.1 Methods

In this study, we investigated the neural encoding of action features across sensory modalities by applying a comprehensive taxonomic model to fMRI data collected during naturalistic stimulation. We analyzed part of a naturalistic fMRI dataset from our previous study (Setti et al., 2023). First, a full-model Canonical Correlation Analysis (CCA) assessed how well the entire action model predicted brain activity, quantifying the proportion of variance explained within cortical regions of interest. Second, a variance partitioning approach measured the unique contribution of each domain. This framework disentangled domain-specific and shared effects in brain representations of action features across experimental conditions. Thirty participants were divided into three groups and exposed to the audiovisual, visual-only, or auditory-only versions of

the same movie. Action events in the stimuli were identified and annotated by three independent raters, according to an ad-hoc action model consisting of six conceptual domains: Space, Effector, Agent & Object, Social, Emotion, and Linguistic. The final model was validated through inter-rater agreement and served as the final input for the multivariate encoding fMRI analysis, offering a detailed and structured representation of action-related information from the movie stimulus as perceived consistently across multiple raters. First, a full-model Canonical Correlation Analysis (CCA) assessed how well the entire action model predicted brain activity, quantifying the proportion of variance explained within cortical regions of interest. Second, a variance partitioning approach measured the unique contribution of each domain. This framework allowed us to disentangle domain-specific and shared effects in brain representations of action features across experimental conditions.

Participants

Thirty healthy and typically developing participants were assigned to one of three experimental conditions consisting of three versions of naturalistic stimulation: audiovisual (AV) (N=10, 35±13 years, 8 females); visual-only (V) (N=10, 37±15 years, 5 females); or auditory-only (A) (N=10, 39±17 years, 7 females). All subjects were right-handed and native Italian speakers. Participants had no history of neurological

or psychiatric conditions, normal hearing, normal or corrected vision, and were drug-free.
Each participant was instructed about the nature of the research and gave written informed consent. The study was approved by the Ethical Committee of the University of Turin (protocol number 195874/2019) and conforms to the Declaration of Helsinki.

Stimuli and Experimental Conditions

Naturalistic stimulation consisted of the presentation of a shortened and edited version (~54 minutes) of the movie "101 Dalmatians" (S. Herek, Great Oaks Entertainment & Walt Disney, 1996), split into six runs (~8 minutes each).
The stimulation was delivered in three different versions: (1) audiovisual (AV), including both visual and auditory features of the movie stimulus; (2) visual (V) features only with no auditory stimulation; (3) auditory (A) features only with no visual stimulation.
In the A and AV versions, an Italian voice-over serving as a narrator's verbal description of the movie was superimposed. The audiodescription included all the aspects of the visual scenery that can not be conveyed through dialogue, music, or environmental sounds. Likewise, in the V and AV versions, Italian subtitles transcribing the whole soundscape (dialogues, narrator's voice-over, environmental sounds) were added. For further

details on the editing procedure, we refer the reader to the original paper (Setti et al., 2023).
In all conditions, participants were instructed to simply enjoy the movie, and in the A condition, participants were instructed to keep their eyes closed.

Taxonomic Action model

Features selection and domains creation

To create a detailed and systematic representation of action-related events, the taxonomic model was constructed through an iterative process that integrated multiple conceptual frameworks ranging from embodied cognition to ecological psychology, and action semantics. We first identified candidate features based on their relevance to action perception and representation in the brain, as supported by prior studies (e.g., Tarhan & Konkle, 2020; Haxby et al., 2020; Wurm et al., 2017; see Table 1 and Fig. 11A). Each feature was operationalized to capture specific aspects of actions.

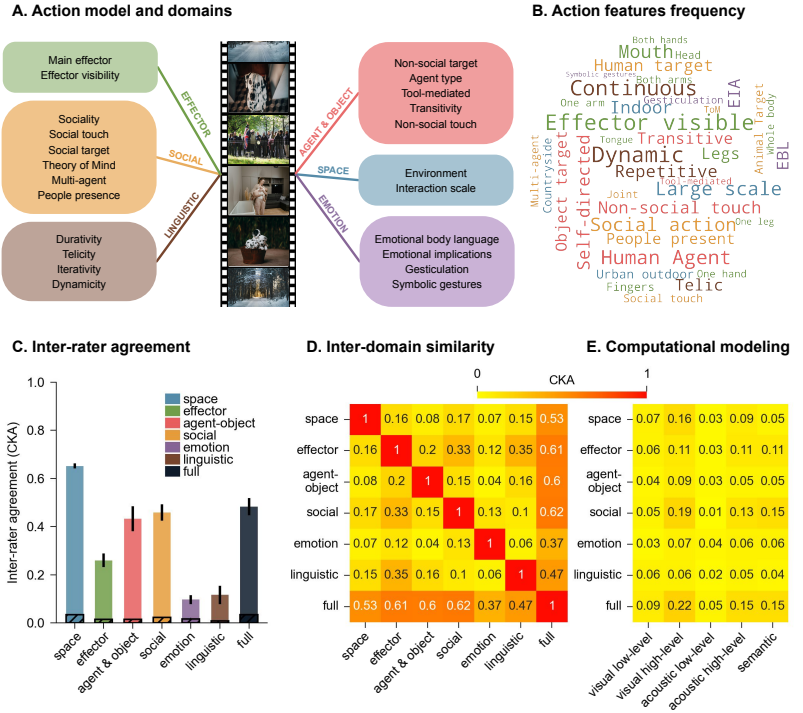

Fig. 13: **Taxonomic Action Model**

(A) Schematic representation of the six action domains (*Space, Effector, Agent & Object, Social, Emotion, and Linguistic*) and corresponding features used to annotate actions of the movie stimulus. (B) Frequency of action features occurrence in the stimulus. Features are color-coded to denote the domain they pertain to, with a larger font size indicating a greater frequency. *EBL*: Emotional Body Language. *EIA*: Emotional Implication of Action. *ToM*: Theory of Mind. (C) Mean Inter-rater agreement across domains and for the full model. Error bars indicate standard deviation. Dashed bars represent the 99th percentile of the

null distribution obtained via permutation testing ($p < 0.01$). (D) Pairwise similarity between domains and the full model, indicating the amount of shared variability. (E) Similarity between model domains and computational models, measuring the extent to which the action model captures low- and high-level perceptual and semantic information in the stimulus. Visual low-level: motion-energy model; visual high-level: ReLU6 layer of the VGG-19 network; acoustic low-level: spectral descriptors; acoustic high-level: ReLU5.1 layer of the VGGish network; semantic: GPT-4 word embeddings.

The selected features were grouped into six domains to minimize overlap and ensure orthogonality: *Space*, *Effector*, *Agent & Object*, *Social*, *Emotion*, and *Linguistic*. This organization reflects both theoretical coherence and empirical evidence suggesting distinct neural representations for these feature sets. These domains were derived from a comprehensive review of existing literature on action representation, combining theoretical insights and practical considerations to ensure their relevance to naturalistic stimuli. Each domain includes features that define specific dimensions of actions, allowing their characterization at multiple levels of granularity.

The *Space* domain refers to the type of environment where an action occurs and the extent of movement required to complete it. It captures the environment's nature - e.g., indoor, urban outdoor, or countryside outdoor (Dima et al., 2022) - and quantifies the interaction scale, defined by the movement scope of the actor within the scene (Tarhan &

Konkle, 2020). The *Effector* domain characterizes the main body part involved in performing the action and whether this effector is visible or hidden from the observer's view during the action (Tarhan & Konkle, 2020). The *Agent & Object* domain addresses the attributes of the action's agent and its non-social target. It distinguishes actions based on whether the agent is human or non-human (Haxby et al., 2020) and whether the action is directed toward an object or the self (Tarhan & Konkle, 2020). This domain also identifies whether the action involves interactions or physical contact with inanimate objects (Masson & Isik, 2021) and whether it requires the use of tools (Gallivan et al., 2013). The *Social* domain includes a set of features related mainly to human or animal interactions (Wurm et al., 2017). In particular, this domain specifies whether the action has a social target (Tarhan & Konkle, 2020), i.e., directed toward another individual (human or animal), or involves interaction or physical contact with another agent (Masson & Isik, 2021). Furthermore, it considers the number of individuals visible in the scene (Dima et al., 2023), whether the action is performed by multiple agents in a coordinated or concurrent fashion (Sebanz et al., 2006; Sinigaglia and Butterfill, 2020), and, finally, whether the action requires inferring another's mental state (Masson & Isik, 2021). The *Emotion* domain describes the emotional dimensions of actions, including the presence of emotional implications of the action (Goldberg et al., 2014), the use of

gesticulation and symbolic gestures (Schippers et al., 2010), and whether the action expresses portrayed emotions through body language, conveyed for instance by movement velocity, acceleration, pitch in the voice or urge (Barliya et al., 2013; Tipper et al., 2015). Lastly, the *Linguistic* domain characterizes actions through features derived from verbal descriptors, focusing on semantic properties that contribute to their representation. These features include durativity (the temporal extent of an action), telicity (whether an action has a clear and intrinsically defined endpoint), dynamicity, (the degree of motion involved in an action - Zarcone & Lenci, 2010), and iterativity (whether the action consists of repetitive cycles - Xu & Huang, 2013). Altogether, these domains provide a structured taxonomic model, grounded in the literature, along with a comprehensive set of descriptors for classifying action-related events.

Domain	**Feature**	**Annotation**	**References**	**Annotation Example**
Space	*Environment*: the type of environment where the action occurs.	• Indoor • Urban Outdoor • Country side Outdoor	Dima et al., 2022	"Dog sniffs the ground": Countryside Outdoor "Girls jogging":

				Urban Outdoor "Man hits friend with a stick": Indoor
	Interaction Scale: the magnitude of movement in the space required to complete the action.	Minimal / Extensive	Tarhan & Konkle, 2020	"Dog sniffs the ground": Minimal "Girls jogging": Extensive "Man hits friend with a stick": Extensive
Effector	*Main Effector*: the primary body part involved in performing the action.	• One hand • Both hands • One arm • Both arms • One leg or paw • All legs	Tarhan & Konkle, 2020	"Dog sniffs the ground": Head or nose "Girls jogging": Whole body "Man hits friend with a stick": Both arms

		or paws ● Mouth ● Head or nose ● Tongue ● Eyes ● Whole body ● Fingers		
	Effector Visibility: whether the effector is visible or hidden from the observer's view.	Visible / Not Visible	Tarhan & Konkle, 2020	"Dog sniffs the ground": Not visible "Girls jogging": Visible "Man hits friend with a stick": Visible
Agent & Object	*Agent Type*: whether the agent performing the action is human or	Human / Non-Human	Haxby et al., 2020	"Dog sniffs the ground": Non-Human "Girls jogging": Human

	non-human.			"Man hits friend with a stick": Human
	Non-social Action Target: whether the action is directed to an object or to the self.	• Object • Self	Tarhan & Konkle, 2020	"Dog sniffs the ground": Object "Girls jogging": Self "Man hits friend with a stick": Social
	Tool-Mediated: whether the action requires the use of tools.	Yes / No	Gallivan et al., 2013	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": Yes
	Transitivity: whether the action involves	Yes / No	Wurm et al., 2017	"Dog sniffs the ground": Yes "Girls jogging":

	interaction with inanimate objects.			No "Man hits friend with a stick": Yes
	Non-Social Touch: whether the agent makes physical contact with an object or self during the action.	Yes / No	Masson & Isik, 2021	"Dog sniffs the ground": Yes "Girls jogging": No "Man hits friend with a stick": Yes
Social	*Sociality*: whether the action involves social interaction with another individual (human or animal).	Social / Non-Social	Wurm et al., 2017	"Dog sniffs the ground": Non-social "Girls jogging": Yes "Man hits friend with a stick": Yes
	Social Action Target:	• Human • Animal	Tarhan & Konkle,	"Dog sniffs the ground": non-

	whether the action is directed to another individual (human or animal).		2020	social target “Girls jogging": Non-social target “Man hits friend with a stick”: Human
	Multi-Agent: whether the action is performed by multiple agents in a coordinated or concurrent manner.	• Single-agent • Multiple agents concurrent • Joint actions	Sebanz et al., 2006; Sinigaglia & Butterfill, 2020	“Dog sniffs the ground”: single-agent “Girls jogging": Multiple agents concurrent “Man hits friend with a stick”: Single-agent
	People Present: whether there are more than 2 people present in the scene.	Yes / No	Dima et al., 2023	“Dog sniffs the ground”: No “Girls jogging": Yes

				"Man hits friend with a stick": No
	Theory of Mind: whether the action involves inferring another individual's mental state.	Yes / No	Masson & Isik, 2021	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": Yes
	Social Touch: whether the agent makes physical contact with an individual (human or animal) during the action.	Yes / No	Masson & Isik, 2021	"Dog sniffs the ground": No "Girls jogging": Yes "Man hits friend with a stick": Yes
Emotion	*Emotional Body Language*: whether the	Yes / No	Barliya et al., 2013; Tipper et al., 2015	"Dog sniffs the ground": No "Girls jogging":

	action expresses emotions through body language.			No "Man hits friend with a stick": Yes
	Emotional Implications: whether the action has emotional significance due to its narrative or context.	Yes / No	Goldberg et al., 2014	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": Yes
	Gesticulation: whether the actor gesticulates during the action.	Yes / No	Schippers et al., 2010	"Dog sniffs the ground": No "Girls jogging": No "Man hits friend with a stick": No
	Symbolic Gesture: whether the	Yes / No	Schippers et al., 2010	"Dog sniffs the ground": No

	action executes culturally specific gestures with symbolic meaning (e.g., handshakes, salutes).			“Girls jogging": No "Man hits friend with a stick”: No
Linguistic	*Iterativity*: whether the action involves repetitive cycles or a single occurrence.	Repetitive / Single	Xu & Huang, 2013	“Dog sniffs the ground”: Repetitive “Girls jogging": Repetitive "Man hits friend with a stick”: Single
	Dynamicity: the degree of motion involved in the action.	Dynamic / Static	Zarcone & Lenci, 2010	“Dog sniffs the ground”: Static “Girls jogging": Dynamic

				"Man hits friend with a stick": Dynamic
	Durativity: whether the action is continuous (e.g., speaking) or instantaneous and bounded (e.g., striking).	Continuous / Bounded	Zarcone & Lenci, 2010	"Dog sniffs the ground": Bounded "Girls jogging": Continuous "Man hits friend with a stick": Bounded
	Telicity: whether the action has a clear and defined endpoint.	Telic / Atelic	Zarcone & Lenci, 2010	"Dog sniffs the ground": Atelic "Girls jogging": Atelic "Man hits friend with a stick": Telic

Table 1. Characterization of the features and domains of the taxonomic action model

Stimulus feature space

Based on our taxonomic model of actions, we created a stimulus feature space to capture detailed action-related aspects present in the movie stimulus. For this purpose, three independent raters (1F, a were recruited and instructed to watch the movie in the AV condition, identify each action in the naturalistic stimulus, and describe it in terms of the taxonomic model. Each rater began with segmenting the stimulus into discrete events. This involved independently detecting individual actions and marking their specific onset and offset times with a temporal resolution of 0.1 seconds. Raters were instructed to annotate as many actions as they could detect, including both primary actions - which occur predominantly in the foreground of a scene - as well as secondary actions, which may (co)-occur in the background. To preserve the ecological validity of the naturalistic stimulation, no specific instructions were given regarding the level of granularity for defining actions. Rather, the raters were free to apply their criteria, ensuring the annotations reflected their subjective interpretations of the action boundaries. For each identified action, the raters then evaluated each of the 22 features of the model by manually annotating the corresponding value. As a result, we obtained a matrix of actions by features for each rater.

Model preprocessing

These matrices were then preprocessed to create a single, time-resolved feature space. First, we converted categorical annotations of specific features into a binary format applying one-hot encoding, obtaining a total of 38 features. We then transformed the datasets from an event-based to a time-resolved format. Specifically, based on each action onset and offset, we mapped the respective features to a continuous time sequence with 0.1 seconds resolution, obtaining a time by features matrix for each rater.

The resulting time-based feature matrices from individual raters were aggregated to create a group-level model. To ensure consistency among raters, a value of 1 was assigned to a feature at each time point only if at least two out of three raters agreed on its presence. The group-level model was then down-sampled to match the temporal resolution of the fMRI data (2 seconds). To account for events shorter than the fMRI sampling interval, each feature was assigned a value of 1 at each time point if it was present in any sample within the 2-second bin. For the multivariate encoding procedure, each column of the model was convolved with a canonical hemodynamic response function (gamma function, duration = 12 seconds, $p = 8.6$, $q = 0.547$) to account for the delayed hemodynamic response of the fMRI signal.

Inter-rater agreement

To verify the consistency of feature annotations, we quantified inter-rater agreement across individual domains and the full model. We employed Centered Kernel Alignment (CKA) (Kornblith et al., 2019) with a debiasing step (Murphy et al., 2024) as a similarity index. CKA measures the shared variance between two multi-dimensional feature spaces, providing a robust method for quantifying similarity between datasets of different size. Debiasing was applied to account for potential confounds arising from significant differences in matrix ranks (i.e., domains' dimensionality ranged from 4 to 12, with the full model comprising 38 descriptors), ensuring a more reliable assessment of similarity.

For each rater pair, we computed CKA scores for individual domains and the full model using their respective downsampled (i.e., 2-seconds temporal window) binary matrices. To establish statistical significance, a permutation test was implemented: a null distribution (n=1,000) was obtained by randomly shuffling chunks of 15 timepoints of one rater's matrix prior to CKA estimation, to account for temporal autocorrelation in the data. The observed similarity scores, averaged across rater pairs, were then compared against the 99^{th} percentile of the null distribution to identify significant agreements ($p<0.01$). This analysis ensured that the group model accurately reflected consistent action feature annotations across raters.

Inter-domain similarity

To evaluate the relationship between domains, we assessed how features from different domains uniquely or redundantly captured aspects of the described actions. This was accomplished by calculating pairwise similarity through CKA between all domains of the downsampled binary group-level model.

Computational modeling

As a final step, we sought to determine the extent to which perceptual or semantic information from the stimulus was captured by the features of the action model. To achieve this, we leveraged a series of computational models to extract diverse representations of the movie stimulus, encompassing: 1) low-level visual descriptors (i.e., motion energy derived from the spatiotemporal integration of Gabor-like filters); 2) high-level visual descriptors (i.e., VGG-19 convolutional deep neural network architecture; Simonyan & Zisserman, 2014); 3) low-level auditory descriptors (spectral and envelope-based properties capturing frequency modulations); 4) high-level auditory descriptors (VGGish deep neural network architecture; Hershey et al., 2017); 5) a high-level semantic representation derived from GPT-4 embeddings (Achiam et al., 2023) extracted from the movie's subtitles.

In detail, low-level visual descriptors were obtained for each two-second segment of the movie videoclip using a comprehensive set of 4,715 motion energy descriptors

derived from space-time Gabor filters. These filters included Gabor wavelets varying in spatial frequencies, orientations, and integrated across three distinct temporal frequencies: 0 Hz (static energy), 2 Hz, and 4 Hz (Nishimoto et al. 2011). Regarding high-level visual descriptors, we utilized the VGG-19 (Simonyan & Zisserman, 2014) convolutional deep neural network architecture to extract a comprehensive set of 4096 visual properties from the central frame of each two-second segment. Specifically, we used the output from ReLU6, the final layer in the stack of convolutional layers, which captures high-level visual information essential for object recognition and image classification. To model low-level auditory characteristics, we extracted for each two-second segment 449 spectral descriptors in the 0-15k Hz frequency range, following the methodology outlined by de Heer and colleagues (2017). Instead, for high-level auditory descriptors, we employed the VGGish (Hershey et al., 2017) model, a convolutional neural network based on the VGGNet architecture and adapted for audio classification. Specifically, we extracted the output of the ReLU5.1 layer, obtaining 4,096 descriptors for each two-second segment. These features captured more abstract properties of the movie's audio track, including contextual information about the soundscape, such as the presence of background noise, music, or speech. Finally, regarding semantic descriptors, we extracted contextual word embeddings

from each subtitle sentence using the pre-trained GPT-4 model (Achiam et al., 2023) (i.e., text-embedding-3-small) via the OpenAI API (https://openai.com/). This process generated a 1,536-dimensional vector for each two-second segment.

We then compared the representations of individual domains as well as the full model to these computational representations using CKA as a similarity measure. This allowed us to evaluate whether specific domains aligned more closely with perceptual or semantic features, offering insights into the nature of the information encoded by each domain.

fMRI data acquisition and preprocessing

Structural and functional data were acquired in the same session with a Philips 3T Ingenia scanner equipped with a 32-channel head coil. For anatomical images, a magnetization-prepared rapid gradient echo sequence use employed (TR=7 ms; TE=3.2 ms; FA=9°; FOV=224 mm, acquisition matrix=224×224; slice thickness=1 mm; voxel size 1×1×1 mm; 156 sagittal slices). Functional images were acquired using gradient recall echo planar imaging (GRE-EPI; TR=2000 ms; TE=30 ms; FA=75°; FOV=240 mm; acquisition matrix (in-plane resolution)=80×80; acquisition slice thickness=3 mm; acquisition voxel size=3×3×3 mm; reconstruction voxel size=3×3×3 mm; 38 sequential axial

ascending slices; total volumes 1,614 for the six runs of the movie).

Acquired fMRI data were preprocessed with the AFNI 17.1.12 software package (Cox, 1996), following the standard steps. First, scanner-related noise was corrected through spike removal (3dDespike). Then, all volumes underwent run-wise temporal realignment (3dTshift) and head motion correction using as base the first run (3dvolreg). Spatial smoothing was then performed with a Gaussian kernel (3dBlurToFWHM, 6mm, full width at half maximum), followed by run-wise percentage normalization. Next, in order to smooth time series and remove unwanted trends and outliers, the normalized runs underwent detrending through Savitzky-Golay filtering (sgolayfilt, polynomial order=3, frame length=200 timepoints) in MATLAB R2019b (MathWorks Inc., Natick, MA, USA) and were concatenated into a single time series.

Afterward, signals related to head motion parameters and framewise displacement were regressed-out through multiple regression analysis (3dDeconvolve). Lastly, the anatomical volumes were aligned to the reference functional image and non-linearly registered (3dQWarp) to the MNI-152 standard space (final voxel size 3 mm iso; Fonov et al., 2009).

fMRI data analysis

To investigate the relationship between action features and brain activity, we employed a variance partitioning framework based on multivariate analysis, using a region of interest (ROI)-based approach.

Parcellation and PCA

First, fMRI data were parcellated into 200 distinct cortical ROIs using Schaefer's atlas (Schaefer et al., 2018). The parcellation, including grey matter cortical voxels only, was applied in the original anatomical space of each subject.

To measure the association between the action model and brain activity we employed a multivariate encoding procedure based on CCA. Since CCA extracts canonical components that maximize the linear association between two matrices, the algorithm limits the dimensionality of the final components to the lower rank of the input matrices. Thus, to accurately reconstruct the action model and account for variability in the number of voxels across ROIs (ranging from 21 to 492; see Appendix Fig. S1), we set the dimensionality of the ROIs to match the rank of the action model. Specifically, for each ROI, principal component analysis (PCA) was performed voxel-wise, retaining a number of components equal to the 38 predictors in the action model. To address cases where an ROI contained fewer than 38 voxels, additional voxels were added by replicating the average time series across existing voxels

within that ROI. This approach ensured consistent dimensionality across ROIs and subjects while preserving the minimum number of components required for subsequent analyses, without excluding brain regions or altering the original informational content of the ROI signals. Importantly, while the amount of variance explained by the resulting 38 components varied across ROIs (see Appendix Fig. S1), it consistently exceeded 76% for all subjects and conditions.

Full-model Canonical Correlation Analysis

Then, to quantify how well the predictors in the action model explained brain activity within each ROI, we employed CCA. CCA is a multivariate statistical technique designed to measure linear relationships between multidimensional variables (Hotelling, 1992; Bilenko & Gallant, 2016). Specifically, CCA finds pairs of canonical components - linear projections of each dataset - that maximize the correlation between the two datasets in a shared canonical space.

In our analysis, CCA was iteratively applied to each ROI, to relate the action model (X, time points × 38 features) to the brain components (Y, time points × 38 components). This process yielded two sets of canonical components U and V, along with the canonical coefficients A and B, for X and Y, respectively. These were used to compute the proportion of variance in Y that could be predicted by X, through the following steps. First, we used U to predict the canonical

projections of V (V_{pred}) by fitting a set of coefficients (W), obtained through least squares regression mapping $U \rightarrow V$ (eq. 1).

$$V_{pred} = U \cdot W \quad (1)$$

Next, we reconstructed predicted brain activity (Y_{pred}) by mapping back the predicted canonical components (V_{pred}) to the original brain activity space, using the canonical coefficients B derived during the CCA process (eq. 2).

$$Y_{pred} = V_{pred} \cdot B \quad (2)$$

Finally, we calculated the coefficient of determination R^2, which quantifies the proportion of variance in the observed Y that was explained by the reconstructed Y (Y_{pred}) (eq. 3,4,5).

$$SS_{residual} = \sum \left(Y - Y_{pred} \right)^2 \quad (3)$$

$$SS_{total} = \sum \left(Y - \underline{Y} \right)^2 \quad (4)$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \quad (5)$$

Here, R^2 represents the proportion of variance in the brain activity of a given ROI that can be explained by the predictors of the action model.

Statistical significance and Permutation Tests

To assess the significance of R^2, we generated null distributions by permuting the temporal order of the fMRI data. For each ROI, the time points of the brain components (in chunks of contiguous 30 seconds, to account for

temporal autocorrelation of the signal) were shuffled randomly across 1,000 iterations while keeping the structure of the action model intact. For each permutation, the CCA procedure was repeated, generating a null distribution of R^2 values for each ROI. Each of these R^2 values was then compared against the null distribution to determine their *p*-value. Once this process was done for all subjects, *p*-values were combined across subjects through Fisher's sum (Fisher, 1970) (eq. 6), yielding a null distribution of combined statistics for each ROI.

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \ln(p_i) \quad (6)$$

The group-level p-value for each ROI was obtained by comparing the observed combined statistics with the relative null distribution, using Pareto tail approximation (Winkler et al., 2016) and t-max correction for multiple comparisons ($p_{corr}<0.05$; Westfall & Young, 1993). Specifically, we obtained the t-max distribution by taking the maximum value across all ROIs, for each permutation. Then, we modelled the right tail (90^{th} percentile) of the t-max distribution by fitting a generalized Pareto distribution (Winkler et al., 2016; Pickands, 1975), from which we derived the *p*-values.

First, this process was done for all subjects of the AV condition. Then the same pipeline was repeated in the A and V conditions, this time including only ROIs that

obtained significant R^2 ($p_{corr}<0.05$) in the AV condition. As a result, we obtained a comprehensive map of the task-related variance in brain activity for each experimental condition, highlighting cortical regions where brain activity was significantly predicted by the full action model.

Variance partitioning

To disentangle the contributions of individual domains in the action model, a variance partitioning approach was employed. This method quantified the unique explanatory power of each domain by measuring the drop in R^2 when its predictors were shuffled.

For each domain, the corresponding predictors were shuffled across timepoints while all other domains were left intact. The partially shuffled action model was then used in the CCA to estimate R^2, reflecting the model's explanatory power after disrupting the specific domain of interest. To obtain a stable estimate of the impact of shuffling, we repeated the process 50 times for each domain and took the average R^2 from the 50 shuffled models. Lastly, the shuffled R^2 was subtracted from the full-model R^2, obtaining a domain-specific R^2. These calculations were performed for each domain, ROI, subject, and condition, generating domain-specific maps of unique variance contributions to brain activity.

We then computed the average across subjects to obtain a group R^2 and, for each domain, the regression coefficient between conditions and the Spearman correlation

coefficients with associated *p*-values were calculated pairwise.

Domain Specificity

To evaluate the spatial specificity of domain contributions to brain activity, we employed a non-parametric rank-based approach. This method tested whether the representation of specific domains within individual ROIs was significantly greater than others, based on the distribution of R^2 values across ROIs.

For each subject and condition, R^2 values were ranked within each domain across all ROIs, identifying regions where the domain exhibited the highest and lowest explanatory power. Converting R^2 values into ranks allowed us to assess the relative, rather than absolute, representation of each domain across ROIs. By focusing on rank-based comparisons, this approach mitigated the influence of global differences in R^2 magnitudes between domains, emphasizing the relative contribution of each domain within a given ROI.

To compare the rank-based representation of domains within each ROI, we applied the Wilcoxon signed-rank test, a non-parametric method for paired data. Pairwise one-tailed tests were conducted for all domain combinations within each ROI and condition. This test determined whether the ranks of one domain consistently exceeded those of another, reflecting greater relative specificity in that region. Then, *p*-values from the three conditions were

aggregated across modalities using Fisher's method (Fisher, 1970). To control for multiple comparisons across ROIs and domain pairs, the aggregated *p*-values were corrected using False Discovery Rate (FDR) control (α=0.05). A significant result for a domain pair within an ROI indicates that one domain consistently ranks higher than the other across subjects. This implies that the ROI exhibits a relatively stronger representation of the more dominant domain, independent of absolute R^2 values. To identify the dominant domain(s) within each ROI we adopted a "maximum-takes-all" procedure. Each domain was compared based on the number of significant pairwise tests it won within the ROI. The domain(s) with the highest number of wins were flagged as dominant. This analysis produced spatial maps of domain-specific representations, highlighting ROIs where specific action-related features were most prominently encoded relative to other brain regions and other features.

6.2 Results

Model validation

Our action model was constructed by selecting a comprehensive set of features from the literature on action representation and asking three raters to identify action events and then manually annotate each feature.

Stimulus segmentation to identify action occurrences was performed independently by each of the three raters, who identified 1,635, 1,304, and 1,025 events, respectively (average event count = 1,321). Event duration ranged from 0.1 to 85.7 seconds and averaged 3.5 seconds. After downsampling to the fMRI temporal resolution and aggregating individual raters' models, the number of timepoints in which at least one action event was present was 1,562 out of 1,614 (Appendix Fig. S2).

The 38 action features comprising the final group model had variable frequency counts (Fig. 11B), ranging from 11 timepoints (0.7% for Symbolic gestures) to 1,530 timepoints (97.95% for Dynamicity; see Appendix Table S1). Frequencies of *Space* features highlighted that the actions in the dataset predominantly occurred in indoor environments and at an extended spatial scale. The *Effector* was almost always visible and most actions were performed using the mouth (e.g., speaking), followed by legs/paws, and fingers. In the *Agent & Object* domain, the most frequent feature was Human Agent, and roughly half

of the time actions were targeted toward the agents themselves and involved contact with an inanimate object. Transitive and Intransitive actions were mostly balanced, while Tool-mediated actions were underrepresented. More than half of the tagged events were considered social, that is, they involved some sort of interaction between agents; of these, most were directed towards humans rather than animals, and only one-fifth were concurrently or jointly performed by multiple agents. An emotional component was frequently present in the stimulus, either conveyed through the actor's body language or implicit in the action itself. Concerning the *Linguistic* domain, the actions in the dataset were predominantly dynamic and continuous rather than static and instantaneous.

To assess the validity of the tagging procedure, the inter-rater agreement was first evaluated by measuring the similarity between individual raters' action models: considering that these were multidimensional binary matrices differing in size, Centered Kernel Alignment (CKA) was chosen as an appropriate metric. CKA is an index of multivariate similarity and ranges from 0 to 1, where 0 means the two matrices are completely dissociable and 1 denotes maximal similarity. The average CKA value across all parings of raters was 0.48 ± 0.04 for the full model, indicating an adequate level of consistency among raters (Fig. 11C). The same analysis was performed also for each domain separately (*Space*: 0.65 ± 0.01; *Effector*: 0.26 ±

0.03; *Agent & Object*: 0.43 ± 0.05; *Social*: 0.46 ± 0.03; *Emotion*: 0.1 ± 0.02; *Linguistic*: 0.12 ± 0.04). We observed the lowest similarities in the *Emotion* domain, which requires the rater to extrapolate and interpret implicit information from the scene, and the *Linguistic* domain, which captures attributes conveyed by the meaning of the action verb. Nonetheless, all CKA scores were statistically significant when compared against the null distribution (p<0.01).

To ensure consistency in the final aggregated model, only annotations agreed upon by at least two out of three raters were retained.

Since multiple features from different domains may co-occur at the same timepoint, we assessed collinearity between domains by measuring the similarity structure between each domain and with the full model (Fig. 11D). CKA scores between domains varied from a minimum of 0.04 to a maximum of 0.35, with the *Effector* and *Emotion* domains exhibiting the highest and lowest similarity, respectively. CKA values between domains and the full model were 0.62 for the *Social* domain; 0.61 for *Effector*; 0.60 for *Agent & Object*; 0.53 for *Space*; 0.37 for *Emotion* and 0.47 for the *Linguistic* domain. Hence, the adopted domains capture sufficiently unique representational content and unbiasedly contribute to the full model.

As our action model was built based on hypothesis-driven taxonomic features derived from the literature, some degree of collinearity with visual, acoustic, and verbal

properties is expected. To measure the impact of these descriptors on both the full model and individual domains, we computed CKA between different computational models and the action model matrices (Fig. 11E). As for the full model, stronger correlations were observed with computational models capturing higher-level perceptual descriptors, both visual and auditory (0.22 and 0.15, respectively), and semantic descriptors (0.15), rather than low-level perceptual characteristics (0.09 for the visual and 0.05 for the auditory model). This effect was also present when considering individual domains, though it was more evident for *Space*, *Effector*, and *Social*, compared to the other domains.

fMRI results in the audiovisual modality

Next, the ability of our action model to predict brain activity was assessed. Since the aim was to measure the association between two sets of multivariate data, i.e., the multidimensional action model and each multi-voxel brain region (i.e., ROI), we used Canonical Correlation Analysis (CCA, Fig. 12A). From CCA, we computed the R^2 as a metric of brain-model association, which expresses the proportion of variance in ROI patterns explained by the action model (see Methods section *Full-model Canonical Correlation Analysis*). Importantly, to quantify the unique contribution of each domain, we also performed variance

partitioning, thus obtaining a domain-specific R^2 for each ROI.

In the AV modality, we identified an extended network of occipital, temporal, posterior parietal, and prefrontal areas (127 ROIs in total out of 200 cortical parcels defined in Schaefer et al., 2018), in which the full action model could significantly explain neural activation (average R^2 = 0.061 ±0.017, p<0.05, t-max correction for multiple comparisons). Posterior temporal and lateral occipital areas showed the highest R^2 values (Fig. 12B; peak R^2=0.129 in right posterior superior temporal sulcus -STS- at x=56, y=-51, z=14, LPI), consistent with previous literature denoting these brain regions as critical for action recognition (Rizzolatti & Craighero, 2004; Orban et al., 2021; Wurm & Caramazza, 2022; Karakose-Akbiyik et al., 2023).

As for the contribution of individual domains (Fig. 12C), the *Effector* domain had on average the highest explanatory power (0.0086 ± 0.0031, 14% of full model R^2), followed by the *Social* domain (0.0059 ± 0.002, 9.6% of full model R^2), *Agent & Object* (0.0041 ± 0.0017, 6.7% of full model R^2), *Linguistic* (0.0034 ± 0.0019, 5.5% of full model R^2), *Space* (0.0034 ± 0.0014, 5.5% of full model R^2), and *Emotion* (0.0028 ± 0.0011, 4.6% of full model R^2). Therefore, 54.1% of the brain variance explained by the full model was not attributable to the unique contributions of single domains, but rather to shared variance across combinations of multiple domains.

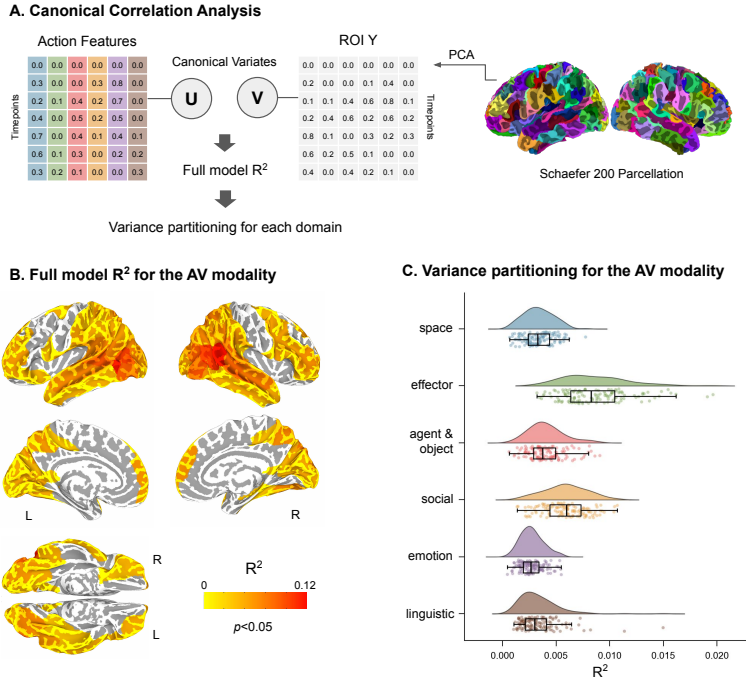

Fig. 14: Brain encoding of action features in the audiovisual modality.

(A) Analysis pipeline for assessing the relationship between action features and brain activity. fMRI data were parcellated into 200 cortical regions of interest (ROIs) using Schaefer's atlas. Principal Component Analysis (PCA) was applied to voxels time series within each ROI and Canonical Correlation Analysis (CCA) was then used to map the action model onto the fMRI data, identifying canonical components that maximized the correlation between the two datasets. The proportion of variance in brain activity explained by the action model was quantified using R^2, which reflects model fit within each reduced ROI. The same process was repeated in a variance partitioning framework to isolate

contributions of individual domains. (B) Brain regions showing significant (pcorr < 0.05, t-max correction) R^2 values for the full action model in the audiovisual (AV) modality. L: left, R: right. (C) Unique contributions of individual domains. Individual points reflect domain-specific R^2 values for each ROI.

Impact of sensory modality

To test the degree of invariance of identified areas to sensory modality, we replicated the same analysis in the V and A modalities, quantifying the amount of task-related variance in brain activity when subjects were presented with visual-only or auditory-only versions of the stimulus.
In the V modality, brain activity was significantly predicted by the full model in 80 ROIs out of 127 (p<0.05, t-max correction, average R^2 = 0.06 ± 0.015, highest R^2 = 0.109 in right LOTC, at x=47, y=-75, z=-1, LPI), while in the A modality, only 12 ROIs survived statistical thresholding (p<0.05, t-max correction, average R^2 = 0.065 ± 0.005; peaking in middle right STS at x=61, y=-30, z=-4, LPI, R^2 = 0.073), reflecting a partial loss of information when one modality is neglected. The average full model R^2 across all tested ROIs (N=127) for the V and A modalities was 0.055 ± 0.014 and 0.05 ± 0.009, respectively (Fig. 12B). The decrease in explanatory power as compared to the multisensory stimulation was mainly observed in frontal and primary visual areas for the V modality. The statistically significant ROIs in the A stimulation all pertained to the superior and middle temporal cortices (Fig. 12A).

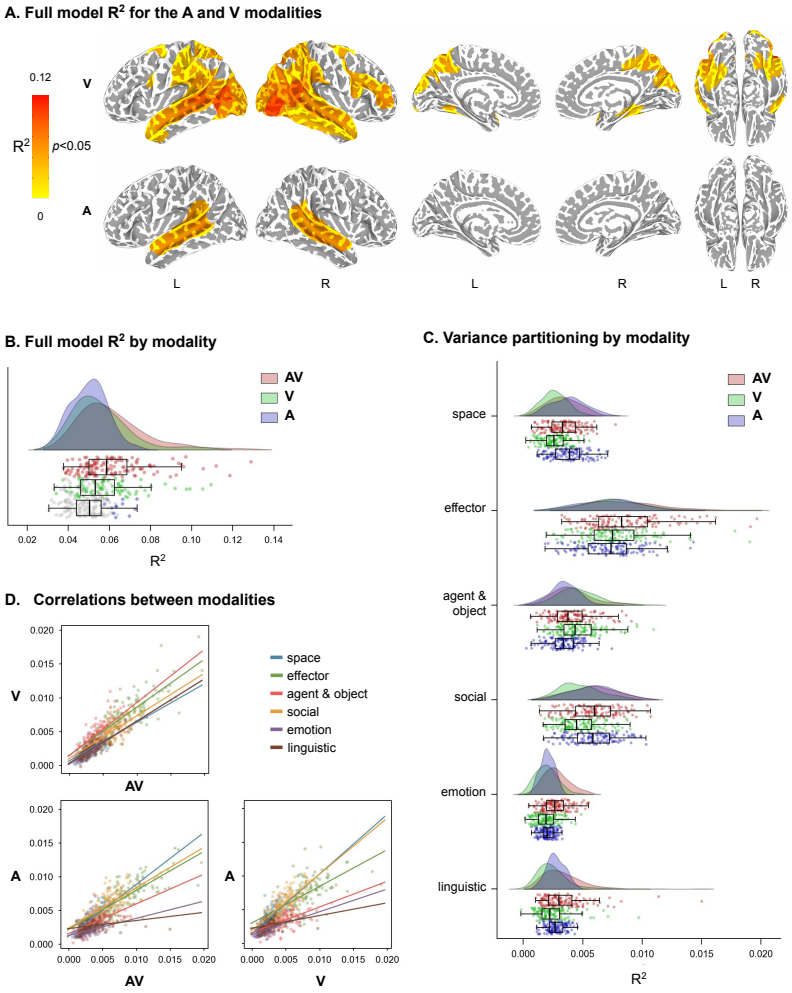

Fig. 15: Impact of sensory modality on action representation in the brain.

(A) Brain regions showing significant (*pcorr* < 0.05, t-max correction) R^2 values for the full action model in the visual (V, top) and auditory (A, bottom) modalities. *L*: left, *R*: right. (B) Distributions of full model R^2 values across all tested ROIs for the three conditions (AV, V, A). Each point represents an ROI full model R2 value; ROIs not surviving statistical thresholding in each modality are colored in grey. (C) Unique contributions of individual domains across sensory modalities. (D) Domains correlations between different modalities. Each point represents an individual ROI domain-specific R^2. Solid lines represent the least square fit between modalities pairings for each domain.

Variance partitioning for the V and A modalities (Fig. 13C) revealed a similar trend in individual domain contributions as for the AV stimulation. For all three conditions, the *Effector* domain explained the highest amount of variance in brain activity (V: 0.0085 ± 0.0029, 14.2% of full model R^2; A: 0.0113 ± 0.0016, 17.4% of full model R^2), followed by the *Social* domain (V: 0.0051 ± 0.0017, 8.5% of full model R^2; A: 0.0077 ± 0.0014, 11.9% of full model R^2). Likewise, the *Emotion* domain had the least explanatory power for all three modalities (V: 0.002 ± 0.0009, 3.3% of full model R^2; A: 0.0022 ± 0.0003, 3.4% of full model R^2).

In the V modality, the domain-specific R^2 was 0.005 ± 0.0017 (8.3% of full model R^2) for *Agent & Object*, R^2 = 0.0028 ± 0.0011 (4.7% of full model R^2) for *Space*, and R^2 = 0.0027 ± 0.0016 (4.5% of full model R^2) for the *Linguistic* domain. In the A modality, the remaining domains R^2 were: 0.0055 ± 0.0009 (8.5% of full model R^2) for *Space*, 0.0035 ± 0.0007

(5.4% of full model R^2) for *Agent & Object*, *and* 0.003 ± 0.0005 (4.6% of full model R^2) for the *Linguistic* domain. Similarly to the AV modality, a substantial portion of brain variance explained by the full model (56.4% for V and 48.8% for A) was shared across combinations of multiple domains.

The invariance of action information organization across sensory modalities was further explored by directly comparing domain-specific R^2 distributions between modalities. For each domain, Spearman's rho was computed between ROIs R^2 of each pair of sensory modalities. The correlation was significant for all domains and modalities combinations (Table 2), indicating that information was similarly organized across the cortex for all three sensory inputs (Fig. 12D). The AV and V modalities had the most similar spatial distribution, with all domains having a correlation coefficient higher than 0.7. For the A modality, the most dissimilar domains were *Emotion*, whose features were predominantly expressed through visual attributes, and *Linguistic*.

	Space	*Effector*	*Agent & Object*	*Social*	*Emotion*	*Linguistic*
AV - V	0.76**	0.77**	0.78**	0.74**	0.77**	0.82**
AV - A	0.78**	0.75**	0.64**	0.63**	0.5**	0.47**
V - A	0.69**	0.67**	0.6**	0.73**	0.51**	0.49**

Table 2. Spearman correlations between modalities. ** p<0.001

Overall, the distribution of explained variance across cortical regions and modalities indicates that the multidimensional framework of action representation is stable across sensory inputs.

Domain Specificity

We observed that ROIs showing a high fit to the full model have high R^2 values for all domains, leading to overlapping spatial maps across domains. Therefore, to assess each domain's spatial specificity, we employed a non-parametric rank-based approach: rather than comparing the absolute magnitude of domain-specific R^2, we focused on the relative contribution of each domain across regions (see Methods section *Domain Specificity*, Appendix Fig. S3-5 and Appendix Table S2). Data from all conditions were aggregated to evaluate tuning across sensory modalities. From this analysis, we obtained spatial maps of domain-

specific representations, highlighting ROIs predominantly tuned to a domain versus the others independently of the sensory modality (Fig. 13). In the following paragraph, we report and briefly discuss the results for each domain, which expand on the current knowledge of domain specificity in AON.

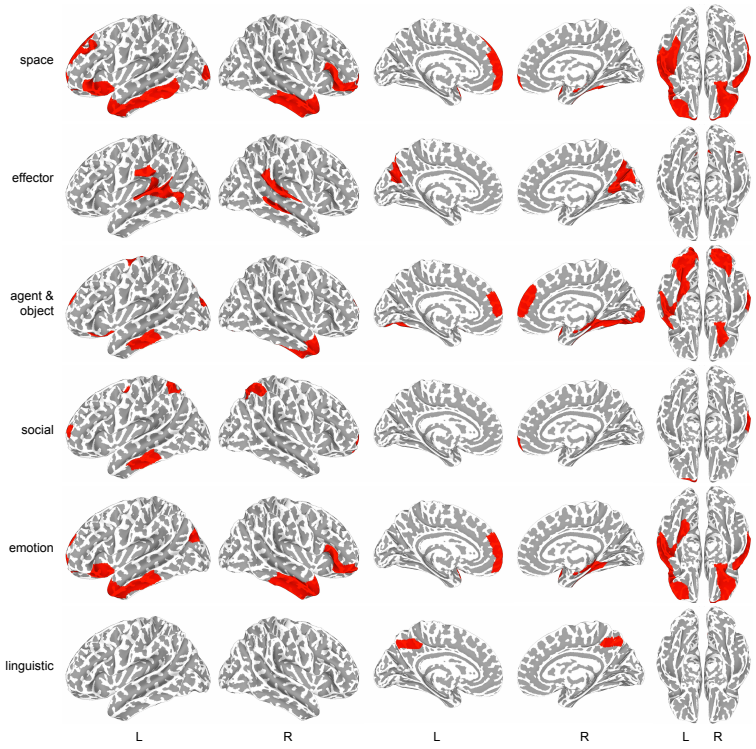

Fig. 16: Spatial domain specificity in cortical action representations.

Cortical regions predominantly tuned to each action domain. ROIs were identified using a rank-based approach and aggregating results across modalities. *L*: left, *R*: right.

Space

This domain describes the context in which the action takes place and its spatial extent. Research on scene representation has highlighted the role of the retrospenial cortex (RSC) and the parahippocampal cortex (PHC) in processing indoor vs outdoor and natural vs urban environments (Henderson et al., 2007; Henderson et al., 2011; Stobbe et al., 2024). In the context of action representation, the setting of an action is often conflated with other perceptual features (Masson & Isik, 2021; Dima et al., 2022), contributing to brain activity in early visual areas and fusiform gyrus (Masson & Isik, 2021). A previous study investigating the representation of actions spatial scale found regions tuned to fine-scale actions in the intraparietal sulcus (IPS) extending to the transverse occipital sulcus (TOS), LOTC, and PHC; areas tuned to intermediate/near space and large interaction scale were identified in posterior portions of middle (MTG) and inferior temporal cortices (ITG), medial parietal cortex and RSC (Tarhan & Konkle, 2020).

Consistently, we showed tuning to the *Space* domain in inferior and middle temporal cortices. Other areas were identified in bilateral orbitofrontal cortex (OFC), pars

orbitalis of the inferior frontal gyrus (IFG*orb*), left pars triangularis of the inferior frontal gyrus (IFG*tri*), left anterior medial prefrontal cortex (mPFC), and left dorsomedial prefrontal cortex (dmPFC).

Effector

Tarhan & Konkle (2020) underlined the importance of the effector in action representation and found that effector-related features modulate responses in occipital, temporal, and parietal cortices, with dorsal areas preferentially encoding effector visibility rather than identity.

Here, effector-specific ROIs were identified in the left parietal operculum (OP), which is considered part of the extended human Mirror Neuron System (Bonini, 2017), and parieto-occipital sulcus (POS). We also found tuning in areas associated with auditory and language processing (i.e., STS and MTG); this response was likely driven by the mouth effector, engaged during speaking.

Agent & Object

This domain predominantly identifies actions involving or directed towards non-social targets, such as transitive actions. Observation of object manipulation has been shown to activate parietal regions (Gallivan & Culham, 2015; Urgen & Orban, 2021). Moreover, ventral LOTC represents actions based on transitivity (Wurm et al., 2017; Tucciarelli et al., 2019).

Here, we found that *Agent & Object* features were preferentially encoded in ventral occipito-temporal areas, part of the ventral pathway for object recognition (Goodale & Milner, 1992). Other identified areas included: bilateral anterior mPFC, right temporal pole (TP), left dorsal premotor area (PMd), left inferior temporal sulcus (ITS), and right primary visual cortex (V1).

Social

Studies investigating the social dimension of actions have typically implicated STS (Isik et al., 2017; Masson et al., 2018; Tarhan & Konkle, 2020), the temporoparietal junction (TPJ; Arioli & Canessa, 2019; Arioli et al., 2021; Masson & Isik, 2021), and mPFC (Wagner et al., 2016; Masson & Isik, 2021). Furthermore, some evidence suggests that sociality is encoded in the dorsal parts of LOTC (Wurm et al., 2017; Han et al., 2024), while the superior parietal lobule (SPL) has been shown to respond to observed socio-affective touch (Masson et al., 2018).

Observation of joint actions recruits areas in the temporal poles, STS/MTG, precuneus, and TPJ (Leube et al., 2012; Eskenazi et al., 2015); these activations overlap with areas comprising the Theory of Mind network (Schurz et al., 2014; Arioli & Canessa, 2019; Arioli et al., 2021), underlying mentalizing processes necessary for inferring others' state of mind.

Adding to the existing literature, we found tuning to the *Social* domain in left ITS, frontal pole (FP), left frontal eye

field (FEF), and SPL and IPS. IPS and FEF are part of the Dorsal Attention Network, which has been shown to encode embodied aspects of social cognition (Lahnakoski et al., 2012). Moreover, a recent meta-analysis (Zhao et al., 2024) suggested that SPL may be part of an action observation pathway dedicated to the processing of non-social actions, therefore capturing modulation of social features.

Emotion

Previous research on naturalistic viewing indicates that affective features, including valence and arousal, partly predict responses in STS, TPJ, anterior temporal lobe (ATL), and mPFC (Masson & Isik; 2021). Moreover, lateral OFC has been shown to map emotional content regardless of the sensory modality (Lettieri et al., 2024). Emotional information conveyed specifically through body language activates ATL (Tipper et al., 2015) and other areas specialized in emotional processing, such as amygdala, OFC, anterior cingulate cortex, and anterior insula (de Gelder, 2006; Sokolov et al., 2020).

Other features included in the domain were Gesticulation and Symbolic gestures, which have been shown to evoke responses in frontal (e.g., IFG) and temporal (e.g., MTG, STG extending to supramarginal gyrus) cortices (Andric et al., 2013; Möttönen et al., 2016; Papeo et al., 2019).

Consistent with these findings, our results revealed areas exhibiting specificity to the *Emotion* domain in OFC, left anterior mPFC, bilateral ATL, and right IFG*orb* and IFG*tr*.

Linguistic

The literature on the neural correlates of action verb processing has highlighted that Telicity modulates activity in the left posterior MTG (Romagno et al., 2012) and bilateral precuneus (Malaia & Newman, 2014), while Iterativity engages IPS (Lai et al., 2023), and Dynamicity MTG/STS (Peelen et al., 2012).

In the present study, only the precuneus showed specific tuning to the *Linguistic* domain.

6.3 Discussion

This study assessed a comprehensive taxonomic model of action representation, designed to capture six key domains of action descriptors (i.e., *Space*, *Effector*, *Agent & Object*, *Social*, *Emotion*, and *Linguistic*) in naturalistic contexts. Using fMRI data collected across different sensory modalities (i.e., audiovisual, visual-only, and auditory-only presentations of the same movie), we demonstrated that this action feature-based model effectively predicts a significant portion of brain activity. Furthermore, the results showed that domain representations were consistently maintained across sensory modalities, both in

their cortical spatial distribution and in their relative contribution to explaining brain activity.

Naturalistic approach to model design and stimulation

The taxonomic model of naturalistic action representation was developed through a multi-step process. First, we identified and categorized critical action features into six conceptual domains, drawing on existing literature on action representation. Next, multiple raters were engaged to extract and annotate action-related information from the movie stimulus. This step comprised two main stages: event segmentation, during which raters were asked to detect discrete actions from the continuous movie stream and define their temporal locations; and feature tagging, i.e, raters described each identified action according to the features outlined in the taxonomic model.

For the segmentation step, raters were instructed to freely identify the temporal boundaries of distinct actions. Event segmentation research suggests that actions follow a partonomic organization: they can be defined at multiple levels of granularity, ranging from fine-grain atomic movements to broader, coarse-grain events based on actor intentions and goals (Shipley & Zacks, 2008; Gu et al., 2018). We deliberately avoided imposing strict constraints or providing explicit instructions about the grain size of actions. This approach aligns with findings indicating that

partitioning a continuous stream of events into discrete units is a subjective and non-trivial process for human observers. In fact, event segmentation emerges from the interplay of perceptual cues and inferential processes involving prior knowledge, observer goals, and task instructions (Shipley & Zacks, 2008). Notably, instructions can modulate segmentation density by guiding observers toward coarser or finer-grain segmentation criteria (Zacks et al., 2001). By refraining from providing specific instructions, raters were allowed to adhere to their own criteria for defining action, fostering the emergence of a more ecologically valid and naturalistic segmentation unit. At the same time, the absence of a common criteria inevitably led to greater variability across raters, as reflected in differences in the number of identified events. Nevertheless, overall inter-rater agreement on the annotated features was found to be robust, and, by including only annotations that were common to at least two out of three raters, we constructed an average group model that effectively captures a structured representation of action domains. Importantly, the domains were designed to be orthogonal, and analyses of inter-domain correlations confirmed their minimal overlap and independent contributions to the final model.

Notably, most of the features incorporated in our model have previously been explored within the literature, often with neural correlates identified for each of them (Grafton

& Hamilton, 2007; Van Elk et al., 2014; Kemmerer, 2021). However, traditional paradigms usually focus on isolated actions, employing highly-controlled single clips with low ecological validity. These approaches often overlook the complex interplay and contextual dependencies between multiple actions. Therefore, to enhance ecological validity, we evaluated our taxonomic model using a naturalistic stimulation paradigm which enables a more comprehensive understanding of how the brain interprets and processes actions within the dynamic context of everyday life, thus ultimately improving the reliability and generalizability of research findings (Hasson et al., 2004).

Model goodness of fit and invariance to sensory input

Testing whether our model could predict fMRI data during AV stimulation, we found that most of the cortex encodes action features. Statistically significant ROIs spanned prefrontal, temporal, parietal, and temporal cortices, substantially overlapping with the extended AON. On average, the full model explained around 6% of the variance in neural activation, reaching a maximum of 13% in posterior temporal and lateral occipital areas. While this effect size is consistent with previous literature (Huth et al., 2012; Cichy et al., 2021; Lahner et al., 2024), a considerable proportion of variance in neural activity remained unexplained. In the present study, we employed a high-

level taxonomic action model which showed higher collinearity with high-level perceptual and semantic rather than low-level properties. Yet, the AON might be preferentially tuned to lower-level perceptual information. In such a case, our results would be driven by the marginal collinearities between our model and perceptual lower-level features. The existing literature, however, appears to confute this possibility: studies comparing or estimating unique contributions of perceptual vs. higher-level models demonstrated an advantage of the latter in explaining brain activations (Urgen et al., 2019; Masson & Isik, 2021; Dima et al., 2022). These findings are paralleled by behavioral investigations suggesting that action categorization and similarity judgments are driven by higher-level dimensions such as sociality and action goals rather than visual characteristics and kinematics (Tarhan et al., 2021; Kabulska & Lingnau, 2023; Dima et al., 2024).

The action model's ability to explain brain activity independently of stimulation modality was also assessed. Results in the video-only modality were similar to the multimodal stimulation, as the model significantly predicted brain activity in most temporal, occipital, and parietal areas. On the other hand, auditory-only stimulation resulted in a loss of explanatory power across the cortex, with only a few areas in the temporal cortex surviving statistical thresholding for multiple comparisons. It should be noted that, for all three modalities, we tested

the same action model, which was developed by tagging events in the audio-video version of the stimulus. While raters were instructed to tag every action independently of the sensory modality of stimulus presentation, the action descriptors inevitably encompass multisensory information. Indeed, the effect size in the multisensory modality was greater than in the visual-only and auditory-only modalities. The increased responsivity of the AON to multimodal audio-video stimulation as compared to unimodal conditions has already been proved (Kaplan & Iacoboni, 2007; McGarry et al., 2012; Bischoff et al., 2014), with evidence pointing toward an advantage of the visual over the auditory modality (Copelli et al., 2022). Equally, while action processing as a whole results to be modality-independent (Ricciardi et al., 2009; 2013), whether the representation of individual action features is due to concurrent additive modality-dependent responses or to a truly shared, modality-independent coding is still uncertain (Alaerts et al., 2009; Rezk et al., 2020).

Domains representation across modalities

Previous investigations into action representation have usually adopted univariate approaches, testing action properties in isolation (Kemmerer, 2021). Therefore, the potential interactions and the complex manner in which such properties may jointly contribute to brain activity have been overlooked. To address this limitation, we

employed a variance partitioning approach, which enabled us to disentangle the unique contribution of each action domain, while accounting for collinearity between features. This approach yielded equivalent results across sensory modalities, both in terms of the magnitude of domain-specific R^2 and their cortical distribution. Specifically, we observed a stability of domain contributions across brain regions and sensory modalities.

In this regard, the *Effector* domain explained the highest amount of variance in all modalities, highlighting the importance of this feature in action perception (Beurze et al., 2007; Abdollahi et al., 2013; Tarhan & Konkle, 2020). The *Social* domain was the second most contributing domain, consistent with previous accounts proposing the socio-affective dimension as an organizing principle for action representation (Wurm et al., 2017; Tarhan & Konkle, 2020; Dima et al., 2022; Kabulska & Lingnau, 2023; Zhao et al., 2024; Han et al., 2024). On the other hand, we found that the *Emotion* domain had the least impact on the full model across all modalities, which may be due to the higher level of complexity of our *Emotion* features with respect to previous studies (e.g., emotional information relayed by the narrative rather than scene valence; Masson & Isik, 2021; Kabulska & Lingnau, 2023). Altogether, these findings suggest that the representational organization of action features remains consistent despite differences in sensory input.

The absolute magnitude of domain-specific R^2 revealed an imbalance toward the *Effector* domain, which was consistent across ROIs. Therefore, to highlight potential preferences in feature tuning, we employed a rank-based approach assessing domains' relative specificity within ROIs. We found weak evidence of domain specificity, with multiple regions preferentially recruited by more than one domain. Nonetheless, as discussed above in the Results section, the spatial distribution of domain preference is consistent with previous literature. Indeed, the lack of specificity of the AON is corroborated by lesion studies, which provide sparse evidence of associations between these specific action features and brain regions (Kalénine et al., 2010; Kalénine et al., 2013; Bonivento et al., 2014; Urgesi et al., 2014).

In the audio-video, visual-only, and auditory-only conditions, single domain contributions added up to 54%, 57%, and 49% of the full model R^2, respectively. Thus, around half of the neural activity explained by the action model can be attributed to variance shared across domains rather than to their unique contributions. Again, the lack of domain specificity in our results is consistent with this notion. Because of computational constraints, it was not feasible to estimate common sources of variance between all possible domain combinations. Interestingly, our action features were grouped into conceptually meaningful domains so that collinearities between domains were

minimized. Despite model orthogonality, we found that shared information across domains accounted for the majority of explained brain activity, suggesting that our taxonomic categorization may not reflect how the brain organizes action-related information. Rather than relying on theoretically motivated stimulus descriptors, recent works (Zheng et al., 2019; Hebart et al., 2020) have employed data-driven approaches to stimulus feature extraction in object coding, leveraging human similarity judgments of stimuli. Such approaches may provide a better approximation of the dimensions governing mental representations (Hebart et al., 2020) and may reveal critical features that previous literature might have missed (Kabulska & Lingnau, 2023), thus increasing the ecological validity of the investigated dimensions. Implementations of such approaches in the framework of action representation may identify action features tailored to describe naturalistic stimuli and explain their neural representations.

6.4 Conclusions

In conclusion, this study has extended the understanding of action representation in the human brain across varying sensory modalities, by taking advantage of a naturalistic stimulation and a comprehensive taxonomic model of action features. The findings reveal that the AON is capable of robustly encoding action-related information across

sensory modalities, highlighting the modality-independent nature of action processing. Furthermore, by means of variance partitioning, we demonstrated that specific action domains contribute distinctly to neural representation, with some domains, such as *Effector* and *Social*, showing stronger influence than others. Importantly, the work underscores the value of a naturalistic approach to neuroscience research, providing a more ecologically valid insight into how actions are processed in real-world settings. This study not only enhances our understanding of neural mechanisms underlying action perception but also opens avenues for further research into how actions are integrated across different sensory inputs to form a cohesive understanding of others' behaviors.

Conclusions

This thesis set out to bridge longstanding theoretical debates with empirical investigation in order to advance our understanding of how human actions are represented in the brain. Drawing from interdisciplinary perspectives—including philosophy, cognitive neuroscience, and computational modeling—the work has developed a comprehensive framework that reconsiders traditional action taxonomies and illuminates the neural underpinnings of action comprehension.

At the theoretical level, the thesis has proposed a deflationary pluralism approach to action taxonomy. Rather than adhering to a single, narrowly defined model, the work emphasizes the necessity of embracing multiple representational formats—sensorimotor, perceptual, and semantic—to capture the multifaceted nature of actions. By delineating how each format contributes uniquely to our understanding of actions, the framework challenges conventional theories that reduce actions to mere motor acts or abstract intentions. Instead, it posits that the richness of human behavior lies in the dynamic interplay between precise motor control, sensory integration, and higher-order conceptual processing. This theoretical contribution not only redefines the "mark" of action but also provides a structured language for future

investigations, thereby encouraging a more nuanced exploration of phenomena such as goal representation, hierarchical structure, and the interface problem.

Empirically, the study leveraged naturalistic fMRI paradigms to validate these theoretical insights. The analysis of multiple film datasets—ranging from narratives in 101 Dalmatians to the complex storytelling in Forrest Gump and the distinctive stylistic features of Grand Budapest Hotel—revealed that neural representations of action are remarkably stable across varying sensory modalities and narrative contexts. Voxel-wise encoding models demonstrated that key action features, when appropriately abstracted, reliably predict brain activity patterns. This consistency across datasets underscores the robustness of the proposed action models and supports the idea that the brain organizes actions in a high-dimensional semantic space, where both low-level sensory inputs and higher-order conceptual cues are integrated to inform prediction and comprehension.

The methodological innovations introduced in this work have further enriched our understanding of action representation. By combining data-driven approaches with theory-guided modeling, the thesis illustrates how computational techniques—such as multivariate pattern analysis and clustering—can be harnessed to reveal the latent structure of action representations in cortical maps.

This integrative approach not only enhances the interpretability of fMRI data but also opens new avenues for the development of predictive models that may eventually have applications in robotics, cognitive rehabilitation, and artificial intelligence.

Despite these advances, several challenges remain. The current framework, while comprehensive, is not exhaustive. Future research could benefit from refining the granularity of action descriptions and exploring additional representational formats that may capture context-dependent variations more fully. Moreover, expanding the empirical paradigm to include diverse cultural contexts or real-life interactive tasks could provide further insights into how environmental factors shape neural coding of actions. Such endeavors would not only test the generalizability of the proposed models but also contribute to a more holistic understanding of the interplay between perception, motor control, and higher-level cognition.

In synthesizing theoretical discourse with empirical evidence, this thesis offers a novel perspective on one of the most fundamental aspects of human behavior. It demonstrates that action representation is a dynamic, multilevel process that cannot be fully appreciated by any single methodological lens. Rather, it is through the convergence of diverse scientific traditions—each contributing its unique insights—that we can begin to

unravel the complexity of how the brain perceives, encodes, and executes actions.

Ultimately, the contributions of this work extend beyond the immediate realm of cognitive neuroscience. They invite a broader rethinking of how we conceptualize actions in both naturalistic and controlled settings, suggesting that a truly integrative model of action must accommodate the inherent diversity and flexibility of human behavior. As the field moves forward, the integration of theoretical pluralism with advanced neuroimaging and computational techniques promises not only to refine our models of action but also to inspire new lines of inquiry into the very nature of intentionality, perception, and motor control.

In conclusion, this thesis lays the groundwork for a unified framework that captures the nuanced interplay between the physical and the conceptual aspects of actions. By reconciling theoretical debates with rigorous empirical analysis, it offers a fresh perspective on how the human brain transforms abstract intentions into concrete motor outcomes. This work thus marks an important step toward a more comprehensive and elegant understanding of the neural architecture underlying action, with implications that reverberate across cognitive science, artificial intelligence, and beyond.

Appendix

Domain	Feature		Count	Percentage
Space	Environment	Indoor	853	64.61
		Urban Outdoor	456	29.19
		Countryside Outdoor	277	17.73
	Interaction Scale	Extensive/Minimal	1101	70.48
Effector	Effector visibility	Visible/Not visible	1498	95.90
	Main effector	One hand	72	4.61
		Both hands	29	1.86
		One arm	95	6.08
		Both arms	97	6.21
		One leg or paw	36	2.31
		All legs or paws	651	41.68
		Mouth	971	62.16
		Head or nose	145	9.28
		Tongue	15	0.96
		Whole body	48	3.07
		Fingers	296	18.95
Agent & Object	Agent Type	Human/Non-human	1186	75.93
	Action target	Self-directed actions	832	53.27
		Inanimate target	554	35.47
	Tool-Mediated	Yes/No	42	2.69
	Transitivity	Yes/No	634	40.59
	Object Touch	Yes/No	783	50.13
Social	Sociality	Social/non-social	1035	66.26
	Action target	Human Target	714	45.71
		Animal Target	321	20.55
	Social touch	Yes/No	201	12.87
	People Present	>2/<=2	527	33.74
	Theory of Mind	Yes/No	24	1.54
	Multi-Agent	Joint actions	114	7.30
		Multiple agents concurrent	211	13.51
Emotion	Emotional Body Language	Yes/No	739	47.31
	Emotional Implications	Yes/No	705	45.13
	Gesticulation	Yes/No	56	3.59
	Symbolic Gesture	Yes/No	11	0.70
Linguistic	Iterativity	Repetitive / Single	799	51.15
	Dynamicity	Dynamic / Static	1530	97.95
	Durativity	Continuous / Bounded	1414	90.53
	Telicity	Telic / Atelic	649	41.55

Table S1. Feature frequency counts. For each domain, the number of

timepoints (and corresponding percentage) in which a feature was tagged is reported. The count refers to the feature level underlined (e.g., 1101 timepoints tagged as extensive rather than minimal interaction scale).

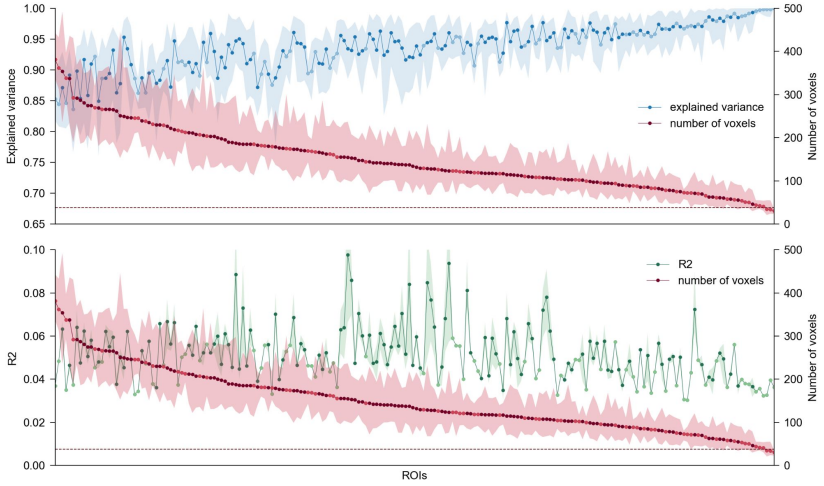

Figure S1. PCA Explained Variance, R2, and number of voxels for each ROI. The top plot shows the relation between ROIs dimensionality and retained variance after PCA. The bottom plot shows the relation between ROIs dimensionality and full model R2 . Data points are average values across subjects, shaded areas represent the min-max range.

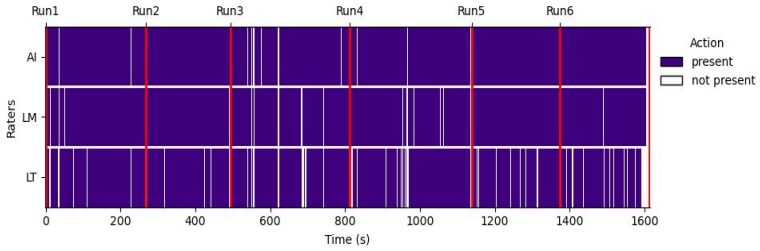

Figure S2. Action presence over time points across raters. Stimulus segmentation was performed independently by each rater (in rows), who identified a variable number of events. Timepoints in which an event is present are identified after downsampling to TR resolution, therefore event counts are inflated with respect to the original tagging temporal resolution.

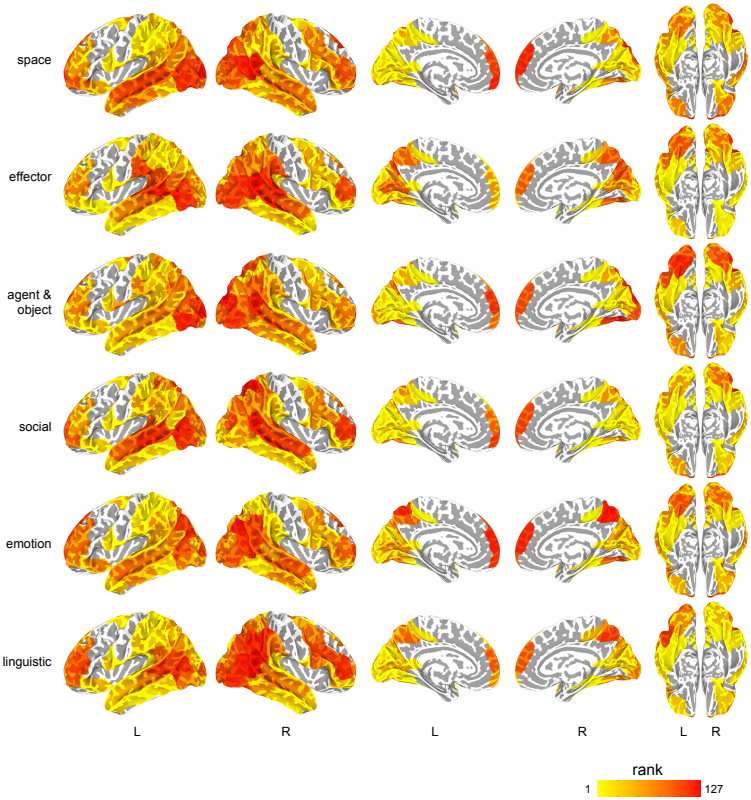

Figure S3. Average ranks in the AV modality. For each subject and domain, ROIs that survived statistical thresholding for the full model were ordered and assigned a rank (from 1 to 127) based on the R2 obtained from variance partitioning.For each ROI, ranks are averaged across subject. *L*: left**.** *R*: right.

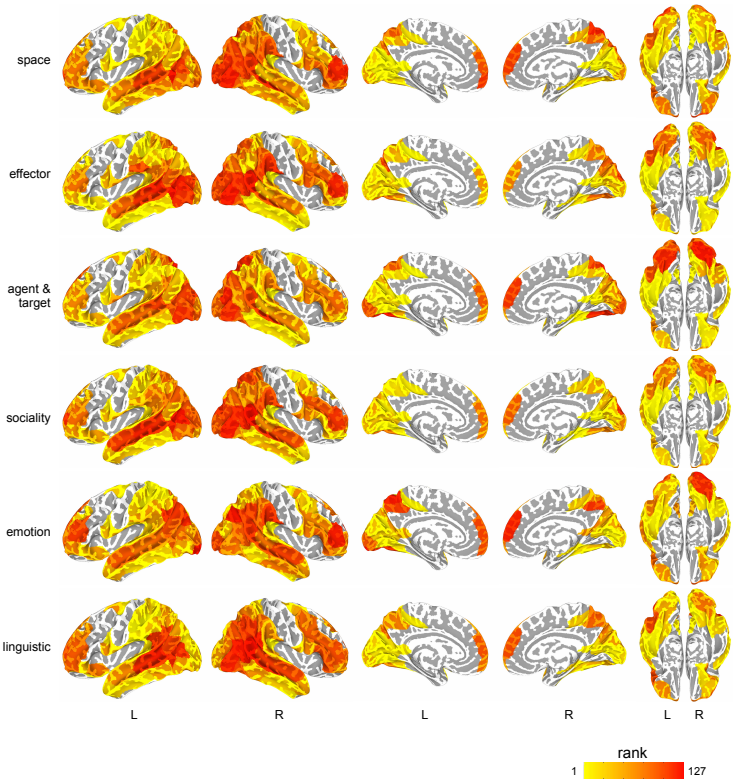

Figure S4. Average ranks in the V modality. For each subject and domain, ROIs that survived statistical thresholding for the full model were ordered and assigned a rank (from 1 to 127) based on the R2 obtained from variance partitioning.For each ROI, ranks are averaged across subject. L: left. R: right.

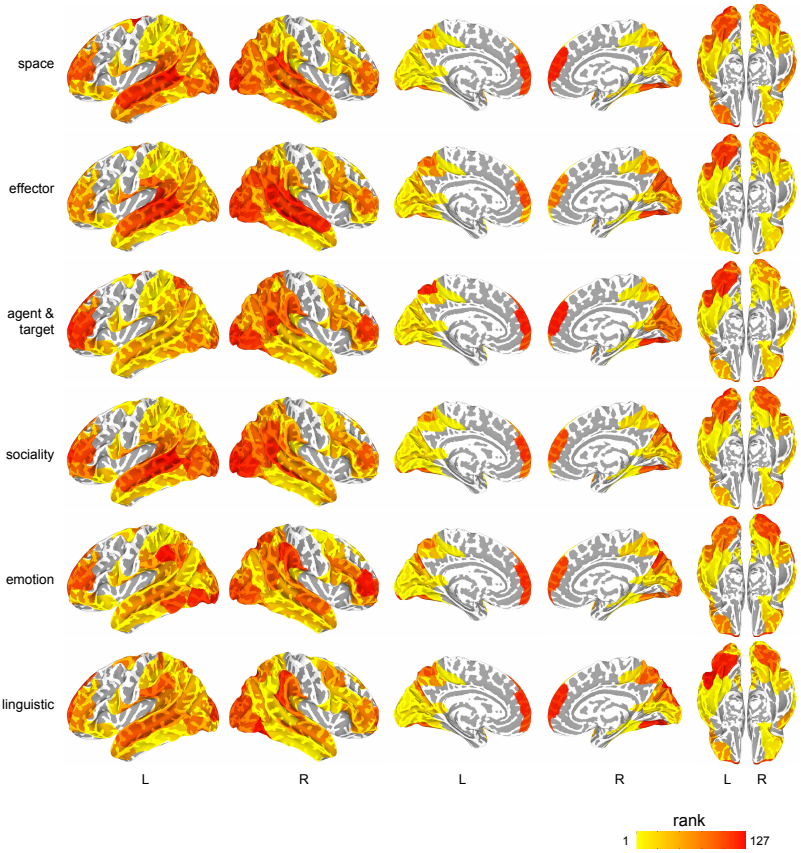

Figure S5. Average ranks in the AV modality. For each subject and domain, ROIs that survived statistical thresholding for the full model were ordered and assigned a rank (from 1 to 127) based on the R2 obtained from variance partitioning.For each ROI, ranks are averaged across subject. L: left. R: right.

	Domain	R^2	x	y	z
AV	space	0.0078	-24	-98	6
	effector	0.0196	56	-51	14
	agent & object	0.0091	19	-90	22
	social	0.0107	26	-63	58
	emotion	0.0055	8	-57	61
	linguistic	0.015	56	-51	14
V	space	0.0056	-56	-65	2
	effector	0.019	-56	-65	2
	agent & object	0.011	-18	-72	56
	social	0.009	56	-51	14
	emotion	0.0049	-26	-97	-10
	linguistic	0.0097	56	-51	14
A	space	0.0056	-56	-65	2
	effector	0.019	-56	-65	2
	agent & object	0.011	-18	-72	56
	social	0.009	56	-51	14
	emotion	0.0049	-26	-97	-10
	linguistic	0.0097	56	-51	14

Table S2. Peak R2 and coordinates of the corresponding ROI for each domain and modality. Center-of-mass coordinates are defined in MNI space.

Bibliography

Abdollahi, R. O., Jastorff, J., & Orban, G. A. (2013). Common and segregated processing of observed actions in human SPL. Cerebral Cortex, 23(11), 2734-2753.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. Nature Neuroscience, 11(9), 1109–1116. https://doi.org/10.1038/nn.2182

Alaerts, K., Swinnen, S. P., & Wenderoth, N. (2009). Interaction of sound and sight during action perception: evidence for shared modality-dependent action representations. Neuropsychologia, 47(12), 2593-2599.

Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S. P., & Wenderoth, N. (2011). Action and emotion recognition from point light displays: An investigation of gender differences. PLoS ONE, 6(6), e20989. https://doi.org/10.1371/journal.pone.0020989

Ambrosini, E., Reddy, V., de Looper, A., Costantini, M., Lopez, B., et al. (2013). Looking ahead: Anticipatory gaze and motor ability in infancy. PLoS ONE, 8(7), e67916. https://doi.org/10.1371/journal.pone.0067916

Andric, M., Solodkin, A., Buccino, G., Goldin-Meadow, S., Rizzolatti, G., & Small, S. L. (2013). Brain function overlaps when people observe emblems, speech, and grasping. Neuropsychologia, 51(8), 1619-1629.

Arioli, M., & Canessa, N. (2019). Neural processing of social interaction: Coordinate-based meta-analytic evidence from human neuroimaging studies. Human Brain Mapping, 40(13), 3712-3737.

Arioli, M., Cattaneo, Z., Ricciardi, E., & Canessa, N. (2021). Overlapping and specific neural correlates for empathizing, affective mentalizing, and cognitive mentalizing: A coordinate-based meta-analytic study. Human Brain Mapping, 42(14), 4777-4804.

Arthur, W. (2006). D'Arcy Thompson and the theory of transformations. Nature Reviews Genetics, 7(5), 401–406. https://doi.org/10.1038/nrg1835

Avenanti, A., Annella, L., Candidi, M., & Urgesi, C. (2013). Compensatory plasticity in the action observation network: Virtual lesions of STS enhance anticipatory simulation of seen actions. *Cerebral Cortex*, *23*(3), 570–580. https://academic.oup.com/cercor/article-abstract/23/3/570/312714

Bach, K. (1978). A representational theory of action. Philosophical Studies, 34(4), 361-379.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. Social and Personality Psychology Compass, 11(7), e12312.

Bach, P., Peelen, M. V., & Tipper, S. P. (2010). On the Role of Object Information in Action Observation: An fMRI Study. *Cerebral Cortex*, *20*(12), 2798-2809. https://doi.org/10.1093/cercor/bhq026

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20-28. https://doi.org/10.1016/j.cobeha.2020.07.002

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. Neuron, 95(3), 709-721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Barliya, A., Omlor, L., Giese, M. A., Berthoz, A., & Flash, T. (2013). Expression of emotion in the kinematics of locomotion. Experimental Brain Research, 225(2), 159-176. https://doi.org/10.1007/s00221-012-3357-4

Beauchamp, M. S. (2005). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. Current Opinion in Neurobiology, 15(2), 145-153.

Becchio, C., Sartori, L., & Castiello, U. (2010). Toward you: The social side of actions. Current Directions in Psychological Science, 19(3), 183-188. https://doi.org/10.1177/0963721410370131

Becchio, C., Manera, V., Sartori, L., Cavallo, A., & Castiello, U. (2012). Grasping intentions: From thought experiments to empirical evidence. Frontiers in Human Neuroscience, 6, 117. https://doi.org/10.3389/fnhum.2012.00117

Beddiar, D. R., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: A survey. Multimedia Tools and

Applications, 79, 30509-30555. https://doi.org/10.1007/s11042-020-09004-3

Beurze, S. M., De Lange, F. P., Toni, I., & Medendorp, W. P. (2007). Integration of target and effector information in the human brain during reach planning. Journal of Neurophysiology, 97(1), 188-199.

Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. Frontiers in Neuroinformatics, 10, 49.

Bischoff, M., Zentgraf, K., Pilgramm, S., Stark, R., Krüger, B., & Munzert, J. (2014). Anticipating action effects recruits audiovisual movement representations in the ventral premotor cortex. Brain and Cognition, 92, 39-57.

Bonini, L. (2017). The extended mirror neuron network: anatomy, origin, and functions. The Neuroscientist, 23(1), 56-67.

Bonivento, C., Rothstein, P., Humphreys, G., & Chechlacz, M. (2014). Neural correlates of transitive and intransitive action imitation: An investigation using voxel-based morphometry. NeuroImage: Clinical, 6, 488-497.

Blackwelder, R. E. (1967). A critique of numerical taxonomy. Systematic Zoology, 16(1), 64-72. https://doi.org/10.2307/2411518

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199-231.

Bruno, V., Ronga, I., Fossataro, C., Capozzi, F., & Garbarini, F. (2019). Suppressing movements with phantom limbs and existing limbs evokes comparable electrophysiological inhibitory responses. Cortex, 117, 64-76.

Caggiano, V., Fogassi, L., Rizzolatti, G., Casile, A., Giese, M. A., & Thier, P. (2012). Mirror neurons encode the subjective value of an observed action. Proceedings of the National Academy of Sciences, 109(29), 11848-11853. https://doi.org/10.1073/pnas.1205553109

Calvert, G. A., Spence, C., & Stein, B. E. (2004). The multisensory mind: A new approach to the study of perception. Nature Reviews Neuroscience, 5(8), 632-640.

Campbell, M. E. J., & Cunnington, R. (2017). More than an imitation game: Top-down modulation of the human mirror system. Neuroscience & Biobehavioral Reviews, 75, 195-202. https://doi.org/10.1016/j.neubiorev.2017.01.035

Cangelosi, A., & Stramandinoli, F. (2018). A review of abstract concept learning in embodied agents and robots. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1752). https://doi.org/10.1098/rstb.2017.0131

Cassis, G., & Schuh, R. T. (2010). Systematic methods, fossils, and relationships within Heteroptera (Insecta). Cladistics, 26(3), 262-280.

Catmur, C., Walsh, V., & Heyes, C. (2007). Sensorimotor learning configures the human mirror system. Current Biology, 17(17), 1527-1531. https://doi.org/10.1016/j.cub.2007.08.006

Catmur, C., & Heyes, C. (2019). Mirroring 'meaningful' actions: Sensorimotor learning modulates imitation of goal-directed actions. Quarterly Journal of Experimental Psychology, 72(2), 322-334.

Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., & Becchio, C. (2016). Decoding intentions from movement kinematics. Scientific Reports, 6. https://doi.org/10.1038/srep37036

Chambon, V., Sidarus, N., & Haggard, P. (2014). From action intentions to action effects: How does the sense of agency come about? Frontiers in Human Neuroscience, 8, 320.

Charniak, E. (1975). A partial taxonomy of knowledge about actions. In Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1. Morgan Kaufmann Publishers Inc., pp. 91–98.

Chung, C. K., Kim, H. J., & Kim, J. S. (2023). Identification of cerebral cortices processing acceleration, velocity, and position during directional reaching movement with deep neural network and kinematic data. NeuroImage. https://doi.org/10.1016/j.neuroimage.2022.118955

Cichy, R. M., Dwivedi, K., Lahner, B., Lascelles, A., Iamshchinina, P., Graumann, M., ... & Oliva, A. (2021). The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. arXiv preprint arXiv:2104.13714.

Copelli, F., Rovetti, J., Ammirante, P., & Russo, F. A. (2022). Human mirror neuron system responsivity to unimodal and multimodal presentations of action. Experimental Brain Research, 240(2), 537-548.

Cortese, A. (2021). Hands position influence on embodiment during action observation: An fMRI study. *PhD Thesis, University of Ferrara.* https://sfera.unife.it/handle/11392/2488084

Costantini, M., Ambrosini, E., Tieri, G., et al. (2010). Where does an object trigger an action? An investigation about affordances in space. Experimental Brain Research, 207, 95–103. https://doi.org/10.1007/s00221-010-2435-8

Cracraft, J. (1978). Science, philosophy, and systematics. Systematic Zoology, 27(2), 213–216. https://www.jstor.org/stable/2412975

Cracraft, J., & Donoghue, M. J. (2004). Assembling the tree of life. Oxford University Press.

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. Nature Neuroscience, 16(6), 763–770. https://doi.org/10.1038/nn.3381

Cutkosky, M. R. (1989). On grasp choice, grasp models, and the design of hands for manufacturing tasks. IEEE Transactions on Robotics and Automation, 5(3), 269-279. https://doi.org/10.1109/70.34763

De Gelder, B. (2006). Towards the neurobiology of emotional body language. Nature Reviews Neuroscience, 7(3), 242-249.

De Gelder, B., & Van den Stock, J. (2011). The bodily expressive action stimulus test (BEAST). Construction and validation of a stimulus basis for measuring perception of whole-body expression of emotions. Frontiers in Psychology, 2, 181.

De Kleijn, R., Kachergis, G., Hommel, B., & van der Weiden, A. (2014). Assessing human mirror activity with EEG mu rhythm: A meta-analysis. Psychological Bulletin, 140(2), 212-237. https://doi.org/10.1037/a0036551

De Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. Journal of Neuroscience, 37(27), 6539-6557.
Decety, J., & Grèzes, J. (1999). Neural mechanisms subserving the perception of human actions. Trends in Cognitive Sciences, 3(5), 172-178.

Della Gatta, F., Garbarini, F., Rabuffetti, M., Viganò, L., Butterfill, S. A., & Sinigaglia, C. (2017). Drawn together: When motor representations ground joint actions. Cognition, 165, 53-60.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics, 6(5), 361-375.

Dima, D. C., Hebart, M. N., & Isik, L. (2023). A data-driven investigation of human action representations. Scientific Reports, 13(1), 5171.

Dima, D. C., Janarthanan, S., Culham, J. C., & Mohsenzadeh, Y. (2024). Shared representations of human actions across vision and language. Neuropsychologia, 202, 108962.
Dima, D. C., Tomita, T. M., Honey, C. J., & Isik, L. (2022). Social-affective features drive human representations of observed actions. eLife, 11, e75027.

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. Proceedings of the IEEE International Conference on Computer Vision, 2, 726–733. https://doi.org/10.1109/ICCV.2003.1238420

Ereshefsky, M. (2007). Foundational issues concerning taxa and taxon names. Systematic Biology, 56(2), 295-301. https://doi.org/10.1080/10635150701317401

Eskenazi, T., Rueschemeyer, S. A., de Lange, F. P., Knoblich, G., & Sebanz, N. (2015). Neural correlates of observing joint actions with shared intentions. Cortex, 70, 90-100.

Fanghella, M., Era, V., & Candidi, M. (2021). Interpersonal motor interactions shape multisensory representations of the peripersonal space. Brain Sciences, 11(2), 1-23. https://doi.org/10.3390/brainsci11020255
Farris, J. S. (1967). Definitions of taxa. Systematic Biology, 16(2), 174-175. https://doi.org/10.2307/2411414

Feix, T., Romero, J., Schmiedmayer, H.-B., Dollar, A. M., & Kragic, D. (2016). The grasp taxonomy of human grasp types. IEEE Transactions on Human-Machine Systems, 46(1). https://doi.org/10.1109/THMS.2015.2470657

Ferretti, G., & Caiani, S. Z. (2019). Solving the interface problem without translation: The same format thesis. Pacific Philosophical Quarterly, 100(1), 301-333.

Ferretti, G., & Zipoli Caiani, S. (2023). How knowing-that and knowing-how interface in action: The intelligence of motor representations. Erkenntnis, 88(3), 1103-1133.

Fisher, R. A. (1970). Statistical methods for research workers. In Breakthroughs in statistics: Methodology and distribution (pp. 66-70). New York, NY: Springer New York.

Fitts, P. M., & Posner, M. I. (1967). Human performance. Belmont, CA: Brooks/Cole.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage, 47, S102.

Fridland, E. (2014). They make it look so easy: The cognitive science of skilled action. Philosophical Psychology, 27(4), 524-544. https://doi.org/10.1080/09515089.2012.727251

Fridland, E. (2017). Skill and motor control: Intelligence all the way down. Philosophical Studies, 174(6), 1-22. https://doi.org/10.1007/s11098-016-0771-7

Gallagher, S. (2005). How the body shapes the mind. Oxford University Press.

Gallese, V., & Rochat, M. J. (2018). Forms of vitality: Their neural bases, their role in social cognition, and the case of autism spectrum disorder. Psychoanalytic Inquiry, 38(2), 154-164. https://doi.org/10.1080/07351690.2018.1405672

Gallese, V., & Sinigaglia, C. (2011). How the body in action shapes the self. MIT Press.

Gallese, V., & Sinigaglia, C. (2011). What is so special about embodied simulation? Trends in Cognitive Sciences, 15(11), 512-519. https://doi.org/10.1016/j.tics.2011.09.003
Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (2004). Action recognition in the premotor cortex. Brain, 127(2), 632-646.

Gallivan, J. P., & Culham, J. C. (2015). Neural coding within human brain areas involved in actions. Current Opinion in Neurobiology, 33, 141-149.

Gallivan, J. P., McLean, D. A., Valyear, K. F., & Culham, J. C. (2013). Decoding the neural mechanisms of human tool use. eLife, 2, e00425.

Giese, M. A., & Rizzolatti, G. (2015). Neural and computational mechanisms of action processing: Interaction between visual and motor representations. Neuron, 88(1), 167-180.

Goldberg, H., Preminger, S., & Malach, R. (2014). The emotion–action link? Naturalistic emotional stimuli preferentially activate the human dorsal visual stream. NeuroImage, 84, 254-264.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. Trends in Neurosciences, 15(1), 20-25.

Grafton, S. T., & Hamilton, A. F. de C. (2007). Evidence for a distributed hierarchy of action representation in the brain. Human Movement Science, 26(4), 590-616.

Grafton, S. T., & Hamilton, A. F. de C. (2007). Understanding action: Mapping action representations in the human brain. Annual Review of Psychology, 58, 69-89. https://doi.org/10.1146/annurev.psych.57.102904.190054

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., ... & Malik, J. (2018). AVA: A video dataset of spatio-temporally localized

atomic visual actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6047-6056).

Han, J., Chauhan, V., Philip, R., Taylor, M. K., Jung, H., Halchenko, Y. O., ... & Nastase, S. A. (2024). Behaviorally relevant features of observed actions dominate cortical representational geometry in natural vision. bioRxiv, 2024-11.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. Science, 303(5664), 1634-1640.

Haxby, J. V., Gobbini, M. I., & Nastase, S. A. (2020). Naturalistic stimuli reveal a dominant role for agentic action in visual representation. NeuroImage, 216, 116561.

Haynes, J. D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. Neuron, 87(2), 257-270.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. Nature Human Behaviour, 4(11), 1173-1185.

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2007). Cortical activation to indoor versus outdoor scenes: An fMRI study. Experimental Brain Research, 179, 75-84.

Henderson, J. M., Zhu, D. C., & Larson, C. L. (2011). Functions of parahippocampal place area and retrosplenial cortex in real-world scene analysis: An fMRI study. Visual Cognition, 19(7), 910-927.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 131-135). IEEE.

Hotelling, H. (1992). Relations between two sets of variates. In Breakthroughs in Statistics: Methodology and Distribution (pp. 162-190). New York, NY: Springer New York.

Houghton, G., & Tipper, S. P. (1996). Inhibitory mechanisms of neural and cognitive control: Applications to selective attention and sequential action. Brain and Cognition, 30(1), 20-43.

Hu, Y., Yang, Y., Huang, P., Ai, D., Sun, H., Zhou, D., et al. (2023). More evidence, greater generalization? The relation between the prevalence of observed action and the strength of generalization depends on action properties. Journal of Experimental Psychology: Human Perception and Performance, 49(3), 306.

Huber, L., Finn, E. S., Handwerker, D. A., Bönstrup, M., Glen, D. R., Kashyap, S., Ivanov, D., Petridou, N., Marrett, S., Goense, J., Poser, B. A., & Bandettini, P. A. (2020). Sub-millimeter fMRI reveals multiple

topographical digit representations that form action maps in human motor cortex. *NeuroImage*, *208*, 116463. https://doi.org/10.1016/j.neuroimage.2019.116463

Humphreys, G. F., Newling, K., Jennings, C., & Gennari, S. P. (2013). Motion and actions in language: Semantic representations in occipito-temporal cortex. Brain and Language, 125(1), 94-105.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron, 76(6), 1210-1224.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. Proceedings of the National Academy of Sciences, 114(43), E9145-E9152.

Jackson, R. R., & Cross, F. R. (2011). Spider cognition. Advances in Insect Physiology, 41, 115-174.

James, T. W., VanDerKlok, R. M., Stevenson, R. A., & James, K. H. (2011). Multisensory perception of action in posterior temporal and parietal cortices. Neuropsychologia, 49(1), 108-114.

Jamrozik, A., McQuire, M., Cardillo, E. R., & Chatterjee, A. (2016). Metaphor: Bridging embodiment to abstraction. Psychonomic Bulletin & Review, 23(4), 1080-1089. https://doi.org/10.3758/s13423-015-0861-0

Jax, S. A., & Rosenbaum, D. A. (2007). Hand path priming in manual obstacle avoidance: Evidence that the dorsal stream does not only control visually guided actions in real time. Journal of Experimental Psychology: Human Perception and Performance, 33(2), 425-441. https://doi.org/10.1037/0096-1523.33.2.425

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. Behavioral and Brain Sciences, 17(2), 187-245. https://doi.org/10.1017/S0140525X00034026

Jeannerod, M. (1997). The cognitive neuroscience of action. Blackwell Publishers.

Kabulska, Z., & Lingnau, A. (2023). The cognitive structure underlying the organization of observed actions. Behavior Research Methods, 55(4), 1890-1906.

Kalénine, S., Buxbaum, L. J., & Coslett, H. B. (2010). Critical brain regions for action recognition: Lesion symptom mapping in left hemisphere stroke. Brain, 133(11), 3269-3280.

Kalénine, S., & Buxbaum, L. J. (2016). Thematic knowledge, artifact concepts, and the left posterior temporal lobe: Where action and object

semantics converge. Cortex, 82, 164-178. https://doi.org/10.1016/j.cortex.2016.06.008

Kalénine, S., Shapiro, A. D., & Buxbaum, L. J. (2013). Dissociations of action means and outcome processing in left-hemisphere stroke. Neuropsychologia, 51(7), 1224-1233.

Kaplan, J. T., & Iacoboni, M. (2007). Multimodal action representation in human left ventral premotor cortex. Cognitive Processing, 8(2), 103-113.

Karakose-Akbiyik, S., Caramazza, A., & Wurm, M. F. (2023). A shared neural code for the physics of actions and object events. Nature Communications, 14(1), 3316.

Kaufmann, A., & Newen, A. (2023). Animal thought exceeds language-of-thought. Behavioral and Brain Sciences, 46, e279.

Kauttonen, J., Hlushchuk, Y., & Tikka, P. (2015). Optimizing methods for linking cinematic features to fMRI data. *NeuroImage*, *110*, 136-148. https://doi.org/10.1016/j.neuroimage.2015.01.063

Kemmerer, D. (2021). What modulates the mirror neuron system during action observation?: Multiple factors involving the action, the actor, the observer, the relationship between actor and observer, and the context. Progress in Neurobiology, 205, 102128.

Kilner, J. M. (2011). More than one pathway to action understanding. Trends in Cognitive Sciences, 15(8), 352-357.

Kilner, J. M., Marchant, J. L., & Frith, C. D. (2006). Modulation of the mirror system by social relevance. Social Cognitive and Affective Neuroscience, 1(2), 143-148. https://doi.org/10.1093/scan/nsl017

Kirsch, L. P., & Cross, E. S. (2015). Additive routes to action learning: Layering experience shapes engagement of the action observation network. Cerebral Cortex, 25(12), 4799-4811.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, May). Similarity of neural network representations revisited. In International Conference on Machine Learning (pp. 3519-3529). PMLR.

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., & Nummenmaa, L. (2012). Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Frontiers in Human Neuroscience, 6, 233.

Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., ... & Cichy, R. (2024). Modeling short visual events through the BOLD moments video fMRI dataset and metadata. Nature Communications, 15(1), 6241.

Lai, Y. Y., Sakai, H., & Makuuchi, M. (2023). Neural underpinnings of processing combinatorial unstated meaning and the influence of individual cognitive style. Cerebral Cortex, 33(18), 10013-10027.

Land, W. M., Volchenkov, D., Bläsing, B. E., & Schack, T. (2013). From action representation to action execution: Exploring the links between cognitive and biomechanical levels of motor control. Frontiers in Computational Neuroscience, 7, 127.

Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 64(2-3), 107-123. https://doi.org/10.1007/S11263-005-1838-7

Lettieri, G., Handjaras, G., Cappello, E. M., Setti, F., Bottari, D., Bruno, V., ... & Cecchetti, L. (2024). Dissecting abstract, modality-specific and experience-dependent coding of affect in the human brain. Science Advances, 10(10), eadk6840.

Leube, D., Straube, B., Green, A., Blümel, I., Prinz, S., Schlotterbeck, P., & Kircher, T. (2012). A possible brain network for representation of cooperative behavior and its implications for the psychopathology of schizophrenia. Neuropsychobiology, 66(1), 24-32.

Liepmann, H. (1908). Über die Funktion des Balkens beim Handeln und die Beziehungen von Aphasie und Apraxie zur Intelligenz1) 2). In Drei Aufsätze aus dem Apraxiegebiet (pp. 51-78). Karger Publishers.

Liu, Y., Jiang, L., Liu, H., & Ming, D. (2021). A systematic analysis of hand movement functionality: Qualitative classification and quantitative investigation of hand grasp behavior. Frontiers in Neurorobotics, 15, 46. https://doi.org/10.3389/fnbot.2021.658075

Liu, Y., Zeng, B., Jiang, L., Liu, H., & Ming, D. (2021). Quantitative investigation of hand grasp functionality: Thumb grasping behavior adapting to different object shapes, sizes, and relative positions. Applied Bionics and Biomechanics, 2021(1), 2640422.

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. Current Biology, 23(20), 2023-2027. https://doi.org/10.1016/j.cub.2013.08.035

Malaia, E., & Newman, S. (2014). Neural bases of event knowledge and syntax integration in comprehension of complex sentences. Neurocase, 21(6), 753–766.

Martins, M. J. D., Bianco, R., Sammler, D., & Villringer, A. (2019). Recursion in action: An fMRI study on the generation of new hierarchical levels in motor sequences. https://doi.org/10.1002/hbm.24549

Masson, H. L., & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. NeuroImage, 245, 118741.

Masson, H. L., Van De Plas, S., Daniels, N., & de Beeck, H. O. (2018). The multidimensional representational space of observed socio-affective touch experiences. NeuroImage, 175, 297-314.

Mayr, E. (1969). Principles of systematic zoology. McGraw-Hill.

Mayr, E. (1969). The biological meaning of species. Biological Journal of the Linnean Society, 1(3), 311-320.

McGarry, L. M., Russo, F. A., Schalles, M. D., & Pineda, J. A. (2012). Audio-visual facilitation of the mu rhythm. Experimental Brain Research, 218, 527-538.

Mizuguchi, N., Nakata, H., & Kanosue, K. (2016). The right temporoparietal junction encodes efforts of others during action observation. *Scientific Reports*, *6*, 30274. https://www.nature.com/articles/srep30274

Mollo, D. C., & Vernazzani, A. (2024). The formats of cognitive representation: A computational account. Philosophy of Science, 91(3), 682-701.

Morales, S., Bowman, L. C., Velnoskey, K. R., Fox, N. A., & Redcay, E. (2019). An fMRI study of action observation and action execution in childhood. *Developmental Cognitive Neuroscience*, *37*, 100655. https://doi.org/10.1016/j.dcn.2019.100655

Möttönen, R., Farmer, H., & Watkins, K. E. (2016). Neural basis of understanding communicative actions: Changes associated with knowing the actor's intention and the meanings of the actions. Neuropsychologia, 81, 230-237.

Murphy, A., Zylberberg, J., & Fyshe, A. (2024). Correcting biased centered kernel alignment measures in biological and artificial neural networks. arXiv preprint arXiv:2405.01012.

Murphy, G. L. (2015). Is there an exemplar theory of concepts? Psychonomic Bulletin & Review, 23(4), 1035-1042. https://doi.org/10.3758/s13423-015-0834-3

Nakawala, H., Gonçalves, P. J. S., Fiorini, P., Ferrigno, G., & de Momi, E. (2018). Approaches for action sequence representation in robotics: A review. IEEE International Conference on Intelligent Robots and Systems, 5666-5671. https://doi.org/10.1109/IROS.2018.8594256

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. NeuroImage, 56(2), 400-410.

Nastase, S. A., Connolly, A. C., & Oosterhof, N. N. (2017). Attention selectively reshapes the geometry of distributed semantic representation. Cerebral Cortex, 27(8), 4277-4291. https://doi.org/10.1093/cercor/bhx138

Newell, K. M. (1991). Motor skill acquisition. Annual Review of Psychology, 42(1), 213-237. https://doi.org/10.1146/annurev.ps.42.020191.001241

Nelissen, K., Borra, E., Gerbella, M., & Rozzi, S. (2011). Action observation circuits in the macaque monkey cortex. *Journal of Neuroscience*, *31*(10), 3743–3756. https://www.jneurosci.org/content/31/10/3743.short

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping, 15(1), 1–25. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.1058

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology, 21(19), 1641-1646.

Niki, C., Kumada, T., Maruyama, T., Tamura, M., & Muragaki, Y. (2019). Role of frontal functions in executing routine sequential tasks. Frontiers in Psychology, 10, 169.

Ohl, M. (2007). 4 principles of taxonomy and classification: Current procedures for naming and classifying organisms. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-33761-44

O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., ... & Cirranello, A. L. (2013). The placental mammal ancestor and the post–K-Pg radiation of placentals. Science, 339(6120), 662-667.

Orban, G. A., Lanzilotto, M., & Bonini, L. (2021). From observed action identity to social affordances. Trends in Cognitive Sciences, 25(6), 493-505.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. Cognition, 107(1), 179-217. https://doi.org/10.1016/j.cognition.2007.09.003

Pacherie, E. (2018). Motor intentionality. In The Oxford Handbook of 4E Cognition (pp. 369-388).

Papeo, L., Agostini, B., & Lingnau, A. (2019). The large-scale organization of gestures and words in the middle temporal gyrus. Journal of Neuroscience, 39(30), 5966-5974.

Peelen, M. V., Romagno, D., & Caramazza, A. (2012). Independent representations of verbs and actions in left lateral temporal cortex. Journal of Cognitive Neuroscience, 24(10), 2096-2107.

Pickands III, J. (1975). Statistical inference using extreme order statistics. The Annals of Statistics, 3(1), 119-131.
Pouget, A., & Snyder, L. H. (2000). Computational approaches to sensorimotor transformations. Nature Neuroscience, 3(11), 1192-1198.

Raut, R. V., Snyder, A. Z., & Raichle, M. E. (2020). Hierarchical dynamics as a macroscopic organizing principle of the human brain. Proceedings of the National Academy of Sciences, 117(34), 20890-20897. https://doi.org/10.1073/pnas.2003383117

Rezk, M., Cattoir, S., Battal, C., Occelli, V., Mattioni, S., & Collignon, O. (2020). Shared representation of visual and auditory motion directions in the human middle-temporal cortex. Current Biology, 30(12), 2289-2299.

Richards, R. A. (2010). The species problem: A philosophical analysis. Cambridge University Press.

Ricciardi, E., Bonino, D., Sani, L., Vecchi, T., Guazzelli, M., Haxby, J. V., ... & Pietrini, P. (2009). Do we really need vision? How blind people "see" the actions of others. Journal of Neuroscience, 29(31), 9719-9724.

Ricciardi, E., Handjaras, G., Bonino, D., Vecchi, T., Fadiga, L., & Pietrini, P. (2013). Beyond motor scheme: A supramodal distributed representation in the action-observation network. PLOS ONE, 8(3), e58632.

Rizzolatti, G., Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience, 27(1), 169-192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fabbri-Destro, M. (2010). Mirror neurons: From discovery to autism. Experimental Brain Research, 200(3-4), 223-237. https://doi.org/10.1007/s00221-009-2002-3

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience, 2(9), 661-670. https://doi.org/10.1038/35090060

Rizzolatti, G., Sinigaglia, C. (2008). Further reflections on how we interpret the actions of others. Nature, 455(7213), 589. https://doi.org/10.1038/455589b

Rizzolatti, G., Sinigaglia, C. (2016). The mirror mechanism: A basic principle of brain function. Nature Reviews Neuroscience, 17(12), 757-765. https://doi.org/10.1038/nrn.2016.135

Romagno, D., Rota, G., Ricciardi, E., & Pietrini, P. (2012). Where the brain appreciates the final state of an event: The neural correlates of telicity. Brain and Language, 123(1), 68-74.

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. Human Movement Science, 26(4), 525-554. https://doi.org/10.1016/j.humov.2007.04.001

Rubin, T. N., Koyejo, O., Gorgolewski, K. J., Jones, M. N., Poldrack, R. A., & Yarkoni, T. (2017). Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. PLOS Computational Biology, 13(10), e1005649. https://doi.org/10.1371/journal.pcbi.1005649

Ryle, G. (1945). Knowing how and knowing that: The presidential address. In Proceedings of the Aristotelian Society, 46, 1-16. Aristotelian Society, Wiley.

Saif, S., Tehseen, S., & Kausar, S. (2018). A survey of the techniques for the identification and classification of human actions from visual data. Sensors (Switzerland, 18(11), 3979. https://doi.org/10.3390/s18113979

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., ... & Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cerebral Cortex, 28(9), 3095-3114.

Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. Proceedings of the National Academy of Sciences, 107(20), 9388-9393.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neuroscience & Biobehavioral Reviews, 42, 9-34.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. Trends in Cognitive Sciences, 10(2), 70-76.

Setti, F., Handjaras, G., Bottari, D., Leo, A., Diano, M., Bruno, V., & Ricciardi, E. (2023). A modality-independent proto-organization of human multisensory areas. Nature Human Behaviour, 7(3), 397-410.

Shahdloo, M., Çelik, E., Urgen, B. A., Gallant, J. L., & Çukur, T. (2022). Task-dependent warping of semantic representations during search for visual action categories. Journal of Neuroscience, 42(35), 6782-6799. https://doi.org/10.1523/JNEUROSCI.1372-21.2022

Shepherd, J. (2021). Intelligent action guidance and the use of mixed representational formats. Synthese, 198(Suppl 17), 4143-4162.

Shipley, T. F., & Zacks, J. M. (Eds.). (2008). Understanding events: From perception to action. Oxford University Press.

Simonelli, F., Handjaras, G., Benuzzi, F., Bernardi, G., Leo, A., Duzzi, D., ... & Lui, F. (2024). Sensitivity and specificity of the action observation network to kinematics, target object, and gesture meaning. Human Brain Mapping, 45(11), e26762.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27.

Sinigaglia, C. (2013). What type of action understanding is subserved by mirror neurons? Neuroscience Letters, 540, 59-61. https://doi.org/10.1016/j.neulet.2012.10.016

Sinigaglia, C., & Butterfill, S. A. (2020). Motor representation and action experience in joint action. Minimal Cooperation and Shared Agency, 181-193.

Smith, C. S. (1997). The parameter of aspect (Vol. 43). Springer Netherlands. https://doi.org/10.1007/978-94-011-5606-6

Sokolov, A. A., Zeidman, P., Erb, M., Pollick, F. E., Fallgatter, A. J., Ryvlin, P., ... & Pavlova, M. A. (2020). Brain circuits signaling the absence of emotion in body language. Proceedings of the National Academy of Sciences, 117(34), 20868-20873.

Soomro, K., & Zamir, A. R. (2015). Action recognition in realistic sports videos. In Computer Vision in Sports (pp. 181-200). Springer. https://doi.org/10.1007/978-3-319-09396-3_9

Spunt, R. P., Kemmerer, D., & Adolphs, R. (2016). The neural basis of conceptualizing the same action at different levels of abstraction. Social Cognitive and Affective Neuroscience, 11(7), 1141-1151.

Stobbe, E., Forlim, C. G., & Kühn, S. (2024). Impact of exposure to natural versus urban soundscapes on brain functional connectivity, BOLD entropy and behavior. Environmental Research, 244, 117788.

Tarhan, L., & Konkle, T. (2020). Sociality and interaction envelope organize visual action representations. Nature Communications, 11(1), 3002. https://doi.org/10.1038/s41467-020-16846-w

Tarhan, L., De Freitas, J., & Konkle, T. (2021). Behavioral and neural representations en route to intuitive action understanding. Neuropsychologia, 163, 108048.

Thompson, D. W. (1917). On growth and form. Cambridge University Press.

Thornton, M. A., & Tamir, D. I. (2021). People accurately predict the transition probabilities between actions. Science Advances, 7(9), eabd4995. https://doi.org/10.1126/sciadv.abd4995

Tipper, C. M., Signorini, G., & Grafton, S. T. (2015). Body language in the brain: Constructing meaning from expressive movement. Frontiers in Human Neuroscience, 9, 450. https://doi.org/10.3389/fnhum.2015.00450

Toschi, R. G. (2017). Representation of actions and outcomes in medial prefrontal cortex during delayed conditional decision-making: Population analyses of single neuron activity. Brain and Neuroscience Advances, 1, 2398212818773865. https://doi.org/10.1177/2398212818773865

Turella, L., Rumiati, R., & Lingnau, A. (2020). Hierarchical action encoding within the human brain. Cerebral Cortex, 30(5), 2924-2938. https://doi.org/10.1093/cercor/bhz284

Tucciarelli, R., Wurm, M., Baccolo, E., & Lingnau, A. (2019). The representational space of observed actions. eLife, 8, e47686.

Urgen, B. A., & Orban, G. A. (2021). The unique role of parietal cortex in action observation: Functional organization for communicative and manipulative actions. NeuroImage, 237, 118220.

Urgen, B. A., Pehlivan, S., & Saygin, A. P. (2019). Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. Neuropsychologia, 127, 35-47. https://doi.org/10.1016/j.neuropsychologia.2019.02.006

Urgesi, C., Candidi, M., & Avenanti, A. (2014). Neuroanatomical substrates of action perception and understanding: An anatomic likelihood estimation meta-analysis of lesion-symptom mapping studies in brain-injured patients. Frontiers in Human Neuroscience, 8, 344.

van Elk, M., van Schie, H., & Bekkering, H. (2014). Action semantics: A unifying conceptual framework for the selective use of multimodal and modality-specific object knowledge. Physics of Life Reviews, 11(2), 220-250.

Vegas, S., Juristo, N., & Basili, V. (2009). A characterization scheme for software reading techniques. Journal of Systems and Software, 82(11), 1879-1894.

Vegas, S., Juristo, N., & Basili, V. R. (2009). Maturing software engineering knowledge through classifications: A case study on unit testing techniques. IEEE Transactions on Software Engineering, 35(4), 551-565.

Vendler, Z. (1957). Verbs and times. The Philosophical Review, 66(2), 143-160. https://doi.org/10.2307/2182371

Wagner, D. D., Kelley, W. M., Haxby, J. V., & Heatherton, T. F. (2016). The dorsal medial prefrontal cortex responds preferentially to social interactions during natural viewing. Journal of Neuroscience, 36(26), 6917-6925.

Wang, S., Hou, Y., Li, Z., Dong, J., & Tang, C. (2018). Combining convnets with hand-crafted features for action recognition based on an HMM-SVM classifier. Multimedia Tools and Applications, 77, 18983-18998.

Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment (Vol. 279). John Wiley & Sons.

Wilkins, J. S. (2014). Do species exist? Principles of taxonomic classification. — By Werner Kunz. Systematic Biology, 63(2), 280-281. https://doi.org/10.1093/sysbio/syt074

Winkler, A. M., Ridgway, G. R., Douaud, G., Nichols, T. E., & Smith, S. M. (2016). Faster permutation inference in brain imaging. NeuroImage, 141, 502-516.

Wurm, M. F., & Caramazza, A. (2022). Two 'what' pathways for action and object recognition. Trends in Cognitive Sciences, 26(2), 103-116.

Wurm, M. F., Caramazza, A., & Lingnau, A. (2017). Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. Journal of Neuroscience, 37(3), 562-575. https://doi.org/10.1523/JNEUROSCI.1717-16.2016

Wurm, M. F., & Schubotz, R. I. (2012). Squeezing lemons in the bathroom: Contextual information modulates action recognition. NeuroImage, 59(2), 1551-1559. https://doi.org/10.1016/j.neuroimage.2011.08.038

Xu, D., Bondugula, R., Popescu, M., & Keller, J. (2006). Bioinformatics and fuzzy logic. 2006 IEEE International Conference on Fuzzy Systems, 817-824. https://doi.org/10.1109/FUZZY.2006.1681805

Xu, H., & Huang, C. R. (2013, September). Primitives of events and the semantic representation. In Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013) (pp. 54-61).

Yang, Y., Liu, Y., Liang, H., Lou, X., & Choi, C. (2021). Attribute-based robotic grasping with one-grasp adaptation. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6357-6363). IEEE.

Yunbo Tang, Dan Chen, and Xiaoli Li. 2021. Dimensionality Reduction Methods for Brain Imaging Data Analysis. ACM Comput. Surv. 54, 4, Article 87 (May 2022), 36 pages. https://doi.org/10.1145/3448302

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. Journal of Experimental Psychology: General, 130(1), 29.

Zarcone, A., & Lenci, A. (2010). Priming effects on event types classification: Effects of word and picture stimuli. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 32, No. 32).

Zech, P., Renaudo, E., Haller, S., Zhang, X., & Piater, J. (2019). Action representations in robotics: A taxonomy and systematic classification. The International Journal of Robotics Research, 38(5), 518-562.

Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., & Li, Z. (2017). A review on human activity recognition using vision-based method. Journal of Healthcare Engineering, 2017. https://doi.org/10.1155/2017/3090343

Zhao, M., Li, R., Xiang, S., & Liu, N. (2024). Two different mirror neuron pathways for social and non-social actions? A meta-analysis of fMRI studies. Social Cognitive and Affective Neuroscience, 19(1), nsae068.

Zheng, C. Y., Pereira, F., Baker, C. I., & Hebart, M. N. (2019). Revealing interpretable object representations from human behavior. arXiv preprint arXiv:1901.02915.

Zhou, X., Stehr, D. A., Pyles, J., & Grossman, E. D. (2023). Configuration of the action observation network depends on the goals of the observer. *Neuropsychologia*. https://www.sciencedirect.com/science/article/pii/S00283932230023 8

Zhuang, T., & Lingnau, A. (2021). The characterization of actions at the superordinate, basic, and subordinate level. Psychological Research, 86(6), 1871-1891. https://doi.org/10.1007/s00426-021-01624-0

Zunino, A., Cavazza, J., Volpi, R., Morerio, P., Cavallo, A., Becchio, C., & Murino, V. (2020). Predicting intentions from motion: The subject-adversarial adaptation approach. International Journal of Computer Vision, 128, 220-239.