COGNITIVE SCIENCE

SOCIAL DICTIONARIES

Series edited by

Wit Pasierbek and Bogdan Szlachta

COGNITIVE SCIENCE

Edited by Józef Bremer

Other volumes:

RELIGIOUS PEDAGOGY, Edited by Zbigniew Marek and Anna Walulik

PUBLIC POLICIES, Edited by Artur Wołek

POLITICAL ETHICS, Edited by Piotr Świercz

GEOPOLITICS, Edited by Jacek Kloczkowski

ENTREPRENEURSHIP, Edited by Wit Pasierbek and Krzysztof Wach

MULTICULTURALISM, Edited by Bogdan Szlachta

GLOBALIZATION AND INTERDEPENDENCE, Edited by Alicja Malewska and Mateusz Filary-Szczepanik

CULTURAL STUDIES, Edited by Leszek Korporowicz, Agnieszka Knap-Stefaniuk and Łukasz Burkiewicz

PUBLIC SECURITY, Edited by Tomasz W. Grabowski

MORAL DISPUTES, Edited by Piotr Duchliński

CONTEMPORARY HUMANITIES, Edited by Bogusława Bodzioch-Bryła

POLITICS OF MEMORY, Edited by Joanna Lubecka and Maciej Zakrzewski

SOCIAL DICTIONARIES

COGNITIVE SCIENCE

EDITED BY

Józef Bremer

IGNATIANUM UNIVERSITY PRESS

KRAKOW 2024

© Ignatianum University in Cracow, 2024

Publication financed within the commissioned tasks programme
from the funds of the Minister of Science and Higher Education
AGREEMENT No. MEiN/2021/DPI/178

Ministry of Science and Higher Education
Republic of Poland

Reviewers
dr hab. Marcin Trybulec
prof. dr hab. Urszula Żegleń

Translated into English by
Ewa Pałka

Proofreading
Michael Timberlake

Editing and proofreading
Klaudia Bień

Cover design
Studio Photo Design – Lesław Sławiński
Layout and typesetting
Jacek Zaryczny

ISBN 978-83-7614-634-8 (Print)
ISBN 978-83-7614-626-3 (Online)

Ignatianum University Press
ul. Kopernika 26 • 31–501 Krakow, PL
tel. +48 12 39 99 620
wydawnictwo@ignatianum.edu.pl
http://wydawnictwo.ignatianum.edu.pl

Printing and binding: HORN PRINT S.J. • Poznań

Contents

Foreword 7

Introduction 9

ANDRZEJ DĄBROWSKI
Intentionality 13

PAWEŁ FORTUNA, ZBIGNIEW WRÓBLEWSKI
The soul 37

EWA ODOJ
The mind-body problem 57

MAREK POKROPSKI
The naturalisation of phenomenology 77

JACEK BIELAS
Embodied cognition 97

JÓZEF BREMER
The cognitive social sciences 115

EWA JAGLARZ
Speech and language disorders 133

SŁAWOMIR SZTAJER
The cognitive science of religion 153

MICHAŁ WIERZCHOŃ
Sensory substitution 171

PIOTR KONDERAK
Cognitive semiotics: the turn towards meaning 185

ŁUKASZ WIRASZKA
Cognitive linguistics 201

MATEUSZ HOHOL, NINA BAŻELA
Mathematical cognition 223

MAREK LECHNIAK
Contemporary logic and its role in the sciences and the humanities 243

ANITA PACHOLIK-ŻUROMSKA
Nonmonotonic logic in mind modelling 265

ANDRZEJ BIELECKI
Cybernetics 281

ANDRZEJ BIELECKI
Artificial neural networks 299

MARIUSZ FLASIŃSKI
Artificial Intelligence 315

DARIUSZ ADAMEK
Emotional reactions 331

MATEUSZ JARMUŻEWSKI
The neurobiological basis of moral responsibility 351

ALEKSANDRA KUZIOR
Cognitive technologies 371

List of authors 389

Foreword

In 2019, the team members of the Ignatian Social Forum decided to continue the work that was initiated by the publication of the *Social Dictionary* in 2004. Scientists from both Polish and foreign academic centres contributed to this publication, which contains over one hundred extended essays that discuss the findings of recent humanities and social science research.

This new project is more extensive than the original *Social Dictionary*: over twenty volumes present the state of humanistic and social knowledge in the third decade of the 21st century. This knowledge concerns man, who is developing within diverse civilizations, cultures and societies, who adheres to many religions, and who exhibits diverse patterns of behaviour. Like the first four volumes (already published in Polish and in English; electronic versions are also available), each new volume is devoted to a research area that is considered particularly important to the humanities and social sciences: each investigates man and his social environment, political and public affairs, and international relations. The analyses of these areas are undertaken from diverse research perspectives; thus, they lead to a more thorough presentation of the problems typically addressed by only one discipline and substantially broaden the scope of the reflections offered by the Authors of the articles. These Authors look for an 'interpretative key' that will allow them to present the most significant issues related to each of the

volumes' main research areas, which are sometimes controversial or debatable among scientists. These research areas give the titles to the volumes of the new *Social Dictionary*. This 'interpretative key' would not be important if the articles published in each volume resembled succinct encyclopaedic entries; however, it becomes significant because the entries take the form of 20-page articles that follow a uniform pattern. The considerations presented by the Authors focus on the essence of the concepts they analyse, including their history, subject matter, and practical aspects. Written by Polish scientists representing not only different academic centres and scientific disciplines but also different 'research sensibilities', the twenty volumes are based on theoretical reflection accompanied by practical considerations. We also treat Catholic social teaching as an element of the 'interpretative key' because it is impossible to ignore twenty centuries of the legacy and richness of Christianity.

We hope that this volume will satisfy Readers as it offers not only an opportunity to learn about scientific approaches to the vital problems faced by contemporary man, states, and societies, but also an insight into sometimes difficult aspects of modernity as viewed from a Catholic perspective. We also hope that Readers will appreciate the effort of Polish scientists who, while undertaking original reflection on these issues, go beyond the mere presentation of other people's thoughts as they are aware of the importance of the intellectual achievements of Polish science.

Series editors
Wit Pasierbek and Bogdan Szlachta

Introduction

Since the time of Plato and Aristotle, philosophy – and, later, other sciences – have sought answers to questions about what mental states such as thoughts, will, and emotions are and how they connect with our physical bodies. At the turn of the 20th century, philosophical reflections on this topic were supported by research in experimental psychology that was initiated by Wilhelm Wundt. Another group of experimental psychologists, influenced by Ivan Pavlov and other physiologists, proposed a revolutionary redefinition of mental states as observable behaviours. The rise of behaviourism brought about a significant shift in perspectives on the human psyche, moving away from earlier introspection-based approaches. Behaviourism dismissed the value of first-person studies of the conscious mind and adopted the position that psychology – as a scientific discipline focused solely on behaviour – should be conducted from a third-person perspective.

Cognitive science emerged in the 1950s as a product of an era when psychology, anthropology, and linguistics redefined their research methods and scopes, while computer science and neuroscience surfaced as distinct scientific and technical disciplines. The cognitive revolution in psychology acted as a counter-revolution against behaviourism, rekindling questions on the status of first-person research. Important changes also occurred in other disciplines: Noam Chomsky, co-founder of transformational-generative grammar, made substantial contributions to psycholinguistics and computer science; Norbert Wiener's cybernetics gained popularity; Marvin Minsky and John McCarthy worked on artificial intelligence, while Alan Newell and Herbert Simon used computers

to simulate human cognitive processes. Alan Turing proposed a test to determine a machine's ability to use natural language, indirectly assessing whether it could think in a human-like manner.

The traditionally monodisciplinary scientific research on the conscious mind and cognition has underscored the need for an interdisciplinary approach. Machines have been developed that, through the straightforward application of algorithms, can solve complex computational tasks far more efficiently than humans. Advances in neuroimaging have allowed researchers to observe the functioning of the living brain and correlate it with a person's mental states. In linguistics and philosophy, there is renewed appreciation of the use and function of everyday language, specifically the language in which we speak of ourselves, our inner states, and the states of others from a first-person perspective. Cognitive science emerged against the backdrop of these and other scientific advancements. When asked what is studied by cognitive science, artificial intelligence ChatGPT responded:

> Cognitive science is an interdisciplinary field that studies the nature of the mind and its processes. It integrates knowledge from various disciplines to understand how people think, learn, remember, and perceive. Here are some key areas of study within cognitive science: perception, attention, memory, language, problem-solving and decision-making, learning, cognitive development, consciousness, artificial intelligence, and neuroscience. Cognitive science employs methodologies from psychology, neuroscience, artificial intelligence, linguistics, anthropology, and philosophy, aiming to create a comprehensive understanding of how the mind works.

Refining questions about specific sciences and their roles within cognitive science yields increasingly precise answers.

Initially, cognitive science research relied mainly on advances in computer science, later transitioning to a focus on neurocognitive science, while social and cultural factors have gained importance in the study of cognitive systems in recent years (leading to the emergence of social cognitive science). The research scope of cognitive science covers topics discussed in the articles published in this volume of *Social Dictionaries*. This includes fields shared by the philosophy of mind, cognitive psychology, computer science, artificial intelligence, neuroscience, and others. The interdisciplinary nature of cognitive science is reflected in its methodological interdisciplinary approach to its research

area, as highlighted by the Authors of the articles. It combines the analytical research methods of the humanities and the social sciences (e.g., in logic and theoretical linguistics) with the experimental and natural-science-based approaches of biology, psychology, and neuroscience. This is further augmented by studies on the development and increasing application of artificial intelligence in our individual and social lives, as well as in industry.

The predecessor of cognitive science research was cybernetics, which assumed that – regardless of whether a system is a machine, a living organism, or a social structure – the processes linked with controlling and processing information and energy in a system share common characteristics. Contemporary cognitive science research on the intentionality and emotionality of mental states combines, for instance, analyses of respondents' observable behaviours in social contexts with investigations of changes occurring in the corresponding regions of their brains. Often, these regions are difficult to identify precisely, so it is usually assumed that specific anatomical parts are involved. Research into mathematical cognition has shown that it is based on older, evolutionarily developed systems of innate knowledge that are culturally universal and develop early in individual growth. The cognitive science of religion examines cognitive processes and mechanisms involved in creating, adopting, and transmitting religious ideas, drawing on knowledge of the evolution of the human mind, the neural basis of religious experience, and studies on the mind's interaction with the surrounding world. Cognitive studies use linguistic grammar theories to build an understanding of expressions used in natural language (everyday language) and verify psycholinguistic hypotheses about data processing through experimental studies. Drawing on experimental neuroscience research and computer-linguistic programming techniques (based on artificial intelligence), they create simulation models (known as cognitive modelling).

The thematic scope of the *Cognitive Science* volume aligns with the characteristics of research in the field, forming an interdisciplinary platform for the exchange of ideas. It features articles that explore key areas of cognitive science – listed by ChatGPT above – including philosophical perspectives (*Intentionality*, *The soul*, *The mind-body problem*, and *The naturalization of phenomenology*), social sciences (*Embodied*

cognition, *The cognitive social sciences*, *Speech and language disorders*, *The cognitive science of religion*, and *Sensory substitution*), linguistics (*Cognitive semiotics: the turn towards meaning* and *Cognitive linguistics*), logic and mathematics (*Mathematical cognition*, *Contemporary logic and its role in the sciences and the humanities*, and *Nonmonotonic logic in mind modelling*), cybernetics and artificial intelligence (*Cybernetics*, *Artificial neural networks*, and *Artificial Intelligence*), neuroscience (*Emotional reactions* and *The neurobiological basis of moral responsibility*) and modern technologies (*Cognitive technologies*).

Distinguished scholars from several national academic institutions contributed to the creation of the *Cognitive Science* volume: Jagiellonian University, AGH University of Krakow, John Paul II Catholic University of Lublin, Adam Mickiewicz University in Poznań, Maria Curie-Skłodowska University in Lublin, Nicolaus Copernicus University in Toruń, the University of the National Education Commission in Krakow, Silesian University of Technology in Gliwice, Ignatianum University in Krakow, Windesheim University of Applied Sciences, and the University of Warsaw.

We live in a society where the digitisation of data profoundly impacts every aspect of our lives: how we communicate and maintain social relationships, work, learn, take care of our health, and participate in political, social, and economic life. On one hand, digitisation promises great benefits, such as better healthcare, more efficient mobility, effective energy use, and enhanced business prosperity. On the other hand, it raises complex questions, including concerns about unauthorised access to sensitive data and control over such information. What does it mean to be a rational, free, and conscious person in a world we share with plants, animals, and artefacts created by artificial intelligence? How significant is first-person knowledge, often referred to in discussions of ethical matters? Questions also arise about identifying reliable knowledge and seeking truth amid an information deluge. Addressing these and related issues requires collaboration among cognitive science researchers and the continued systematic development of this scientific discipline.

Volume editor
Józef Bremer

Andrzej Dąbrowski
University of the National Education Commission in Krakow
https://orcid.org/0000-0002-3998-9608
https://doi.org/10.35765/slowniki.447en

Intentionality

Summary

DEFINITION OF THE TERM: Intentionality is direction toward something. It is a fundamental feature of our mental acts, consciousness, and the mind in general, characterised by the mind's direction toward an object: thinking involves thinking about something, and consciousness involves consciousness of something. Collective intentionality (we-intentionality), which governs the actions of groups based on shared intentions, differs from individual intentionality understood in this way.

HISTORICAL ANALYSIS OF THE TERM: The concept of intentionality has its roots in ancient philosophy, evident in the works of thinkers like Parmenides, Plato, and Aristotle, and it continues through scholastic and modern philosophy. Franz Brentano introduced intentionality into modern philosophy as the key distinction between mental and physical phenomena. Later studies on this category, undertaken by representatives of various schools and currents, have led to different interpretations and lines of thought.

DISCUSSION OF THE TERM: The study of intentionality raises five key questions: 1) What is an intentional relationship? 2) What are mental contents? 3) What are (or can be) the objects of intentional reference of acts? 4) What are the characteristic features of intentionality in the social sciences, i.e., collective intentionality? 5) How does intentionality relate to cognitive science?

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Dedicated research programmes have contributed to

the systematic study of intentionality across different disciplines. This section explores 1) the phenomenal intentionality programme, which delves into intentionality in sensory experiences, and 2) the semantical programme, which highlights the deep connection between language, semantics, and cognition. Additionally, the notion of collective intentionality is examined, offering insights into the nature of human social life. The concept of intentionality is also shown to be intellectually stimulating and a source of inspiration for researchers in the area of cognitive science.

Keywords: Brentano, intentionality, mental intentionality, language intentionality, collective intentionality

Intentionality is a core concept in the philosophy of mind, based on the premise that mental life and individual consciousness are intentional. It also plays a significant role in the philosophy of language and semiotics, where signs and linguistic expressions point to or refer to something. However, the scope of intentionality extends beyond these areas. In contemporary discussions, intentionality is examined within social philosophy and the social sciences, where the focus shifts from individual intentionality to collective intentionality, i.e., collective intentionality. The article primarily explores intentionality from the perspectives of the philosophy of mind and language, organising the issue within these frameworks. However, given its inclusion in the Social Dictionaries series, it also incorporates the social science perspective and discusses the concept of collective intentionality.

Definition of the term

Intentionality is a defining feature of our mental acts (such as wishing, believing, imagining, or thinking), consciousness, and the mind in general; it consists in orienting them toward an object: wishing is always wishing for something, believing is believing something, imagining involves imagining something, thinking entails thinking about something. This can be succinctly defined as “intentionality is the direction of mental acts toward something”. The term itself has linguistic roots in the Latin word *intentio*, which was used by the Scholastics to translate the Greek *noema* (concept). This etymology suggests one interpretative approach to intentionality as the property of the mind being directed toward something (or identifying something) by way of conceptual content. In this sense, intentionality is deeply intertwined with both the functioning of the mind and our capacity for conceptual thinking.

A key distinction must be drawn between individual intentionality, as described above, and collective intentionality, which pertains to social actors and is crucial for understanding the nature of social and institutional life. Collective intentionality occurs when two or more people share common intentions or goals, and their actions are coordinated to reflect this shared commitment. The social sciences study societies (including social changes) and human interactions within broader

societal contexts, including cultural, economic, and political interactions. The study of collective intentionality is interdisciplinary and focused on psychological foundations (such as desires, intentions, and various forms of reasoning) that drive joint or shared actions.

Historical analysis of the term

The issue of intentional reference and mental inexistence has its roots in ancient and medieval philosophy. In antiquity, it can be traced to Aristotle's philosophy, particularly his work *On the Soul*, where he explores the functioning of the senses and the nature of the intellect. According to Aristotle, every object in the world is composed of matter and form. Sensory cognition involves perceiving the form of an object without its matter and transmitting this form to the intellect for further processing. These sensory forms already contain mental forms, which are first stored in the passive intellect and later utilised by the active intellect. In Aristotle's view, cognition requires an intermediary between the object and thought, which is the form. In medieval philosophy, Anselm further refined this idea by distinguishing between existence in the intellect (*in intellectu*) and existence in reality (*in re*). While that which exists in reality also exists in the intellect, not everything in the intellect necessarily exists in the external world. The human mind can generate abstract ideas, notions, or concepts that have no direct counterpart in reality. Thomas Aquinas expanded on this, stating that "the object which is thought is intentionally in the thinking subject, the object which is loved in the person who loves, the object which is desired in the person desiring" (Brentano, 2015, p. 93). In his efforts to clarify the workings of the theoretical and practical mind (including the will), Aquinas frequently used the polysemous term *intentio* to denote a form, concept, or extrasensory imagination, anticipating several themes later associated with intentionality.

Dominic Perler systematically explored the issue of intentionality in medieval thought, proposing five key models: (1) Thomas Aquinas's model of formal identity; (2) Peter Olivi's and Dietrich von Freiberg's model of constitutive intentionality, which is crucial for the growth of knowledge; (3) John Duns Scotus's model of intentional objects; (4) Hervaeus

Natalis's and Peter Auriol's model of intentional presence within the mind; and (5) William Ockham's and Adam Wodeham's model of natural signs (Perler, 2002). Contemporary scholars who continue to investigate various aspects of intentionality in the epistemology, metaphysics, and philosophy of language in the thought of ancient and medieval thinkers include Joel Biard, Myles Burnyeat, Victor Caston, Dorothea Frede, Richard Gaskin, Elizabeth Kruger, Cyrille Michon, Dominic O'Meara, Claude Panaccio, Robert Pasnau, Christof Rapp, Peter Simons, Richard Sorabji, and Hermann Weidemann (Perler, 2001).

In modern philosophy, the study of intentionality was reintroduced by Austrian psychologist and philosopher Franz Brentano, who aimed to reestablish the focus of descriptive psychology and develop a more refined research method. He emphasised experience, observation, and the precise description of mental processes as the foundation of psychological inquiry: "Psychology, like the natural sciences, has its basis in perception and experience" (Brentano, 2015, p. 29). Central to his approach was the notion of mental acts, which are defined by their intentionality, i.e., their direction toward specific objects. This characterisation of mental phenomena is articulated in the following passage from his influential work *Psychology from an Empirical Standpoint*:

> Every mental phenomenon is characterised by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on (Brentano, 2015, pp. 92–93).

From the outset, it is evident that intentionality distinguishes mental phenomena from physical phenomena – only mental acts are characterised by intentionality. However, the passage raises several important questions: What exactly are mental acts, and what types of these acts exist? What is mental inexistence? What does it mean to be directed toward an object? What kind of relationship is this, and what types of objects are involved? How do these objects influence the reference relation? Various thinkers have provided different answers to these questions, leading to the emergence of diverse paths of thought.

Discussion of the term

To begin, I propose focusing on three key problems and addressing them briefly, while incorporating some later developments: 1) What is an intentional relation? 2) What are mental contents? 3) What are or can be the objects of an intentional reference?

1) What is an intentional relation?
Intentionality is 'direction toward' or 'reference to' an object, which makes it a form of relation. A relation, in the basic sense, is the connection between two elements. Different types of relations exist. Some, like physical or biological relations, occur between objects or organisms; others, such as personal or social relations, exist between individuals or social groups. These relations presuppose the existence of objects interacting with one another and determine specific objects or events that can be verified in space and time; all members of these relations are spatio-temporal. Logical relations, however, are of a different kind. Their members include the contents and elements of logical sentences or expressions, connected by logical connectives like conjunction, disjunction, implication, equivalence, and negation. Logical relations and logical objects are beyond time and space; they are eternal and immutable, and the members involved belong to the same category. The subject–object relation in intentionality is unique. Its uniqueness lies in the fact that the object of the intentional relation – the second member – can vary widely. It might be a real spatio-temporal object, an event, a situation, or a state of affairs. It can also include the beliefs of others; for example, a subject might be convinced that someone else holds certain beliefs. In some cases, the object may not exist at all or may be purely thought about (abstract objects) or imagined (fictitious objects). Consequently, intentional relations are often viewed as unique, quasi-relations, or pseudo-relations.

Intentionality characterises our minds and all mental phenomena in a way that physical phenomena do not exhibit. As a result, intentionality eludes scientific explanation because it does not fit within the natural world. It cannot be accounted for in the same manner as natural phenomena and physical relations. This challenge has prompted philosophers to explore various approaches to naturalising intentionality, content, or

linguistic meaning. Notable attempts to naturalise intentionality include Daniel C. Dennett's instrumentalism, Fred Dretske's information-theoretic approach, and Ruth Millikan's teleosemantics.

2) What are mental contents?

Franz Brentano made an important distinction between transcendent and immanent objects, the latter of which can be identified – though not necessarily – with mental content. Kazimierz Twardowski further clarified this by establishing a clear separation between the act, content, and object of representation (Twardowski, 1965). In his view, an act refers to a mental activity, while content is what directs an act toward a specific object. An object exists independently of the act itself. Twardowski also emphasised the interdependence of act and content (there can be no act without content, nor content without an act), while asserting the independence of an object, which remains distinct from both the act and the content. Alexius Meinong later defended the idea of an object's independence from an act and the independence of an act itself. Mental content plays a crucial role in the intentionality of the mind as it is through content that subjects can direct toward something, regardless of whether it exists in external reality or within their own interiority. While mental content is often examined in the context of psychology and the philosophy of mind, it also plays a role in logic and semiotics. In these fields, content refers to what is expressed by a proposition or logical sentence. However, we will set this aspect aside for the moment.

There is a general consensus that mental content represents the informational content of various mental states – such as perceptions, beliefs, desires, feelings, and thoughts – that exist in an individual's mind. The subject uses this content to interpret and identify objects and events in the world. It is also widely accepted that mental content has multiple sources. The primary source is sensory experience and sensory cognition, where information from the external world is processed through the senses and represented in the mind as mental contents. Another significant source of contents is discursive thinking and reflection, which can generate new conclusions and ideas, thus creating new cognitive contents. Language plays an equally vital role in shaping mental contents as it influences how we think, the beliefs we hold, how we categorise reality, and how we organise our experiences.

Recognising the multiple sources and dimensions of mental content, theorists have introduced several important distinctions. Among the most prominent and widely debated are the distinctions between narrow and broad mental content, and between conceptual and nonconceptual content. Narrow mental content refers to representations that depend solely on an individual's internal mental states, without reference to external context, the characteristics of objects, or the relationships between them. Broad mental content, on the other hand, refers to representations that account for the external context, including the mind's interaction with the external world. In short, broad mental content is richer and more complex than narrow mental content.

The distinction between conceptual and nonconceptual content concerns how information is represented in the mind. Conceptual content involves representations that rely on cognitive categories or concepts and use language or other symbolic systems to process information. For instance, thoughts about complex objects or ideas expressed through linguistic expressions exemplify conceptual content. Nonconceptual content, by contrast, refers to something given to us in a more direct way, such as sensory impressions or perceptual experiences. These experiences, like the perception of colours (qualia), are not expressed through concepts and are often difficult to articulate in language.

3) What are or can be the objects of an intentional reference?
An object of an intentional reference is another central issue that has been extensively explored in the phenomenology of Edmund Husserl (Husserl, 1901) and Roman Ingarden (Ingarden, 1931). Ingarden introduced the idea of an intentional object as both a correlate and product of the act of consciousness, though we will not delve into this theory here. Husserl began by defining the essence of consciousness: it is intentional, meaning that consciousness is always consciousness of something (*Bewusstsein-von*…). However, the intentionality of consciousness is not a simple and uniform phenomenon as it is composed of various acts – such as perceiving, imagining, recalling, and planning – each with its own specific character and corresponding objects. For example, the object of perception is a real external object; the object of imagination is a fictional or imaginary object; the object of recollection is a past memory; and the object of planning is a projected future event or anticipation.

When addressing the general question of what kinds of objects we refer to in conscious acts, the following are not correct answers: a) transcendent objects, b) objects that exist only within the mind, or c) objects that exist both immanently and transcendently. According to Husserl, the correct answer is that, in each conscious act, we refer to an external object, even if that object does not exist in reality. Intentional thinking is indifferent to the actual existence of objects. Referring to objects is only one of the many functions of intentionality in conscious life. Husserl identified other critical functions, such as interpretation, synthesis, and constitution. The process of constitution is particularly significant, as it involves the formation of the world in consciousness. While intentionality was a foundational concept in Husserl's phenomenology, he did not provide a clear, singular theory of it. Instead, through successive analyses, particularly from the perspective of transcendental idealism, Husserl revealed additional aspects and functions of intentionality. Far from being a simple feature distinguishing mental phenomena from physical ones, intentionality is dynamic and creative in our consciousness (Husserl, 1988).

Twardowski sought to grasp the nature of an object independently of the act and content of representation. He approached the concept of the object in a relatively broad manner, positing that every representation necessarily has an object; there are no objectless representations. According to Twardowski, the essence of an object is its representability. An object can be real or unreal, possible or impossible, existing or non-existing. His analyses inspired Alexius Meinong, who defined an object as anything that can be grasped, recognised, or evaluated through a mental act (Meinong, 1904). For Meinong, an object could be anything that is both possible and impossible, leading him to develop his own theory of objects (*Gegenstandstheorie*). Meinong distinguished between several types of objects: 1) real objects (which exist), 2) ideal objects (which subsist, in Meinong's terminology), 3) objects that have non-being (including both contradictory ones, like a square circle, and noncontradictory ones, like the present King of Poland), and 4) objects that are not determined with respect to being (i.e., incomplete objects, such as fictional characters, e.g., Hamlet). It is important to note that 3 and 4 are not mutually exclusive.

Meinong's object theory, particularly his inclusion of objects that do not have being, faced criticism from Bertrand Russell but also inspired

further developments in the field. Various theorists have built on or responded to Meinong's ideas, including Dale Jacquette, Ernst Mally, Terence Parsons, Richard Routley (also known as Richard Sylvan), and Edward N. Zalta. Additionally, several reductionist approaches have emerged: 1) Deontologisation strategy, which reduces beings to the level of language and sentences. According to this view, non-existent objects do not exist, but true sentences about the non-existence of objects do exist (e.g., Tim Crane and Charles Crittenden). 2) The possible worlds strategy, which argues that while non-existent objects do not exist in the actual world, they may exist in other possible worlds (e.g., Francesco Berto and Graham Priest). 3) Fictionalism, which maintains that although non-existent objects do not exist in reality, they possess a form of fictional existence and fulfil their roles within the fictional worlds where they exist (e.g., Mark Kalderon, Frederick Kroon, Kendall Walton, and Stephen Yablo). In Poland, the topic of objects has been explored by scholars such as Arkadiusz Chrudzimski, Jacek Paśniczek, and Maciej Sendłak.

Systematic reflection with conclusions and recommendations

The primacy of individual mental intentionality. John Searle's conception of the intentionality of consciousness is perhaps the most well-known. Searle views consciousness as a biological function that evolved through natural selection and is the result of the brain's specific organisation. He argues that consciousness arises from neurophysiological processes in the brain. However, Searle stresses that merely understanding these processes does not fully explain the nature of consciousness. Consciousness is fundamentally a subjective experience unique to each individual; only the person who experiences consciousness can truly understand it, and this experience is accessible only to that person. Searle also highlights the qualitative aspect of subjective experiences, which are challenging to conceptualise and articulate. This includes experiences such as seeing a rainbow, smelling a red rose, or feeling pain – elements often referred to as qualia. Furthermore, Searle emphasises that consciousness is inherently intentional: it always refers to something and is always about something (Searle, 1983).

Referring to Searle's concept of intentionality, it is worth discussing at least four points. First, in his opinion, not all conscious mental states are intentional – there exist non-intentional experiences, such as joyful elation, depression, or melancholy, as well as unconscious intentional states such as unconscious perceptions, beliefs, and thoughts. Second, it is primarily the various mental states that constitute consciousness that are intentional. Only consciousness and the states that form it can be attributed intrinsic (or primary) intentionality. Moreover, while language and other products of human creativity are intentional, they can only be attributed secondary intentionality. Third, particular mental states have intentional content, which represents specific objects, events, states of affairs, etc., and a psychological mode, which affects how this intentional content is perceived and experienced by the subject. Fourth, intentional mental states are characterised by a direction of fit between the mind and the world (either mind-to-world or world-to-mind) and have the following conditions of satisfaction: 1) Truth conditions (for a belief): for a belief to be considered true, there must be a corresponding state of affairs that matches the belief. For instance, if John believes that the sun is shining, this belief is true only if it is indeed sunny. 2) fulfilment conditions (for a desire): a desire is fulfilled when its object is available. For example, if John desires to drink tea, the desire is fulfilled when tea is available for him to drink. 3) Carrying-out conditions (for an intention): an intention is fulfilled when the intended action is carried out. For example, if John intends to go for a walk to Jordan park, the intention is fulfilled when he actually takes that walk in the park.

Two fundamental features of consciousness are commonly identified: 1) qualitative sensation or subjective experience (such as the experience of seeing the colours of a rainbow, smelling a rose, or feeling an itch or pain), and 2) intentionality. Some consciousness researchers argue that all conscious states fall into one of these two categories, leading to a distinction between phenomenal and intentional aspects of consciousness (as proposed by scholars like Ned Block). However, counterexamples – such as vivid and emotionally charged thinking aimed at problem-solving, which exhibits both intentional and phenomenal qualities – complicate the notion of a strict division between the two types. Given these complexities, it may be more accurate to consider these as two interrelated aspects of consciousness rather than distinct

types. The relationships between consciousness and intentionality, consciousness and subjective experience, and intentionality and subjective experience are intricate and warrant further examination, especially in light of recent research, such as the phenomenal intentionality research programme, which explores how intentionality integrates with conscious experience.

The phenomenal intentionality research programme was developed to investigate the nature of intentionality of phenomenal experience. Its core premise is that all sensory experiences are intentional. Proponents of this programme generally support at least three key positions. The first is the primacy of phenomenology, or more precisely, the primacy of experience in explaining, including explaining intentionality. Sensory experiences are largely intentional, and their intentionality is rooted in their phenomenology. Second, they support phenomenal internalism and reject externalism, asserting that phenomenal intentionality is determined internally, meaning it does not rely on external conditions. Third, they endorse phenomenal liberalism, which posits that, in addition to sensory phenomenology, cognitive phenomenology also exists and contributes to the grounding of cognitive states. This perspective aims to address challenges related to the determination of content in the philosophy of language and cognition. Overall, the phenomenal intentionality research programme is seen as a significant shift in the philosophy of mind, potentially revolutionising our understanding of intentionality and conscious experience (Kreigel, 2013; Pautz, 2013).

The primacy of linguistic semantics. Many theorists of intentionality, including Brentano, Twardowski, Meinong, Husserl, Ingarden, and Searle, have recognised the role of intentionality in language, particularly in relation to the meaning and reference of linguistic expressions. Intentionality in language primarily manifests in how language relates to the external world – specifically, how names correspond to objects and how sentences refer to states of affairs or facts. Some critics have challenged, either sporadically or systematically, the mentalist perspective, questioning concepts such as ‘introspection’, ‘subject’, ‘subjectivity’, ‘acts’, and ‘mental content’, as well as the very notion of mental intentionality. Other theorists bypassed these debates and turned directly to logic, semantics, and the philosophy of language, seeking greater precision, reliability, and objectivity. Notably, the semantic contributions

of Gottlob Frege and Bertrand Russell have significantly influenced the exploration of language and intentional thinking through language.

Frege made a crucial observation regarding the identity of propositions such as “a = a” and “a = b”. For instance, the proposition “Evening Star = Evening Star” is a tautology, whereas “Evening Star = Morning Star” is not. Despite both names referring to the same object (the planet Venus), they differ in their sense – the way in which the object is presented. According to Frege, linguistic signs have both a reference (*Bedeutung*) and a sense (*Sinn*). Specifically, names denote an object and convey a sense, with the sense of a name determining its referent. Frege’s semantics extends beyond names to include predicates, which denote concepts (ideas, or purely mental representations) and sentences. A sentence expresses a logical thought and denotes a logical value: true or false. The thought expressed by a sentence determines its logical value. Frege’s semantics is underpinned by several key principles: 1) The principle of substitutivity: if two expressions share the same sense, they can be exchanged in any context without altering the sentence’s logical value. 2) The principle of compositionality: the meaning (sense) of a compound expression is determined by the meanings of its individual parts and the rules for their combination. Understanding a compound expression requires understanding its constituent parts. 3) The principle of contextuality: the meaning of an expression depends on the contexts in which it is used. Since words can have different meanings depending on their context, the meaning of a word must be assessed within the context of the entire sentence.

Frege’s work established a foundational framework for the study of language semantics. His distinctions between sense and reference, his principle of compositionality, and his analysis of linguistic structure have been pivotal in shaping modern semantic theories. Bertrand Russell built upon Frege’s ideas in his exploration of logic and semantics, critically examining the sense-reference distinction and developing his own theories of denotation and descriptions. Ludwig Wittgenstein was influenced by Frege and engaged with his ideas in his *Tractatus Logico-Philosophicus*, where he both challenged and expanded upon Frege’s concepts. Rudolf Carnap advanced the semantics of formal language, while Donald Davidson contributed to the semantics of natural language

and theories of meaning and truth. Frege's impact also extended to the philosophy of mind, as his semantic analyses have enhanced our understanding of thinking, conceptualisation, and mental representation.

Key issues in the philosophy of language and semantics include the nature of linguistic meaning and reference. Sentences in natural language possess meaning and, as a result, direct attention to entities beyond themselves, much like mental states. The issue of meaning raises further questions, such as the rules and conventions that confer meaning on specific expressions and the rules for interpreting these expressions by recipients. While the broader problem of meaning is important, here we will focus on reference, which concerns how linguistic terms relate to their referents. Understanding how terms connect to the objects they denote is crucial for comprehending the structure and function of language. This understanding aids in accurately describing and understanding the world, ensuring effective communication. Various theories address the mechanisms of reference, depending on the type of terms considered. Some theories suggest that referencing depends on grasping the meaning of a term, while others emphasise the significance of context or environment. Additionally, some propose more complex mechanisms, such as intention-based theories. A particularly compelling approach is the theory of direct reference.

According to David Kaplan, direct reference theory posits that certain linguistic expressions – such as proper names, indexicals, and demonstratives – point directly to specific objects in the world without relying on indirect descriptions or features. For example, the proper name 'Warsaw' directly refers to the city itself, not through a detailed description but as a direct reference. This notion of direct reference forms the foundation of causal theories of reference proposed by philosophers like Saul Kripke, Keith Donnellan, and Hilary Putnam, who developed their theories in response to shortcomings in Russell's descriptive approach.

In *Naming and Necessity*, Kripke argues that proper names, unlike most descriptions, function as rigid designators. This means they refer to the same object across all possible worlds (Kripke, 1980). For instance, the proper name 'Richard Nixon' always denotes the same individual, regardless of the context. Rigidity is not limited to proper names; it also applies to species names referring to natural kinds, such as 'tiger', 'gold', or 'water'. These terms denote objects that possess certain essential

and necessary characteristics, though they may also have incidental properties. For example, the term 'water' as a natural kind refers to any substance with the essential chemical composition of H_2O. Despite variations in state or impurities, a substance remains water as long as it retains this essential characteristic.

The concept of intentionality as a form of reference has significantly influenced many philosophers, particularly those who focus on language analysis. Notably, Wilfrid Sellars and Donald Davidson have contributed to this discussion through their semantic theories, which are based on the thesis of a primary nature of a semantic relationship and seek to explain the intentionality of consciousness through the semantic properties of language. Sellars, for instance, argued that the semantics of language and the intentionality of discursive thinking are intricately linked. He posited that socially shaped language plays a crucial role in developing Rylean conception of language, thereby shaping our understanding and interpretation of the world (Sellars, 1956).

Davidson's position on language and cognition is particularly compelling. He argues that language is a necessary condition for thinking and having beliefs; essentially, one cannot think or cognise without language. His argumentation is based on the idea that to have beliefs and thoughts, one must be able to interpret language. This process of interpretation involves understanding the content of others' statements and attributing meaning to linguistic expressions. Since we interpret through language, it is fundamental and indispensable. Davidson asserts that without language, thinking and cognition are impossible; animals lack language and thus cannot think or cognise in the same way humans do. Only humans possess language, which means only humans are capable of complex thought and cognitive processes (Davidson, 1982).

Collective intentionality. Humans are inherently social beings who use language and engage in social interactions with other members of a group within their species. They develop the ability to think collectively, form groups, and work together toward common goals. Community serves as the central framework within which individuals establish and uphold social norms and live by shared values. Human social life fosters the development of cultural and institutional systems, as well as political structures where decisions are made about communal living. Politics, in essence, is the capacity to govern a community (or state) for the

common good. These social, cultural, and political dimensions of human life are all rooted in collective intentionality.

The modern philosophical discussion of collective intentionality traces its roots to the works of ancient and later modern philosophers such as Plato, Aristotle, Hobbes, Locke, and Rousseau, as well as sociologists like Durkheim, Simmel, and Weber. Many contemporary scholars also draw inspiration from W. Sellars' ideas (Tuomela & Miller, 1988). Significant contributions to this field have come from thinkers like Margaret Gilbert, John Searle, and Raimo Tuomela. Margaret Gilbert tackled the question of the meaning of the word 'we', analysing various contexts where its use is inappropriate and eventually identifying a core meaning – one that refers to a plural subject that cannot simply be reduced to the properties of individual subjects. John Searle, through his exploration of mental intentionality, delved into the workings of the human mind before introducing the concept of collective intentionality, along with his theory of status functions and constitutive rules. These concepts help explain how institutional realities and social structures are created and sustained through human agreements and shared actions. Raimo Tuomela, a Finnish social philosopher, presents a sophisticated analysis of how individuals navigate between the I-mode and the we-mode, showing how people operate from a 'we' perspective that embodies collective intentionality. According to Tuomela, this form of intentionality is deeply embedded in human thought and behaviour, emerging from both biological and cultural adaptation processes.

Collective intentionality refers to the shared intentionality of a group and manifests when individuals within the group share common beliefs, intentions, undertake joint tasks, or pursue a common goal. While individual intentionality is expressed in sentences such as "I see that x", "I think that y", "I hope that z", or "I intend to achieve this or that", collective intentionality is reflected in statements like "we see that x", "we think that y", "we hope that z", or "we intend to achieve this or that". Collective intentionality is fundamental to group activities and cooperative efforts. For example, in a football team, while each player may have personal objectives and roles during training and matches, the team as a whole works towards the common goal of winning the game. Similarly, in a team working on a project, individual members may have specific tasks, but all efforts are directed towards the shared goal of successfully

completing the project. Research into collective intentionality sheds light on how groups establish goals, make decisions, organise their activities, and ultimately achieve their objectives.

The concept of collective intentionality posits that social groups can possess intentions and goals that are distinct from, and often exceed, the aggregate of the individual intentions and goals of their members. This means that a group can act as a distinct entity with its own interests and objectives, which may not necessarily align with those of its individual members. Notably, the roots of this concept can be traced back to Émile Durkheim, who argued that social reality is an influential force that shapes individual lives. According to Durkheim, society exerts a significant impact on its members through mechanisms such as language, norms, and institutions, which mould behaviour, expectations, and opportunities. He asserted that individual decisions are frequently influenced by social pressures rather than solely by personal preferences. In this view, society creates a framework that individuals navigate from the outset, which demonstrates society's inherent autonomy.

Not all researchers of collective intentionality agree that social groups possess intentions and goals entirely independent of those of their individual members. Michael Bratman, for instance, argues that individual intentions play a crucial role in explaining collective behaviour. According to Bratman, three key elements are necessary for the existence of shared intention: 1) Both I and you, as participants of shared intentionality, must have our own individual intentions toward a given goal or action. 2) Both I and you must also have intentions to act as a group and pursue a shared goal. These group intentions should be coherent and align with one another, avoiding contradictions. 3) Both I and you must have a common understanding of our intentions: each of us should know that I have an intention, you have an intention, and that these intentions are mutually acknowledged by both of us.

But where does the capacity for shared intentionality originate? Evolution provides a partial answer, suggesting that shared intentionality may be fundamental to the development of cooperative societies. While many animals, such as polar bears, leopards, rhinos, koalas, and sloths, lead solitary lives (except during mating seasons), apes exhibit more socialisation. They show interest in one another, communicate through signals like sounds, gestures, and facial expressions, engage in ritual

grooming, and hunt together. Within their groups, they establish hierarchies led by an alpha male, coexist relatively peacefully, and share food sources. However, they also maintain clear group boundaries, do not spend a lot of time together, do not cooperate tightly, and can exhibit hostility. In contrast, bonobo chimpanzees are known for their more cooperative and sociable behaviour. Members of different groups interact closely, share meals, and treat outsiders similarly to their own group members. Humans, as the closest relatives to chimpanzees, have inherited and greatly expanded upon this social character. Today, humans create complex cultures and social structures, characterised by intricate hierarchies, norms, and systems of cooperation. While biological evolution alone cannot fully explain this leap, cultural evolution offers a more comprehensive understanding.

Human development is deeply embedded in a socio-cultural reality that evolves and profoundly shapes human cognitive processes, communication, and creativity. Michael Tomasello, among others, has extensively studied these issues. He highlights that humans uniquely possess the ability to 1) conceptualise the same situation from multiple social perspectives, 2) make socially recursive and self-reflective inferences about the intentional states of others, and 3) evaluate and regulate their own thoughts in relation to the normative standards of other groups. Tomasello refers to this distinctive way of thinking as objective-reflective-normative. This mode of thinking arises from what Tomasello identifies as ‘shared intentionality’, i.e., the human capacity to collaborate and align intentions with others. Various forms of collaboration require people to recognise the goals and plans of others, which in turn stimulates the development of language, abstract thinking, and more sophisticated forms of communication. This enables people to plan for the future, organise larger social groups, and establish more complex normative systems (Tomasello, 2014).

To fully understand human intellectual and social development, it is essential to integrate both biological and cultural evolutionary perspectives. Biological evolution provides the foundational conditions for the emergence of initial forms of collective action that are crucial for developing basic cognitive and social abilities. However, cultural evolution plays an equally vital role. The dynamic interaction between biology and culture fosters the advanced thought processes, communication

methods, and social behaviours that define human uniqueness. The interplay between these two evolutionary spheres creates a level of sophistication that cannot be achieved by either alone.

Analysis of collective intentionality helps clarify several important questions, such as 1) What constitutes group interaction? 2) What is the nature of collective decision-making? 3) What defines collective responsibility? 4) How do social structures emerge and evolve? 5) What is the ontological status of these structures? (This list is not exhaustive and serves only as an illustration.) Although there is not enough space to delve deeply into each of these questions, we can briefly outline their significance.

1) What constitutes group interaction? Collective action involves a group working together toward a shared goal, with a mutual understanding of both the means and the values involved. This requires joint planning, effective communication, and cooperation, which includes task-sharing and mutual support. Trust among group members is crucial; when trust is present, communication becomes more fluid, and cooperation becomes more efficient and productive.

2) What is the nature of collective decision-making? Decisions involving household budgets, child-rearing, or family event planning necessitate collective decision-making. Similarly, public institutions, governmental bodies, and non-governmental organisations also rely on collective decision processes. Such decision-making requires open communication and negotiation, where group members strive to reach an agreement or a viable compromise. These negotiations often address priorities, resource allocation, or potential consequences. While individual decisions can be straightforward and quick, they tend to be based on limited information. In contrast, collective decisions are more complex and time-consuming but provide a more thorough evaluation of the situation. However, this complexity does not guarantee that collective decisions are free from errors.

3) What defines collective responsibility? Group members share responsibility for decisions and goals pursued together, even if not everyone contributes directly to every decision, action, or outcome. They collectively benefit from successful outcomes and may also bear moral and legal obligations to address damage resulting from poor decisions or harmful actions.

4) How do social structures emerge and evolve? Social structures begin with simple human interactions that, over time, become fixed in repetitive patterns. When these actions are regularly repeated and recognised as normative, they can evolve into formal institutions. Such communities continually evolve, adapting to changes such as demographic shifts, social conflicts, natural disasters, or technological advances.

5) What is the ontological status of social structures? There are three predominant positions regarding the ontological status of social structures. Nominalists view these structures as mere products of human actions and interactions, asserting that their 'reality' is secondary to that of individuals. Constructivists argue that social structures result from the social construction of reality, while social realists maintain that they possess an intrinsic nature and exist independently of individuals. They contend that these structures continue to exist even if all their members are replaced.

Intentionality versus cognitive science. Cognitive science is an interdisciplinary field that explores the mind (or cognitive system), including its origins, structure, and functions. It encompasses philosophy (especially the philosophy of mind and epistemology), psychology (particularly cognitive psychology), neurobiology, cognitive linguistics, artificial intelligence, and anthropology. Researchers in these disciplines employ a range of scientific and experimental methods to investigate various facets of mental processes and aim to integrate their findings. A central approach in cognitive science is the use of mental representations to explain cognitive functions, despite significant opposition from anti-representationalists. These mental representations are closely linked to intentionality. The mind's intentionality is manifested through these representations, which serve as the fundamental units of information processing. They function similarly to symbols or mental images, reflecting objects and events in the external world while also having the capacity to represent things that do not exist in the external world.

First, intentionality provides a framework for studying how the mind represents and manipulates information to achieve cognitive goals, such as pursuing truth or solving problems. Second, cognitive scientists can model various cognitive processes, such as desires, beliefs, and intentions, by examining how intentionality influences action planning and decision-making. As mentioned earlier, intentionality is crucial

for understanding the semantics of natural language. Third, cognitive scientists can explore how intentionality impacts interpretation, communication, and the processing of linguistic utterances, as cognitive systems engage in cognition and communication to enable action. Fourth, cognitive scientists can investigate how intentionality shapes social interactions. Additionally, understanding intentionality can advance the development of artificial intelligence by modelling mental processes and behaviours based on intentions and practical reasoning. Cognitive scientists focused on artificial intelligence and robotics can use these insights to design intelligent agents and systems capable of understanding and interpreting human intentions.

***

The current discourse on intentionality spans various fields and disciplines, from the philosophy of mind to social and interdisciplinary sciences, including cognitive science. This reflects its complexity, intellectual depth, and interdisciplinary nature. Philosophically, it is important to note that adopting the concept of intentionality does not necessarily commit one to a specific ontological, epistemological, or semantic stance. For instance, Husserl's phenomenological analysis of intentionality led him toward idealism, while Ingarden's phenomenological realism resulted in a robust ontology that includes real, abstract, and ideal entities. Thus, the study of intentionality opens the door to a wide array of philosophical perspectives, fostering further reflection and debate on the nature of thought, cognition, and truth. Collective intentionality, in particular, raises significant and timely research questions in social philosophy and the social sciences. By examining both individual and collective intentionality, we gain deeper insights into human thinking and action at both personal and societal levels. Consequently, the exploration of intentionality not only enriches our understanding of the human mind but also inspires continued inquiry into the nature of society and cognition.

References

Brentano, F. (2015). *Psychology from an Empirical Standpoint*. Trans. A.C. Rancurello, D.B. Terrell, & L.L. McAlister. London and New York: Routledge Classics.

Davidson, D. (1982). Rational Animals. *Dialectica*, *36*(4), 318–327.

Harman, G. (1999). Intentionality. In: W. Bechtel, & G. Graham (ed.), *A Companion to Cognitive Science* (pp. 602–610). Oxford: Blackwell Publishing Ltd.

Husserl, E. (1901). *Logische Untersuchungen*. Zweiter Teil: *Untersuchungen zur Phänomenologie und Theorie der Erkenntnis*. Vol. 2. Halle: Max Niemeyer Verlag.

Husserl E. (1998). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. First Book*. Trans. F. Kersten. Dordrecht, Boston, and London: Kluwer Academic Publishers.

Meinong, A. (1904). Über Gegenstandstheorie. In: A. Meinong (ed.), *Untersuchungen zur Gegenstandstheorie und Psychologie* (pp. 1–51). Leipzig.

Ingarden, R. (1931). *Das literarische Kunstwerk. Bine Untersuchung aus dem Gremgebiet der Ontologie. Logik und Literaturwissenschaft*. Halle: Niemeyer.

Ingarden, R. (1987). *Spór o istnienie świata*. Vol. I–III. Warszawa: PWN [English edition: Ingarden, R. (2016). *Controversy over the Existence of the World*. Vol. I-III. Trans. A. Szylewicz. Frankfurt am Main: Peter Lang].

Kriegel, U. (Red.). (2013). *Phenomenal Intentionality*. Oxford: Oxford University Press.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, Massachusetts: Harvard University Press.

Pautz, A. (2013). Does Phenomenology Ground Mental Content? In: U. Kriegel (ed.), *Phenomenal Intentionality* (pp. 194–234). Oxford: Oxford University Press.

Perler, D. (ed). (2001). *Ancient and Medieval Theories of Intentionality*. Leiden: Brill.

Perler, D. (2002). *Theorien der Intentionalität im Mittelalter*. Frankfurt am Main: Vittorio Klostermann.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Searle, J. (1990). Collective Intentions and Actions. In: P. Cohen, J. Morgan, & M.E. Pollack (ed.), *Intentions in Communication*. Cambridge, Mass.: Bradford Books, MIT Press.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. In: H. Feigl, & M. Scriven (ed.), *Minnesota Studies in the Philosophy of Science*. Vol. I (pp. 253–329). Minneapolis, MN: University of Minnesota Press.

Tomasello, M. (2014). *Natural History of Human Thinking*. Cambridge – Massachusetts – London: Harvard University Press.

Tuomela, R. (2007*). The Philosophy of Sociality: The Shared Point of View*. Oxford: Oxford University Press.

Tuomela, R., & Miller, K. (1988). We-Intentions. *Philosophical Studies*, *53*, 367–389.

Twardowski, K. (1965). O treści i przedmiocie przedstawień. Trans. I. Dąmbska. In: K. Twardowski, *Wybrane pisma filozoficzne* (pp. 3–91). Warszawa: PWN [English edition: Twardowski, K. (1977). *On the Content and Object of Presentations*. Trans. R. Grossman. The Hauge: M. Nijhoff].

Paweł Fortuna
The John Paul II Catholic University of Lublin
https://orcid.org/0000-0002-0633-4453
Zbigniew Wróblewski
The John Paul II Catholic University of Lublin
https://orcid.org/0000-0003-4477-6903
https://doi.org/10.35765/slowniki.422en

The soul

Summary

DEFINITION OF THE TERM: The concept of the soul (ψυχή [psyché] in Greek and *anima* in Latin) is integral to diverse philosophical, religious, and cultural systems. It refers to the principle of life, which is the principle that unifies and organises the functioning of plants, animals, humans, and the entire world.

HISTORICAL ANALYSIS OF THE TERM: Three main understandings of the soul can be distinguished: (a) colloquial (S1) – rooted in folk traditions and folk knowledge; (b) religious (S2) – grounded in mythical and religious tradition; (c) philosophical (S3) – a cognitive category subject to scientific and philosophical reflection.

DISCUSSION OF THE TERM: Since the advent of modern philosophy, the notion of the 'soul' has gradually been phased out of scientific discourse, theology, and even colloquial language. It has largely been replaced by terms like 'psyche', 'mind', 'self', 'identity', or 'personality'. This shift away from S3 has been contested by proponents of neoscholastic psychology, researchers studying the unconscious, and psychologists who prioritise religiosity in their reflection.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Today, S1 retains the strongest relevance, while S3 is often viewed by the scientific community as ontologically vacant and

epistemologically redundant. However, interest in the soul is being reignited in discussions on human uniqueness, particularly within the frameworks of transhumanism, posthumanism, and the anthropomorphisation of artificial intelligence systems. The prominence of S1 justifies the need for further study in folk anthropology, the development of typologies of beliefs linked with S1, and an exploration of the factors influencing how these beliefs are shaped and understood.

Keywords: soul, monism, dualism, trialism, folk anthropology

Definition of the term

The concept of the soul (ψυχή [psyché] in Greek and *anima* in Latin) is integral to diverse philosophical, religious, and cultural systems, where it refers to the principle of life that unifies and organises the functioning of various beings such as plants, animals, humans, and the entire world (the soul of the world). Depending on the type of being, the soul is attributed different functions (faculties), such as vegetative, sensory, or rational. Its ambiguity stems from a rich and diverse cultural, religious, and philosophical tradition, which is also reflected in language. The soul is frequently equated with the spirit (πνεῦμα [pnéuma] or νοῦς [noús] in Greek; *spiritus* or *animus* in Latin), defined as an immaterial and immortal substance (an essence, an independently existing being) capable of surviving the death of the physical body, with its fate extending into an afterlife or another state of being. In the anthropological dimension, the soul symbolises the hidden dimension of human existence, representing emotional and spiritual depth, a transcendence of sensory experience, and a connection with a higher being. In this sense, the soul is sometimes linked to consciousness, self-awareness, identity, moral values, conscience, and free will. The concept of the soul is embedded in folk traditions, folk psychology, and religious systems, and it continues to feature in scientific discourse, although it has become a subject of criticism and debate in modern times. Within Christian anthropology, the notion of the soul is shaped by Revelation (in both the Old and New Testaments) and is further interpreted through theological and philosophical inquiry. Key aspects of this interpretation include the nature of the soul (a spiritual substance), its origin (creationism), and its functions (cognitive, volitional, and transcendent).

Using a genetic criterion to organise the diverse meanings associated with the concept of the soul, three distinct interpretations can be identified: a) Colloquial (S1) – rooted in folk traditions and folk psychology; this view approaches the soul from a first-person perspective, seeing it as an immaterial force necessary to explain the functioning of man and natural phenomena. b) Religious (S2) – grounded in mythical and religious traditions; this understanding assigns a transcendent significance to the soul and explores its origins, man's contact with spirits, the supernatural realm, a higher being, and existence beyond biological death;

c) Philosophical (S3) – a third-person perspective that treats the soul as a cognitive category subject to scientific and philosophical reflection.

These different interpretations of the soul can also be ranked based on their theoretical complexity. This ranges from a naive theory (S1), as seen in folk psychology, to partially theorised concepts (S2), which may be found in religious discourses that do not rely on theology, which is based on cognitive schema used in philosophy, and finally to fully developed theories (S3) that emerge in philosophical and theological reflection. Additionally, these interpretations can be aligned with stages of personal development. The naive understanding of the soul (S1) typically arises in early childhood, while the more reflective and theoretical perspectives (S2 and S3) are acquired during adulthood.

Historical analysis of the term

Cultural sources of the concept of the soul

In tribal beliefs, the colloquial concept of S1 is characterised by several key properties: (a) undefined life force present in all living beings and throughout nature (e.g., among the Aborigines); (b) a close bond between the soul and the body (e.g., the Yakuts) or separation between them (e.g., the Hurons); (c) invisible mental states (e.g., desires), which are the causes of visible behaviours (e.g., the Hmong tribes); d) the existence of the soul without the body (e.g., the Santals) (Frazer, 1920).

In primitive cultures, the idea of S1 arises from three main sources: (1) The experience of life symbolically contained in breath (the soul – the breath): the process of breathing explains biological life and is interpreted as a sign of inner existence that extends beyond the physical dimension (e.g., in Sumerian, Mesopotamian, Egyptian, and Greek cultures). The close association between the breath, the soul, and the spirit is present at the linguistic level, e.g. the Greek word ψυχή [psyché], meaning ‘soul’, is derived from the verb ψύχω [psychō], meaning ‘to breathe’ or ‘to give breath’. Similarly, the Latin word *animus* (from which the French *âme*, Italian *anima*, and Spanish *alma* are derived) and the Sanskrit *ātman*, meaning ‘spiritual self’, have roots in an Indo-European root that relates to the breath. Likewise, the Slavic words *duša* and *duchъ* are

connected to the word *dých*, meaning ‘breath’. 2) The inevitability of death: the finitude of the body’s existence and the difficulty of imagining non-existence prompt people to transcend these limitations in the idea of the soul, which transcends the normal course of life. 3) The experience of sleep and dreams: while the body is asleep (apparently dead), the soul continues to live, function, dream, think, and meet with deities and the spirits of ancestors. 4) Pagan religious experience as a source of intuitions about the existence of non-material realities, although it is debated whether animism – the belief that all animate and inanimate natural phenomena possess a soul – was the earliest form of religiosity (Bremer, 2010).

Religious sources of the concept of the soul

The religious understanding of the soul (S2) has its roots in early religious systems such as animism and polytheism and later evolved in doctrinal religions like Judaism, Christianity, and Islam. In ancient Egyptian civilisation, the soul was believed to consist of multiple aspects: *Ka* (K’a) – the vital life force unique to each individual, which remained with the body after death (spirit – a person’s double); *Ba* (B’a) – the archetype of the immortal soul, capable of existing independently of the body; *Akh* (Aḫ) – the element of the soul associated with union with the divine; *Kaibit* – the soul of the shadow; *Ren* (Rn) – a person’s name. In Orphic beliefs, the soul was seen as immortal and destined to undergo a series of reincarnations to purify itself from original sin and earthly ties. The idea of reincarnation originated earlier in Hinduism (and later influenced Buddhism, Jainism, and Sikhism), where it is believed that the soul (*atman*) experiences a cycle of births and deaths based on actions (*karma*) in previous lives.

In ancient Israel, the soul was understood as deeply connected to the body. Old Testament anthropology emphasises man’s psycho-physical unity (MacDonald, 2017). In the Hebrew Bible, the foundational text of Judaism, the word נֶפֶשׁ [*nefeš*], meaning ‘soul’, is closely associated with breath and represents the desire for life, life itself, and the person as a whole. *Nefeš*, as the life-giving element, is seen as a divine gift from God, animating the human body and shaping the human personality.

This anthropology identifies three key dimensions of man: בָּשָׂר [*basar*] – the physical body, transient and fragile; רוּחַ [ruah] – the spirit, the life-giving force; נְשָׁמָה [*neshamah*] – the breath of life, the divine gift from God that animates and sustains life.

In New Testament anthropology, written in Greek, the soul and body are juxtaposed in a nuanced way. The word ψυχή [*psyche*], meaning 'soul', refers to the essence of life, endowed with moral consciousness and destined for eternal communion with God, and to the ideas of eternity and the soul's salvation. By contrast, σῶμα [*soma*] refers to the human body, both in its perishable physical condition and in a deeper theological sense, connected to spiritual and Eucharistic meanings. This distinction is further paired with πνεῦμα [*pnéuma*], meaning 'spirit', which signifies the divine Spirit flowing within a person, and σὰρξ [*sarks*], meaning 'flesh', which is the opposite of 'spirit' and represents human imperfection, the negative principle, and the realm of sin.

In Islam**,** the Qur'an introduces the concept of رُوح [*ruh*], meaning 'soul', as a divine gift that exists both before birth and after death. It is the source of consciousness and the ability to discern good from evil. The soul can be perfected through prayer, religious practices, studying the Qur'an, and living a moral life. As in Judaism and Christianity, Islam teaches the existence of an afterlife where the soul is judged based on its deeds and faith.

The Catechism of the Catholic Church describes the soul as the immortal part of man created by God and infused into the body at the moment of conception. The essence of human nature is the unity of body and soul. According to the Catechism:

> The unity of soul and body is so profound that one has to consider the soul to be the "form" of the body: i.e., it is because of its spiritual soul that the body made of matter becomes a living, human body; spirit and matter, in man, are not two natures united, but rather their union forms a single nature (CCC, 1993, point 365).

Through the soul, man possesses the capacity to discern between good and evil, moral freedom, and responsibility for his actions. At the very moment of his death, "each man […] in his immortal soul" (CCC, 1993, point 1022) is judged in a particular judgment and is sent to heaven, hell, or purgatory. At the Last Judgement**,** in which "the truth of

each man's relationship with God will be laid bare" (CCC, 1993, point 1039), there will be a resurrection, restoring the unity of soul and body for all the dead.

Philosophical sources of the concept of the soul

Throughout history, several key problem areas have shaped the philosophical understanding of the soul (S3) (Judycki, 2010). These include the soul's existence, origin, individuality, substantiality, nature, and its relationship with the body (the mind-body problem). The debate over the soul's existence is divided between two main positions: (a) the soul does not exist, and there is no need for the concept to explain human nature (a view supported by materialism and naturalism), and (b) the soul does exist, and it is essential for understanding human functioning, with its existence demonstrated either directly through phenomenological experience or indirectly through philosophical argumentation. The question of the soul's origin has been addressed through various hypotheses, notably by Saint Augustine: (a) traducianism (the soul of a child comes from the souls of the parents); (b) creationism (souls are directly created by God); (c) pre-existence (souls exist before being united with a human body); (d) germinal principles (*rationes seminales*) (souls are created within the human embryo and are actualised over time). Proponents of individuality claim that the soul is inseparable from a single organism and possesses its own distinct characteristics and identity. In contrast, the holistic view treats the soul as part of a larger whole – the soul of the world – or as an emanation of the divine soul. The substantialist position holds that the soul exists as an independent substance (being or essence), distinct from the body and other elements of reality. The non-substantialist position, however, views the soul as a function, feature, or aspect related to the organism's life – a source of life forces, a seat of primordial drives, or an organising force and substratum for human experiences. While defining the nature of S3, properties such as its simplicity, spirituality, and indestructibility (immortality) are considered. The debate over the soul's simplicity involves its unity, complexity, and divisibility. Its spirituality is defined by its immateriality and serves to explain higher mental acts (e.g., cognition, love, and abstract comprehension of

relationships), while indestructibility (immortality) pertains to its ability to exist after the death of the body.

The differences between these positions become evident in the debate on the relationship between the soul and the body (MacDonald, 2017). Since the early days of philosophy, three major positions have emerged: monism, dualism, and hylomorphism (the latter sometimes considered a form of dualism). Monism asserts the unity of the fundamental substance or principle underlying all reality. There are two main types of monism: materialist (physicalist) monism, which reduces all existence to organised matter, and idealist monism, which considers only non-material aspects of reality. The roots of monism can be traced back to the Presocratic philosophers. For example, Thales of Miletus viewed the soul as a moving force, Leucippus and Democritus regarded the soul as a component of matter, Heraclitus of Ephesus equated the soul with the mutable nature of arche (fire). Later, this idea was further developed by philosophers such as George Berkeley (subjective idealism), Gottfried Leibniz (pluralistic idealism), and Georg Hegel (absolute monism).

Dualistic anthropology, profoundly influenced by Plato and Descartes, posits a clear distinction between the soul and body. Plato presented man as a being that is fundamentally divided, with the spiritual and the corporeal realms in opposition. In his anthropology, man is the rational soul, which has a divine and eternal nature, while the mortal body is the instrument of the soul's action. The soul – self-contained, immaterial, and immortal – pre-exists the body, and its union with the body (which Plato metaphorically describes as a 'prison') is transient and contingent (death separates them). According to Plato, the soul is tripartite and consists of the following three parts: the rational soul (*noús*), which distinguishes humans from other living beings and enables cognising the truth, understanding ideas, and attaining wisdom (this aspect of the soul is often equated with the mind); the spirited soul (*thymos*), responsible for emotions and present in both humans and animals; the appetitive soul (*epithymia*), which governs desires and is found not only in humans and animals but also in plants.

While Plato and other ancient and medieval philosophers distinguished between living beings (possessing a soul) and non-living beings, Descartes introduced a new distinction into conscious and

unconscious beings. In his view, the soul is an independent substance, i.e., a 'thinking thing' (*res cogitans*), while the body is an 'extended thing' (*res extensa*). He proposed that the point of interaction between these two distinct substances is the pineal gland, which he believed to be the site of their union. Descartes was the first philosopher to equate the soul with thinking and thinking with consciousness. He sharply separated mental activity from the structure and functioning of the biological body, which he reduced to mere sensory processes. In his ontological framework, the human body is a complex machine, similar to animals' bodies. This marked a radical departure from earlier conceptions of the soul as the principle of life, as Descartes believed that the body has no need for the soul as the principle of life; plants lack a vegetative soul, and animals are mere mechanisms. Later interpretations of the concept of S3 as *res cogitans* evolved toward a desubstantialised and functional understanding of the soul: rather than being a substance, the soul is seen as a collection of mental functions (encompassing the full scope of mental life) and cognitive, moral, and aesthetic activities.

Hylomorphism posits the existence of the soul as the substantial form of the body, which lives in unity with it. This notion was introduced by Aristotle and later developed by Saint Thomas Aquinas. According to Aristotle, the soul is the first act of an organic body that has life in potency. In this view, the soul and body are inseparably united: the soul is not the body, it did not pre-exist it and is not equipped with eternal motion. Without the soul, the body is lifeless, and the soul cannot manifest its functions without the organs of the body. As a non-autonomous principle that organises matter for life, the soul constitutes a set of behavioural and mental functions. In plants, the soul governs reproduction and nutrition; in animals, it enables perception, movement, nutrition, growth, and reproduction; while in humans, the soul is rational, comprising both active and passive parts. These different types of soul are not distinct entities but rather aspects of one unified human soul that encompasses vegetative and sensory souls. Aquinas also understands the soul as the substantial form of the body – the first and autonomous act of its existence. The soul not only organises matter but creates it and gives it specific existence. As the principle of corporeal life (vegetative soul), sensual life (sensory soul), and intellectual life (rational soul), it also possesses faculties – such as intellect, will, and memory – that are not

reducible to it. Moreover, the soul can perform actions that have non-material aspects, such as cognition, decision-making, or love. Despite its profound unity with the body, Aquinas maintains that the soul is immortal; that is why it is not the body that is its source but the spiritual principle – the Creator (creationism). After biological death, the soul continues its existence as an incomplete substance, which requires the body to fulfil all its functions – hence Aquinas's emphasis on the belief in the resurrection of the body.

The version of hylomorphism developed in the main current of Christian anthropology is compositional dualism, distinct from Platonic and Cartesian dualism. In this framework, the soul and body are so closely united that the soul is considered the form of the body in all manifestations of human life, from basic physiological functions to intellectual acts. Because of the soul's highest functions, it is referred to as the 'rational soul', and due to the immaterial nature of its intellectual acts, it is also called the 'spiritual soul'.

Discussion of the term

The debate over the existence of the soul is one of the most fundamental and enduring questions in philosophy and science. It is likely to persist as long as people continue to explore the ontological foundations of their nature and identity. In the absence of empirical evidence for the soul, most researchers advocate abandoning it as an object of scientific inquiry, with methodological naturalism currently being the dominant approach in the natural sciences. However, the challenge of empirically verifying the soul does not diminish its significance in the self-understanding of many individuals worldwide. The concept of the soul profoundly influences personal identity, spiritual and religious life, moral choices, relationships, and creativity. This deep-rooted influence invites ongoing interdisciplinary research into people's beliefs about the soul, even if it is regarded merely as a "useful illusion". Regardless of its scientific standing, the soul remains an existentially meaningful concept.

The dispute over the status of the soul in philosophy and science

Since the advent of modern philosophy, there has been a gradual displacement of the term 'soul' from scientific discourse, theology, and everyday language, in favour of concepts such as 'psyche', 'mind', 'self', 'identity', or 'personality'. This shift has been influenced by several factors: (a) the difficulty of defining S3, as it is an ambiguous concept with deep cultural, historical, and theological connotations, making it problematic to incorporate into research as a measurable variable; (b) methodological materialism – natural sciences and empirical psychology (which has emerged from philosophy) focus on measurable aspects of reality, postulate freedom from metaphysical speculation, and reject the soul hypothesis due to its immeasurability; (c) lack of empirical evidence – advances in neuropsychology have led to the soul being regarded as a redundant archaic concept, often dismissed as myth, illusion, or noble lie, resulting in the working assumption that the soul does not exist; (d) secularisation – the growing acceptance of a scientific worldview based on naturalism, coupled with society's shift away from *sacrum* and increasing skepticism about supernatural elements and immortality (Kugelmann, 2023).

Since the late 19th century, Friedrich A. Lange's term "psychology without a soul" (in German: *Psychologie ohne Seele*) has gained traction. The works of Franz Brentano and William James played a significant role in the removal of S3 from empirical psychology, symbolically rejected by James R. Angell at the 1910 meeting of the American Psychological Association. Meanwhile, the determination to defend S2 has also weakened in Christian theology, reflecting a move away from philosophical conceptions that deviate from contemporary ontological beliefs (cf. Possenti, 2007).

The exclusion of S3 was resisted by proponents of neo-scholastic psychology, researchers of the unconscious, and psychologists whose work focused on religiosity. Neo-scholastic psychology, developed from the second half of the 19th century to the mid-20th century, sought to integrate discoveries from the natural sciences with Thomistic thought. This movement was shaped by Cardinal Désiré-Joseph Mercier, who viewed the doctrine of the soul as a matter of reason rather than faith. Other 20th-century figures who recognised S3 in their work include

Hubert Gruender SJ, Thomas Verner Moore, and Stephan Strasser. Carl G. Jung used the term 'soul' to describe how individuals relate to their inner psychic processes (inner personality). An intriguing contribution to this discourse came from James Hillman (1997), who proposed understanding the soul metaphorically – not as a substance but as a symbol and a point of view, representing a reflective space between a person and events. In contrast to scientific psychology, Hillman saw the soul's indefinability as a strength rather than a weakness, akin to fundamental concepts in fields like physics, biology, or philosophy that form the basis of human thinking about the world (e.g., matter, life, nature, or energy).

The exclusion of S3 from scientific debate has coincided with a shift in formulating philosophical (anthropological) questions into historical, cultural, sociological, biological, or computational ones. However, this shift has faced opposition from some philosophers and psychologists, who criticise the dangers of reductionism, naturalism, and mechanism, namely the tendency to address mind-related problems solely through the lenses of physics, biology, or chemistry (Possenti, 2006). They argue that while materialism is often supported by empirical research, traditional theories addressing the mind-body relationship (including idealism and panpsychism) continue to play a role in the debate on the mind-body problem (Nannini, 2004). Given that materialism is not a unified research programme and that the scientific consensus against the soul hypothesis is more a result of denying its existence than of producing positive evidence, some argue that disbelief in the soul should be viewed as an assumption of the scientific worldview rather than a discovery. As such, it is suggested that the soul hypothesis be treated like other scientific hypotheses, subject to ongoing debate, refinement, and testing (Baker & Goetz, 2011). The application of empirical methods to study S3 is also questioned. Critics note that such phenomena may elude empirical investigation, and just because something is not detected through empirical methods does not imply it does not exist (Possenti, 2007). In addressing anthropological questions, methodological pluralism – combining empirical and speculative approaches – is advocated. This is especially pertinent, as contemporary scholars considering the soul hypothesis include figures from a range of disciplines, such as biologists (Ettore Boncinelli), psychologists (Robert Kugelmann), linguists (Mark C. Baker), and philosophers (Richard Swinburne and Alvin Plantinga).

Despite the unsuccessful efforts of Karl Popper and John C. Eccles to introduce the concept of S3 (understood in line with Cartesian interactionist dualism) into neuroscientific research, a new hypothesis is emerging within the cognitive sciences. This hypothesis, inspired by the Aristotelian tradition, suggests that the soul could serve as an auxiliary concept in brain research (Bremer, 2010). The idea is to harness the 'synthesising' potential of the traditional concept of S3 (rational and sensory souls), thereby integrating different mental and biological states under a single principle. The unity of these states (philosophically explained by the concept of the soul) finds empirical confirmation, which makes it possible to cautiously attribute a heuristic function to S3 in research, while acknowledging that certain properties attributed to it (such as immortality and immateriality) lie beyond the scope of the empirical sciences.

A critical evaluation of S3, as understood from the third-person perspective of science, encounters the first-person phenomenological experience of individuals. The postulate of a 'science without a soul' is not equivalent to the postulate of a 'world without a soul'. Studies across various cultural contexts reveal that between 50% and 90% of adults believe in the existence of the soul (e.g., Halman et al., 2008; Richert & Harris, 2008). Moreover, both children and adults spontaneously distinguish the soul from the mind. According to Kugelmann (2023), science cannot ignore the significance of S1 in everyday experience, cultural practices, language, and the understanding of the self, others, and nature. He emphasises that the soul remains a category deeply embedded in our understanding of the world, as reflected in language through the noun 'soul', the adjectives 'soul-like', 'soulful', 'souled', 'soulless', 'soulish', and the verbs and verb phrases 'to ensoul', 'souling', or 'doing the soul'.

The noun 'soul' refers to the central or innermost aspect of a human being – the spiritual essence of a living organism. Research suggests that essentialism is a fundamental and universal feature of human cognition (Bloom, 2004). The adjectives derived from 'soul' tap into the profound dimensions of human existence, identity, love, and passion. They refer to higher mental activities beyond mere sensory perception or emotional responses, relating to our understanding of ethics, morality, and art. For instance, soul music, with its themes of searching for transcendence, consolation, healing, and spiritual connection to something

greater, expresses faith, hope, and love. The verbs derived from 'soul' convey a form of self-activity – an identity – a sustaining mode of being in the world and becoming oneself by integrating others into one's individuality. As Maurice Merleau-Ponty suggested, the soul is linked to the dynamic living process of being in the world. It actualises itself through action, animating and shaping the person. By defining the meaning of a situation, the soul contributes to the creation of sense.

Despite the rise of secularisation, contemporary man still seeks the truth about the soul. Jung (1933) warned that abandoning the traditional understanding of the soul risks creating feelings of inner emptiness, alienation, and disorientation. He argued that such a disconnection could separate individuals from their true nature, leading to a loss of connection with deeper values. Ignoring spiritual needs, he suggested, could undermine one's sense of meaning and purpose in life. The importance of both self-understanding and integrity of identity for mental health and a sense of fulfilment is widely recognised in psychological theories. In this context, preserving the concept of the soul is seen as nurturing the inner source of unity and coherence in thoughts, feelings, and actions, and as protection against identity fragmentation and inner chaos.

The current surge in non-religious spirituality can be interpreted as a response to the void left by the diminishing role of the soul in modern life. Belief in the existence of immortal S1 that is separate from the mind (in a God-centred or cosmic-spiritual form) enables individuals to consider their lives in the context of transcendence. It also affirms the soul's independence from physical, biological, psychological, and social forces. The belief in the 'inviolability' of S1 provides a sense of inalienable inner freedom that empowers people to protect their inner independence from torture, hunger, pain, indoctrination, or stigmatisation. This is why the soul remains a central anthropological concept, particularly in situations involving violations of dignity, such as during wars, migration, poverty, disease, or addiction. Furthermore, anthropological reflections on the concept of the soul are prompted by ongoing debates in various countries surrounding the death penalty, abortion, and in vitro fertilisation, as well as by confrontation with anthropomorphised systems based on artificial intelligence.

Beliefs about the soul as a research object

The presence of the soul as a fundamental category in people's lives has prompted scholars from various disciplines to explore contemporary understandings of S1, supported by empirical research. The primary focus of analyses is the conceptual structures that individuals use to understand the ontological foundation of their nature – concepts like the soul, spirit, body, and mind in Western culture – which are considered expressions of 'folk anthropology (FA)' (Fortuna et al., 2023a). People's beliefs about S1 and its connection to other elements that define man's identity and nature serve as an informal conceptual and explanatory framework people rely on in everyday life.

The widespread and varied nature of beliefs related to FA is reflected in studies examining how children and adults form their conceptions of self. Research has explored individuals' beliefs about the role of the soul, brain, and mind in performing biological, psychobiological, and mental functions; where they locate the soul, mind, and self within the body, and how these beliefs influence moral decision-making (Bloom, 2004; Richert and Harris, 2008). Systematic reflections on this topic focus on (1) the origins of FA, (2) individual differences in how S1 and other aspects of the ontological basis of one's nature are understood, and (3) the role these common-sense beliefs play in an individual's life.

FA arises from psychological essentialism, shaped by individual phenomenological experiences and culturally transmitted ideas about the ontological basis of human nature. Research has shown that intuitive dualism develops in early childhood, as children naturally understand mental states and mental beings as fundamentally different from physical objects and real-world events (Bloom, 2004). Regardless of cultural context, children recognise that bodies and minds function according to distinct causal principles, operating under the assumption that one – and only one – mind is responsible for the actions of one – and only one – body. This early childhood impression of dualism is thought to stem from children's ability to imagine what is absent or unreal, to experience dreams, to reflect on mental images, and infer the mental states of themselves and others.

The issue of the soul inevitably emerges in individual development as a topic for reflection, since – as Czaja notes – "the 'question of the

soul', concerning its existence or non-existence and its various interpretations, is a fundamental question for any anthropology, both scientific and everyday" (Czaja, 2006, p. 26). Engaging with diverse perspectives prompts both philosophical and psychological reflection, shaping beliefs related to FA, often unconsciously and to varying degrees subject to critical scrutiny. This process is influenced by several factors, including narratives that explain the non-material aspects of the world; philosophical explorations of identity aimed at understanding one's inner self; religious beliefs focusing on life after death and communion with a higher power; teleological reasoning that seeks purpose and meaning in life; experiences of love, suffering, and other boundary situations. Additionally, spiritual and transcendental experiences, such as those found in meditation, contemplation, mystical encounters, or contact with art, also play a crucial role.

Research identifies three types of FA: (1) unipartite anthropology, as seen in naïve monists who base their understanding of human existence either solely on corporeality (physicalist monism) or on the mind-spirit element (spiritualist monism or idealism); (2) bipartite anthropology, exemplified by naïve dualists who view the body and mind/soul as distinct essences; (3) tripartite anthropology, articulated by naïve trialists who distinguish between body, mind, and soul, using this framework to explain their own functioning. Naïve trialism emerges from differentiating the soul from the mind both ontologically and functionally. Research shows that children in the U.S. can identify the soul's spiritual function, as opposed to cognitive or biological functions, while U.S. and Polish adults who believe in the soul typically view it as emerging earlier than the mind, remaining more stable throughout life, and persisting longer after biological death (Richert & Harris, 2008). Respondents tended to associate cognitive functions like problem-solving, thinking, memory, and discerning between good and evil with the mind, while life force, the afterlife, connection to a higher being, and spirituality were associated with the soul.

Within similar FA, people may offer varied answers to the questions, "What is the soul?" and "What is the soul for?". S1 is described both literally – as an immaterial essence or life force – and metaphorically, using imagery from natural phenomena (such as fire or breath), technology (rudder, battery), invisibility (transparent man), physics (energy,

magnetic field), interiority (autonomous inner realm), mental processes (consciousness), and religious concepts (God's spark, seed of immortality) (Czaja, 2006). It has been observed that naïve monists may view both the mind and soul as functions of the body, specifically the brain, while naïve dualists and trialists often consider the mind as a function of the soul or regard the soul as a construct layered upon the mind. Moreover, believers in the soul's existence can be placed on a 'soul spirituality continuum', which reflects the strength of their belief in various attributes of S1: its existence before birth, its survival after biological death, its role as the source of man's life force and spiritual essence, and its ability to enable life after death or facilitate man's connection with a higher power (Richert & Harris, 2008).

The individual understanding of S1 is reflected in the following areas of life: (a) defining membership in the category of 'man' by identifying the soul as a component of man's ontological structure ("The soul is 'me' because 'I' am not my body – when I die, the body remains, but I am not there"; Czaja, 2006, p. 241); (b) shaping subjectivity and identity, when S1 is treated as a key factor that defines the self and gives personal meaning to experiences; (c) explaining human functioning, including the sources of illness – naive trialists assign a greater role than monists and dualists to the soul in biological (life, growth), psychobiological (e.g., sensations of hunger, sleep needs), and psychological (perception, desires, emotions, and beliefs) processes; (d) moral reasoning – identifying how respondents view S1 allows researchers to predict their ethical decisions regardless of declared beliefs, as the belief in the existence of the soul influences considerations related to the morality of actions, such as the ethical assessment of decisions regarding abortion, stem cells, or euthanasia. Individuals who believe in the existence of a mind and soul often base the moral status they assign to other entities (e.g., humanoid robots) on these beliefs (Fortuna et al., 2023b); (e) religiosity – better differentiation between the soul and the mind correlates with stronger religiosity and the belief that the soul is distinct from the mind and sovereignly 'governs' moral judgments and transcendence. Belief in the existence of the soul influences reflections on immortality and a sense of connection with the deceased (Kugelmann, 2023); (f) spirituality – Christians rate their spirituality higher than Buddhists, who also hold a strong belief in the existence of the soul, yet they are more likely

than Christians to attribute cognitive functions to the soul; (g) building relationships – many people refer to S1 when explaining the experience of a deep, emotional connection with others (e.g., using expressions like 'soulmate' or 'meeting the other half of your soul'); (h) creativity – S1 is associated with both the process of creation and the reception of art (e.g., "Every artist dips his brush in his own soul and paints his own nature into his pictures" – Henry Ward Beecher).

Systematic reflection with conclusions and recommendations

Currently, S1 holds the strongest status as a category embedded in the everyday reflections of ordinary people, while S3 is regarded by much of the scientific community as ontologically empty and epistemologically unnecessary. Reflection on the soul gains renewed relevance in discussions about human uniqueness, especially in the contexts of transhumanism, posthumanism, and the anthropomorphisation of AI-based systems. These discussions raise questions of identity (Who am I?) and anthropology (What is man?), which in turn provoke inquiry into FA: What makes me human? Both the biotechnological enhancement of humans, as advocated by transhumanism (posthuman design), and the posthumanist idea of subjectivising all forms of life (a post-anthropocentric community of life) clash with the anthropocentric view of man as a being superior to other beings. Research indicates that individuals with anthropocentric beliefs tend to condition the attribution of moral status to non-human entities or figures subjected to cyborgisation based on whether they possess mind and soul. However, this tendency diminishes as the strength of their anthropocentric beliefs increases (Fortuna et al., 2023b).

Kugelmann advocates for reintroducing the concept of the soul into academic discourse, emphasising that the focus should not only be on "soul making" by people but also on – based on the assumption on its existence – "how meaning is made, how lives become, how we suffer, love, live, and die" (Kugelmann, 2023, p. 276). The significance people attach to S1 drives its exploration within FA, with a priority on identifying the psychocultural factors that shape FA. Research involving children

suggests that the experience of duality arises independently of the cultural transmission of religious ideas, though it remains unclear whether this is a universal phenomenon. Moreover, concepts such as 'soul', 'spirit', and 'mind' used in different languages are not universal and are deeply influenced by the specific ethno-psychology and ethno-philosophy of each culture. Relying on Western ethno-psychological categories as a foundation for scientific research risks a distorted perspective (e.g., in Chinese culture, the primary ontological subject is not the individual but the multi-generational family). Thus, conducting research in other cultural contexts requires methods that avoid imposing beliefs based on the Christian tradition about the ontological basis of human nature.

The confirmed diversity among believers in the existence of S1 underscores the need to develop a typology of FA, along with a deeper understanding of the psychological factors that shape how meanings are attributed to S1 and the factors that shape the architecture of FA. It seems interesting to explore the connections between FA and philosophical and religious traditions as well as the ideological currents and pop culture narratives. Given the assumed hierarchical relationship between essences in naïve trialism, it is likely that distinct categories such as naïve spiritualists (who prioritise the soul), naïve mentalists (who prioritise the mind), and naïve somatists (who prioritise the body) will emerge. A key area for future research involves identifying the functions that S1 – or FA more broadly – plays in human life, both within and outside the context of faith. For instance, it may influence attribution processes (e.g., How can the dead convey messages to the living?), theories of mind (e.g., What is it like to be dead?), concept acquisition (e.g., conceptualising the supernatural properties of the deceased), the formation of beliefs (e.g., regarding the causes of illness and engagement in health-promoting behaviours), teleological reasoning (e.g., viewing one's life as having a purpose to fulfil before one's death), formulating moral judgments (e.g., decisions about medical diagnoses and treatment), and experiences of community or awe for beauty.

References

Baker, M.C., & Goetz, S. (2011). Introduction. In: M.C. Baker & S. Goetz (ed.), *The soul hypothesis. Investigations into existence of the soul* (pp. 1–20). New York: Continuum International Publishing Group.

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York: Basic Books.

Bremer, J. (2010). Pojęcie duszy w naukach kognitywnych. *Filozofia chrześcijańska*, *7*, 37–63.

Czaja, D. (2006). *Dusza. Figury wyobraźni i gry językowe*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Fortuna, P., Wróblewski, Z., & Gut, A. (2023a). Dusza w potocznej antropologii: przyczynek do identyfikacji typów ontologicznej architektury naiwnego spirytualizmu. In: T. Maziarka (ed.), *W poszukiwaniu duchowości człowieka* (pp. 111–132). Kraków: Wydawnictwo CCPres.

Fortuna, P., Wróblewski, Z., Gut, A., & Dutkowska, A. (2023b). The relationship between anthropocentric beliefs and the moral status of a chimpanzee, humanoid robot, and cyborg person: the mediating role of the assignment of mind and soul. *Current Psychology*, *43*. DOI: 10.1007/s12144-023-05313-6.

Frazer, J.G. (1920). *The Golden Bough. A Study in Magic and Religion*. New York & London: MacMillan and Co.

Hillman, J. (1997). *Suicide and the Soul*. Woodstock, Connecticut: Spring Publications, Inc.

Judycki, S. (2010). Istnienie i natura duszy ludzkiej. In: S. Janeczek (ed.), *Antropologia* (pp. 121–177). Lublin: Wydawnictwo KUL.

Jung, C.G. (1933). *Modern man in search of a soul*. Trans. W.S. Dell & C.F. Baynes. New York: Harcourt Brace Jovanovich.

Catechism of the Catholic Church (CCC) (1993). https://www.vatican.va/archive/ENG0015/_INDEX.HTM. Accessed 27.08.2024.

Kugelmann, R. (2023). *Soul in soulless psychology*. Cambridge: Cambridge University Press.

McDonald, P.S. (2017). *History of the Concept of Mind*. New York: Routledge.

Possenti, V. (2007). Spór o jedność człowieka i wyzwania nowego naturalizmu. Dusza – umysł – ciało. Trans. A. Gudaniec. In: A. Maryniarczyk & K. Stepień (ed.), *Dusza, umysł, ciało. Spór o jedność bytową człowieka* (pp. 97–145). Lublin: Polskie Towarzystwo Tomasza z Akwinu.

Richert, R.A., & Harris, P.L. (2008). Dualism revisited: body vs. mind vs. soul. *Journal of Cognition and Culture*, *8*(1–2), 99–115. DOI: 10.1163/156770908X289224ht.

Ewa Odoj
Ignatianum University in Cracow
https://orcid.org/0000-0003-0821-9317
https://doi.org/10.35765/slowniki.484en

The mind-body problem

Summary

DEFINITION OF THE TERM: The mind-body problem concerns the question of how interaction between mental and physical states is possible. The solution depends on how the nature of the mind is understood.

HISTORICAL ANALYSIS OF THE TERM: The most significant historical positions on mind–body problem are substance dualism and reductive physicalism (in a form of logical behaviourism, eliminativism, and type identity theory).

DISCUSSION OF THE TERM: The failure of projects that are types of reductive physicalism primarily stems from their omission of the phenomenal aspect of conscious states (*qualia*). That is why the most popular contemporary positions regarding the nature of mental states are types of non-reductive physicalism. However, they become entangled with the accusation of epiphenomenalism – the position according to which, although mental states are not reducible to physical states, they cannot exert causal effects in the physical world. Another important position is functionalism, but it faces similar objections to those of competing positions.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Contemporary reflection on the mind–body problem should take into account the results of cognitive sciences, which concern, among other things, the nature and genesis of phenomenal consciousness.

Keywords: mind-body problem, concepts of the mind, reductive physicalism, non-reductive physicalism, functionalism

Definition of the term

The mind-body problem, which is the problem of interaction between mind and body, is a central issue in the philosophy of mind. The distinction between mind and body has its original source in the experience of oneself as a conscious, autonomous self ('I'), which seems non-reducible to the material body. The self appears as the subject of mental states, which include perceiving, thinking, wishing, feeling, and imagining. The question arises here of whether and how the mind, which is immaterial and acts in an autonomous and free manner, can influence and be influenced by the material body, which remains under the strict rule of the laws of nature. Historically, the mind-body problem was first explicitly posed in this way by Descartes as the problem of the interaction between spiritual substance (in Latin *res cogitans*) and corporeal substance (in Latin *res extensa*), and it subsequently became one of central issues of 17th-century philosophy (among others, for Malebranche, Leibniz, and Spinoza). The development of the natural sciences led to emphasising another aspect of the mind-body problem: the question of the synthesis of our everyday understanding of mental phenomena, which is called folk psychology, and the picture of ourselves that emerges from the natural sciences, in which we are viewed in a naturalistic way (that is, exclusively as elements of the natural world). From this perspective, in the 20th and 21st centuries, primarily in analytical Anglo-Saxon philosophy of mind, the problem in question was called the mind-body problem, or even the mind-brain problem, to emphasise the role of the central nervous system.

Philosophers of mind try to identify a certain set of properties (predicates), that characterise mental states and distinguish them from physical (material) states. In the tradition of substance dualism, whose most important representative is Descartes, the radical ontic difference between mind and body was emphasised, and non-spatiality was considered an elementary property of mental states (as opposed to spatial matter). Today, privacy, intentionality, and consciousness of mental states are pointed at, although none of these properties is unquestioned (Bayne, 2022, pp. 6–23). The privacy (subjectivity) of mental states means that they are cognitively accessible in a direct, privileged way only to the 'I' experiencing that state. For all other subjects, these states are

accessible only indirectly (for example, by observing the facial expression of a person who is experiencing pain or through verbal description). This asymmetry does not exist in cases of cognitive access to physical states as they are potentially given to different persons in the same way. Following Franz Brentano, it is assumed that the characteristic property of mental states – which differentiates them from physical states – is intentionality, that is, their being directed toward something (referring to something). Acts such as thinking and judging, along with desires, emotions, perceptions, and so on, refer to something different than themselves: to an object (not necessarily existing in reality), an event, or a state of affairs. Another property of mental events is their conscious character. This is, on the one hand, intuitively their most obvious property, and, on the other hand, the most difficult to describe. There are different types of consciousness (for example, minimal consciousness or self-consciousness), but from the point of view of the mind-body problem, the most important one is referred to as phenomenal consciousness. It is impossible to define this type of consciousness in an unambiguous and non-circular way, which is why in the literature on the philosophy of mind the intuitive meaning of phenomenal consciousness is indicated by the use of the phrase "*what it is like*" (Nagel, 1974). Conscious subjects experience a specific phenomenal qualities (also referred to as *qualia*, singular *quale*) that accompany their own conscious states, allowing them to know what it is like to, for example, perceive the colour red, experience an itch, and so on. This knowledge is unattainable except through subjects' own phenomenal consciousness (cf. Hutto, 2011).

The natural sciences – at least those based on classical physics – provide a picture of the material world as subject to deterministic laws of nature (cf. Penrose, 2000). Specific causes necessarily produce specific effects, and every event in the material world has its causal explanation in another event of a material world. However, the pre-scientific but elementary intuitions of folk psychology suggest that we are dealing with four types of causal interactions: mental–physical, physical–mental, mental–mental, and physical–physical. In the mind-body problem, the first of these plays a key role. We experience that the conscious 'I' can, in an undetermined way, causally influence the body it belongs to through acts of free will (for example, by making a decision to raise a hand). In other words, everyday experience suggests that mental acts can initiate a new

causal chain in the physical world in an undetermined way. The question arises as to how to reconcile these two seemingly incompatible perspectives – the scientific and the everyday. This problem can be framed as a trilemma, that is, a set of three theses, each of which appears to be true, yet the conjunction of any two entails the necessity of rejecting the third:

1. Mental phenomena are distinct from physical phenomena, meaning they cannot be fully explained in terms of physical phenomena.
2. Mental phenomena causally affect physical phenomena.
3. The physical world is causally closed, meaning every event in the physical world has a necessary cause in another event within the physical world.

Thesis (3) is an elementary assumption of the natural sciences, while (1) and (2) reflect our everyday experience. Statements (1), (2) and (3) are incompatible because (1) and (2) imply that the physical world is not causally closed; (1) and (3) imply that, while mental phenomena exist, they do not causally affect physical phenomena; in turn, (2) and (3) imply that there exist no mental phenomena distinct from physical phenomena, as the former are reduced to the latter. Solving the mind-body problem by rejecting (1) and (2) characterises physicalist (also called naturalist or materialist) positions, while the rejection of thesis (3) is proposed by proponents of interactionist dualism (for example, Descartes), who emphasise thesis (2). In turn, the rejection of (2) is supported by both proponents of psychophysical parallelism and representatives of contemporary epiphenomenalism. The former accept the parallelism of the course of mental events and physical events while denying the possibility of interaction between them (for example, Malebranche).The latter assert that mental phenomena exist but these phenomena do not influence physical phenomena even though they are themselves influenced by them (Bremer, 2010, pp. 30–32).

Historical analysis of the term

The historically dominant view characteristic of the modern era was substance dualism. In accordance with the assumption typical of this period – that the order of cognition delineates the ontic order – modern

thinkers, beginning with Descartes, assumed the existence of two radically different kinds of substances: bodies and minds. This assumption was based on the idea that in cognition we are given two radically different types of properties: physical (primarily spatiality) and mental (primarily non-spatiality). Religious inspirations also played an important role in the dominance of the position of substance dualism, as this position aligns with the religious conception of the soul. Descartes is considered the classical substance dualist. He argued that the human being is a peculiar agglomeration of body (matter), which has a mechanical nature, and soul (mind), which is immortal and capable of existing independently of the body.

The problem of mind-body causation is particularly difficult (or even impossible) to solve for substance dualists due to their emphasis on the radical ontic distinction between mind and body and due to their attribution of opposite properties to these substances. A sort of substance dualism, whose representatives maintain that although this interaction is extremely difficult to explain (or may even be an inexplicable mystery) it is nonetheless a fact (both in terms of mental–physical and physical–mental interaction), is referred to as interactionist dualism. An attempt to overcome the problem of mind-body interaction, while preserving the thesis of the ontic separateness of spiritual and bodily substances, is the proposition of psychophysical parallelism. Proponents of this position assume either that only two types of effective interactions exist – mental–mental and physical–physical – or that there are no direct interactions at all. This view was popular in the modern era among followers of Descartes' philosophy. Representatives of psychophysical parallelism from this period believed that God was responsible for correlating changes in the physical and mental orders: either God actively causes these changes each time (for example, Malebranche) or He foresaw all changes and established their harmony at the beginning of the world's existence (for example, Leibniz). Thus, modern psychophysical parallelism paid a very high price for its proposal to solve the problem of interaction – namely, it adopted a very strong metaphysical assumption regarding the existence and nature of God, which itself requires philosophical justification (Iwanicki, 2012, pp. 37–41, 48–51; cf. Kim, 2011, pp. 31–60).

Substance dualism negates a fundamental assumption of the natural sciences and, for this reason, was largely replaced by the positions of

reductive physicalism in the 20th century. Nevertheless, substance dualism still seems an attractive philosophical conception as it has certain advantages over materialist theories, which assume a metaphysical monism stance about the ultimate nature of reality. Substance dualism better explains the grave problem of personal identity and enables the preservation of our pre-scientific intuitions about possessing free will, which allows us to initiate new causal chains in the world, countering the determinism of the natural world. It is worth emphasising that this issue is far from trivial, as our belief that we possess free will is a foundational assumption of social life (including democratic systems), in which we assign rights and responsibilities to individuals.

Reductive physicalism positions that gained popularity in the 20th century include (1) logical behaviourism, (2) type identity theory, and (3) eliminative materialism (eliminativism). These positions assume that it is possible to reduce mental states to physical states, after which it turns out the mind-body problem is not a real problem as the interaction it refers to is simply the interaction of what is physical with what is also physical. Common to these positions is the assumption of metaphysical monism regarding the ultimate nature of reality, according to which there exists only one type of substance, namely material substance.

In the 1920s and 1930s, under the influence of the Vienna Circle, the representatives of logical behaviourism (for example, Gilbert Ryle), a view considered the strongest form of reductive physicalism, argued that mental phenomena could be described in intersubjectively accessible behavioural categories as dispositions to specific behaviours. They advocated elimination of the term ‘consciousness’ and all other mental terms by replacing them with terms referring to observable events (for example, taking a painkiller and saying it hurts in the case of feeling pain). This reduction was supposed to be of an analytical necessity nature (by virtue of definition) (Heil, 2020, pp. 50–70). Representatives of eliminative materialism (for example, Paul Churchland), starting from the premise that only science – ultimately physics – determines what truly exists (so-called scientific realism), argued that mental language should be replaced by physical language as the former refers to something that does not really exist. In their opinion, folk psychology is a pre-scientific proto-theory that will be entirely eliminated over time by a scientific description of human neurophysiology, much like the

theory of combustion as oxidation replaced the 17th-century phlogiston hypothesis, according to which phlogiston was supposed to be a material substance that vaporises during combustion. According to the representatives of eliminative materialism, just as the phlogiston hypothesis became completely redundant in light of modern physicalist theories, so too will mental language be phased out from our discourse as neuroscience advances (Heil, 2020, pp. 166–177). Another refined version of reductive materialism was type physicalism or type-type identity theory. Its representatives (for example, John Smart), inspired by examples of successful identifications in natural science (for example, lightning being an electrical discharge), postulated the reduction of mental events to physical events by establishing the identity of certain types of mental states and certain types of neurophysical states. For example, in all human beings, pain K is identical to an increase in C_4 fibre activity. This reduction was intended to be of an empirical necessity nature (by virtue of the law of nature). According to Smart, 'sensations' (mental states) and 'brain processes' are two different names that refer to the same object (analogous to the terms 'lightning' and 'electrical discharge'), with one being a colloquial expression and the other a scientific term (Heil, 2020, pp. 71–87).

Programmes of reductive physicalism faced harsh criticism. For example, representatives of logical behaviourism were accused of proposing a reduction that is unworkable because not every mental event is linked to a behaviour. We can imagine beings that are completely immune to pain and, despite experiencing it, do not exhibit any behavioural response. Similarly, a person who is not experiencing pain could perfectly simulate the behaviour of someone experiencing it. Eliminativism can be criticised as a self-defeating position: if eliminativists are convinced that no beliefs exist, they thereby undermine their own view. Hilary Putnam (Putnam, 1975) presented a famous critique of the type-type identity theory, which is called the charge of multiple realisability. Type-type identity theorists equate a particular type of mental events with exactly one type of physical events. However, it is highly doubtful that there exists exactly one type of neuronal process that is associated with, for example, all instances of pain in all beings that feel pain. For example, such a primitive animal as the octopus also appears to feel pain, although in the case of these creatures it is realised on the basis

of different neuronal processes than those we encounter when human beings feel pain (Bayne, 2022, pp. 38–55, cf. Kim, 2011, pp. 61–127).

Discussion of the term

Reductive physicalism programmes formulated in the 20th century faced serious criticism, resulting in a loss of popularity. The most significant objections to these positions relate to the previously mentioned issue of the phenomenal qualities of conscious states (*qualia*). Of particular importance in the discussions on this topic were articles published in the second half of the 20th century by Thomas Nagel and Frank Jackson. In his text *What Mary Didn't Know* (Jackson, 1986), Jackson presented a famous thought experiment describing a neuroscientist called Mary, who has been locked in a black and white room since birth. She possesses all available knowledge of the physical aspects of colour vision but has never seen colours other than black and white. After leaving the room, Mary experiences colour vision for the first time and thus becomes acquainted with the phenomenal qualities of colours (*qualia*). Intuitively, it seems that Mary has acquired new knowledge that she did not (and could not) have had before; namely, she learned what it is like to perceive colours. Jackson's reasoning can be reconstructed as follows:

1. Physicalist knowledge is not all knowledge.
2. If physicalism is true, then physicalist knowledge is all knowledge.
3. Physicalism is false (Iwanicki, 2012, p. 66).

The thought experiment involving Mary aims to demonstrate the truth of premise 1.

In a similar vein, Nagel argues in his article *What Is It Like to Be a Bat?* (Nagel, 1974) that none of us knows what it is like to orient ourselves in space through the ability to echolocate, as bats do. Physiologists may know all about bats, but they will never possess first-person, phenomenal knowledge of what it is like to be a bat. According to Nagel, physical descriptions of the world and all scientific knowledge are always non-exhaustive because they leave out the phenomenal aspect of conscious states, which can only be cognised from a first-person perspective

(Chalmers, 2003; Bayne, 2022, pp. 133–153; Iwanicki, 2012, pp. 54–63, 65–68; cf. Bremer, 2005, pp. 193–212).

These arguments are referred to as arguments from knowledge because they defend the irreducible character of knowledge concerning one's own conscious states. At the same time, they are treated as strong arguments in favour of property dualism, which assumes the reality and uniqueness of both physical and mental properties. Thus, they serve as important counter-arguments against reductive physicalism, whose representatives advocate for the reduction of mental properties to physical properties. In light of, among others, the criticism mentioned here, positions classified as non-reductive physicalism have gained popularity. Their representatives claim that, although only material bodies exist in a substantive way (thus they assume a materialistic metaphysical monism), some bodies also possess, in addition to physical properties, mental properties that cannot be reduced to the former. Thus, under non-reductive physicalism, on the one hand, the fundamental assumption of natural science regarding the ultimate material structure of the natural world is maintained; on the other hand, mental states are acknowledged as real, although they are not ascribed the ontic status of substance.

Non-reductive physicalism assumes three key theses: 1. "There are mental properties that are distinct from any physical properties". 2. "Mental properties depend on physical properties". 3. "Mental properties make a causal contribution to what happens [in the world – E.O.]" (Baker, 2009, pp. 110–111).

As far as the first thesis is concerned, it means that non-reductive physicalists accept the aforementioned property dualism. In their view, material beings may, but need not (as is the case with inanimate beings), simultaneously possess two types of properties: physical and mental. Thesis (2) is formulated in very general terms, and its proper specification constitutes the most important and challenging task for representatives of non-reductive physicalism. Following Donald Davidson, it is commonly accepted to refer to the relationship mentioned in this thesis as supervenience (from the Latin *super*, meaning 'on' or 'above', and *venire* meaning 'to come'). This term suggests the superstructure of second-order properties (or phenomena), in this case mental, on top of first-order properties (phenomena), which serve as the basis – in this case physical properties. Advocates of this term seek to define it in such

a way as to denote a relationship strong enough to imply a dependence of mental states on physical states, while at the same time weak enough not to entail a reduction of the one to the other. For example, according to Davidson, a set of properties A supervenes on another set B if and only if an object that changes in terms of mental properties also changes in terms of physical properties (cf. Davidson, 1970, p. 212).

Thesis (3) implies the rejection of epiphenomenalism by non-reductive physicalists. Epiphenomenalism is the view that while mental properties exist and differ from physical properties, they are merely correlates (epiphenomena) of physical phenomena (neural processes) and never play a causal role in the physical world. Non-reductive physicalists aim to defend the assumption that is present in everyday experience and scientific psychology that mental states (and events) possess causal power. Furthermore, most non-reductive physicalists want to remain consistent with the scientific picture of the world (or, rather, to enhance this picture with a philosophical theory of mind) and thus unquestionably accept the assumption of the causal closure of the physical world – that is, the assumption that every event in the physical world has its sufficient cause in another event in the physical world (see thesis (3) of the trilemma outlined above) (Baker, 2009, pp. 109–113; Heil, 2020, pp. 178–184).

The programme of non-reductive physicalism has faced severe criticism from Jaegwon Kim, who has put forward, among other arguments, the so-called argument from overdetermination. In this argument, he contends that if a particular mental event, let us call it M1, causes another mental event, M2, then there exists such a physical event P1 that serves as the basis on which M1 supervenes; simultaneously, there exists another physical event P2 that serves as the basis for M2. If the material world is causally closed, which is the aforementioned fundamental assumption of the natural sciences, then P2 has its sufficient cause in P1. Consequently, it seems that M2 is overdetermined, meaning it has two causes. On the one hand, it is caused by M1; on the other hand, if M2 supervenes on P2, then the sufficient cause of P2, which is P1, also serves as a sufficient cause for M2. According to Kim, if mental events are not identical to physical events, then any event that has its cause in a mental event is overdetermined – it has two sufficient causes. Kim compares this situation to a scenario in which two assassins

independently fatally shoot the same victim at exactly the same moment. In his opinion, it is improbable that this type of situation would occur in every case of mental causation (Baker, 2009, p. 114). Kim's argument applies to any type of mental causation. While it was presented above using mental–mental causation as an example, a similar situation arises concerning mental–physical causation. If M represents the decision to raise one's hand, P denotes the corresponding physical neural event that serves as the basis for M, and E refers to the movement of the hand, then Kim's reasoning can be reconstructed as follows:

1. P is causally sufficient for E (COMPLETENESS).
2. If P is causally sufficient for E, then nothing that is distinct from P causes E unless E is overdetermined (EXCLUSION).
3. M is distinct from (i.e., non-identical to) P (DISTINCTNESS).
4. This is not a case of overdetermination (NON-OVERDETERMINATION).

Therefore:

5. M doesn't cause E. (Bayne, 2022, p. 181).

In the light of Kim's argument, mental causation appears questionable because all events have their own sufficient physical causes. If Kim's reasoning, including the premises he has accepted, is correct, then non-reductive physicalism entails epiphenomenalism, regardless of the declarations and intentions of its proponents. This conclusion could be rejected if any of the premises were implausible. The completeness thesis is closely related to the aforementioned assumption of the natural sciences regarding the causal closure of the natural world. While it is theoretically possible to have a situation of overdetermination (as in the case of death from two simultaneous gunshots), such an explanation for every mental event seems unlikely. Therefore, accepting the theses of exclusion and non-overdetermination seems justified. To preserve the thesis of mind-body causation, it remains necessary to reject the distinctness thesis. This, in turn, implies a return to reductive physicalism in the form of a theory that posits the identity of mental states and physical states. According to such a reductionism, just as a forest fire is caused by an electrical discharge and a lightning bolt, which is not a situation of overdetermination because they are exactly the same phenomenon, so too is the movement of my hand caused by physical neural processes that are identical to my mental states. The conclusion of Kim's argument seems to be this: either we accept

the identity of mental and physical states, or we recognise that they are not identical but only physical events have causal power. In both cases, our belief that mental phenomena (that is, acts of free will) can induce new non-determined causal chain in the physical world (see thesis (2) of the trilemma), remains a pipe dream (Bayne, 2022, pp. 180–183; Kim, 2009, pp. 38–46; Kim, 2011, pp. 214–223; Baker, 2009, 113–116; Heil, 2020, pp. 184–193).

The intensive research on artificial intelligence (AI) suggests that special attention should be paid to the position of functionalism, the first formulations of which date back to the mid-20th century. According to the basic idea of functionalism, a particular mental state should be equated with the function (or role) it performs in a properly organised system. Functionalists view physical states and mental states as two kinds of descriptions – material and functional – of a single substantial being. This position emerged from inspirations drawn from computer science, and its sources can be found in the works of the pioneer of this scientific area, mathematician Alan Turing. He proposed an understanding of minds in terms of calculating machines and understanding of mental states as results of computational processes in the brain. This approach forms the basis of machine functionalism, otherwise known as computationalism. The relationship between brain and mind is conceived in this form of functionalism analogously to the relationship between hardware and software. According to the original idea of functionalism proposed by Putnam, it is irrelevant what the material substrate of mental states is – while performing certain mental functions (for example, multiplication), the system can be either a biological organ, such as the brain, or an electronic system. This stems from the functionalists' acceptance of the thesis of the multiple realisability of mental states. Consequently, this conception enables the possibility of attributing mental properties not only to human beings but also to other living beings and even to devices (for example, computers). On the grounds of computationalism, the brain is identified with a computational machine that executes specific programmes, which are algorithms that can be expressed in numbers. Representatives of this position equate these algorithms with mental states. This conception has provided a very fruitful paradigm for cognitive research, including research on artificial intelligence. Computationalism assumes the so-called strong AI thesis, according to

which calculating machines can (at least potentially) think, just like living beings with elaborate biological brains (as opposed to the so-called weak AI thesis, which posits that machines can only simulate thinking) (Kim, 2011, pp. 129–167; Van Gulick, 2009, pp. 128–131; Heil, 2020, pp. 88–105).

A typical critique of functionalism points out, among other things, that performing specific functions – understood as producing the same effects as mental states – is not sufficient to attribute conscious states, which, according to this critique, is a fundamental property of mental states. This conclusion is supported by two highly influential arguments presented in the form of thought experiments: the Chinese room by John Searle, and the China brain, popularised by Ned Block. These arguments also undermine the thesis of the multiple realisability of mental states as they point out that not every type of matter can provide a basis for the emergence of consciousness and intentionality, which are the characteristic properties of mental states.

Against the strong AI thesis, and thus against computationalism, Searle formulated the famous Chinese room argument. In it, he presented the situation of a man locked in a room who has no understanding of the Chinese language at all. He is given queries in Chinese from the outside, to which he formulates responses based solely on an instruction book specified which characters of the Chinese alphabet to use in response to specific texts. This scenario is meant to simulate the so-called Turing test, the passing of which, according to functionalists, is supposed to prove the actual thinking of a particular system (a Turing machine). According to Searle, a person locked in a Chinese room can provide perfect answers and yet not understand Chinese in the slightest. In this situation, this person merely simulates the conscious mental states involved in understanding the language. Moreover, despite operating with Chinese characters – that is, the syntax of the language – the semantics, i.e., the reference of language to the external reality, remains completely beyond this person's reach. This person thus lacks what was earlier called intentionality, which is the characteristic property of mental states.

In the China brain thought experiment, it is argued that, by virtue of the principle of multiple realisability, a particular mental functional state can be realised by the entire population of China in such a way that

every Chinese person performs the same task that is a task of a single neuron. Every Chinese person receives certain inputs, to which they react according to the instructions they have, which specify what output they should give to the next Chinese person in a given situation. According to this critique, the execution of such commands by the entire population of China corresponds to how functionalists understand mental states. However, we are unable to identify where in this simulation of the brain composed of Chinese persons we are dealing with conscious phenomenal states (*qualia*). These states cannot be attributed to any single Chinese person (just as we cannot attribute, for example, the conscious perception of the colour green to a single neuron), let alone to the entire population of China involved in this experiment. This indicates that a functional understanding of mental states is insufficient as it overlooks their phenomenal nature (Braddon-Mitchell & Jackson 2007, pp. 107–128).

The critique formulated by Searle and Block remains to some extent consistent with Jackson's and Nagel's aforementioned arguments against reductive physicalism. Both the reduction of mental states to physical states and their reduction to functional states (especially if understood in a cause-and-effect way) overlook the conscious nature of mental states and thus fail to provide a sufficient description of what they are (cf. Van Gulick, 2009, pp. 143–149). This insufficiency of the functional conception of mind means that mental phenomena are not adequately captured on its grounds, consequently it is impossible for representatives of this position to give a satisfactory answer to the problem of mind-body interaction.

Functionalism, however, faces further difficulties related to the problem of interaction. Namely, like non-reductive physicalism, functionalism seems to entail epiphenomenalism. Because mental properties are captured functionally within this conception and the properties of the material basis are captured structurally, the relationship between the former and the latter is referred to as realisation (or implementation or execution). Just as a computer program is realised in a specific computer system, mental states are, in humans, realised in the biological matter of the brain. The functional properties (computer program or mental states) are here second-order properties that are realised on first-order properties, that is, on the material properties of the system

(of a computer or brain). From this perspective, it can be seen how functionalism is subject to a critique analogous to the one formulated by Kim toward the supervenience theory: the entirety of causal power falls to the material properties of the system. The functional state realised is merely a derivative of the material properties of the system which are responsible for the emergence of physical cause-and-effect chains. Consequently, functionalists should either recognise that mental causal power is only apparent and thus declare the position of epiphenomenalism, or they should capture the relationship between mental and physical properties not only as the realisation of a particular function but also as some kind of identity (of types or tokens). In both cases, functionalists will fail to resolve the mind-body trilemma, that is, to reconcile our three aforementioned intuitions. Instead, they will reject the second thesis of the trilemma in the first case and the first thesis of the trilemma in the second case (Van Gulick, 2009, pp. 142–143; Kim, 2009, pp. 46–48).

Systematic reflection with conclusions and recommendations

The solution to the mind-body problem depends on how mental states are understood. If they are reduced to physicalist states, the mind-body interaction will turn out to be illusory, as from this perspective it will simply be a physical–physical interaction. However, such an approach to mental states is unsatisfactory for a number of reasons, the most important being the omission of the conscious characteristic of mental states, especially experiential phenomenal qualities, known as *qualia*. Attempts to define mental states solely by the functions they perform in a complex system face a similar problem as these theories also seem to be 'blind' to the consciousness dimension of mental states. A system can realise mental functions without the phenomenal and intentional characteristics typical of mental states. If mental states are assumed to be irreducible and unique properties built on physical properties as second-order properties, then while their uniqueness can be preserved, the possibility of their causal agency is seriously questioned. In this approach, both defining the relationship between mental and physical

states as a relationship of supervenience and as a relationship of realisation (implementation) of a specific function leave no room for the real causality of mental states. This is the case because the principle of causal closure of the physical world means that the entirety of causal influence falls on first-order physical properties.

In spite of these difficulties, the vast majority of philosophers of mind do not attempt to defend the rather historical position of substance dualism. Some, like Daniel C. Dennett, advocate epiphenomenalism and thus deny the causal power of mental states. They regard folk psychology, within which it seems to us that such interaction occurs, solely as a useful fiction that allows us to predict and explain human behaviour, which in turn is essential for leading a normal social life. The functional description of mental states (especially when the behavioural aspect is emphasised in it) is useful from the perspective of scientific development. Therefore, it is still readily accepted by those researchers, who deal with the mind (not only philosophers), but the idea of functionalism is adapted to advances in the cognitive sciences. For example, it is pointed out that the functional level of consciousness and intentionality does not emerge through a simple division of the system into structural and functional levels. Sophisticated manifestations of the activity of the mind, such as high intelligence, intentionality, and self-awareness, emerge thanks to the integration of the more elementary subsystems that make up the human brain. Thus, the idea of functionalism should be considered at numerous interdependent levels, rather than expecting the emergence of such advanced mental functions in a single 'step' that would constitute the entire transition from the physical to the mental level (Van Gulick, 2009, pp. 138–141; cf. Bremer, 2005, pp. 151–192). Today, the thesis of the multiple realisability of mental states is also limited – the emergence of conscious states seems to be closely related to the properties of the biological organic matter of neuronal systems. Moreover, in the cognitive sciences, under the banner of the 'embodied mind', the indispensable role of the human body in the creation and functioning of all mental phenomena is emphasised. Contemporary understandings of the mind thus dissociate from Cartesian substance dualism and the foundational assumptions of this position. The problem of consciousness, especially its advanced forms, including phenomenal consciousness (*qualia*), seems to pose a fundamental puzzle today, the

resolution of which is essential for providing the answer to the mind-body problem (cf. Bremer, 2005).

References

Baker, L.R. (2009). Non-Reductive Materialism. In: A. Beckermann, B.P. McLaughlin, & S. Walter (ed.), *The Oxford Handbook of Philosophy of Mind* (pp. 109–127). Oxford: Oxford University Press.

Bayne, T. (2022). *Philosophy of Mind*. *An Introduction*. Abingdon: Routledge.

Braddon-Mitchell D., & Jackson F. (2007). *The Philosophy of Mind and Cognition*. Oxford–Cambridge: Blackwell Publishing.

Bremer, J. (2005). *Jak to jest być świadomym. Analityczne teorie umysłu a problem neuronalnych podstaw świadomości*. Warszawa: Wydawnictwo IFiS PAN.

Bremer, J. (2010). *Wprowadzenie do filozofii umysłu*. Kraków: Wydawnictwo WAM.

Chalmers, D.J. (2003). Consciousness and its Place in Nature. In: S.P. Stich, & T.A. Warfield (ed.), *The Blackwell Guide to Philosophy of Mind* (pp. 102–142). Malden: Blackwell Publishing.

Davidson, D. (1970). Mental Events. In: *Essays on Actions and Events: Philosophical Essays Volume 1* (pp. 207–224). Oxford: Clarendon Press.

Heil, J. (2020). *Philosophy of Mind. A Contemporary Introduction*. New York: Routledge.

Hutto, D. (2011). *Consciousness*. In: J. Garvey (ed.), *The Bloomsbury Companion to Philosophy of Mind* (pp. 35–53). London: Bloomsbury.

Iwanicki, M. (2012). Dualizm psychofizyczny. Odmiany, argumenty, zarzuty. In: M. Miłkowski, & R. Poczobut (ed.), *Przewodnik po filozofii umysłu* (pp. 37–84). Kraków: Wydawnictwo WAM.

Jackson, F. (1986). What Mary Didn't Know. *The Journal of Philosophy*, *83*(5), 291–295.

Kim, J. (2009). Mental Causation. In: A. Beckermann, B.P. McLaughlin, & S. Walter (ed.), *The Oxford Handbook of Philosophy of Mind* (pp. 29–52). Oxford: Oxford University Press.

Kim, J. (2011). *Philosophy of Mind*. Boulder: Westview Press.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450.

Penrose, R. (2000). *The Large, the Small and the Human Mind*. Cambridge: Cambridge University Press.

Putnam, H. (1975). Philosophy and Our Mental Life. In: H. Putnam, *Mind, Language and Reality: Philosophical Papiers*. Vol. 2 (pp. 291–303). Cambridge: Cambridge University Press.
Van Gulick, R. (2009). Functionalism. In: A. Beckermann, B.P. McLaughlin, & S. Walter (ed.), *The Oxford Handbook of Philosophy of Mind* (pp. 128–151). Oxford: Oxford University Press.

Marek Pokropski
University of Warsaw
https://orcid.org/0000-0002-7423-6951
https://doi.org/10.35765/slowniki.413en

The naturalisation of phenomenology

Summary

DEFINITION OF THE TERM: This section provides a concise description of the term 'phenomenology' and explores various interpretations and approaches to understanding it. Additionally, it identifies and delineates the related research area.

HISTORICAL ANALYSIS OF THE TERM: This section presents the process of defining the concept of phenomenology. It traces the evolution and emergence of the research area and highlights the historical determinants and key debates that have shaped its development.

DISCUSSION OF THE TERM: This section applies the previously discussed arguments to address current theoretical and practical challenges identified by the Author.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: This section outlines various ways of addressing contemporary theoretical problems from different perspectives within science. It also discusses potential solutions proposed by researchers, emphasising those that should be considered for practical implementation.

Keywords: phenomenology, Husserl, naturalisation, cognitive science

Definition of the term

The naturalisation of phenomenology refers to the non-reductive integration of phenomenological insights into the cognitive sciences. This interpretation of naturalisation is endorsed by the editors of the volume *Naturalizing Phenomenology* (Roy et al., 1999), which compiles various theoretical approaches on the subject. It is important to emphasise that this integration is intended to occur at the level of explanation as the goal of naturalisation is to align phenomenology with explanations that adhere to the standards of the natural sciences and to demonstrate continuity between phenomenology and naturalistically oriented cognitive science. Proponents of this approach believe that naturalising phenomenology will advance our understanding of mental phenomena. While a fully satisfactory theory of naturalisation has yet to be formulated, several intriguing proposals have been made regarding how this integration might be achieved.

Two clarifications are necessary regarding the above definition. First, in this context the term 'phenomenology' refers to the philosophical movement initiated by Edmund Husserl in the early 20th century, which emphasises the description and analysis of acts of consciousness and lived experience. The goal of phenomenological research is to uncover the structure and origin of acts of consciousness. Apart from Husserl, key figures often associated with the naturalisation of phenomenology include Martin Heidegger, Jean-Paul Sartre, and Maurice Merleau-Ponty. Merleau-Ponty in particular holds a prominent position with his phenomenology of the body, which has significantly influenced the concept of embodied cognition in contemporary cognitive science. It should be noted that in analytic philosophy the term 'phenomenology' is used of mind in a slightly different sense where it refers to the subjective nature of experienced sensory qualities, often termed *qualia* (e.g., Chalmers, 1996). The naturalisation of phenomenology in this context would involve explaining how the phenomenal properties of mental states arise from physical processes in the brain. Although sensory qualities are not a central focus of phenomenological philosophy as derived from Husserl, this does not preclude phenomenology from contributing to the exploration of these issues.

The second point concerns the understanding of naturalism. Broadly speaking, naturalism can be divided into two key positions: ontological

naturalism, which asserts that everything that exists is part of nature; and epistemological or methodological naturalism, which upholds the validity of the investigative and explanatory methods used in the natural sciences. In the context of the naturalisation of phenomenology, ontological naturalism is understood in a non-reductive sense. This means that the proposed continuity between phenomenology and cognitive science should preserve the properties identified by phenomenology rather than eliminating or replacing them with those studied in cognitive science. However, the project of naturalising phenomenology primarily addresses methodological concerns, advocating a shift from reductionism to integrationism. This approach emphasises that the naturalisation of phenomenology should not entail abandoning its methods or reducing explanations of the mind solely to terms of natural science. Instead, it aims to methodologically bridge these research traditions by integrating phenomenological insights with findings from the cognitive sciences, particularly neuroscience.

Historical analysis of the term

The relationships between phenomenology and cognitive science began to solidify in the 1990s. Prior to this period, phenomenology primarily acted as a critic of cognitive science, particularly challenging the computational theory of mind that was dominant at the time. This theory proposed that cognitive processes could be understood as the processing of symbolic representations, analogous to the functioning of computers. A notable figure in this critique was Hubert Dreyfus, who, drawing from Heidegger, criticised the artificial intelligence project and emphasised the need for embodying and situating AI in the world. Dreyfus was also among the first philosophers to engage Husserl's philosophy within cognitive science discussions, interpreting Husserl as a precursor to the computational theory of mind (e.g., Dreyfus and Hall, 1982). In this interpretation, Husserl's concept of a noema could be seen as analogous to the notion of mental representations produced by computational processes. In the 1990s, the classical computational paradigm began to decline, giving rise to an interest in topics like consciousness and the embodiment of cognitive agents, which are the core concerns of phenomenology.

Consciousness has entered cognitive science as the 'hard problem', which is linked to the explanatory gap: a challenge that, according to some, lies beyond the current capabilities of cognitive science and may require a solution from a yet-to-be-defined future science (e.g., Chalmers, 1996). Chalmers argued that the phenomenal properties of mental states – such as the experience of pain, the sensation of the colour red, or the smell of coffee – resist causal and functional explanations within cognitive science. Although Husserl's phenomenology does not primarily focus on sensory qualities (termed hyletic data by Husserl, which are imbued with meaning by intentionality), it offers significant insights into the intentional structure of experience, which is crucial for explaining consciousness. In this context, phenomenology was brought into cognitive science by Francisco J. Varela, the founder of neurophenomenology and a key proponent of the naturalisation project. Varela regarded phenomenology as a "remedy for the hard problem" (Varela, 1996), convinced that naturalising phenomenology would pave the way for a scientific method capable of investigating subjective experiences.

The second key issue is the impact of embodiment on cognitive processes, such as how an organism's body – its structure and motor capacities – affects perception, memory, and other cognitive functions. Alongside James Gibson's work in ecological psychology, phenomenology has significantly contributed to the development of the embodied cognition paradigm. Phenomenology, particularly through Husserl's later works, such as *Ideas II* (1989) and Merleau-Ponty's *Phenomenology of Perception* (2012), explores the relationship between the body and consciousness and how this relationship influences cognitive processes and our being and acting in the world. Phenomenological analyses of the body emphasise the crucial role of tactile and kinaesthetic (movement-related) sensations in shaping our experience of our own corporeality and of material objects. The structure of perceptual experience reveals that our body provides a unique perspective from which we perceive the world. Spatial relations – such as near/far and left/right – are projected from the body's location, from which we see a certain profile of objects; it is through our body's motor capacities that we can change perspectives and perceive different profiles of objects. Our bodily interactions with objects, especially those used as tools, demonstrate how these objects quickly become natural extensions of our capabilities. In

general, phenomenologists argue that our corporeality fundamentally shapes consciousness, influencing perceptual processes, spatial orientation, emotional experiences, and more. Moreover, it is through the body, which is both expressive and intentional in its interactions with the world, that we come to understand other subjects, i.e., other embodied consciousnesses.

In addition to its insights on consciousness and corporeality, phenomenology brings to cognitive science a sophisticated analysis of intentionality. Originating in Franz Brentano's descriptive psychology, intentionality refers to the property of conscious acts that consists in being directed toward an object, or, in another interpretation, the possession by mental states of representational content, through which they relate to the objects they represent. Intentionality was a central theme in Husserl's phenomenology. He attributed to it the function of imbuing sensory data with meaning and distinguished its various forms, such as perceptual intentionality, intentionality in emotions, signitive intentionality in speech acts, and memory intentionality in recollection. For instance, one can see a bear, feel afraid of it, talk about encountering it in the Tatras, or recall the experience later. In Husserl's phenomenology, recognising and describing these varieties of intentional acts is essential to understanding consciousness and its relationships to the world. Merleau-Ponty's (2012) phenomenology extends the concept of intentionality to the body, positing that the body itself manifests a primary intentionality in its actions and existence in the world. While cognitive science has drawn on phenomenology's insights in this area, it appears that these contributions have not been fully integrated or explored so far.

Husserl extended beyond the static analysis of intentionality, which focuses on the typology and structural description of intentional acts, by introducing a genetic perspective aimed at understanding the genesis of acts of consciousness. This approach is well illustrated by the observation that intentional acts of consciousness unfold and change over time, e.g., sensory content varies, and unconscious factors, such as bodily experienced affective arousal, influence these acts. For this reason, Husserl (1964) placed great emphasis on the study of time-consciousness, which he saw as consisting of two key functions: retention and protention. Retention refers to the preservation in consciousness of what has just passed, while protention involves the anticipation of what

is yet to come. These functions enable acts of consciousness to extend over time, maintaining the continuity and identity of both the act of consciousness and its object. This analysis of time-consciousness is among phenomenology's most significant and frequently cited contributions to cognitive science research on consciousness. Concepts derived from Husserl's phenomenology of time-consciousness have been applied to various areas, including the study of perception, bodily movement, social interaction, and the understanding of mental disorders such as depression.

Husserl's critique of naturalism

To fully appreciate the relationship between phenomenology and naturalistic philosophy, it is worth revisiting the first half of the 20th century and considering Edmund Husserl's own stance on the matter. In his works, Husserl openly criticised attempts to reductively explain consciousness using purely naturalistic methods and the terminology of the natural sciences. He argued that proponents of naturalism treat every object of study as merely part of nature, governed by strict laws that can be described mathematically, as is the case with mathematical natural sciences. However, Husserl saw this approach as fundamentally flawed, stemming from the mistaken belief in a universal mathematical scientific method that could be applied across all fields, including the study of consciousness. For Husserl, this was an absurdity: firstly, because consciousness is not a mere part of nature like stones or stars; secondly, because the structure and content of experience cannot be fully captured by causal or formal laws. Therefore, he argued that a phenomenological method which is capable of grasping the unique structure of experience is necessary to explain consciousness in a meaningful way.

Importantly, Husserl did not deny the connections between consciousness and its physical, organic substratum. He acknowledged the relationship between consciousness and the body, as evidenced in extensive passages in the second volume of *Ideas* (1989). Moreover, Husserl was not opposed to empirical research in psychology and psychophysics, such as that conducted by Wilhelm Wundt, Gustav Fechner, or proponents of Gestalt psychology. On the contrary, he recognised

the value of such research. However, he contended that empirical psychology lacked a sufficient theoretical foundation and therefore only captured consciousness in a superficial manner, thus being incapable of fully explaining it. Husserl believed that a reform of psychology was necessary that would ground the discipline in phenomenology. In his 1925 lectures, later published as *Phänomenologische Psychologie* (1962), Husserl proposed a new psychology – phenomenological psychology – which he envisioned as an intermediary discipline between transcendental phenomenology and empirical psychology. The task of phenomenological psychology, according to Husserl, would be to study the essence (or structure) of acts of consciousness, such as forms of perceptual consciousness, and to provide empirical psychology with concepts that cannot be developed within a naturalistic framework, thereby offering a solid theoretical foundation for the discipline.

In summary, Husserl's position is not anti-naturalistic but firmly anti-reductionist. He criticised a reductionist form of naturalism that seeks to study consciousness exclusively through the methods of mathematical natural sciences while disregarding the subjective dimension of experience. Husserl argued for the necessity of reforming psychological research by grounding it in a phenomenological theoretical framework. However, this does not imply that Husserl would support contemporary efforts to naturalise phenomenology. On the contrary, it seems likely that he would oppose any attempt at naturalisation that diminishes phenomenology's foundational and guiding role. In many current proposals for a naturalised phenomenology, this foundational role is precisely what is compromised, reducing phenomenology to just one of many disciplines contributing to the interdisciplinary field of cognitive sciences.

Discussion of the term

In the debate on the possibility of naturalising phenomenology, three main positions can be distinguished. The first position firmly rejects this possibility, arguing that the essence of phenomenology lies in its transcendentalism (e.g., Moran, 2013), which seeks to illuminate the conditions of possibility for cognition itself. This interpretation positions phenomenology as a foundational discipline that cannot be naturalised.

The second position explores the potential for collaboration between phenomenology and cognitive science, suggesting that phenomenology could be adapted by stripping away its transcendental aspects (e.g., Pokropski, 2021). This general perspective offers a wide range of possibilities for methodological collaboration. For instance, Dan Zahavi (2010) advocates for a loosely defined partnership based on mutual inspiration that could lead to collaborative research projects. Similarly, Shaun Gallagher (2003) discusses the "mutual enlightenment" of phenomenology and cognitive science, proposing a concept called front-loaded phenomenology. This approach suggests that phenomenological concepts could be integrated early in the design of cognitive science experiments. More rigorous approaches within this collaborative framework focus on the idea of the constraints that phenomenology and cognitive science might impose on each other. The nature of these constraints remains a subject of ongoing debate (e.g., Pokropski, 2021; Roy et al., 1999). Some scholars suggest mapping the structures of experience onto neural activity using formal relations like isomorphism or homomorphism. Others, such as those working in neurophenomenology, explore the concept of generative passages – a somewhat ambiguous idea – to describe a new kind of reciprocal relationship (e.g., Lutz & Thompson, 2003). Still others (e.g., Pokropski, 2021) argue that the constraints provided by phenomenology should be understood through the mechanistic model of explanation that is dominant in neuroscience. This approach involves using phenomenology to narrow down the range of possible mechanisms responsible for a given phenomenon by identifying the relevant properties of these mechanisms.

A third perspective in the debate on the naturalisation of phenomenology suggests that the goal should not be to naturalise phenomenology, but rather to phenomenologise nature: essentially, to redefine the concept of nature through a phenomenological lens and thereby initiate a paradigm shift in the natural sciences (e.g., Gallagher, 2018). This view begins with Husserl's critique of the natural sciences, which often overlook the subjective dimension of cognitive processes and mistakenly treat nature as something 'ready-made', to be discovered by 'impersonal' and neutral scientific methods. In contrast, phenomenology, in the spirit of Kant, emphasises the active role of the cognising subject in constituting the nature that he cognises. In other words, the nature we cognise is one that reveals itself to a subject or a community

of subjects. Proponents of this view argue that the concept of nature should incorporate aspects of subjectivity. This could be achieved by introducing terms such as intentionality, embodiment of consciousness, agency, and the situating of the subject within a practical context into the vocabulary of the natural sciences.

Evaluating the prospects for the naturalisation project in light of the positions discussed above, it is clear that the first position effectively closes the debate by asserting that transcendentalism cannot be integrated with the natural sciences. While it is true that Husserl's phenomenology evolved towards transcendental philosophy (e.g., in the first volume of *Ideas)*, limiting Husserl's project solely to its transcendental aspects overlooks its considerable thematic diversity and evolution over time. Furthermore, even if naturalising Husserl's phenomenology is deemed impossible, this does not preclude the naturalisation of other phenomenological approaches, such as Merleau-Ponty's phenomenology of the body. The third position appears too radical as it advocates for a substantial reform of the natural sciences – a change that may be difficult for the scientific community to accept. Additionally, Husserl's critique of naturalism, which the proponents of this view rely on, is somewhat outdated. It is based on a deterministic view of nature governed by universal laws, whereas modern science, including fields like quantum mechanics, acknowledges a more complex non-deterministic nature. Furthermore, contemporary science has largely moved beyond the idea of a reductive unification of scientific disciplines based on a strict deductive-nomological model. Modern cognitive science, especially through the concept of embodied cognition, increasingly incorporates aspects such as embodiment, causation, and situated cognition. Therefore, among the three positions, the second – advocating for a mutually constraining relationship between phenomenology and cognitive science – seems the most viable and is supported by the majority of proponents of naturalisation.

Mathematisation

A recurring topic in the discussion of naturalising phenomenology is the potential application of mathematics to analyse phenomenological descriptions of the structures of experience. Some argue that this

application might be crucial for the success of naturalisation efforts (e.g., Roy et al., 1999). The benefit of this approach lies in its ability to rigorously express the properties of experience in a format accessible to other researchers. Mathematisation could also facilitate the translation of phenomenological and natural scientific representations into a universal language of numbers.

Husserl's own stance on applying mathematics to phenomenology was notably ambivalent. Despite being trained as a mathematician, he criticised the mathematisation of the natural sciences for contributing to what he saw as a crisis in rationality (e.g., *The Crisis of the European Sciences and Transcendental Phenomenology*). Husserl argued that phenomenology, unlike mathematics, is not a formal discipline but a material one, concerned with acts of consciousness and their correlated objects of experience. He believed that the nature of this correlation and intentional contents cannot be adequately captured by strict mathematical formulas, suggesting that applying mathematical methods to consciousness might be misguided (*Ideas I*). Nevertheless, Husserl also recognised the significant role of mathematical methods in the success of modern science, particularly the application of geometry to the description of the material world and the formalisation of the laws of nature. He frequently employed quasi-mathematical notations in his own work. For instance, in *The Logical Investigations*, he used symbolic notation to explore the hierarchical structure of acts of perception and recollection, and in *The Phenomenology of Internal Time-Consciousness*, he employed diagrammatic representations to illustrate the flow of consciousness. Thus, Husserl's view on the relationship between mathematics and phenomenology appears to be somewhat inconsistent.

Roy, Petitot, Pachoud, and Varela (1999) suggest that Husserl's reluctance to apply mathematics to the study of consciousness stemmed from the limitations of the sciences available in his time. For instance, during Husserl's era, dynamical systems theory was primarily confined to physics, while today this theory extends its application to a range of fields, including ecology, biology, and cognitive science. Dynamical systems theory focuses on describing and analysing the behaviour of complex systems, particularly how this behaviour evolves over time. It employs formal tools such as differential equations, where variables represent key properties of the systems under study and parameters

denote external conditions. The system's behaviour is formally captured through a set of these equations, which can also be visualised as trajectories in an *n*-dimensional state space, where *n* corresponds to the number of variables. Varela argues that dynamical systems theory offers a promising avenue for the naturalisation of phenomenology. According to Varela, dynamical models are well-suited to representing the time-varying properties of subjective experience.

Selected methodological proposals: neurophenomenology

Neurophenomenology, introduced by Francisco Varela (1996), a Chilean biologist and neuroscientist, aims to create a research method that integrates phenomenology and neuroscience. Its primary objective is to address the hard problem of consciousness, i.e., clarifying the relationship between subjective experience and the underlying neural processes. Inspired by Husserl's phenomenology, particularly the concept of phenomenological reduction, Varela posits that it is possible to reliably investigate the structures and properties of experience from a first-person perspective and that these descriptions can be combined with neuroimaging data, such as electroencephalography (EEG). Neurophenomenology is interested in identifying general structures of experience that are common across multiple subjects, rather than describing individual experiences or their qualitative aspects. As Varela explains:

> The novelty of my proposal is that disciplined first-person accounts should be an integral element of the validation of a neurobiological proposal, and not merely coincidental or heuristic information. This is why I choose to describe the situation by the hypothesis that both accounts be mutual constraints on each other (Varela, 1996, p. 344).

Thus, the role of phenomenology extends beyond merely describing experience; it is also essential in building a scientific theory, for example, by validating the results of neurocognitive research.

Neurophenomenology is based on three pillars:

1. (NPh1) First-person data from the careful examination of experience with specific first-person methods.
2. (NPh2) Formal models and analytical tools from dynamical systems theory, grounded in an embodied-enactive approach to cognition.
3. (NPh3) Neurophysiological data from measurements of large-scale, integrative processes in the brain. (Lutz &Thomson, 2003, p. 34).

The method of neurophenomenology is grounded in the reliability of introspective reports (point 1 above). Instead of relying on traditional introspective methods, neurophenomenology employs a unique first-person approach. Here, an experimenter and a participant collaboratively develop descriptive categories that most accurately capture the structural aspects of the experience being studied. Participants then describe their experience using these categories, thus ensuring that these descriptions are not influenced by externally imposed concepts or theories but are rooted in terms that suit participants. Participants are trained in selected elements of the phenomenological method, including a simplified version of phenomenological reduction, in order to suspend common-sensical notions and theoretical beliefs about the essence of experiences and mental states. Another component of the neurophysiological method is the recording of participants' brain activity using EEG (point 3). This data is then analysed to identify correlates with descriptions of experience and the neurophysiological data. Another component is dynamical systems theory (point 2), which is used to establish relationships between first-person descriptions and third-person data. This theory is based on the assumption that both phenomenological phenomena and neuronal processes exhibit dynamic properties that can be effectively captured by dynamical systems theory.

One of the earliest (and still scarce) neurophenomenological experiments was conducted by Lutz and Thompson (2003) and investigated perceptual readiness. In this study, participants were shown a blurred image from which a geometric shape gradually emerged. They were asked to indicate the moment they became aware of the shape. To categorise their perceptual readiness, pre-determined categories were used: steady readiness, fragmented readiness, and unreadiness. The goal was to correlate these categories with EEG patterns recorded during the experiment. This pilot study demonstrated that phenomenological methods could enhance the quality of first-person data. However,

the integration of the second pillar of the neurophenomenological method – formal dynamic models – was not fully achieved. The integration of descriptive categories and EEG patterns was based on simple correlations over time, without examining how specific EEG parameters relate to the phenomenological categories of perceptual experience. Moreover, the study did not present a dynamic model that would integrate the experiential and neural aspects.

In summary, neurophenomenology represents a non-reductionist approach to naturalised phenomenology. Its aim is to provide reliable descriptions of experiences under study and identify their structural properties, while also potentially validating empirical studies of consciousness. However, until neurophenomenology develops dynamic models that effectively integrate both phenomenological and neuroscientific domains, it remains a promising yet theoretical proposal.

Selected methodological proposals: front-loaded phenomenology

Front-loaded phenomenology proposes a way to naturalise phenomenology by incorporating key phenomenological concepts and theories into experimental practice (e.g., Gallagher, 2003). This approach to naturalisation is relatively flexible as it does not necessitate extensive training in phenomenological methods nor a deep understanding of phenomenological theories by researchers or participants. Research teams use phenomenological concepts, notions, and distinctions early in the research process, during the experiment design phase. Examples of phenomenological concepts used in this phase include Merleau-Ponty's ideas and their contemporary elaborations on the body schema and body image, the sense of agency, the sense of ownership of one's body, as well as Husserl's notions of time-consciousness and intentionality. Gallagher explains:

> the phenomenology that is front-loaded might be developed in pure phenomenological analysis (as in Husserl), or in neurophenomenological experiments. Second, it is a natural consequence of front-loading phenomenology that, as in the neurophenomenological model, the phenomenology becomes part of the analytic framework for interpreting the results, and not just part of the data to be analysed. (Gallagher, 2003, p. 92).

An illustrative example of a phenomenologically inspired study is found in research on the sense of agency. In a study (see Gallagher, 2003), Gallagher's phenomenological concept of the sense of agency was employed to create experimental conditions where subjects experienced a disturbance in their sense of agency related to a bodily action – specifically, moving a point on a screen using a joystick. Concurrently, the participants' brain activity was monitored using fMRI. The findings indicated that when participants perceived the action as being caused by someone else, there was a correlation with increased activity in the right inferior parietal cortex. Conversely, when the action was perceived as self-generated, increased activity was observed in the anterior insula.

In evaluating the project of front-loaded phenomenology, it should be observed that this approach prioritises research pragmatism over detailed methodological rigour. Advocates of this method might argue that if collaboration between phenomenologists and cognitive scientists proves effective in the laboratory, then it is a viable approach worth pursuing. However, the specific role of phenomenological concepts within the research process remains somewhat ambiguous. One possibility is to treat phenomenological concepts as analogous to other scientific or colloquial psychological concepts that inform the conceptualisation of research phenomena and the design of experiments. Yet, this approach does not clarify why phenomenological theories would be preferable or superior to other conceptual frameworks. Another option, as suggested by Gallagher, is to use phenomenological concepts to formulate research hypotheses that can then be empirically tested. However, this use of phenomenological concepts to generate hypotheses does not align with the current applications of front-loaded phenomenology, which often use phenomenological concepts more to define *explanandum* than to generate testable hypotheses, as illustrated in the sense of agency study. A separate and more complex issue is whether phenomenological claims can be empirically tested at all, but Gallagher contends that it is possible and suggests that naturalisation could lead to 'mutual enlightenment' between phenomenology and cognitive science. The third option, which the quote above hints at, assigns to phenomenology an analytical role and a role in interpreting findings. In this case, phenomenological concepts would be used not just to describe phenomena but to provide explanatory insights. In summary, while front-loaded phenomenology

offers a straightforward and natural approach to integrating phenomenological concepts into empirical research, it is crucial to further clarify how these concepts should be applied throughout the entire research process.

Systematic reflection with conclusions and recommendations

Evaluating the current state of the debate on the naturalisation of phenomenology, it is evident that while there has been notable interest and some innovative theoretical proposals, significant progress remains elusive. Despite increasing engagement with phenomenology among cognitive scientists and neuroscientists and attempts to apply phenomenological theories in cognitive science research, the naturalisation process has not achieved any clear progress. A key challenge is the methodological integration of phenomenology with the cognitive sciences, a task that remains inadequately addressed. One reason for this lack of progress may be the insufficient use of the tools offered by philosophy of science to reflect on the issue. Specifically, the relationship between phenomenology and various types of scientific explanations needs further examination. For instance, the exploratory nature of naturalisation as integration – a feature central to its definition – is often overlooked or inadequately addressed in proposed solutions. It is crucial to assess how phenomenology aligns with different types of explanations used in the cognitive sciences, such as functional, dynamic, and mechanistic explanations, and to determine whether phenomenology converges with these explanatory models. Additionally, it should be determined whether phenomenology offers its own specific type of explanation, and if so, what it entails (for a detailed discussion, see Pokropski, 2021).

A critical issue is phenomenology's lack of explanatory potential. Phenomenology, using a method of reduction, often abstracts from causal relations in the world and replaces them with motivational relations in the case of mental phenomena (e.g., Husserl, 1962). This poses a challenge for its naturalisation, which aims to demonstrate continuity between phenomenological descriptions of lived experience and the causal processes described by the natural sciences. The resolution of

the problem of mental causality may significantly impact the success of naturalisation efforts. While phenomenology cannot offer scientific causal explanations, it provides valuable tools for first-person description and analysis of mental phenomena. For instance, phenomenological research has enhanced our understanding of cognitive processes by incorporating the aspect of embodiment. Additionally, phenomenological methods and concepts have been applied to gain insights into mental disorders such as depression and schizophrenia.

A potential solution to the problem of causality in the naturalisation of phenomenology is to approach it as an explanatory integration within the framework of the mechanistic model of explanation (Pokropski, 2021). The mechanistic model is widely recognised in cognitive science and neuroscience as a dominant explanatory framework (e.g., Craver, 2007). According to this model, explaining a phenomenon involves identifying the causal mechanisms responsible for producing it. Therefore, to explain mental phenomena, one would seek the underlying neuronal mechanisms that give rise to them. Given the complexity of mental phenomena, a comprehensive explanation requires integrating knowledge from diverse fields such as neurophysiology, psychology, and computational modelling. The mechanistic model facilitates this integration by constraining the space of possible mechanisms through the identification of key properties and hypothetical mechanisms, ultimately aiding in the characterisation of the appropriate mechanism (see, e.g., Craver, 2007). Despite its strengths, the mechanistic model has limitations, particularly when addressing complex mental phenomena like consciousness and its disorders, which necessitate first-person methods. Here, phenomenology can play a crucial role. While phenomenology may lack direct explanatory power, it offers detailed descriptions and analyses of the structure of experience, which can help identify the limitations of explanatory models. Thus, integrating phenomenological and mechanistic approaches presents a promising avenue for the naturalisation of phenomenology.

The project of naturalising phenomenology remains ongoing and open to further development. Should a robust and effective proposal for naturalised phenomenology emerge, it could mark significant progress in the quest to construct a comprehensive science of consciousness. Regardless of whether this project ultimately succeeds, exploring the

relationship between phenomenology and the natural sciences proves to be highly valuable. This understanding is crucial not only for phenomenology, which has historically played a role as both a guiding and foundational discipline in scientific rationality and as a side-discipline, but also for the natural sciences, which, despite their roots in modern mathematical naturalism, still grapple with the enigmatic nature of phenomena such as consciousness.

References

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

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

Dreyfus, H., & Hall, H. (1982). Introduction. In: H.L. Dreyfus, & H. Hall (ed.), *Husserl, intentionality and cognitive science* (pp. 1–27). Cambridge: MIT Press.

Gallagher, S. (2003). Phenomenology and Experimental Design: Toward a Phenomenologically Enlightened Experimental Science. *Journal of Consciousness Studies, 10*(9-10), 85–99.

Gallagher, S. (2018). Rethinking nature: Phenomenology and a non-reductionist cognitive science. *Australasian Philosophical Review*, *2*(2), 125–137.

Husserl, E. (1962). *Phänomenologische Psychologie: Vorlesungen Sommersemester 1925*, The Hague: Nijhoff.

Husserl, E. (1989). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy: Second Book Studies in the Phenomenology of Constitution*. Trans. R. Rojcewicz & A. Schuwer. Dordrecht, Boston & London: Kluwer Academic Publishers.

Husserl, E. (1964). *The Phenomenology of Internal Time-Consciousness*. Trans. J.S. Churchill. Bloomington, Indiana: Indiana University Press.

Lutz, A., & Thompson, E. (2003). Neurophenomenology: Integrating Subjective Experience and Brain Dynamics in the Neuroscience of Consciousness. *Journal of Consciousness Studies, 10*(9-10), 31–52.

Merleau-Ponty, M. (2012). *Phenomenology of Perception*. Trans. D.A. Landes. New York: Routledge.

Moran, D. (2013). 'Let's look at it objectively': Why phenomenology cannot be naturalized. *Royal Institute of Philosophy Supplements*, *72*, 89–115.

Pokropski, M. (2021). *Mechanisms and consciousness: integrating phenomenology with cognitive science*. New York: Routledge.

Roy, J.-M., Petitot, J., Pachoud, B., & Varela, F.J. (1999). Beyond the gap: An introduction to naturalizing phenomenology. In: J. Petitot, F.J. Varela, B. Pachoud, & J.-M. Roy (ed.), *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science* (pp. 1–80). Stanford: Stanford University Press.

Varela, F. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies 3*(4): 330–49.

Zahavi, D. (2010). Naturalized phenomenology. In: S. Gallagher, & D. Schmicking (ed.), *Handbook of phenomenology and cognitive science* (pp. 3–19). Dordrecht: Springer.

Jacek Bielas
Ignatianum University in Cracow
https://orcid.org/0000-0001-9771-4162
https://doi.org/10.35765/slowniki.457en

Embodied cognition

Summary

DEFINITION OF THE TERM: The concept of embodied cognition refers to the crucial role played by the various aspects of the body, not only the nervous system, in interaction with the environment as a constitutional component of the cognitive system.

HISTORICAL ANALYSIS OF THE TERM: Contemporary research on embodied cognition originates primarily from a critique of the classical view of cognitive processes formulated within computationalism. Computationalism posits that cognitive processes, which occur mainly in the brain, involve the computational processing of information through symbolic representations of the objects of cognition. In response to this critique, embodied cognition has become the central idea of an interdisciplinary research project known as embodied cognitive science.

DISCUSSION OF THE TERM: Despite its scientific foundation, the concept of embodied cognition remains ambiguous, and the embodied cognitive science project lacks coherence. Importantly, the role of corporeality in cognitive activity is still debated, with positions ranging from various versions of computationalism at one end, to 'minimal' and 'extreme' embodied cognition at the other.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: The current exploration of embodied cognition includes various approaches, leading to subject-matter and methodological diversity. This diversity presents challenges, such as difficulties in replicating research. Nonetheless, the development of embodied cognitive science

has the potential to broaden the range of issues addressed and enhance the coherence of these issues within cognitive science. Achieving this requires ongoing theoretical and empirical efforts.

Keywords: cognitive processes, corporeality/body, computationalism, embodied cognition, 4E

Definition of the term

In the literature, the notion of embodied cognition is associated with the view that cognitive activity in living organisms (including humans) and robots is conditioned by the functioning of their entire structure, not just isolated parts, like the brain or processor. Approaches to embodied cognition emphasise the role of various aspects of corporeality in cognitive processes, with particular emphasis on the sensorimotor properties of the body acting to achieve specific goals in its environment. For example, orienting covert visual attention involves motor preparation and motor readiness to move the eyeball towards a specific location or object in space. Similarly, the emotion of fear can be understood as the experience of preparing and readying internal organs and muscles to flee from, e.g., a bear. The concept of the 'self' is seen as the experience of being extended, yet limited, in space, which configures the body schema to 'reach' for objects in the world and use them for its purposes. Cognitive activity, understood in terms of corporeal mobility, is experienced through the body's resistance and gravity, which are inherent in its physical nature. Through its own resistance and gravity, the body experiences the resistance and gravity of other objects in the world, which necessitates force and effort.

The motor activity of the corporeal structure in the environment requires the coordination of its components within the internal space of the body in relation to the external environment, often under time pressure. The brain plays a central role in this coordination, but neuronal processes are necessarily influenced by what they coordinate. This view of corporeality which remains in close relation to the environment is a constitutive element of the cognitive system: body – brain – environment. However, the dynamics of the interactions between these components remain open to question.

It is suggested that the sensorimotor and utilitarian nature of cognition may be rooted in evolutionary determinants (Wilson, 2008). In phylogenetic development, humans emerged relatively recently; before that, the primary role of living organisms was to coordinate their movements to achieve specific effects in space, such as approaching and contacting or moving away from specific objects. While humans have unique capacities for conscious reflection, long-term planning, abstract

thinking, and out-of-context reasoning, the cognitive value and specific nature of the concept of embodied cognition within an evolutionary framework is compelling because it helps explain the development of human cognitive abilities based on sensorimotor skills.

The issue of embodied cognition is addressed either independently or in an interdisciplinary manner across various fields, including philosophy, linguistics, psychology, neuroscience, sociology, computer science, and robotics. In recent decades, interest in this subject has intensified, and embodied cognition has become a mainstream concept in research on cognition and the functioning of both living organisms and machines. It is now a leading idea in the interdisciplinary research program known as embodied cognitive science, although the many theoretical research approaches and proposed solutions to the issue have their roots in the early beginnings of European thought. However, it is essential to emphasise the development of new methods, particularly experimental ones, for verifying different approaches to embodied cognition, primarily within the empirical sciences. Moreover, new circumstances and contexts have emerged that allow this issue to be addressed in novel ways: for example, fresh opportunities for exploring embodied cognition are provided by advances in computational intelligence algorithms, including avatars and chatbots, as well as developments in robotics.

Historical analysis of the term

The anti-Cartesian roots of the concept of embodied cognition

Although the issue of the status of the body in the entirety of cognitive functioning has roots in Platonic idealism and Aristotelian hylemorphism, it is generally considered that modern reflections on this issue began with Cartesian dualism, or, more precisely, its critique. Descartes viewed the human body as an extended material object, perceived by the non-extended mental agent as merely an object that always exists in a specific place and time, yet separate from the mind. In his *Notes Directed Against a Certain Programme*, Descartes suggested that the body transmits stimuli from the external world to the mind via sensory

organs. These stimuli are not ideas but are converted by the mind into ideas, implying a 'sign' character of sensations as a "reconstruction of signals transmitted by the nervous apparatus" (as quoted in: Czerniak & Michalski, 2008, p. 14). In modern cognitive science, the ideas, i.e., 'signs' resulting from the body's processing of stimuli, could be interpreted as representations, and the stimulus processing could be seen as computational (although, obviously, Descartes himself did not provide the foundation for such an interpretation).

At the turn of the 19th century, the work of French thinker Maine de Biran offered an approach that challenged the rigid boundaries of Cartesian dualism. In Cartesian thought, the relationship between *res cogitans* (soul, mind, or intellect) and the body as *res extensa* is essentially a cause-and-effect relationship. In contrast, Maine de Biran observed that the inner space of one's body is given in acts of direct apperception and that within this inner space the self extends as a force expressed through effort and movement. This effort is perceived as a self that 'strives to move', while simultaneously encountering resistance from the body and, through it, resistance from other objects in the world" (Maine de Biran, 1970, p. 55). Thus, the body is no longer seen merely as a lifeless extension of the mind, a physiological fact, or an object in the world but is instead internally moved and experienced viscerally and kinaesthetically.

Corporeality in phenomenology

In Maine de Biran's description of corporeality, it is easy to notice an analogy to the intentional structure of consciousness later articulated in phenomenology. Prominent phenomenologists, such as Husserl, Heidegger, and Merleau-Ponty, emphasised the integral role of corporeality in the totality of experience. They analysed how the body, including its physical aspects, influences cognitive processes related to acts of a temporal, spatial, attentional, kinaesthetic, and social nature. Edmund Husserl (1983) introduced a key distinction between the body as a physical and biological entity (*Körper*) perceived from the third-person perspective, and the lived body (*Leib*) experienced from the first-person perspective.

Maurice Merleau-Ponty (2012) is recognised for his most advanced treatment of the body understood not only in spatial terms but also through experiential dimensions. Central to Merleau-Ponty's characterisation of the status of corporeality within the totality of cognitive activity is the notion of the body schema. This concept encompasses spatial and temporal orientation in the world and the motor functions required for concrete practical tasks. Merleau-Ponty builds his concept of corporeality on the fundamental notion of 'pre-reflective intentionality', which, as spatial motility, underpins all forms of man's (i.e., body-subject) activity while he directs himself towards concrete objects in the environment in order to experience, understand, and manipulate them. Thus, the space of the body and the external space form an inseparable practical system where the body serves as the common ground for the cognition of objects in space as targets for action. The initial Cartesian 'I think' transforms into a corporeal 'I can', which is always situated within the world.

In his neurophenomenology project, Francisco Varela (1996) proposed combining the physical and biological characterisation of the body from the third-person perspective with insights into the subjective sphere of the individual's experiences from the first-person perspective. This proposal aimed to address the difficult problem of consciousness, which involves the explanatory gap identified by Levine (1983) between the phenomenal nature of experience and the physical and biological nature of the body (brain). Regardless of the effectiveness of neurophenomenology as a project for naturalising phenomenology, it is worth emphasising its significant impact on the development and practical application of methods that integrate phenomenological approaches with experimental psychology and neurology. An example is the neurofeedback method (see, e.g., Petitmengin, Navarro, & Le Van Quyen, 2007), which allows agents to gain insight into their neuronal processes as correlates of mental states, thereby enhancing their capacity for self-monitoring psychological processes. Additionally, it is worth mentioning here that techniques for controlling mental states by influencing bodily states have been known and practiced for thousands of years, as evidenced by the yoga system.

A critique of computationalism as a contemporary context for the origin of the concept of embodied cognition

In the mid-20th century, Alan Turing's work on artificial intelligence led to a paradigm shift in our understanding of cognition. This new framework likened the operation of the mind to that of a computer, reducing cognitive processes to arithmetic operations on symbols representing properties of the external world. This perspective became known as computationalism. A key characteristic of computationalist approaches is the depiction of cognition as a process comprising two consecutive stages: transducing the energy of a stimulus into a symbolic representation, then processing these symbols according to specific rules. This leads to specific outcomes in the form of producing grammar structures of languages, extracting sequences of words from other words, solving logical problems, or recognising stimuli in relation to stored patterns or three-dimensional representations of the world. These operations occur within the nervous system, which enables stimuli to be encoded as a language of thought, akin to programming languages in computers. Thus, within the computationalist framework, which dominates cognitive science, the mind (or more specifically, the brain) is described as a biological computer that algorithmically operates with symbols that represent elements of the external world.

In the mid-twentieth century, the approach to cognitive processes in computational terms dominated in cognitive science in areas such as language theory, attention, visual perception, and problem-solving theories. However, several limitations of this computationalist perspective were identified, e.g., critics raised concerns about how algorithmic processes in the brain could lead to conscious and qualitatively distinct experiences and how these internal processes could represent states of the external world. One significant aspect of this critique was the overly narrow view of the cognitive system's boundaries, as is evident in both classical computationalism and its connectionist variant. This critique refers to the 'brain-centric' approach, which assumes that cognition occurs exclusively 'in the head'.

Ecological psychology

James Gibson's (1966) work on ecological psychology led to a radically different approach to cognition, drawing heavily on concepts from phenomenology, particularly Maurice Merleau-Ponty's idea of 'pre-reflective intentionality' and Martin Heidegger's notion of 'ready-to-hand'. Gibson's central premise is that bodily structure and motor capacities for orientation and action form the foundations for cognitive processes in both humans and animals. His focus is primarily on visual perception, described as direct perception of the environment that is not mediated by other cognitive processes but is dependent on the organism's location and motor activity. This perspective addresses the challenge posed by the limited stimuli reaching sensory organs within a given timeframe, which can hinder adequate cognitive interpretation. While computational approaches argue that this information is processed in the brain by integrating bottom-up and top-down processes with cognitive structures stored in it, ecological psychologists argue that adequate interpretation of stimuli occurs through the organism's continuous movement and integration of elements of the cognised object in its entirety. Thus, perception here is not merely the reception of stimuli from the environment which become data subject to computational operations (i.e., brain operations occurring entirely in the head), it is a holistic process realised through the entire body interacting within the environment. Interestingly, the concept of cognition as fundamentally motor-dependent is also reflected in contemporary experimental research on visuospatial attention in the category of action for perception. For example, Craighero et al. (1999) demonstrated that the efficiency of object perception is influenced by the preparation of movements of the hands through their appropriate positioning for grasping objects.

The concept of affordances has become a key aspect of ecological psychology. Affordances are the abilities of an organism resulting from its physical properties (e.g., the state of matter or the shape of the body), which are determined by the features of an environment. This framework underscores that the totality of psychological functioning, including perception processes and the meaning of concepts in both humans and animals, are intricately linked to interactions between physical properties of objects and physical properties and motor capabilities dictated by

an organism's body structure. For example, the perception and the concept of an umbrella do not depend solely on the computational capacity for processing information in the head but also on what can be done with the umbrella, depending on the interactions between water and human skin and the fact that the human hand has a specific structure. Consequently, for other organisms (e.g., bats or fish) an umbrella may exist as something totally different or may even go unnoticed.

The corporeality of artificial intelligence in robotic systems

The development of robotic systems was initially dominated by the belief that cognitive processes could be reduced to operations performed on abstract symbols within the central modules of these devices. A notable example is Shakey, one of the first robots, developed at the Artificial Intelligence Laboratory at Stanford. Shakey's task was to navigate a room by avoiding certain objects and moving others. These tasks were executed using the STRIPS program, which operates on the sense-model-plan-act sequence. A camera registers elements of the environment, the computer transforms them into symbolic, i.e., abstract, models; the program compares them with predefined models stored in the computer's memory and then executes the device's action program, which is also pre-defined. This mechanism of robotic operation, based on the principle of creating representations, has been found to be inefficient and possibly inadequate. In his article *Intelligence without Representation* (1991), Rodney Brooks criticised this model of cognitive processes used in artificial intelligence and robotics research. According to the model, the central module performs computational operations on data that represent elements of the environment, after transforming them based on input received through external modules. This process is inefficient as it requires additional resources to create representations (Shakey was very slow). Brooks argued that instead of creating representations it is "better to use the world as its own model" (1991, p. 139) as this is sufficient when a system draws information directly from its environment, thanks to which it is able to adapt dynamically to changes in this environment. In designing and constructing his

Creatures, Brooks adopted an approach based on principles from Gibson's ecological psychology rather than a computational mechanism. His robots function through an elaborate system of sensors that register the environment and are directly linked to mechanisms responsible for performing specific moves using a system of effectors, without intermediate steps of creating environmental representations. In these devices, their structure and resulting physical capabilities are crucial, highlighting the need for software that handles motor activities. Brooks claimed that a representation system is superfluous and that a direct "perception to action" connection is sufficient (he often referred to action as behaviour). This approach bears a strong resemblance to behaviourism in psychology and the S-R (stimulus-response) formula. However, despite the effectiveness of Brooks' robots compared to previous machines based on computing and the significant impact of his conception on the robotics industry, the claim that his solutions lack a representation system continues to be debated. Generally, attention is drawn to the need for software to select information from the environment and to choose and trigger particular actions, which, depending on the understanding of the term adopted, may possess characteristics of a representation.

Discussion of the term

In the cognitive sciences, the current popularity of the idea of embodied cognition is commonly linked to the publication of *The Embodied Mind* in 1993 by Francisco Varela, Eleanor Rosch, and Evan Thompson. In this work, the authors challenged the formal view of cognitive processes as numerical operations on symbolic representations performed only in the brain. Instead, they proposed a multidimensional analysis of the mind – and thus also cognitive processes – including not only the brain but also the non-neural aspects of the body (physical, biological, and phenomenological) that interact closely with the environment. These researchers' proposal was instrumental in the emergence of the project of embodied cognitive science based on the notion of embodied cognition. Proponents of this view focus on the body itself and its entirety as crucial factors in cognition, which they understand in the broadest possible sense (Dąbrowski, 2021).

In metatheoretical contrast to opposing computationalist approaches, the field of embodied cognition involves several key issues. One issue is the proportion of computational and bodily factors in the totality of cognition and the dynamics of these proportions at different stages of the cognitive process. For instance, are purely computational factors dominant in the initial phase, while the body only assumes a leading role in later stages? Additionally, can defining the movement parameters for effectors such as the eyeball or the hand at the premotor stage of attentional orientation be described as purely computational programming, or does this stage already exhibit features of embodiment? If so, what would such embodiment consist in?

Nowadays, many researchers agree that embodiment is essential for any form of intelligence, whether natural or artificial (Ziemke, 2015). It is challenging to conceive of intelligence that exists solely as an abstract algorithm without its physical (bodily) concretisation. However, the precise nature of embodied cognition remains an open question that is closely tied to other questions, such as what is a body and what types can it assume? For instance, human cognitive processes are thought to be unique due to the specific characteristics of the corporeal agents realising them. This raises the question of what type of body AI systems (as models of natural intelligence) should possess. It is worth emphasising that despite the popularity of the project of embodied cognitive science, not everyone agrees that intelligence requires a body (Ziemke, 2015). Thus, the status of corporeality in the entirety of cognitive processes continues to be debated, with positions ranging from various forms of computationalism to ‘minimal’ and ‘extreme’ embodied cognition. This ongoing debate means that computational theories of mind are still relevant. A concrete example of this debate is the ongoing dispute in experimental psychology about the nature of visual attention. Proponents of the premotor theory of attention argue that it is linked to motor processes, while others believe it is independent of motor processes, at least in the case of endogenous visual attention. The categorial difference between theses about the mind and theses about cognition should be emphasised here.

A crucial issue in determining whether cognition needs to be embodied, and if so, to what extent, is the understanding of the body/embodiment itself. Ziemke (2015) identifies six understandings of embodiment in the

literature, noting the difficulties in defining them precisely: 1) structural coupling between agent and environment, 2) historical embodiment as a result of the history of structural coupling, 3) physical embodiment, 4) organismoid embodiment, i.e., organism-like bodily forms such as humanoid robots, 5) organismic embodiment of autopoietic living systems, and 6) social embodiment.

Additionally, Dąbrowski (2021) argues that a complete characterisation of corporeality should include its phenomenal aspects. This includes the image of the body, the body schema (which is primarily experienced as a feeling) as well as purely somatic sensations and sensations of an emotional and thus also motivational nature.

These various definitions of corporeality represent different understandings and approaches to the concept of embodied cognition.

Systematic reflection with conclusions and recommendations

Contemporary conceptions of embodied cognition: 4E cognition

The different understandings and aspects of corporeality form the basis for four leading approaches to cognition within the field of embodied cognitive science, collectively referred to as '4E cognition'(4E from the first letters of the four adjectives that describe these approaches): enactive (linked to action), embedded (rooted in the world), extended, and embodied (the most general, and encompassing all the others).

Enactive cognition

Enactive cognition views cognition as constituted through sensorimotor activity. This cognition is an emergent process resulting from the interactions between the brain, body, and environment. Initially proposed by Varela, Rosch, and Thompson (1993) as 'autopoietic enactivism', this approach equates cognition with life, suggesting that all living organisms, even bacteria, exhibit cognitive capacities to some extent, dynamically

reorganising their functioning in response to environmental conditions. 'Sensorimotor enactivism' emphasises the intentionality and experiential nature of perception, aligning with Merleau-Ponty's *Phenomenology of Perception*. In this view, perception involves motor exploration of the environment, which leads to establishing patterns of relationships between an organism's movements, the mental content of perception, and the world. Thus, what is perceived is determined by what an organism is capable of doing – perception is what we do. In 'radical enactivism', the concept of mental content as mental representations is eliminated, deemed adaptively ineffective and explanatorily redundant.

Embedded cognition

The relationship between cognitive activity and the environment assumes a different form in the perspective of cognition embedded in the world. Here, the environment is seen as a constitutive component of the cognition process itself, extending beyond the internal processes within the cognitive agent. The functioning of organisms, understood as open systems, depend heavily on the conditions of the environment in which they develop and function. Thus, these conditions influence their physiological structure, realisation of potential, and action patterns. In humans, this perspective considers not only natural determinants but also social and cultural realities.

Extended cognition

Within extended cognition, which is distinct from other forms of cognition through its explicitly philosophical nature, the environment is treated as a tool that potentially enhances task performance. This approach evolved from distributed cognition (Hutchins, 1996), which highlights the demanding nature of cognitive processes under time constraints, resulting in the need to shift the burden of cognitive activity to the environment, which thus becomes part of the cognitive system. For example, operating complex equipment like a large ship or aircraft requires distributing activities among several specialists. Also, organising books

on a shelf according to a specific system makes it easier to find the one that is needed later. Such scenarios demonstrate that an agent's cognitive capacities increase within a physically or socially organised environment. The concept of extended cognition, presented by Andy Clark and David Chalmers (1998) in *The Extended Mind*, is characterised by functionalism. It posits that the same cognitive function can be realised in different ways. If an action leads to a particular goal, it qualifies as a cognitive process, even if its realisation extends beyond the boundaries of the brain. Examples of this include tools that enhance the body's capabilities, and external media that serve as equivalents of mental content. The extended mind, understood functionally, is a cognitive system that transcends the traditional mind-body dualism.

Embodied Cognition

This concept serves as a common denominator for various perspectives that challenge formalist and brain-centrist views that explain cognition in terms of computational processes and symbol manipulation taking place only in the head. Instead, they emphasise the significant role of extracerebral and extra-neural aspects of the body that are closely tied with the environment as constitutive elements of the cognitive system. In *Beyond the Brain: How Body and Environment Shape Animal and Human Minds*, Louise Barrett (2011) writes that an organism's cognitive activity reflects "the kind of nervous system it possesses, which in turn reflects the kind of body it has, which in turn is influenced by the kind of ecological niche it occupies" (p. 11). Cognitive activity should be understood not only in its behavioural aspect but also in its phenomenal aspect. Thus, human cognitive activity fundamentally differs from that of animals due to differences in their bodies, nervous systems, and habitats. Consequently, even if the exploratory behaviours of humans and animals appear similar, a human cannot comprehend what it is like to explore the world as, for example, a bat.

The four leading approaches to embodied cognition (4E) within embodied cognitive science address and resolve constituent issues differently. This diversity has led to criticism of the 4E framework as heterogeneous, incoherent, and incomplete (Dąbrowski, 2021). Various

criteria can be used to classify these approaches, such as the dynamics of the relationships between the embodied agent and the environment, the boundaries of the cognitive system, the specific aspects of the entirety of cognitive functioning they focus on, and the degree of the embodiment of cognitive processes.

Shaun Gallagher (2010) proposed one such classification based on the degree to which the body influences cognitive processes. He distinguished five degrees: 1) weak or minimal embodiment, in which cognitive processes are primarily influenced by neural conditioning, largely separated from the environment; 2) biological embodiment, in which the organism's overall constitution and motor capacities influence cognition; 3) embodied semantics, in which the experience of corporeality is reflected in linguistic structures and the conceptualisation of the world (mainly through metaphors); 4) embodied functionalism, in which the body is part of a broad cognitive system that includes the brain, body, and world, though some situations involve non-corporeal internal (purely mental) representations; and 5) strong or radical embodiment, in which cognition relies entirely on dynamic relations between the brain, body, and world.

Contemporary research on the embodiment of cognitive processes encompasses an increasing spectrum of issues, including perception, attention, language, memory, learning, emotions, problem-solving, social cognition, and moral reasoning. The fact that there are various approaches to operationalising these issues, particularly in psychology, poses challenges in interpreting research findings. This complicates the replication of empirical studies in this area, although it reflects a broader replication crisis in the cognitive sciences. This difficulty likely stems from the subtle nature of the area of the mind as a sphere of private events.

Despite these challenges, there is growing encouragement to apply the conceptual framework of embodied cognition in designing scientific experiments and interpreting research findings (Barrett, 2011), which is what is happening. The development of embodied cognitive science has rekindled interest in concepts previously deemed too ephemeral in the empirical sciences, broadening the scope of research topics and enhancing the conceptual apparatus within their explanatory systems. An illustrative example is the project of neurophenomenology proposed

by Francisco Varela (1996) to naturalise phenomenology and address the crisis in the sciences of consciousness. Varela's proposal methodologically combines first-person perspectives, studied through phenomenological reduction, with third-person methods, such as modern neuroimaging techniques. This approach aims to discover the neurophenomenological circulation between previously unknown elements of the structure of experience and neuronal processes. While the efficacy of this approach in resolving the difficult problem of consciousness remains controversial, Varela's research program has led to practical applications, such as research into the neuropsychological determinants of the dynamics of epileptic seizures and potential self-control mechanisms (Petitmengin, Navarro & Le Van Quyen, 2007).

Integrating the various perspectives offered by sciences that explore mental phenomena and neural processes within the framework of embodied cognitive science represents a step towards a more coherent and comprehensive theory of mind. However, achieving this goal requires further theoretical and empirical efforts.

References

Barrett, L. (2011). *Beyond the Brain: How Body and Environment Shape Animal and Human Mind*. Princeton: Princeton University Press.

Brooks, R. (1991). Intelligence without Representation. *Artificial Intelligence*, *47*(1–3), 139–59. DOI: 10.1016/0004-3702(91)90053-M.

Clark, A., & Chalmers, D.J. (1998). The extended mind. *Analysis*, vol. 58, no. 1 (Jan., 1998), pp. 7–19.

Craighero, L., Fadiga, L., Rizzolatti, G., & Umiltà, C. (1999). Action for perception: A motor-visual attentional effect. *Journal of Experimental Psychology Human Perception Performance*, *25*(6), 1673–1692. DOI: 10.1037/0096-1523.25.6.1673ht.

Czerniak, S., & Michalski, R. (2008). *Cielesność, Kompensacja, Mimesis: Wokół pojęciowego instrumentarium współczesnej antropologii filozoficznej*. Warszawa: Wydawnictwo IFiS PAN.

Dąbrowski, A. (2021), Ucieleśnione poznanie – założenia, tezy i wyzwania. *Argument Biannual Philosophical Journal*, *11*(1), 13–32.

Gallagher, S. (2010), Interpretations of Embodied Cognition. In: W. Tschacher & C. Bergomi (ed.), *The Implications of Embodiment: Cognition and Communication* (pp. 59–71). Exeter: Imprint Academic.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.
Heidegger, M. (2004). *Being and Time*. Trans. J. Macquarrie & E. Robinson. New York: Harperperennial.
Husserl, E. (1983). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*. Trans. F. Kersten. Hingham, Massachusetts: Martinus Nijhoff Publishers.
Hutchins, E. (1996). *Cognition in the Wild*. Cambridge: MIT Press.
Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*, 354–361.
Maine de Biran, P. (1970). *The influence of habit on the faculty of thinking*. M.D. Boehm (ed.). Westport: Greenwood Press.
Merleau-Ponty, M. (2012). *Phenomenology of Perception*. Trans. D.A. Landes. New York: Routledge.
Petitmengin, C., Navarro, V., & Le Van Quyen, M. (2007). Anticipating seizure: Pre-reflective experience at the center of neuro-phenomenology. *Consciousness & Cognition*, *16*, 746–764.
Wilson, M. (2008). How did we get from there to here? An evolutionary perspective on embodied cognition. In: P. Calvo, T. Gomila (ed.), *Handbook of cognitive science: An embodied approach* (pp. 375–388). San Diego: Elsevier.
Varela, F., Thompson, E., & Rosch, E. (1991). *The Embodied Mind*, Cambridge: MIT Press.
Varela, F. (1996). Neurophenomenology. A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3(4): 330–349.
Ziemke, T. (2015), Czym jest to, co zwiemy ucieleśnieniem? *AVANT*, *3*, 161–174. DOI: 10.26913/60202015.0112.0014.

Józef Bremer
Ignatianum University in Cracow
https://orcid.org/0000-0001-9664-8896
https://doi.org/10.35765/slowniki.411en

The cognitive social sciences

Summary

DEFINITION OF THE TERM: This section presents a definition of the cognitive social sciences, understood as an interdisciplinary science that draws on research conducted within philosophy and neuroscience as well as within symbolic interactionism, which is a leading strand of the social sciences.

HISTORICAL ANALYSIS OF THE TERM: This section describes the emergence of two research areas: 1) the cognitive sciences, which include theoretical linguistics, psychology, philosophy, neuroscience, cybernetics, and artificial intelligence; 2) sociology, with a particular emphasis on the development of symbolic interactionism.

DISCUSSION OF THE TERM: In recent decades, the social sciences and the cognitive sciences have moved closer together. In this section, the cognitive social sciences are characterised by enlisting examples of emergence theory and neuroscience as components of cognitive science. This is done by referring to explanatory grounding and treating the cognitive sciences as having laid the foundations for the development of the cognitive social sciences. In contrast to solutions offered by the subject literature, mechanistic explanations are avoided in the cognitive social sciences, the theoretical framework of which offers an insight into the functioning of the individual mind and human behaviour, including integrative analyses that cover different levels of organisation: starting with genes and leading to social relations and cultural phenomena.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: The article discusses key issues related to the social

aspects of cognition, which allows social neuroscience concerning direct human interactions to be broadened to cover the entire range of areas where social diversity interacts with cognitive diversity. The emphasis is placed on conceptual and empirical problems that arise from serious attempts to analyse these issues, on new resources, and on considerations associated with such an approach. However, it is also necessary to revise theoretical social concepts to make use of these resources. The article offers several conclusions regarding both the understanding of the process of socialisation in categories which are cognitive sciences-friendly and the ethical problems associated with the development of the cognitive sciences.

Keywords: symbolic interactionism, social psychology, theories of emergence, social neuroscience

Definition of the term

The cognitive sciences is an interdisciplinary study of the mind that combines research conducted within philosophy, neuroscience, psychology, theoretical linguistics, and artificial intelligence. Its aim is to understand human reasoning and behaviour, and then to use this knowledge in, e.g., medicine, educational programmes, robotics, and building increasingly intelligent machines that simulate our natural intelligence. The origins of the cognitive sciences can be traced back to the mid-1950s, when researchers from several scientific disciplines combined their efforts in the development of theories of mind based on complex mental representations and computational procedures.

The social sciences focus on behaviours of individuals and entire communities, their interactions with one another, and their relationships with the natural and technological environments. Their aim is to understand the evolving processes in which we are involved, an example of which is the process of sustainable development for our planet. When we talk about the evolution of contemporary social structures, we mean their presence in the information society of the digital age.

At the intersection of these two research fields, new directions and programmes have emerged and are developing, such as cognitive sociology, political psychology, behavioural economics, cognitive anthropology, and the cognitive social sciences. According to Tuukki Kaidesoja and his colleagues:

> The cognitive social sciences can be defined as disciplines and research programs that aims to integrate the social sciences (e.g., sociology, anthropology, political science and economics) with the cognitive sciences (e.g., cognitive psychology, cognitive neuroscience and cognitive science) (Kaidesoja, Sarkia & Hyyryläinen, 2019, p. 1).

After briefly outlining the historical development of the cognitive sciences and the social sciences, we will then take a closer look at symbolic interactionism as a representative of currently leading sociological theories. On the one hand, it can be used in extending the scope of interdisciplinarity of the research conducted within the cognitive sciences; on the other hand, the results of this research can be used in conducting sociological research. Two related cognitive social science

projects will be discussed with reference to interactionism: one from the emergently understood philosophy of mind, and one from the social neurosciences.

Historical analysis of the term

The contemporary development of the cognitive sciences is closely linked to the cognitive turn in the studies of the mind (1950–1960). Before this period, behaviourism took centre stage in psychology and the philosophy of mind, which was a reaction to questioning the scientific validity of introspection used as the preferred method of research in psychology. It was assumed that reports on a person's inner mental life based on introspection and conducted in the first-person narrative elude external intersubjective verification.

The symposium on information theory held on 10–11 September 1956 at Massachusetts Institute of Technology (Cambridge, USA) was a milestone in the cognitive turn. It brought together the linguist Noam Chomsky (1928–) and pioneers of artificial intelligence research: Allen Newell (1927–1992), Herbert Simon (1916–2001), and Marvin Minsky (1927–2016). The basic premise of artificial intelligence is the use of algorithms capable of simulating human thought processes and behaviours and of performing tasks using human-like intelligence. Our intelligence enables us to learn, plan, adapt to and influence new social environments.

Chomsky, who is a nativist, severely criticised the behaviourist conception of language and instead developed his theory of universal and transformational grammar. Allen Newell, Herbert Simon, and Cliff Shaw presented the first computer program that performed automatic reasoning, which was considered the first artificial intelligence program. The possibilities for the multifaceted use of computer science inspired neurophysiologist and cybernetician Warren McCulloch (1898–1969) and logician Walter Pitts (1923–1969) to conduct in-depth research on a digital model of neurons, which led to a description of the functioning of the first neural network. In addition, important elements in revitalising this kind of interdisciplinary research included the assumptions of cybernetics formulated by Norbert Wiener (1894–1964) and the mathematical

model designed by Alan Turing (1912–1954) – the Turing machine – to implement algorithms, which contributed to the development of a computational model of the mind and a modular model of the brain. With time, the use of artificial intelligence has transformed numerous aspects of our social life: from the advent of mobile telephony to the development of language models based on GPT architecture (Generative Pre-trained Transformer), i.e., several new versions of ChatGPT.

The research agenda outlined during the aforementioned symposium was closely linked to work on the development of artificial intelligence. Its authors almost uncritically adopted from academic philosophy of the time the dominant theories about the functioning of human thinking (e.g., functionalism and the theory of the identity of mental and physical states), treating them as tools enabling the modelling of intentional states of the human mind. In doing so, they failed to notice that their views on the human psyche may be limited in scope (Bremer & Flasiński, 2022). The contemporary cognitive sciences, supported by non-invasive neuroimaging of the brain, is returning to questions about first-person conscious experiences and about the impact of society and culture on neural systems. This requires more-complex research programmes which would include the cognitive social sciences.

In most textbooks, Max Weber (1864–1920), Émile Durkheim (1858–1917), and Karl Marx (1818–1883) are presented as the founders of sociology as a distinct scientific discipline. Nevertheless, Charles A. Ellwood (1873–1946) and George Herbert Mead (1863–1931), whose works were inspired by the philosophy of social pragmatism, also exerted a huge impact on the development of the social sciences. In 1909, Ellwood wrote:

> The truth is, sociology and all the other social sciences must wait upon the development of biology and psychology; and these antecedent sciences are, even today, in an unsettled condition. How absurd, then, to demand that sociology shall have a settled body of theory, or else deny to it the name of science! None of the social sciences possesses a settled body of theory (Ellwood, 1909, p. 108).

Both researchers contributed to the development of Herbert Blumer's (1900–1987) sociological thought and thus also to the paradigm of sociological research he formulated, namely 'symbolic interactionism'

(Blumer, 1986), which constitutes a specific theoretical perspective for studying collective life and human behaviours. The continuous and evolving exchange of meanings and symbols is the key mechanism in the formation of social structures, which occurs in the course of all social interactions between actors who are aware that they are partners. This paradigm is often treated as research theory and practice in the fields of both social psychology and sociology. The pioneers of symbolic interactionism include Polish-American sociologist Florian Znaniecki (1882–1958) and William Thomas (1863–1947), who together published the monumental work *The Polish Peasant in Europe and America* (Thomas & Znaniecki, 1995).

The 1990s witnessed the development of qualitative research methods which contributed to a deeper understanding of interactionism in different social contexts. This theory continues to evolve in the 21st century, including the exploration of new interpretations and applications that integrate symbolic interactionism with contemporary sociological issues, thus testifying to the continued relevance and adaptability of interactionism (cf. Carter & Fuller, 2016). Socialisation, understood as the processes by which personalities and self-concepts are formed, values and attitudes are transmitted, and the culture of one generation is entrusted to the next generation, is another important area of research undertaken within the framework of symbolic interactionism.

From theoretical and practical perspectives, studies conducted within symbolic interactionism are of an interdisciplinary nature. Within this interactionism, both psychologists and sociologists use the term 'social psychology' to refer to their disciplines' fields of specialisation. The subdisciplinary differences between them at a conceptual level are evident: sociological social psychology draws its perspective from theories such as Blumer's symbolic interactionism and emphasises the impact of variables at an individual level on macroscopic social analyses in the study of topics such as social structure and group dynamics. Psychological social psychology emphasises the impact of the social context on variables at an individual level and studies effects such as social influence, group perception, and the change of an individual attitude. Over the past decades, the theoretical basis of these and other effects have changed many times and have emphasised a range of approaches, including functionalist, behaviourist, attributional, social-cognitive, and, most

recently, multimethodological and polytheoretic. Methodological differences are also present between these sub-disciplines: sociologically oriented social psychologists employ sociometric data, unstructured interviews, surveys, observational and archival techniques, and qualitative research methods based on the analysis of data collected during interviews, while psychologically oriented social psychologists prefer to rely on experimental methodologies. However, despite differences in the methodologies and theoretical foundations, social psychologists from both disciplines largely share the same research interests, namely individual and social human behaviours (Uldall, 2013). As we will see below, social neuroscientists refer to both disciplines: they use observable neural (physiological) processes and behavioural responses to draw conclusions about unobservable mental states. They assume that all levels of analyses are correlated (e.g., the states of the nervous system causally affect mental states, and these may in some way interact with responses occurring at the level of neurons). Since its inception, symbolic interactionism has opened up an empirical social science perspective on the study of human behaviours and group life (Blumer, 1986).

Polish sociologists worth particular mention for their research within the paradigm of symbolic interactionism include, first and foremost, Elżbieta Hałas, whose work includes the significant study *Interakcjonizm symboliczny. Społeczny kontekst znaczeń* (2012). In this study, as in her previous ones and numerous articles, she systematises issues related to the processes of symbolisation in the phenomena of social life and focuses on, among others, the links between culture and society. The development of symbolic interactionism in Poland is thoroughly discussed by Krzysztof Konecki and Anna Kacperczyk (Konecki & Kacperczyk, 2020). In the section below, we will use this theory as one example of the wide range of social sciences.

The end of the 20th century and the beginning of the 21st century brought tremendous developments in the biological, psycholinguistic, and neuropsychological sciences that Ellwood had looked forward to. These were possible thanks to the discovery of the structure of the nervous system (for which Santiago Ramón y Cajal and Camillo Golgi were awarded the Nobel Prize in medicine in 1906), the invention of new neuroimaging technologies, and the rapid development of artificial

neural networks that are capable of learning from examples and able to automatically generalise acquired knowledge. Thus, the social sciences have become an indispensable element of interdisciplinary cognitive sciences research. Recognition of the interactions between agents – or rather, the ability to recognise social interactions – is usually associated with aspects of higher-level social cognition, including mentalisation, i.e., the ability to understand one's own and others' behaviours based on the mental states and brain states correlated with them.

Discussion of the term

In the aforementioned article by Kaidesoja and his colleagues, these authors emphasise that their aim is to reconstruct the arguments for the cognitive social sciences, to specify their premises, inferential structures, and conclusions, and to evaluate these arguments by analysing their presuppositions, the plausibility of their premises, the validity of their conclusions, and the potential conceptual ambiguities. These arguments are (1) the argument from explanatory grounding, (2) the argument from theoretical unification, (3) the argument from constraints, and (4) the argument from complementarity. All four arguments justify the need for the cognitive social sciences, although, unlike us, the authors of the article do not discuss the results of different research programmes conducted by cognitive scientists working within the cognitive social sciences (Kaidesoja, Sarkia & Hyyryläinen, 2019). They conclude that they have managed to arrange the arguments in such an order that the argument from explanatory grounding justifies the strongest form of interdisciplinary integration of the cognitive and social sciences, while the argument from complementarity is its weakest form of justification.

Therefore, the argument from explanatory grounding is extensively analysed below, focusing – unlike Sarkia and his colleagues (Sarkia, Kaidesoja & Hyyryläinen, 2020) – on symbolic interactionism and its cognitive extension within non-reductionist theories of emergence and social neuroscience. Emergence is understood here as the emergence of qualitatively new social forms and behaviours which are the result of interactions between individuals and their minds and brain processes. Social neuroscience is a complex intersection of the methods and

studies conducted within social psychology and neuroscience, both of which analyse how neural and physiological processes affect social behaviours. Both social behaviours and the related neural processes can be observed and measured empirically.

Emergence of social behaviours

In recent decades, the interdisciplinary overlap of the scopes and methodologies of the cognitive sciences and social sciences has intensified. New disciplines and research programmes have emerged and are developing at the intersection of these two areas of science. These include cognitive sociology, political psychology, behavioural economics, and new research programmes in cognitive anthropology.

Following the line of the argument from explanatory grounding, we assume that the cognitive sciences directed at social problems deserve a place in the research field dedicated to symbolic interactionism because it reveals their ontological basis and thus deepens our understanding of social phenomena in general. This means that if we assume that, from the explanatory perspective, symbolic interactionism is rooted in the cognitive sciences, then the cognitive sciences cannot be based on issues of interest to the social sciences. In doing so, we adopt a non-reductionist, emergentist explanatory model that is known from the philosophy of mind of the following type: the component processes of the lower ontological level (physical, chemical, biological, etc.) of a complex system form the basis for the processes of the higher level that emerge from this lower level (general systemic features – mental, social, etc.). In theories of emergence – following the common-sense principle – it is usually assumed that higher-level features are not reducible to lower-level objects, processes, and features: the whole (a system) is more than the sum of its parts. The emergent features of a given system are ontologically new features in relation to the basic features. Although it is generally accepted that lower-level objects and processes causally interact with higher-level processes, objects, and features, it remains debatable whether emergent features of a higher level causally interact with lower-level objects; it is a question of downward causation or mental causation.

By applying the same reasoning as used in analyses of the biological phenomenon of life treated as an emergent trait and its relationships with organisms' physical-chemical foundations, we can identify three types of downward causation in social phenomena: strong, moderate, and weak (Bremer, 2015). A strong type of downward causation requires a clear separation of the higher and the basic levels and their autonomy. This corresponds to classical biological vitalism, which assumes the existence of creative forces that defy description and mechanistic explanation. Examples of such theories can be found in dualistically oriented psychology and philosophy. The immaterial 'life force' can causally influence the lower level, i.e., the matter from which the body of an organism is built.

The main characteristic of moderate types of downward causation is that they exclude situations in which higher-level features exert a direct causal influence on objects and laws from the lower level. Causation takes place by limiting the conditions for development at the lower level. Since the higher level has already emerged (or is emerging), it constrains the lower level and affects its functioning in a certain area without exerting actual causal physical influence. Studies in rodents conducted by neuroscientist James Curley and his colleagues reveal that the quality of the social environment (i.e., the higher level) can have a profound effect on the development and activity of neural circuits (i.e., the lower level), which further influences numerous behavioural and physiological responses, including the expression of emotions in that environment (Curley et al., 2011). The authors of this study examined behaviours that illustrate social modulation during the development of neural circuits and focused on those circuits in which a long-term effect is observed. Similar results regarding the effects of social conditions on neural systems have been obtained in studies on primates, including humans. Adult males who have been maltreated in childhood and who have a low-activity form of the MAOA gene (Monoamine oxidase A) have been found to be significantly more prone to violent and disruptive behaviours than males who either have a high-activity form of this gene and have been maltreated, or have a low-activity form of this gene but have not been maltreated (Champagne & Curley, 2005). The results of these studies do not support a strong type of causation but confirm the moderate impact of social conditions on the development of the nervous and biological systems of a given individual.

A weak type of biological downward causation refers to the conceptual instrumentarium of the phase spaces for non-linear dynamic systems whose trajectories are attracted by relevant attractors. The theory of higher levels built on attractors does not explain these levels but merely allows them to be described.

In further analyses, we accept the thesis of a moderate impact of social relations on our nervous system. Returning to the cognitive social sciences, we can summarise that we rely on an emergent, i.e., a (non-reductionist and non-mechanistic version of) physicalist ontology, or, in other words, on an emergent materialist ontology (called weak materialism). The key assumption in this ontology is that social processes are dependent on lower-level processes but cannot be reduced to them and – from the causal perspective – moderately affect the processes in the brain from which they have emerged.

Social neuroscience

This section presents several examples of studies conducted within social neuroscience, which is a relatively new specialised part of the cognitive sciences that combines interactional symbolism with the brain sciences. John T. Cacioppo and his colleagues' research reveals that by definition, social individuals (e.g., humans and certain animal species) form structures that transcend the individual, ranging from pairs and families to larger groups and entire cultures. Social structures have evolved alongside neuronal, hormonal, cellular, and genetic mechanisms which support these mechanisms. The social behaviours that emerged thanks to these structures have helped organisms to survive, reproduce, and, in the case of some species, care for their offspring for as long as is needed.

Social neuroscience is an interdisciplinary field of science that studies neural, hormonal, cellular, and genetic mechanisms and investigates the connections and influences between communities of organisms and the biological levels of their organisation (Cacioppo, Berntson & Decety, 2010). As humans, we are a unique social species in this respect in that our social institutions, civilisations, and cultures are highly developed and our territorial reach knows no narrow boundaries. With reference

to emergence, we can say that our biology has helped to shape the society and environment we have created, and our social environment has helped to shape our bodies and brains. Social neuroscience should be viewed as a sub-discipline that aims to analyse human behaviours and human biology and to explore where we are as a species that belongs to a wider chain of biological organisms. As mentioned above, social psychology is also the study of social behaviours, but here the focus is on determining how the thoughts, feelings, and behaviours of individuals affect what is socially real, what is imagined, and the presumed presence of others. Social perception and social cognition cover the intrapersonal level (psychological processes) as well as interpersonal and group social interactions and social influences.

Social neuroscience makes use of integrated, multi-level conceptual analyses that range from the neural level through the level of the subject to the social level. It is based on the premise that new insights and new knowledge can be gained by incorporating findings and theories drawn from other levels of analysis and other studies. Following the development of social neuroscience, psychoneuroimmunology, and psychophysiology, the explanatory power of integrative analysis has become increasingly evident. This can be seen in attempts to model complex phenomena that take place on different levels of neural organisation. Efforts aimed at integrating information from different levels of analysis are particularly challenging given, on the one hand, the inherent complexity of neurobiological systems, and on the other, interactional symbolism. But it is only through such studies that a fuller cognitivist understanding of both mind-body relationships and symbolic interactionism will be possible.

Michael P. Kelly and his colleagues formulated a hypothesis regarding the impact of social interactionism on models of predictive processing in the brain (Kelly et al., 2019). In their opinion, predictive coding assumes that the brain actively anticipates incoming sensory stimuli rather than merely registering them passively. Predictive coding is efficient in the sense that the brain does not need to store multiple versions of the same information at different levels of the processing hierarchy. These models integrate conceptions of perception from neuroscience and sociology and highlight divergences in descriptions of how thinking subjects interact with the social world. These authors combine the concepts

of predictive processing models with Weberian ideal types and social practices, including the theses of symbolic interactionism. The result is a relatively unified hypothetical structure that integrates perceptions of the social world and hypothetical brain processes. This hypothesis combines aspects of neuroscience and psychology with social theory, with the aim of demonstrating how social behaviours can be cognitively 'mapped' onto brain processes. This provides a conceptual framework that links the two disciplines, opens up a creative dialogue between them, and outlines a perspective for future studies.

The advances that are taking place in social psychology and neuroscience as a result of neuroimaging technologies have made it possible to study diverse activation in different brain areas during normal and disturbed thinking, thus the complex structures of the human mind might be revealed. Furthermore, brain-computer interfaces have reached a level of sophistication at which surgically implanted intracerebral implants (also called neural implants or neuroprostheses) are capable of 'translating' thought-induced brain activation into movement in external devices. Electrodes placed in the brain to record or stimulate neural activity are opening up an exciting new area of research in neuroscience, biomedical engineering, and the social sciences. Such innovative techniques are revolutionising our understanding of the brain and pointing to new possibilities for improving the functioning of neurons. Thanks to being directly connected to the neural circuits of the brain, intracerebral implants offer a variety of applications, including restoring a person's sensory functions or enhancing cognitive abilities. Anubhav Singh discusses the advances in the field of brain implants and their potential impact on human behaviours, well-being, and thus also social interactions. These innovations hold great promise for people with visual and speech impairments or paralysis as they offer them the opportunity for effective and independent social communication (Singh, 2023).

These and other advances in cognitive sciences research represent a new era in understanding the biological and psychological basis of our human existence. Using an analogy from physics, we can say that observations of wave-particle duality as a feature of quantum objects (e.g., photons or electrons) – which, depending on the situation, involve the manifestation of either wave characteristics (diffraction, interference) or particle characteristics (specific location, momentum) – have

revealed the inadequacy of classical physics and initiated the emergence of quantum physics. Similarly, it is becoming increasingly clear that the conventional mono-disciplinary models in both biology and social psychology are proving inadequate or at least incomplete when attempting to explain the complexity of individual and social behaviours.

For several years, the Polish Academy of Sciences, in cooperation with other entities, has been running the Social Neuroscience Laboratory in Warsaw, whose main aim is to study the neural and behavioural processes that shape the ways in which we process social signals (https://psych.pan.pl/struktura/pracownie-naukowe-ip-pan/pracownia-neuronauki-spolecznej/). Ongoing empirical research addresses the question of what factors influence our social cognitive processes and the social brain activity that underlies these processes. Social cognitive abilities and the activity of the 'social brain' are being studied both in individuals from the general population and in patients with neuropsychiatric disorders in order to identify the factors that influence a person's social functioning. In particular, these studies focus on the impact of the feeling of loneliness (or, to use a technical term, 'perceived social isolation') on behavioural and physiological processes related to social cognition.

On the one hand, the results of studies conducted within social neuroscience, including brain implants, offer numerous opportunities to advance medicine and improve the quality of people's lives, thereby returning healthy – or at least healthier – citizens to society. On the other hand – and unfortunately this is rarely addressed in scientific discussions or studies – their use raises significant ethical concerns. These include, in particular, issues related to the privacy and safety of the participants of such studies. In the case of brain-computer interfaces, there is the risk of unauthorised access to patients' neural data occurring without their knowledge. This raises questions about the protection of data acquired in this way and consent for its use. It becomes crucial to strike a balance between the benefits of sharing information to further scientific progress and the protection of privacy. Furthermore, enhancing human abilities and cognitive capacities with brain implants leads to questions about social justice and equality. If neuroprostheses become readily available, there is the risk of a widening social gap between those who can afford or have access to such enhancements and those who cannot or do not.

Ensuring socially equitable distribution and the financial accessibility of such technologies will be essential to prevent the widening of the social gaps that already exist.

Systematic reflection with conclusions and recommendations

On the one hand, there has been a social turn in the cognitive sciences since the 1970s, as evidenced by the emergence of social neuroscience and other experimental research programmes; on the other hand, many theoretical and practical social scientists have become more interested in the cognitive sciences and have drawn information from them, as evidenced by the emergence of cognitive sociology and social psychology. This article discusses the cognitivist approach to symbolic interactionism treated as a representative of the social sciences. The basic concepts and theoretical insights regarding interactionism are indeed relevant for broadly defined sociology and social psychology. The traditional disciplines of the sciences of living organisms deal with complex phenomena occurring at their given level of description and explanation, including their molecular, anatomical, psychological, and social levels. The results obtained from research in these areas provide key insights into biological, psychological, and social processes. The cognitivist interdisciplinary approach is characterised by multi-level conceptual analyses combined with empirical research and carries additional explanatory value.

Social behaviours can be understood based on philosophical emergence theory, which assumes a moderate version of downward causation. Just as this theory prevents the reduction of mental states to physical (mechanical) states, its social version prevents the reduction of social behaviours to physical states. We can treat theories of emergence as surrogates for neuroscientific theories. The emergentist approach to social states corresponds to theses formulated within social neuroscience, the interpretation of which also does not have to lead to the reduction of social states to neural states. Social neuroscience indicates the methodological necessity of acknowledging the relevance of the relationship between events occurring at the neurobiological level and social behaviours (described in the language of the social sciences and

the humanities, which uses third- and first-person linguistic categories). The latter can be based on events described in the language of interactional symbolism. These descriptions have prompted researchers in biology, the cognitive sciences, and the social sciences to work together in a more systematic way. This has been accompanied by a widespread belief that understanding the workings of the individual mind and human behaviour could be broadened by means of integrative analysis that covers various levels of organisation, ranging from genes to social relations and cultural phenomena. Recent years have brought an expansive development of neurosciences, which include social cognitive neuroscience, social affective neuroscience, social cultural neuroscience, social computational neuroscience, social developmental neuroscience, and social comparative neuroscience.

Social neuroscience uses data from brain studies in a methodological and theoretical way to describe and explain social processes. The general approach of social neuroscience, which is part of the cognitive sciences, aims to understand the psychological processes that underlie human social behaviours. Since the psychological processes that occur in a person are phenomena that cannot be directly observed, social neuroscientists simultaneously rely both on a combination of measurable or observable neuronal and physiological responses, and on the observation of our everyday behaviours, drawing conclusions about our mental states from both these sources. In the future, by studying neuronal reactions, social neuroscientists may be able to tell us more not only about social behaviours or the brain areas correlated with these behaviours, but also about how social processes affect the functioning of neural systems.

The aforementioned studies on brain implants offer a better understanding of the workings of the brain and thus have revolutionised treatment methods that are intended to restore people's physical and mental abilities and enable their social integration. From restoring lost functionality to enhancing cognitive abilities, neural interfaces offer new possibilities for human enhancement and interaction. However, several pertinent questions arise: What are, or will be, the social and cultural implications of using technologies that makes it possible to manipulate our brains and thus, indirectly, our minds and social behaviours? What impact will social neuroscience have on our self-understanding and on

the general social notion of man? Following the implementation of new technologies, it is vital to consider the ethical issues related to respect for privacy, equality, and the accessibility of such tools. Progress in this area should be accompanied by increasing interdisciplinary collaboration between scientists, ethicists, policy makers, and representatives of community groups to ensure that brain implants are developed and implemented in a way that is responsible, ethical, and beneficial to society as a whole. By striking the right balance, we can unlock the extraordinary potential of brain implants, while protecting the core values and well-being of both individuals and society as a whole.

References

Blumer, H. (1986). *Symbolic Interactionism: Perspective and Method*. Berkeley and Los Angeles, California: University of California Press.

Bremer, J. (2015). Przyczynowość skierowana ku dołowi i jej rozumienie w biologii. *Poznańskie Studia z Filozofii Nauki*, *24*(1), 93–115.

Bremer, J., & Flasiński, M. (2022). The Turing Test, or a Misuse of Language when Ascribing Mental Qualities to Machines. *Forum Philosophicum*, *27*(1), 6–25. DOI: 10.35765/forphil.2022.2701.01.

Cacioppo, J.T., Berntson, G.G., & Decety, J. (2010). Social Neuroscience and Its Relationship to Social Psychology. *Social Cognition*, *28*(6), 675–685. DOI: 10.1521/soco.2010.28.6.675.

Carter, M.J., & Fuller, C. (2016). Symbols, Meaning, and Action: The Past, Present, and Future of Symbolic Interactionism. *Current Sociology*, *64*(6), 931–961. DOI: 10.1177/0011392116638396.

Champagne, F.A., & Curley, J.P. (2005). How Social Experiences Influence the Brain. *Current Opinion in Neurobiology*, *15*(6), 704–709. DOI: 10.1016/j.conb.2005.10.001.

Curley, J.P., Jensen, C.L., Mashoodh, R., & Champagne, F.A. (2011). Social Influences on Neurobiology and Behavior: Epigenetic Effects during Development. *Psychoneuroendocrinology*, *36*(3), 352–371. DOI: 10.1016/j.psyneuen.2010.06.005.

Ellwood, Ch.A. (1909). The Science of Sociology: A Reply. *American Journal of Sociology*, *15*(1), 105–110. https://www.jstor.org/stable/2762623. Accessed 06.03.2024.

Hałas, E. (2012). *Interakcjonizm symboliczny. Społeczny kontekst znaczeń*. Warszawa: Wydawnictwo Naukowe PWN.

Kaidesoja, T., Sarkia, M., & Hyyryläinen, M. (2019). Arguments for the Cognitive Social Sciences. *Journal for the Theory of Social Behaviour*, *49*(4), 480–498. DOI: 10.1111/jtsb.12226.

Kelly, M.P., Kriznik, N.M., Kinmonth, A.L., & Fletcher, P.C. (2019). The Brain, Self and Society: A Social-Neuroscience Model of Predictive Processing. *Social Neuroscience*, *14*(3), 266–276. DOI: 10.1080/17470919.2018.1471003.

Konecki, K.T., & Kacperczyk, A. (2020). Symbolic Interactionism in Poland: Inspirations and Development. *Qualitative Sociology Review*, *16*(4), 8–34. DOI: 10.18778/1733-8077.16.4.02.

Patel, H. (2021). Fundamental Principles of Artificial Intelligence. *International Research Journal of Modernization in Engineering Technology and Science*, *3*(1), 318–323. https://www.irjmets.com/uploadedfiles/paper/volume3/issue_1_january_2021/5605/1628083227.pdf. Accessed 06.03.2024.

Sarkia, M., Kaidesoja, T., & Hyyryläinen, M. (2020). Mechanistic explanations in the cognitive social sciences: Lessons from three case studies. *Social Science Information*, *59*(4), 580–603. DOI: 10.1177/0539018420968742.

Singh, A. (2023). Exploring the Potential of Brain Implants: Unlocking the Secrets of Neural Enhancement. *Journal of Neuroscience and Neuropharmacology*, *9*(1), 001–002. DOI: 10.4172/24699780.2023.9.1.189.

Thomas, W.I., & Znaniecki, F. (1995). *The Polish Peasant in Europe and America.* Champaign, Illinois: University of Illinois Press.

Uldall, B.R. (2013). Social Psychology. In: A. Runehov, & L. Oviedo (ed.), *Encyclopedia of Sciences and Religions* (pp. 2171–2179). Dordrecht: Springer. DOI: 10.1007/978-1-4020-8265-8_1047.

Ewa Jaglarz
Ignatianum University in Cracow
https://orcid.org/0000-0001-8754-8770
https://doi.org/10.35765/slowniki.455en

Speech and language disorders

Summary

DEFINITION OF THE TERM: The article explores various approaches to understanding speech and language disorders, including medical, behavioural, symptomatic, and cognitive perspectives. This section clarifies the distinction between language, speech, and speaking. In psycholinguistic and cognitive contexts, language is viewed as a sign system – a mental entity used for communication and social interaction – while speech is seen as the multifaceted individual act of employing this system.

HISTORICAL ANALYSIS OF THE TERM: Early interest in these disorders can be traced to the medical practices of phoniatrists. Over time, thanks to the development of scientific centres, the study of disorders evolved into a distinct scientific discipline. Today, the theory of speech and language disorders also plays a significant role in cognitive science, with several research centres in Poland contributing to this field.

DISCUSSION OF THE TERM: The complexity of language and speech, the heterogeneity of the convergence and divergence of symptoms and causes, and the interdisciplinary nature of the field make it challenging to precisely categorise different types of speech and language disorders. This section offers a critical review of theoretical models of disorders, revealing their diverse nature, particularly from a cognitive perspective.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: The analysis presented in the article underscores the importance of accurate diagnosis and understanding of the varied manifestations and forms of these disorders. Language, with its multiple

functions, is a vital tool for individuals' participation in social life. The cognitive approach offers practical insights into applying knowledge of speech and language disorders.

Keywords: language, speech, speech and language disorders, psycholinguistics, cognitive science

Definition of the term

Language is a complex communication system that defines human uniqueness and facilitates social interaction. It serves as a tool for expressing thoughts, emotions, beliefs, and intentions through well-defined syntactic, semantic, and phonological structures. Language can be examined from various perspectives – linguistic, psychological, anthropological, neurobiological, and cognitive. In cognitive science, language is primarily viewed as a tool of the mind, enabling information processing, thinking, and communicating. Cognitivists study how language is represented and processed in the brain, as well as how it is used in thinking. Some emphasise the role of language in shaping thought processes; for example, the Sapir-Whorf hypothesis suggests that the language we use influences how we perceive and interpret the world. Psycholinguistics investigates the mental processes involved in language production, comprehension, and the generation of speech and writing. The usage-based approach posits that language acquisition and language structures are shaped by language experience, making language adaptable and dynamically evolving in response to communicative contexts. Neurolinguistics explores the biological basis of language processing, focusing on how various brain regions contribute to language comprehension and production. As an interdisciplinary field, cognitive science offers a comprehensive understanding of speech and language disorders.

Language, as both a product and tool of the human mind, is a multifaceted phenomenon that integrates symbolic, cultural, cognitive, and biological dimensions. Research into language not only deepens our understanding of language itself but also sheds light on the workings of the human mind, society, and culture, making it inherently multidisciplinary and interdisciplinary.

Language disorders, broadly understood as disorders of linguistic communication, manifest as communicative dysfunctions that impair an individual's ability to understand, process, and produce spoken and written language across one or more of its core domains: phonology, morphology, syntax, semantics, and pragmatics. These disorders can stem from a variety of causes and manifest differently depending on the person's age, language experience, and cultural background.

Examples include acquired disorders like aphasia, often resulting from brain damage, and developmental disorders such as specific language impairment (SLI). The effects of these disorders can range widely, leading to difficulties in education, social interactions, and professional settings.

Although speech and language disorders are related, some perspectives distinguish them based on the specific aspects of communication they affect. The distinction between 'language' and 'speech' is often credited to Ferdinand de Saussure, the Swiss linguist regarded as the father of modern structural linguistics. In his seminal work, *Course in General Linguistics*, de Saussure defined 'language' (*langue*) as a shared system of signs and rules used by members of a linguistic community, whereas 'speech' (*parole*) is defined as the individual use of language by humans, i.e., how people and groups employ this system for communication. De Saussure focused on the structural and semiotic aspects of language, introducing key concepts such as the 'linguistic sign', which he broke down into two components: the 'signifier' (*signifiant*), the form of a word or expression, and the 'signified' (*signifié*), the meaning it conveys. He also noted that meaning in language arises from the differences between its elements: words acquire meaning through their relationships to one another rather than in isolation.

Roman Jakobson expanded on de Saussure's structural approach by introducing functional aspects of language, particularly through his holistic model of communication, which incorporated elements like phonology and the various functions of language in different contexts. Jakobson did not challenge the distinction between linguistic norms (*langue*) and individual speech acts (*parole*). He viewed linguistic norms as social, fixed, unified, and homogeneous, but speech as individual, variable, heterogeneous, and physiological. For Jakobson, linguistics is divided into two areas: one focusing on the social norms of language, a static system that functions as an interconnected whole, and the other examining individual speech as random and personal (Jakobson, 1983). Unlike de Saussure, who analysed language as it exists at a specific moment in time (synchronic analysis), Jakobson developed a method that combined synchronic and diachronic analysis. This allowed him to study language both in its immediate use and as it evolves over time. His exploration of the 'virtuality' of language norms alongside the actuality of

individual speech led to deeper insights into speech disorders, a topic Jakobson himself studied, particularly aphasia.

In this context, Noam Chomsky's contributions from the 1960s are also noteworthy. This American linguist is best known for his theory of universal grammar, which posits that all humans are born with an innate capacity to acquire language. According to this theory, all languages share a fundamental structure, with the differences between them being superficial. Chomsky introduced the concept of "linguistic competence", an unconscious inherent ability that allows children to rapidly and effectively learn their native language. He also developed transformational-generative grammar, which explains how grammatical rules can generate a virtually limitless number of sentences from a finite set of principles. Chomsky distinguished between competence, i.e., a language user's internal knowledge of language, and performance, i.e., the actual use of language in real-life situations. Performance encompasses how utterances are constructed and interpreted in everyday communication, often subject to errors and limitations imposed by psychological, social, and environmental factors (Greene, 1972). Whereas de Saussure viewed language as an objective entity, Chomsky's approach sees language as processed and used, appearing in the mind thanks to abstract innate mental structures and mechanisms – which are mental entities.

Polish psycholinguist Ida Kurcz defines language as a system of signs and rules of their use, and speech as the actual acts of use of this system (Kurcz, 2000, p. 14). Kurcz highlights the distinction between language, speech, and speaking. In her view, speech refers to

> the phenomenon of human communication, primarily through sound (but also through gestures, writing, etc.), whereas speaking consists in specific utterances and concrete acts of language use by concrete people, aligning with de Saussure's notion of 'parole'. Language is the system of signs and rules that govern their use – it is something that is applied and used in speech and realised in concrete speech acts (Kurcz, 1987, p. 11).

A similar understanding can be found in Stanisław Grabias's work, although his definition places greater emphasis on the functional role of language. He defines speech as

> a system of activities through which, with the help of language, people learn about the world and communicate knowledge of both themselves and the world to others. This definition reflects the belief that language leads directly to the human mind – his cognitive abilities, emotions, and desires – and that linguistic behaviour offers insights into human activity by revealing how individuals share knowledge and function within social groups (Grabias, 2020, p. 534).

Within this metatheoretical framework, speech and language emerge as multifaceted phenomena, approached by different researchers with varying emphases. These perspectives explore the structural and functional dimensions of language, the distinct elements and roles of this system, its static and dynamic characteristics, and its determinants (social, biological, and cognitive), etc.

This variety of fundamental approaches to speech and speaking leads to a corresponding diversity in how their disorders are understood and addressed. On the one hand, speech disorders typically refer to difficulties in the physical act of speaking, i.e., the physical and biologically determined process of language production. These disorders may involve problems with articulation (i.e., the accurate production of sounds), speech fluency (such as stuttering), and voice modulation (including tone, volume, and quality). While the mechanical aspects of speech production are central to these disorders, their causes are varied and can include neurological damage, developmental defects, trauma, illness, or genetic factors. Diagnosing and treating speech disorders often requires a multidisciplinary approach involving speech therapists, neurologists, psychologists, and psycholinguists. Speech disorders encompass a wide range of diverse pathological conditions that affect an individual's ability to express himself through verbal language.

On the other hand, language disorders refer to difficulties with understanding or using words in context, impacting both spoken and written language. These disorders can affect sentence formation, vocabulary usage, comprehension of instructions, or the ability to build complex language structures. Language disorders can manifest as issues with expression (the ability to convey thoughts and ideas) or comprehension (the ability to understand speech and text).

Stanisław Grabias offers a synthesis of the two notions of language and speech. Through deductive indirect inference, he has concluded that a speech disorder would manifest as difficulty in language acquisition:

> The concept of speech disorders should be derived from an analysis of situations where an individual is unable to acquire language at all or acquires it only partially, or where, despite knowing the language, he is unable to construct linguistic utterances or is forced to produce statements that disrupt social interaction (Grabias, 2012, p. 24).

This perspective highlights the interconnectedness of the two categories. Grabias even equates speech with linguistic communication, suggesting that speech disorders should be understood within the broader context of communication, as "this will allow us to notice not only what exists related to a given dysfunction but also what is missing but should exist for linguistic behaviour to be effective" (Grabias, 2012, p. 49).

Thus, both language and speech are crucial for communication and often overlap. For example, children with speech disorders may struggle to articulate words clearly, which can impede their ability to communicate and learn language. Likewise, difficulties with language comprehension can hinder the acquisition of proper articulation and other speech skills which are crucial when communication is approached also from the perspective of cognitive processes.

Historical analysis of the term

In the history of research on speech disorders, the connection between the brain and speech functions has been recognised since antiquity, with early insights from Hippocrates and Aristotle. However, the foundation of modern research was laid by French neurologist Paul Broca, who identified the brain's functional anatomy as the neural basis of speech. Broca's work demonstrated that damage to specific areas of the brain could lead to speech disorders. Similarly, German neurologist Carl Wernicke is credited with discovering the cerebral localisation of speech comprehension, while Adolf Kussmaul introduced the term 'aphasia' and classified its various types. The study of brain-speech relationships advanced significantly in the 20th century through the work of Wilder Penfield, who made contributions to brain mapping. This period also witnessed rapid developments in psycholinguistics and neuropsychology. Noam Chomsky introduced the concept of generative grammar, positing that language acquisition is an innate ability. Ferdinand de Saussure's

analytical framework for understanding language as a system greatly influenced the development of psycholinguistics. Additionally, Roman Jakobson's research on language and phonology played a key role in advancing the study of speech disorders within linguistics. Technological innovations also transformed the field during the 20th century, with the emergence of tools such as electroencephalography (EEG), computed tomography (CT), and magnetic resonance imaging (MRI).

In Poland, Władysław Ołtuszewski was instrumental in establishing the first institute of phoniatry (*Instytut Odchyleń Mowy*), a medical unit in Warsaw dedicated to speech disorders. He also worked to popularise knowledge in the field, outlining his findings in his 1906 publication *O mowie i jej zaburzeniach* [*On Speech and Its Disorders*]. Ołtuszewski classified speech disorders into two main categories: central disorders such as aphasias (perceptual aphasia, speech impairments in deafness and in intellectual disabilities) and dysarthrias and peripheral disorders (only dysarthrias).

A pioneering figure in the study of speech and language disorders in Poland was psychologist and physician Mieczysław Choynowski. His contributions to speech therapy, particularly in the classification of speech disorders, had a profound impact on the development of the field in Poland. In 1913, he published *Zarys logopedii czyli nauki o zaburzeniach mowy u dzieci* [*An Outline of Speech Therapy, or the Study of Speech Disorders in Children*], a work that is widely considered a milestone in the development of Polish speech therapy. Choynowski identified several key types of speech disorders, including alalia – a complete absence of speech in children; dyslalia – difficulties in articulating speech sounds correctly; stammering – disruptions in the fluency of speech; speech blocks; aphasia – the loss of speech ability due to brain damage; bradylalia – an abnormally slow speed of speech production; tachylalia – an excessively fast speed of speech production; rhinolalia – nasal speech caused by improper airflow through the nasal and oral cavities. Choynowski emphasised the wide variety of speech disorders, relying primarily on symptomatic criteria for classification. His work remains foundational in the Polish study and treatment of speech and language disorders.

In post-war Poland, research and treatment of speech disorders developed across several key centres throughout the country, including

Lublin, Kraków, Warsaw, and Poznań. Among these, Lublin is particularly notable as the birthplace of Polish speech disorders science. The speech therapy unit at the John Paul II Catholic University of Lublin (KUL) was established in 1962, becoming one of Poland's first academic units dedicated to this discipline. Halina Jastrzębowska, the pioneering force behind speech therapy at KUL, played a crucial role in the development of the field. In 1969, Leon Kaczmarek, often referred to as the 'father of Polish speech therapy', spearheaded the establishment of the Chair of Speech Therapy and Applied Linguistics at Maria Curie-Skłodowska University in Lublin. Moreover, the Polish Speech Therapy Association (*Polskie Towarzystwo Logopedyczne*) was founded in Lublin in 1963, thanks to Kaczmarek's initiative. By the 1980s, Poland had four primary centres training speech therapists. Besides the aforementioned KUL, the University of Warsaw was notable, initially led by Halina Mierzejewska. In 1974, the Postgraduate Speech Therapy Studies (*Podyplomowe Studium Logopedyczne*) was established under the direction of Irena Styczek and later Maria Przybysz-Piwkowa. Gdańsk also saw the development of a centre founded by Bronisław Rocławski at the University of Gdańsk. Another important centre was the Studium in Opole, founded by Tadeusz Gałkowski and Grażyna Jastrzębowska. Today, several research centres continue to advance the study of language, particularly within the cognitive science framework. The Institute of Linguistics at Adam Mickiewicz University in Poznań is renowned for its strong research tradition in cognitive linguistics, offering diverse research programs and initiatives focused on language and cognition. The Centre for Research on Culture, Language, and Mind (Q Centre) (*Centrum Badań nad Kulturą, Językiem i Umysłem (Centrum Q)*) at the University of Warsaw is another key player, conducting interdisciplinary research that integrates psychology, computer science, philosophy, and linguistics.

Discussion of the term

The study of speech and language disorders spans multiple scientific disciplines. Initially, these issues were primarily examined within the medical sciences. The International Statistical Classification of Diseases

and Related Health Problems (ICD-10) categorises speech disorders under the Disorders of psychological development section, specifically in code F80, which covers 'Specific developmental disorders of speech and language'.

> Disorders in which normal patterns of language acquisition are disturbed from the early stages of development. The conditions are not directly attributable to neurological or speech mechanism abnormalities, sensory impairments, mental retardation, or environmental factors. Specific developmental disorders of speech and language are often followed by associated problems, such as difficulties in reading and spelling, abnormalities in interpersonal relationships, and emotional and behavioural disorders (ICD-10, https://icd.who.int/browse10/2016/en#/F80-F89).

As of January 1, 2022, the ICD-11 classification has been introduced, although it is not yet implemented in Poland, where there is a five-year transition period before the new classification system takes effect.

In both linguistic and cognitive streams, speech disorders are recognised as complex communication issues that can impact an individual's ability to express thoughts and communicate effectively. Language serves dual intertwined functions – representational and communicative – making it challenging to separate their distinct meanings and influences. As Roman Jakobson observed, "Thanks to psychological, neurological, and, above all, linguistic research, it is becoming increasingly clear that language is a vehicle not only for interpersonal but also for intrapersonal communication" (Jakobson, 1989, p. 60). Notably, the cognitive approach enables the synthesis of these distinct functions. As language users, we simultaneously and concurrently employ both functions – as tools for communication with others and as tools for thinking and self-reflective communication.

In considering the communicative function from a cognitive science perspective, Jakobson's classification of language functions is particularly relevant. He identified six basic functions of language, each of which can be associated with specific speech disorders:

1. Referential function: Linked to the context of the message. Disorders of this function manifest as difficulties in accurately describing reality, as seen in conditions like autism or sensory aphasia.
2. Poetic function: Focuses on the form of the message and its aesthetic aspects. Disorders in this function reveal as problems

with using language in a creative or artistic way, typical in individuals with autism spectrum disorders.

3. Emotive (also called expressive) function: Reflects the speaker's attitude toward the content of the message. Disorders of this function involve difficulties in expressing emotions and attitudes towards the topics being discussed, as observed in conditions such as motor aphasia.
4. Phatic function: Related to maintaining communication between the speaker and the listener. Disorders of this function cover difficulties in initiating or maintaining conversations, which are characteristic of individuals with apathy or severe psychological disorders.
5. Metalinguistic function: Enables communication about language itself. Disorders in this function are seen as difficulties in analysing and discussing language, as in dyslexia or certain types of aphasia.
6. Conative function: Related to the influence of the message on the listener, aimed at prompting a specific action or response. Disorders of this function may involve difficulties in using language to persuade or give commands, which can be observed in various types of aphasia (Jakobson, 1989).

Roman Jakobson separately explored aphasia as a quintessential language disorder. As a linguist, he had the expertise to examine language in all its states: in action, in motion, in emergence, and in decay (Jakobson, 1989, p. 150). He recognised the dual nature of language, where systems of combination and selection are continuously at work, and he analysed the effects of aphasic conditions along two axes: disruption in the domain of similarity and disruption in the domain of contiguity. In the domain of similarity, one of the two main types of linguistic operations according to Jakobson (the other being metaphor and metonymy), disruption manifests through the use of unexpected comparisons or metaphors, and difficulties in transitioning from words to their synonyms, descriptive expressions, and heteronyms (which is linked to a loss of polyglotic capacities). This results in the patient's internal 'idiolect' becoming isolated, with a diminished ability to switch between different meaning codes. Consequently, the individual may struggle to perceive relationships between objects or concepts that are typically

seen as similar. Disruption in the domain of contiguity affects the ability to produce well-formed longer utterances. According to Jakobson,

> This type of aphasia, characterised by an impaired ability to establish context and often described as a disruption in contiguity, limits both the range and variety of sentences. The syntactic rules that structure words into coherent units are lost, a condition known as agrammatism, reducing sentences to a mere 'pile of words' (Jakobson, 1989, p. 164).

Jakobson's identification of two specific types of aphasic disorders, as well as others related to language function, underscores his view of language:

> Language is a system of systems, a holistic code containing various individual subcodes. Different speech styles form a correct hierarchy of codes rather than a random, mechanical collection (Jakobson, 1989, p. 145).

Today, Jakobson's conception can be used in the field of natural language processing (NLP), a branch of artificial intelligence that focuses on the interaction between computer programs and human language. Applying Jakobson's ideas could enhance algorithms' ability to understand, interpret, and generate human language in a way that is useful, accurate, and natural.

Cognitive science offers an interdisciplinary, comprehensive approach to understanding and treating speech disorders, integrating both their neurobiological and cognitive aspects. This perspective emphasises the importance of understanding how various cognitive processes and brain structures collaborate to enable language production and comprehension. Notable cognitive models of speech disorders include the Wernicke-Lichtheim-Geschwind model, the interactionist model, and the neural network model.

Moreover, the behaviourist perspective, which was introduced in the 1960s by Polish researcher Leon Kaczmarek, provides valuable insights into speech disorders. Kaczmarek approached speech from a structuralist angle as a collection of distinct components that function somewhat autonomously. According to this model, each of the three components he identified based on the symptoms of disorders they lead to – content, linguistic form, and phonemic substance – could generate specific difficulties. For content disorders, Kaczmarek listed difficulties in

generalisation and abstraction, illogical text construction, and abnormalities in thinking orientation. Linguistic form disorders included conditions like muteness, aphasia, alalia, agrammatism, and slips of the tongue. He divided disorders of substance into suprasegmental issues, such as stammering, rhinolalia, laryngeal pronunciation, aphonia, and bradylalia, and segmental issues like dyslalia. Disorders at both levels – suprasegmental and segmental – include palatolalia and muteness (Kaczmarek, 1975).

Stanisław Grabias (2020) critically analysed Kaczmarek's theory, challenging the notion that the components of language (i.e., speech acts) operate independently and sequentially: first we think, then then we clothe the thought in a linguistic form, and finally we give it substance. Grabias argued that the brain functions simultaneously across these processes. He also questioned some of Kaczmarek's detailed claims, particularly those related to psychological disorders and stuttering. Despite these criticisms, Grabias acknowledged Kaczmarek's contribution to terminological clarity and his effort to move beyond the biologism prevalent in symptomatic classifications.

Following the recommendations of the International Association of Logopedics and Phoniatrists and the European Union of Phoniatrists, Tomasz Zalewski distinguished between two main types of speech disorders in the context of speech development: specific speech development delays, which do not indicate pathology, and speech development disorders, which are deviations from the norm caused by factors such as hearing impairments, central speech centre dysfunctions, intellectual disabilities, cerebral palsy, or emotional issues (Zalewski, 1993, p. 48). It can be inferred that Zalewski acknowledged the cognitive nature of speech disorders. First, he posited that the gateway to the mind leads through auditory processes, including neural-level auditory processing. Second, he claimed that speech has a neural foundation, noting deviations from the norms that accompany it. Third, he referred to mental aspects of language, such as behavioural disorders, emotional issues, and intellectual disabilities.

Irena Styczek, the author of the first monograph on speech therapy in Poland, emphasised etiology as a primary criterion for classifying speech and language difficulties. She categorised disorders into two main groups: extrinsic (exogenous, environmental factors that she considered

disorders rather than impediments) and intrinsic (endogenous, inherent speech impediments). Within intrinsic disorders, Styczek identified the following types based on their causes: dysglossia (distortion of speech sounds or inability to produce them due to abnormal structure of the speech organs or reduced hearing), dysarthria (also called anartria, distortion or inability to produce speech sounds due to damage to the neural pathways innervating the speech organs), dyslalia (also called alalia, delayed language acquisition due to the delayed development of specific brain structures), aphasia (partial or total loss of language abilities resulting from damage to certain brain areas), stuttering (a fluency disorder characterised by irregularities in rhythm and tempo, with an unclear cause), speech neuroses (also called logoneuroses, disorders such as mutism, aphonia, stammering, and irregular speech rate or modulation associated with neuroses), oligophasia (underdevelopment of speech skills due to mental retardation), and schizophasia (disordered speech seen in individuals with psychological disorders and thought disorders) (Styczek, 1983, pp. 250–251). While Styczek's classification laid a crucial foundation, it is now considered somewhat outdated. The basic categories she proposed, such as dyslalia and aphasia, are still in use but have been expanded and refined with advances in diagnostic and therapeutic technologies, such as neuroimaging and artificial intelligence. Styczek's work remains valuable for its pioneering role in shaping modern therapeutic approaches.

Stanisław Kania offered a linguistic perspective on speech disorders, focusing primarily on their formal aspects. He proposed that the essence of a speech disorder lies in the form and structure of utterances (linguistic signs). If this is the case, the individual elements of utterances – ranging from isolated sounds to syllables, words, and entire utterances – would represent the segmental aspect of the disorder. Disorders of utterance segments could manifest both at the level of the sound inventory (paradigmatic) and within the phonetic structure of the sounds themselves (syntagmatic). Kania argued that syntagmatic and paradigmatic disorders are interconnected and can manifest in both primary and secondary forms. He wrote that "the same patient can suffer from 1) primary paradigmatic disorders, 2) primary syntagmatic disorders, 3) secondary paradigmatic disorders, and 4) secondary syntagmatic disorders" (Kania, 1982, p. 10). He believed that secondary

disorders are influenced by the structure of a specific language or individual pronunciation patterns, which he considered internal causes. For primary paradigmatic disorders, Kania adapted a traditional phoniatric classification (mogilalia, paralalia, and deformation) and proposed his own categories: elision (mogilalia), substitution (paralalia), and deformation. Kania's structural approach is further reflected in his classification of syntagmatic disorders, which he divided into quantitative, qualitative, sequential, combined, and non-systemic. For each type of disorder, he specified subcategories.

Kania's classification narrows the exploration of disorders to a purely formal aspect. However, his models of natural language analyses can be utilised to develop cognitive models that more accurately reflect the mechanisms of language processing. With advances in technology, particularly AI tools, these models could automate diagnostic processes.

A comprehensive approach to understanding the diverse forms of speech pathology has been undertaken by Halina Mierzejewska and Danuta Emiluta-Rozya, who created a graphical representation of causes and symptoms. Their classification, which has evolved since its initial proposal in 1997, highlights the complexity and variety of speech disorders. While the authors did not explicitly address the heterogeneity of convergence and divergence in speech disorders, these concepts are central to their framework. Heterogeneity of divergence refers to the phenomenon where various elements (causes and symptoms of disorders) diverge in different directions or exhibit diverse trajectories of change. Conversely, heterogeneity of convergence is a situation or process in which diverse elements (causes or symptoms of disorders) merge despite their initial differences. As these authors observe,

> analysis of the graphical form of our overview reveals (…) that the same range of symptoms can be caused by different endogenous, exogenous, or endo-exogenous causes (Emiluta-Rozya, 2012, p. 76).

While Mierzejewska and Emiluta-Rozya present the broadest possible range of symptoms and causes, they also acknowledge that

> we have not found a way to visually differentiate between the impairments of verbal comprehension and speech production in the different forms of speech pathology (Emiluta-Rozya, 2012, p. 75).

Thus, for functional reasons, they note that the typology they propose is particularly useful for practitioners – speech therapists. Their detailed illustrations, akin to a logarithmic slider, enables therapists to navigate through various causes, forms, and symptoms of speech disorders and can also serve as a foundation for developing algorithms to create an information 'sieve' for analysing individual utterances.

Stanisław Grabias (2020) offers a comprehensive classification of speech disorders by integrating cognitive and neurobiological perspectives. His framework is grounded in several key premises: first, language is uniquely human due to the complex structure of the human brain; second, children acquire language spontaneously and internalise it through neural pathways until around the age of six, after which language acquisition shifts to a more programmed process; third, language acquisition relies on hearing, which is essential for properly channelling language into the mind. Based on these premises, the author identifies three key research assumptions that underpin his detailed classification: 1) the meaning and structure of utterances reflect the systemic, social, situational, and pragmatic interactional competences; 2) there exist distinct linguistic and communicative competencies (i.e., spontaneous entities in the mind that organize sentences and dialogic and narrative utterances; 3) formative skills manifested in biological processes that determine language encompass aspects such as physical hearing, auditory processing at the neuronal level with phonemic awareness, brain plasticity, and the skeletal and muscular systems. Thus, speech disorders are viewed as a conceptual amalgamation of the causes and effects of dysfunction, with their diversity influenced by the anticipatory nature of corrective actions. In response to this understanding of language and its impact on speech disorders, Stanisław Grabias proposed the following classification:

1. Speech disorders arising from underdeveloped biological functions necessary for language: deafness and hearing impairment, alalia and dyslalia (where competencies are not developed due to central auditory processing disorders), oligophasia, autism, childhood epilepsy, and foetal alcohol syndrome.
2. Speech disorders resulting from inadequate development of executional abilities: dysglossia, cluttering, stuttering, and dysarthria (which are associated with damage to the central or peripheral nervous system).

3. Speech disorders due to the breakdown of linguistic and communicative abilities: aphasia (characterised by the loss of the ability to understand and produce speech caused by brain damage, with difficulties linked to impaired speech programming mechanisms), pragnosia (resulting from damage to the 'non-linguistic' hemisphere of the brain, leading to difficulties in understanding metaphorical expressions, longer utterances, and recognising the speaker's emotional states), schizophasia (a result of schizophrenia), and dementia (caused by the degeneration of neurons and synaptic systems, as seen in conditions like Alzheimer's disease).

Grabias's approach situates speech disorders within the interdisciplinary research realm of cognitive science. He posited that these disorders often stem not just from anatomical or executive dysfunctions but from disruptions in cognitive processes such as memory, attention, perception, and thinking. In his work, Grabias emphasised the crucial role of cognitive processes in both linguistic communication and speech disorders. He explored a comprehensive range of factors influencing human functioning, including cognitive, emotional, and social dimensions, all of which affect communicative ability.

Systematic reflection with conclusions and recommendations

Language is viewed as an integral component of the human cognitive process, serving not only as a communication system but primarily as a tool for thinking and organising knowledge. Linguistic structures are represented in the mind and utilised in the processes of thinking and understanding the world. Consequently, speech and language disorders are recognised as complex communication issues that necessitate an interdisciplinary approach to both diagnosis and therapy. Stanisław Grabias emphasises that understanding the functions of language is crucial in shaping therapeutic interventions. A detailed analysis of linguistic behaviour reveals the areas of influence and directs appropriate interventions. Language serves as a gateway to the cognitive dimensions of humanity, which encompass knowledge about oneself and the

world, as well as the ability to apply that knowledge. It functions as a tool for expressing human emotions and desires and situates individuals within the fabric of social life, thus providing direction within the model of standard procedures (Grabias, 2020). This conceptualization of the cognitive aspect of language has practical applications in both diagnosis and therapy.

Advancing diagnostic and therapeutic methods is essential for enhancing the quality of life for individuals with communication disorders. A thorough diagnosis should involve a comprehensive assessment that includes medical examinations, evaluations of speech and language, and psychological testing. Effective therapy requires a multifaceted approach tailored to the specific type and severity of the disorder. This may encompass speech and language therapy, behavioural therapy, and the use of augmentative and alternative communication (AAC) techniques.

The interdisciplinary field of cognitive science examines the mechanisms of language processing, including both production (such as speaking and related disorders) and comprehension (listening). Insights from research into speech and language disorders can be applied to psycholinguistic experiments and to the development of computer models for natural language processing.

References

Emiluta-Rozya, D. (2012). Formy zaburzeń mowy. In: S. Grabias, M. Kurkowski (ed.), *Logopedia. Teoria zaburzeń mowy. Podręcznik akademicki* (pp. 73–87). Lublin: Wydawnictwo UMCS.

Grabias, R. (2012). Teoria zaburzeń mowy. Perspektywy badań, typologie zaburzeń, procedury postępowania logopedycznego. In: S. Grabias, M. Kurkowski (ed.), *Logopedia. Teoria zaburzeń mowy. Podręcznik akademicki* (pp. 15–71). Lublin: Wydawnictwo UMCS.

Grabias, S. (2020). Logopedia polska w stuleciu Polski Niepodległej. Trzy perspektywy rozwoju dyscypliny. *Logopedia*, *47*(2), 517–537.

Greene, J. (1972), *Psycholinguistics. Chomsky and psychology*. Harmondsworth: Penguin.

International Statistical Classification of Diseases and Related Health Problems (ICD-10). (2016). https://icd.who.int/browse10/2016/en. Accessed 25.09.2024.

Jakobson, R. (1989). *W poszukiwaniu istoty języka 1, 2. Wybór pism*, M.R. Mayenowa (ed.). Warszawa: Państwowy Instytut Wydawniczy.
Kaczmarek, L. (1975). Korelacyjna klasyfikacja zaburzeń słownego i pisemnego porozumiewania się. *Logopedia*, *12*, 5–13.
Kania, J.T. (1982). *Szkice logopedyczne*. Warszawa: Wydawnictwo Szkolne i Pedagogiczne.
Kurcz, I. (1987). *Język a reprezentacja świata w umyśle*. Warszawa: Wydawnictwo PWN.
Mayenowa, M.R. (1989). Roman Jakobson – uczony i człowiek. In: R. Jakobson, *W poszukiwaniu istoty języka*. Warszawa: Państwowy Instytut Wydawniczy.
Styczek, I. (1983). *Logopedia*. Warszawa: Państwowe Wydawnictwo Naukowe.
Zalewski, T. (1993). Klasyfikacja zaburzeń mowy. In: T. Gałkowski, Z. Tarkowski, T. Zalewski (ed.) *Diagnoza i terapia zaburzeń mowy* (pp. 7–39). Lublin: Wydawnictwo UMCS.

Sławomir Sztajer
Adam Mickiewicz University in Poznań
https://orcid.org/0000-0001-8975-2994
https://doi.org/10.35765/slowniki.423en

The cognitive science of religion

Summary

DEFINITION OF THE TERM: The cognitive science of religion is an interdisciplinary approach within the study of religion that focuses on the cognitive determinants of thoughts and behaviours. It examines the cognitive processes and mechanisms involved in the formation, assimilation, and transmission of religious ideas, as well as the processes and mechanisms that underpin ritual behaviours. The cognitive approach also draws on knowledge about the evolution of the human mind, the neural basis of religious experience, and the interactions between the mind and the world.

HISTORICAL ANALYSIS OF THE TERM: Systematic research within the cognitive science of religion began in the 1990s, when a number of theories were formulated which determined the direction of research in the following decades. The development of the cognitive science of religion was partly motivated by dissatisfaction with the findings in the traditional study of religion.

DISCUSSION OF THE TERM: The cognitive science of religion is based on the assumption that religion results from ordinary, natural cognitive mechanisms. Religions are based on mental mechanisms that have evolved in response to the environmental challenges faced by our ancestors over many millennia. The thesis of the naturalness of religious ideas has a significant methodological implication: religious phenomena should be studied in the same ways as all other natural phenomena.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: The main advantages of the cognitive science of religion

are its interdisciplinarity and its commitment to accounting for the diverse determinants to which religious beliefs and behaviours are subjected and to integrating knowledge from cognitive science, neuroscience, evolutionary theory, and cultural studies.

Keywords: religion, ritual, cognitive processes, evolution

Definition of the term

The cognitive science of religion is a relatively new area of research established in the 1990s within the broadly understood field of cognitive science. The contemporary cognitive science of religion represents a multidisciplinary and sometimes interdisciplinary approach to the study of religion. Apart from cognitive scientists, it encompasses scholars from diverse disciplines such as the study of religion, history, archaeology, cultural studies, biology, anthropology, and cognitive, evolutionary, social and developmental psychology (White, 2021, p. 1). The term 'cognitive science of religion' is translatable in Polish into two different terms. One is '*kognitywistyka religii'* [the cognitive science of religion], which is legitimate given the origins of this approach and constitutes a sub-discipline of the cognitive sciences. However, given the increasing role of the cognitive approach in the study of religion, it can also be translated into '*religioznawstwo kognitywne*' [the cognitive study of religion], which is in line with the names used in other disciplines, e.g., cognitive anthropology or cognitive sociology. In this case, it constitutes either a sub-discipline of the study of religion or a theoretical and methodological approach within the study of religion. The two different understandings of cognitive research on religion are not unusual because other sub-disciplines of contemporary the study of religion, such as psychology of religion, sociology of religion, or anthropology of religion, also function as sub-disciplines of their respective sciences, i.e., psychology, sociology, or anthropology.

The cognitive approach adopted within the study of religion is based on the assumption that religious thinking and religious behaviours can be explained through the natural processes and mechanisms that constitute the human cognitive system. The cognitive processes and mechanisms that underpin religion are explained in evolutionary terms. Cognitive scholars of religion are interested in not only how the workings of the human mind influence religion but also how the human mind was shaped by natural selection in the environment inhabited by our ancestors. It is argued that at least some mental mechanisms have been shaped in response to the recurring problems people faced in their evolutionary past. An evolutionary perspective, especially evolutionary psychology, has been present in the cognitive science of religion from

its very beginning (Liddle & Shackelford, 2021). Research on the human mind to which scholars of cognitive science of religion refer also includes searching for the neural correlates of religious thinking and religious behaviour and examining the complex interactions between the mind and the natural and social environment (Schjoedt, 2009).

The cognitive science of religion began with examining the acquisition, representation, and processing of religious concepts and beliefs, as well as the impact of religious representations on behaviour. Cognitivists also seek answers to fundamental questions such as the origins of religion, the widespread occurrence of religious ideas and behaviours across cultures worldwide, as well as the ease with which religious ideas are acquired by human individuals and spread within populations. The cognitive science of religion encompasses key dimensions of religion: religious concepts and beliefs, rituals, religious experiences, various forms of social organisation of religious life, and religious ethics. Additionally, phenomena traditionally studied by religious scholars (such as magic or myth), although not constitutive of religious traditions, are also of interest.

The cognitive science of religion is multidisciplinary and involves not only religious scholars, but also psychologists, anthropologists, sociologists, neuroscientists, biologists, historians, and archaeologists. The goal of the cognitive science of religion moves beyond merely juxtaposing the achievements of various scientific disciplines: it is directed at creating interdisciplinary projects that integrate knowledge produced within these disciplines and obtaining a comprehensive understanding of research problems. Broad interdisciplinarity prevents the reduction of religious phenomena to a single type, which is a procedure that in the history of the study of religion has led to the formulation of simplistic, one-dimensional theories of religion. Instead, cognitivists construct multi-causal models and consider different aspects of the phenomena they study.

Historical analysis of the term

Although systematic research within the cognitive science of religion began in the 1990s, the first article heralding this research strand was published by Stewart Guthrie in 1980 under the title “A Cognitive Theory

of Religion", in which a cognitive approach to religious beliefs and practices is proposed and the key role of anthropomorphisation in religious thinking is emphasised. Guthrie expanded his theory in a monograph published in 1993. In the 1990s, several other cognitive theories of religion were proposed that marked key research directions. They included the cognitive theory of religious ritual systems (Lawson & McCauley, 1990); Pascal Boyer's theory of the cultural transmission of religious ideas, which emphasised the role of cognitive constraints on the emergence and spread of religious ideas (Boyer, 1994); and Harvey Whitehouse's early version of his theory of two modes of religiosity – imagistic and doctrinal, which he later developed and published in a monograph (Whitehouse, 2004). Subsequent decades saw the formulation and empirical testing of new hypotheses, including the use of experimental studies.

The development of the cognitive science of religion was driven by dissatisfaction with the findings of the traditional study of religion, where phenomenological, historical-comparative, and socio-cultural approaches were prominent, and by doubts regarding the methods used in the study of religion. Critics noted that these approaches often prioritised the interpretation of religious phenomena over their explanation and concluded that this limitation would be overcome not by replacing interpretation with explanation but by restoring the balance between the two by formulating explanatory theories (Lawson & McCauley, 1990, p. 14). Some interpretive approaches treat the explanation of religious phenomena as an attempt to reduce them to non-religious phenomena. Phenomenological researchers, in particular, have advocated recognizing religion as a *sui generis* reality that is irreducible to the non-religious phenomena studied within other sciences. The cognitive approach challenges the belief in the autonomy and irreducibility of religion, viewing religious phenomena as part of the human world and subject to natural determinants. As such, they can be explained in terms of non-religious phenomena.

Cognitivists have also criticised various socio-cultural approaches, particularly those that emphasized the diversity and incomparability of religious phenomena and focused on the cultural determinants of human behaviour. Cognitive scientists argue that behind the diversity of religious phenomena are recurrent ideas and behaviours that can be

explained by the universal structure of the human mind. Cultural factors are important, but diverse forms of religious life cannot be explained by cultural influences alone. Moreover, the dynamics of cultural systems are largely shaped by the cognitive constraints inherent in human minds that are related to memory, attention, and information processing. These constraints can be explained by well-established biological theories, such as evolutionary theory and the findings of neuroscience.

Cognitive scientists also highlight those aspects of religion that have been rarely investigated before, such as the fact that religious ideas and beliefs are not limited to the ideas and beliefs that make up religious doctrines in specific religions. An official doctrine may be normative, i.e., it defines what to believe in, but it certainly does not exhaust the stock of ideas endorsed by religious people. Cognitive scientists argue that the study of religion has focused too much on official doctrines and declared beliefs, but not enough on how religious people actually think. Experimental methods are often used in the cognitive science of religion to uncover this 'unofficial' version of religious thinking.

Discussion of the term

The fundamental thesis formulated within the cognitive approach is the naturalness of religion thesis. Religion is a natural phenomenon not only in the sense that it depends on universal, non-cultural cognitive determinants, but also because it is the result of the operation of natural, ordinary cognitive mechanisms (Boyer, 1994). Religious thinking does not rely on cognitive abilities that are unique to religion; rather, religions utilise mental mechanisms that have evolved over millions of years of environmental challenges faced by our ancestors. The thesis of the naturalness of religion has significant methodological implications: religious phenomena should be studied in the same way as all other natural phenomena. Explaining religion does not require postulating the existence of any supernatural reality.

In explaining religious phenomena, many cognitive scholars of religion adopt an evolutionary approach based on the assumption that the mind, which is composed of specialized cognitive mechanisms, is a product

of evolution. Each of these mechanisms is a response to a specific adaptive problem faced by our ancestors. The cognitive mechanisms that underpin religious beliefs and convictions are explained based on evolutionary theory, although researchers can come to no consensus regarding defining religious ideas and behaviours from the perspective of evolutionary theory. Adaptationists argue that at least some religious beliefs and behaviours are adaptive in terms of enhancing survival and the reproduction of human individuals or groups, and in terms of the adaptation of certain individuals or groups compared to other human individuals or groups. The adaptive nature of at least some religions is argued to be supported by their role in coping with various problems. Among other things, religions can have a positive impact on physical and mental health, provide individuals with meaningful, holistic, and optimistic visions of the world, reduce the rationally motivated deception of others (thus they stabilise social exchange), and sustain cooperation between individuals from the same social group. On the other hand – and this is an argument against the adaptive nature of religion – religious behaviours involve significant costs for believers (in terms of, e.g., time, energy, and material resources); in some cases, they can even pose a threat to the life and well-being of individuals and social groups (e.g., extreme asceticism, ritual suicides, and religious conflicts).

Most cognitive scientists of religion question the adaptive nature of religious behaviours. They argue that religious beliefs and behaviours are a by-product of adaptive cognitive mechanisms but do not themselves have any adaptive value. For example, the concept of God as a being who possesses a mind and thinks, feels, and acts in an intentional manner is not an adaptation itself but rather a product of a cognitive mechanism which is an adaptation that underlies the human capacity to represent the mental states of others and to predict their behaviour. This adaptation has evolved in response to the complexity of collective life, irrespective of religion. Moreover, attempts to explain religion in terms of cultural evolution are also put forward within the cognitive approach. In this case, the unit of evolution is a certain cultural trait or a meme.

In the cognitive science of religion, considerable importance has been attached to the beliefs that underpin the doctrine of a religion that are verbalised, systemised, and rationalised by religious experts. Such beliefs constitute official doctrine and certainly have a crucial impact

on believers' thinking. However, as cognitive researchers observe, religious thinking is not limited to doctrinal concepts and beliefs as religious people in everyday life tend to use religious ideas that deviate significantly from those preached in official doctrines. According to Justin L. Barrett (2004), the behaviour of religious people shows a clear tendency towards a peculiar theological correctness: they declare beliefs that align with official doctrine; at the same time, in situations that require rapid information processing, they use concepts and beliefs that deviate significantly from the doctrinal norm. For example, while praying, a religious person may imagine God not as an omniscient, omnipresent, immutable, and transcendent being but as a very human-like person who understands the ethnic language in which the believer is praying and who is subject to various constraints: temporal, spatial, cognitive, and emotional. Even though some religions' doctrines prescribe thinking of God as a being fundamentally different from man, in various situations of everyday life believers use an anthropomorphic image of God.

People have diverse concepts of supernatural beings as a consequence of the existence in their minds of various information processing systems. There is no agreement among researchers on how these systems should be differentiated. Some postulate the existence of two processing systems, and, for the sake of simplicity, it can be assumed that the mind functions by relying on two processing systems: intuitive and reflective. Intuitive processing is fast, automatic, and unconscious, while reflective processing is slow, controlled, and conscious. Information processing based on these two different systems is not specific to religious thinking but it plays a particularly important role in religions because it enables religious concepts to be applied in various circumstances, including advanced theological thinking and colloquial religious interpretation of everyday experiences.

There is also a tendency in religious thinking to create anthropomorphic representations. The anthropomorphisation of supernatural beings, understood as the attribution of human characteristics to these beings, is a phenomenon long recognised by philosophers and scholars of religion. However, cognitive theories of religious anthropomorphisation have been developed rather recently and include the theory proposed by Stewart Guthrie (1993), according to which anthropomorphisation is a natural and universal tendency of the human mind. The propensity

to anthropomorphise is probably an adaptation that enables the rapid, though not always accurate, detection of threats and opportunities in the environment. The use of an anthropomorphic model to understand the world is done, as it were, automatically, in a non-reflective way. Humans have an intuitive ability to understand other humans, and they use this ability to understand the non-human world. While the propensity to anthropomorphise is not limited to religious thinking, anthropomorphic religious beliefs are more systematised and integrated than non-religious ones and are sometimes sanctioned within doctrine.

Cognitive researchers argue that religious concepts are not formed in a haphazard manner but follow a universal pattern determined by cognitive inclinations and the constraints of the human mind (Boyer, 1994; Boyer, 2001). A cursory look at the enormous stock of religious ideas generated in thousands of known world religions gives the impression that religious imagination has no limits. However, this impression is mistaken, as in-depth analyses of the structure of religious representations demonstrate. The ideas of supernatural beings are created using a limited catalogue of characteristics. A key role in this process is played by the conceptual schemata with which human minds categorise the world, based on intuitive expectations of what things are in the world. According to many researchers, such intuitive knowledge is largely independent of culture. As a result, religious concepts, although not culturally universal, arise from an intuitive understanding of the world that is the product of universal cognitive processes (Boyer, 1994, p. viii).

Although notions of supernatural beings are based on intuitive knowledge, they also contain counterintuitive features, which means they violate at least some intuitive expectations about how things are in the world (Boyer, 2001). Cognitivists point out that our common-sense thinking about the world is based on an intuitive ontology. This is knowledge that, unlike scientific knowledge, is not the product of reflection but arises from automatic inference mechanisms. Usually, we are not aware that our everyday behaviours are based on 'intuitions' which allow us to act efficiently without reflective thinking: when leaving a room, we head towards the door and not towards the wall because our intuitive physics tells us that a solid body does not penetrate another solid body. Religious ideas also arise from intuitive knowledge of the world – such as that offered by folk physics, folk biology, or folk psychology – but they also

have properties which contradict intuitive ideas. A ghost, for example, is a person that, like other persons, possesses mental states, such as perceptions, beliefs, desires, and intentions. Thus, in this respect, our thinking about ghosts aligns with the intuitive psychology that underlies our thinking about persons. However, a ghost also possesses properties that contradict intuitive knowledge, such as the ability to penetrate walls. Properties that are incompatible with intuitive knowledge make the notion of a ghost attract attention, while properties consistent with an intuitive ontology trigger mental systems of inference.

Research on how people remember concepts indicates that concepts that contain a single property which is incompatible with intuitive knowledge are remembered better than concepts which are compatible with intuitive knowledge and better than concepts which contain many properties that are incompatible with intuitive knowledge. Since ideas that are easy to remember are more likely to be disseminated through the human community, it is likely that they gain an advantage over less well-remembered ideas (Atran, 2002). According to researchers who work on the psychological determinants of culture, the workings of the mind have a significant impact on cultural transmission. In particular, cognitive constraints, including those linked to memory, allow certain types of mental representations to gain an advantage in cultural transmission and thus become more easily disseminated. Religious ideas belong to such representations.

However, it should be remembered that cognitive models of cultural transmission tend to overlook non-cognitive determinants of transmission, such as the technology of writing, in which dissemination significantly alters the way in which religious ideas are spread. Writing and the various material artefacts used in religions constitute memory reservoirs that are external to the organism and substantially extend biological memory. The religions of literate cultures enable the creation and dissemination of highly counterintuitive religious ideas and elaborate theological systems. Concepts of God created within rationalised theological systems are characterised by a high degree of counterintuitiveness, which means they include a large number of counterintuitive traits.

Religious concepts and beliefs are an important dimension of religion, but not the only one. Religious ritual acts (i.e., symbolic behaviours subject to strictly defined rules, which are repetitive, dramatised, not

oriented towards practical goals, and usually refer to beliefs in supernatural beings and powers) play a crucial role in religions. To answer the question of why people engage in ritual behaviours and what functions they perform in human societies, various approaches, including functional, symbolic, phenomenological, structural, and psychoanalytic, have been developed in studies on rituals. These studies focus on, e.g., the communicative nature of ritual behaviours, based on the premise that rituals serve to convey information. Consequently, much attention has been given to the interpretation of rituals and less to the explanation of ritual acts, particularly to questions such as: Why do people undertake ritual activities? Why are some rituals repeated frequently (e.g., Mass), while others are performed only once (e.g., baptism)? Why do the effects of some ritual acts last, while others fade over time? Finally, why do some rituals evoke strong emotional responses, while others are monotonous, even boring? Cognitive theories of ritual seek answers to these questions by formulating empirically testable hypotheses, which are verified through observations of rituals in different religions.

Religious ritual acts differ from instrumental activities undertaken in everyday life, where a practically defined goal is achieved through the selection of appropriate measures. To open the door of my house, I have to perform a series of rational activities: find a bunch of keys, choose from them the one that fits the lock, make the appropriate movements, press the handle, push the door, and so on. In a ritual, the rationality of the action recedes into the background; activities and their results are not causally linked, and the participants in the ritual are usually unable to explain how their activities have led to the expected results, or they give vague and mysterious answers when asked this question. Another vital feature that distinguishes rituals from activities in everyday life is pertinent here: the fact that a ritual act does not achieve its intended purpose is not a reason to question it and abandon it in the future. A ritual intended to bring rain is very likely to be repeated in the future, even if has repeatedly failed to produce the intended effect in the past. In everyday life, our activities are governed by a different set of rules: if a key does not fit a lock, another key is selected, and the ill-fitting one is cast aside as useless. Consequently, a ritual is hardly susceptible to change.

Robert N. McCauley and Thomas Lawson (2002) proposed a cognitive theory of ritual actions that explains rituals in terms of mental

representations. This theory postulates the existence of a specific ritual competence, similar to the linguistic competence studied by Noam Chomsky, among others. Ritual actions are represented in the mind based on the ordinary action representation system. In other words, people intuitively understand ritual as a particular type of action that has a specific form. The form of action consists of three elements: the agent, the action itself, and the object. The understanding of a religious ritual is based on our intuitions about actions in general but also covers specific elements such as the notion of supernatural agency and the supernatural effects of the action. Various tools are used in religious ritual actions, and their place in the structure of these actions facilitates the understanding of rituals by those who participate in them. A ritual is thus structured as follows: person A acts on person or object B through tool T in order to cause certain supernatural effects through supernatural causality. The presence of culturally postulated superhuman agency distinguishes religious ritual acts from other acts. Rituals differ in terms of elements in their structure to which supernatural causality is attributed. Using this criterion, two basic types of ritual can be distinguished: one in which superhuman agency is associated with the agent of the act, and one in which this agency is associated with the object or tool of the ritual act. Information about the location of superhuman agency in the structure of the ritual is important because it allows predictions about the characteristics of the ritual to be made that can refer to its repetitiveness and the frequency of repetition, the level of emotional arousal of the participants, the permanence of the results of the ritual acts, and the location of the ritual (central or peripheral) in a given religion. For example, McCauley and Lawson observe that rituals in which agency is attributed to the agent of the act are less frequently repeated, their effects are considered enduring, they elicit strong emotional responses from participants, and they are perceived to be central within a religion. A good example of such an act is baptism in the Christian tradition, during which the agent of the act (the priest) administers the sacrament. This ritual does not need to be repeated many times, its effects are permanent, and it is a key Christian rite. Conversely, rituals in which superhuman agency is linked to the object or tool of the act are repeated more often and their results are less enduring; they do not evoke strong emotions and usually do not play a key role. An example is making an offering to a god.

Rituals can also be differentiated by the modes of religiosity they manifest, as described by the British anthropologist Harvey Whitehouse. He identified two basic modes of religiosity – imagistic and doctrinal – which, although they can co-exist in various religions, are based on different cognitive mechanisms, in particular different types of memory (Whitehouse, 2004). Whitehouse paid particular attention to how particular modes of religiosity influence the cultural transmission of religious ideas and behaviours. The imagistic mode of religiosity includes rituals that are repeated infrequently but are highly spectacular. They evoke strong impressions in participants and are remembered through episodic memory. They also increase social cohesion and substantially contribute to the individual's identification with a social group. Religions with a predominantly imagistic mode do not develop strong religious leadership, hierarchy, nor doctrinal orthodoxy. Because they are grounded in a local cultural context, they hardly disseminate outside this context. An example of a ritual typical of this type of religiosity is the initiation ritual practised in traditional cultures. During such a ritual, the initiate is subjected to acts that they usually remember for a lifetime. At the same time, the initiation ritual, as a result of which a boy becomes a man, is a one-off, unique ritual. The doctrinal mode of religiosity is characterised by the frequent repetition of rituals and the doctrinal message. While the doctrinal message is based on semantic memory, the repetition of rituals is based on implicit procedural memory, with the result that rituals are rarely reflected on and modified. This is conductive to the preservation of orthodoxy, which is further sustained by religious experts, who provide an authoritative interpretation of doctrine and ritual. In this mode of religiosity, frequently repeated rituals – such as in Lawson's and McCauley's conception – are not accompanied by intense emotions. Since religious content is communicated in the form of doctrine, sometimes using writing, religions based on this mode of religiosity disseminate more easily in diverse cultural environments. Organised Christian churches are a good example of institutions based on a doctrinal mode of religiosity. However, it should be emphasised that the two modes of religiosity – imagistic and doctrinal – often co-occur within specific religions, and the predominance of one allows a given religion to be described as either imagistic or doctrinal.

The cognitive processes that underpin religion play a key role in shaping religious behaviours. However, the study of cognitive processes

is not only concerned with what happens in our heads. Some researchers argue that cognitive processing also encompasses elements of the environment from outside the organism. They treat human cognition as a process that takes place not only in the brain but also in the body and in the world that surrounds the organism. Humans have developed the ability to use material artefacts as tools for thinking. There is no consensus as to whether these artefacts, which are external to the organism, should be regarded as elements of the cognitive system or whether we should instead speak of the interaction of the cognitive system with the environment. However, there is no doubt that human cultures make extensive use of the ability to delegate certain cognitive tasks externally and to process information based on various elements of the man-made environment. In this context, religious thinking can be understood as a process involving both the workings of the cognitive system and its interaction with the environment.

Material objects – both those that are part of the natural world and those that are man-made – are an intrinsic dimension of almost every religion. The former include sacralised objects of nature, such as sacred mountains, rivers, trees, stones, celestial bodies, and animals; the latter include material artefacts, such as paintings, icons, relics, liturgical vestments, temples, etc. In principle, any material object can be sacralised. Indeed, in the history of world religions, various objects have been considered sacred. Material artefacts are not merely the background to or decorations of religious life but perform a number of important cognitive functions (Sztajer, 2018). They can be treated as tools that support and sometimes enable the creation, understanding, processing and communicating of religious ideas. Material representations of supernatural beings 'embody' supernatural worlds which are inaccessible to the senses by presenting what was previously the object of indeterminate imaginings in a concrete, tangible form.

What cognitive functions do material objects perform in religion? First, they make it possible to externalise an inherently subjective religious consciousness and to bring the content of this consciousness into public circulation. This is an extremely important function because religions are collective endeavours that are shared by the members of a community. The representation of religious ideas in material form means they are perceived as belonging to a reality that is external to the subject, thus

they are no longer merely subjective but now intersubjective. Second, the externalisation of religious ideas involves anchoring them in more tangible material representations. Conceptual structures obtain greater permanence by being linked to their corresponding material structures. Believers may use different images of a deity, but tying these images to a material representation which is external to the mind, such as a sculpture of the deity, gives these images greater permanence, especially considering the multi-generational continuity of religions. Third, material artefacts expand the possibilities of biological memory. Religions, like other cultural domains, rely on memory reservoirs external to the organism, which broaden both individual and collective memory. Fourth, material artefacts make it possible to give concrete form to abstract religious ideas by condensing the supernatural world that is inaccessible to the human senses into a tangible form that is accessible to these senses. In this way, the processing of religious representations involves sensorimotor processes. Fifth, material artefacts make the supernatural world appear accessible, present in the here and now and, above all, more real. A sense of the reality of supernatural beings and powers is induced by linking religious ideas to practically tangible objects that are not only accessible to the senses but can also be manipulated in ritual acts. Some scholars further point out that in the early stages of the development of religion, material artefacts, especially physical representations of supernatural beings, may have functioned as prototypes of religious ideas rather than reflections of them. In other words, religious ideas may have arisen as representations of material objects by emerging from human contact with such objects.

Systematic reflection with conclusions and recommendations

In recent decades, the cognitive science of religion has undergone numerous modifications which testify to the expansiveness of this research field. The main advantages of this approach are its interdisciplinarity and its commitment to accounting for the diverse determinants to which religious beliefs and behaviours are subjected. Currently, the cognitive approach seeks to integrate knowledge from cognitive science,

neuroscience, evolutionary theory, and cultural studies. Another advantage is its capacity to formulate new hypotheses that are subjected to systematic and reproducible empirical tests, including experimental tests. The application of methods used in the cognitive and evolutionary sciences to the study of religious phenomena brings the study of religion closer to the accepted standards of science.

Critics of the cognitive approach enumerate a number of limitations. The objections they formulate primarily concern reductionism, which is an essential component of the cognitive explanation of religion, the use of an approach that is too narrow, the restriction to one type of factor, and the marginalisation of interpretation and understanding in favour of explaining religious phenomena in terms of causal mechanisms (see Cohen et al., 2008). Some of these objections seem valid when applied to early cognitive theories on religion, but they are unjustified when considering more recent research directions and developments in the cognitive science of religion which account for the multidimensionality and high diversity of religious phenomena, as well as the issues of understanding and interpretation. Accusing cognitive theories of religion of reductionism is justified in the sense that explaining religious phenomena involves reducing them to phenomena of a different type. The creators of such theories reject the assumption – widespread among phenomenologists of religion – of the autonomy of religion and the irreducibility of religious phenomena. However, it should be borne in mind that the reductionism present in the cognitive approach fundamentally differs from the reductionism present in certain classical theories of religion, which attempted to fully explain religion by relying on lower-order social, psychological, economic, etc. phenomena. In its extreme form, such reductionism led to questioning not only the distinctiveness and specificity of religion but also whether religion is real. Cognitive theories of religion make it possible to develop multidimensional models of religious phenomena that take into account diverse determinants. An important role here is played by explanatory pluralism (i.e., the multiplicity of ways of explaining), which, when applied to religion, means that it is impossible to explain such a complex and multidimensional phenomenon as religion in a complete and definitive way. Instead, it is proposed to consider different levels of analysis of the phenomenon and to use a variety of methods.

The cognitive science of religion not only deepens our understanding of religious phenomena but also provides knowledge that can be used in practice, especially regarding religious education, the activities of religious groups, or religious policy-making. The cognitive science of religion also constitutes a point of reference in worldview disputes about religion, but it is highly unlikely that the knowledge it provides will make it possible to settle such disputes.

References

Atran, S. (2002). *In Gods We Trust: The Evolutionary Landscape of Religion*. Oxford: Oxford University Press.

Barrett, J.L. (2004). *Why would anyone believe in God?*. Walnut Creek, CA: AltaMira Press.

Barrett J.L. (ed.). (2022). *Oxford Handbook of the Cognitive Science of Religion*. Oxford: Oxford University Press.

Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion*. Berkeley: University of California Press.

Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Cohen, E., Lanman, J., McCauley, R., & Whitehouse, H. (2008). Common Criticisms of the Cognitive Science of Religion-Answered. *The CSSR Bulletin*, *37*, 112–115.

Guthrie, S.E. (1993). *Faces in the clouds: A new theory of religion*. New York: Oxford University Press.

Lawson, E.T., & McCauley, R.N. (1990). *Rethinking religion: Connecting cognition and culture*. Cambridge, UK: Cambridge University Press.

Liddle J.R., & Shackelford T.K. (ed.). (2021). *The Oxford Handbook of Evolutionary Psychology and Religion*. Oxford: Oxford University Press.

McCauley, R.N., & Lawson, E.T. (2002). *Bringing Ritual to Mind: Psychological Foundations of Cultural Forms*. Cambridge: Cambridge University Press.

McCauley, R.N. (2011). *Why Religion is Natural and Science is Not*. Oxford: Oxford University Press.

Schjoedt, U. (2009). The religious brain: A general introduction to the experimental neuroscience of religion. *Method & Theory in the Study of Religion*, *21*(3), 310–339.

Sztajer, S. (2018). *Pojęciowa konstrukcja świata religijnego. Przekonania religijne a procesy poznawcze*. Poznań: Wydawnictwo Naukowe Wydziału Nauk Społecznych UAM.

White, C. (2021). *An Introduction to the Cognitive Science of Religion: Connecting Evolution, Brain, Cognition, and Culture*. Milton: Taylor and Francis.

Whitehouse, H. (2004). *Modes of religiosity: A cognitive theory of religious transmission*. Walnut Creek, Calif.; Oxford: Altamira Press.

Michał Wierzchoń
Jagiellonian University in Krakow
https://orcid.org/0000-0002-7347-2696
https://doi.org/10.35765/slowniki.486en

Sensory substitution

Summary

DEFINITION OF THE TERM: In this article, the concept of sensory substitution is discussed. This section presents considerations on the definition of the concept and demonstrates the conditions necessary for substitution to occur, with analysis of selected examples.

HISTORICAL ANALYSIS OF THE TERM: This section includes a historical overview of sensory substitution technology, emphasising the importance of developing non-invasive prosthetic devices tailored to the needs of visually impaired individuals.

DISCUSSION OF THE TERM: Three areas of sensory substitution research are discussed in this section: 1) research on the development of substitution devices, 2) neurobiological research that investigates neuroplasticity resulting from the use of substitution devices, and 3) research on the subjective experience of substitution.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: This section proposes directions for further research on sensory substitution, which should aim to enhance the integration of sensory substitution technology into everyday life.

Keywords: sensory substitution, experience, perception, neuroplasticity, vision rehabilitation

Definition of the term

Sensory substitution enables individuals deprived of access to environmental information through a specific sense (e.g., visual information for visually impaired individuals) to receive it through another sensory modality (e.g., hearing or touch). A parking sensor is an example of substitution. In this case, visual information about the distance to an obstacle is substituted with auditory information. Thus, hearing (a substituting sense) replaces vision (a substituted sense).

Today, the term 'sensory substitution' refers to the use of prosthetic technologies, known as sensory substitution devices, which provide information. These devices typically consist of a sensor (e.g., a camera), a stimulator (e.g., a headset), and a device that implements an algorithm to translate the sensor's signal. The goal is for the sensor's signal to become accessible and intelligible for the substituting sense. By using sensory substitution devices, information normally perceived through the substituted sense can be presented to the substituting sense. As previously mentioned, a parking sensor serves as the simplest example of such devices: visually inaccessible information about objects behind a car becomes accessible via a parking sensor (i.e., a simple algorithm that translates information about distance into sound frequency) and an effector in the form of a loudspeaker that emits sounds generated by the algorithm.

The development of sensory substitution technologies is primarily driven by their potential for rehabilitation. These devices can provide visually impaired individuals with information about environmental characteristics typically perceived through vision, such as an object's distance, shape, or colour (Kałwak, Reuter, Łukowska, Majchrowicz, & Wierzchoń, 2018). Most research in this field has been conducted with individuals who have been blind since birth. However, sensory substitution is also applicable to those who lost their sight later in life, as well as sighted individuals whose access to sensory information is temporarily restricted (Łukowska et al., 2021).

Historical analysis of the term

The term sensory substitution was introduced into the scientific literature by Paul Bach-y-Rita in the 1980s. It builds on earlier research on developing various technological solutions aimed at assisting visually impaired individuals in perceiving their environment. Sensory substitution is possible thanks to the creation and advancement of new non-invasive prosthetic devices that use computer-assisted processing of sensory information that is inaccessible to the visually impaired. Unlike traditional aids, such as a white cane, sensory substitution relies on advanced technology. These devices operate independently of other widely used assistive technologies, such as mobile phone applications. The goal of sensory substitution devices is to make visual information accessible to visually impaired individuals through another functional sense, such as hearing or touch.

The first device described in the international literature as a sensory substitution device was an electrophthalm, an artificial eye developed by Polish ophthalmologist Kazimierz Noiszewski in the late 19th century. This device converted light stimuli into tactile and auditory stimuli. In the 1960s, another Polish ophthalmologist, Witold Starkiewicz, worked on a modified prototype of Noiszewski's device (Farmer, 1978). Starkiewicz's electrophthalm utilised a set of three hundred photocells that, when stimulated, activated vibrators attached to the visually impaired person's forehead. This technology enabled the creation of a simplified image of the visually impaired person's environment. While the concept of an electrophthalm and the findings of Starkiewicz's research have been published both in Poland and internationally, contemporary scientific publications on sensory substitution often overlook this information, typically identifying BrainPort as the first sensory substitution device.

The first version of the BrainPort device was described in the journal *Nature* in 1969 (Bach-y-Rita & Kercel, 2003). BrainPort is a sight-to-touch substitution device that converts images from an embedded low-resolution camera into patterns of mild electrical impulses. It provides information about the visual stimulus by encoding the brightness of each image pixel and translating it into the corresponding vibration strength of a stimulator placed on the tongue. Analogous to the camera's resolution, the stimulator's resolution is 20 by 20 pixels. BrainPort users can sense

patterns of brightness on the tongue's surface and learn to recognise the shape, size, and location of presented objects. The device operates in an online mode, which means it can also detect objects' movement in the environment; however, it does not provide information about depth, such as the distance of objects. BrainPort has been commercialised and available on the market since 1998. In 2015, the V100 version was approved for use by the Food and Drug Administration (FDA). Although commercially available, it is not widely used.

Originally, BrainPort was developed to support balance, particularly in post-stroke patients. Pioneering research into the neuroplasticity of individuals blind from birth has been conducted using this device. Research in this area has laid the foundations for many experiments in cognitive neuroscience, demonstrating that the brain's functional organisation is not immutable but can undergo significant reorganisation, even in adults (Heimler & Amedi, 2020).

Discussion of the term

Undoubtedly, the most important research challenge is the development of sensory substitution device technology. This task demands an interdisciplinary approach, requiring knowledge of, among others, the brain's sensory information processing, prosthetic device design, methodologies for testing their usability, programming skills, as well as experience with experiments using visually impaired subjects and methods for testing devices in the context of environmental interaction. Research in this field is typically conducted by interdisciplinary teams comprising psychologists, neuroscientists, computer scientists, engineers, doctors, and rehabilitation therapists. It is worth mentioning here that developing a universally acceptable technological solution for sensory substitution would significantly impact a large group of people. In 2012, the World Health Organization estimated that 40 million people were blind and 285 million had significant visual impairments. Many of these individuals live in developing countries, so the solution's usability and affordability are crucial for its widespread adoption. Apart from BrainPort and Noiszewski's electrophthalmic device, the literature describes a number of technological solutions in this area, including the Enactive Torch,

the vOICe, and the Colorophone (Łukowska, Kałwak, Osiński, Janik, & Wierzchoń, 2022).

The Enactive Torch is a sight-to-touch substitution device proposed by Tom Froese and colleagues (Froese, McGann, Bigge, Spiers, & Seth, 2012). It converts information about an object's distance into vibrations. In standard settings, the vibration intensity increases as an object is approached, similarly to parking sensors. The device employs an infrared or ultrasonic sensor and operates in an online mode, providing continuous information about the distance to obstacles. This functionality allows users to locate objects in space and estimate their size, enabling navigation in the environment despite the lack of visual information.

The vOICe, developed by Peter Meijer (1992), is a sight-to-hearing substitution device. It converts images from an embedded camera into auditory information, encoding details about the colour and position of objects in the perceptual scene. The vOICe operates offline, which means that it presents static perceptual scenes, such as photographs or still frames from a camera image. The perceptual scene is represented as a soundscape, where the pitch, loudness, and temporal sequence of sounds convey information about colour and the position of an object. The pitch of the sound encodes information about the vertical position of the object (high-pitched sounds represent an object located in the upper part of the visual field); the loudness of the sound encodes information about the brightness (a loud sound represents a very bright object), and the temporal sequence of the sound encodes information about the horizontal position of the object (sounds representing the image on the left are presented before those on the right side of the visual field).

The encoding algorithm for the vOICe device was developed by Amir Amedi in his EyeMusic technology, which includes information on the five primary colours – white, blue, red, green and yellow – alongside the rules for encoding object position and brightness. Colours are represented by different musical instruments. Peter Meijer, the creator of the vOICe, believes that, despite their apparent similarity, the vOICe and EyeMusic serve different purposes and cater to different user needs. In his opinion, the primary functions of sensory substitution devices are to navigate space, avoid obstacles, and recognise objects, none of which requires colour perception. The basis for Meijer's position is not clearly stated, but it significantly influences his technological approach.

This observation confirms that the development of sensory substitution devices relies not only on effective algorithms but also on a precise description of the target users' needs.

Among the many sensory substitution devices described in the literature, it is worth mentioning a sight-to-hearing device called the Colorophone, developed by Dominik Osiński (Osiński, Łukowska, Bizoń-Angov, Hjelme & Wierzchoń, 2021). This device translates colour information into auditory information. The Colorophone uses an algorithm to convert camera image signals into sounds delivered via open headphones (i.e., headphones that do not cover the users' ears). In the latest version, Colorophone 2.2, the algorithm employs the RYGB colour model, where each of the four primary colours – red, yellow, green, and blue – is assigned a different sound frequency (red corresponds to high frequency, yellow and green to medium frequencies, and blue to the lowest frequency). Following the rules of the RYGB colour model, all other colours are represented by combinations of these primary colours, with the intensity of each component encoded by the loudness of the corresponding sound. Users hear multiple sounds simultaneously and must identify the specific colour of an object.

Colorophone 2.2 operates in two modes: pointing and panoramic (cf. Łukowska, Kałwak, Osiński, Janik, & Wierzchoń, 2022). In the pointing mode, colour information is averaged from a set of pixels at the centre of the visual field (e.g., 50 x 50 pixels). Thus, scanning the entire perceptual scene requires head movements to shift the camera's central view. In the panoramic mode, information on colour is encoded from a strip of pixels across the entire visual field (left to right) at mid-height (e.g., 50 pixels high). Information from the left visual field is presented in the left headphone, from the right field in the right headphone, and objects in the centre generate stereophonic sound. This mode allows users to understand the spatial arrangement of objects without needing to scan the scene. Extensive research on the neural and cognitive mechanisms shaping the sensory substitution experience using the Colorophone has been conducted at the Consciousness Lab, affiliated with Jagiellonian University (Łukowska et al., 2021; Łukowska, Kałwak, Osiński, Janik, & Wierzchoń, 2022).

Research on developing sensory substitution devices can be categorised into two interrelated areas: behavioural research and usability

studies. Behavioural research focuses on evaluating the effectiveness of sensory substitution devices for specific tasks, aiming to measure task performance. Common tasks include object recognition (e.g., geometric shapes, letters, and faces), object localisation, estimating the number of objects, spatial navigation, obstacle avoidance, reading, and colour recognition (cf. Ptito et al., 2005 and Froese et al., 2012). Researchers assess both the accuracy and speed of task performance to determine how well a device can enhance the functioning of individuals deprived of a particular sense in real-world settings. Typically, behavioural tasks are designed to resemble everyday activities during which a device could assist visually impaired users. However, it is notable that researchers seldom consult with visually impaired individuals to identify their actual needs when developing these tasks. Instead, they usually rely on their own tacit assumptions regarding the needs of the visually impaired and the resulting assumptions about tasks that could be facilitated by sensory substitution devices. This issue seems vitally important because the visually impaired tend to develop effective adaptations, i.e., coping strategies, which they already use in their daily lives. Thus, it is crucial for sensory substitution devices to target tasks for which effective adaptations have not yet been developed, rather than duplicating existing solutions or integrating multiple adaptations into a single technological solution.

Developing a sensory substitution device that is genuinely useful to end users requires comprehensive usability studies. These studies aim to determine whether potential users are willing to utilise the device, focusing on its effectiveness, efficiency, and user satisfaction. Satisfaction is a crucial aspect of this evaluation, as even the most advanced technology will not be adopted if it fails to meet users' expectations, whether due to a non-intuitive interface, unpleasant sounds that substitute image, or the appearance or design of the device, which may be unacceptable according to users. Usability studies assess which features of sensory substitution devices need improvement to encourage more frequent use (Osiński, Łukowska, Bizoń-Angov, Hjelme, & Wierzchoń, 2021). They also provide recommendations on how to present the devices to end users, create user instructions, and design training programs to maximise their effectiveness.

These studies are crucial because market analyses indicate that despite the availability of numerous technological solutions, visually

impaired individuals seldom use sensory substitution devices in their daily lives. This may be due to researchers focusing primarily on the technological aspects rather than on testing usability under real-world conditions and considering the specific needs of end users and their existing adaptations to disability. In our research, we accounted for these issues when proposing a model for an interactive sensory substitution system (Łukowska, Kałwak, Osiński, Janik, & Wierzchoń, 2022). We emphasised that a sensory substitution system should encompass not only the device itself but also the environmental context in which it is used and the individual expectations and skills of the user.

From a cognitive science perspective, the design of sensory substitution devices themselves is as interesting as the identification of the neuronal mechanism responsible for the occurrence of the described effect. Research in this area has largely focused on brain neuroplasticity in persons using sensory substitution devices. Neuroimaging studies focus on tracking neuronal activation in both the sensory cortex of the substituted modality (e.g., the visual cortex) and the substituting modality (e.g., the auditory or somatosensory cortex) during sensory substitution. These studies investigate how neuronal activation changes in these areas after training with specific sensory substitution devices (Poirier, De Volder, & Scheiber, 2007). fMRI and PET studies have documented activation in areas typically involved in processing stimuli of the substituted modality in both visually impaired and sighted individuals, e.g., the visual cortex is activated during the presentation of sounds linked to visual information in the substitution algorithm. Increased connectivity between regions responsible for processing substituting and substituted information has also been observed. These neurobiological findings are considered evidence that sensory substitution can lead to neuroplasticity in the brain. However, results vary depending on the group; for example, different outcomes have been observed between individuals who lost their vision early versus later in life (Kałwak et al., 2018 and Poirier, De Volder, & Scheiber, 2007). Our team's study obtained results that indicate the importance of sensorimotor integration, manifested by increased engagement of the dorsal network when sensory substitution devices are used by sighted individuals who are temporarily deprived of visual information after extensive substitution training with the Colorophone device (Łukowska et al., 2021). These results reveal that while

the device was used effectively in practice, neuroplasticity did not extend to the activation of regions associated with visual experience.

Although some neuroimaging studies suggest that long-term training (involving super-users) with various types of visual sensory substitution devices leads to activation in brain areas typically associated with vision, such as the visual cortex, these results do not necessarily clarify the nature of the perceptual experience, i.e., whether users of sensory substitution devices truly see with another sense. The inconsistent findings across studies highlight the need for a unified theoretical model to interpret these results effectively. Poirier, De Volder, & Scheiber (2007) have proposed a model distinguishing between two processes that might account for the observed activation of the visual cortex during sensory substitution: cross-modality and mental (visual) imagery. The former process is multimodal integration, where neuroplasticity links the activation of visual regions to the sensory input provided by substitution devices. This mechanism is likely prevalent in individuals with early vision loss. The latter process, observed in sighted individuals using substitution systems in experimental settings, may instead reflect mental imagery. In this approach, sensory substitution can manifest as a visual experience, a multimodal experience, or an imagined image, depending on the experimental group.

Philosophers have also explored the issue of sensory substitution and its impact on mental representations. Focusing on the experience of substitution, they asked: What do subjects actually experience during sensory substitution? Does substituting visual information with auditory input resemble sight, hearing, or some other sensory experience? The literature indicates that our understanding of the nature of sensory substitution experiences remains limited (Auvray & Myin, 2009; Proulx, 2010). Discussions on this issue generally centre on 1) the qualitative characterisation of subjective experiences and perceptions when using sensory substitution devices (Block, 2003; Proulx, 2010); 2) analyses of how environmental interactions influence experience, inspired by an enactive approach (Kałwak et al., 2018); and 3) the development of methods for measuring and describing subjective characteristics of sensory substitution (Froese et al., 2012). Notably, most of these analyses remain within the realm of philosophical discourse. Empirical studies specifically examining the phenomenology of sensory substitution are

rare and often address dimensions of experience in a somewhat arbitrary or random manner.

Despite the lack of systematic analyses, it is worth mentioning a few pioneering studies in this area (cf., e.g., Froese et al., 2012). Notably, research conducted at the Consciousness Lab at Jagiellonian University has also investigated this topic. Our studies have focused on examining the boundary conditions of the substitution experience. Specifically, we tested whether it is possible for sighted individuals to experience sensory substitution in a way that extends beyond the characteristics of the substituting sense. For example, could one "see" through touch? We also explored whether experiences beyond the directly involved sensory modalities could occur (cf. studies postulating that sensory substitution can lead to synthetic synaesthetic experiences, e.g., Proulx, 2010). The findings of our study with sighted participants (Reuter, Kałwak, Sznajder, & Wierzchoń, under review) revealed that even short-term training with the Enactive Torch device could effectively assist users with specific tasks involving environmental interaction. However, we did not observe systematic changes in the qualitative characteristics of the reported experiences, and training did not lead to a fundamentally new form of substitution experience. Participants primarily reported different strategies for using the device to accomplish set tasks (such as finding an object). The only notable qualitative change observed was in participants' use of analogies to describe their experiences, with some reporting a sensation akin to approaching an object (e.g., the metaphor of approaching a stove). These findings, combined with a review of the literature, suggest that developing a qualitatively distinct experience of substitution may require long-term, systematic, and customised training.

Systematic reflections with conclusions and recommendations

In summary, current research on sensory substitution has demonstrated that individuals can effectively use sensory substitution devices to perform simple tasks in both experimental and real-world settings. Neuroimaging studies reveal that sensory substitution provides a valuable model for exploring neuroplasticity, showing that it can change the

functional organisation of the cortex and overcome the strong specialisation of individual cortical regions by enhancing multimodal integration. These findings underscore the potential benefits of sensory substitution technology for rehabilitating individuals with visual impairments. However, despite these advancements, the widespread practical use of such devices remains limited.

Future research should prioritise usability studies to ensure that sensory substitution devices meet the specific needs of users and address the qualitative characteristics of the substitution experience. Usability studies should adopt a personalised approach to understanding the requirements of visually impaired individuals and consider the entire substitution system, including the device, the user's characteristics, and the environment in which the device will be used. Such comprehensive analysis is essential for developing devices that truly address the practical needs of visually impaired individuals. Given that the first sensory substitution device was introduced nearly half a century ago, it is clear that more targeted efforts are needed to integrate these technologies into everyday life (Łukowska, Kałwak, Osiński, Janik, & Wierzchoń, 2022).

The importance of describing the subjective individual experience of substitution and the possible individual differences in this experience is less obvious but seems equally crucial for the development of this research area. Research should not only attempt to characterise the phenomenal properties of the experience of substitution but also address the relevance of this experience for environmental interaction and everyday functioning. It would be useful to explore not only how people experience substitution but also to what extent the acquisition of new phenomenal characteristics is relevant to them. For example, Neil Harbisson (2019), a super-user of sensory substitution, reports that he not only utilises sensory substitution in his daily life but also experiences its phenomenal properties in dreams (i.e., when the sensory substitution device does not affect his sensory experience). This suggests that sensory substitution can become an inherent characteristic of our sensory experience. If so, the visually impaired would certainly be more likely to use substitution devices actively.

References

Auvray, M., & Myin, E. (2009). Perception with compensatory devices: from sensory substitution to sensorimotor extension. *Cognitive Science*, *33*(6), 1036–1058. DOI: 10.1111/j.1551-6709.2009.01040.x.

Bach-y-Rita, P., & Kercel, S.W. (2003). Sensory substitution and the human-machine interface. *Trends in Cognitive Sciences*, *7*(12), 541–546.

Block, N. (2003). Tactile sensation via spatial perception. *Trends in Cognitive Sciences*, *7*(7), 285–286.

Farmer, L.W. (1978). Mobility devices. *Bulletin of Prosthetic Research*, 47–118.

Froese, T., McGann, M., Bigge, W., Spiers, A., & Seth, A.K. (2012). The Enactive Torch: A New Tool for the Science of Perception. *IEEE Transactions on Haptics*, *5*(4), 365–375.

Harbisson, N. (2019). Hearing colors: My life experience as a cyborg. In: *Creativity, Imagination and Innovation. Perspectives and Inspirational Stories* (pp. 117–125).

Heimler, B., & Amedi, A. (2020). Are critical periods reversible in the adult brain? Insights on cortical specializations based on sensory deprivation studies. *Neuroscience & Biobehavioral Reviews*, *116*, 494–507.

Kałwak, W., Reuter, M., Łukowska, M., Majchrowicz, B., & Wierzchoń, M. (2018). Guidelines for quantitative and qualitative studies of sensory substitution experience. *Adaptive Behaviour*, *26*(3), 111–127. DOI: 10.1177/1059712318771690.

Łukowska, M., Kałwak, W., Osiński, D., Janik, J., & Wierzchoń, M. (2022). How to teach a visually impaired person to hear colours? Multidimensional training for a colour-to-sound sensory substitution device – design and evaluation. *International Journal of Human – Computer Studies*. DOI: 10.1016/j.ijhcs.2022.102925.

Łukowska, M., Osiński, D., Bizoń-Angov, P., Ciupińska, K., Hat, K., Gwiaździński, P, del Pin, S.H., Kałwak, W., Korczyk, M., Reuter, M., Szwed, M., & Wierzchoń, M. (submitted). Better act than see: individual differences in sensorimotor contingencies and (meta)cognitive strategies while using a colour-to-sound sensory substitution device. *Adaptive Behavior*.

Meijer, P.B. (1992). An experimental system for auditory image representations. *IEEE transactions on biomedical engineering*, *39*(2), 112–121.

Osiński, D., Łukowska, M., Bizoń-Angov, P., Hjelme, D.R., & Wierzchoń, M. (2021). Colorophone 2.0: A wearable color sonification device

generating live stereo-soundscapes – design, implementation and usability audit. *Sensors*, *21*, 7351. DOI: 10.3390/s21217351.

Proulx, M.J. (2010). Synthetic synaesthesia and sensory substitution. *Consciousness and cognition*, *19*(1), 501–503.

Poirier, C., De Volder, A.G., & Scheiber, C. (2007). What neuroimaging tells us about sensory substitution.*Neuroscience & Biobehavioral Reviews*, *31*(7), 1064–1070. DOI: 10.1016/j.neubiorev.2007.05.010.

Ptito, M., Moesgaard, S.M., Gjedde, A., & Kupers, R. (2005). Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. *Brain: A Journal of Neurology*, *128*(Pt 3), 606–614. DOI: 10.1093/brain/awh380.

Reuter, M., Kałwak, W., Sznajder, M., & Wierzchoń, M. (under review). How to measure sensory substitution experience? Enactive Torch studies with qualitative methods. *Analiza i Egzystencja*.

Piotr Konderak
Maria Curie-Skłodowska University in Lublin
https://orcid.org/0000-0002-9161-4870
https://doi.org/10.35765/slowniki.470en

Cognitive semiotics: the turn towards meaning

Summary

DEFINITION OF THE TERM: Cognitive semiotics is the transdisciplinary study of meaning structures and meaning-making processes. It integrates concepts and methods from semiotics, cognitive science, and linguistics, with significant reference to phenomenology.

HISTORICAL ANALYSIS OF THE TERM: Cognitive semiotics has emerged from the collaboration of semioticians, linguists, cognitive scientists, and philosophers searching for answers to questions about the essence of meaning. The starting point included a semiotic perspective on cognition and a cognitive perspective on semiosis.

DISCUSSION OF THE TERM: Cognitive semiotics aims to develop a conceptual framework that provides a unified account of phenomena discussed in semiotics, linguistics, and cognitive science. The key concept of meaning is understood as the dynamic relationship between an embodied subject and the environment, shaped by the value that elements of the environment represent to that subject. This understanding forms the basis for the concept of multi-level meaning, or the Semiotic Hierarchy, which includes four levels: animation, consciousness, sign function, and language. Methodologically, cognitive semiotics employs pheno-methodological triangulation, advocating the use of first-, second-, and third-person methods in studying meaning structures and meaning-making processes, with an emphasis on the epistemological primacy of first-person methods. Addressing the relationship between cognitive semiotics and cognitive science involves referencing enactive and embodied concepts of cognition.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Cognitive semiotics provides a view of agency that extends beyond the concept of an information-processing agent. The proposed conceptual framework, including the Semiotic Hierarchy, enables a coherent account of the diverse everyday activities of cognitive subjects in terms of meaning-making processes. Despite the progress made in cognitive semiotics, it continues to face challenges in integrating methods from various approaches.

Keywords: cognitive semiotics, Semiotic Hierarchy, pheno-methodological triangulation, enactivism, embodiment

Definition of the term

Cognitive semiotics is a transdisciplinary study of meaning-making which integrates concepts and methods from semiotics, cognitive science, and linguistics, with significant contributions from phenomenology. This field emerged from the idea of integrating the humanities, social sciences, and natural sciences in an attempt to explain the phenomenon of meaning and meaning-making processes. Influential thinkers in philosophy (Gottlob Frege, Hilary Putnam, and Maurice Merleau-Ponty), semiotics (Charles Sanders Peirce and Ferdinand de Saussure), and linguistics (George Lakoff, Ronald Langacker, and Eleanor Rosch) provided the theoretical foundations for examining these processes and meaning structures. Cognitive semiotics also draws on the components of cognitive science research, namely psychology and neuroscience, in order to offer an empirical account of the subject's interaction with the world and other subjects. Meaning-making processes are understood as dynamic: they emerge in evolution and change over the course of ontogenetic development.

The term 'cognitive semiotics' might suggest that this research area is merely a sub-discipline of semiotics, traditionally understood as the theory of signs or the study of sign systems (Zegleń, 2000). However, cognitive semioticians adopt a broader perspective and define semiotics as the study of meaning structures and meaning-making processes. In this broader view, signs – classically defined within traditional semiotics – represent only one type of meaning structure. Semiosis, which in cognitive semiotics is understood as a meaning-making process, covers a wider array of phenomena beyond the mere creation, use, and interpretation of signs. This broader view is encapsulated in the conceptual framework known as the Semiotic Hierarchy, which describes various levels of meaning-making.

Similarly, the term 'cognitive' might imply that cognitive semiotics is simply another branch of cognitive science, akin to cognitive psychology or cognitive linguistics. However, the relationship between these fields is more nuanced. Cognitive semiotics emerged from dissatisfaction with tendencies in classical cognitive science to reduce cognitive subjects to their functional dimension. It is considered inadequate to treat cognitive subjects merely as 'agents', i.e., systems that perform functions in a specific

environment in response to input data based on their internal functional organisation. In other words, cognitive semioticians approach cognitive processes not only in terms of information processing but also in terms of meaning-making processes and the meaning structures that emerge as a result. Thus, while cognitive semiotics incorporates some concepts and methods from standard cognitive science, it relies more heavily on 4E approaches (enactive, embodied, embedded and extended), particularly those emphasising enactive and embodied cognition.

Given its transdisciplinary nature, cognitive semiotics should not be considered as a research discipline. The issue of meaning-making transcends the boundaries of individual disciplines and research areas, extending beyond academic activities to include insights from practitioners. Both academic and non-academic studies in this field share a common goal: to explore meaning-making processes.

Historical analysis of the term

Cognitive semiotics has emerged from the collaboration of semioticians, linguists, cognitive scientists, and philosophers involved in the research program at the Centre for Cognitive Semiotics at Lund University. A significant milestone in this development was the 2011 7th Conference of the Nordic Association for Semiotic Studies, titled *Towards Cognitive Semiotics: A Semiotic Perspective on Cognition – A Cognitive Perspective on Semiosis*. The concept of integrating semiotic, linguistic, and cognitive studies was discussed at this conference and has historical roots in the work of various semioticians. One prominent figure, Göran Sonesson, in his monograph *Pictorial Concepts* (1989), adopts a semiotic-cognitivist perspective to construct a conceptual framework for analysing iconicity. Sonesson draws on Husserl's writings and Piaget's research to formulate a conception of the sign, in which a key element is "a double asymmetry between the two parts, because one part, expression, is more directly experienced than the other; and because the other part, content, is more in focus than the other" (Sonesson, 2010, pp. 24–25). Thomas Daddesio, often regarded as the informal spiritual father of cognitive semiotics, proposes a "cognitive theory of symbols" (Daddesio, 1995, p. 2) and advocates for a "cognitive approach to semiosis" (Daddesio,

1995, p. 10). In the Polish research tradition, Jerzy Pelc (2013) argues for closer integration between research on cognition and communication and research within semiotics. Umberto Eco (1999) does not object to semiotics being linked with the cognitive sciences, although he does not state whether semiotics is a branch of cognitive science or vice versa.

The historical origins of cognitive semiotics can be traced back to three main research areas: semiotics, cognitive science, and phenomenology. Understanding meaning-making processes necessitates the concept of the subject. Instead of considering 'functionally equivalent' agents, i.e., agents performing the same function regardless of their physical constitution, cognitive semiotics introduces the notion of an embodied subject who interacts with an environment, both physical and social, that holds meaning for him. A pivotal publication in this context is *The Embodied Mind: Cognitive Science and Human Experience* by Francisco Varela, Evan Thompson, and Eleanor Rosch. This work inspires phenomenological perspectives and introduces enactivist and embodied frameworks for thinking about cognition. The dynamic nature of meaning-making processes is further inspired through evolutionary and developmental approaches. Merlin Donald's *Origins of the Modern Mind* (1991) and Terrence Deacon's *The Symbolic Species: The Co-evolution of Language and the Brain* (1997) have been instrumental in shaping ideas on the evolution of meaning-making. From a developmental perspective, specifically the stages through which intersubjectivity emerges in children, cognitive semioticians have been significantly influenced by the research of Lev Vygotsky, Jean Piaget, and Michael Tomasello (*Origins of Human Communication*, 2008), as well as studies by Colwyn Trevarthen, and Ulrike Lüdtke.

Cognitive semiotics has since become an institutionalised field. In 2013, the International Association for Cognitive Semiotics (IACS) was established, bringing together researchers focused on meaning-making. IACS has organised several conferences: Lund (2014), Lublin (UMCS, 2016), Toronto (2018), Aachen (2022), with the fifth edition scheduled for 2024 at Lund University. Since 2007, the journal *Cognitive Semiotics*, published by de Gruyter, has provided a platform for scholarly contributions. Additionally, cognitive semiotics is increasingly being integrated into education, either as a dedicated field of study or as a course within degree programmes.

Discussion of the term

Research in cognitive semiotics begins with an ambitious goal: to integrate findings from diverse disciplines into a coherent framework. Like any inter-, multi- or trans-disciplinary approach, cognitive semiotics faces the challenge of integrating research results from significantly different disciplines, particularly due to the differing conceptual grids and methodologies across fields. A major difficulty lies in addressing the concept of meaning. Therefore, cognitive semiotics should aim to establish a sufficiently general conceptual framework that provides a common ground for solutions regrading meaning.

The multilevel nature of meaning: the Semiotic Hierarchy

The study of cognitive semiotics is rooted in the belief in the multidimensional nature of meaning, as reflected in the term's inherent ambiguity. In an attempt to construct a unified framework for studying meaning and meaning-making, Zlatev (2009) and later Zlatev and Konderak (2023) proposed the Semiotic Hierarchy. This conceptual framework captures various types of meaning and processes of meaning-making: biological meaning, phenomenal meaning, cultural (though not linguistic) meaning, and linguistic meaning. The starting point involves defining meaning as the dynamic relationships between a subject and the environment in which the subject is embedded, determined by the value that elements of the environment represent to the subject. In other words, meaning-making is a relational phenomenon consisting of value-based interactions between subjects and their 'world'. This approach leads to an understanding of meaning in terms of layers or levels, where successive layers – animation, consciousness, sign function, and language – emerge to form a hierarchy of meaning-making. The relationship between these layers is dynamic and bidirectional. On the one hand, each successive level requires and is founded on the lower level: the activity of a living organism is necessary for the emergence of consciousness; consciousness is required for the use of signs, and semiotic capacities form the basis of linguistic meaning-making. On the other hand, each higher level

transcends ('sublimes') the level on which it is founded. This relationship model is based on Merleau-Ponty's (1962) concept of *Fundierung*, which describes a two-way relationship that emphasises the primacy of the originator over the originated in the sense that the originated is a specific or distinct form of the originator. Simultaneously, the originator manifests itself in the originated (Merleau-Ponty's, 1962, p. 458).

The lowest layer in the Semiotic Hierarchy, located above the "lower semiotic threshold" (cf. e.g., Nöth, 2001), emerges in the activity of living organisms. In other words, only animate beings are capable of intrinsic meaning. Semiosis occurring above this threshold is the focus of biosemiotics (Jakob von Uexküll, Kalevi Kull). An embodied subject, such as a living organism, operates in an environment known as the Umwelt, where the value that gives meaning to environmental elements is tied to the organism's survival.

Life is the condition for the emergence of the next layer: the level of experience or agency. Subjects at this level can consciously experience the world, detecting conditions favourable or unfavourable to survival and taking appropriate actions. The experiencing subject is intentionally directed toward experienced objects. At this level, phenomenal meaning emerges, and the subject becomes embedded in Husserl's *Lebenswelt*, i.e., the life-world. Thus, what is meaningful at this level is directly experienced.

Phenomenal consciousness is a necessary, though not sufficient, condition for the third level of meaning-making, where the sign function emerges. This function emerges in the processes of using and creating intersubjectively accessible signs. According to concepts developed in cognitive semiotics, a sign emerges when a subject distinguishes between the sign itself, the form it takes (e.g., a pointing gesture, an image, or a symbol), and what the sign represents. At this third level of meaning-making, the subject is able to link an expression and its content based on grounding relations: selected cultural phenomena from the

Lebenswelt become meaningful through relations of proximity (indices), similarity (icons), or convention (symbols).

At the fourth level, subjects aim to use symbols in a correct and systematic way, adhering to the accepted norms of their linguistic community. The environment in which subjects are embedded constitutes 'the universe of discourse', encompassing everyday conversations,

novels, poems, myths, scientific theories, and other spoken or written 'texts'. Language at this level requires not only a set of symbols but also a grammar, understood as a system of semiotic norms.

Subjects capable of meaning-making at higher levels can also engage in meaning-making at lower levels. For instance, a subject evaluating scientific publications on the climate crisis can understand the language of a text (level IV of the Semiotic Hierarchy), interpret thermometer readings as signs of atmospheric conditions (level III), consciously experience the world (high temperatures might feel unpleasant, level II), and regulate their behaviour to stay alive (seeking shelter from the sun, level I). Each level involves distinct meanings and corresponding meaning-making processes: temperature readings are meaningful for an organism's survival, and the organism assigns meaning to this environmental factor by regulating its behaviour. The conscious experience of temperature constitutes meaning at the second level, and consciousness allows subjects to give meaning to this experience. At the third level, a thermometer reading is a sign, and interpreting this sign in the categories of other signs constitutes the meaning-making process.

The Semiotic Hierarchy provides a general framework for discussing meaning-making processes and meaning structures. In research practice, the study of these processes often focuses on specific levels (e.g., the conscious experience of spontaneous gestures, iconic and pointing gestures, the relationship between sign language and pantomime) or the relationships between levels (e.g., whether and how sign language emerges from pantomimic activity). The Semiotic Hierarchy is subject to change. For example, the latest approach to this hierarchy (Zlatev & Konderak, 2023) stratifies the second level into subjective and intersubjective levels of meaning-making, with the latter accounting for the phenomenon of shared intentionality.

The dynamics of meanings: structures versus processes

Originally influenced by enactivist and embodied approaches to meaning-making subjects, cognitive semioticians emphasised the dynamism of meaning-making processes. However, the emergence of relatively permanent meaning structures – especially at the levels of

sign function (e.g., the consolidation of 'OK' gestures or a raised thumb) and language (e.g., the formation of specific grammar structures) – necessitated a more 'dialectical' approach to processes and structures. This approach was inspired by Merleau-Ponty's concept of the world's structure as sedimentation, i.e., the ability to consolidate specific structures, and spontaneity, i.e., the ability to create new structures. Consequently, meaning possesses a dual nature: it can be characterised as both a dynamic process and a stable structure. On the one hand, situated embodied actions give rise to meaning structures; on the other hand, pre-existing structures constrain the actions taken. Thus, at each level of the Semiotic Hierarchy, there are two aspects of meaning: relatively stable structures (norms), and processes that develop over time and can lead to these norms changing. Research within cognitive semiotics may adopt either a processual or structural perspective or seek to establish rules for the coexistence of processes and structures in real-world interactions with the environment. The dynamic interplay of sedimentation and spontaneity is evident at every level of meaning.

As with the levels of meaning, the distinction between meaning structures and meaning-making processes provides a general framework for discussing semiosis. These concepts are useful for describing both the formation of linguistic norms and the emergence of behavioural patterns. They are applicable in both ontogenetic descriptions and evolutionary approaches. Research within cognitive semiotics encompasses both the characterisation of processes (sometimes excluding the structural dimension) and the characterisation of emergent structures.

Pheno-methodological triangulation

Cognitive semiotics, in addressing the issue of meaning, draws on the conceptual frameworks and methodologies of three research areas: semiotics, cognitive linguistics, and cognitive science. The diversity of methods used within these fields poses significant methodological challenges for cognitive semiotics.

The primary methodological approach adopted in cognitive semiotics is pheno-methodological triangulation, also known as phenomenological triangulation (previously known as methodological

triangulation). This approach involves integral description of the phenomenon of meaning from three perspectives: first-person, second-person, and third-person. Unlike standard cognitive science, cognitive semiotics recognises the epistemological priority of the first-person approach, asserting – in line with the phenomenological perspective – that all knowledge originates from experience. The concept of triangulation has two dimensions: ontological and epistemological. The ontological dimension pertains to the nature of the phenomena under study, whether they involve first-person data (e.g., in studies on agency) or third-person data (e.g., in neuroscience). The epistemological dimension involves capturing the phenomenon from three perspectives: the involved researcher (the first-person perspective); interactions with others (the second-person perspective); observation by an unbiased, uninvolved observer (the third-person perspective). Examples of first-person methods include phenomenological analyses and systematic intuitions. Second-person methods encompass participant observation, interviews, and intersubjective validations. Third-person methods include experiments and statistical analyses. Pheno-methodological triangulation advocates integrating these three perspectives as much as the nature of the phenomenon under study allows.

This postulate supports the realisation of the second methodological component of cognitive semiotics, i.e., the conceptual-empirical loop. Research often begins with analyses of concepts characteristic of semiotics or cognitive linguistics. The idea of the conceptual-empirical loop reflects the interdependence between the conceptualisation of a phenomenon (e.g., what is spontaneous gesticulation?) and its empirically-studied manifestation (e.g., how does spontaneous gesticulation manifest in the behaviour of lecturers?), thus generating subsequent research steps. Conceptual distinctions are reflected in empirical research, which can, in turn, lead to further reconceptualisation.

Cognition versus meaning-making: 4E cognitive science

Defining the reciprocal relationship between cognitive semiotics and cognitive science is challenging due to historical and theoretical differences between the two approaches. This is further complicated by

ongoing debates about the nature of cognitive science itself. Cognitive semiotics has emerged in clear opposition to the computational, representationalist approach of traditional cognitivism, which has been criticised for its functionalist assumptions and contrasted with phenomenological approaches. Cognitive semiotics has developed alongside the development of alternative, i.e., 4E approaches to cognition. Discussions on the understanding of meaning-making processes and meaning structures have been driven by the idea of cognition as embodied interactions with the environment and the idea of cognitive processes that extend beyond the boundaries of the brain and body.

In 1991, the publication of *The Embodied Mind: Cognitive Science and Human Experience* by Varela, Thompson, and Rosch introduced two key concepts of 4E approaches: enactivism and embodiment. Later, in his 2007 book *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Thompson systematically presented the principles of this "new cognitive science". Cognitive semiotics engages with these approaches, particularly in its recognition that only living beings can engage in genuine meaning-making processes. The lowest level of the Semiotic Hierarchy corresponds to the minimal requirements formulated within enactivism. In both these approaches, meaning-making processes are constituted by environmental and corporeal factors. The dynamics of semiosis can only be explained by referring to experiencing, embodied subjects that enact their own domain of meaningful phenomena. Thus, enactivism and embodiment form the cognitive foundation for the first two levels of the Semiotic Hierarchy: animation and consciousness.

The meaning-making processes in the Semiotic Hierarchy are based on a dynamic interaction between the subject and the environment. This interaction highlights three key phenomena: the dynamic nature of meaning-making processes, the role of agency in these processes, and the embodied nature of subject-environment interactions. As Thompson (2007) observes, living beings actively generate and sustain their existence. As they explore their surroundings, their actions shape what is meaningful to them. Meaning-making subjects are not passive agents waiting for information and processing it later; rather, in the spirit of James Gibson's psychology, they actively seek out information while exploring their environment. Consequently, the understanding of the meaning-making subject diverges from the agent in traditional

cognitivism. The role of the subject is not limited to operations on amodal symbols performed according to a set of rules. Instead, the subject is treated as a purposefully acting autonomous system with a sense of agency.

The distinction between quasi-subjects and proper subjects illustrates different aspects of agency at various levels. Both make meanings but do so in different ways. Quasi-subjects are living beings with quasi-experience who exhibit biologically determined, goal-directed actions. The behaviour of quasi-subjects can be explained at the first level of the Semiotic Hierarchy.

In contrast, only proper subjects are capable of subjectively experiencing the world, and their activities are explained at the second level of the Semiotic Hierarchy. The functional dimension, although necessary, is not sufficient to understand the subject at this level. The subject's experiences enact the meaningful elements of the environment. From Merleau-Ponty's perspective, the subject's experience is embodied: the subject's body is the condition of experience and is in a two-way relationship with the environment. On the one hand, the environment continuously influences and regulates our bodies; on the other hand, our bodies continuously influence the environment. As a result, a pattern of dynamic, embodied interaction emerges.

The roles of the other two components – extended and embedded cognition – in explaining meaning-making processes are far less clear. On the one hand, the extended approach can be reconciled with enactivism and the embodied approach. The claim that meaning-making processes cannot be confined to the head or even to the body seems to align with the idea of the relational nature of meaning-making processes. On the other hand, the functionalist foundations of Clark and Chalmers' (1998) conception lead cognitive semioticians to seek an alternative conception of extended cognition. Another challenge associated with extended and embedded cognition is the notion of an acting subject. Extended meaning-making systems, which emerge from the coupling of a subject and elements of the environment, are often seen as too transient to underpin stable agency. The ease with which these meaning-making systems can couple and uncouple complicates efforts to conceptualise them as subjects within the framework of cognitive semiotics. From this perspective, the extended approach fails to fully capture the significance

of embodiment. Treating the body as merely a tool does not capture the role of the body in meaning-making processes, especially in establishing what is meaningful while experiencing the world, i.e., in the transition from subjectivity to intersubjectivity.

Systematic reflection with conclusions and recommendations

Cognitive semiotics appreciates the multiplicity and diversity of our interactions with the world, through which our actions are not only purposeful and effective but also meaningful. Migrating in response to climate change, navigating environments in search of friends, tracking a familiar figure with one's gaze, using gestures to communicate, speaking a broken foreign language, dancing spontaneously at a concert, performing ballet on stage, and preparing both tasty and elegant dishes – these just a few everyday activities that can all be examined through the lens of cognitive activities. Many, if not all, of these activities can be described as purposeful behaviours guided by incoming information and governed by various rules. However, each of them also represents the activity of an embodied, sentient, and acting subject interacting with both physical and social environments. In striving to present a comprehensive understanding of living beings, cognitive semioticians seek to extend the notion of an information-processing agent by recognising subjective experience and embodied interaction as indispensable elements in explaining human actions in the world.

A number of scholars could challenge the above statement, arguing that the agent from standard cognitive science should be r e p l a c e d by a meaning-making subject. In this vein, there has been much discussion within 4E approaches calling for the replacement of representational and information-processing approaches with enactive, embodied, and extended approaches. While these 4E approaches initially received enthusiastic – albeit somewhat uncritical – support, there is now a growing recognition among cognitive semioticians of their significant limitations. This growing recognition has led to increasing acknowledgement of the valuable contributions of traditional cognitive science. Thus, cognitive semiotics seeks a 'new science of mind', i.e., a multifaceted

description of a subject who is both acting and feeling. One promising approach that transcends the limitations of functionally understood cognitive science while preserving its achievements is the concept of the Amalgamated Mind, developed by Rowlands (2010; cf. also Konderak, 2018).

The theoretical frameworks adopted by researchers are valuable as they expand our understanding of a wide range of phenomena. To what extent can the Semiotic Hierarchy or pheno-methodological triangulation meet researchers' expectations? The increasing number of applications of the Semiotic Hierarchy – such as analyses of the evolution of language, the development of communication in children, the polysemiotic communication, and narrativity – demonstrates the potential of this framework. However, the still-ambivalent attitude of some researchers towards standard cognitive science, their scepticism about the psychological idea of internal representation, and their questioning of neurocentrism in cognitive science underscore the challenges in applying the principles of pheno-methodological triangulation. The limitations of the neurophenomenology programme are evidence of these difficulties.

References

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*, 10–23.

Daddesio, T. (1995). *On Minds and Symbols: The Relevance of Cognitive Science for Semiotics*. Berlin: De Gruyter Mouton.

Eco, U. (1999). *Semiotics in the Next Millennium*. http://www.umbertoeco.it/CV/Semiotics%20in%20the%20next%20millennium.pdf. Accessed 16.08.2024.

Konderak, P. (2018). *Mind, Cognition, Semiosis: Ways to Cognitive Semiotics*. Lublin: UMCS Press.

Merleau-Ponty, M. (1962). *Phenomenology of Perception*. London: Routledge & Kegan Paul.

Nowakowski, P., & Komendziński, T. (2014). Cognition as shaking hands with the world. Introduction. *Avant*, *V*(2), 10–16.

Nöth, W. (2001). Protosemiotics and physicosemiotics. *Sign System Studies*, *29*(1), 13–25.

Pelc, J. (2013). Dwa pojęcia komunikacji. *Studia Semiotyczne*, *XXVIII–XXIX*, 107–119.

Rowlands, M. (2010). *The New Science of the Mind*. Cambridge, MA: The MIT Press.
Sonesson, G. (1989). *Pictorial Concepts: Inquiries into the Semiotic Heritage and Its Relevance to the Interpretation of the Visual World*. Lund: Lund University Press.
Thompson, E. (2007). *Mind in Life: Biology, Phenomenology and the Sciences of Mind*. London: Belknap Press.
Varela, F.J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge: The MIT Press.
Zlatev, J. (2009). The semiotic hierarchy: life, consciousness, signs and language. *Cognitive Semiotics*, *4*, 169–200.
Zlatev, J., & Konderak, P. (2023). Consciousness and Semiosis. In: J. Pelkey (ed.), *Bloomsbury Semiotics Volume 1: History and Semiosis*. London: Bloomsbury Academic.
Żegleń, U. (2000). *Wprowadzenie do semiotyki teoretycznej i semiotyki kultury*. Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.

Łukasz Wiraszka
Jagiellonian University in Krakow
https://orcid.org/0000-0002-3653-5926
https://doi.org/10.35765/slowniki.448en

Cognitive linguistics

Summary

DEFINITION OF THE TERM: Cognitive linguistics is a strand in modern linguistics that encompasses a number of complementary theories and methodologies based on common assumptions about the nature of language. The most fundamental of these assumptions is that language is an integral part of human cognition.

HISTORICAL ANALYSIS OF THE TERM: The beginnings of broadly understood cognitive linguistics date back to the mid-20th century and are linked to the cognitive revolution. Noam Chomsky's generative theory of language concerned the linguistic ability inherent in the human mind; however, unlike modern cognitive linguistics, he viewed it as a mechanism that generates grammatical sentences in a given language. The pioneers of modern cognitive linguistics were Chomsky's former students, who – noticing the limitations of the generative approach to language – laid the theoretical foundations of 'second-generation' cognitive linguistics at the turn of the 1980s.

DISCUSSION OF THE TERM: The key postulates of cognitive linguistics include the alignment of the theory of language with current knowledge of the workings of the human mind (which is called cognitive commitment); the primacy of meaning in linguistic inquiry; the extension of the scope of research to all forms of language use, including metaphorical language; and acknowledgement of the role of embodiment. Cognitive linguistics treats metaphor as a cognitive phenomenon (a way of thinking) and adopts a subjectivist view of linguistic meaning.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: By studying linguistic phenomena, cognitive linguists formulate claims about mental processes. The full realisation of this task requires interdisciplinary research conducted within cognitive science.

Keywords: cognitive linguistics, cognitive science, cognitive grammar

Definition of the term

Cognitive linguistics is a strand in modern linguistics that encompasses a number of distinct but complementary theories and methodologies based on shared fundamental assumptions about both the nature of language and ways of doing linguistics. The beginnings of modern cognitive linguistics date back the turn of the 1980s, when its first key postulates were formulated in reaction to some of the claims of Chomsky's generative school, which was then dominant in American linguistics, and in the wake of new research developments in cognitive psychology. These postulates laid the foundations for the main theoretical concepts developed in the 1980s by the 'founding fathers' of cognitive linguistics, the most prominent of whom were G. Lakoff, R. Langacker, and L. Talmy.

Along with other disciplines, such as cognitive psychology, philosophy, anthropology, neuroscience, and artificial intelligence research, cognitive linguistics contributes to cognitive science, which – as an interdisciplinary research programme – mainly focuses on the mechanisms of human cognition and thinking. Cognitive linguistics is part of the cognitive 'enterprise' because of its treatment of language as a kind of knowledge that resides in the minds of its users. Although the basic aims of cognitive linguistics do not fundamentally differ from the aims pursued by linguists outside cognitivism (i.e., the description of linguistic phenomena and the formulation of theories to account for them), cognitive linguists look for answers to questions about the nature of linguistic knowledge "in what they consider to be the source and at the same time *raison d'être* of language: in the capacities of the human mind" (Tabakowska, 1995, p. 5).

This approach stems directly from one of the fundamental assumptions of cognitive linguistics, according to which "language is an integral part of human cognition" (Langacker, 1987, p. 12). An important element of this assumption is the emphasis on the 'integration' of linguistic ability with other aspects of the workings of the human mind. The knowledge that makes up linguistic ability is not an autonomous conceptual component (module) in the human mind: it is based on the same general mental processes that underlie other types of knowledge and it closely interacts with other cognitive abilities, such as categorisation, memory, perception, reasoning, and attention. In cognitive linguistics, it is assumed that

language not only uses but also reflects the mechanisms of human cognition, so the study of linguistic phenomena is ultimately reduced to the study of the mental mechanisms that make up the human cognitive apparatus (Tabakowska, 1995, p. 12).

The conceptual and methodological framework of modern cognitive linguistics is marked by its two fundamental commitments, articulated by Lakoff (1990) in the first issue of the journal *Cognitive Linguistics*: the cognitive commitment and the generalisation commitment. The cognitive commitment requires that any claims related to linguistic phenomena accord with the findings of other cognitive sciences regarding the workings of the human mind. For example, the structure of conceptual categories attributed to linguistic expressions should take into account the prototypical effects demonstrated in psychological experiments, while postulating linguistic categories based solely on the classical Aristotelian model of categorisation (in which category membership is defined by a set of necessary and sufficient features) would require special empirical justification (Lakoff, 1990, p. 41). This understanding of the cognitive commitment is thus intended to maximise the psychological plausibility of linguistic theories.

The generalization commitment is a commitment to searching for high-level generalisations. This is not about seeking the maximum economy of description, which is typical of formal approaches, but rather about identifying shared mechanisms and principles underlying different levels of language structure and various aspects of language use. If language is an ability based on the general properties of the human mind, the same applies to all its elements. This means that, for example, syntactic phenomena can – and should – be largely explained by the same mechanisms and conceptual constructs as semantic phenomena, and that language structure is divided into different levels based on solely practical factors (Tabakowska, 1995, p. 12).

Recognizing language as an integral part of human cognition naturally leads to emphasising its cognitive function. From this perspective, language is not merely a tool for interpersonal communication: being essentially a structured inventory of conceptual categories through which we perceive reality and understand our experiences, language constitutes a cognitive ability that is involved in processing and organising information about the world (Geeraerts & Cuyckens, 2007, p. 5).

If we view language as a repository of world knowledge, then semantic motivations should be considered in all linguistic inquiries.

The key theoretical frameworks of cognitive linguistics include R. Langacker's cognitive grammar, G. Lakoff and M. Johnson's conceptual metaphor theory, G. Fauconnier and M. Turner's mental spaces and conceptual blending theory, and several variants of construction grammar developed by, among others, Ch. Fillmore, W. Croft, and A. Goldberg.

Historical analysis of the term

The cognitivist trend in linguistics was initiated in the mid-1950s by the American linguist N. Chomsky, who – together with several pioneers in artificial intelligence research – participated in the first cognitive revolution (also called the cognitive turn). This revolution rejected behaviourism (see, e.g., Chomsky, 1959) and directed research towards the functioning of the human mind with the intention of describing the processes occurring in it in such a way that they could be simulated by computers (cf. Bremer, 2023, pp. 71–72). The revolution in linguistics initiated by Chomsky involved a move away from a purely descriptive approach to language seen as an abstract, static system (i.e., as a set of linguistic structures and relationships between them), which was characteristic of the American school of structuralism (descriptivism). Instead, Chomsky and his followers started putting forward claims that actually amounted to a theory of the human mind or, more specifically, the part of the mind responsible for language. For this reason, the generative current initiated by Chomsky can be classified as broadly understood cognitive linguistics (Taylor, 2002, pp. 6–8).

One of the assumptions underpinning generativism was the autonomous nature of human linguistic ability. According to this approach, language constitutes a distinct module in the human mind which is governed by its own laws, and linguistic knowledge is a distinct, special type of knowledge acquired through the operation of a unique, largely innate mechanism of language acquisition (Chomsky, 1969). By adopting the assumption of the autonomy of language, Chomsky's linguistics became independent from other scientific disciplines concerned with the human

mind, which allowed him to postulate essentially arbitrary mechanisms for explaining linguistic phenomena.

In keeping with a popular tendency at the time to view the human mind as a binary calculating machine (Bremer, 2023, p. 72), Chomsky developed an 'operational' conception of language, which consisted in a set of grammatical rules resembling an algorithm that described actions performed on linguistic structures (Heinz, 1983, pp. 405–406). That kind of grammar was supposed to make it possible to 'generate' (hence the name of this theory) all correct and only correct (i.e., grammatical) sentences of a language. In order to ensure algorithmic precision, grammatical rules had to be strictly formalised, preferably along the lines of arithmetic rules, while individual elements of generative grammar took the form of recursively applied operations on (often semantically empty) symbols. Chomsky's theory was thus characterised by far-reaching formalism and a desire, borrowed from the hard sciences, for maximum economy of description, so that the greatest possible number of syntactic structures could be generated using the fewest possible rules.

The consequences of treating linguistic ability as a formalised mechanism that generates grammatical sentences were, first, the primacy of form over meaning (the first versions of generative grammar were completely anti-semantic); and second, emphasising the fundamental importance of the correctness of generated syntactic structures (Heinz, 1983, pp. 402, 407). For this reason, syntax (treated as an autonomous 'submodule' of language) played the most important role in Chomsky's generative theory, while the requirement of correctness led him to introduce a cardinal distinction between linguistic competence and performance, i.e., how people actually use language. Competence, which according to Chomsky was the proper subject matter of linguistics and was thus modelled by generative grammar, was defined as the linguistic ability of an imaginary ideal language user who produces only grammatical sentences. The everyday speech of real language users, which is full of all sorts of ungrammaticalities and phenomena that cannot be subjected to rules, was relegated to the domain of performance and, as such, placed outside the purview of linguistic investigations. Focusing on the criterion of correctness and adherence to the principle that theories must be constructed with mathematical elegance ultimately led to a situation in which successive versions of generative grammar were

able to describe only a small percentage of the structures that occur in actual linguistic communication (Tabakowska, 1995, p. 10).

That generativist detachment from authentic linguistic data was one of the manifestations – motivated by Cartesian dualism – of removing from linguistic considerations any factors that arise from the physical nature of human beings. As E. Tabakowska puts it, according to generativists,

> linguists investigate the linguistic competence of an ideal language user, who is free of any physical or mental imperfections, always obedient to the rules of grammar, guided by reason rather than emotion, consistent in his linguistic performance and resistant to the influence of the extra-linguistic context of language use; in a word, one who is devoid of the complex "human factor" that defines the individuality of each of us (Tabakowska, 1995, p. 8).

In Chomsky's theory, this 'human factor' was absent even from the component responsible for the semantic layer of language. The generative grammar from the period of Chomsky's pioneering 1957 treatise *Syntactic structures* did not include any elements responsible for the analysis of the meanings of generated syntactic constructions (which is illustrated by the famous sentence coined by Chomsky, 'Colourless green ideas sleep furiously', whose key property was its grammaticality). The next version of Chomsky's theory was called transformational-generative grammar, which was developed in the first half of the 1960s and led to the semantic turn in generativism (Heinz, 1983, p. 402). However, although the semantic component played an increasingly important role in each subsequent version of generativism, it was a semantics based on formal logic and truth conditions, and thereby referring to (a model of) the objective world, detached from human language users and their perception of reality (an objectivist theory of meaning, see Johnson, 1987, p. xxii ff. and Lakoff, 1987, ch. 11). In addition, the transformational-generative model of language permitted purely formal operations (i.e., transformations) based on generated syntactic structures, the application of which led to the construction of sentences with a different syntactic structure, which nevertheless had the same logical semantic structure as the initial sentence. This meant that a language user's choice of a particular syntactic structure did not affect the meaning of linguistic utterances, as long as they referred to the same state of affairs in objective reality.

What Chomsky and his followers found most challenging was metaphor. No matter how complex and formalised the operations postulated within the framework of their theory were, they were unable to generate metaphorical expressions, which are ubiquitous in linguistic utterances. This task was further hampered by the conceptual apparatus of formal logic adopted for the semantic module, which was based on truth values (i.e., truth or falsity). Thus, for the generativists, figurative use of language constituted "an obvious case of violation of grammatical rules and, within the framework of the model being created, it could only be treated as an atypical and marginal phenomenon: a chimerical ornament hooked at the periphery of the 'normal' fabric of language" (Tabakowska, 1995, p. 9). For this reason, metaphor was usually overlooked and treated as uninteresting from the perspective of the tasks of linguistic theory. However, generative theory reached such a level of abstraction that it posited operations and objects that could not be found in any linguistic material (Taylor, 2002, p. 7). Postulating 'invisible' elements, such as traces or empty categories, obviously aimed to increase the range of correct syntactic structures of a language that could be generated. In this way, transformational-generative grammar somehow imposed on general theory of mind mental processes and objects that were unique to language and had no direct confirmation in either actual linguistic data or in the findings of other cognitive sciences.

Chomsky's generative school, viewed as an early form of broadly understood cognitive linguistics, was characterised by modularity, formalism, the primacy of syntax, and a kind of rationalism manifested in researchers' focus on an idealised competence without taking into account human language users and the linguistic data they produce. These tendencies and the resulting constraints led several of Chomsky's students and colleagues, including R. Langacker and G. Lakoff, to reject the generative approach and develop their own linguistic theories. As a result, the cognitive grammar created by Langacker and the theory of conceptual metaphor proposed by Lakoff and his collaborator, the philosopher M. Johnson, are among the key concepts of modern cognitive linguistics and part of 'second-generation' cognitive science (Bremer, 2023, pp. 72–73).

Discussion of the term

The key postulates underlying the theoretical frameworks developed by Lakoff, Langacker, and other pioneers of today's cognitive linguistics included the rejection of both formalism and the dogma of language autonomy, the shift of focus from syntax to semantics, the inclusion of figurative language in linguistic research, and attention to the role of embodiment in the development of the cognitive apparatus and, consequently, of human linguistic ability.

Abandoning the principle of modularity and the assumption that language is independent of other cognitive abilities led to the acceptance of the two aforementioned cognitive commitments (first, to make sure that any claims related to linguistic phenomena accord with the findings of the other cognitive sciences regarding the workings of the human mind; second, to search for manifestations of general mental processes in linguistic structure and to explore the close links between language and other cognitive abilities). One of the cognitive abilities that is key for language is the process of categorisation, i.e., the construction of conceptual categories and the assignment of particular elements of our experience to these categories. The link between linguistic ability and the ability to categorise is quite obvious, as the words of a given language serve as access points to the conceptual structures associated with different categories. Moreover, the linguistic units identified by linguists are also conceptual categories subject to the same processes. For example, speakers of a given language categorise physically distinct sounds as examples of the same phoneme (cf. Taylor, 2002, pp. 9–10). In line with the commitment to cognitivism, in their linguistic inquiries into various aspects of language cognitive linguists use the prototype-based model of categorisation developed in cognitive psychology (e.g., Taylor, 2001).

Another example of breaking down the barriers between language and other cognitive abilities and identifying the effects of their workings in language structure can be found in how people divide the objects they perceive into figure and ground, which was demonstrated by Gestalt psychology. As a result of this process, some elements of the perceptual field attract the perceiving subject's attention much more than other elements, thereby constituting the perceptual figure, while other elements represent the perceptual ground. This cognitive mechanism seems to

operate in various aspects of the functioning of language, a fact that has been incorporated into the theories of R. Langacker and L. Talmy. In these theories, the phenomenon of figure and ground is conceptual: it does not refer to sensory perception but to the division of the conceptual contents symbolised in a linguistic utterance into a cognitively salient figure and an inconspicuous ground. Thus, in the spirit of the generalisation commitment, the figure/ground opposition finds application in analyses of both semantic and grammatical phenomena. For example, the semantic structure of the noun *arm* includes a reference to the entire human body, which constitutes the conceptual ground for the part of the body that is denoted by this noun, and which therefore constitutes the conceptual figure. Of the two nouns in the expression *the clouds over the city*, the word *clouds* is more conceptually prominent as the language user focuses on this word, or more precisely on the object it symbolises. Thus, this object functions as the figure, as opposed to *the city*, which in this example functions as the ground. However, the opposite figure/ground alignment is in place in the case of the phrase *the city under the clouds*, where the noun *city* is the more prominent element (figure). Let us note here that, objectively speaking, both expressions refer to the same situation. In a similar way, the grammatical subject of a sentence is the figure, while the complement represents the ground, which is why the sentences *the police have apprehended the car thief* and *the car thief has been apprehended by the police* (which objectively refer to the same situation) reflect different figure/ground configurations, so that their semantic interpretations are also different.

The rejection of the ideal of mathematical-logical formalism, which accompanied the development of modern cognitive linguistics, stemmed, on the one hand, from growing scepticism in cognitive science about the validity of understanding the human mind in terms of a Turing machine that operates on symbols; on the other hand, this rejection stemmed from the conviction that logical formulas typical of the exact sciences are incompatible with natural language because it “escapes the strict rules of logic and is generally inconsistent in their treatment” (Tabakowska, 1995, p. 11). Moreover, the imperative that guides cognitive linguists to include all linguistic phenomena within the scope of their inquiry, even those that do not follow rigid rules, justifies the abandonment of the desire for maximum economy of description and limits the importance

of the requirement for the elegance of theory. This is exemplified by the fact that cognitive grammar allows and even predicts cases in which the language user knows general grammatical rules (e.g., the rule for the formation of plural forms of nouns) and at the same time stores in their memory specific linguistic forms that result from the application of these rules (as is the case with frequently used plural forms, such as *eyes* or *toes*). When the grammar of a language includes both the description of a general rule and the specific results of its application, this is – due to the obvious redundancy of information – undesirable from the perspective of the elegance of theory, and the typical exclusion of such cases in formal approaches is what Langacker calls the "the rule/list fallacy" (Langacker, 1987, p. 29; Taylor, 2002, p. 14). This example shows that economy of description gives way to psychological plausibility in cognitive linguistics.

Langacker programmatically inscribed the opposition to formalism into his theory of language in the form of the content requirement (Langacker, 1987, p. 53; 2008, p. 24). This requirement imposes important constraints on the type of linguistic units admissible in the cognitive model of language: they can only be semantic and phonological units directly occurring in users' utterances, their associations in the form of symbolic units, and units that emerge from all three as a result of mental processes confirmed in psychological research, such as categorisation and schematisation.

However, it should be observed that, in contrast to the generative approach, cognitive linguistics rejects the possibility of introducing semantically empty symbols into a theory of language. As Langacker emphasises:

> Meaning is what language is all about; the analyst who ignores it to concentrate solely on matters of form severely impoverishes the natural and necessary subject matter of the discipline and ultimately distorts the character of the phenomena described (Langacker, 1987, p. 12).

Thus, whereas generative linguistics is characterised by formalism linked with the primacy of syntax, the main feature of cognitive linguistics is semanticism, which also manifests itself in the assumption that any, even slight, change in the linguistic form of an utterance entails a change in its meaning.

The approach of modern cognitive linguistics to linguistic meaning is characterised by subjectivism. According to this view, the meaning of linguistic utterances cannot be captured in terms of truth and falsity, nor can it be expressed using the formalised language of logic because meaning is not a reflection of the state of affairs in the real world but a reflection of the way in which this state of affairs is perceived by the human language user (Tabakowska, 1995, p. 17). More specifically, linguistic meaning is equated with conceptualisation, explicated as cognitive processing (Langacker, 1987, p. 5). In the case of the recipient of a linguistic message (i.e., the listener/reader), it is about the totality of cognitive processing triggered by someone's utterance or by the reading of a passage of text in a specific context; on the other hand, in the case of the author of an utterance (i.e., the speaker/writer), it is about the cognitive processing which leads to the use of a particular linguistic form (e.g., an expression, a sentence, etc.) for the purpose of linguistic communication. Importantly, conceptualisation is processual in nature: it is not merely about the activation of single static concepts that are part of the language user's knowledge; it is about a dynamic process of activating conceptual content in real time, which also covers the integration of this content and the creation of new concepts, as well as emotions and sensory experiences. Conceptualisation understood in this way is ultimately a function of specific patterns of neural activity and is likely to be imagistic rather than propositional in nature (Langacker, 2008, pp. 31–36).

Cognitive linguistics adopts an encyclopaedic conception of linguistic semantics (Langacker, 1987, pp. 154–166; Langacker, 2008, pp. 39–41). Encyclopaedic semantics is contrasted with the traditional conception of linguistic meaning, according to which semantic content resembles a dictionary definition. The dictionary view of linguistic semantics assumes that the meaning of a lexical unit (i.e., a word) is a limited piece of information, usually expressed in terms of a small set of features. It is based on the assumption that it is possible to precisely draw boundaries between purely linguistic meaning and knowledge about the world or contextual knowledge, which in traditional linguistics is the focus of pragmatics. However, a dictionary approach to semantics poses a number of problems in analyses of the meaning of linguistic expressions, if only because individual elements of the meaning of a word may turn out to

be more or less relevant , or even totally irrelevant, in different instances of its use and the meaning of complex lexical units cannot always be determined solely on the basis of the meanings of their constituent elements. Furthermore, defining the boundaries of linguistic (semantic) knowledge is an unfeasible task in practice, since every word is in fact a mental access point to a certain body of knowledge, and it is impossible to predict all the directions in which the activation of conceptual content might propagate in the conceptualisation associated with the use of a word; thus, this body of knowledge must be considered open-ended. In view of the above, the approach to the analyses of meaning adopted in cognitive linguistics is based on the assumption that conceptual content associated with lexical units constitutes a structured set of broad knowledge which is encyclopaedic in nature and ultimately crystallises in a specific context, and no sharp boundary can be drawn that separates strictly linguistic (semantic) content from extra-linguistic (pragmatic) content.

Blurring the boundaries between traditionally opposed branches of linguistics and their corresponding elements of language structure – in a fashion similar to the fuzzy boundaries in the prototype model of categorisation – is a fairly common practice in modern cognitive linguistics. In addition to the aforementioned blurring of the distinction between semantics and pragmatics, it is worth noting Langacker's seemingly controversial position regarding lexical units and the grammatical elements of language. In his view, the boundary between lexicon and syntax is fluid, unlike the clearly defined boundary between them in the generative model, where they have completely different statuses (a system of rules versus an unordered list of objects and their features) and belong to distinct components of language. More precisely, the words of a language and its grammatical elements form a continuum of symbolic units, each of which is a link between two poles: phonological and semantic (Langacker, 2008, p. 15).

The notion of continuum here refers to the level of detail of information concerning the pronunciation (in the case of the phonological pole) and the meaning (in the case of the semantic pole) of a symbolic unit. One end of this continuum is occupied by specific lexical units (e.g., the noun *book*, the verb *read*, or the preposition *on*), in which pronunciation and meaning are specified in detail; at the other end are grammatical

units and structures whose both meaning and pronunciation are specified in a schematic way (e.g., a part of speech such as a noun, a verb, or a preposition). On the one hand, their phonological pole contains very little information (it is schematic), but it cannot be said to be completely empty as it includes information on, for example, the phonological patterns and restrictions present in a specific language (e.g., every Polish word contains at least one vowel and simple Polish prepositions are usually monosyllabic, etc.); on the other hand, the semantic pole of these parts of speech also includes schematically defined contents (in cognitive linguistics, parts of speech are defined based on conceptual/semantic rather than formal criteria, which distinguishes it from structuralist and generative approaches), for example, a noun denotes a thing, understood as an area in a specific conceptual space, a verb – a process, and a preposition – a relation which does not change in time.

Syntactic structures also possess meanings generally characterised in a similar way, e.g., the ditransitive construction that underlies sentences such as *I sent you an e-mail*, *the company paid the employees a bonus*, or *the students asked the lecturer questions*. The semantic structure of this construction could be described as follows: one thing acts on another thing with a certain force in such a way that, as a result, the other thing passes into the area of control of a third thing (assuming, of course, the aforementioned understanding of the concept of *thing*, which is specific to cognitive grammar). As for intermediate cases, i.e., those that occupy the middle of the aforementioned continuum of symbolic units, they include constructions in which the semantic and phonological poles are simultaneously characterised in a less detailed way than concrete linguistic expressions but in a more concrete way than highly schematic grammatical elements. As an example, let us consider the morphological structure of a noun with the suffix {*-er*}. The phonological pole of such an entity is defined, on the one hand, in a more general way than the phonological pole of specific words created based on this structure (e.g., singer, farmer, cooker); on the other hand, it is defined in a more concrete way than the general category of the noun. The same applies to its semantic pole: it is more general compared to the concrete examples of such a word, but it also contains more information than a schematic unit, such as a noun; it is used to express a person (or, less frequently, thing) that performs a specific action or has a specific attribute.

These simple examples illustrate the blurring of the conventional boundaries between lexicon, morphology, and syntax postulated by Langacker. One important implication of the thesis of the symbolic nature of language as a whole should be emphasised: elements that are traditionally identified as strictly grammatical carry meaning just like lexical units do, although this meaning is more general. Thus, cognitive linguistics rejects not only seeing grammatical structures in purely formal terms, but also the possibility of attributing the same meaning to utterances that refer to the same situation but are based on different syntactic constructions, as is typical of generative grammar. In the light of the assumptions of cognitive linguistics, grammatical structures are "conventionally sanctioned (and therefore objectified) ways" of conceptualising conceptual content; in other words, "they are linguistic images of the world as seen through the eyes of the 'conventional' observer" (Tabakowska, 1995, p. 17).

The study of metaphor is among the most important research areas in cognitive linguistics. According to Taylor (2002, p. 487): "[p]robably the one publication which has done more than any other to popularise the Cognitive Linguistics movement was Lakoff and Johnson's short and very readable book, *Metaphors We Live By* (1980)". However, it is also significant that this was the first of several monographs which played a key role in the development of the theoretical foundations of modern cognitive linguistics. The approach proposed in the book deals with – in a supremely cognitivist manner – metaphor, which generativists reduced to the status of literary ornament and thus consistently dismissed. Since the 1980s, this approach to metaphor has undergone a number of modifications and extensions which are not discussed here, but its fundamental tenets remain largely valid.

First and foremost, Lakoff and Johnson draw attention to the ubiquity of figurative language, including metaphor – not only in poetry, but also in everyday life, as the very title of their seminal monograph suggests. This fact justifies and even mandates the inclusion of metaphor in the subject area of linguistics.

The cognitivist understanding of metaphor is evident in the very name of the theory proposed by Lakoff and Johnson: conceptual metaphor theory. The essence of this theory boils down to viewing metaphor as a mental, or conceptual, rather than a strictly linguistic phenomenon. In

other words, metaphor is treated here not just as a way of expressing one's thoughts on a linguistic level but primarily as a way of thinking. As the authors put it, "our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature" (Lakoff & Johnson, 1980, p. 3). The metaphorical nature here consists in conceiving a set of concepts related to one area of human experience in terms of a set of concepts related to some other area of experience. In this approach, the primary function of metaphor is cognitive and not stylistic.

Understood in this way, metaphor is usually expressed in the form of the formula: X IS Y (in cognitive linguistics, the notation in capital letters symbolises conceptual objects), where X denotes the set of concepts related to the experience that constitutes the actual object of reflection (called the target domain), while Y denotes the set of concepts through the prism of which the experience is conceived (called the source domain). The mechanism that underlies the classical version of conceptual metaphor is the metaphorical mapping (in the mathematical sense of the word 'mapping') of certain elements from the source domain structure onto the target domain. It is worth emphasising that the source domain, which provides the conceptual framework for the conceptualisation of elements from the target domain, usually encompasses experiences that are more concrete and perceptible to the senses, whereas the target domain usually refers to experiences that are more abstract and thus more difficult to perceive.

An example of a conceptual metaphor is LOVE IS A JOURNEY, in which the source domain consists of the elements of knowledge concerning travelling, while the target domain consists of knowledge concerning the romantic relationship between two people. This metaphor is based on establishing the following mapping between the elements from the source and target domains: people in the target domain who are in a relationship correspond to travellers from the source domain; the partners' shared goal corresponds to the destination of the journey; the shared experiences to date correspond to the distance travelled so far; problems in the relationship correspond to obstacles on the road, etc. As a cognitive mechanism, this metaphor predisposes us to think of the relationship between two people in terms of a journey, which in turn translates into a particular way of describing love linguistically: "Look how *far we've come*. We're *at a crossroads*. We'll just have to *go our*

separate ways. We can't *turn back* now. I don't think this relationship is *going anywhere*" (Lakoff & Johnson, 1980, p. 44). According to the cognitive approach to metaphor, concrete figurative linguistic expressions such as those above are, in a certain sense, a by-product of the workings of a particular conceptual metaphor, which, as a way of thinking, is primary to language. It is worth noting that the above examples of expressions based on the metaphor LOVE IS A JOURNEY do not belong to poetic language – they represent ordinary everyday speech. However, it is always possible to creatively develop a conceptual metaphor by creating an innovative expression in which the meaning is in harmony with its general sense. For example, the statement *A fleeting affect has lifted us into a state of weightlessness from which we are falling, aware of our inevitable collision with the ground* could probably be considered a poetic description of a romantic relationship based on the aforementioned conceptual metaphor. However, this example also shows that the boundary between everyday language and literary language is essentially a matter of degree, and that "all uses of language draw from the same conceptual resources" (Tabakowska, 1993, p. 27).

Other frequently mentioned examples of conceptual metaphors include LIFE IS A JOURNEY (*He has reached the end of his journey*, *I need to slow down a bit and catch my breath*); ARGUMENT IS WAR (*someone's line of defence*, *bring out the big guns*, *to shoot down someone's arguments*); TIME IS A VALUABLE COMMODITY (*to waste time*, *to gain time*, *to save some time*); HAPPY IS UP (*to lift someone's spirits*, *ears up*, *to be down*). There is also the 'conduit metaphor', which is our way of thinking about linguistic communication and consists of a set of three conceptual metaphors: MEANINGS ARE OBJECTS, LINGUISTIC EXPRESSIONS ARE CONTAINERS, and COMMUNICATION IS SENDING (Lakoff & Johnson, 1980, p. 10). The examples of our everyday understanding of linguistic communication by means of the conduit metaphor include *to give someone information*, *empty words*, *words full of contempt*, *to find no sense in someone's speech*, *to perceive someone's words in a certain way*, etc.

Our physical experience constitutes the basis for a large number of conceptual metaphors. At a very early stage in our ontogenetic development, recurring patterns of pre-conceptual experience, called image schemata, emerge from this experience (Johnson, 1987). For example,

one of the pre-conceptual structures that continuously appears in our everyday physical experience is the image schema SOURCE-PATH-GOAL, which is an element of the experience of moving from one place to another. Another example is the CONTAINER schema, which emerges from humans' physical experience almost from birth: in our infancy we are put into various containers (e.g., a cot, a pram) which form a barrier for our body, and later we place various objects in containers ourselves (e.g., blocks in a box) or we find ourselves in them (e.g., in a building or a car). Another example of an image schema is UP-DOWN, which we experience, for example when our mother picks us up from the cradle and puts us back down, or when we build a tower by adding new blocks to it and then we knock it down and observe how its height decreases. These kinds of image schemata, which arise from the specificity of our bodies and our experience of the physical environment in which we live, give structure to our concepts and therefore influence the way we speak. Thus, the SOURCE-PATH-GOAL schema underpins our metaphorical thinking about life and love. The UP-DOWN schema provides the experiential basis for many orientational metaphors (Lakoff & Johnson, 1980, p. 14 ff.), such as the aforementioned metaphor HAPPY IS UP or MORE IS UP (e.g., *rising/lowering prices*, *high/low wages*, *high/low birth rates*, etc.). The CONTAINER schema is the basis for understanding such abstract domains of our experience as linguistic communication (WORDS ARE CONTAINERS FOR MEANING) or states and situations, which we also comprehend in terms of containers (e.g., *to be in a difficult situation*, *to be in crisis*, *to fall into depression*, *to enter a state of euphoria*, etc.).

Image schemata are merely one of numerous manifestations of the impact of the specificity of the human body and its physical interaction with the environment on the human cognitive apparatus and, consequently, also on language. The conviction that such an impact exists is expressed in cognitive linguistics in the form of the thesis of the embodiment of mind and language. The relevance of this thesis to cognitive linguistics is expressed by the title of M. Johnson's 1987 monograph *The Body in the Mind. The Bodily Basis of Meaning, Imagination, and Reason*. Factors related to the embodiment of the human mind in language theory and research are thus programmatically embedded in modern cognitive approaches to language.

Systematic reflection with conclusions and recommendations

As (necessarily briefly and selectively) outlined above, the assumptions, postulates, and research areas in modern cognitive linguistics show how and why the study of language fits into a broad agenda of cognitive science research. A key reason why the study of language is part of cognitive science is the belief that the ability to use language is based on general properties of the human cognitive apparatus and that the structure of language reflects these properties in many ways. Thus, the fundamental task of linguistics that emerges from such an approach is to discover and describe phenomena of a mental nature based on authentic linguistic data. However, meeting such a broad objective poses a number of serious challenges.

First, despite its cognitivist bias and its focus on language as a cognitive ability, cognitive linguistics cannot forget that the mental processes responsible for language are embedded not only in the human body but also in a complex socio-cultural context, which is consequently reflected in the structure of language, as is embodiment. Thus, the significance of the theory of language developed within cognitive linguistics and its contribution to the body of cognitive science depends crucially on the extent to which it takes socio-cultural factors into account. Paying more attention to these issues and further developing the methodological basis for their analyses within cognitive linguistics will undoubtedly constitute a welcome element of the humanistic turn in cognitive inquiry (cf. Bremer, 2023).

The second fundamental challenge faced by cognitive linguistics concerns a frequently formulated objection to it, namely that it is impossible to falsify its claims regarding the semantic structure of language. The problem here stems from the adoption of a subjectivist approach to meaning, within which meaning is equated with conceptualisation. In the absence of tools that give us direct access to the mental world of language users, such an approach runs the risk of unconditional acceptance of postmodern ‘methodological relativism’ in semantic analyses, which “may indeed seem to reduce a scholarly pursuit to an impressionistic exercise in interpretation” (Tabakowska, 1993, p. 15). The situation is also not improved by the still rather common practice of formulating

semantic claims based on introspective data. One way to cope with this challenge is to increase the replicability of semantic studies, for example by means of semantic research based on quantitative methods (this is already being dynamically developed within cognitive linguistics) and the use of statistical tools (e.g., exploratory techniques and statistical hypothesis testing). Another response to both challenges is the intensification of interdisciplinary cognitive research on language, including research programmes involving the use of advanced techniques for studying neurophysiological processes.

Finally, it is important to emphasize the significance of confronting linguists' claims with relevant findings (including the most recent) from other cognitive sciences. If cognitive linguistics is to be truly cognitive, it must remain faithful to its fundamental commitment: the cognitive commitment.

References

Bremer, J. (2023). *The cognitive humanities*. In: B. Bodzioch-Bryła (ed.), *Contemporary Humanities*, (pp. 69–84). Kraków: Wydawnictwo Naukowe Uniwersytetu Ignatianum w Krakowie.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1959). Review of B.F. Skinner's *Verbal Behavior*. *Language*, *35*(1), 26–58.

Chomsky, N. (1969). *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Geeraerts, D., & Cuyckens, H. (2007). Introducing Cognitive Linguistics. In: D. Geeraerts, & H. Cuyckens (ed.), *The Oxford Handbook of Cognitive Linguistics* (pp. 3–21). Oxford/New York: Oxford University Press.

Heinz, A. (1983). *Dzieje językoznawstwa w zarysie*. Warszawa: PWN.

Johnson, M. (1987). *The Body in the Mind. The Bodily Basis of Meaning, Imagination, and Reason*. Chicago/London: The University of Chicago Press.

Lakoff, G. (1987) *Women, Fire, and Dangerous Things*. Chicago: The University of Chicago Press.

Lakoff, G. (1990). The Invariance Hypothesis: is abstract reason based on image-schemas?. *Cognitive Linguistics*, *1*(1), 39–74.

Lakoff, G., & Johnson, M. (2008). *Metaphors We Live By*. Chicago: University of Chicago Press.

Langacker, R. (1987). *Foundations of Cognitive Grammar*. Vol. 1. *Theoretical Prerequisites*. Stanford: Stanford University Press.
Langacker, R. (2008). *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.
Taylor, J. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. Oxford: Clarendon Press.
Taylor, J. (2002). *Cognitive Grammar*. Oxford: Oxford University Press.
Tabakowska, E. (1993). *Cognitive Linguistics and Poetics of Translation*. Tübingen: Gunter Narr Verlag Tubingen.
Tabakowska, E. (1995). *Gramatyka i obrazowanie. Wprowadzenie do językoznawstwa kognitywnego*. Kraków: Wydawnictwo Oddziału Polskiej Akademii Nauk.

Mateusz Hohol
Jagiellonian University in Krakow
https://orcid.org/0000-0003-0422-5488
Nina Bażela
Jagiellonian University in Krakow
https://orcid.org/0009-0003-1147-1086
https://doi.org/10.35765/slowniki.436en

Mathematical cognition

Summary

DEFINITION OF THE TERM: Mathematical cognition is an area of science that deals with mathematical activity in the mind (such as processing numbers and geometry), and the biological basis for human mathematical competence. Mathematical cognition is an area of interdisciplinary research founded on psychology and neuroscience.

HISTORICAL ANALYSIS OF THE TERM: The study of mathematical cognition follows a tradition of philosophical inquiry concerning mathematical intuition. While philosophers like Plato, Descartes, Kant, and Poincaré viewed this intuition as innate, Helmholtz thought it is shaped by experience. In his pioneering research on the development of mathematical cognition in children, Piaget continued to use the category of intuition and pointed out that knowledge of numbers and space is not built through observation but in action.

DISCUSSION OF THE TERM: Mathematical cognition is based on systems of core knowledge that are evolutionarily old, develop early in the course of individual development, and are culturally universal. These systems enable people to accurately assess the size of small sets of objects, estimate the size of larger sets, process shapes, and navigate spatially from an early age. However, these systems have limitations which are overcome – with the support of language – when learning numerals and spatial vocabulary. The research findings further indicate that the habit

of finger-counting plays an important role in the transition from core knowledge systems to symbolic mathematics.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: The study of mathematical cognition provides knowledge about one of the main research areas in the social sciences: the interactions between nature and culture. It also has implications for education, where it aids the creation of guidelines for supporting students with learning difficulties.

Keywords: geometry, numbers, spatial-numerical associations, innateness, core knowledge

Definition of the term

From the perspective of cognitive science, 'mathematical cognition' can be understood in two ways: as the *explanans* and as the *explanandum* of cognitive processes (Hohol, 2020). In the former, the mind is explained with the aid of mathematics, particularly modelling various cognitive processes using mathematical tools. In the latter, cognitive activity associated with the use of mathematics is explained by enlisting the methods and theories used in empirically oriented cognitive science. This approach involves interdisciplinary research grounded in experimental psychology and cognitive neuroscience, complemented by computational modelling, cognitive anthropology, cognitive ethology, and educational research.

In this article, we focus primarily on 'mathematical cognition' understood as *explanandum*, using the approach proposed by Rafael Núñez and George Lakoff (2005). Within this approach, the subject matter of studies on mathematical cognition includes

> some aspect of the psychological, neurological, or educational reality involved in some mathematical behaviour, performance, or competence of a person. The subject matter is defined at the level of an individual or at the level of an individual's nervous system. Mathematics per se is untouched (Núñez & Lakoff, 2005, p. 110).

The scope covers "what is usually meant by 'mathematics', [i.e.,] in general, simple arithmetic, number processing, or numerical calculation", as well as spatial and quantitative aspects of the environment, the modes of justification and proof practices used in mathematics, and the processing of mathematical objects and structures taught at the academic level. The scope of research on mathematical cognition is broad and includes studies on different populations, from infants to professional mathematicians. The research methods "are mainly standard empirical methods used in behavioral studies in psychology, studies with neuropsychological syndromes, and computational models of numerical processing" (Núñez & Lakoff, 2005, p. 110). Typical questions asked by researchers of mathematical cognition include: How does the mind represent numbers? Is there an innate basis for human numerical and geometrical competence, and if so, what is it?

Which brain structures are involved in the elementary cognitive processing of numbers and geometric properties? Is the cognitive processing of mathematical structures acquired at the academic level, built on elementary numerical-geometric cognition?

Historical analysis of the term

One of the earliest and most seminal works that has shaped our understanding of mathematical theories is Euclid's *Elements* (4th-3rd century BC), in which he presented patterns of logical reasoning and mathematical deduction. He summarised the mathematical knowledge accumulated over centuries by Greek mathematicians who drew on Egyptian and Babylonian achievements. Euclid's systematisation included plane geometry, stereometry, geometric algebra, and number theory, the latter also being geometric in nature. Numbers were treated as segments, while multiplication was seen as a measurement. Hence, geometry remained synonymous with mathematics for centuries, and geometric proofs were regarded a model of rigorous thinking. A frequently occurring term in *The Elements* is 'construction', which is clearly evident in Book I, where the first theorem concerns the construction of an equilateral triangle on a given finite straight line. Although the term 'construction' suggests that mathematical objects do not exist before reasoning is conducted, Euclid's work does not contain philosophical declarations of this kind.

Plato (4th century BC) believed that it was possible, through many years of training, to acquire mathematical knowledge thanks to an innate cognitive ability to grasp elementary knowledge about points, segments, angles, and relations between figures. This ability, traditionally called 'mathematical intuition', can be considered a historical predecessor to 'mathematical cognition' (Hohol, 2020). Intuition, also called 'insight', is linked to the metaphor of 'thinking is seeing', which dates back at least to Plato. Plato's thesis of the intuitiveness of elementary mathematics is illustrated in, among others, his dialogue *Meno*, in which an uneducated slave, guided by questions posed by Socrates, is capable of deductive reasoning, which leads him to the conclusion that a square built on the diagonal of a given square is twice as large (see Hohol, 2020). According

to Plato, constructing a geometric figure creates a representation of a mathematical object that is 'seen' by the mind but exists independently of it (see Hohol, 2020; Murawski, 2013). This thesis was defended by Plato's followers, such as Speusipppus of Athens (4th century BC). However, not all Hellenistic philosophers were Platonists. For instance, Aristotle defended the thesis of the empirical nature of mathematical knowledge that arises from the abstraction of mathematical objects from physical objects, while Menaechmus (4th century BC), a mathematician associated with the Platonic Academy, argued that geometrical constructions should be interpreted literally as operations that result in the creation of mathematical objects.

Mathematical intuition became one of the classical themes of modern philosophy. Descartes (1596–1650) claimed that mathematical theorems concerning eternal truths are discovered through intuition, which he called the 'natural light'. Significantly, according to him, the intuition that makes it possible to recognise the clarity and distinctness of mathematical theorems is innate. It is completely independent of the testimony of the senses and can be expressed by the metaphor 'thinking is seeing' (see Hohol, 2020, Murawski, 2013). Discussions on intuition continued in subsequent eras, most notably in the conceptions of Immanuel Kant (1724–1804), who believed that all mathematical theorems have the status of synthetic a priori judgements, which extend our knowledge (in contrast to tautologies) and are completely independent of empirical justification as their truth is recognised intuitively (see Hohol, 2020; Murawski, 2013). Synthetic a priori judgements are possible by linking mathematical constructs to space and time, understood as forms of pure intuition, inherent in our faculty of mind, which filter sensory impressions and organise our cognition. The internal representation of space enables the subject to construct geometry, while the representation of time underpins his ability to construct numbers, which are understood as units attached to one another sequentially in time (see Hohol, 2020).

Henri Poincaré (1854–1912) directly referred to Descartes' and Kant's conceptions of mathematical intuition and substantially modified its understanding (see Hohol, 2020). Poincaré called mathematical intuition the 'innate creativity' of the human mind which allows man to recognise fundamental relations and mathematical theorems as true. Noticing that mathematical cognition involves both conscious and

unconscious processing was Poincaré's innovative contribution to the understanding of this concept. He believed that many mathematical theorems arise at the unconscious level but must be supplemented by consciously controlled reasoning. Contrary to Kant, he claimed that the axioms that serve as the starting point of geometrical proofs are not a priori judgements but are adopted by mathematicians by virtue of convention, convenience, utility, and empirical facts. This led Poincaré to recognise that the geometry defined by Euclid's axioms was not the only valid version of geometry, although he acknowledged it as the most convenient. Innate mathematical intuition thus guarantees that reasoning remains correct, and the final results depend on the assumptions made by convention (see Hohol, 2020).

Hermann von Helmholtz (1821–1894) went even further in modifying the previous understanding of mathematical intuition by denying its innate nature and leaning towards radical empiricism. Although he agreed with Kant that sensual experiences and impressions are filtered through the mind, he rejected the transcendentalist viewpoint entirely. According to Helmholtz, the cognitive ability traditionally called mathematical intuition is shaped by accumulating experiences and reinforcing the associated memory traces. This means that the shaping of mathematical intuition can, and even should, be 'broken down' into simpler components; then, it should be shown how mathematical thinking is shaped during individual development. Helmholtz's theses on the empirical origins of mathematical intuition were still philosophical in nature but were inspired by his psychological research (it is worth mentioning here that he was one of the pioneers of experimental psychology), which laid the foundations for empirical research on learning mathematics (see Hohol, 2020).

The term 'mathematical intuition' did not fall out of use with the rise of experimental psychology, although its meaning underwent another significant change. Jean Piaget (1896–1980) played an important role in developing modern empirical research on mathematical cognition. He argued that cognitive development proceeds according to ordered stages, beginning with the sensorimotor stage (from birth to two years), when children begin to understand object permanence. The next stage (ages two to around six), the preoperational stage, is characterised by the inability to take a viewpoint other than one's own. Then (ages seven to twelve), children go through the concrete

operational stage, during which they master the basics of logic and become able to take another's viewpoint, which enables them to understand cause-and-effect relationships. Finally, at age twelve, children enter the formal operational stage, which is characterised by the ability to think abstractly.

Piaget devoted other works to the genesis of numerical and geometrical competence, based on a constructivist and action-oriented approach to all mathematical concepts (see Semadeni, 2023). In his opinion, mathematical concepts are not innate and do not emerge suddenly. They are gradually constructed in a child's mind and are based on experience. They are shaped not merely by perceiving the physical world, but by exploring the environment and actively manipulating objects. Piaget argued that arithmetical competence develops in parallel with logical abilities. The concept of number arises as a result of abstraction (i.e., disregarding the physical characteristics of objects, such as colour, size, or shape) based on actions performed on objects, such as the assignment of objects to sets (see Szczygieł, 2017). The development of the concept of number requires internalising the principle of permanence and, in particular, understanding that changing the position of objects in sets does not change their number. In a typical study conducted by Piaget, a child was shown two rows of the same number of objects (e.g., balls), located at the same distance; then, the objects in one of the rows were moved apart, which meant a change in their spatial arrangement but not in their number. Piaget noticed that even five-year-old children incorrectly claimed that pulling objects apart led to an increase in the number of objects. This error does not subside until the age of six or seven, when it marks the complete formation of the concept of number. However, later studies demonstrated that tasks that test the understanding of the principle of object permanence are correctly solved by children at a younger age. This discrepancy was explained by, among other things, the unnaturalness of Piaget's testing conditions (see Szczygieł, 2017). Moreover, using other research paradigms made it possible to study the presence of number representations even in newborns (see the *Discussion of the term* section).

Piaget was convinced that geometrical concepts do not simply derive from observing the world. For example, 'equality' derives from internalising the action of aligning two objects; a 'straight line' has at its root the

internalisation of the action of following something, e.g., with one's eyes, without changing direction; an 'angle' derives from the internalisation of two intersecting movements of, e.g., hands (see Hohol, 2020; Semadeni, 2023). Regarding the cognitive origins of geometry, Piaget – as was already mentioned above – did not shy away from the classical notion of intuition. As he wrote in a book he co-authored with Bärbel Inhelder:

> The 'intuition' of space is not a 'reading' or apprehension of the properties of objects, but from the very beginning, an action performed on them. It is precisely because it enriches and develops physical reality instead of merely extracting from it a set of ready-made structures, that action is eventually able to transcend physical limitations and create operational schemata which can be formalised and made to function in a purely abstract, deductive fashion. From the rudimentary sensori-motor activity right up to abstract operations, the development of geometrical intuition is that of an activity, in the fullest sense (Piaget & Inhelder, 1967, p. 449).

Piaget's concept, although largely undermined by later research (see Spelke, 2022; Szczygieł, 2017), was original and historically important as it represented the first attempt to define the stages of a child's arrival at the concept of Euclidean space (see Hohol, 2020). According to Piaget, individual development occurs in reverse order to that of the historical development of geometry as a branch of mathematics: while Euclidean geometry was established in antiquity, projective geometry developed in the 19th century, and topology emerged at the turn of the 20th century, Piaget argued that the most primordial concept of space that is formed in the mind of a child at around the age of three is 'topological space' (although it must be emphasised here that what Piaget meant by topology is far from its strictly mathematical definition (see Hohol, 2020; Semadeni, 2023)). He formulated his conclusions based on experiments in which children distinguished shapes, e.g., by touch, and did well when one of the objects was closed and the other was not. However, the youngest children did not manage to distinguish between a square and a triangle or even between rectilinear and curvilinear figures. On this basis, Piaget concluded that three-year-old children did not yet have the concepts of lengths or angles. While certain Euclidean concepts seem to form at around six or seven years of age, children still see all geometrical relations relative to their viewpoint (which corresponds to the still egocentric preoperational stage of general cognitive development). On this basis, Piaget concluded that the child at this

age has a concept of projective space. Finally, at around the age of twelve (which coincides with entry into the formal operational stage), children develop the concept of Euclidean space, which does not refer to a particular arrangement of objects that occupy perceived positions but rather to the organisation of space itself as an invisible 'container' independent of the mobile objects contained within it in which the relations between these objects are independent of the current viewpoint (Piaget & Inhelder, 1967). As is the case with the development of the concept of number, today the thesis of the primacy of topology is contested. For example, Piaget's observation that children first distinguish between open and closed figures is explained by their greater familiarity with certain objects rather than by the most basic nature of topology (see Hohol, 2020). What seems to be still valid, though, is Piaget's thesis that mathematical competence does not derive from mere visual perception of the world but from movement and action in the world.

Discussion of the term

The interdisciplinary scientific study of cognition, today called cognitive science, emerged at the turn of 1960s, and initially covered primarily cognitive psychology, linguistics, and computer science, which were gradually joined by neuroscience, anthropology, and philosophy. The pioneers of cognitive science, such as Allen Newell and Herbert Simon, were vitally interested in mathematical and logical reasoning, which is reflected in what were considered the first artificial intelligence programs: The Logic Theorist and the General Problem Solver, which were capable, among other things, of successfully proving the theorems from Bertrand Russell and Alfred North Whitehead's famous mathematical treatise *Principia mathematica*. Other programs from this period successfully proved theorems of Euclidean geometry. As critics observed, computer programs could achieve analogous mathematical results in a completely different way from humans. To eliminate this problem, designers and programmers gradually began to use psychological data (originally verbal protocols, later reaction time data). Hence, geometric programs started to simulate the interactions that people have with external representations e.g., diagrams (see Hohol, 2020).

However, proving theorems, which is a complex activity involving multiple cognitive processes, did not become the main focus of research on mathematical cognition. Instead, researchers focused on more elementary issues, such as the way the mind represents numbers, which, as it turned out, has much in common with the analogue processing of physical quantities in general. The results of numerous studies demonstrate that, when comparing two perceived physical objects, both humans and animals perform better (i.e., they achieve higher correctness and faster reaction times) when the objects differ significantly in size than when the difference in size is not as large. When subjects compare pairs of objects with a fixed size difference, they find it easier to compare pairs of small objects than pairs of larger objects (see Butterworth, 2022). The former is called the distance effect and the latter the size effect. Importantly, these effects occur when different physical quantities are processed, including the length of a segment or the brightness of an object.

In 1967, Moyer and Landauer described a similar distance effect for numbers (see Brożek & Hohol, 2017; Dehaene, 2011). When people are asked to choose the larger of two numbers (represented by Arabic symbols), they respond faster and make fewer errors when the numbers being compared are separated by a greater distance (e.g., 5 and 9) than when the distance is smaller (e.g., 4 and 5). A size effect for numbers has also been discovered: when the distance between the numbers being compared is constant (e.g., the difference between 3 and 5 and 5 and 7 is 2), people respond faster and make fewer errors when the numbers are smaller. Based on the discovery of these two effects, the concept of the mental number line has been proposed. The distance effect indicates that, at the level of automatic and unconscious cognitive processing, representations of numbers are spatially structured; comparing two numbers requires a cognitive 'zooming in' of a particular part of the axis, which is easier for numbers that are further apart. The size effect indicates that, unlike the number line taught at school, the mental number line is compressed logarithmically: as the numerical magnitudes increase, the distances between them decrease; thus, comparing smaller numbers is easier due to their greater resolution (see Cipora & Nęcka, 2012). Some studies suggest that during the course of individual development – specifically with the acquisition of school

mathematical competence – the degree of logarithmic compression of the number line decreases, which is associated with a reduction in the strength of the distance and size effects. However, most researchers today agree that the cognitive representation of magnitude, including numerical magnitude, can be activated independently of space, which means that the distance and size effects cannot be used to confirm the existence of the mental number line.

Nevertheless, the theoretical construct of the mental number line has gained substantial support from the SNARC effect, first observed by Stanislas Dehaene and colleagues in 1993 and replicated several hundred times since (Dehaene, 2011; see also Cipora & Nęcka, 2012; Brożek & Hohol, 2017). The SNARC (Spatial-Numerical Association of Response Codes) effect, i.e., the spatial relationship between a number and a response type, is visible in experiments in which the respondent's task involves making decisions about the properties of numbers presented one after another (e.g., indicating whether a number is even or odd): shorter reaction times are observed for the left hand for relatively small numbers, and for the right hand for relatively large numbers. The numbers are relatively small or large because this effect can be observed even when the set of test stimuli includes numbers from 1 to 9 (the number 5 being omitted for methodological reasons). This pattern suggests that, as in the case of the number line taught in school, the mental representation of numbers is structured in a directional way: smaller numbers being more to the left and larger numbers to the right. Currently, debate is ongoing regarding which factors affect the strength of the SNARC effect and even its direction (see Butterworth, 2022). On the one hand, the role of cultural factors in the direction of the mental number line is emphasised. The standard SNARC effect is observed in cultures where writing and reading runs from left to right, while the reverse SNARC effect is observed in cultures where writing and reading runs from right to left. Moreover, when people are asked to count from 1 to 10 on their fingers, those who start with the left hand manifest a stronger effect than those who start with the right hand. On the other hand, the results of studies conducted on animals indicate that the association between numbers and space has deep evolutionary roots.

The distance and size effects indicate that numbers are cognitively represented in an analogue way, i.e., like physical quantities. The

SNARC effect is usually interpreted as an argument for linking mental representations of numbers and space. These conclusions are supported not only by the results of numerous behavioural studies but also by the results of neuroscience studies conducted using functional magnetic resonance imaging. For example, the intraparietal sulcus of the brain responds in a similar way when people process symbolic numbers and segments. Based on these and other results, Vincent Walsh has formulated a theory of magnitude called ATOM (A Theory of Magnitude), which assumes that in the parietal lobe cortex there is a common area of the metric mechanism which is involved in processing time, space, and numbers (see Butterworth, 2022). It is worth remembering that this does not mean that all these quantities are processed in exactly the same way or that their mechanisms do not involve additional components. Interestingly, the evidence supporting the role played by the parietal lobe in the processing of numbers and space emerged long before neuroimaging techniques were used. In the first half of the 20th century, the Austrian neurologist Josef Gerstmann described a set of symptoms (called Gerstmann syndrome) associated with damage to the angular gyrus. These include difficulties in processing numbers, distinguishing sides, and writing, as well as finger agnosia (e.g., impaired recognition of which of the patient's fingers is currently being stimulated).

Contemporary research aims to understand the mechanism of number processing not only at the level of brain structures but also the level of individual neurons. Studying macaques, Andres Nieder (2019) found numerical neurons located in the intraparietal sulcus area, i.e., an area specialised in size processing, and in the lateral prefrontal cortex, i.e., a structure involved in general cognitive processes. The name of these neurons is associated with the fact that they respond selectively to small non-symbolic numbers: a given neuron responds most strongly when a macaque perceives one dot; another neuron responds when the animal perceives two dots, and so on, up to four. Four (plus or minus one element) is a barrier to what is known as subitising, which is the effortless, rapid, and precise assessment of the number of elements exhibited. Subitising must not be confused with estimation, which involves approximating the number of elements in a set greater than four (plus or minus one) nor with counting. While subitising and estimation are independent of symbolic representations such as Arabic numbers

or numerals, when we have learned counting, we are able to precisely determine the number of elements in a set that exceeds – and even significantly exceeds – the range of subitising. Although many studies have revealed that representatives of some animal species can be taught to count, researchers agree that this ability is part of man's cognitive ecology (see Butterworth, 2022).

In the context of the considerations presented in the *Historical analysis of the term* section, two questions arise: given current cognitive knowledge, can the elementary numerical abilities of subitising and estimation be considered the foundation on which counting, arithmetic, and other mathematical abilities are superstructured during ontogeny? Are these elementary abilities innate? In other words: were the philosophers who searched for the innate basis of mathematical intuition right, or was Hermann von Helmholtz, who questioned the existence of anything innate apart from the learning mechanisms themselves, right? Although debate concerning these questions is still ongoing (see Butterworth, 2022; Haman & Gut, 2016; Nieder, 2019), many researchers of mathematical cognition are inclined to support the thesis of the innateness of elementary numerical cognition and the thesis that it forms the basis for learning mathematics.

Regarding the innateness of elementary numerical competence, human infants manifest the ability to distinguish sets of objects that differ in number without the need for any training. This has been demonstrated by studies based on the habituation paradigm. When children aged about six months observe a series of pictures with an equal number of black dots that differ in size or placement, the period of their interest in the stimulus is shortened, which indicates habituation. When they are shown pictures with different numbers of dots, regardless of their spatial arrangement, they regain interest in the picture, as manifested by an extended period of attention to the stimulus (see Brożek & Hohol, 2017). Other studies have shown that infants possess an intuition of equinumerosity. For example, when hearing a sequence of four syllables, infants looked longer at a set of four dots than at a set of twelve dots. Furthermore, studies conducted on various animal species – ranging from non-human primates to rodents, birds, and even fish – indicate that the ability to perceive numerosity is common in nature, at least among vertebrates (see Butterworth, 2022; Nieder, 2019).

According to the influential and still developing theory of core knowledge formulated by Elizabeth Spelke (2022), our cognitive abilities are based on core knowledge systems. According to Spelke (2022), these systems

> operate on a limited domain of entities, and they capture only a limited subset of the properties that our perceptual systems deliver. These systems are also encapsulated, automatically activated and unconscious (…). These core knowledge systems emerge early in ontogeny and function throughout life, and they are present in people of all ages and in all cultures. Both during and after infancy, they guide our thinking and learning. Moreover, the same abilities and limits are found in a wide range of animals and depend on homologous brain systems and processes across those animals, providing evidence that core systems emerged deep in our evolutionary past (p. 190; see also Haman & Gut, 2016).

According to core systems theorists, two distinct systems located in the parietal lobe (called the object tracking system and the approximate number system), are responsible for subitising and estimation. Together, they serve as our 'number sense', which is a prerequisite for mastering numbers expressed symbolically and performing actions on them (see Dehaene, 2011).

As indicated above, there are arguments that support the thesis that representations of numbers are spatial in nature. But where do our spatial intuitions come from and are they, as Kant claimed, the basis of Euclidean geometry? According to Spelke (2022), as is the case with the 'number sense', our geometric abilities are grounded in two phylogenetically old, ontogenetically early, and culturally universal core systems, which are called the system of layout geometry and the system of object geometry (Hohol, 2020). The former is located in the hippocampus area and processes distances and directions, enabling navigation in the environment. The latter is located in the lateral structures of the occipital lobe and processes angles and lengths, enabling the recognition of two-dimensional shapes and three-dimensional objects. As Spelke (2022) observes, all core systems have certain limitations. The object tracking system provides accurate knowledge on numbers, but only up to about four objects. The approximate number system allows operations on larger numbers to be performed, but at the expense of precision. The system of layout geometry is sensitive to distance and direction, but not to angles; the system of object geometry processes lengths and

angles, but not directions. How is it possible, then, that the cognitive representations generated by such limited systems are combined into a constraint-free system of mathematical concepts?

Although many details are still unknown, according to Spelke (2022), combining accurate representations of small numbers with approximate representations of larger numbers is possible thanks to the mastery of numerals. Children begin to use basic numerals at a very young age but cannot assign proper meanings to them until the age of two. A significant change occurs in the third year of life, when children begin to grasp that the numeral 'one' represents the abstract value of 1. This process continues for subsequent numbers and their corresponding numerals up to the number 4 in the fourth year of life. From this point onwards, the ability to use numerals develops rapidly, which goes hand in hand with learning how to solve more complex arithmetic tasks. Children formulate a universal rule that allows them to assign a corresponding numeral to each number, which promotes their understanding of both the cardinal and ordinal aspects of numbers. Language thus functions as a scaffolding to overcome the limitations of core number systems. Research further indicates that finger counting habits play an important role in the transition between knowledge contained in core systems and symbolic mathematics. Fingers are concrete objects that are easy to manipulate and are suitable for representing the numerosity of different objects. Moreover, due to their natural structuring, finger counting habits promote mastery of the ordinal aspect of numbers, whereas core knowledge systems only encode the cardinal aspect (see Szczygieł et al., 2015).

Research indicates that, in geometry, spatial vocabulary learning plays a similar role to numerals (see Hohol, 2020). This process is also supported by children's growing experience with map-like sketches, through which children learn to represent the system of layout geometry, originally represented by the system of object geometry (direction, distance), as two-dimensional structures whose key geometric properties are angles and lengths.

Systematic reflection with conclusions and recommendations

Contemporary research on mathematical cognition can be seen as a continuation of philosophical inquiries into the nature of mathematical intuition. This is evidenced by the fact that Stanislas Dehaene and Elizabeth Brannon (2010) do not hesitate to call one of the main currents of research a 'Kantian research programme', which emphasises the continuity between the biopsychological foundations of mathematical cognition and how our minds process time and space, where Kantian 'filters' are replaced by core cognitive systems with a long evolutionary past. Dehaene and Brannon do not even hesitate to claim that if Kant were alive today, he would have become a neurocognitive scientist rather than a philosopher (p. 519). Compared to earlier philosophical inquiries, contemporary research on mathematical cognition makes it possible to go beyond the realm of rational speculation and to build theories that can be tested in experiments using methods that have also been proven to work well in the study of other areas of the mind. However, it would be rash to conclude that the empirical study on mathematical cognition simply replaces the philosophy of mathematics. While it is hardly conceivable today to build plausible philosophical theories about mathematical knowledge that would openly contradict the current understanding of neurocognitive mechanisms, it is equally hard to imagine neurocognitivists solving the problem of the ontological status of mathematical objects and structures. In other words, contemporary research on mathematical cognition cannot resolve the aforementioned dispute between Speusipppus and Menaechmus.

Another important strand is that empirical research on mathematical cognition sheds new light on a number of problems that basically all social sciences face. A key one relates to the question of 'nature or culture'. On the one hand, the foundations of mathematical cognition are extra-linguistic and directly related to the numerosity processing abilities and spatial skills possessed by many animals, which leads many researchers to consider them innate (see Butterworth, 2022; Haman & Gut, 2016). On the other hand, the innate form of mathematical cognition, i.e., core knowledge, is highly limited (e.g., the issue of the narrow scope of subitising). Transcending the limitations of core knowledge is

possible thanks to a specifically human cultural product, i.e., language, as evidenced by the fact that speakers of languages characterised by narrow numeral systems, e.g., the Munduruku people, perform less precise calculations outside the scope of subitation (see Brożek & Hohol, 2017; Butterworth, 2022). The classic 'nature or culture' question thus seems misplaced to the extent that, without innate core systems, it could be argued that we would acquire no mathematical practices through education at all. However, these practices are deeply immersed in culture, the products of which act as 'cognitive tools' that significantly modify core knowledge.

There are also a number of important practical implications for education associated with the study of mathematical cognition. First and foremost, experts in the field of mathematical cognition recommend introducing mathematical abstractions at school by grounding them in spatial experience, which is common to all humans due to the structure of our bodies. Particularly important in this respect is the promotion of the formation of spatial-numerical associations. However, the thesis 'more space, better mathematics' is not universally true. While stronger spatial-numerical associations in children (expressed, e.g., by the aforementioned stronger SNARC effect) are usually correlated with better competence in school mathematics, the relationship may be reversed in adulthood. This is particularly evident in the case of professional mathematicians, who do not manifest the SNARC effect at all, which may indicate that prolonged training leads to the construction of more abstract, or less embodied, representations of numbers (Cipora et al., 2016).

Moreover, research on 'number sense' is valuable for understanding the nature of developmental dyscalculia, which hinders and, in extreme cases, even prevents the acquisition of school mathematical knowledge and is a barrier to entry into numerous vocational pathways. As cyclically published educational reports indicate, learning difficulties in mathematics have far-reaching consequences for the well-being of entire societies (see Butterworth, 2022, pp. 30–33). Thus, psychological interventions are recommended as early as possible to help those with diagnosed dyscalculia or at risk of dyscalculia in order to stimulate the fullest possible development of their number sense, which is a prerequisite for the acquisition of school mathematical knowledge (see Dehaene, 2011; Butterworth, 2022).

Finally, research in the field of mathematical cognition demonstrates that the processing of mathematical knowledge, which is considered the most abstract object of human cognition, is, at its core, based on the mechanisms used in everyday life by everyone, regardless of their educational and professional paths.

References

Brożek, B., & Hohol, M. (2017). *Umysł matematyczny*. Kraków: Copernicus Center Press.

Butterworth, B. (2022). *Can fish count? What animals reveal about our uniquely mathematical mind*. London: Quercus.

Cipora, K., Hohol, M., Nuerk, H.-C., Willmes, K., Brożek, B., Kucharzyk, B., & Nęcka, E. (2016). Professional mathematicians differ from controls in their spatial-numerical associations. *Psychological Research*, *80*(4), 710–726. DOI: 10.1007/s00426-015-0677-6.

Cipora, K., & Nęcka, E. (2012). Kontinua a przestrzeń – przegląd badań nad przestrzennym komponentem poznawczej reprezentacji wielkości i nasilenia. *Psychologia – Etologia – Genetyka*, *26*, 7–21.

Dehaene, S. (2011). *The number sense*. New York: Oxford University Press.

Dehaene, S., & Brannon, E.M. (2010). Space, time, and number: A Kantian research program. *Trends in Cognitive Sciences*, *14*(12), 517–519. DOI: 10.1016/j.tics.2010.09.009.

Haman, M., & Gut, A. (2016). Wiedza wrodzona. In: J. Bremer (ed.), *Przewodnik po kognitywistyce* (pp. 681–712). Kraków: Wydawnictwo WAM.

Hohol, M. (2020). *Foundations of geometric cognition*. London: Routledge.

Murawski, R. (2013). *Filozofia matematyki: Zarys dziejów*. Poznań: Wydawnictwo Naukowe Uniwersytetu Adama Mickiewicza.

Nieder, A. (2019). *A brain for numbers: The biology of the number instinct*. Cambridge, MA: The MIT Press.

Núñez, R., & Lakoff, G. (2005). The Cognitive Foundations of Mathematics: The Role of Conceptual Metaphor. In: J.I.D. Campbell (ed.), *Handbook of Mathematical Cognition* (pp. 109–124). Psychology Press. DOI: 10.4324/9780203998045.109.

Piaget, J., & Inhelder, B. (1967). *The child's conception of space*. New York: W.W. Norton & Company.

Semadeni, Z. (2023). *Różne oblicza matematyki*. Toruń: Wydawnictwo Naukowe UMK.

Spelke, E.S. (2022). *What babies know: Core knowledge and composition. Volume 1*. Oxford University Press. DOI: 10.1093/oso/9780190618247.001.0001.

Szczygieł, M. (2017). Konstruktywizm Jeana Piageta i koncepcja zmysłu liczby a edukacja matematyczna. *Edukacja*, *140*(1), 7–26. DOI: 10.24131/3724.170101.

Szczygieł, M., Cipora, K., & Hohol, M. (2015). Liczenie na palcach w ontogenezie i jego znaczenie dla rozwoju kompetencji matematycznych. *Psychologia rozwojowa*, *20*(2), 23–33. DOI: 10.4467/20843879PR.15.014.3803.

Marek Lechniak
The John Paul II Catholic University of Lublin
https://orcid.org/0000-0002-0768-7963
https://doi.org/10.35765/slowniki.475en

Contemporary logic and its role in the sciences and the humanities

Summary

DEFINITION OF THE TERM: Logic can be approached from two perspectives: as a scientific discipline and through its applications. In a broader sense, logic consists of logical semiotics, formal logic, and methodology. In a narrower sense, it refers specifically to formal logic.

HISTORICAL ANALYSIS OF THE TERM: The history of logic dates back to antiquity, its evolution has not been linear but has progressed through various successive forms. Contemporary logic encompasses classical logical calculus (along with metalogic and the philosophy of logic) and systems of non-classical logics that either deviate from or extend classical logic.

DISCUSSION OF THE TERM: In the sciences, logic is applied in mathematics, where classical logical calculus plays a central role, and in physics, where both classical calculus and non-classical logics are used. In the humanities, logic helps address methodological issues specific to the field, such as defining concepts and offering humanistic interpretations, often through non-classical logics and formal ontologies.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Logic, at least *de jure*, underpins the construction of scientific theories, ensuring precision in language and clarifying the relationships between theoretical elements. It also guides the explanatory and justificatory processes within those theories. Researchers leveraging

contemporary non-classical logics can apply them to the development of interdisciplinary methods, particularly in fields like cognitive science.

Keywords: contemporary logic, logic in the sciences, logic in the humanities, non-classical logics, cognitive science

Definition of the term

There is no single, universally accepted definition of logic. Broadly speaking, logic can be understood as

> the analysis of language and research activities aimed at establishing rules for the effective use of language and the performance of these activities (Marciszewski, 1970, p. 126).

It is important to differentiate between a broad understanding of logic (general logic) and a narrower one. In the broader sense, logic revolves around two central themes: the need for clear, consistent, rigorous, and ordered thinking and speaking, and the issue of correct inference. In the narrower sense, logic focuses on studying schemas, i.e., forms of infallible reasoning which possess the highest degree of generality and are applicable across all sciences and in everyday language (Marciszewski, 1970, p. 133). The broader conception of logic figuratively encompasses the common usage of the adverb 'logically', which, similarly to the Greek *logos*, is most often applied to verbs like 'say', 'think', and 'act.' To speak logically means to express thoughts clearly and precisely, conveying them unambiguously and intelligibly. Logical thinking involves avoiding self-contradiction, drawing valid conclusions from known information, and reasoning correctly without moving from true to false statements. Acting logically implies well-considered, efficient action, setting appropriate goals, and achieving them with minimal effort.

Textbooks typically divide logic into three sub-disciplines: the logical theory of language (logical semiotics), formal logic, and the methodology of science. Logical semiotics examines language as a tool for efficient communication. Formal logic addresses objective relationships between propositions of specific forms, grounded in the relationships between states of affairs. These relationships are expressed through the laws of logic, which ensure the reliability of inferences across various fields of knowledge. Finally, the general methodology of science is concerned with the methods used to understand the world, particularly how we justify the claims we make about it (Kamiński, 1992).

Since the earliest days of thought, two distinct perspectives on logic have emerged:

- logic as a field of knowledge (*logica ut doctrina* or *logica docens* in medieval terms),
- logic as a tool for acquiring knowledge and a set of practical logical skills (*logica ut disciplina* or *logica utens*).

This duality can be observed in Aristotle's approach to logical problems. He treated logic as a propaedeutic to knowledge (science), emphasising its role in establishing general scientific laws. However, when Andronicus of Rhodes was organising the *corpus Aristotelicum*, he placed Aristotle's logical works at the beginning of the collection, collectively naming them *Organon* (meaning 'tool'). This distinction remains relevant in modern views of logic. On one side, there are sophisticated scientific systems, rigorously studied using formal methods; on the other, there is an ongoing concern about how these complex theories can be applied as practical tools in the sciences and the humanities. Both aspects of logic – as *doctrina* and *disciplina* – play a crucial role in the interdisciplinary field of cognitive science. Logic, along with its various sub-disciplines, is deeply involved in the study of information, representation, and the processing of knowledge.

Historical analysis of the term

The periodisation of the history of logic is challenging. J.M. Bocheński suggests dividing the history of European logic into five key periods: (a) the ancient period, (b) the early medieval period (7th–11th centuries), (c) the scholastic period (12th–15th centuries), (d) the period of modern classical logic (16th–19th centuries), (e) the period of mathematical logic. In his opinion, the early medieval and modern classical logic periods were largely uncreative, leaving the ancient period, the scholastic period, and the mathematical logic period as the most significant (Bocheński, 1961). Bocheński emphasises that the development of logic is not continuous; new forms of logic do not build directly on old logic. Each new logic arises from different assumptions, employs distinct tools, and addresses previously unexplored problems. Logic evolves within divergent forms, where sharper problems, better methods, and additional laws emerge, but later forms are not necessarily superior to earlier ones. This holds true even for mathematical logic, which, despite its technical sophistication, lacks

the subtlety of medieval semiotics and tends to equate logic with logical calculus. According to Bocheński, three fundamental historical forms of logic can be identified:

- The ancient form, in which theorems are usually formulated in subject-specific language, the semantics of language is underdeveloped, and theories are constructed in simplified everyday language with variables. Research begins with expressions formulated in natural language, from which laws and rules are abstracted.
- The scholastic form, in which theorems are formulated in metalanguage and semantics is sophisticated and elaborate. Theories are constructed in everyday language without limiting semiotic functions, and the laws and rules are derived through abstraction from natural language.
- The mathematical form, which marks a return to certain aspects of antiquity, where theorems are formulated in subject-specific language, but with more extensive use of variables and strict semiotic definition of signs. Prior to the 1930s, the semantics of language was largely neglected. The novel aspect of this form is the use of artificial language, where formalism is prioritised over interpretation. This approach underpins many artificial intelligence systems used in cognitive science to model cognitive processes.

Discussion of the term

Sub-disciplines of logic

The division of logic into sub-disciplines cannot be made in a clear-cut manner. However, it is commonly categorised into the following three main areas:

- Logical semiotics, in which language is examined at three levels: the syntactic level (relations between expressions); the semantic level (relations between a language expression and the object of reality it refers to; the pragmatic level (relations between language expressions and their users). As part of the general theory of language, logical semiotics plays a key role in cognitive science by

helping to analyse how meaning emerges in perception, thought, and language as a system of signs.
- Formal logic studies the formal relationships between language expressions based on their structural correctness. It examines how expressions with a particular structure lead to other expressions of a similar structure.
- Methodology of science studies broadly understood scientific methods that aim to produce cognitively valuable results.

The scope of logical semiotics includes several key issues: theories of syntactic categories and logical grammars; various types of logical divisions, including classification and typology theories; analyses of errors in expressing thoughts verbally; inquiries into definitions and semantic issues, such as the concept of denotation (e.g., denotation by singular term or descriptions); and analysis of truth in both formal and natural languages. It also includes pragmatic conceptions and problems related to the use of language (such as meaning, recognition, and conversational implicatures).

In formal logic, scholars construct various logical systems (Marciszewski, 1987), study the formal properties and relationships between these systems (metalogic), and engage in philosophical inquiries into the nature and applicability of formal systems (philosophy of logic).

Finally, the methodology of science (often referred to today as the 'philosophy of science') deals with general inquiries into the nature of science, both as a functional process and as a product, i.e., a scientific theory. It also covers the theory of questions, various theories of reasoning and argumentation, as well as the exploration of the structure and dynamics of scientific theories.

Contemporary formal logic

The term 'formal logic' is derived from its focus on the formal aspects of reasoning rather than the content of expressions. In formal logic, the specific meaning of expressions is abstracted away and attention is placed on their structure. As a result, the emphasis is on propositional forms rather than individual propositions because formal logic deals with the relationships between expressions of a particular form,

not their specific truth values. A characteristic feature of contemporary formal logic, though not essential, is its use of symbolic language. This symbolic representation ensures the unambiguous expression of logical rules and enables the mechanisation of reasoning procedures. The combination of symbolic language and the calculative ('mathematical') nature of formal logic systems makes them models of scientific rigor and precision. These systems provide a solid foundation for the justification of logical theorems (Łukasiewicz, 1961; Borkowski, 1970).

Classical logical calculus

At the core of formal logic lies classical logical calculus, which consists of two main components: propositional calculus and the more specific predicate calculus (or calculus of quantifiers) with identities. Some scholars also include the algebra of sets and relations as part of classical logic. Under certain assumptions, the theory of categorical propositions can be considered a subset of predicate calculus. Additionally, Stanisław Leśniewski's systems, especially protothetics and ontology, are sometimes classified within classical logic.

Propositional calculus can be structured in various ways. At the syntactic level, it can be represented through systems like the axiomatic system or natural deduction systems (e.g., assumptional systems or the Genzen systems). At the semantic level, tools like truth tables (0–1 tables) or analytic tableaux can be employed. Regardless of the construction method, the propositional calculus rests on two fundamental assumptions: the principle of bivalence (every proposition is either true or false) and the assumption of truth-functional operators (each logical operator has a clear relationship between the truth values of its arguments and the overall truth value of the expression it forms). These assumptions involve the idealisation of natural language, where not every expression is semantically determined to be true or false (e.g., vague or incomplete sentences), and where there are many conjunctions which are not truth-functional (non-extensional functors).

A natural deduction system can take the form of, for example, an assumptional propositional calculus, which is based on two types of rules: rules for creating assumptional proofs by contradiction, and

primary rules for adding new lines to a proof that define the formal behaviour of basic logical operators (negation, conjunction, disjunction, implication, and equivalence). This system can be extended to a more specific predicate calculus with identity by expanding its vocabulary, adding formation rules for new constants, and introducing rules that define the formal behaviour of quantifiers and the identity symbol.

Non-classical logics

It has been recognised that classical logical calculus may be insufficient for adequately capturing certain types of expressions found in natural language and the language of its scientific variants. Although propositional calculus is typically sufficient at the level of propositions, there are reasons to modify the classical system. One such reason is the presence of numerous non-extensional phrases in natural language, such as modal expressions. Another reason is the growing scepticism about the universal applicability of certain laws of propositional calculus. In response, two main directions of non-classical logics have emerged. In the first direction, non-extensional expressions are incorporated into the language of propositional calculus. In the second direction, deviant logics are developed where the set of symbols in the new system remains the same as in propositional calculus, but certain theses of propositional calculus are not theses in these new logics.

Modal logics

Among the extensions of propositional calculus, several systems of modal logics stand out, each designed to capture different types of expressions. These include alethic modal logics (e.g., ‘it is necessary that’, ‘it is possible that’), epistemic logics (e.g., ‘he knows that’, ‘he believes that’), deontic logics (e.g., ‘it is obligatory that’, ‘it is permitted that’, ‘it is forbidden that’), temporal logics, and causal logics. It has been observed (Von Wright, 1951) that there are numerous formal analogies between the behaviour of operators used in these logics (e.g., ‘it is necessary that’, ‘he knows that’, and ‘it is obligatory that’), as well

as certain disanalogies (e.g., while the proposition 'it is necessary that *p*' in many systems often implies '*p*', the phrase 'it is obligatory that *p*' does not carry the same implication). Due to these analogies, all extensions of classical logic are broadly referred to as modal logics. There exists a continuum of systems of non-classical logics. By the mid-20th century, intuitive difficulties were noted regarding certain syntactically formulated laws of modal logics, such as those involving the iteration of modal expressions (von Wright, 1951; Świrydowicz, 2004; Hughes & Cresswell, 1968). However, the discovery in the late 1950s of simple formal semantics for modal logics (e.g., relational semantics, known as Kripke semantics) marked a new phase in the development of these logics.

Relational semantics is based on Leibniz's idea that a proposition is necessary if it is true in every possible world. The key innovation that allows this idea to be applied to various systems of modal logics is that the quantification of possible worlds is restricted through the introduction of an accessibility relation between worlds. According to this view, a proposition is necessary if it is true in every possible world accessible from the current world. The concept of possible worlds makes it evident that many properties of modal operators that lead to the construction of different systems of modal logics in formal semantics are reflected in various properties of accessibility (alternativeness) relations between possible worlds. Such semantics serve as a tool to determine whether certain formulas in modal logics are theses of specific systems of modal logics.

Many-valued logics

There are many systems that can be treated as deviations from classical logic. The most well-known are systems of many-valued logics, where each operator is treated as a function. By retaining the matrix-based characterisation of operators, the number of values in these matrices is increased, thus expanding the range of possible characteristics of operators (e.g., in 3-valued logic alone, there are 27 possible unary operators and 3^9 binary operators). In the initial phase of developing these logics, it appeared that, due to the revision of the principle of

bivalence, they were competitive with propositional calculus (e.g., the laws of non-contradiction and excluded middle did not hold). However, it was subsequently discovered that the matrix-based characterisation of operators could be maintained without revising the classical notion of truth. Many-valued matrices could be created by intersecting the division of propositions into true and false with other divisions (e.g., determined vs. undetermined, informed vs. uninformed). An example is the matrices used in systems of relevant logics, where the goal is to preserve a content-related connection (the relevance criterion) between premises and conclusions – a criterion that classical logic does not meet, particularly due to the principle of explosion (e.g., $p \wedge \neg p \rightarrow q$). An undeniable benefit of using a matrix-based characterisation is that it provides a simple criterion for decidability and an easy way to characterise new operators. Unfortunately, the downside is that the results obtained can sometimes be counterintuitive.

Applications of logic in the sciences

The interdisciplinary nature of logic is reflected not only in its capacity to integrate various aspects of different disciplines (such as the sciences and the humanities) but also in its ability to identify shared elements across fields that, from the standpoint of those individual disciplines, may seem disparate.

Applications of logic in mathematics

It is acknowledged that there are three main types of concepts in mathematics (Grzegorczyk, 1981):

1. Logical concepts: these include truth-functional operators, quantifiers (expressions that specify for how many objects the proposition expressed in these expressions holds; for every x, for some x), and identity expressed as $x=y$.
2. Concepts related to set theory: these include sets, membership (which defines whether an element belongs to a set), relations, and functions.

3. Arithmetical concepts: these include natural numbers, real numbers, and arithmetical operations (such as addition, multiplication, and others).

The role of the concepts from each of these three groups in classical mathematics can be briefly characterised as follows: logic primarily provides methods for constructing concepts and building proofs; set theory supplies the most general concepts (such as the concept of a set); arithmetic concepts form the basis for specific mathematical content.

> Logic captures the methods of reasoning used in the sciences that are regarded as valid and organises them into logical systems. These systems are sets of laws and rules which, when followed, allow us to replicate all forms of reasoning that we intuitively consider sound. In mathematics, there is broad consensus regarding the acceptance or rejection of certain methods of reasoning. There is no situation where some view certain rules of reasoning as valid while others deem them completely incorrect. The differences lie only in the interpretation of certain rules and views regarding the scope of their application. Practically speaking, the set of all reasoning methods used in mathematics yields the classical system of logic (Grzegorczyk, 1981, p. 65).

Following Grzegorczyk, whose views reflect a widely held opinion on the matter, it can be stated that classical logic is sufficient for reasoning in classical mathematics, and there is no need to introduce other types of logic (e.g., non-classical logics).

A complication in this area emerged at the beginning of the 20th century with the advent of the intuitionist programme. According to its proponents (e.g., L.E.J. Brouwer and A. Heyting), classical mathematics should not be based on classical logic, which is characterised by non-constructiveness. This non-constructiveness arises from the acceptance of the law of the excluded middle $p \vee \neg p$ and the law of double negation $p \equiv \neg \neg p$ in classical logic. According to the law of excluded middle, either the state of affairs asserted by the proposition or the state of affairs asserted by its negation must hold. The law of double negation states that a proposition is equivalent to the negation of its negation. However, intuitionism, starting from a different perspective than classical logic, understands propositions and negation differently: a propositional variable represents not just a state of affairs (a description of a mathematical object) but also the procedure for its construction. Therefore, the law of excluded middle, interpreted as “every state of a mathematical object

is either constructed or its negation is constructed", cannot be accepted in intuitionism. In any richer calculus, there are propositions for which neither the proposition nor its negation can be proven. Similarly, the law of double negation, which in the intuitionistic interpretation reads as "the absurdity of the absurdity of a proposition implies its truth (constructability)", is unacceptable for intuitionists. Thus, within the intuitionist programme, logic becomes a part of mathematics, specifically a theory of proof procedures used in mathematics that is abstracted from the actual activities of mathematicians. The calculus used to control reasoning in intuitionistic mathematics is intuitionistic logic, which is a non-classical logic. In an intuitionistic interpretation, this logic is competitive with classical logic, while in a more precise interpretation it is not competitive but rather pertains to a different non-ontological, research perspective.

Applications of logic in physics

According to W. Heisenberg, it was historically believed that the role of language in the natural sciences was minimal. However, this perspective shifted dramatically in the early 20th century with the advent of relativity theory and quantum mechanics, particularly with the discovery of radioactive elements. These breakthroughs fundamentally altered our understanding of matter and its transformation processes.

> The 20th-century physicists had to adjust to a new understanding of the microcosm of elementary particles, moving away from modelling them after the macrocosm of large, tangible bodies. Unlike macroscopic objects, which are directly accessible to the senses, elementary particles cannot be assigned straightforward geometric and kinematic properties like shape or motion in space. The concept of the elementary particle in modern physics is far more abstract than Democritus' atom. Even in ancient Greek thought, atoms lacked properties such as smell, colour, taste, viscosity, and temperature. Furthermore, Heisenberg added that when conceptualising elementary particles, we can no longer disregard the physical processes through which our knowledge of these particles is obtained (Kiczuk, 2001, p. 143).

In light of the conceptual complexity of modern physics, Heisenberg observed that physicists rely on two distinct languages. On the one hand, the laws of physics are articulated through mathematical language,

consisting of equations that provide a precise mathematical representation of various physical phenomena. On the other hand, the terms of physical theory defined by these equations – such as time, mass, space, velocity, force, position, and acceleration in Newton's mechanics – carry illustrative phenomenal content that can be integrated into everyday language. As new physical theories emerge, they introduce additional concepts (expressed in the language of mathematics) into everyday language, such as the concept of entropy from thermodynamics, thereby enriching our imaginative understanding of the world. Thus, according to Heisenberg, it is impossible to avoid using a second language (in addition to the mathematical language) of physical theories, namely, an imaginary language, which is a precisely refined form of everyday language.

> According to Heisenberg, physics consists not only of experimentation and measurement on one side and mathematical apparatus on the other. The physicist's duty is also to explain, using everyday language, everything that actually happens in the interplay between experiment and mathematics (Kiczuk, 2001, p. 148).

The use of these two types of language necessitates using two types of logic to control physical reasoning. On the side of the mathematical apparatus, the language of classical propositional calculus is used, while on the side of an imaginary language, non-classical logics are employed to define the concepts and guide reasoning.

It has been observed that various non-classical logics can enhance our understanding of physical theories expressed in an imaginary language. These include temporal logics (particularly tense logics, which rigorously express different properties of time, as well as interval temporal logics or chronologies), causal logics, and various logics of counterfactual conditionals (often used to express physical laws). Additionally, efforts have been made to develop specific non-classical logics that have operators designed to capture the unique phenomena observed in physics (e.g., quantum mechanics logics) or to adapt existing systems (such as many-valued logics) for this purpose.

Applications of logic in the humanities

To discuss the application of logic in the humanities, it is essential first to define what is meant by the humanities. In the most general terms, the humanities are considered a set of disciplines dealing with the various manifestations and products of human psycho-spiritual life (Kamiński, 1992). These disciplines can be categorised into three main groups: disciplines focused on individuals and society (such as psychology, cultural anthropology, and socio-economic sciences); disciplines concerned with cultural products (including legal studies, the study of religion and morality, and the arts and languages); disciplines examining the history of human society and its products (the historical sciences). An additional consideration is whether philosophy should be classified as a human science (as suggested by the division of scientific disciplines adopted in Poland). Philosophy is deeply intertwined with logic, which may not necessarily be the case in other non-philosophical humanities.

Logic in philosophy

The role of logic in philosophy is pivotal. Philosophical discourse often serves as a model for scientific discourse, but there is a great diversity of ways in which philosophy is practiced, including the permissible sources of cognition and the methods of philosophical explanation and justification. The significance of logic in explaining and justifying philosophical claims depends on the philosophical approach being employed. In systematic philosophy, for example, deductive reasoning and reliable inference schemes are fundamental. Philosophical propositions are often the logical (or substantive) consequences of base statements that are directly justified, such as those based on intellectual intuition. Thus, philosophical propositions are indirectly justified, relying on argumentation where schemas of reliable reasoning play a fundamental role (Janeczek, 2018). Philosophical discourse typically employs natural language, though it may also be technical. Argumentation is also relevant to cognitive science as it pertains to the human as a corporeal-mental being. Moreover, the development of algorithms capable of argumentation is a key area in artificial intelligence. Formal logic is instrumental

in the formal reconstruction of philosophical reasoning, aiding in the uncovering of implicit assumptions within philosophical systems and assessing the formal validity of reconstructed reasoning. Such analysis sometimes reveals the need for non-classical logics or demonstrates that specific logical tools with particular inferential strengths are required to justify certain propositions.

Some 20th-century philosophers advocate for limiting the cognitive ambitions of philosophy, emphasising that it should focus on well-founded conceptual analysis (analytic philosophy, Bocheński, 1993). Analytic philosophers use the tools of logic to conduct linguistic analyses aimed at achieving simplicity and clarity. This approach explores how words in natural language are used to capture their meanings, based on the idea that language reflects reality. The core of conceptual analysis involves identifying the constitutive (essential) and accidental properties of a concept and understanding how it (as an abstract entity) relates to other concepts. To achieve this, philosophers analyse the uses of the relevant word in natural language, supplemented by a regulative element, i.e., the constructive moment (Łukasiewicz, 1961). Due to the logical nature of its analyses, philosophy takes the form of applied logic. In such analyses, logic is considered the most general ontology, "a theory of the most general and simplest relationships between all things in general, and therefore an excellent tool not so much for reasoning as for analysis" (Bocheński, 1993, p. 191). This understanding of logical analysis aims to define the analysed concept and capture its subtypes and the relationships between them. As such, it can become a tool for use in the humanities.

Conceptual analysis often involves referencing the meaning of the analysed expression in natural language; organising intuitions related to the usage of the expression by distinguishing different meanings of the word and different contexts in which it is used; characterising the nearest genre of the defined concept; determining the formal properties of the concept, including its attributes and relations; introducing a classification of the concept based on a principle of division; analysing individual members of the division by identifying their essence and properties (Bocheński, 1993). The traditional logical tools used in this analysis include categorisation of concepts (the domain of reality being analysed), definition, and logical division. These actions are often

supported by symbolisation (rather than formalisation) and the use of first-order logical tools. The results of Bocheński's application of these relatively simple methods have been noted in the literature (e.g., works on freedom, authority, industrial enterprise, etc.).

Applications of logic in the specific human sciences

Often, when seeking to identify the features that distinguish the humanities from the natural sciences, attention is drawn to the former's distinctly cultural subject matter, characterised by two features: firstly, culture consists of actions and products of the human psyche (considered from both individual and social perspectives), whereas the reality examined by the natural sciences exists independently of such operations; secondly, culture has a value aspect associated with human actions and creations. This cultural dimension shapes the understanding of the subject matter of the humanities as inherently individual – as cultural phenomena, through their association with the activities of specific individuals and social groups, are marked by individuality, uniqueness, and temporality.

This individuality of the subject matter entails a number of logical difficulties in its characterisation. These difficulties may include:

- the logical peculiarities of humanistic concepts (vagueness, imprecision, and ambiguity), and challenges in defining them in a typical way (e.g., typological concepts, concepts with families of meanings, and the persuasive nature of concepts); the need to use partial definitions or persuasive definitions (Pawłowski, 1986);
- a different type of explaining (humanistic interpretation).

Logical problems of defining in the humanities

Due to the specificity of concepts in the humanities, the logical theory of definition plays a special role in explicating these concepts, whose particular feature is that they are predominantly typological or involve families of meanings (Pawłowski, 1986).

In the case of typological concepts, the boundaries of their scopes are often blurred. These concepts, which frequently refer to gradable

attributes (e.g., intelligent, diligent), tend to have an ordering rather than a strict classifying nature. In the humanities, typological concepts are most often defined based on empirically given orderings. For example, in typology we use orderings based on extreme (ideal) types, indicating the minimum and maximum points on a scale of gradable attributes.

Concepts with families of meanings (e.g., art, avant-garde, novel, tragedy) are characterised by the fact that their scope is not defined by a set of objects sharing common features, but rather by a series of subsets of objects connected by partial similarities. This family of subsets corresponds to a family of meanings that constitutes the meaning of a given concept. It has been pointed out (Wittgenstein) that it is impossible to define such concepts fully; however, this position should be modified: these concepts cannot be given an equivalence definition (*per genus proximum et differentiam specificam*), but they can be defined through a set of partial definitions which provide partial criteria for the applicability of the concept (by indicating necessary or sufficient conditions). A characteristic feature of concepts that can only be partially defined is that no number of partial definitions can provide a complete definition, and the scope of such a concept remains open (meaning that there will always be objects for which it remains uncertain whether they fall within the scope of the concept.

Humanistic interpretation

The unique nature of the subject matter in the humanities necessitates a different mode of explanation compared to the natural sciences. Ajdukiewicz defines explanation in the natural sciences as a process where, for a given accepted proposition, we seek arguments either solely among accepted propositions (universalising explanation) or among both accepted and non-accepted propositions (theoretical explanation). In reasoning that is termed explanation, the direction of reasoning is reversed compared to that in logical implication because the conclusion is the sought-after reason for the stated state of affairs. The proposition obtained at the conclusion of the explanation is not certain but merely probable, as explanation proceeds according to a reductive inference scheme (from consequences to reasons). In the humanities, however,

explanation often involves statements directly based on understanding someone's utterances or behaviour, or the existence of products of someone's activity. The logical analysis of the method of explanation in the humanities has led to the concept of humanistic interpretation. Broadly speaking, this is the activity of formulating answers to questions such as: "Why did *x* undertake such and such an action?", or "Why did *x* give his product such and such a form?" The answer consists of the following: 1) an assumption of *x*'s rationality; 2) a description of *x*'s knowledge that allows him to identify possible actions he might take and their results; 3) a description of *x*'s order of values derived from these results, such that the result of the considered action is a preferred value. Fundamental to this type of explanation is the assumption of the rationality of the acting subject (for actions taken under conditions of certainty), which states that:

> If *x* is to choose one of the actions $c_1, ..., c_n$, which, based on his knowledge, are mutually exclusive and exhaustive and reliably lead to results $s_1,..., s_m$, and these results are ordered by a value preference relation characteristic of *x*, then *x* will choose action c_i associated with the most preferred result s_j (Kmita, 1971).

Of course, as the authors emphasise, humanistic interpretation is not the only method of explanation used in the humanities (an example of others is genetic explanation). However, it does allow for capturing, to some extent, the specificity of the subject matter of the humanities.

Non-classical logics and formal ontologies

The subject matter of the humanities, due to its high degree of complexity, is difficult to capture precisely as most concepts 'escape' standard definitions and require researchers to engage in various cognitive efforts. Against this backdrop, the possibility of using different types of non-classical logics arises. These logics can be viewed as definitions by postulates for various concepts related to human cognition (e.g., epistemic logics), man's actions (e.g., logics of norms or deontic logics), or choices (e.g., logics of preferences).

Examples include logics of knowledge and belief, formal theories of belief change, or dynamic logics of belief (Lechniak, 2011). These are

typically systems of standard modal logics, where modal operators are interpreted as phrases like '*x* is convinced that', '*x* accepts that', '*x* knows that', or 'in the face of knowledge of a state of affairs *p*, *x* changes his belief', etc. For instance, the successive axioms of modal logic, when interpreted epistemically, can be seen as theses about properties of knowledge (e.g., the truthfulness of knowledge, the non-contradictory nature of knowledge, infallibility regarding one's own knowledge, completeness concerning one's own knowledge, etc.). Combining different systems of non-classical logics, such as the logic of knowledge with the logic of belief or the logic of belief with dynamic (change) logic, also yields interesting results. Although each system is based on simple axioms, the integration of concepts described by the axioms of one system with those of another reveals the complexity of concepts related to human knowledge and beliefs (e.g., logics of public announcements, where a subject's (or group of subjects') belief is derived from information obtained from the subject's perceptions and information received from other subjects).

Formal ontologies, like non-classical logics, can make a valuable contribution to the analysis of concepts in the human sciences, particularly in cognitive science. These are formal systems (of axioms and definitions) often built upon classical logic – possibly enriched by additional theories such as mereology (i.e., the theory of sets in a collective sense) – aimed at formally characterising concepts within a specific subject domain. Models derived in this way can serve as a conceptualisation, expressed in formal language, of a knowledge domain that forms the basis for constructing a knowledge base. Unlike formal logic systems, which specify consequence relations for a set of propositions, ontologies aim to provide a possibly adequate formal model of the domain. Therefore, ontologies may consist of dozens of axioms and definitions (as opposed to logic systems, which usually have a small number of axioms). Examples include ontologies of beliefs, ontologies of historical objects, and so on. It has also been proposed (e.g., by Jerzy Perzanowski) to apply formal ontologies in cognitive science (psychoontologies).

Systematic reflection with conclusions and recommendations

Specialised language and knowledge-generating operations, such as the explanation and justification of claims, are crucial in any scientific discipline. Logic, serving as both a theory of language for the effective formulation and communication of ideas and as a formal theory of reasoning, should play a fundamental role in providing internal principles for organising scientific theories (logical calculus as a basis for scientific reasoning and argumentation) and in serving as an external basis (logic as a tool for meta-reflection on scientific theory) in every scientific discipline. This dual role of logic is applicable across both the sciences and the humanities, albeit in different ways. Cognitive scientists search for a methodology that integrates these two fields of study. The task of logic/logics is to refine the principles that distinguish correct reasoning from incorrect reasoning, which cognitive science places within the realm of reasoning representation. Information processing in the human mind and its underlying nervous system includes reasoning; in this context, logic contributes to cognitive science.

In the sciences, propositional calculus plays a predominant role as the foundation for mathematical concepts and reasoning. The logical theory of reasoning is also important in the methodology of these sciences as it facilitates the creation of models for scientific theories and the description of research procedures used in these fields. Non-classical logics provide tools for the formalisation of reasoning in the notional language of physics.

In the humanities, logical tools of analysis make it possible to capture the methodological specificity of these sciences, where typological concepts or concepts with families of meanings are prevalent, and explanations must account for the individuality of the subject matter while assuming the rationality of human acting or creating subjects. The role of non-classical logics (which provide tools for analysing concepts in the humanities) and formal ontologies (which benefit from the achievements of both classical and non-classical logics) is significant.

Logic plays a particularly crucial role in cognitive science, providing models for human cognitive activities (e.g., non-classical logics such as dynamic logics, logics of knowledge and belief, logics of action, and

deontic logics). These models allow for formal reconstructions of cognitive activities or models of the mind.

It should be noted that, in practice, interest in non-classical logics among physicists has been relatively low, whereas such applications are widespread in the philosophy of the natural sciences. On the other hand, the use of logical methods is widespread in the methodology of the humanities, and some ideas from logic (e.g., the semantics of possible worlds) have gained significant attention among humanists who are not logicians due to the multitude and diversity of philosophical interpretations of the concept of possible worlds.

References

Bocheński, I.M. (1961). *A History of Formal Logic*. Notre Dame: University of Notre Dame Press.
Bocheński, I.M. (1993). *Logika i filozofia. Wybór pism*. Warszawa: PWN.
Borkowski, L. (1970). *Logika formalna*. Warszawa: PWN.
Grzegorczyk, A. (1981). *Zarys logiki matematycznej*. Warszawa: PWN.
Hughes, G.E., & Cresswell, M.J. (1968). *An Introduction to Modal Logic*. London: Methuen Publ.
Janeczek S., Starościc, A., & Tkaczyk, M. (ed.) (2018). *Logika*. Vol. I. *Natura logiki*. Lublin: Wydawnictwo KUL.
Kamiński, S. (1992). *Nauka i metoda. Pojęcie nauki i klasyfikacja nauk*. Lublin: TN KUL.
Kiczuk, S. (2001). *Przedmiot logiki formalnej oraz jej stosowalność*. Lublin: Wydawnictwo KUL.
Kmita, J. (1971). *Z metodologicznych problemów interpretacji humanistycznej*. Warszawa: PWN.
Lechniak, M. (2011). *Przekonania i zmiana przekonań*. Lublin: Wydawnictwo KUL.
Łukasiewicz, J. (1961). *Z zagadnień logiki i filozofii*. Warszawa: PWN.
Marciszewski, W. (ed.) (1987). *Logika formalna. Zarys encyklopedyczny*. Warszawa: PWN.
Pawłowski, T. (1986). *Tworzenie pojęć w naukach humanistycznych*. Warszawa: PWN.
Świrydowicz, K. (2004). *Podstawy logiki modalnej*. Poznań: Wyd. UAM.
von Wright, G.H. (1951). *An Essay in Modal Logic*. Amsterdam: North Holland.

Anita Pacholik-Żuromska
Nicolaus Copernicus University in Toruń
https://orcid.org/0000-0002-8674-9412
https://doi.org/10.35765/slowniki.424en

Nonmonotonic logic in mind modelling

Summary

DEFINITION OF THE TERM: Nonmonotonic logic departs from classical rules of inference by not adhering to the monotonicity principle, which states that if a sentence follows from a given set of premises, it must also follow from the union of this set with any other set. The role of nonmonotonic logic in models of the mind is to reflect the way people think. This logic is a useful tool for creating formal models of the mind because it adequately captures the nuances of everyday reasoning, such as the omission of certain premises or the introduction of new assumptions that may contradict earlier ones.

HISTORICAL ANALYSIS OF THE TERM: Since the emergence of contemporary philosophy of mind in the second half of the 20th century, models of the mind have primarily been grounded in computability. However, classical logic alone proved insufficient to fully describe the mechanisms underlying complex mental processes. As a result, analogue information processing gave way to distributed processing, and algorithms designed to simulate human cognitive abilities, such as learning, became increasingly sophisticated. Today, predictive coding is undergoing a revival, often in combination with nonmonotonic logic.

DISCUSSION OF THE TERM: Nonmonotonic logic can be used to model both neuronal information processing and the formation of propositional attitudes. This is because nonmonotonic inferences within a neuronal network are patterns defined in terms of neural activity vectors, while they can also be treated as propositional structures because they are linked through nonmonotonic consequence relations.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: The key advantage of using nonmonotonic logic in modelling the mind lies in its ability to facilitate the creation of formal mechanisms required for the accumulation of conceptual (i.e., propositional) knowledge. Such models illustrate how a subject's dynamic belief network is formed, particularly in processes like metacognition.

Keywords: mind model, nonmonotonic logic, representations, prediction, uncertainty

Definition of the term

Nonmonotonic logics are a type of logic in which reasoning does not proceed in the classical manner, characterised by the principle of monotonicity, which states that if a sentence follows from a certain set of premises, it also follows from the union of that set with any additional set (Makinson, 2008). An inference is considered nonmonotonic when a conclusion *p* follows from a set *Z* but does not follow from a superset of *Z* (Makinson, 2005). Unlike deductive reasoning, nonmonotonic inferences are drawn on the basis of partial information (Moore, 1985), intuition, and expectations (Gärdenfors & Makinson, 1994). Moreover, new information can lead to a change in earlier conclusions, while retaining the previous premises and without falling into contradiction.

Models of the mind, however, are a deeply interdisciplinary concern, drawing on fields such as philosophy, psychology, logic, neuroscience, and linguistics. These models are a classic problem in cognitive science, i.e., the interdisciplinary (or transdisciplinary) study of cognition. An example of the practical application of modelling cognitive processes with formal tools – algorithms – is the deployment of artificial intelligence in autonomous vehicles.

The role of nonmonotonic logic in models of the mind is to accurately reflect how information is processed, particularly in decision-making, where flexible thinking, accounting for incomplete information, perspective shifting, and reconciling conflicting beliefs are essential. The goal is for such a model to genuinely mirror the workings of the human mind.

Earlier models of the mind also incorporated algorithms, although they were based on classical logic. These models emerged relatively late, at a time when algorithms used in computing and cryptographic machines were already established. Even before this, models of the mind often drew inspiration from contemporary technical advances. Historically, the mind was frequently compared to mechanical systems, from the pneumatic toys of ancient Greece to the clockwork mechanisms that influenced Leibniz's philosophical reflections on the mind.

Historical analysis of the term

The modern history of models of the mind began in 1948 with the first interdisciplinary symposium on "Cerebral Mechanisms and Behavior", which featured prominent figures like John von Neumann, Paul Weiss, and Warren McCulloch – pioneers in the field of artificial intelligence. The early models of the mind were grounded in the computational paradigm of cognitive science, which is based on automata theory, i.e., mathematical models of devices that process discrete information. Automata are generally classified into two main types: acceptors (recognisers) and transducers. They are designed to recognise languages and compute partial functions over the set Σ^*, where Σ and Δ are finite alphabets, and Σ^* and Δ^* are the sets of all possible words over the relevant alphabets (cf. Marciszewski, 1987, p. 200).

Since the emergence of contemporary philosophy of mind (or analytic metaphysics of mind) and cognitive science in the second half of the 20th century, three main computability-based models have been developed. The first, based on the computer metaphor, posits that the mind functions much like a computer program – or even a literal computer – where mental processes are computational and operate according to algorithms. This approach gave rise to Jerry A. Fodor's Representational Theory of Mind, which postulates that the content of mental states is determined by the internal relationships between mental representations, which serve as carriers of information. Computations are performed on these representations, often referred to as the 'symbols' of the language of thought (Fodor, 1992). Fodor's understanding of computability was closely aligned with classical logic, specifically the classical propositional calculus and predicate calculus, which he believed underpinned rational reasoning.

However, it soon became evident that the mind's complexity exceeded what classical logic could capture, particularly in describing the mechanisms behind complex mental processes. Consider the phenomenon of akrasia, or weakness of will, where individuals hold contradictory beliefs about what they should do and what they want to do, such as knowing that smoking is harmful and should be given up immediately, yet simultaneously lacking the desire to quit. Models of reasoning that depicted the subject as simply moving from one true belief to another, as

Fodor (1992) suggested, were overly idealised. In response, the brain metaphor emerged, shifting the perspective from modelling the natural mind after a computer to understanding the brain's functioning through simulations on artificial neural networks. This shift marked the transition from the computer metaphor to connectionism. Unlike the serial information processing of earlier models, connectionism emphasised distributed information processing, which better aligned with the brain's network-like structure, where "almost every part of the brain is being 'watched' by some other part of the brain, often by several other parts at once" (Churchland, 1996, p. 17). As a result, artificial neural networks (ANNs) were developed to simulate mental processes, such as face recognition. In a connectionist architecture, information processing occurs as patterns of neuronal activation in the sensory organs are monitored by successive layers of neurons. These layers transform the information and pass it along to other brain structures (Churchland, 1996). The brain encodes information as activation patterns that can be strengthened through, for example, cognitive training. Connectionist models enable the definition of learning procedures, allowing them to adapt to the environment.

However, while connectionist models applied to artificial neural networks offered valuable insights, they fell short of fully explaining how the mind works. One significant limitation was their insufficient consideration of the role that the body and environment play in cognition. To address these gaps, models began to emerge within the framework of 4E cognition. According to this perspective, the mind is embodied, extended, embedded (situated within the environment), and cognition is enactive (i.e., arises from direct interaction with the environment). By emphasising the role of the body in cognition, 4E theories draw on the phenomenological tradition.

This shift in understanding raised the question of how to model cognition involving not just the brain but the entire organism interacting with its environment. Explanations rooted in statistics and probability theory seemed best suited for this purpose. However, theories based on these approaches reignited the longstanding internalism-externalism debate in the philosophy of mind, echoing earlier disputes, such as those between Fodor and Putnam (Fodor, 1992). This debate revisits an even older question about the direction of cognition: does it proceed from the mind

to the world or from the world to the mind? According to 4E theories, not only cognition and meaning are extended but also representations themselves, as stimuli from the world are integral components of mental representations, not merely their activators.

Models within the 4E theories are characterised by two key features: the incorporation of environmental feedback; the requirement for representations to be dynamic or, in some cases, the complete rejection of representations. To move beyond the traditional notion of representation, some researchers turned to cybernetics, adopting concepts such as invariants to replace the role of representations within the brain-body-environment interaction loop (Seth, 2015, p. 15).

The unifying element for models that incorporate feedback, preserve representations, and address the workings of the brain – thus bridging connectionism and the 4E theories – turned out to be predictive coding. This computational model, grounded in Bayesian inference theory and the free energy principle (Friston, 2009), offers a mechanistic explanation of the mind's functioning. According to Bayesian inference, the probability of a given hypothesis is updated as new data becomes available. This is a posteriori probability, which aligns with the intuitive idea that "the result of an observation confirms a hypothesis when it is highly probable on its grounds and unlikely otherwise" (Marciszewski, 1987, p. 412). In the Bayesian brain, experience (feedback from the world) is compared with prior predictions (hypotheses) about expected situations. These hypotheses can be interpreted as representations, which are models of the anticipated (expected) state of the world. Predictive coding is used, for example, to formalise unsupervised learning in artificial neural networks, enabling systems to self-organise in relation to both their components and their environment. This mechanism has been extended to natural cognitive systems to explain how living organisms process information. In this framework, self-organisation is an ongoing process of adjusting mental representations to align with the world, aiming to minimise predictive error and optimise action. Perception and attention are two distinct but correlated aspects of the brain's fundamental mechanism for minimising predictive errors. Perceptions result from the brain's neuronal model of the world (this model reflects how the brain 'sees' reality), where expectations and predictions regarding the states of the world are compared with stimuli from the environment.

While Bayesian inference is effective at adjusting preconceived beliefs, it faces challenges in situations where the subject lacks complete knowledge of the world. Hence, researchers are seeking better formal tools without excluding or dismissing predictive coding. One promising direction involves models of information processing based on nonmonotonic logic.

Discussion of the term

Researchers generally assume that human reasoning is nonmonotonic (Malinowski, 1997). Bayesian reasoning, although it handles the modification of previously accepted beliefs, encounters difficulties in contexts where the subject does not have complete knowledge of the world–contexts that closely resemble real-life scenarios (Mota, Sridharan, & Leonardis, 2021). Additionally, in human–robot interactions, combining predictive processing with nonmonotonic logic has proven effective in programming accurate decision-making and improving the speed and precision of responses (Rilay & Sridharan, 2019).

Nonmonotonic logic can be defined as formal structures designed to capture defeasible reasoning, i.e., typically human reasoning that is rational but not strictly deductive. Unlike deductive reasoning, which is based on complete information and follows strict logical rules, nonmonotonic reasoning operates on partial information (Moore, 1985), intuition, and expectations (Gärdenfors & Makinson, 1994). This approach allows for the integration of new information without discarding previous premises or introducing contradictions. For instance, adding new premises can alter the conclusions drawn, reflecting a key aspect of everyday reasoning. Moreover, new information can revise previous inferences without leading to contradictions. Specific extensions, such as adding new premises, make logic nonmonotonic. This is frequent in colloquial reasoning:

Most birds can fly.
An ostrich is a bird.
Therefore, an ostrich can fly.

However, the conclusion in this case is false because ostriches cannot fly. Nonmonotonic rules of inference allow us to resolve the error created by generalisation:

An ostrich is a bird, so an ostrich can fly.
$\{p\} \vdash z$
An ostrich is a bird. But an ostrich cannot fly.
$\{p, q\} \nvdash z$

The classical consequence relation ($\vdash$) is stronger than the nonmonotonic consequence relation (|~). As an operation, the classical consequence relation, denoted as Cn, functions as a closure operation (Makinson, 2005, p. 4). This means it fulfils the following conditions of deductive inference (Malinowski, 1997):

Reflexivity (inclusion): $A \subseteq Cn(A)$
(Set A is a subset of the set of all consequences of set A).
Idempotence: $CnCn(A) = Cn(A)$
(The set of logical consequences of the consequences of A is equal to the set of consequences of A).
Monotonicity: Always when $A \subseteq B$, then $Cn(A) \subseteq Cn(B)$
(A consequence of a given set of premises is also a consequence of any extension of that set).

In contrast, nonmonotonic logic does not satisfy the monotonicity condition. In nonmonotonic logic, it is not necessarily the case that if a sentence follows from a set of premises, it will also follow when additional premises are added:

$x \in Cn(A) \wedge x \notin A \cup B$

where x is a consequence of the set A of sentences but is not a consequence of superset $A \cup B$ (Makinson, 2005, p. 2; 2005a, p. 4).

Usual everyday reasoning is not always deductive and is accompanied by the weaker relation of consequence. People are not Turing machines, and their judgments and opinions often blur the line between deduction and induction – a feature inherent to nonmonotonic reasoning. While classical consequence effectively handles deductive reasoning involving semantics and syntax, it falls short in contexts involving pragmatics, such as communicative acts, where intentions, beliefs, and psychological factors play a crucial role.

Nonmonotonic logic has demonstrated its effectiveness in modelling concrete cognitive actions implemented within artificial systems, such as robotic software (Riley & Sridharan, 2019; Mota, Sridharan, & Leonardis, 2021). These systems utilise nonmonotonic logic to facilitate decision-making, predict behaviour, model learning processes, and program responses, beliefs, and experiences (Riley & Sridharan, 2019). The advantage of using nonmonotonic logic to model the mind lies in its ability to support the creation of formal mechanisms for the accumulation of conceptual (i.e., propositional) knowledge. Models of creating this knowledge based on nonmonotonic logic illustrate how expanding a set of premises, i.e., incorporating new knowledge, changes its consequence. Thus, conclusions are drawn based on the available information in a way that allows them to be revised with the introduction of new information without discarding any previous premises (cf. Makinson, 2005; Makinson, 1994).

Nonmonotonic logic can be used for modelling both neuronal information processing and the formation of propositional attitudes (Balkenius & Gärdenfors, 1991). At the intentional (i.e., propositional) level, nonmonotonic logic can model the exchange of beliefs in conversations because it addresses both symbolic and subsymbolic levels of information processing. This means it can simultaneously describe neural activity vectors and the propositional forms of beliefs (Balkenius & Gärdenfors, 1991, p. 37; Blutner, 2004). This is possible because, although nonmonotonic inferences in neural networks are schemes defined in terms of neural activity patterns, they can also be treated as propositional structures because they are connected through the nonmonotonic consequence relation |~ (Balkenius & Gärdenfors, 1991, p. 36). Nonmonotonic logic effectively bridges the gap between different modes of computation: symbolic models (such as Fodor's Representational Theory of Mind) and subsymbolic (neuron-like) models. As such, it can be used to model the behaviour of artificial neural networks because the actions of connectionist networks can be interpreted as nonmonotonic inferences. In other words, nonmonotonic logic can serve as both a descriptive and analytical tool for examining the emergent properties of connectionist networks.

Balkenius and Gärdenfors (1991) demonstrated that it is possible to define a nonmonotonic inference relation that reflects asymptotic updates of information states.

Let <*S,* ≥> be a partially ordered set of activation states, and *w* a connection matrix. Then, the notion of asymptotic updates naturally leads to a nonmonotonic inference relation between information states:

$$s \mid\sim_w t \text{ if } s' \geq t \text{ for every } s' \in ASUP_w(s)$$

Nonmonotonic logic asymptotically describes how neural activity propagates through a connectionist network. The correspondence between the connection matrix and the weight-annotated system of Poole (1988) involves the application of a translation mechanism in which node *i* is transformed into an atomic symbol p*i*. An activating connection in the network is transformed into a two-way logical equivalence ↔, and an inhibiting connection is transformed into a logical equivalence ↔ with a negation applied to one of its arguments. Implicit weights should be treated as the absolute value of the corresponding matrix elements (Blutner, 2004). When nonmonotonic logic is applied, there is a strict correspondence between the knowledge representation in weight-annotated Poole systems – used in various artificial intelligence applications and employing default logic (e.g., the premise “Birds can fly”) – and the coding of knowledge in Hopfield networks, which are recurrent networks used in connectionist models of associative memory.

As mentioned above, while predictive coding is a robust formal tool for modelling cognitive functions and abilities, it is insufficient for fully capturing human reasoning. On the one hand, predictive coding effectively formalises the relationship between internal representation and the external world, explaining how the mechanism of correcting representations and adapting behaviours to the environment operates at the neural level (Chanes & Barrett, 2016). It should also be noted that predictive coding is used to model conversational situations, such as determining the order of turn-taking. This can be explained by a functional prediction mechanism that enables a role change from speaker to listener (Riest, Jorschick & de Ruiter, 2015). On the other hand, nonmonotonic logic is suitable for modelling predictions based on hypotheses or expectations about the future state of the environment (Gärdenfors & Makinson, 1994). However, the problem lies not merely in capturing predictability but also in dealing with uncertainty (Talamadupula et al., 2010).

The term ‘uncertainty’ is often used incorrectly to describe a situation where we do not know exactly what will happen when we take action or make other changes. In fact, this situation is more accurately termed ‘ambiguity’, which is also known as ‘risk’ in economics and probability theory. Uncertainty arises when we do not know all the alternatives, outcomes, or probabilities. The world where no one has complete knowledge of a system is referred to as the Open World, while the world in which the knowledge of a system is complete is termed the Closed World. “The Closed World Assumption (CWA) is the assumption that what is not known to be true must be false. The Open World Assumption (OWA) is the opposite. In other words, it assumes that what is not known to be true is simply unknown” (Sequeda 2012). Therefore, appropriate types of Open World Semantics are needed to represent partial knowledge; researchers use heuristics to develop them but they may not employ the correct formal models in the process. Although much effort has been devoted to adapting Bayesian inference methods to handle uncertainty (Riley & Sridharan 2019), this approach is inadequate for the task. According to Riley and Sridharan, “modeling uncertainty by attaching probabilities to logic statements is not always meaningful” (Riley & Sridharan 2019, p. 4). Therefore, to create action models used for reasoning or learning in dynamic environments, nonmonotonic logical reasoning needs to be combined with relational reinforcement learning and inductive learning (Riley & Sridharan 2019, p. 4).

Due to its nature, nonmonotonic logic enables challenging preconceived assumptions because of its inherent flexibility and has considerable potential for coding and performing the type of reasoning needed to cope with uncertainty, especially with some training (Mota, Sridharan & Leonardis, 2021). Predictive coding based on Bayesian inference is insufficient for creating an architecture that adequately describes human reasoning. This is not only because it does not address uncertainty, but also because it is problematic in the context of induction (Kyburg, 2001). Nonmonotonic logic performs well at encoding complex expressions and subtle differences, such as the difference between “¬ a” (which implies that “a is believed to be false”) and “not a” (which only implies only that “a is not believed to be true”) (cf. Mota, Sridharan & Leonardis 2021, p. 242). This distinction makes it possible to differentiate between negation and an epistemic alternative. Adding new premises can reduce the

set of consequences, which helps in recovering from inferential errors more quickly (Mota, Sridharan & Leonardis 2021, p. 242).

Systematic reflection with conclusions and recommendations

Examples of practical applications of nonmonotonic logic in artificial neural networks, including their mutual integration, include the development of systems capable of reliably and efficiently recognising and responding to various stimuli, such as different traffic signs (Riley & Sridharan 2019). This capability is crucial for applications such as programming autonomous vehicles. An attempt to transfer the mechanism of artificial neural networks to the mechanisms of the mind aligns with the principles of connectionism.

Given that nonmonotonic logic is particularly effective for modelling propositional knowledge, it can be concluded that it is also a powerful tool for modelling metacognition, especially propositional metacognition, i.e., self-knowledge understood as the subject's first-person knowledge about his mental state, to which he has privileged access. Examples include judgements such as "I am convinced that I am convinced that spring begins on March 21". In other words, these are first-person accounts of, e.g., one's beliefs, desires, decisions – beliefs about beliefs. In this sense, self-knowledge represents second-order beliefs that constitute the subject's understanding of the contents of his mind.

Researchers began to reconsider the idea of modelling specific cognitive actions using nonmonotonic logic because they wanted to implement a form of self-knowledge in robots, or in this case, an "ability to provide explanatory descriptions of their decisions" (Mota, Sridharan & Leonardis, 2021, p. 241). Although inferences about the world are Bayesian – predictive processing can be a tool for modelling first-order beliefs – second-order beliefs (i.e., beliefs about beliefs that are characteristic of self-knowledge) must be modelled using nonmonotonic logic because it "helps recover from errors caused by reasoning with incomplete domain knowledge" (Mota, Sridharan & Leonardis 2021, pp. 242–243). Probabilistic elements do not allow efficient inferences to be made within common-sense knowledge (Riley & Sridharan 2019).

The application of nonmonotonic logic to modelling self-knowledge is not a new concept. For instance, it was used in autoepistemic logic to create a model of a perfectly rational subject who reasons about his own beliefs (Moore, 1985). This demonstrates that nonmonotonic logic, which can represent knowledge, is also well suited to modelling self-knowledge. Indeed, self-knowledge often involves scenarios where the subject holds seemingly contradictory beliefs, such as believing that smoking is detrimental to health while also having no desire to quit smoking. The issue of incoherent beliefs in self-knowledge is linked to how their contents are determined because a subject's self-perception can be influenced by external factors from the social environment. A change in the environment or the group of which the subject belongs can have a profound effect on his beliefs. Sometimes, as a result of such a change, a person may adopt entirely new beliefs, which in turn can alter their self-view. This phenomenon in itself is nothing new, as individuals often undergo changes as they adapt to new life situations. The interesting question is how we maintain psychological coherence and a sense of identity despite these changes. In the course of life, we adapt to new conditions and challenges. To maintain our sense of 'being ourselves', we shape our own narrative according to our evolving beliefs, which may change due to various factors. These changes in our beliefs affect our self-perception. We strive to integrate new beliefs in a way that maintains coherence, although we may not always be able to justify precisely why we accept some beliefs and reject others.

Przymusiński (1997) proposed a method for unifying various nonmonotonic formalisms by integrating nonmonotonic logic with decision trees, offering an alternative to predictive coding. This approach exemplifies inductive learning methods, including reinforcement learning. To manage contradictory beliefs and paradoxes, nonmonotonic logic incorporates practicality as a component of rationality (cf. Makinson, 2005, pp. 129–130). By applying nonmonotonic logic to the architecture of the mind, particularly regarding propositional attitudes, we can model how a dynamic network of beliefs evolves. In this model, a change in one belief can trigger adjustments across other beliefs, and the introduction of a new belief may influence the entire network.

The linguistic form of self-knowledge expressed as first-person accounts such as "I am convinced that I am convinced that *p*" consists

of two basic representations: one of the Self and one of the experienced state. These are linked by an intentionality relation, which captures the subject's attitude towards a given judgment about himself. Given its representational nature, self-knowledge can be computationally modelled (cf. Lewis, 2015) and implemented in artificial neural networks using nonmonotonic logic.

Nonmonotonic logic also proves valuable in modelling psychological disorders and addressing issues such as self-illusions and false beliefs, particularly when individuals reflect on their own beliefs and encounter contradictions or 'bizarre' notions (Moore, 1985). For instance, schizophrenic patients may recount experiences from psychotic episodes, recognising them as hallucinations yet still firmly believing in the reality of these events. In other words, they may resist acknowledging the fallibility of their inferences. Nonmonotonic logic can account for this phenomenon by allowing for the possibility that a subject may hold beliefs that deviate from his initial premises (Moore, 1985). Thus, models of self-knowledge based on nonmonotonic logic provide insight into why individuals may misjudge their own mental states while still preserving the consistency of the beliefs that constitute their self-knowledge.

If we consider self-knowledge as a form of inner speech due to its formulation in the form of judgements, then nonmonotonic logic is particularly well-suited for its formalisation. Since nonmonotonic logic, particularly in its properly supraclassical parts, is not subject to the substitution rule, it is possible to draw conclusions that extend beyond classical logic frameworks (cf. Jarmużek, 2007). This properly enables the recognition of seemingly contradictory judgements and the continuous updating of self-knowledge. "This property is in contrast to classical logic, whose inferences, being valid, can never be "undone" by new information" (Jarmużek, 2007, p. 259). Importantly, new information often comes from external sources during interactions, influencing how we adjust our self-knowledge. Therefore, in models of mind and cognition it is necessary to move beyond the predictive coding paradigm, which supports internalist models of mind, particularly when addressing issues related to artificial intelligence.

In summary, nonmonotonic logic provides a powerful framework for modelling propositional attitudes within the architecture of the mind, as conceptualised by 'folk psychology' (Fodor, 1992). It illustrates how

a dynamic network of beliefs is structured, where changing one belief can prompt changes in others, and introducing a new belief that can impact the entire network. Nonmonotonic logic is particularly effective in developing formal models of the mind, including metacognition, because it captures the complexities and nuances of belief changes, such as the omission of premises or the introduction of new assumptions that contradict previous ones, which also actually occurs in the process of belief change. Nonmonotonic inferences reflect how people reason, make decisions, and revise their knowledge. From the perspective of human-robot interaction, models based on nonmonotonic logic enable artificial systems to manage uncertainty in the Open World, where many factors remain unknown due to incomplete information.

References

Balkenius, C., & Gärdenfors, P. (1991). Nonmonotonic Inferences in Neural Networks. In: J. Allen, R. Fikes, & E. Sandewall (ed.), *Principles of Knowledge Representation and Reasoning. Proceedings of the Second International Conference (KR91)* (pp. 32–39). California: Morgan Kaufmann Publishers.

Blutner, R. (2004). Nonmonotonic Inferences and Neural Networks. In: W. van der Hoek (Red.), *Information, Interaction and Agency* (pp. 203–234). Berlin: Springer, Dordrecht. DOI: 10.1007/1-4020-4094-6_7.

Churchland, P.M. (1996). The *Engine of Reason*, *the Seat of the Soul*. A *Philosophical Journey into the Brain*. Cambridge, Massachusetts: MIT Press.

Fodor, J.A. (1992). Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum. In: Fodor, J.A . *A Theory of Content and Other Essays* (pp. 3–30). Cambridge, Massachusetts: MIT Press.

Friston, K.J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. DOI: 10.1016/j.tics.2009.04.005.

Gärdenfors, P., & Makinson, D. (1994). Nonmonotonic inference based on expectations. *Artificial Intelligence*, *65*(2), 197–245. DOI: 10.1016/0004-3702(94)90017-5.

Jarmużek, T. (2007). Book Review: David Makinson, "Bridges from Classical to Nonmonotonic Logic", King's College Publications, London, 2005, pp. 216, ISBN 1-904987-00-1. *Logic and Logical Philosophy*, *16*, 259–262. DOI: 10.12775/LLP.2007.007.

Makinson, D. (2005). Bridges from Classical to Nonmonotonic Logic. *Text in Computing*, *5*. London: King's College Publications.

Malinowski, J., (1997). Logiki niemonotoniczne. Czy logika jest nauką kognitywną? *Przegląd Filozoficzny – Nowa seria*, R. VI, *1*(21), 31–53.

Marciszewski, W. (ed.) (1987). *Logika formalna. Zarys encyklopedyczny z zastosowaniem do lingwistyki i informatyki*. Warszawa: PWN.

Moore, R.C., (1985). Semantical considerations on nonmonotonic logic. *Artifical Intelligence*, *25*(1), 75–94.

Mota, T., Sridharan, M. & Leonardis, A. (2021). Integrated Commonsense Reasoning and Deep Learning for Transparent Decision Making in Robotics. *SN Computer Science*, *2*(4), 242. DOI: 10.1007/s42979-021-00573-0.

Poole, D. (1988). A logical framework for default reasoning. *Artificial Intelligence*, *36*(1), 27–47.

Przymusinski, T.C. (1997). Autoepistemic logic of knowledge and beliefs. *Artificial Intelligence*, *95*(1), 115–154. DOI: 10.1016/S0004-3702(97)00032-5.

Riley, H., & Sridharan, M. (2019). Integrating Non-monotonic Logical Reasoning and Inductive Learning With Deep Learning for Explainable Visual Question Answering. *Frontiers in Robotics and AI*, *6*(125). DOI: 10.3389/frobt.2019.00125.

Seth, A.K. (2015). The Cybernetic Bayesian Brain – From Interoceptive Inference to Sensorimotor Contingencies. In: T. Metzinger, & J.M. Windt (ed.). *Open MIND*, *35*, 1–24. DOI: 10.15502/9783958570108.

Sequeda, J. (2012). *Introduction to: Open World Assumption vs Closed World Assumption, Dataversity*. https://www.dataversity.net/introduction-to-open-world-assumption-vs-closed-world-assumption/#. Accessed 30.11.2022.

Andrzej Bielecki
AGH University of Krakow
https://orcid.org/0000-0002-0192-3785
https://doi.org/10.35765/slowniki.417en

Cybernetics

Summary

DEFINITION OF THE TERM: Cybernetics is an interdisciplinary science focused on the control mechanisms within systems that interact with their environment. These systems receive signals from their environment through input modules, in some cases also extracting energy and matter. They then interact with their surroundings via output modules. A fundamental principle of cybernetics is that common characteristics are shared by processes involved in controlling, transmitting and processing signals, information, and energy in a system, regardless of whether the system is a machine, a living organism, or a social structure. Cybernetics places special emphasis on autonomous systems, which are capable of setting their own goals, devising algorithms to achieve them, and maintaining functional stability.

HISTORICAL ANALYSIS OF THE TERM: The evolution of cybernetics is intricately linked to advances in mathematics, computer science, robotics, and biology as it intersects with critical research areas across these disciplines. Starting in the 1930s, the theoretical foundations of cybernetics were laid by mathematicians with concepts such as the Turing machine, and by physicians with ideas like homeostasis. The field has also been profoundly shaped by the progress in the development and continual improvement of computers. Today, key areas of cybernetic research focus on artificial intelligence, robotics, and the cybernetics of biological systems.

DISCUSSION OF THE TERM: The broad applicability of cybernetic tools, including the universal concept of a cybernetic system, allows for

its use across a wide range of fields. Cybernetic methods are employed to analyse economic, social, technical, biological, political, and even cultural systems. The discipline offers methods for examining the structure and dynamics of various systems, irrespective of their carrier, highlighting its potential for diverse applications.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Cybernetics plays a crucial role in modelling and analysing complex phenomena. Both in the past and now, Polish scientists have been making significant contributions to the field.

Keywords: cybernetic system (agent), control, information, homeostasis, feedback

Definition of the term

Cybernetics (from the Greek κυβερνήτης, meaning helmsman) is a theoretical interdisciplinary science focused on information and signal processing within cybernetic systems, which are often referred to as agents in contemporary literature. These agents can be either isolated software modules or processes within computer systems (known as disembodied agents), or physical systems (embodied agents). A cybernetic system is a dynamic and individual unit which acts within a specific environment. Through specialised input modules, the agent receives signals that carry information from the environment. For physical (i.e., embodied) agents, particularly biological ones, they may also receive energy and, occasionally, matter. The analysis of information and signal processing in cybernetic systems – and for physical agents, the processing of energy and potentially matter – focuses on how these resources are transferred between the different modules of the agent. This analysis is conducted within the framework of control studies, which examine how systems interact with their environment to achieve specific goals.

At the heart of cybernetics lies the observation that control processes share certain common characteristics that can be studied regardless of whether they involve machines, living organisms, or social systems. The field is divided into several areas of focus, including technical cybernetics (Arbib, 1964; Ashby, 1958; Wiener, 1948; Mazur, 1966), biocybernetics (Arbib, 1964; Ashby, 1958; Wiener, 1948; Tadeusiewicz, 1994), and social cybernetics. Cybernetics employs theoretical tools such as systems theory, modelling theory, information theory, game theory, and decision theory. Key mathematical branches integral to cybernetics include control theory, dynamical systems theory, probability theory, and mathematical logic. The discipline overlaps with many fields, including mathematics, physics, chemistry, biology, medicine, computer science, automation, electronics, mechatronics, robotics, bionics, philosophy, economics, psychology, cognitive science, and sociology. The boundaries of cybernetics are fluid and dynamic. For instance, in the 1950s and 1960s, artificial intelligence was considered part of cybernetics, but it is now predominantly categorised under computer science. Nonetheless, in a broader sense, fields like computer science and automation can be classified as branches of technical cybernetics.

Historical analysis of the term

Although the modern understanding of cybernetics is often traced back to 1948, when Norbert Wiener (1894–1964) published *Cybernetics: Or Control and Communication in the Animal and the Machine* (Wiener, 1948), its roots extend much further. Plato (427–347 BC) first used the word 'cybernetics' in his dialogue *Gorgias*. André Ampère (1775–1836) later adapted the term from its literal meaning of 'steering a ship' to describe the broader scope of politics, which covers the art of governing the state and aligns closely with what is now known as social cybernetics. The term 'cybernetics' was used in a similar sense for the first time in written Polish by philosopher Bronisław Trentowski (1808–1869) in his 1843 work *Stosunek filozofii do cybernetyki, czyli sztuki rządzenia narodem* [*The relation of philosophy to cybernetics or the art of governing the nation*].

The 1930s marked a period of intensive research that laid the groundwork for what would become cybernetics. Although the widely recognised beginning of cybernetics (in 1948) was still several years away, significant and often groundbreaking results emerged from studies in biology, medicine, psychology, psychiatry, technology, and mathematics. Key concepts that later became fundamental to cybernetics were also defined during this time. For instance, the concept of homeostasis, which describes the process of maintaining functional balance within an organism, was introduced by the American physiologist, neurologist, and psychologist Walter Cannon (1871–1945) in 1932. Alan Turing (1912–1954) began systematic research that would become the mathematical foundations of cybernetics and technical cybernetics. His work on the Turing machine (Turing, 1936) established a formal model of computational efficiency and revealed that some functions are inherently uncomputable, highlighting that machines' capabilities are limited. Similarly, Kurt Gödel (1906–1978), an Austrian mathematician who lived in the USA from 1940, conducted research on the logical foundations of mathematics and metamathematics in the 1930s. In 1929, as part of his doctoral thesis, he proved the completeness theorem for first-order predicate calculus. In 1931, he proved two other fundamental theorems in the foundations of mathematics: the incompleteness theorem for any axiomatic deductive system that includes the axiomatic system

of natural numbers, and the consistency theorem for such a system. Gödel's theorems are of fundamental importance for the problem of proving mathematical theorems using computers (cf. Arbib, 1964).

A significant milestone in the development of technical cybernetics was the creation of early calculating machines that had some functionality resembling modern computers. ABC (Atanasoff-Berry Computer), built in the USA in 1939 by John Atanasoff and Clifford Berry, was a specialised machine designed to solve linear algebraic equations. While not programmable, ABC is considered a prototype of the modern computer due to its pioneering use of electron tubes for arithmetic operations and capacitors for memory. In 1941, the Z3 computer, developed by Konrad Zuse in Germany, represented the first programmable computing machine. Unlike ABC, Z3 used electrical relays instead of electron tubes. It featured program input via perforated tape, data entry through a keyboard, and results displayed on a screen.

A further step towards the development of a computer was the creation of the Colossus machine in the UK in 1943. Designed to break German ciphers, Colossus was the first electronic, digital, programmable computing machine. It performed arithmetic and logical operations using electron tubes, and its programming was implemented through the setup of numerous relays rather than being stored in memory or read from punched tape or cards.

The next major leap came with ENIAC (Electronic Numerical Integrator and Computer), completed in the United States in 1945. ENIAC was built using approximately 18,000 electron tubes and over 1,000 switches. Initially, it lacked working memory and was programmed by manually switching cable plugs. Later, it was upgraded with a punch card reader. Originally designed for calculating ballistic trajectories for long-range artillery, ENIAC's applications expanded to include nuclear weapons design, wind tunnel testing, weather forecasting, and space radiation research.

A major theoretical advancement in biocybernetics was achieved in 1943 by neuropsychiatrist Warren McCulloch (1898–1969) and logician Walter Pitts (1923–1969), who developed a formal model of the neuron and neural networks that incorporated logical principles.

Another significant theoretical contribution came from biologist Ludwig von Bertalanffy (1901–1972), who first presented his general

system theory in a 1937 philosophy seminar at the University of Chicago. A significant impetus for the development of cybernetics came from the computational challenges associated with directing anti-aircraft artillery fire. As aircraft speeds increased exponentially, calculations needed to predict their future position and continuously adjust in real-time to account for their manoeuvres. This challenge led to the development of the concept of negative feedback in control systems, which later became a foundational principle of cybernetics.

In the UK in the early 1940s, psychiatrist Ross Ashby (1903–1972) explored analogies between adaptive processes in animals/humans and equilibrium states in mechanical systems, continuing the work on homeostasis initiated by Walter Cannon. His research culminated in the creation of the electromechanical homeostat in 1948. In 1946, following McCulloch's initiative, annual interdisciplinary conferences entitled "Control and Communication in Machines and Animals" began in New York. In 1947, Norbert Wiener (1894–1964) named the newly established field 'cybernetics'. A year later, Claude Shannon (1916–2001) published his seminal work on the mathematical theory of communication (Shannon, 1948), which he expanded to include information theory and coding in 1949. As previously mentioned, the publication of Wiener's famous book, along with the introduction of communication and information theory, is generally regarded as the birth of cybernetics as a new science. Additionally, John von Neumann (1903–1957), a Hungarian mathematician, physicist, and computer scientist working in the United States, made significant contributions to cybernetics. His work on automata theory (von Neumann1951), including research on self-replicating automata and cellular automata, further shaped the field. Since these foundational developments, cybernetics has evolved rapidly, with several research strands emerging and expanding the scope of the discipline.

The first strand, closely tied to advances in electronics, was the development of computing technology. Computers are commonly classified into the following generations based on the technology used in their construction: zero-generation computers (relay technology), first-generation computers (vacuum tubes), second-generation computers (transistors), third-generation computers (integrated circuits with small-scale integration), fourth-generation computers (integrated circuits with medium and large-scale integration), and fifth-generation computers

(technologies not based on silicon, such as quantum, optical, chemical, or biochemical computing). Currently, fifth-generation computers are still in the early experimental stages.

Following the construction of ENIAC, vacuum tube computers were systematically refined and became widely used in academic institutions. Concurrently, advances in semiconductor technology, particularly the development of the semiconductor diode and transistor, spurred the creation of second-generation computers. The first transistor, known as the point-contact transistor, was developed in 1947 at Bell Telephone Laboratories in the US. It was soon replaced by the junction (layer) transistor, introduced in 1950. One of the first computers to incorporate second-generation technology was built at the University of Manchester in 1953. This machine utilised both transistors and vacuum tubes. Additionally, in 1954, Bell Telephone Laboratories developed an airborne computer for the B-52 bomber that also combined solid-state and vacuum tube technologies.

The first second-generation computer that was fully solid-state was constructed in the UK at the Atomic Energy Research Establishment in 1955. Despite its advanced technology, it had less computing power compared to the vacuum tube computers of that era. The transition to second-generation computers gained momentum with the introduction of the IBM 7090 in 1959, which offered superior computing power and remained in use into the 1980s.

The IBM 360, released in 1965, marked the advent of third-generation computers. This generation benefited from the miniaturisation of integrated circuits, which reduced power consumption and enabled the development of minicomputers. These advances also spurred the creation of new operating systems.

Fourth-generation computers are distinguished by their use of microprocessors – single integrated circuits with a high level of integration that execute instructions. The first microprocessor, developed in 1971, can serve as the central processing unit (CPU) of a computer. This technology paved the way for the development of microcomputers and personal computers. Multi-core processors, introduced in 2001, further advanced computing power. Concurrently, innovations in liquid crystal displays and thin-film transistors facilitated the creation of laptops. The enhanced computing speed and progress in telecommunications

technology led to the development of computer networks and the rise of the internet.

The second major strand of cybernetic research was linked to artificial intelligence (AI) (see the article Artificial Intelligence in this volume). During the 1950s, when computing power was limited compared to the demands of AI systems, research primarily focused on theoretical aspects, with prototype implementations being quite rudimentary. Initially, the field concentrated on creating formal models of human inference and the way in which humans reason. In 1955, the Logic Theorist system was developed to prove mathematical theorems, successfully demonstrating its capability by proving dozens of established theorems. By the late 1950s, efforts expanded to include abstract inference methods based on formal logic rather than purely heuristic approaches.

A significant advance in the neural networks domain (see the article Artificial neural networks in this volume) occurred in 1957 with the development of the first electromechanical neural network system, known as the perceptron, by American psychologist Frank Rosenblatt (1928–1971). The neurons in the perceptron were designed based on the McCulloch and Pitts model, and the network itself was structured in layers. This means that the neurons were organised into subsets corresponding to each layer, where each neuron in one layer was connected to all the neurons in the subsequent layer, but not to neurons in other layers. In this context, a connection between neurons signifies that the output signal of one neuron serves as the input signal for another. The perceptron modelled the functioning of a visual system, where neuronal connections form a layered structure. Rosenblatt's research primarily focused on the learning capabilities of this type of network, specifically on the algorithm for adjusting the weights of individual neurons, which are the system's parameters. There was also research into the potential for performing logical operations using both single neurons and layered networks. The 1980s and 1990s marked a period of rapid development in artificial neural network research. During this time, new types of networks were proposed, along with numerous training algorithms. Significant progress was made in understanding neural network learning, the approximation properties of networks, the dynamics of recurrent neural networks, and their practical applications. Additionally, new models of the neuron were described.

The 21st century has seen the development of advanced types of neural networks, including deep networks and graph networks. Research on artificial intelligence, which gained momentum in the late 1960s and 1970s, expanded significantly with the exploration of knowledge representation. During this period, several types of semantic networks were introduced, in which knowledge was modelled as graphs whose vertices represent objects and edges represent the relationships between them. Additionally, script systems were developed as formal models of natural language, which describe the structure and semantics of the world. Research into natural language processing in cybernetics also led to the emergence of syntactic methods for object recognition and scene analysis. These methods are based on a formal description of the rules that govern object construction. The complete set of these rules forms a descriptive grammar that generates all possible patterns within a scene. The theoretical foundation of grammar provides a formal description of a language. This analysis is carried out using a formal automaton, which determines whether a given object belongs to the language generated by the defined grammar. Syntactic methods are applied in various fields, including vision systems for autonomous robots and the analysis of medical images.

A fascinating area of biology-inspired research is the modelling of evolutionary processes, which gained significant traction in the late 1950s. John Holland (1929–2015), an American psychologist, computer scientist, and mathematician from the University of Michigan, pioneered this field by implementing the first genetic algorithms. These models, like most models that draw from biological principles, have several functions: first, they provide theoretical tools for analysing biological phenomena; second, they serve as a starting point for creating artificial systems with characteristics similar to biological prototypes (a strictly cybernetical aspect); third, they often lead to the development of algorithms that address a wide range of computational problems. These models include artificial immune systems, developed in 1986 by American physicists J. Doyne Farmer (b. 1952), Norman Packard (b. 1954), and computer scientist and biologist Alan Perelson, which have been explored for their potential as artificial intelligence systems. Today, artificial immune systems are employed as pattern classifiers, associative memory models, and anomaly detectors. It is worth

emphasising that research in artificial intelligence extends beyond the realm of cybernetics, encompassing significant contributions from philosophy and psychology (Flasiński, 2011).

Another significant area of cybernetic modelling of biological processes and structures is membrane computing (also known as membrane systems or P-systems), introduced in 1998 by Romanian computer scientist Gheorghe Păun (b. 1950). In this computational framework, a system comprises multiple membranes arranged hierarchically based on containment relations. Computations occur in parallel within individual membranes, with results transferred between membranes at the completion of each computational step. The structure and dynamics of computational processes in membrane systems are organised similarly to biochemical and biophysical processes in biological systems. Beyond artificial intelligence, biologically inspired cybernetic research also explores scenarios related to the origin of life and the study of minimal systems that qualify as living systems. A notable example of this research is the concept of the chemoton, a theoretical model proposed in 1974 by Hungarian biologist and chemist Tibor Gánti (1933–2009). The chemoton represents a minimal network of chemical processes that exhibit characteristics of a living system (Gánti, 1975).

The third major strand of cybernetic research is linked to robotics, particularly in the realm of biologically inspired autonomous robots. In the early days of cybernetics, efforts were made to create robots that mimicked biological reflexes. A pioneering example was the cybernetic turtles developed by British neurophysiologist and cyberneticist William Grey Walter (1910–1977) in 1949. These robots were among the first to display autonomous behaviour. Equipped with basic touch and light sensors, they were programmed to react to light stimuli by moving toward a light source, a behaviour known as positive phototaxis in biology. Walter's initial models, named Elmer and Elsie, were used to explore the effects of stimulus amplification on performance, revealing that excessive amplification could impair the robots' ability to locate the light source efficiently.

Subsequent developments in robotics included research into robot vision systems, which began in the 1960s, and the integration of industrial robots into factory production lines in the 1970s. Since the 1970s, robots, which were previously only operated remotely, were used for lunar exploration by the Soviet Union and, since the 1990s, for Mars

exploration. Today, robots – both humanoid and animal-like – which exhibit highly refined motor skills are being developed, with the US company Boston Dynamics leading the field.

The Polish school of cybernetics merits special attention for its significant contributions and unique historical context. Until the mid-1950s, in the USSR and, consequently, in Poland, cybernetics was considered a reactionary pseudoscience. Thus, research in this field had to be conducted under the constraint of avoiding the term 'cybernetics' and faced severe limitations on publishing results. However, from 1956 onwards, cybernetics began to flourish in Poland despite the challenging conditions imposed by the communist regime, including restricted international scientific communication and limited funding. A pioneering Polish cyberneticist was Henryk Greniewski (1903–1972), a logician and mathematician who specialised in financial mathematics. Starting in 1948, Greniewski was instrumental in developing Poland's first computers and co-founded the *Grupa Aparatów Matematycznych* [*Mathematical Apparatus Group*] at the State Mathematical Institute in Warsaw with mathematicians Kazimierz Kuratowski and Andrzej Mostowski. He introduced the concept of a 'relatively isolated system, which is akin to today's terms 'cybernetic system' or 'agent'. His work encompassed modelling conditioned reflexes, studying codes and languages, and exploring information and network theories. Greniewski also developed models of animal behaviour, referred to as synthetic animals, and was a key figure in establishing the Polish Cybernetics Society in 1962.

Another prominent figure was Mieczysław Choynowski (1909–2001), a psychologist and philosopher renowned for founding Polish psychometrics. From 1946, Choynowski worked at Jagiellonian University, where he applied cybernetic methods to model the nervous system, focusing on concepts like negative feedback and homeostasis. From 1948 to 1950, Lem (1921–2005) worked under his supervision at the *Konserwatorium Naukoznawcze* [*Science Studies Seminar*] at Jagiellonian University during its operation between 1945 and 1950. Lem's experiences under Choynowski significantly shaped both his scientific essays on cybernetics and his renowned science fiction literature.

Choynowski also collaborated with Antoni Kępiński (1918–1972), who developed the innovative concept of information metabolism, according to which information processing is a crucial aspect of metabolism in

biological systems. He argues that traditional analyses of metabolism, which typically focus on chemical reactions and energy transformations in biological cells, should also include an informational component. Kępiński also explored cognitive representations (or cognitive frameworks) and the dynamics of psychological reactions. However, he did not align his theory with cybernetics, partly because he viewed cybernetics as a technical discipline and was critical of reductionist approaches to human behaviour, favouring personalism instead.

Another key figure in Polish cybernetics was Oskar Lange (1904–1965), who significantly contributed to economic cybernetics. Lange was the only member of the first generation of Polish cyberneticists whose work gained considerable recognition in the Western world. His research focused on centralisation versus decentralisation in management, the regulation and self-regulation of economic systems, and the interplay between cybernetics and philosophy. Stanisław Lem (1921–2006), mentioned above, also made notable contributions, particularly in exploring the limitations and possibilities of modelling biological systems, including their psychological aspects. His concept of 'phantomatics' (Lem, 1974) addressed virtual reality, while his work examined the properties of the mind and its biological constraints, as well as the potential for surpassing these limits with artificial intelligence. Lem's writings also touched on the sociological implications of cybernetics, and his cybernetic analysis of literary works remains unique to this day. He also incorporated original ideas and concepts from the field of cybernetics in his novels. Despite his primary role as a writer rather than a scientist, Lem's contributions to cybernetics are frequently cited in both philosophical and scientific contexts, reflecting the depth of cybernetic thought in his essays and novels.

Marian Mazur (1909–1983) developed a pioneering theory of autonomous systems and qualitative information theory (Mazur, 1966). He applied his concept of autonomous systems to explore the structure and dynamics of the human psyche. Mazur's work extended into economic and social cybernetics, a field further advanced by Józef Kossecki (1936–2005), who conducted research on the cybernetics of culture.

Neurophysiologist Jan Trąbka (1931–2012), the head of the Department of Biocybernetics at Jagiellonian University Medical College from 1991 to 2002, made significant contributions to neurocybernetics. From

1959 to 1960, as a Rockefeller Foundation Fellow, he collaborated with Norbert Wiener at Harvard University. Trąbka's research focused on the brain, including the use of cybernetic methods to analyse consciousness. He approached cybernetics as a broader philosophical and worldview phenomenon, extending beyond control and information theory.

Ryszard Tadeusiewicz (b. 1947), a biocyberneticist and computer scientist at AGH University of Krakow, where he served as rector from 1998 to 2005, has conducted extensive research in artificial intelligence systems, particularly artificial neural networks. His work has also encompassed vision systems for industrial robotics. Tadeusiewicz has achieved significant results in modelling biological systems, especially in neurocybernetics (Tadeusiewicz, 1994), and in simulating insect swarm behaviour. His contributions include advances in pattern recognition and scene analysis, with notable applications in robotics and medical image analysis. He also investigated the autonomic system model proposed by Marian Mazur, examining its implementation and behaviour in dynamic environments.

Zdzisław Pawlak (1926–2006) made significant contributions to various fields, including the development of rough set theory, the creation of a formal model for a calculating machine known as Pawlak's machine, and the mathematical modelling of DNA genetic codes. He also developed a unique mathematical framework for conflict theory. In 1950, Pawlak constructed the GAM-1 computer, based on relay technology, which was the first computer developed by the Mathematical Apparatus Group at the State Mathematical Institute.

Today, Polish scientists are actively advancing research in artificial intelligence, biocybernetics, and information theory. In the field of artificial intelligence systems, the Polish scientific community is particularly prominent in investigating the theoretical foundations of artificial intelligence, artificial neural networks (see the article Artificial neural networks in this volume), and syntactic methods.

Discussion of the term

Cybernetics investigates how certain types of systems function within their environments by utilising theoretical models and creating artificial systems. At the core of cybernetic research, scholars examine

fundamental problems such as stimulus-response patterns within these systems. One prominent methodology is the concept of the b l a c k b o x, which was initially introduced in electrical engineering in 1941 and later adapted for cybernetics by Ross Ashby in 1956 (Ashby, 1956). In this context, the black box represents a system whose internal workings, goals, and algorithms are unknown to the researcher. However, the researcher can interact with the system through its inputs and observe its outputs. By collecting a sufficient amount of data on how the system responds to various stimuli, researchers can attempt to deduce the rules governing the system's stimulus-response relationships or determine whether these responses are deterministic.

Cybernetics offers general concepts and formal techniques to study the properties, dynamics, and structures of cybernetic systems across different domains, regardless of the medium in which these systems are embedded – whether technological, biological, or social. This includes analysing aspects such as system stability, the structure of signal and energy flows, how signals constitute and encode information in the system, particularly in relation to control, as well as how the system interacts with its environment, primarily in the context of goal achievement. For instance, cybernetic methodologies enable the analysis of a homeostat's structure, its interaction with other system modules, and the mechanisms by which homeostasis is maintained, i.e., the structure and strength of feedback loops that ensure stability. These methodologies have facilitated the development and detailed analyses of a universal structure for autonomous systems (Mazur, 1966).

Cybernetics plays a crucial role in contemporary scientific research, particularly within biology. It has significantly influenced the development of systems biology, a rapidly advancing field that utilises cybernetic methods to study biological processes and the dynamics of biological structures. A key focus in this area is understanding the flow and encoding of information within biological systems. The prevailing view in contemporary biology is that information processing is fundamental to life, with statements like "life = matter + information" and "nothing in biology makes sense except in the light of information flow" being widely accepted as foundational truths. Cybernetic methods also address fundamental questions about the nature and origins of life. Researchers continue to explore cybernetic definitions of life and analyse the phenomenon of life

through a cybernetic lens (Korzeniewski, 2001), with the understanding that life extends beyond purely biological paradigms and thus requires approaches that surpass traditional biological methods. This extends to research on the role of information in biology and chemistry (Bielecki & Schmittel, 2022).

A related branch of cybernetics, molecular cybernetics, intersects with supramolecular chemistry (Schmittel, 2019), exploring topics such as chemical switches, chemical logic gates, and molecular machines. The growth of the internet has spurred cybernetic research in computer science, leading to the development of intelligent assistants that enable advanced web search capabilities. Notable advancements include intelligent dialogue systems like ChatGPT and the expansion of multi-agent systems. This research encompasses 'emotional agents', which focus on two main aspects: first, the ability of artificial intelligence systems to detect human emotions and adapt their responses accordingly; second, the modelling of emotions within AI systems, such as autonomous robots, to mimic the emotional functionality found in biological systems. For instance, an autonomous robot equipped with an anxiety simulator could avoid potentially dangerous situations. As autonomous robotics evolves, the need for increasingly sophisticated sensors grows. A major challenge in cybernetics is integrating stimuli from various types of sensors to create a comprehensive understanding of the environment, allowing artificial cybernetic systems to 'understand' and effectively navigate their surroundings.

Systematic reflection with conclusions and recommendations

Cybernetics is an expansive field that has given rise to distinct scientific disciplines such as automation, computer science, and robotics. Although these areas are often treated as separate fields, they share significant overlaps with cybernetics. This interconnectedness is evident in the fact that cybernetics is not formally recognised as a distinct scientific discipline in Poland. From 2005 to 2018, the term 'biocybernetics and biomedical engineering' was used within the engineering sciences, but it has since been replaced by 'biomedical engineering'. Despite

this, cybernetics remains well represented in prestigious scientific journals and research institutes at universities, particularly in the field of biocybernetics.

In summary, cybernetics is experiencing rapid development and maintains extensive intersections with numerous other scientific disciplines. Along with the expanding functionalities of the internet, the proliferation of robotics and smart technologies – such as smart cars, smart homes, and smart energy grids –are driving advances in both theoretical and applied cybernetics. The widespread integration of cyber systems also have cultural implications, with terms like 'cybersecurity', 'cybernetwork', and 'cybersex' becoming part of everyday language. Furthermore, issues such as addiction to computer games and cybersex are becoming significant social and medical concerns, highlighting the darker side of cybernetics.

References

Arbib, M.A. (1964/1987). *Brains, Machines and Mathematics*. New York: Springer-Verlag.

Ashby, W.R. (1958). *An Introduction to Cybernetics.* London: Chapman & Hall LTD.

Bielecki, A., & Schmittel M. (2022). The information encoded in structures: Theory and applications to molecular cybernetics. *Foundations of Science*, *27*(2), 1327–1345.

Flasiński, M. (2016). *Introduction to Artificial Intelligence.* New York: Springer-Verlag (Polish edition: Flasiński, M. (2011). *Wstęp do sztucznej inteligencji.* Warszawa: Państwowe Wydawnictwo Naukowe).

Gánti, T. (1975). Organisation of chemical reactions into dividing and metabolizing units: the chemotons*. BioSystems*, *7*(1), 189–195.

Korzeniewski, B. (2001). Cybernetic formulation of the definition of life. *Journal of Theoretical Biology*, *209*(3), 275–286.

Lange, O. (1970). *Introduction to Economic Cybernetics*. Oxford: Pergamon Press (Polish edition: Lange, O. (1965). *Wstęp do cybernetyki ekonomicznej.* Warszawa: Państwowe Wydawnictwo Naukowe).

Lem, S. (2013). *Summa Technologiae.* Minneapolis, MN: University of Minnesota Press (Polish edition: Lem, S. (1974). *Summa technologiae.* Kraków: Wydawnictwo Literackie).

Mazur, M. (1966). *Cybernetyczna teoria układów samodzielnych*. Warszawa: Państwowe Wydawnictwo Naukowe.

Shannon, C.E. (1948). The Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Schmittel, M. (2019). Dynamic functional molecular systems: From supramolecular structures to multi-component machinery and to molecular cybernetics. *Israel Journal of Chemistry*, *59*(3), 197–208.

Tadeusiewicz, R. (1994). *Problemy biocybernetyki*. Warszawa: Państwowe Wydawnictwo Naukowe.

Turing, A.M. (1936). On computable numbers with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*(1), 230–265.

von Neumann, J. (1951). The general and logical theory of automata. In: *Cerebral Mechanisms and Behavior: The Hixton Symposium*. New York: John Wiley and Sons.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine.* Cambridge, Massachusetts: MIT Press.

Andrzej Bielecki
AGH University of Krakow
https://orcid.org/0000-0002-0192-3785
https://doi.org/10.35765/slowniki.418en

Artificial neural networks

Summary

DEFINITION OF THE TERM: An artificial neural network (ANN) is a computational system designed to mimic the structure and function of the human nervous system. The basic functional unit of an ANN is the artificial neuron, which is the model of the biological neuron. In a network, neurons are interconnected, with the output from some neurons serving as the input for others. Depending on the connection pattern, different network structures are created.

HISTORICAL ANALYSIS OF THE TERM: The foundation for modelling neuronal systems lies in discoveries made by neurophysiologists since the late 19th century. Recognising the neuron as the fundamental unit of the nervous system enabled the development of the first theoretical neuron model in 1943, leading to the creation in 1957 of the first artificial neural network, which was a layered network. The 1980s saw significant advances in neural network theory, including new network types, learning methods, and improved mathematical analysis, which fuelled the rapid progress of ANNs into the late 20th and early 21st centuries.

DISCUSSION OF THE TERM: Artificial neural networks can be implemented as either software or electronic circuits, in both analogue and digital forms. A crucial aspect of ANNs is their ability to learn, typically through iterative algorithms that adjust the network based on input examples provided by the programmer. ANNs are widely used to model various aspects of the nervous system and cognitive processes.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Beyond their biological inspiration, artificial neural

networks are a form of artificial intelligence capable of solving a wide range of problems. These include optimisation, clustering, control, pattern recognition, diagnosis, and prediction.

Keywords: artificial neural network, neuron model, neural network learning

Definition of the term

An artificial neural network (ANN) is a software or electronic system designed to mimic the structure and function of the nervous system.

The foundational concept behind the modelling of neural systems (more on modelling can be found in the article Cybernetics in this volume), and thus the construction of ANNs, is The Neural Theory. This theory posits that neurons are the fundamental functional units of the nervous system. Accordingly, the first step in creating an ANN is to develop a formal model of a neuron. The first such model was developed in 1943 by two Americans: neurophysiologist Warren McCulloch (1898–1969) and logician Walter Pitts (1923–1969). They proposed both a model for an individual neuron and a model for artificial neural networks built from these neurons. Additionally, they explored the ability of these networks to perform logical computations consistent with classical propositional calculus (McCulloch & Pitts, 1943).

In its original form, McCulloch and Pitts's neuron model functioned as a threshold unit with several weighted inputs and a single output. When the total excitation of the neuron exceeded a certain threshold, a normalised signal would be produced at the output. The total activation of a neuron is calculated as the sum of the products of each input signal and its corresponding weight. These weights are parameters of the artificial neuron and must be set so that the implemented system, whether an individual neuron or a network of such neurons, functions – as a computational system – according to the designer's expectations, effectively solving the specific task for which it was designed.

An artificial neural network is a system comprised of interconnected artificial neurons, where the output signal of one neuron serves as the input of other neurons. The first successful implementation of an ANN was achieved in 1957 by American psychologist Frank Rosenblatt (1928–1971), a pioneer in the field of artificial intelligence (Rosenblatt, 1958). Rosenblatt developed a layered network known as the perceptron, where neurons are organised into distinct layers, with each neuron in one layer connected only to neurons in the subsequent layer. The perceptron was inspired by the structure of the visual system, particularly the layered arrangement of neurons in the retina.

Since the 1970s, research in the field of artificial neural networks has advanced rapidly. A critical area of study is neural network learning, or training, which involves developing algorithms to adjust the weights of neurons within the network. Some learning algorithms also enable dynamic modification of the network's structure by adding or removing neurons, resulting in what are known as ontogenic neural networks. This modification mirrors biological processes related to neuroplasticity in the brain, which not only involves the expansion of connections between neurons but also changes the number of neurons. Another key research area focuses on developing new types of ANNs based on different neuron models and varying structural designs. The mathematical foundations of ANNs are intensively studied, along with their applications. These include modelling various biological neuronal structures and, extending beyond their biological origins, employing neural networks as artificial intelligence systems to solve a wide range of complex problems.

Historical analysis of the term

The study of the nervous system forms the foundation for the development of artificial neural networks. Nervous tissue is composed of three main components: neuronal, glial, and vascular. The neuronal component consists of neurons, which are specialised biological cells responsible for processing and transmitting signals. The basis for all neuron models is rooted in the work of neurophysiologists, whose discoveries have a rich historical background. The Spanish neurohistologist and neuroanatomist Santiago Ramón y Cajal (1852–1934) is widely regarded as the founder of The Neuron Theory. Ramón y Cajal was awarded the Nobel Prize in Physiology or Medicine in 1906, sharing the honour with Camillo Golgi (1843–1926) "in recognition of their work on the structure of the nervous system". Prior to their discoveries, the brain was believed to be a homogeneous mass. Golgi's development of a staining method for histological preparations in 1873 allowed scientists to visualise the brain's structure. However, Golgi hypothesised that the brain was composed of a continuous network of tubules. Ramón y Cajal, using Golgi's staining technique, observed that the brain is made up of distinct

individual cells that do not physically touch one another. He particularly noted that the protoplasm of each neuron remains separate. Additionally, the English physician and scientist Richard Caton (1842–1926) made significant contributions by reporting his findings on the brain's electrical activity in 1875.

The English physician and physiologist Charles Sherrington (1857–1952) made significant contributions to the understanding of the nervous system, earning the Nobel Prize in Physiology or Medicine in 1932 alongside Edgar Adrian (1889–1977) "for their discoveries regarding the function of neurons". In 1887, Sherrington introduced the concept of the synapse as a connection point between two neurons or between a neuron and another excitable cell, such as a muscle cell. A synapse consists of three components: the presynaptic terminal (the end of the signal-transmitting neuron, called the presynaptic neuron), the synaptic cleft (the gap between the presynaptic neuron and the signal-receiving, or postsynaptic cell), and the receptor on the postsynaptic cell. There are two types of synapses: chemical and electrical. In a chemical synapse, signals are transmitted between cells through neurotransmitters, which are electrically neutral chemicals. The synaptic cleft in chemical synapses is approximately 20 nanometers wide. In contrast, an electrical synapse has a narrower synaptic cleft, around 4 nanometers, and transmits signals through ion transport. Electrical synapses allow for much faster signal transmission – nearly a thousand times faster than chemical synapses – but the signals are weaker, less controllable, and bidirectional. The functioning of chemical synapses was elucidated by British physiologist and pharmacologist Henry Dale (1875–1968) and Austrian pharmacologist Otto Loewi (1873–1961), who were jointly awarded the Nobel Prize in 1936 "for their discoveries relating to chemical transmission of nerve impulses". Further advances in understanding neural mechanisms came from English scientists Alan Hodgkin (1914–1998), a biochemist, and Andrew Huxley (1917–2012), a physiologist and biophysicist. They described the mechanisms of ion transport in neurons that underlie the generation of action potentials. Hodgkin and Huxley developed a mathematical model of ionic conductance in neurons, which were represented by a system of ordinary differential equations, as well as a functionally equivalent electronic model (Hodgkin & Huxley, 1952). Their work laid the groundwork for mathematical

and electronic models of neurodynamic phenomena, which later served as the basis for many types of artificial neural networks. For their pioneering research, Hodgkin and Huxley were awarded the Nobel Prize in Physiology or Medicine in 1963.

The second major component of neural tissue, alongside neurons, is glial cells, which were first discovered in 1851 by German anatomist Heinrich Müller (1820–1864). In the late 1850s, German pathologist Rudolf Virchow (1821–1902), using a staining technique developed by Golgi, identified astrocytes, a specific type of glial cell. Until the late 20th century, glial cells were thought to only form the myelin sheaths of axons (i.e., single neuronal projections that conduct impulses the synapse) and the spatial scaffolding for neurons. However, in the 21st century, it was discovered that certain types of glial cells also play a role in signal transmission by exchanging signals with neurons (Fields, 2009). Despite these discoveries, glial cells have yet to be incorporated into cybernetic models of the nervous system or artificial neural networks, which are currently based solely on models of chemical synapses.

The development of artificial neural networks began with two key milestones. The first was the creation of the threshold neuron model in 1943 by McCulloch and Pitts. The second was the implementation of the perceptron in 1957, which was built of artificial neurons based on this model. Initially designed to recognise alphanumeric characters, the perceptron proved sensitive to variations in the scale of the input patterns. For the single-layer perceptron, Rosenblatt developed an iterative weighting algorithm (network learning algorithm) and demonstrated its convergence. A similar structure was the ADALINE network, created in 1959 by Bernard Widrow (b. 1929) and Tedd Hoff (b. 1937). This single-layer network, implemented as an electronic circuit, was designed for adaptive signal processing (Widrow & Hoff, 1960).

In the late 1960s, Marvin Minsky (1927–2016) critically analysed the capabilities of existing artificial neural networks, highlighting their significant limitations. He provided examples of problems that single-layer networks could not solve and discussed the challenges associated with training networks of multiple layers. In particular, Minsky pointed out fundamental issues related to the training of ‘hidden layers’, which do not directly send signals to the network’s output but instead transmit them to other layers (Minsky & Papert, 1969). Based on his analysis,

Minsky concluded that artificial neural networks were not a promising avenue for advancing artificial intelligence.

After an initial surge of interest in artificial neural networks in the second half of the 1950s and 1960s, Minsky and Papert's book led to a significant decline in research throughout the 1970s. However, despite this downturn, several important developments occurred during this period. One key breakthrough that helped revive the field was the proposal of the backpropagation algorithm for training multilayer networks, introduced in 1974 by Paul Werbos. The success of backpropagation was based on the careful development of a formal model of learning, where the learning cybernetic system has access to a set of correct outputs in a given context. In the case of artificial neural networks, the context is the vector of the input signal given to the network, and the result is the vector of the output signal. The learning sequence is formed by a set of pairs, which consists of an input signal vector and the corresponding output signal vector, which represents the correct response of the network to a given input. This sequence includes a set of the input signals for which the correct system responses are known, and the corresponding set of output signals that serve as reference points for the weight adjustment algorithm. This method of learning is called supervised learning. By having a measure of the network's correctness, it is possible to develop an algorithm that adjusts the weights of neurons so that the network's performance improves over time. One of the simplest measures of this type is the sum of the squares of the differences between the vectors of the correct responses and the vectors of the network's responses to a given stimulus from the learning string. Here, the square of a vector is understood as the scalar product of the vector with itself. Since this measure is a non-negative number that equals zero only if all the vectors of the current network responses match the vectors of the correct responses, it will be zero only if the network is performing absolutely correctly. A weight adjustment algorithm is therefore used to ensure that this measure, which is a function of all weights in the network, gets smaller with each learning step. Since this measure is a function with real values and arguments that are the weights of the network, gradient algorithms can be used to find its minimum. The simplest of these is the largest gradient descent algorithm, known since the mid-19th century. When applied to the context of layered network

learning with the quadratic error function described above, it yields the backpropagation algorithm.

To apply gradient-based methods for training layered neural networks, it was necessary to modify the original McCulloch-Pitts model as its threshold-based activation function was not differential. To address this, a differentiable activation function was introduced. In the modified model, the total excitation of a neuron is transformed by a differentiable function, which produces the neuron's output signal. While the steepest gradient descent method is effective, it is relatively slow compared to other gradient-based methods, hence faster gradient-based optimisation techniques are now commonly used. The introduction of a suitable learning process model made it possible to apply existing numerical methods for training neural networks. However, the backpropagation algorithm did not gain immediate popularity. The authors of this algorithm did not use the results obtained in the 1970s, and it wasn't until the 1980s that backpropagation became widely recognised and adopted as a result of several publications on this algorithm. This resurgence of interest in artificial neural networks spurred the development of new neuron models, new types of neural networks, and new learning algorithms.

Finnish computer scientist Teuvo Kohonen (1934–2021) introduced a new type of network, now known as Kohonen networks (Kohonen, 1982) or self-organising maps. These networks model associative memory and are trained using self-organising learning maps; in turn, these maps are trained using unsupervised learning. This learning algorithm groups input signals based only on their similarity. As a result, the neuron weights are adjusted so that similar signals excite the same neurons.

Another notable neural network type is the Hopfield network, introduced by American physicist John Hopfield (b. 1933), who also explored neural networks as computational systems (Hopfield, 1982). The Hopfield network is composed of McCulloch-Pitts neurons, with each neuron connected to every other neuron in the network. As a result, due to the loop structure of these connections, signals circulate within the network when some or all the neurons are activated. The network's response is determined by the final state reached after this iterative process of neuron excitation. As an artificial intelligence system, the Hopfield network is used as an associative memory model and as an optimisation system (Mańdziuk, 2000). Additionally, Hopfield networks have become

important tools for modelling the dynamics of excitations in biological neural systems.

Cellular networks, introduced in 1988, are networks that create local connections, where neurons are arranged at the nodes of a grid (typically a square grid, though other structures are possible) and are only connected to nearby neurons (usually within a direct neighbourhood, sometimes within a given radius). In this context, a neighbourhood is understood according to the structure of the network. Neurons at the edges of the network disrupt the symmetry of interneuronal connections, so opposite edges are often connected. From a geometrical perspective, this means that the network spans the surface of a torus instead of lying flat on a plane. The dynamics of state changes in a cellular network are described by differential equations if the network is made of analogue neurons; otherwise, these dynamics are described by difference equations. The state dynamics of a single neuron can be described by an ordinary differential equation, while the dynamics of the entire network are described by partial differential equations. Therefore, cellular networks are also used to numerically solve partial differential equations and to simulate them in continuous time when using cellular networks implemented with analogue electronics.

Since the 1980s, there has been a rapid proliferation of new neural network types and neuron models. Additionally, multi-module hybrid systems that integrate neural networks as one of several components have become increasingly common. These advances have led to a wide range of applications for neural networks across various domains.

Discussion of the term

As mentioned earlier, one of the core challenges in artificial neural networks is the learning process, specifically the algorithm for adjusting the parameters of individual neurons. Numerous algorithms have been developed for this purpose, but listing them all would be beyond the scope of this article. Instead, the general taxonomy of neural network learning methodologies will be provided.

In practice, calculating the parameters of an artificial neural network based on theoretical analysis of the network and the problem it aims to

solve is rare. This approach is typically feasible only for small networks and straightforward problems. In some cases, advanced mathematical techniques, adaptations of physical models, or specific network learning algorithms may yield algebraic formulas for setting weights in a single step. For instance, some problems addressed by Hopfield networks fall into this category. However, for most neural networks, learning is an iterative process, where the parameters of the neurons are adjusted incrementally, with each step making small changes to refine the network's performance.

Iterative network learning methods generally fall into two main categories: supervised learning (or training) and unsupervised learning. Supervised learning algorithms are used when information about the desired output of the network is available; this information guides the training of the neural network. Depending on their nature and application, supervised methods are characterised by various strategies. In these methods, we have a set of input signals for which we know the correct responses that the network should generate at the output. A learning sequence consists of pairs of vectors, where the first vector represents the input signal, and the second vector encodes the correct response of the network The learning algorithm involves comparing the known correct response with the network's actual response and adjusting the network parameters so that the network's response is closer to the correct one when the same input signal is presented again. Commonly used methods of this type include the aforementioned gradient methods, where a known and fixed learning sequence is used to define a differentiable error function with a global minimum corresponding to network weights that produce responses matching the correct ones. Gradient methods are numerical techniques for solving ordinary differential equations or optimisation problems adapted for training neural networks. These methods are deterministic, meaning that the current network parameters uniquely determine their change at each learning step. In contrast, stochastic methods involve randomly setting the network parameters at each learning step according to probability distributions defined based on information about the solution. Examples of this type of strategy include genetic algorithms and evolutionary strategies, which can also be used to train neural networks.

The second major group of neural network learning methods is unsupervised learning. These methods are characterised by the absence of

information about the potential solution to the problem. Consequently, the network's task is to classify similar signals into the same category. From a mathematical perspective, this approach establishes an equivalence relation: two input signals are related if they are similar. Of course, for a given problem, this similarity must be well defined. The result of the network learning is the partitioning of the input space into equivalence classes. These methods fall into two main groups: those based on competitive learning rules and those based on Hebbian learning rules. The former involve presenting an input signal, identifying the neuron whose weight vector is most similar to the presented signal, and then adjusting the weights of only that neuron to make the two vectors even more similar (winner-take-all algorithms). Alternatively, we can modify the weights of several neurons that are most similar to the presented signal using winner-take-most algorithms. Kohonen networks are trained using this method (Tadeusiewicz, 1993). The second group of unsupervised learning methods is based on the rule developed by Donald Hebb (1904–1985), a Canadian neuropsychologist who studied the relationships between mental functions and brain structure and activity. His observations and experiments led him to conclude that a repeatedly stimulated synaptic connection between two neurons increases its conductivity (Hebb, 1949).

An important aspect of artificial neural networks is their mathematical framework (Bielecki, 2019). Key areas of research include the mathematical properties of the learning process, the dynamic properties of trained recurrent networks, and the computational capabilities of neural networks. Often, mathematical models used to describe physical phenomena are adapted to address these issues (Hertz et al., 1991). Neural network learning algorithms are predominantly iterative implementations of a numerical process. Consequently, questions about the convergence and stability of the learning process are central. When the theoretical model of the learning process is expressed as a differential equation, an additional problem arises regarding how well the dynamics are preserved when discretising the equation by setting a time step. Specifically, it concerns how the dynamics of the continuous differential equation are transferred to the discrete dynamical system generated by the numerical method applied to it. These issues are related to the topological coupling of discrete dynamical systems and to the shadowing

property (also known as the pseudo-orbit tracing property). Another significant mathematical issue involves using neural networks to represent multidimensional mappings, for which dependencies are known only from a limited number of data points in the learning sequence. Intensive theoretical research in this area was conducted in the second half of the 1980s and into the 1990s. Starting with established mathematical results, primarily Kolmogorov's theorem on the approximation of functions of multiple variables by functions of a single variable, it was demonstrated that any continuous function on a compact set with arbitrarily small but non-zero error can be approximated by a layered neural network with one hidden layer, neurons with a logistic activation function in the hidden network, and a linear activation function in the output layer. This implies that such a network is a universal approximator.

Building on the foundational studies of artificial neural networks, another important research area is their application in modelling neurophysiological structures and processes, as well as mental processes – both healthy and pathological (Duch, 2000). Neural networks have proven useful for modelling associative memory and investigating its robustness against damage. For instance, researchers can study how the removal of a certain number of neurons impacts memory. Similarly, damage to sensory pathways, which alters the organisation of cortical maps, can be modelled using self-organising maps, such as those developed by Kohonen. Neuronal models have also been instrumental in understanding how disruptions in excitatory and neuronal inhibition interactions can lead to severe memory impairments. Moreover, the phenomenon of parasitic attractors in recurrent networks has provided insights into certain mental disorders, such as hallucinations and delusions. This research area continues to advance rapidly.

Neural networks have multiple applications as artificial intelligence systems. For example, in technology and industry, they are used for controlling and identifying dynamic objects. In the latter case, a common task is to develop a non-linear representation that approximates a given technical setting. Layered neural networks, which are universal approximators, can effectively address the problem of non-linear identification. Additionally, neural networks are employed for monitoring and diagnosing complex technical devices. A common approach involves constructing neural classifiers based on measurement data to evaluate the state of

these devices in real-time. In the medical field, neural networks are used for medical image recognition and to improve diagnostic processes. In economics, neural networks forecast market trends, such as predicting demand and supply levels for various goods and services. Particularly notable is their application in predicting electricity demand, given the challenges associated with large-scale electricity storage. An additional difficulty is investment in renewable energy sources, which, especially in the case of wind power plants, exhibit a strong dependence on weather conditions. Neural networks are also used to support decision-making and in recognising and classifying market behaviours.

Polish scientists are making significant contributions to the field of artificial neural networks, advancing both theoretical frameworks and practical applications. Jacek Żurada (born in 1944), who has been working in the United States since 1979, focuses on learning algorithms for layered networks and the stability conditions of these networks. Ryszard Tadeusiewicz (born in 1947), affiliated with AGH University of Krakow throughout his academic career, specialises in modelling nervous system structures and applying neural networks to image classification and analysis, particularly in medical imaging and industrial contexts. Józef Korbicz (born in 1951) focuses on the practical applications of neural networks in technology and industry. Leszek Rutkowski (born in 1952) explores the theoretical foundations of neural networks, including neuro-fuzzy hybrid systems and their applications in image recognition and control systems. Włodzisław Duch (born in 1954) investigates the use of neural networks for modelling neurological and mental processes in both healthy and pathological conditions (Duch, 2000).

Systematic reflection with conclusions and recommendations

Artificial neural networks are specific cybernetic systems inspired by well-defined biological prototypes. However, these prototypes are so complex that, given the current state of computer science and electronics, we can only model a small subset of the nervous system's aspects. Despite advances in neurophysiology, our understanding of the nervous system, including individual neurons and the role of glial cells, remains

incomplete. Nevertheless, models of the structure and function of the nervous system as well as mental functions have made significant contributions to neurophysiology and psychiatry, and further research in this area is anticipated. It is important to note, however, that many scientists are sceptical about the feasibility of creating an artificial intelligence system – particularly one based on neural networks – that has capabilities comparable to human intelligence, especially in terms of broad creative functions. On the other hand, regardless of their biological origins, artificial neural networks have proven to be valuable tools for solving a wide range of problems. One notable disadvantage is their cumbersome learning process and the fact that they operate as computational systems where knowledge is not explicitly represented. This lack of explicit representation makes it challenging to trace the reasoning process that connects the system's inputs with its outputs. However, a significant advantage of neural networks is that the knowledge required to begin solving the problem is minimal. It is sufficient to provide a representative set of examples for the learning sequence.

Artificial intelligence systems, including neural networks, enable the rapid analysis of large amounts of diverse data and the implementation of increasingly autonomous automated systems. While this offers great opportunities for technological advancement and solving new civilisational challenges, it also poses risks related to its use as a dangerous tool for surveillance and societal control.

References

Bielecki, A. (2019). Models of Neurons and Perceptrons: Selected Problems and Challenges. *Studies in Computational Intelligence*, *770*. Cham: Springer.

Duch, W. (2000). *Sieci neuronowe w modelowaniu zaburzeń neuropsychologicznych i chorób psychicznych*. In: W. Duch, J. Korbicz, L. Rutkowski, R. Tadeusiewicz (ed.), *Sieci neuronowe* (pp. 589–616). Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Hebb, D.O. (1949). *The Organization of Behaviour, a Neuropsychological Theory*. New York: Wiley.

Fields, R.D. (2009). *The Other Brain*. New York: Simon & Schuster.

Hertz, J., Krogh, A., & Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation.* Boca Raton: CRC Press.

Hodgkin, A.L., & Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, *117*(4), 500–544.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Kononen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Mańdziuk, J. (2000). *Sieci neuronowe typu Hopfielda.* Warszawa: Akademicka Oficyna Wydawnicza EXIT.

McCulloch, W.S., & Pitts, W. (1943). A logical calculus for the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge: MIT Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Tadeusiewicz, R. (1993). *Sieci neuronowe*. Warszawa: Akademicka Oficyna Wydawnicza.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Systems.* Ph.D. Thesis. Cambridge: Harvard University.

Widrow, B., & Hoff, M.E. (1960). Adaptive switching circuits. In: *1960 IRE WESCON Convention Record*, *4* (pp. 96–104).

Mariusz Flasiński
Jagiellonian University in Krakow
https://orcid.org/0000-0002-0910-4928
https://doi.org/10.35765/slowniki.491en

Artificial Intelligence

Summary

DEFINITION OF THE TERM: This section presents various definitions of artificial intelligence and a typology of different AI systems.

HISTORICAL ANALYSIS OF THE TERM: This section offers a chronological overview of the main AI paradigms, beginning with the 1950s. In the domain of symbolic artificial intelligence, various approaches to the construction of AI systems are discussed, including cognitive simulation, a logic-based approach, and a knowledge-based approach. In the domain of computational intelligence, mathematically inspired models, connectionist models, and evolutionary computation are explored.

DISCUSSION OF THE TERM: This section details the most important methods and models of artificial intelligence, including state space search methods, models of mathematical logic, rule-based knowledge models, structural models of knowledge representation, syntactic pattern analysis, image recognition, genetic algorithms, genetic programming, and artificial neural networks.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: This section explores the fundamental methodological challenges in constructing 'intelligent' systems. It introduces two key dichotomies: strong AI vs. weak AI, and general AI (AGI) vs. specific AI. Additionally, it discusses the concept of artificial superintelligence.

Keywords: Artificial Intelligence (AI), symbolic artificial intelligence, computational intelligence, strong artificial intelligence, weak artificial intelligence

Definition of the term

The literature offers numerous definitions of the term ‘artificial intelligence’ (AI), which reflects its historical evolution and the lack of consensus on its scope and subject area. Below, we present two definitions from basic introductory textbooks on AI. The first definition attempts to capture the meaning (essence) of artificial intelligence, while the second aims to generalize various definitions that have emerged over the years as the research field has developed.

In his *Introduction to Artificial Intelligence,* Flasiński (2016) identifies two main meanings of the term. In the first, it means a common research area of computer science and robotics in which the research goal is to develop systems (computer and/or robotic) to perform tasks that require intelligence when performed by humans (sometimes such systems are referred to as intelligent systems). In the second, it is a feature of artificial systems that enables them to perform tasks that require intelligence when they are performed by humans. AI in the second meaning is inherently tied to human intelligence, whose detailed exploration is left to disciplines like philosophy, psychology, and cognitive science.

Russell and Norvig (2020) introduce a taxonomy of AI definitions based on two criteria: 1) whether the focus is placed on the (intelligent) behaviour exhibited by these systems or on the mental-cognitive processes that are simulated within them; 2) whether we benchmark the behaviour of the system and mental-cognitive processes against human beings or against rationality, understood here as ‘ideal’ intelligence (in action). Based on these criteria, we obtain the following typology of AI systems:

- systems that think like humans, e.g., the definition in Haugeland, 1985,
- systems that behave like humans, e.g., the definition in Rich & Knight, 1991,
- systems that think rationally, e.g., the definition in Winston, 1992,
- systems that behave rationally, e.g., the definition in Nilsson, 1998.

To manage this diversity of definitions, Russell and Norvig define artificial intelligence through the concept of an agent as it is understood in computer science. An agent is a system that perceives its environment through sensors and interacts autonomously with it via effectors to

achieve desired goals. If it can improve its behaviour through learning and/or knowledge acquisition, it is termed an intelligent agent. Examples of intelligent agents include, among others, humans, robots, and chatbots. Thus, Russell and Norvig (2020, p. viii) define AI as the study of constructing intelligent agents.

Historical analysis of the term

The historical analysis in this section is based on the work of Flasiński (2016). The origins of artificial intelligence (AI) as a research field date back to the mid-1950s. In 1955, during their research at Carnegie-Mellon University, Allen Newell and Herbert Simon constructed the Logic Theorist system. This system could generate mathematical proofs for nearly forty theorems from Alfred N. Whitehead and Bertrand Russell's seminal work *Principia Mathematica*. The Logic Theorist is considered the first AI system, marking 1955 as the beginning of AI research and the emergence of the cognitive simulation paradigm, which posits that intelligent systems should simulate human mental-cognitive processes.

A year later, in 1956, at a Dartmouth College conference on the mechanisation of mental processes, John McCarthy of MIT proposed another significant paradigmatic approach based on mathematical logic. This approach involves constructing intelligent systems based on various models of mathematical logic. McCarthy suggested using lambda calculus, (developed in the 1930s by Alonzo Church and Stephen C. Kleene) and constructed the LISP programming language based on this calculus, which became the first standard language for AI. In 1973, Alain Colmerauer and Philippe Roussel from the Aix-Marseille University expanded this approach with first-order logic and resolution-based inference as a model for simulating reasoning in AI systems, and with their PROLOG programming language, the second standard programming language for AI.

Both approaches aim to simulate intelligence per se, i.e., general intelligence. In contrast, research conducted at Stanford University in the mid-1960s by Edward Feigenbaum and Joshua Lederberg had a less ambitious goal: it focused on constructing a system for solving complex problems in well-defined application areas (in this particular

research, they wanted to identify the molecular structure of unknown organic compounds through analysis of spectroscopic spectra performed by the DENDRAL system they had constructed). These systems are constructed based on databases of specialist knowledge acquired from experts in the problem domain and are thus called expert systems, and this approach is called a (domain) knowledge-based approach. Two significant approaches with vast practical applications in AI have evolved from this paradigm: syntactic (structural) pattern recognition and structural models of knowledge representation. The former (Flasiński, 2019) builds upon Noam Chomsky's concept of generative grammars and finds applications in diverse fields, such as image and computer vision analysis, bioinformatics, and automatic signal recognition (medical applications such as ECG, EEG, industrial, seismic, etc.). The latter is rooted in Roger Schank's linguistic theory of conceptual dependency. and its structural models are instrumental in constructing large knowledge bases used in natural language processing systems, including chatbots and internet search engines.

The five paradigmatic approaches discussed above belong to one of two sub-areas. One is symbolic AI, which assumes that knowledge can be explicitly represented in an intelligent system, that knowledge can be represented symbolically, and that simulation of mental-cognitive operations can be performed by means of formal operations on symbolic notations. The second sub-area of AI is computational intelligence (CI), in which it is assumed that knowledge can be represented by numerical information, that knowledge processing is primarily based on (numerical) computation, and that knowledge in the system can be implicit. The three paradigmatic approaches discussed below fall under CI.

The oldest computational intelligence approaches are models inspired by mathematics, as the origins of pattern recognition (not to be confused with syntactic pattern recognition), which is the largest group of methods (and practical applications) within these models, dates back to the 1930s, thus pre-dating the formal emergence of artificial intelligence as a research area. A common feature of these models is the use of various branches of mathematics (mainly mathematical analysis and probability calculus) to solve problems: object/event/process recognition; automatic partitioning of large sets of objects into subsets representing specific categories (each subset contains only objects definitively falling under

the corresponding category); inference based on various logics (e.g., multi-valued logics or fuzzy logics), non-monotonic schemes, or statistical-like models when knowledge is uncertain or incomplete or when categories are defined in an ambiguous way. These models are limited by the fact that the objects/events/processes analysed can generally only be characterised by numerical (and not symbolic, categorical, or qualitative) data.

Connectionist models are the most prominent and widely implemented representative of the second approach in computational intelligence; they are artificial neural networks designed to simulate the operations of biological neural networks, such as the brain. Frank Rosenblatt pioneered this approach by constructing the first artificial neural network, the Perceptron, in the early 1960s. There is an incorrect conception, based on the extensive practical applications of neural networks, that the field of artificial neural networks equates to the field of artificial intelligence. Despite the significant success of artificial neural networks in solving numerous problems, this approach also has inherent weaknesses, primarily related to the implicit nature of the knowledge they represent.

The third approach of computational intelligence is evolutionary computing, in which genetic algorithms stand out as its most renowned method. Genetic algorithms experienced rapid development following the 1975 publication of John Holland's monograph, *Adaptation in Natural and Artificial Systems*. The foundational premise of evolutionary computing is twofold: first, it assumes the feasibility of constructing a space encompassing potential solutions to a given problem; second, it posits the ability to simulate evolutionary processes akin to those observed in natural species, in which specimens (better or worse adapted to the 'environment' of this space) correspond to (better or worse) solutions to the problem at hand. The instrumentarium of Darwin's evolutionary theory (populations, genotypes of individuals, phenotypes of individuals, selection processes of individuals, crossover of individuals, mutations of genotypes, etc.) is used in the construction of intelligent systems based on evolutionary computation in implementing relevant computational operations (and data structures). In the framework of evolutionary computing, alongside genetic algorithms, complementary models have been developed, including evolutionary strategies (used in numerical

optimisation tasks), evolutionary programming (in which any way of representing the genotype of an individual is allowed, not only sequential representation), and genetic programming (which is aimed not merely to find the solution to a given problem but primarily to find a solver, i.e., a metaschema, of a given problem (Flasiński, 2016). These concepts will be discussed in the next section.

In summary, the aforementioned paradigmatic approaches in AI research cover all this area. Since the 1950s, in terms of diverse methodologies used as a starting point for the construction of intelligent systems, AI has developed into a heterogeneous research area. These methodologies have been based on different scientific disciplines, including (cognitive) psychology, philosophy (of mind), logic, linguistics, mathematics, neurophysiology, and biology. From a methodological perspective, AI is a conglomeration of divergent research sub-areas unified by the goal of constructing computer systems (robotic systems) capable of solving problems which require intelligence in humans.

Discussion of the term

This section will identify the fundamental issues and challenges encountered when constructing (computer) artificial intelligence systems according to the paradigmatic approaches discussed in the historical perspective above. Before delving into these issues, it is essential to introduce a fundamental methodological principle in AI system construction. AI systems are computer systems designed to simulate (i.e., roughly reproduce) the intelligent action (action is understood here as the performance of certain activities, as well as – purely theoretical – mental-cognitive activity) of a certain object (e.g., a human) or a set of objects (which is the case in, e.g., evolutionary computing). This means that constructors of AI systems design them as computer simulators using models of intelligent activity ‘provided’ by (cognitive) psychologists, philosophers (of the mind), logicians, linguists, mathematicians, neurophysiologists, biologists, and representatives of other scientific disciplines that may be helpful in constructing AI systems. In the article, the problem domain of AI and the basic ‘open problems’ in this area of research will be identified and characterised according to the

aforementioned division into two main subareas, i.e., symbolic artificial intelligence and computational intelligence, with separate reference to each group of methods.

The methods within the cognitive simulation approach, known as state space search methods, are based on the following assumptions: First, we must define the rules for constructing an abstract (representation) model of description for the problem for which we want to simulate a solution. This implies the ability to generate an abstract model for each instance of the problem. For certain problems, this task is trivial, e.g., for games, mathematical problems, etc., because their problems are precisely defined in a highly formalised way. However, problems that are difficult to formalise present a fundamental (methodological) limit to the applicability of this approach. Humans can solve these problems because, according to cognitive psychology, they can create (abstract) mental representations (cognitive structures) to which abstract models of description in artificial intelligence systems correspond. Thus, the future development of cognitive simulation methods in AI will heavily depend on findings in cognitive psychology.

The second major limitation in developing cognitive simulation methods lies within computer science. Once a problem can be represented with an abstract model of description, the next stage involves defining the state space, which, from a mathematical perspective, is a graph. Each node of the graph represents a 'state' (step) in the problem-solving process (these states include the initial states corresponding to situations in which problem solving begins, e.g., in a game of chess there is one initial state which represents the initial configuration on the chessboard, and the final states (goal states) usually correspond to situations in which the problem has been solved). Each edge of the graph (edges connect two 'states'-nodes) corresponds to a single action in solving the problem (in a game of chess, this would be a single move). Of course, the state space is a theoretical concept which, for a non-trivial problem, becomes a huge graph. This graph, or rather its 'skeleton' in the form of a structural tree (i.e., an acyclic graph), is 'partially' generated by a specific cognitive simulation method during the solving of a particular instance of the problem. Due to the potentially enormous size of such a graph, the fundamental challenge lies in developing a strategy for generating this graph in such a way that it is as small as possible. This

issue constitutes the fundamental second 'open problem' within cognitive simulation.

It can be concluded that, in constructing systems for inference based on models of mathematical logic, the development of effective artificial intelligence methods has been remarkably successful. This is largely attributed to the profound progress in mathematical logic made in the 20th century, the outcomes of which have been successfully 'consumed' within the field of artificial intelligence. This does not mean that important and interesting research is not conducted in the area of the applications of logic in artificial intelligence (e.g., research in fuzzy, non-monotonic, and temporal logics, spatial and causal inference, and common-sense reasoning, etc.), but it seems that research in this area does not encounter fundamental barriers to its development, unlike in some other approaches to artificial intelligence.

A parallel can be drawn between methods based on mathematical logic and the sub-area of models based on rule-based knowledge. Artificial intelligence systems constructed based on these models, known as expert rule-based systems, are used to codify knowledge limited to a well-defined area of applications, which in cognitive psychology is called procedural knowledge. From a computer science perspective, however, such systems work in a similar way to systems based on mathematical logic because the set of rules in such a system corresponds to a set of logical formulas structured as implications. Expert rule-based systems are probably the broadest class of artificial intelligence systems implemented for practical applications (in industry, commerce, services, management, etc.). The theoretical assumptions for the construction of such systems were developed in the 1970s and 1980s and are still successfully used today to construct programming languages/environments (or rather languages/environments of knowledge engineering). A certain advantage of this approach over the mathematical logic approach is the greater readability (understandability) of rule-based knowledge bases compared to knowledge bases that consist of mathematical logic formulas.

While rule-based knowledge models are used to represent and infer on the basis of procedural knowledge, structural knowledge representation models are primarily used to represent declarative knowledge as it is understood in cognitive psychology. The approach based on these

models also saw its most dynamic growth in the 1970s and 1980s, and the theoretical foundations developed in collaboration with psychologists, linguists, and psycholinguists are still used in the construction of large knowledge bases widely applied today. The best-known models in this area include the semantic networks by Allan M. Collins and Ross Quillian, used for constructing ontologies (i.e., formal representations of domain knowledge based mainly on relations between concepts/categories); frame systems (structural knowledge representations based on stereotypical schemas) by Marvin Minsky, and scripts (representations of structural procedural knowledge based on stereotypical event sequences) by Roger Schank and Robert P. Abelson. Regarding structural models of knowledge representations, it seems that the possibility of further development in this research area is now 'closed'.

In syntactic pattern analysis grounded in Noam Chomsky's concept of generative grammars, despite the wide range of applications mentioned in the previous section, the fundamental challenge of grammatical inference remains unresolved (even with the development of a number of efficient methods for certain subclasses of problems). This issue epitomises the complexity of problems in artificial intelligence and can be seen as an attempt to construct an algorithm that would reconstruct a generative grammar (and thus a formalism that defines this unknown language) based on a sample of sentences (or other formal patterns) belonging to a formally undefined language. This grammar should generate a sample (a simpler version of this problem) or should generate all the (implicit) sentences of this language (a more difficult version of this problem). The impossibility of constructing such algorithms (for various classes of generative grammars) is currently the most serious fundamental barrier to the further development of methods within this AI approach (Flasiński, 2019).

The foundational models and methods of the oldest sub-area within computational intelligence (i.e., pattern recognition, also called image recognition) were established in the 1960s and 1970s. Methods such as linear discriminant analysis, Fisher's linear discriminant, Bayes classifier, the k-NN method, the support vector method, the decision tree classifier, and the k-means clustering algorithm are not only standard models of artificial intelligence today but are also considered 'classics' of computer science. Although the research in this approach can also

be regarded as 'closed', interesting theoretical and applied results are still obtained in research conducted at the intersection between pattern recognition, machine learning, and artificial neural networks.

In evolutionary computing, genetic algorithms and evolutionary strategies stand as well-established theories which are standard models in this area. However, the situation is more difficult for the other two groups of methods, namely evolutionary programming and genetic programming. The aim of the construction of AI methods within these two groups is not merely to develop – in case of a difficult problem – a method which would make it possible to find a satisfactory (according to the adopted criteria) solution to this problem instance. Rather, the aim is to find a metamethod that would generate a 'solver' for all problems of a given class. Potential applications of such an ambitiously formulated problem would cover an artificial intelligence system that writes computer programs, designs electronic circuits, and 'discovers' new pharmaceuticals. Significant breakthroughs are anticipated in this research area.

Artificial neural networks are the best-known group of computational intelligence models. This group has experienced the most rapid development over the past decade, addressing a hitherto fundamental methodological barrier for artificial neural networks: their ability to process symbolic information. An artificial neuron is, by its very nature as a model of a biological neuron, used to process (continuous) signals, i.e., numerical information. For the first fifty years of using artificial neural networks to solve practical problems in AI, this was a fundamental obstacle which prevented exploitation of the enormous potential of neural networks that stems from their ability to remember huge data sets. While initial attempts to apply neural networks to natural language processing (i.e., information of a symbolic nature) began in the 1990s (e.g., Jeffrey L. Elman's recurrent networks), a significant breakthrough occurred in 2017. In this year, the concept of a transformer based on advanced computational architecture using modules of multilayer neural networks was developed to overcome this fundamental methodological barrier and enable the construction of Large Language Models (LLMs), of which the Generative Pre-trained Transformer (GPT) and the ChatGPT chatbot are good examples (Yurafsky & Martin, 2023). Consequently, in the near future, the field of artificial intelligence is poised for dynamic growth in this area.

Systematic reflection with conclusions and recommendations

Artificial intelligence is the most rapidly developing research area in the discipline of computer science. This is true for all groups of methods mentioned in the previous section (although the history of the development of AI reveals periods in which certain groups of methods were clearly in a state of regression, while others developed at a very fast pace). A fundamental question arises here that has significant implications for not only practical applications of intelligent systems (computer science and robotics) but also philosophy, psychology, and cognitive science: can artificial intelligence match (or surpass) 'natural' intelligence, i.e., human intelligence, given that human intelligence is considered the pinnacle of intelligences known to us empirically, i.e., in the natural world? (Of course, philosophers from certain schools might posits that there are higher intelligences, e.g., separate substances). Scientists working in the field of artificial intelligence are divided into two groups on this issue. Proponents of the strong AI approach argue that there are no obstacles to constructing a computer-based intelligent system capable of acting like the human brain with all its mental and cognitive activities (including mental phenomena, e.g., consciousness, motivation, attentiveness, feelings, etc.). Conversely, advocates of the weak AI approach believe that although an intelligent computer system can simulate some mental-cognitive activities (such simulations are undoubtedly useful in testing hypotheses about the functioning of human perceptual, cognitive and mental processes), we will not be able to fully reproduce these processes in a computer (robotic) system. (An overview of the various strong AI and weak AI perspectives within the philosophy of mind can be found in Flasiński, 2016).

In the context of the above question regarding the possibility of creating artificial intelligence as a full realisation of human intelligence, it is essential to consider whether any fundamental barriers might impede such a simulation. Before attempting to answer this question, it is worth reflecting on the concept of intelligence itself. Flasiński (2016) synthesised various definitions of intelligence from cognitive psychology literature, defining intelligence as a set of abilities that allow one to:

- adapt to a changing environment,
- perform a cognitive activity consisting of creating and operating on abstract structures (which basically correspond to cognitive structures and mental representations as understood in cognitive psychology).

(An overview and characterisation of these abilities based on cognitive psychology literature can be found in Flasiński, 2016).

To identify potential barriers to the development of AI, we should focus on the second component of this definition as modern intelligent systems largely meet the first component, and no fundamental barriers seem to exist that would limit further development of this aspect of AI.

The identification and characterization of fundamental barriers in the development of artificial intelligence (AI) have been explored by, e.g., Flasiński (2016), who used the classification of fundamental cognitive operations based on the epistemological views of Aristotle and Saint Thomas Aquinas. Following their views, three generic mental-cognitive operations (Latin: *tres operationes rationis*) can be identified: comprehending concepts (in Latin: *indivisibilium intelligentia*), pronouncing judgements (in Latin: *componere et dividere*), and performing reasoning (in Latin: *ratiocinare*), i.e., moving from one judgement to another according to the rules of logic.

Of these three generic operations, only reasoning has been satisfactorily implemented in AI systems, largely due to the outstanding development of logic in the 20th century. The results of this development address all fundamental issues relevant to constructing efficient algorithms that simulate various aspects of human reasoning. First, such algorithmisation was possible thanks to the achievement of a very high degree of formalisation in propositional logic, resulting from great attention to the precise definition of the syntax of this logic (which was undoubtedly influenced by the emergence of analytic philosophy, with its strong emphasis on language). Second, models of (mathematical) logic were introduced based on the paradigm of symbolic computation that later greatly facilitated the work of computer scientists who constructed computer tools for the construction of AI systems (e.g., the LISP language, based on Alonzo Church's lambda calculus, and the PROLOG language, based on the method of resolution inference in first-order logic). Third, the development of the semantics of logic/model theory

(thanks to prominent 20th-century logicians such as Alfred Tarski, Thoralf Skolem, and Kurt Gödel) was crucial for the applicability of propositional logic to solve problems related to the physical reality around us.

The situation is much more challenging for the second generic mental-cognitive operation, i.e., pronouncing judgements (propositions). Following Flasiński (2016), we will analyse this issue using Kant's division of propositions. The generation of analytical a priori propositions is performed in AI systems by referring to knowledge stored in knowledge bases, which in AI systems are the equivalent of databases in (standard) computer systems. Knowledge bases are constructed using various artificial intelligence approaches (discussed in the previous section). For example, they can take the form of structured knowledge representations (e.g., semantic networks, frame systems, scripts) or sets of first-order logic formulas (or other models of mathematical logic). Thus, the generation of analytic a priori propositions in an AI system relies on the (non-creative) use of pre-existing knowledge accumulated in the system, which aligns with the Kantian notion of analytic a priori propositions. Synthetic a posteriori propositions are formulated based on prior experiences. In AI systems, the generation of such propositions is done by simulating a logical inductive inference scheme (inference rules). Such a scheme is used, e.g., in training artificial neural networks, training classifiers in pattern (image) recognition, or grammatical inference in syntactic pattern recognition. Hence, we are able to simulate the generation of these propositions in AI systems. However, a fundamental barrier arises with synthetic a priori propositions because neither cognitive psychology nor the philosophy of mind (or epistemology in general) can explain the mechanism behind the formation of such propositions. Typically, the formulation of synthetic a priori propositions is attributed to concepts that are difficult to algorithmicise, such as insight (into the essence of things), intellectual intuition, or pure perception. The situation is most challenging when it comes to the operation of comprehending concepts, understood as a process of abstraction. Neither cognitive psychology nor the philosophy of mind can adequately describe or explain the mechanism of conceptualisation here. In the 1980s and 1990s, some hopes in this area were linked to interesting concepts of 'embodied categorisation' that were introduced in cognitive linguistics by George Lakoff and Ronald Langacker. Unfortunately, although these

concepts seemed promising as a starting point for constructing algorithms to simulate human comprehension, nearly half a century later they have still not been integrated into the field of artificial intelligence.

Summing up, it should be said that the two fundamental barriers related to the comprehension of concepts and the formulation of synthetic a priori propositions remain significant obstacles to realising the postulate of strong artificial intelligence, at least at this stage of AI research.

To complete this reflection on artificial intelligence, two additional issues should be addressed. The first concerns the distinction between artificial general intelligence (AGI) and narrow AI. Narrow AI systems are constructed in narrow areas of simulation of certain human mental-cognitive abilities (e.g., natural language processing, image recognition, expert systems for medical diagnosis). In contrast, AGI systems should be understood as systems characterised by general intelligence, i.e., capable of solving any intellectual-cognitive tasks that a human can solve. It should be noted that (hypothetical) AGI systems are not typically expected to exhibit the other mental phenomena listed in the definition of strong AI, such as consciousness or attentiveness. The second issue relates to the concept of superintelligence, which refers to artificial intelligence that surpasses human intelligence. (Hypothetical) artificial superintelligent systems are often linked to the concepts of technological singularity proposed by Ray Kurzweil and the transhumanism of Nick Bostrom. However, this issue seems to belong to the realm of futurology.

References

Flasiński, M. (2016). *Introduction to Artificial Intelligence*. Cham: Springer.

Flasiński, M. (2019). *Syntactic Pattern Recognition*. New Jersey–London–Singapore: World Scientific.

Haugeland, J. (ed.) (1985). *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press.

Nilsson, N.J. (1998). *Artificial Intelligence. A New Synthesis*. San Francisco: Morgan Kaufmann.

Rich, E., & Knight, K. (1991). *Artificial Intelligence*. New York: McGraw-Hill.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence. A Modern Approach*. Boston: Pearson.
Winston, P.H. (1992). *Artificial Intelligence*. Reading, MA: Addison-Wesley.
Yurafsky, D., & Martin, J.H. (2023). *Speech and Language Processing*. Upper Saddle River: Prentice-Hall.

Dariusz Adamek
Jagiellonian University in Krakow
https://orcid.org/0000-0002-5570-2259
https://doi.org/10.35765/slowniki.481en

Emotional reactions

Summary

DEFINITION OF THE TERM: Emotional reactions[1] are a form of complex and coordinated motor phenomena that are supported by functional changes throughout the entire organism, including the nervous system itself, particularly the autonomic (vegetative) system. These changes are crucial for avoiding survival-critical threats or for achieving important life goals, both in a physical-real and mental/conceptual way.

HISTORICAL ANALYSIS OF THE TERM: The terms 'emotional reactions' and 'emotional brain' have appeared in scientific literature for many years, but the author first encountered these concepts in 2004, when T. Dalgleish published his review article entitled *The emotional brain* (Dalgleish, 2004).

DISCUSSION OF THE TERM: Emotions and emotional reactions are complex and multifaceted issues that necessitate interdisciplinary research spanning fields like neurology, psychology, and zoology. This article focuses on selected neurobiological aspects of emotions and emotional reactions. Even 'primitive' animals, including invertebrates and unicellular organisms, exhibit behaviours that could be interpreted as

[1] In certain contexts, the term 'responses' may be more appropriate, particularly when addressing the conscious aspects of emotions. However, the considerations in the article aim to emphasise the predominant role of subconscious brain and mind processes that continuously shape our actions. Consequently, 'reactions' may better capture the true nature of what we understand by and associate with emotions. Hence, both terms are used in the article.

manifestations of emotions, though the definition of these terms plays a critical role in their interpretation.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Since Paul Broca's discovery of a small region in the left frontal lobe, now known as Broca's area, which, when damaged, causes speech disorders (aphasia), the concept of specific brain regions, or 'centres' responsible for distinct functions, including emotions, has emerged. This discovery laid the foundation for neurology as an independent branch of medicine. When it comes to emotions, it seems impossible to isolate specific 'centres'. Perhaps the entire brain is 'emotional'.

Keywords: types of emotions, motivations, triune brain, Papez circuit, interdisciplinary research

Definition of the term

Emotional reactions are complex and coordinated motor phenomena that are supported by functional changes throughout the entire organism and in the nervous system itself, particularly the autonomic (vegetative) system.

A. First, these reactions enable, facilitate, and enhance an individual's effective response to life-critical needs or challenges, such as avoiding danger, surviving threatening situations (e.g., encountering a predator, rival for food, or a mate), obtaining food, procreation, quenching thirst, and caring for offspring.
B. Second, though more open to debate, emotional reactions enable and facilitate an organism's adaptation to environmental or internal changes in a way that minimises damage or loss and optimises the likelihood of future success.
C. Third, though also debatable, emotional reactions are fundamental to the 'motivational engine'. At a basic level, they are necessary for generating the drive or desire to take action, though they are not sufficient on their own.

There is considerable confusion surrounding the conceptual framework related to emotions and their manifestations. The debate centres on defining different types of emotions and distinguishing between those that are primary or simple and those that are secondary or complex. For instance, questions arise regarding the nature of concepts like affects and drives, or which emotions should be considered primary. Terms such as 'core affect' and 'emotional intelligence' are also subjects of ongoing discussion. According to Jaak Panksepp, mammalian brains contain distinct emotional systems, four of which – seeking, fear, rage, and panic – are present from birth, while others such as lust, care, and play develop later. Notably, this classification does not include emotions like hatred, jealousy, or love. This topic is further explored in the book edited by Davis and Panksepp (2017).

It seems reasonable to argue that much like the study of consciousness – often best approached by examining its correlates through interdisciplinary observation – the optimal approach to understanding emotions is to analyse their 'manifestations', i.e., emotional reactions. This is similar to the challenge posed by consciousness, exemplified by

T. Nagel's famous philosophical dilemma, "What is it like to be a bat?", which was thoroughly examined by J. Bremer (Bremer, 2005). One could reframe this question as "What is it like to be a sad bat?" (or a happy, angry, etc., bat). While emotions are undoubtedly communicable, what do they truly mean to an individual human or animal?

From an evolutionary perspective, when emotional reactions/manifestations are considered within a biological, including neurobiological, context, the thesis or definition given at the beginning of this section seems plausible. I prefer the term 'emotional reactions' rather than 'emotions', as these reactions can be observed and it is only through them that we can infer the emotions of a human or other animal. However, I acknowledge that throughout this text the terms 'emotions' and 'emotional reactions' are frequently used interchangeably. Referring to the aforementioned definition of 'emotional reactions', let us examine the key aspects outlined in points A, B, and C.

While the examples illustrating point A are fairly straightforward, point B can be exemplified by emotions or states such as sadness or anxiety. These emotional states may play a role in survival during or after a highly stressful situation in which the individual cannot actively counteract negative factors or threats. In such cases, individuals may attempt to minimise the likelihood of drawing attention to themselves. For instance, remaining still – a behaviour known as 'freezing', is often an adaptive strategy to avoid detection by predators. Alternatively, displaying sadness may evoke empathetic responses, such as compassion or empathy, ideally soliciting help from others within a group and potentially increasing the chances of future success (e.g., in reproductive opportunities).

The motivational role of emotions (point C) – particularly positive ones – can be complex and is not always straightforward. It is generally accepted that emotions drive behaviours such as returning to familiar childhood places or celebrating Christmas with family. But can the internal compulsion that drives salmon to the Sargasso Sea, birds to migrate to their breeding grounds, or millions of certain butterfly species to undertake annual journeys spanning thousands of kilometres also be classified as emotionally motivated? I introduce another distinction here, noting that common opinion tends to regard these behaviours as driven

more by 'innate' (?) instincts related to reproduction or survival (i.e., the acquisition of water and food). This raises several questions: (a) Is the motivation behind these behaviours fundamentally different from human motivations, such as the pursuit of food, drink, sex, or higher social status? (b) When a human pursues sex for the first time, is this fundamentally different from a salmon seeking to spawn, despite having no personal experience of the act? While the context differs when a drive, such as the instinct to procreate, is realised only once in a lifetime (as in the case of salmon), the situation changes when such acts (conditioned by intuition and drive) can be repeated. Even relatively primitive animals have access to personal experiences that either encourage or discourage certain behaviours. In this way, motivational mechanisms – positive or negative – are formed, which may be amplified or suppressed. A second condition, besides the possibility of repetition, is necessary here: memory. Specifically, acquired memory, which differs from other forms of memory, such as that based on encoded innate 'algorithms'. For example, an innate algorithm is the primary motivation that drives a newborn mammal to seek out nipples. With repeated episodes of nipple-seeking and feeding, memory mechanisms amplify the original reflex, and this memory-dependent amplification of an act (in this case, breastfeeding) becomes an emotion (or emotional response) in the sense of point C. This emotion serves as a motivating or 'driving' force. However, it is important to distinguish these motivations from those behind conscious human actions, especially long-term ones, which are based on multiple memories and directed towards specific goals.

Historical analysis of the term

In addressing questions (a) and (b) regarding the differences between human emotional responses and those of animals, it is worth noting the long historical association between emotions and the heart – initially understood quite literally, but later metaphorically. This linkage likely dates to antiquity when emotions and emotionality were often equated with the heart. Literature and art abound with examples where the heart symbolises emotions, particularly those seen as motivational forces, often contrasted with reason or the mind (the latter typically being viewed

as ‘cool’ compared to the ‘hot’ heart). A favourite example of the motivational role of the heart – or rather, emotions – comes from the opening of Saint Augustine’s *Confessions*, which states that “Thou madest us for Thyself, and our heart is restless, until it repose in Thee” (Augustine, 2001). This statement suggests that God designed emotions as a motivational mechanism within the human ‘project’. The definition of emotional reactions discussed above integrates what are often referred to as ‘drives’ and ‘emotions’ (Ciccarelli & White, 2021).

Paul Broca’s (1824–1880) discovery of a small area in the left frontal lobe of the brain, which, when damaged, causes speech disorders (aphasia), marked the beginning of the belief that specific brain regions (centres) are responsible for distinct functions (LaPointe, 2012). This discovery laid the groundwork for neurology as a distinct branch of medicine and is now a key part of the interdisciplinary neurosciences. For a long time, researchers have been intrigued by the brain structures responsible for emotions, particularly intense ones like love or hate. But a question arises: What constitutes the ‘emotional circuitry’ of the brain? The prevailing theory suggests that, unlike Broca’s area, it might not be possible to pinpoint a specific anatomical centre for any single emotion. Instead, emotions may involve a more diffuse network of anatomically and functionally interconnected brain structures. The first neuroanatomical concept of emotions was the Papez circuit, named after American neuroanatomist James Papez (1883–1958) (Papez, 1937). This circuit includes the following structures: the hippocampus, the fornix, the mammillary bodies, the mammillothalamic tract, the anterior thalamic nucleus, the projection fibres from the thalamus to the cingulate gyrus, and the cingulate gyrus itself. The circuit is completed by the projection fibres from the cingulate gyrus back to the hippocampus, forming a closed loop of connections.

In 2004, T. Dalgleish published a review article entitled *The emotional brain* (Dalgleish, 2004). This article, along with two other books (Purves, 2008; Bear et al., 2016), provides most of the historically established conceptual foundations and serves as a substantial inspiration for my own reflections, which I present below. In this context, a comparison of classical theories of emotion is illustrative.

The first theory, known as the James-Lange theory (proposed by William James and Carl Lange in the late 19th century), posits that

symptoms generated by the autonomic nervous system are perceived as emotions. The second theory, the Cannon-Bard theory, developed by Walter Cannon and Philip Bard in the 1920s, argues that a stimulus perceived as 'emotional' triggers an emotional response, which originates in the central nervous system and leads to changes in various organ functions and motor-somatic reactions (e.g., fleeing from danger) (Cannon, 1927).

Figuratively speaking (though this may sound somewhat simplistic), according to the James-Lange theory, if a person is running with a racing heart, elevated adrenaline levels, and increased metabolism, their brain is essentially 'experiencing' the neuronal correlates of fear. The Cannon-Bard theory suggests a reversed sequence: upon seeing a threatening animal, the brain first 'interprets' the danger and determines that fear is warranted, then activates the autonomic (vegetative) nervous system and other mechanisms necessary for fight or flight, depending on the assessment. At first glance, the James-Lange theory might seem surprising, and even sometimes counterintuitive, but in practical terms it offers a more constructive approach for scientific study in humans and animals. While we cannot know for certain what a monkey feels when it encounters a snake (the classic 'fear' model), if the monkey reacts by quickly fleeing (typically climbing a tree if one is available), we can reasonably infer that it is experiencing fear. This situation is analogous to the study of consciousness, where we cannot directly observe or define it but rely on certain 'correlates' to infer its presence.

Discussion of the term

Preliminary remarks

For the past 20 years, I have dedicated one interdisciplinary lecture in an optional course on the foundations of neuroscience for medical students to the exploration of the 'brain', or more precisely, the 'neural' mechanisms of emotion. I intentionally schedule this lecture during the week immediately before Christmas in the winter semester as I believe that this holiday, especially in Poland, is the most emotionally charged of the year. One of the key examples I present in support of this is Chopin's

Scherzo in B minor, Op. 20, which incorporates the carol *Lulajże Jezuniu [Sleep, my little Jesus]*. This piece, artistically infused with the melody of the carol, symbolises the composer's emotional 'return' to his childhood Christmases. The carol, ingeniously woven into a tumultuous array of melodies and moods that reflect Chopin's emotional state, encapsulates the associations, moods, and feelings tied to Christmas. Chopin's ability to convey his Christmas emotions through this work is remarkable, especially considering that it was composed during Chopin's first Christmas away from his homeland, making this symbolic 'return' all the more poignant. But what does this 'return' truly signify?

We will explore later how emotion serves as the core or essence of art, both in its creation and reception. To conclude our Christmas-themed introduction, let us pose a rhetorical question: What drives us to celebrate Christmas? What else could it be if not a somewhat defined brain reward system that is correlated with our festive feelings? This system is similar to the one activated when satisfying a desire or craving, as well as during emotional states like gaining satisfaction from social or economic success or sexual fulfilment. It is the same brain system that sometimes drives us hundreds or even thousands of kilometres to experience happiness for at least one Christmas day – whether through the joy of a child or the pleasure of a reunion. It may also be the same system that can be artificially stimulated by alcohol, drugs, or delicious food. We seek to experience these feelings not only in the 'real' world but also in virtual reality, dreams, or various substitutes for these profound yet simple joys. It turns out that, on the one hand, we are not so different from other animals, even those considered primitive. On the other hand, we share something with the 'restless heart' mentioned by Saint Augustine.

Macro- and micro-emotions – an attempt at differentiation

The emotional reactions discussed in points A and B above can be categorised as 'macro-emotions'. These are commonly recognised emotional manifestations, such as fear and desire, which are intuitively acceptable and have specific names. However, the scope and definition of these terms are often debated.

Emotional manifestations are not limited to these 'big' emotions. Throughout a single day, we do not simply experience three discrete moments of fear, five moments of sadness, two moments of joy, five moments of desire, and one moment of satisfaction or fulfilment. What about the rest of the time? Are we left with an emotional 'void', a state of 'zeroing out', neutrality, or indifference? From the perspective of interdisciplinary sciences that examine the relationships between behaviour and the nervous system, this is not an accurate representation of how we function. Arguably, there is no single moment when we are completely devoid of emotional activity. Instead, we continuously experience a spectrum of micro-tensions and micro-solutions ('neutralisations'), i.e., micro-emotions that occur and dissipate constantly, whether consciously or unconsciously. This concept of micro-emotions, while originally associated with fleeting facial expressions, more accurately reflects the continuous and dynamic nature of our emotional lives.

I propose the thesis that, using a chemical analogy, these are not indivisible 'atoms' that combine to form more complex emotions from which specific combinations could create the macro-emotions that can be assigned names. To use another analogy, this is not similar to the Ekman and Friesen system, which encodes specific facial expressions as the literal sum of muscle contractions. In their system, the components are distinct anatomical or functionally separate facial muscles (the authors identified 46 'action units' in the facial muscles, whose contractions 'encode' specific emotional expressions, known as the Facial Action Coding System). 'Atomic' micro-emotions are continuously present with us in every moment of life (we also experience emotions, including macro-emotions, during sleep). I conceptualise micro-emotions as fragments or partial versions of macro-emotions that utilise the same neural circuitry but are activated in a minimal or incomplete manner. Despite their partial activation, they are persistently present, creating an emotional backdrop from which 'big' emotions can arise.

The essence of both micro- and macro-emotions lies in the tensions continuously managed by the prefrontal cortex. This brain region, akin to a Bayesian computer, constantly evaluates the probabilities of events and the outcomes of our actions and plans. These tensions arise from the interplay between predictions and actual outcomes, between what is anticipated (expected or desired) and what seems probable.

The ‘players’ in this process are the countless, often unconscious associations between verbally describable events and non-verbal cues, such as smells, colours, shapes, and auditory or material stimuli.

We do not need to search for micro-emotions in a theatrical performance or a sports event, though they certainly exist there too, fluctuating every second, or even fraction of a second – like when we watch a ball in flight, wondering if it will hit the goal or land inside the tennis court lines. Micro-emotions, however, are equally present in far more mundane situations, such as whether we will catch the bus on time. Importantly, emotions, both in their expression and perception, can be conscious or unconscious. These micro-emotions are likely to be mostly unconscious. The expression and perception of emotions will be discussed below.

Arguably, one of the most emotionally charged realms where micro-emotions are present is any form of artistic expression. One could even argue that the true purpose of art – whether literary, musical, or visual – is to play with emotions, to evoke and manipulate them.

The emotional tensions created by an artist can be both macro-emotional, such as those related to a love story or the revelation of a crime’s perpetrator, or micro-emotional, which find expression in the subtle nuances of a sentence, gesture, or phrase. For me, musical examples are particularly resonant. In a piece of music, the macro-emotional play may be elusive, discernible perhaps only to the seasoned listener. However, nearly every note – and certainly every small phrase – contains micro-tensions often linked to the music’s tonality, or more precisely, to a temporary departure from that tonality, creating a desire for resolution back to the tonic. A perfect example of such micro-tension is the musical cadence, where the listener anticipates a specific resolution, often to the tonic. Composers have long played with these micro-emotions – not just those of contemporary music, filled with dissonances that challenge ‘normal’ cultural expectations of note or chord progression. This emotional play of micro-tensions has existed since the earliest days of music, even as it first broke free from rhythmic patterns. These tonal tensions, moving away from and returning to the tonic, are present in both Elvis Presley and Chopin. Studies on event-related potentials (ERPs) using electroencephalography (EEG) coupled with computer analysis have shown that even in musically untrained children the brain exhibits

different patterns of electrical activity (which reflects neuronal activity), depending on whether the musical phrase ends in an 'expected' or 'unexpected' way (Koelsch, 2013). In light of the above, I propose the following conclusions: rather than attempting deductive definitions of 'emotion', it is more fruitful to adopt an inductive, behaviourist approach. This involves an interdisciplinary analysis of behaviours and emotional reactions under specific conditions. Instead of asking "What is a particular form of emotion?", we should focus on the question, "What are its manifestations?".

A comparative neuroscience of emotions

I have already mentioned that even primitive animals, such as invertebrates and more primitive organisms like unicellular animals (depending on how we define them), display manifestations of emotion-like states. In such cases, it may be appropriate to use terms like 'proto-emotions' or even 'pre-proto-emotions'. These categories could relate to the diversity of expression, the complexity of the molecular or cellular mechanisms involved, or the sophistication of their functional systems, with the human nervous system serving as the most complex example. Focusing on animals more closely related to humans, Paul MacLean's concept of the 'triune brain' is a particularly imaginative and influential idea. This theory suggests that the human brain operates on three evolutionary 'layers' or 'floors'. MacLean referred to these as the 'reptilian' brain (which corresponds roughly to the brainstem, cerebellum, basal nuclei, thalamus, and olfactory bulb); the 'paleomammalian' brain (associated with the limbic system located deep within the cerebral hemispheres); and the 'neomammalian' brain (essentially the cerebral cortex).

MacLean's concept has been, and continues to be, the subject of criticism, particularly regarding its comparatistic anatomical and functional assumptions. However, even if the anatomical structures listed above do not strictly correspond to the levels of complexity reached by the brains of the three evolutionary categories of animals (the brain structures involved in emotional roles will be discussed below), the model of the 'triune brain' remains undeniably inspirational. It serves to remind us that even the brain structures of 'primitive' animals like

reptiles are still present within each of us. Similarly, the neuronal circuits and connections that reptiles rely on are embedded in our own brains. What fundamentally sets humans apart is the extensive development of the cerebral cortex, often referred to as the pallium. We can reasonably assume that this expansive cortex regulates and moderates the more 'primitive' mammalian and reptilian parts of our brain, particularly those emotions often considered primitive (or rather primeval) and their manifestations. Yet, this regulation is not simply a matter of suppression or inhibition. It involves an intricate modulation of our socially observable emotional reactions, allowing for heightened sensitivity and the capacity for rapid shifts in emotional states – sometimes within minutes or even seconds. Interestingly, the suppression of reptilian emotions in humans can, in certain circumstances, morph into the opposite: amplified emotional behaviours. As a striking albeit extreme example, consider how an 'enraged' animal such as a crocodile, dog, or buffalo may attack and kill another creature, including humans. Now imagine a human, overtaken by a macro- or micro-emotion, armed with a machine gun…

The most valuable insight from the triune brain model, in my view, is the recognition that these 'primitive' paleomammalian and reptilian brain structures are integral components of our neurological makeup – not as separate, alien entities but as fundamental parts of our brain. The significance of this cannot be overlooked. The cortex's pallium acts as a protective and regulatory layer, managing these 'reptilian-paleomammalian' brain functions and, by extension, the emotions associated with them. However, there are moments – experienced by everyone – when the cortex's control over these emotions is either diminished or bypassed, often unconsciously. In such instances, instead of suppression, we may even witness an intensification of those more primal emotional responses.

To illustrate this point, during lectures for students, I often highlight the anatomical similarity in body posture, such as lordosis seen in female mammals (e.g., rats). This posture, which exposes the external genital area, signals to males that the female is ready to mate. Surprisingly, we see a similar pose in many commercials featuring attractive women modelling intimate lingerie or in erotic scenes from films. This brings us to the concept of body language, which plays a crucial role in many areas, including professional success. Conscious body language

is also a powerful form of macro-emotional expression (the relationship between emotion expression and perception will be discussed below). Here is a question to ponder: What message is being conveyed when a model strikes a pose that, while not overtly sexual, could easily be interpreted as such? This interpretation may arise from a simple mental adjustment, such as imagining the figure rotated by 90 degrees. This mental transformation – likely unconscious in many cases – may still evoke similar connotations. It serves as a reminder that body language is a potent vehicle for communicating intentions and emotional states that is often more revealing than verbal communication. As a side note, political analysts frequently study the postures, reactions, and gestures of politicians, as these often provide more insight into their true emotional states and intentions than their words, which may be carefully crafted to conceal their real sentiments.

Pragmatics of emotions

The ability to accurately assess and control emotional reactions in oneself and others is crucial for achieving social success. This is where the concept of emotional intelligence comes into play. Disrupted or abnormal emotional reactions are characteristic of many psychiatric disorders and are often at the root of psychosomatic illnesses.

Unexpected emotional reactions, sometimes baffling even to the individual experiencing them, can also be linked to notions of spirituality. These reactions may be interpreted as the ‘whisper of the Evil Spirit’ or, conversely, as the influence of the ‘Holy Spirit’. While this is certainly a controversial idea and outside the conventional methodology for studying emotions, it is not far-fetched to claim that emotions often drive decisions that may contradict rational analysis. In fact, the literature frequently points out that many so-called rational decisions are merely illusions. Is the dichotomy between the rational person and the irrational person – underpinned by the supposed divide between a ‘rational brain’ and an ‘emotional brain’ – even real? Any so-called rational choice comes with an awareness of limitations: incomplete knowledge, uncertainty about the validity of assumptions, and doubts regarding the logical integrity of the reasoning process. Yet, the selection of these ‘rational’ premises is

often emotionally charged, whether consciously or not. In moments of indecision, emotions tend to tip the scales, providing the final motivation (e.g., 'this option feels more comfortable or appealing to me').

The pragmatic and social role of emotions is fundamental to forming and maintaining interpersonal bonds as well as determining our ability to cooperate within 'target groups', such as sports teams, flight crews, and more. Emotions also play a significant role in shaping political choices and influencing consumer behaviour through advertising.

Similarly, the notion of purely rational decision-making should be viewed as misleading. Concepts like 'unifying spirituality', especially in phrases like 'national spirit' or 'team spirit', are also misrepresentations. What we often refer to as the 'spirit of the nation' or even the idea of the 'nation' itself is actually a group of people who share similar emotional reactions to symbols and content such as the national anthem, patriotic songs, the flag, national colours, and the memory of heroic figures and their actions. In my view, a true 'Pole', for instance, could be defined as someone who experiences vegetative, 'positive' physiological reactions (like a quickened heartbeat or tears) to such symbols and content. The term 'vegetative', referring to reactions from the autonomic nervous system, is key here as these reactions are either uncontrollable or only minimally so. These reactions are often seen as the most authentic expressions of our emotional bonds, traditionally described as coming from the 'heart' and frequently contrasted with 'reason'.

These issues highlight the importance of how we express and communicate emotions as well as how we perceive and empathise with the emotions of others. We communicate and perceive emotions through emotional reactions, and both communication (expression) and perception (reading) can occur consciously or unconsciously. Additionally, these processes can be either voluntarily controlled or, in many cases, uncontrollable (or only partially controllable).

Emotions are expressed through several primary channels:

a. Reactions of the visceral motor nervous system (autonomic responses) – physiological changes not visible to the naked eye,
b. Speech: (including prosody as well as the content itself),
c. Facial expressions,
d. Body language and non-verbal communication (including dress style and touch).

The perception of emotions can occur both consciously and unconsciously. Unconscious emotion perception is closely linked to the concept of mirror neurons (Rizzolatti, 2004; Gładziejewski, 2011). These neurons activate in the observer's brain in a way that mirrors the motor and vegetative responses of the person being observed (called observables, a term borrowed from quantum mechanics). For instance, when watching someone being kicked in the groin, we might instinctively adopt a protective posture for this part of our own body. Mirror neurons are thought to be the neural basis for empathy, i.e., the ability to experience or understand the emotions and sensations of others. Empathy is not exclusive to humans; it is also observed in some primates and potentially other animals with 'lower' degrees of cognitive development.

Systematic reflection with conclusions and recommendations

Since the pioneering work of Paul Broca and James Papez, numerous brain 'centres' have been identified or at least hypothesised. However, we still cannot ascertain the specific functions of the most distinct 'fields' within the vast expanses of the brain, including both cortical and subcortical structures. While the Papez circuit is now considered a historical piece in the study of brain cognition, it overlaps with structures known as the limbic system, which are recognised for their significant role in regulating emotions. However, the term 'regulation' is used cautiously here as it does not necessarily pinpoint the exact origin of emotion formation or emotional states in the brain. It is generally accepted that not the entire limbic system is involved in emotion. For example, while the parahippocampal gyrus, hippocampus, and hypothalamus are part of the limbic system, their connection to emotional processes is less well-defined.

More specifically, the following brain regions are thought to play critical roles in regulating emotions:

a. the supraorbital gyri and medial prefrontal cortex,
b. the ventral part of the basal ganglia (nucleus accumbens),
c. the medial dorsal thalamic nucleus (nucleus mediodorsalis thalami), distinct from the anterior thalamic nuclei mentioned in the Papez circuit,

d. the amygdala,
e. the cingulate gyrus,
f. the septal nuclei (septal area),
g. a part of the hypothalamus.

It is important to note that these structures should be considered in the plural form as they are present in both hemispheres of the brain. However, for simplicity, we will use the singular term unless the plural is explicitly relevant. Additionally, other areas, such as the periaqueductal gray (PAG) and the ventral tegmental area (VTA), are also significant in emotional mechanisms. The PAG is involved in pain signal transmission, while the VTA plays a crucial role in the reward circuit (which will be discussed below).

Among these structures, the amygdala – a relatively large cluster of nuclei with a diameter of slightly over 1 cm – receives particular attention. It serves as a central hub for forming associations between mental and perceptual events, contributing to the creation and storage of 'emotional memory' and the assignment of valence and salience to events. The amygdala's central role is underscored by its extensive anatomical connections. It receives inputs from all sensory pathways and systems, including the olfactory structures located nearby, suggesting the important role of smell in emotional processing, especially in animals. Moreover, the amygdala is intricately linked with the entire limbic system. It has a notable connection to the hypothalamus through the stria terminalis, which is vital for coordinating endocrine and vegetative functions.

Joseph LeDoux's concept of the 'high road' and 'low road' pathways of amygdala activation (LeDoux, 1998) illustrates the multidirectional connections of the amygdala. The 'low road' is a direct pathway through which sensory experiences of all modalities activate the amygdala almost instantaneously. This rapid reaction can sometimes lead to errors due to its speed, but it can also provide crucial protection against immediate threats. For example, a sudden movement may trigger a reflexive recoil in response to a perceived danger, like mistaking a garden hose for a dangerous reptile. In contrast, the 'high road' involves sensory information passing through the cerebral cortex before reaching the amygdala. This pathway allows for more deliberate processing and can prevent unnecessary or exaggerated reactions. While the 'high road'

might prevent an overly hasty reaction, it might also delay response time, potentially putting us at risk if the perceived threat is real. Importantly, it is not possible to activate only one of these (conceptual) pathways. Even if we remain calm due to the cognitive analysis performed by the cerebral cortex, the amygdala, stimulated by the 'low road', will still send signals to the hypothalamus. This initiates the activation of the vegetative nervous system and the adrenal cortex, increasing adrenaline levels.

It is important to recall that when an animal repeatedly has a personal experience – such as eating or mating – it must be able to remember it, even if this memory is represented as a simple association of valence (positive-negative, pleasant-unpleasant, etc.). Episodic memory has been observed in animals, despite their inability to communicate it verbally. One key neurobiological mechanism (used by the neurons of the amygdala) involved in this process is long-term potentiation (LTP), which is based on glutamatergic signalling and NMDA-type glutamate receptors. For example, blocking NMDA receptors with antagonists impairs fear-conditioning responses, providing indirect evidence of their role in memory formation. The amygdala serves as an associative binder, forming what we call 'emotional memory'. While this memory can be conscious, its unconscious aspects are particularly intriguing. For instance, a deeply unpleasant event, such as being bitten by a dog during childhood, might be 'forgotten' in terms of episodic memory but could persistently influence behaviour, resulting in a heightened lifelong fear of dogs.

The Papez circuit has played a pivotal role as a neurological concept, introducing a research framework for identifying various other 'circuits' of functional brain connections as material 'substrates'. These circuits, if not directly responsible for the 'creation', are at least involved in the 'regulation' of brain-derived emotional processes – a method that remains influential to this day. For example, we now talk of the fear circuitry and the reward circuitry that incorporate additional components not mentioned above (points a–g). The reward circuitry, which acts as the primary 'reinforcer' for what the amygdala attributes positive (pleasurable) significance to, includes not only the amygdala but also the prefrontal cortex, nucleus accumbens, and the ventral tegmental area (VTA).

The nucleus accumbens and the ventral tegmental area (VTA), particularly the latter, merit detailed exploration and interdisciplinary research.

The nucleus accumbens, a sizable cluster of neurons (generally, 'nuclei' in the brain refer to anatomically distinct groups of neurons outside the cerebral and cerebellar cortex), is situated in the lower anterior part of the cerebral hemispheres within the ventral striatum. The VTA is a relatively small nucleus located in the midbrain, close to the substantia nigra, and is visible only under microscopic examination. Like the 'compact' part of the substantia nigra (the pars compacta), it is composed of dopamine-producing neurons that use dopamine as a neurotransmitter to communicate with other neurons. The axons of VTA neurons project to the nucleus accumbens, where they release dopamine at their terminals. Dopamine is often referred to as the 'happiness hormone', though this is only partially accurate. While dopamine influences feelings of pleasure, it also plays various roles in other brain functions, such as regulating prolactin secretion. Indeed, the stimulation of dopamine release by VTA neurons at the synapses they form with neurons in the nucleus accumbens is – unsurprisingly – an important functional element of the reward circuitry because the connection between the VTA and nucleus accumbens is an anatomical part of this circuit. In summary, the reward circuitry is considered the primary structure whose activation signifies a state of pleasure, though it might be better described as a 'desired' or appetitive state.

The fear circuitry can be viewed as the counterpoint to the reward circuitry as it primarily deals with 'aversive' stimuli. This circuitry includes not only the previously mentioned limbic structures but also the hippocampus and the locus coeruleus, the latter being a distinct, darkly pigmented, and easily identifiable small nucleus located in the brainstem near the floor of the fourth ventricle, and primarily using norepinephrine as a neurotransmitter.

In attempting to pinpoint the locations in the brain responsible for emotions, it becomes apparent that while we have a fairly detailed understanding of the specific structures and their connections involved in emotional processing, emotions are ultimately the product of the brain's overall activity. This activity, in turn, can influence our emotional state in real time through feedback mechanisms. For example, acting out or pretending to feel certain emotions can affect the actor's own emotional experience. Emotions simply cannot be neatly separated from other brain functions. According to LeDoux's concept, it is not feasible to

rely solely on either the 'high road' or 'low road' pathways in emotional processing. Paradoxically, even deliberate suppression of emotional expression can still be interpreted as conveying emotional information...

To conclude my 'Christmas' lecture on the interdisciplinary neuroscience of emotion, I circle back to the Christmas theme. The essence of its emotional power lies in empathy, though not merely interpersonal empathy; rather, it is a profound empathy between humans and the God-child. This creates a paradox: an omnipotent God portrayed as a vulnerable child in need of care and concern. While this topic warrants a separate discussion, if we consider the 'omnipotence' of emotions in terms of their undeniable impact on our lives, we find that emotions may not be omnipotent in the traditional sense (i.e., as shaping the individual and collective experiences of individuals and societies/nations), but they are certainly omnipresent in the brain. Research into the neuronal basis of emotional reactions reveals that their neural correlates are pervasive throughout the brain. Functionally, these emotional processes infuse the 'stream of consciousness' with a wide spectrum of colours and intensities – from the enchanting hues of a rainbow to the murky shades of confusion and loss of clarity.

References

Augustine. (2001). *The Confessions of Saint Augustine*. Trans. E.B. Pusey. Project Gutenberg. https://www.gutenberg.org/files/3296/3296-h/3296-h.htm. Accessed 28.08.2024.

Bear, M.F., Connors, B.W., & Paradiso, M.A. (2016). *Neuroscience. Exploring the Brain*. Philadelphia: Wolters Kluwer.

Bremer, J. (2005). *Jak to jest być świadomym. Analityczne teorie umysłu a problem neuronalnych podstaw świadomości*. Warszawa: Wydawnictwo IFiS PAN.

Cannon, W.B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, *39*, 106–124.

Ciccarelli, S., & White, J.N. (2021). Motivation and Emotion. In: Ciccarelli, S., & White, J.N. *Psychology*. Harlow: Pearson, 364–397.

Dalgleish, T. (2004). The emotional brain. *Nature Reviews Neuroscience*, *5*, 583–589.

Davis, K.L., & Panksepp, J. (2017). *The Emotional Foundations of Personality. A Neurobiological and Evolutionary Approach*. London: W.W. Norton & Company New York, London.

Gładziejewski, P. (2011). Czy empatia jest symulacja mentalną? Dyskusje z podejściem reprezentacjonistycznym w koncepcji neuronów lustrzanych. *Diametros*, *27*, 108–120.

Koelsch, S. (2013). *Brain and music*. Chichester: Wiley-Blackwell.

LaPointe, L.L. (2012). *Paul Broca and the Origins of Language in the Brain*. San Diego: Plural Publishing.

LeDoux, J. (1998). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster Paperbacks.

Papez, J.W. (1937). A proposed mechanism of emotion. *Archives of Neurology and Psychiatry*, *38*(4), 725–743. DOI: 10.1001/archneurpsyc.1937.02260220069003.

Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W.C., LaMantia, A.-S., McNamara, J.O., White, L.E. (ed.) (2008). *Neuroscience*. Sunderland: Sinauer Associates Inc.

Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience*, *27*, 169–192.

Mateusz Jarmużewski
Windesheim University of Applied Sciences
https://orcid.org/0000-0002-8398-1398
https://doi.org/10.35765/slowniki.482en

The neurobiological basis of moral responsibility

Summary

DEFINITION OF THE TERM: Responsibility for one's acts is a fundamental aspect of human beings as moral agents. In most ethical systems, the prerequisite for moral responsibility is the freedom to act, which is closely linked to the concept of free will. Both moral acts performed in the physical world and the notion of responsibility itself can be seen as based on natural foundations and patterns, primarily explained at the neurobiological level according to current knowledge.

HISTORICAL ANALYSIS OF THE TERM: Interest in the biological basis of behaviour dates back to antiquity, and awareness of its existence is reflected in Western anthropology and Christian ethics. However, neurobiological justification for (a lack of) moral responsibility is a development of the modern era. In this context, scepticism regarding concepts like free will, although not always leading to the rejection of responsibility, clearly calls into question the natural foundations of moral responsibility.

DISCUSSION OF THE TERM: New trends in neuroscience and systems biology offer a way out of the philosophical impasse in discussions on free will, particularly in the context of deterministic and stochastic explanations offered by cognitive neuroscience. These trends highlight the indirect, interdependent, and long-term nature of an organism's autonomy, which is essential for understanding both the subject and object of moral responsibility.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: Analyses of the neurobiological basis of human behaviour suggest a diffuse yet comprehensive spectrum of moral responsibility. This broader understanding encompasses the entire neurobiological dynamism of the person, embedded in traditional personalist anthropology, with particular reference to the Catholic theological and moral tradition and the socio-ethical context of the present day.

Keywords: neurobiology, moral responsibility, free will, virtue ethics, moral acts

Definition of the term

The issue of responsibility, particularly in a normative-ethical context, remains crucial not only for justifying moral valuations at a personal level but also as a fundamental aspect of human beings as moral agents. Responsibility for broadly defined action – or, more precisely, accountability, i.e., the possibility of attributing action and its consequences to a specific human person – is a basic assumption of most ethical systems. Responsibility for action in the context of its normative assessment in terms of 'good' or 'bad' is called moral responsibility.

The scope of action for which we can be responsible – or for which we can legitimately be held accountable – is extensive. While the primary subject of responsibility is the human being as the person, ethical reflections also consider collective or historical moral responsibility. The object of moral responsibility can encompass specific acts, behaviours, attitudes, intentions, or moral virtues. In this context, Catholic moral theology and personalist ethics focus on moral acts, which presuppose human action in the strict sense (*actus humanus*), as opposed to an action of a human being (*actus hominis*) that occurs in or through humans (Thomas Aquinas, 2017, S. Th. I–II 1.1.).

Although moral acts include conscious and voluntary thoughts or resolutions, they do not encompass the entire spectrum of human activities and attitudes subject to moral evaluation, judgement, and, consequently, moral responsibility. Broadly understood, moral action includes all aspects of human existence for which we can – and perhaps should – be held responsible.

Most ethical systems, along with many folk beliefs within Western culture, emphasise one fundamental condition for moral responsibility: the freedom of one's actions, resolutions, omissions, expressed attitudes, or ways of being. Only what a person *freely* does or pursues can be the object of moral responsibility (Prümmer, 1956, p. 8). While freedom as a basis for moral responsibility seems a fairly universal assumption, the answer to the question of what this freedom actually is varies depending on philosophical, theological, and cultural contexts and eras. The same applies to the question of the degree of freedom necessary (and sufficient) to justify both the moral prerogatives of action and the moral responsibility of the person.

Closely related to the condition of freedom is the question of the causes, sources, and nature – or, more generally speaking, the basis – of human action. Accepting the anthropological assumptions of continuity, or at least of a clear connection between humans and the material world, every human action, thought, or attitude can, in principle, be attributed to certain physical, chemical, biochemical or biological causes and underlying patterns. Among these diverse natural bases, we can distinguish between what directly causes or physically realises a given action or attitude and what (bio)physically enables it. Using terminology from the theology of creation, these can be referred to as first- and second-order causes. Naturally, many more levels of causation can be identified here, and the relationships between them are neither obvious nor easy to explain. This task becomes even more complex when direct and (above all) indirect causes of a supernatural nature are included in reflection on the basis of human action.

Extending this analogy to associate (or to some extent equate) the natural basis of human action with supernatural causes – or with c r e a t e d laws of nature – brings us into the realm of the classic theological dispute about human free will in the context of God's omniscience and providence. Without diminishing the distinctive nature of theological explanations or indiscriminately adopting pan(en)theistic views, we can avoid shallow interventionism and thereby place the theological debate about divine and human freedom within the context of all the essential bases for human action. Crucially, the search for direct causes and indirect patterns that govern human moral action – or, more broadly, those aspects of human activity that are subject to moral judgement – involves understanding both what enables and causes such action as well as what justifies attributing responsibility to the acting subject.

When considering whether an act should be deemed free, we ask several key questions: (a) Who or what is behind a performed act or expressed attitude? (b) Could the person have acted differently? (c) To what extent does what the person does or represents reflect their capacity to respond to their own rationality (reasons-responsiveness)? In other words, does a person act in accordance with their (more or less legitimate) reasons and are they able to recognise, understand, and implement these reasons freely? Colloquially speaking, does a person genuinely 'do what they want' or what they (happen to) think is right at the

moment? Questions (a) and (b) address two fundamental philosophical conditions for the existence of free will, both of which are extensively scrutinised by the empirical sciences. Although question (c) – about the 'capacity to respond to one's reasons – may seem less complex, it also involves inquiries into the natural basis of rationality.

The fact that morally acting human beings primarily function in the physical world as living organisms justifies focusing on biological explanations. The location of the centres of human thinking and acting in the central nervous system, and especially in the brain itself, explains the central position of broadly understood neuroscience. In Polish-language contexts, the terms 'neurobiology' and 'neurobiological' sometimes cover the entire field of modern neuroscience. Originally, however, from a methodological perspective, neuroscience is a biological field that extends beyond the multifaceted study of brain function. Unlike a purely neuroscientific approach, which includes fields such as cognitive neuroscience or neuropsychology, a neurobiological approach emphasises the functioning of the entire organism and the biologically determined role of the brain and nervous system in this process.

The neurobiological basis of moral responsibility encompasses both strictly neural processes and those from the broader biological context of human volitional actions. Related disciplines, such as evolutionary biology, systems biology, ethology, and ecology, are invaluable for studying the processes in the latter group. As in many branches of cognitive neuroscience, understanding the functioning of the nervous system and its interaction with the entire organism (or with a subject of moral action in the case of humans) requires insight into physical, biophysical, and biochemical processes. It is worth noting that, beyond basic knowledge of brain biophysics and chemistry, attempts to explain the volitional aspects of organisms' functioning at the level of, e.g., quantum physics often remain speculative and hypothetical.

Philosophical considerations concerning determinism, free will, and the epistemological status of neuroscience significantly impact the interpretation of neurobiological data on the origins of human action. The issue of free will – whether approached from metaphysical, analytical, theological, or neurophilosophical (philosophy of neuroscience) perspectives – is crucial for properly situating the concept of moral responsibility and outlining its natural justifications. However, the potential thesis of

the non-existence of free will does not necessarily lead to the rejection of moral responsibility (Fisher, 2007).

Similarly, the strictly philosophical mind-brain/mind-body problem, the related problem of mental causation, along with considerations about the existence and nature of self-consciousness and purely psychological descriptions of agency and freedom, often diverge from neurobiological explanations. Furthermore, moral responsibility is also clearly distinct from legal, civil, or professional responsibility.

Analyses of the neurobiological basis of responsibility, as understood in this context, focus on explanations within the natural sciences, often treating other explanations (such as psychological, phenomenological, or theological) in a perfunctory or contextual manner. However, since moral responsibility itself is primarily an ethical and social concept – and thus transcends the methodological framework of the natural sciences – the results of such analyses inherently fall within the realm of interdisciplinary dialogue, as well as the philosophy and theology of science.

Historical analysis of the term

Although systematic knowledge of the nervous system in humans and animals began to develop between the 17th and 19th centuries, the biological basis of human behaviour has intrigued people since antiquity. Exactly the same can be said about cognitive neuroscience and the classical philosophical study of cognition. The widespread understanding that links thinking, emotions, and decision-making to the brain is a product of modern science. However, the relationships between these processes and the functioning of the human body also fascinated early ethicists, metaphysicians, and epistemologists.

Researchers today cannot ignore the historical awareness of biology's influence on human choices, both in terms of *what* we choose and *how* we choose it. Efforts to neurobiologically justify responsibility (or the lack thereof) are linked to the modern development of natural and empirical sciences and their increasingly radical divergence from the humanities and theology. Historically, systematic reflections on human freedom and responsibility have followed a fundamentally different path from scientific attempts to explain human action and its governing

patterns. This divergence persisted until the Enlightenment, after which a gradual shift in this area coincided with the rise of modern neuropsychology, brain science, and the analytical tradition in philosophy.

In antiquity (and not only in the West), cognitive processes were predominantly located in the human heart, most likely due to its clear connection to blood circulation. The heart was commonly regarded as the seat of the rational soul, while emotions and sensory desires were sometimes attributed to other organs, such as the liver, according to Aristotle. One of the few exceptions was Hippocrates (460–377 BC), whose views on the brain's role were not significantly different from contemporary understandings. In Alexandria in the 3rd century BC, Herophilus was the first to differentiate between the cerebellum and the cerebrum, suggesting that the cerebellum, due to its greater density, controls the muscles. Erasistratus accurately located logical thinking in the cerebral cortex, speculating that human intelligence was related to the number of visible brain sulci (Gross, 1987). Later, Galen (130–200 AD) suggested that the brain and spinal cord as its extension were the site of sensation, thought, and behaviour control.

This state of knowledge, including descriptions of neuropsychiatric disorders arising in the brain by Avicenna, persisted until the Renaissance. René Descartes (1596–1650), considered a pioneer in brain and mind research, is best known for his theory of dualism, which, among other things, proposed that the pineal gland is where the mind interacts with the body (after accounting for the mechanisms that direct the circulation of cerebrospinal fluid). However, Descartes' contribution to the neurobiology of action is broader and encompasses his theory of reflex action, which explained reflexive behaviours as reactions to external stimuli that move the skin and the filaments, releasing the 'animal spirit' and innervating the muscles (Elias & Saucier, 2006, p. 6). This mechanistic explanation paved the way for studying the neural basis of animal and, by extension, human behaviour.

The next major advances in neuroscience came with the discovery of the specific functions of the nervous system and their increasingly precise anatomical correlations. The work of Bell (d. 1842) and Magendie (d. 1840), which demonstrated the distinct functions of the spinal nerves, is considered the beginning of this era. While it is challenging to enumerate all the pivotal moments in this rapidly evolving

field, several key developments stand out: Golgi's (d. 1926) microscopic description of nerve cells; Ramón y Cajal's (d. 1934) description of their intricate network connections; Charles Scott Sherrington's (d. 1952) description of the anatomy and functioning of synaptic connections; Loewi's (d. 1969) description of the role of neurotransmitters; Bernstein's (d. 1917) Membrane Theory of action potentials.

The second half of the 20th century brought significant progress with the development of advanced neuroimaging techniques, such as magnetic resonance imaging (MRI), magnetoencephalography (MEG), and positron emission tomography (PET). Additionally, improvements in older techniques like electroencephalography (EEG) and transcranial magnetic stimulation (TMS) also advanced the field. Functional MRI (fMRI) allows researchers to observe brain activity across different regions, though it has limited temporal resolution, while EEG provides real-time measurements of electrical potentials in the brain but has limited spatial resolution.

These promising developments have motivated researchers like Benjamin Libet (d. 2007) to study the neural basis of human action and the universal experience of free will. Since the mid-1980s, hundreds of experiments – primarily employing EEG and fMRI – have sought to uncover the brain processes involved in human will and intentional agency, and the neuroscientific and philosophical interpretations of these studies have generated considerable interest. The findings of Libet's pioneering work suggest that the brain makes a decision to act unconsciously before the acting subject becomes aware of it. According to Libet, while the acting subject can consciously interrupt an already initiated action generation, this ability is limited to a brief window of time (Libet 1985). The results obtained by Libet and other similar findings have led many scholars, such as American neuropsychologist Daniel Wegner, philosopher Derk Pereboom, Belgian philosopher Jan Verplaatse, and Polish neuroscientist Włodzisław Duch, to argue that free will is merely an illusion.

Although dismissing the idea of free will profoundly affects the understanding of moral responsibility, it does not necessarily eliminate it (Fischer, 2007). Some theorists separate the two concepts to an extent that it is possible to talk of moral responsibility even if the idea of free will is completed rejected (Duch, 2010). Additionally, the controversial

notion that free will might exist without implying moral responsibility is also considered by some (Waller, 1998).

Despite the implications of his findings, Libet himself does not exclude "philosophically real individual responsibility and free will" (Libet, 1985, p. 538). Problems often observed in this context – apart from insufficient evidence for conscious veto – include Libet's assumptions about the nature of causality, what conscious and free choice would actually consist in, and whether the acts studied even resemble those for which we would be morally responsible. Nonetheless, the empirical data from Libet's experiments appear to challenge the intrinsic sense of causal agency (Bremer, 2013, p. 65) that underpins modern folk, legal, and philosophical concepts of moral responsibility.

In Christian ethics, the notion of conscious and free choice has traditionally been taken for granted. Discussions about the natural basis of action – where an attitude or act might allegedly lose (partly or completely) its volitional and free character – often focused on emotions and mental disorders, whether real or perceived (Prümmer, 1956, pp. 14–19). Interestingly, discussions on affectivity did not only concern negative phenomena. To some extent, the Thomistic tradition recognises the constitutive role of affect and emotion in moral action (Stump, 2011). However, the impact of modern rationalism has largely marginalised the non-rational elements of human action in discussions of moral responsibility.

At the same time, in this philosophical and theological tradition we find references to Aristotle's theory of the virtues, which, with later modifications, has become permanently embedded in Western ethical and moral-theological terminology. This continuity remains despite the fact that, until recently, the impact of this part of the Aristotelian-Thomistic legacy in ethics was mainly confined to moral rhetoric and to symbolic representations of moral ideals. Originally, the concept of virtue (*arête*) was grounded in the idea of embodied dispositions that significantly and consistently influence human actions. Although the role of reason and rationality (understood in antiquity quite differently than today) in shaping and realising virtue remains central, the subject of these enduring moral dispositions to act (moral virtues) is the human being as a holistic being comprised of vegetative, sensitive, and rational dimensions (Aristotle, 1934).

The idea that the source of human action – which can be subject to moral responsibility – resides within the acting subject is a cornerstone of Christian and Western anthropology. In this classical perspective, will reflects the intrinsic principle of motion, which is something humans share with other animals. This assumption, sometimes referred to as the condition of endogeneity, along with the subject's knowledge of purpose (end, goal) and the possibility of free choice, constitutes the basis for the *actus voluntaris*, which can be the object of individual moral responsibility (Thomas Aquinas, 2017, S. Th. I–II 6. 1–2.). The primary distinction between humans and animals lies in the means by which they (re)cognise the goal of their actions. While animals possess a sensory form of cognition, humans – thanks to their capacity for rational cognition – surpass it in both degree and quality. Nonetheless, the human will remains a mental form of desire, a mode of activity in the physical world that is typical of the human species.

The perception of will as an inner principle of motion has roots in ancient Greek thought. Apart from Aristotle, this idea is present in Stoic philosophy, which influenced later thinkers such as St. Augustine. Some interpretations suggest that Augustine's original use of the term *voluntas* refers to the Stoic concept of *hormê*, meaning a positive impulse or an inclination towards action. Voluntary causes (*hormê* – the inner principle of motion) are attributed to God, angels, humans, and, in an analogous sense, animals (Byers, 2006, p. 174). Augustine's later works emphasise the role of will as the ultimate arbiter, responsible for giving consent either to God or to the body (the world). In both cases, and irrespective of weakening natural inclinations associated with original sin, the acting subject remains morally responsible more in spite of than because of the natural basis of action.

Many centuries later, Descartes explicitly equated will with the mind – *res cogitans* – which he saw as capable of intervening in the mechanistic and deterministic realm of *res extensa*. For Descartes, the will, synonymous with free choice, embodies our capacity to do or not do something. He considered freedom of the will self-evident and equated exercising it to acting voluntarily. According to Descartes, the ability to do *a or b* – a function of the totally free and immaterial mind – was foundational to moral responsibility and human "supreme perfection" (Descartes, 1984, p. 37).

Many later conceptualisations of will, including those in ethical and theological contexts, were increasingly shaped by these early modern ideas. Thinkers such as Spinoza, Hobbes, and Hume placed human (free) will within the broader framework of nature's universal causal relationships. This perspective laid the groundwork for modern empirical naturalism, which soon extended to encompass human will, the mind, and agency. Despite this, it is the libertarian dualism of Descartes, further complemented by Kantian notions of freedom and moral responsibility, that has most significantly influenced contemporary understandings of moral action (Jarmużewski, 2021). Additionally, the political and economic concepts of individual freedom, coupled with enormous success in science and technology, fostered a growing belief in human self-determination, the power of reason, and the reality of free choice.

This impact was not limited to the rationalist tendencies of the (post)Enlightenment. Also in Christian personalism, the emphasis has often been placed on a specific moral act, the moment of choice, and its transcendent and free character. Linking all these elements to the experience of self-consciousness was not coincidental. It represents a synthesis of Thomistic thought and elements of Augustine's theology with a modern approach to freedom and responsibility. In the Polish context, it is important to acknowledge the contributions of the Lublin School, notably Karol Wojtyła and Mieczysław Krąpiec, to the discourse on responsibility, especially regarding its natural and transcendent foundations.

Drawing on both scholastic tradition and personalism, Wojtyła clearly distinguished between what the person does in performing an act (*actus humanus*) and what is simply happening to (in) the person (*actus hominis*). While both phenomena contribute to the broadly understood dynamics of the person, only the former qualifies as an act for which we are morally responsible in the proper sense (*actus voluntaris*). Wojtyła argued that every moral act entails the experience of transcendence (Wojtyła, 2021, pp. 121–122, 127, 161–162). However, as an intrinsic element of a moral act, this experience does not diminish the underlying dynamism of the acting person. Assuming the unity of soul and body in the human person, it must be recognised that this dynamism is largely composed of neurobiological elements.

Discussion of the term

Although the question of moral responsibility in the context of the natural basis of action has been debated since antiquity, the modern era has introduced new challenges that have profoundly altered both the perspective and tone of this discussion. The challenges are linked to both the aforementioned dualism and rationalism of early modern thought and the development of empirical sciences, particularly empirical psychology in the 18th and 19th centuries, which radically placed the issue of responsibility in the context of the biophysical constitution of human beings.

Many of the 19th-century 'fathers of psychology' had extensive knowledge of physiology, while some, such as William James, even argued that the scientific study of psychology should be grounded in an understanding of biological systems. In contrast, Freudian psychoanalysis saw the source of human action primarily in the subconscious, while behaviourism emphasised stimulus-response relationships, which profoundly altered the approach to free will and moral responsibility. The contributions of early psychologists, like William Carpenter and William James, thus marked a significant shift in thinking about moral responsibility.

While the perspective on the justification of individual freedom and responsibility still remains relatively conservative, it has laid the groundwork for an empirical justification of long-standing scepticism regarding freedom of action within classical philosophy and theology. The psychological paradigm has undoubtedly also contributed to the emergence of 'illusionist' views on free will and moral responsibility that are typical of the second half of the 20th century and inspired by several decades of research in neuroscience, cognitive neuropsychology, and other related fields.

Benjamin Libet's approach has faced multifaceted criticism. First, the readiness potential measured – or rather calculated – through EEG recordings may not account for all potential causes of an action. Second, even though a decision regarding simple motor tasks like finger movements in no way resembles decisions for which we might be morally responsible, it is still a neurologically complex and complicated process.

Third, the temporal sequence between brain activity and subjective (introspection-based) awareness of intention does not necessarily establish a causal relationship (Bremer, 2013, pp. 60–61, 65).

Nevertheless, some neurophilosophers, referring to the non-existence of free will, which they claim has been proven, propose rejecting the concept of moral responsibility as well. This stance is often associated with hard determinism, which contrasts with various forms of libertarianism that defend modern notions of free will and moral responsibility. Both hard determinism and libertarianism share the belief that free will is fundamentally incompatible with determinism (incompatibilism). Semicompatibilism does not dismiss the notion of responsibility, but even compatibilist positions often question the universality of freedom and autonomy, thereby narrowing the scope of individual moral responsibility. The standard argument against free will posits that both determinism and stochasticism (the randomness of events) are problematic. However, alternative paradigms within neuroscience, as well as advances in modern evolutionary and systems biology, offer potential pathways out of this philosophical impasse.

In recent decades, many research programmes have focused on exploring the broader foundations of human action. Despite their great diversity, all these new approaches shift the focus from atomised neuronal signals (which supposedly correspond to intentional mental states) in the context of individual – and typically uncomplicated – motor acts, to the complexity and multiplicity of causes underlying intentional actions. A leading researcher in this field, Patrick Haggard aptly observes that the freedom of volitional acts is associated with a range and complexity of information, which triggers action when integrated (Haggard, 2019, p. 12).

Haggard suggests that the neuronal initiation of action should be viewed as a gradual transition from an initial state of fluidity, where various environmental and internal stimuli generate competing effects, to a state of certainty just before the action occurs. The primary motor cortex serves as the convergence point for all these initial stages, which unfold at different times and locations across the motor cortex (including those studied by Libet), the parietal areas, and notably the prefrontal cortex which is involved in memory and conscious intention (Haggard, 2019).

Recent neuroscientific paradigms also emphasise the long-term aspect of free will. For example, about a decade ago, Peter Tse developed the theory of criterial neural causation. Tse observed that the physical conditions necessary for the formation of action potentials at synapses, which constitute the criteria for firing, are subject to continuous resetting. He proposed that conscious thought processes, reflecting the brain's current state, can influence future changes in these criteria. This leads to the concept of a "criterial information code" that allows individuals to exert an indirect influence on their future actions. According to Tse, actions resulting from synaptic stimuli meeting certain neuronal criteria established in the near or distant past are neither entirely random nor fully determined (Tse, 2013). Since conscious impact on these criteria can take place on the scale of milliseconds as well as years, Tse's theory supports moral responsibility for both moral dispositions (virtues) and for specific moral acts.

Another notable approach that addresses the long-term nature of free action is the research led by Stefan Bode. The findings of his team suggest that voluntary decisions to act are akin to perceptual decisions. Both types of decisions are nothing more than the end result of a process called "evidence accumulation", which eventually exceeds a critical threshold. In the absence of sufficient evidence or cues at the time of action, this process proceeds on the basis of a history of previous choices, preferences, and accumulated knowledge. According to Bode, internally generated brain activity can partially be stochastic and partially reflect past choices (Bode et al., 2012). At the same time, due to the multiplicity of elements that make up the accumulation process, the conscious impact on decisions made in the brain also remains partial. Thus, a subject's freedom and responsibility for the resulting action hinge less on absolute freedom from past decisions and more on the gradual process of moral development. Like many other theories (for example, those related to the notion of neuroplasticity), the long-term nature of conscious mental causation may entail the individual's responsibility for the habituation process, if only through the consistent repetition of proper acts (Aristotle, 1934, 1103b).

The interpretation of the brain's generation of intentional actions as "evidence accumulation" highlights another distinctive feature of contemporary neuroscience: a shift from a *computational neuroscience*

model to a *dynamical neuroscience* model, the latter using mathematical formalisms (*dynamical systems)* to analyse global brain activity (e.g., based on EEG recordings) over time. This model emphasises variables such as rate, duration, and dynamic flow of change, while typically rejecting the notion of neural representations of mental states and their spatial localisation (Shapiro, 2013, pp. 358–359). The stochastic elements identified in the dynamics analysed in this way could function as a vehicle for undetermined action alternatives (which is one of the philosophical conditions of free will), while the elements of order (control parameters) could enable volitional control, accountability for an act, and the moral responsibility of the subject.

Aaron Schurger and colleagues argue that intentional actions do not originate from a specific location or process but rather "emerge from a causal web in the brain" (Schurger et al., 2015, p. 761). According to the Stochastic Decision Model (SDM), what we call spontaneous volitional acts result from the integration of random fluctuations and external stimuli, reaching a threshold that triggers a motor response. The readiness potential (RP) measured by Libet does not reflect the processes that lead to a decision but represents general brain activity triggered by various factors, which corresponds to the subject's increasing propensity to act rather than to the brain's ultimate determination (Schurger et al., 2015). Stochastic models of decision-making additionally suggest that consciousness is not essential for controlling action or making decisions. Personal control of action is equated with the interaction of multiple processes that, in their totality, can be considered the neurobiological basis of responsibility.

In this framework, freedom of choice and action is based on the natural autonomy that is inherent in all living organisms, albeit to varying degrees and in different ways. The theory of natural autonomy, also known as biological autonomy, describes these fundamental characteristics of organisms as relatively self-contained systems of autopoiesis or causal closure (Moreno & Mossio, 2015). The concept of 'natural autonomy', i.e., the autonomy that is developed through evolution, explains – within the limits of classical biophysics (without tentative references to the alleged indeterminism of quantum physics) – sufficient freedom (available alternatives of action) and purposefulness (accountability) of actions at a level appropriate to specific animal species. Although it is

difficult to speak of any moral responsibility here, it is also difficult to imagine the accountability of moral acts in humans without this form of evolutionary inheritance.

The concept of organisms' natural autonomy encompasses successive stages of organisation and regulation to promote growth, the maintenance and self-imposition of constraints on one's own boundary conditions, and, ultimately, relatively autonomous causation. The gradual development of behavioural flexibility and intentionality at both individual and species levels can be seen as a precursor to uniquely human action, rooted in the broader autonomy of the human organism. Rationally enhanced and justified autonomy is based on an evolutionarily older "freedom from" direct physical causes and linear cause-effect relationships (Moreno & Mossio, 2015).

Human cognitive capacities complement those present at the somatic and behavioural levels, allowing acting humans to more effectively evaluate available alternatives and to increase the subjectivity and accountability of actions through indirect or long-term mechanisms of conscious control. This integration of a natural propensity for exploration with the ability to adapt and stabilise supports the argument for a naturally developed sense of responsibility (Waller, 1998). For the concept of moral responsibility, this suggests a deeper integration of somatic, affective, and ecological aspects of agency, which is an element of a reemerging (Christian) virtue ethics (Stump, 2011).

Systematic reflection with conclusions and recommendations

Supported by neurobiological research, the complexity of the cause-and-effect relationships that underlie human action and the indirect, relative, and largely long-term understanding of causation and free will suggest a broader perspective on moral responsibility. Since the neurobiological basis of moral agency extends beyond the potential role of self-consciousness in making specific decisions and resists simplistic explanations at the biophysical or biochemical levels, the same may be true of the natural basis for attributing personal responsibility to such action.

Conclusions about the neurobiological basis of moral responsibility are often applied in legal and criminal contexts to address the complexities of criminal responsibility, particularly in atypical cases. At the same time, the notion of legal responsibility tends to be very practical and binary: an individual is, in principle, either accountable or not. Social, psychological, and neurobiological mitigating circumstances are often regarded in this context as deviations from the normal state of the person, not as irregularities inherent in the human constitution. A similar approach is evident in discussions of moral responsibility.

Despite the different meaning of these terms, a person's moral responsibility is often equated with their accountability for specific acts. This clearly includes any conscious moral action (or at least its final observable outcomes). Importantly, however, such accountability for particular acts, when considered in light of their neurobiological causes, has the aforementioned characteristics of natural causality: indirectness, long-term nature, modality, and ecological (social) determinants.

The long-term nature of moral responsibility also impacts the proximal aspect of specific actions. Even if an action is largely determined at the moment of choice, the experience of deliberate decision-making (perceived as if it were fully free) significantly shapes the neurobiological and psychological basis of future actions (Tse, 2013). This understanding of responsibility implies responsibility for shaping intentions and plans, setting goals, integrating and valuing specific neuropsychological processes (such as rational thinking and directing attention), as well as developing virtuous character traits and acting on them (Jarmużewski, 2021).

In this context, the objects of responsibility can include both specific neural pathways and global changes in brain activity, the latter viewed as a dynamical system (Shapiro, 2013). When considering the latter, the exact nature of personal will remains elusive, particularly when neurocognitive explanations are based on random fluctuations in neural activity (Schurger et al., 2015). However, the subject's responsibility for these generated actions can be supported by even minimal evidence of non-accidental cues in the fluctuating dynamics. These cues might represent processes of moral reflection or a personal history of choices and virtue formation, for which we could also be responsible.

Aristotle's concept of virtue provides a useful framework here. According to Aristotle, virtue is a goal-directed disposition or a 'choice-making'

disposition. While virtuous actions may exhibit a trained automaticity (*hexis*), they remain conscious and deliberate choices, driven by an active use of 'rational desire' (*orexis dianoetike*) and 'desiring reason' (*orektikos nous*) (Zagzebski, 1998, p. 217). The capacity to desire the good – as is central to the human volitional sphere and is grounded in practical wisdom (*phronēsis*) – must lead to actual choices of appropriate actions. Every choice and every action start from existing desires or goals (telos), which, through rational deliberation, can shape both the means to achieve them and future desires (ends) of an acting person. Aristotle argued that an absolutely 'free' choice is not possible, and human moral action always results from specific desires and rational deliberation. Nonetheless, this framework supports a robust normative responsibility for learnt and cultivated virtues (moral and intellectual), correct thinking, and the pursuit of goals consistent with human nature.

The natural autonomy of living organisms, including humans, supports this understanding of responsibility. On the normative side, the unique position of the human species in the world of living beings raises the question of responsibility for how we function in that world. In the context of the contemporary ecological crisis, human responsibility for mitigating climate change is usually framed as concern for nature as something external to humans and their capacity to act in the world. Such an approach is often not only ineffective but misaligned with a fundamental characteristic of human action, namely its embodied, ecologically shaped, and interdependent nature.

On this basis, it is possible to argue for a diffuse yet broader and ethically viable notion of moral responsibility by pointing to the myriad causes behind human moral actions. Identifying the exact place of a subject's personal and voluntary agency within the complex network of neurobiological causes and patterns is indeed challenging. Given the current state of knowledge and the multifaceted nature of individual moral acts, free will and intentional actions must be understood as entangled with neurobiological, ecological, and social factors. Nevertheless, the evolutionarily developed autonomy of living beings, combined with uniquely human cognitive capacities, indicates that we can exert a real impact on our present and (in particular) future actions to an extent that justifies moral responsibility.

References

Aristotle (1934). *Nicomachean Ethics*. In: *Aristotle in 23 Volumes*, vol. 19. Trans. H. Rackham. Cambridge–London: Harvard University Press; William Heinemann Ltd. http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0054%3Abekker%20page%3D1107a. Accessed 15.08.2024.

Bode, S. et al. (2012). Predicting Perceptual Decision Biases from Early Brain Activity. *Journal of Neuroscience*, *32*(36), 12488–12498.

Bremer, J. (2013). *Czy wolna wola jest wolna? Kompatybilizm na tle badań inderdyscyplinarnych*. Kraków: Wydawnictwo WAM.

Descartes, R. (1984). *Principles of Philosophy*. Trans. V.R. Miller. Dordrecht: Reidel.

Duch, W. (2010). *Czy jesteśmy automatami? Mózgi, wolna wola i odpowiedzialność*. In: P. Francuz (ed.), *Na ścieżkach neuronauk*. Lublin: Wydawnictwo KUL, 219–264.

Elias, L.J., & Saucier, D.M. (2006). *Neuropsychology: Clinical and Experimental Foundations*. Boston: Pearson/Allyn & Bacon.

Fisher, J.M. (2007). *Compatibilism*. In: J. Fisher, R. Kane, D. Pereboom, & M. Vargas (ed.), *Four Views on Free Will*. Oxford: Blackwell, 44–84.

Haggard, P. (2019). The Neurocognitive Bases of Human Volition. *Annual Review of Psychology*, *70*(1), 9–28.

Jarmużewski, M. (2021). *Neurobiology and the Free Will: Implications for Catholic Theological Ethics*. Doctoral Dissertation. KU Leuven.

Libet, B. (1985). Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *Behavioral and Brain Sciences*, *8*(4), 529–539.

Moreno, A., & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical* Enquiry. Dordrecht: Springer.

Prümmer, D.M. (1956). *Handbook of Moral Theology*. J.G. Nolan (ed.). Trans. Gerald W. Shelton. Cork: The Mercier Press.

Shapiro, L. (2013). Dynamics and Cognition, *Minds and Machines*, *23*(3), 353–375.

Shurger, A., & Uithol, S. (2015). Nowhere and Everywhere: The Causal Origin of Voluntary Action. *Review of Philosophy and Psychology*, *6*(4), 761–778.

Stump, E. (2011). The Non-Aristotelian Character of Aquinas's Ethics: Aquinas on the Passions. *Faith and Philosophy*, *28*(1), 29–43.

Thomas Aquinas *(2017) The Summa Theologiæ.* Trans. Fathers of the English Dominican Province. K. Knight (ed.). http://www.newadvent.org/summa/. Accessed 16.08.2024.

Tse, P.U. (2013). *The Neural Basis of Free Will: Criterial Causation*. Cambridge, MA – London: MIT Press.
Waller, B.N. (1998). *The Natural Selection of Autonomy*. Albany: State University of New York Press.
Wojtyła, K. (2021). *Person and Act and Related Essays*. Trans. G. Ignatik. Washington, D.C.: The Catholic University of America Press
Zagzebski, L.T. (1998), *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press.

Aleksandra Kuzior
Silesian University of Technology
https://orcid.org/0000-0001-9764-5320
https://doi.org/10.35765/slowniki.480en

Cognitive technologies

Summary

DEFINITION OF THE TERM: Cognitive technologies are a group of technologies based on artificial intelligence algorithms, machine learning, deep learning, and artificial neural networks. They are equipped with a variety of functions modelled on human thought processes and are thus capable of learning and performing tasks that traditionally require human intelligence. They imitate and sometimes even surpass human perceptual, cognitive, and decision-making capacities.

HISTORICAL ANALYSIS OF THE TERM: Interest and research in artificial intelligence and cognitive technologies date back to the 1950s and 1960s, when specialists began exploring these fields. Early examples include work on servomanipulators, programming languages, and heuristic programming.

DISCUSSION OF THE TERM: The term 'cognitive technologies' refers to an interdisciplinary fusion of psychological research and technological innovations that significantly enhance our cognitive processes. These technologies emulate the functions of the human brain through tools such as natural language processing, data mining, and pattern recognition. Key areas of focus in cognitive technology research include the development of automated and intelligent environments, as well as the integration of technology with the human body. This article will primarily concentrate on the former.

SYSTEMATIC REFLECTION WITH CONCLUSIONS AND RECOMMENDATIONS: The current trajectory of cognitive technology

development is aimed at creating an artificial mind for humanoid robots and intelligent environments. The artificial mind is intended to imitate the human mind, which requires the implementation of mental processes related to thinking and consciousness. However, this pursuit should be accompanied by systematic ethical reflection and the delineation of boundaries for the implementation of cognitive technologies in order to ensure they contribute to a better, more humane society.

Keywords: cognitive technologies, artificial intelligence, generative artificial intelligence, cognitive science, artificial neural networks

Definition of the term

The terms ‘artificial intelligence’ (AI) and ‘cognitive technologies’ often overlap in scope, but in this article we emphasise that AI focuses on developing effective algorithms that generate the best overall solutions to problems, while cognitive technologies emphasise making the best decisions based on specific circumstance. Cognitive technologies prioritise providing the information that is required to make these decisions rather than making them directly. They are subsets of technologies based on AI algorithms, machine learning, and deep learning. Equipped with various functions modelled on human thought processes, these technologies can perform tasks traditionally requiring human intelligence. They imitate – and sometimes even surpass – human perceptual, cognitive, and decision-making capacities.

Cognitive (or learning) technologies are often referred to as technologies that mimic certain functions of the human brain using tools such as natural language processing, data mining, and pattern recognition. They are considered a subset of the broader artificial intelligence field, which is often seen as part of biomimetics (Rouse, 2018). In broader definitions, cognitive technologies encompass not only artificial intelligence but also advanced analytics, high-performing computing, and cyber-physical systems (Elia & Margherita, 2021).

Cognitive technologies are grounded in cognitive science and are often further supported by insights from biology, neuroscience, biotechnology, and nanotechnology. They leverage bioinformatics and nanoinformatics databases, artificial intelligence, and modern communication tools to enhance and extend human cognitive and decision-making capabilities. While cognitive technologies have yet to replicate the full complexity of human intelligence, they are advanced enough to significantly improve quality of life (Bainbridge, 2006). At the same time, human creativity continues to push the boundaries of technology, potentially leading to innovations that could eventually compete with human intelligence (Lakomy, 2023), raising legitimate concerns about the possible gradual degradation of human cognitive abilities (Kuzior, 2017).

Cognitive technologies have been recognised as megatrends, indicating their rapid spread across the global business and social environment. They are now being applied in numerous areas of the

economy and daily life, including medicine and healthcare, banking, finance, commerce, marketing, media, education, journalism, industry, and the military. Their widespread adoption can provide companies with a significant competitive edge (Lakomy, 2023). The market for cognitive technologies is already estimated to be worth billions of dollars. As the labour market evolves, driven by the development of these technologies, there will be a growing need for education that combines technical and IT expertise with social sciences and humanities, addressing the ethical considerations that cognitive technologies present.

Historical analysis of the term

Ideas for building thinking machines date back to antiquity, when Greek mythologists envisioned such devices, like the carved ivory statue of Galatea that came to life in Ovid's tale of Pygmalion. In the 18th century, a mystifying contraption known as the 'Mechanical Turk' was presented as an ingenious chess-playing machine, though it was later revealed that a human chess player was secretly operating it from within. In 1997, IBM's Deep Blue computer defeated Garry Kasparov, the reigning world chess champion. This victory marked a pivotal moment in the field of artificial intelligence and cognitive technologies, sparking an explosion of research and development. Other key milestones were reached in 2005, when a robot built at Stanford university won the DARPA Grand Challenge race, and in 2011, when IBM's Watson, then a symbol of cognitive technology, triumphed over two grand champions on the quiz show *Jeopardy!* (without having access to the internet). Cognitive technology has also played a crucial role in the success of streaming services like Netflix, significantly contributing to its remarkable market dominance (Rouse, 2018).

Cognitive technologies are often closely associated with artificial intelligence, a topic that has garnered extensive attention in both popular media and academic research. However, fewer scholarly studies focus directly on cognitive technologies. Both AI and cognitive technologies are commonly linked to the futuristic concepts depicted in science fiction films. Many people today may not realise that they are already using solutions based on AI and cognitive technologies, often unconsciously.

In the coming years, cognitive technologies are expected to play a crucial role in shaping human–technology interactions, particularly in areas such as automation and robotics, machine learning, and information technology (Kuzior & Kwilinski, 2022).

The origins of research into AI and cognitive technologies can be traced back to the work of pioneers in the 1950s and 1960s, such as John McCarthy, Marvin Minsky, and Alan Turing. For instance, mid-20th-century studies explored the use of servomanipulators to explore the possibilities of direct interactions between digital computers and the physical world. These devices were programmed to autonomously perform functions typically attributed to humans, such as perceiving the environment and deciding on methods for accomplishing tasks (Ernst, 1962). Subsequent research focused on developing programming languages tailored to the needs of AI research and simulating cognitive processes (Newell, 1963). This era also saw the emergence of heuristic programming for solving complex problems (Minsky, 1961) and the design of collaborative human-machine cognitive systems (Woods, 1985).

Heuristic programming techniques have advanced decision-making automation and created environments where humans (natural cognitive systems) interact with artificial cognitive systems. These advancements have introduced new challenges, including the potential for creating hybrid cognitive systems that integrate both human and artificial cognitive components. Donald Norman, an electrical engineer and cognitive psychologist known for his work in user experience and usability, is also a recognised expert in the interdisciplinary design of such hybrid cognitive systems and co-founder and director of the User Experience/Usability consultancy.

The pursuit of understanding and enhancing the human mind has led researchers to develop highly personalised computer systems that can function as private counsellors, psychotherapists, or spiritual guides. One such example is ANNE (Analogies in Natural Emotion), a personalised computer database system that operates using the principle of analogy. ANNE mirrors the emotional structure of the user's mind, which is deeply rooted in cultural context. Rather than dictating actions or authoritatively telling the user what to do, ANNE engages in a dialogue with the user, simulating an intimate understanding of their emotions

and helping them make their own decisions in specific situations. This system leverages psychological, sociological, and computational methods (Bainbridge, 2006). Additionally, advances in nanotechnology have led to the development of microprocessors that can be inserted into living organisms, including humans. These microprocessors can monitor internal biological states and communicate with external devices.

Cognitive technologies offer innovative approaches to conceptualising (i.e., precisely defining) research processes and facilitating real-world communication (Bainbridge, 2006). What do these new communication methods entail? A key example is the development of artificial intelligence systems that use natural language processing (NLP) technologies. These systems aim to enable machines to understand and respond to human language in a more natural and intuitive way. By analysing the structure of language and how humans use it to communicate, NLP systems can be designed to deliver more accurate and relevant responses to user input.

It is important to note that cognitive technologies have significantly evolved with the advent of the internet, particularly through the development of the web and cloud computing. The IBM's aforementioned Watson is a prominent example of these advances.

Cognitive technologies leverage neural networks to mimic the processes of the human brain. Today, brain mapping serves not only as a tool for diagnosing diseases and dysfunctions but also as a blueprint for developing artificial neural networks and neurocomputers. A key feature of neural networks is their ability to learn from examples and automatically generalise the knowledge acquired. While the human brain serves as the prototype for artificial neural networks, these models are highly simplified representations of the brain's complexity. The development and understanding of artificial neural networks have been significantly influenced by findings in neurophysiological and bionic research.

One of the earliest practical implementations of a neural network was the perceptron, which retains functionality even when some of its components are damaged (Tadeusiewicz, 1993). The perceptron model has proven effective in industrial enterprise management for identifying operational bottlenecks and enhancing performance (Kwilinski & Kuzior, 2020). Other early neural network models include Adaline, Limulus, Avalanche, Cerebellatron, Brain State in a Box, Neocognitron, Adaptive

Resonance Theory, Mark III, Neural Emulator Processor, Mark IV, Odyssey, Crossbar Chip, Anza, Parallon, and Anza Plus (Tadeusiewicz, 1993). Currently, approximately 30 types of neural network models are recognised, and they are categorised as convolutional, recurrent, and generative.

Today, artificial intelligence and cognitive technologies influence nearly every aspect of social and individual life. Their impact extends across various domains, including management, accounting, auditing, finance, banking, social security, economic development and economic growth. They also play a significant role in education, cognitive engineering systems, ethics, and neuroethics. Additionally, these technologies are reshaping the labour market, marketing strategies, and the ways human needs and behaviours are created and understood. Their effects are also evident in medicine and healthcare, natural language processing, machine translation, and many other fields (Kuzior & Kwilinski, 2022).

Discussion of the term

Cognitive technologies are grounded in the large-scale architecture of the brain and models of the functions of specialised brain areas, which are known in neurocognitive computing as cognitive architectures. Given the close correlation between brain processes and mental processes, cognitive architectures can be considered computer models of the functions of the human mind. Many cognitive architectures focus on describing human-machine communication and interaction (Human-Computer Interfaces) and are thus concerned with cultural practices.

Cognitive architectures that are based on a logical approach and simulate higher-order cognitive activities such as thinking and reasoning can be categorised into three main types:

- symbolic architectures;
- emergent architectures,
- hybrid architectures (Duch, 2010).

In cognitive architectures, cognitive functions are supported by different types of memory:

- perceptual memory: identifies familiar objects and detects deviations from the familiar,

- associative memory: automatically draws simple conclusions,
- procedural memory: stores information related to manual skills and actions,
- semantic memory: interprets meaning
- working memory: combinatorially integrates pieces of information into a larger whole.

However, human brains and computer 'brains' differ fundamentally (Duch, 2010). Humans can make independent decisions and store vast amounts of information, whereas computers require programming to perform specific functions and have limited data storage capacities. Additionally, the underlying mechanisms of the human brain and computer processes are distinct, with the former relying on complex neuronal networks and the latter on silicon-based circuitry.

Cognitive technologies enhance human cognitive abilities, making it crucial to understand how the human mind functions. The complexities of the mind are progressively being uncovered through cognitive science, an interdisciplinary field that integrates insights from cognitive psychology, neuroscience, philosophy of mind, artificial intelligence, linguistics, logic, physics, and other disciplines (Bremer, 2016; Duch, 1998). Research in cognitive science leads to a better understanding of the properties and functions of the human brain. It explores learning, language, thinking, perception, consciousness, decision-making, and intelligence, both natural and artificial. Cognitive science emerged in 1956 from the field of cybernetics, pioneered by Norbert Wiener. Key figures such as Claude Shannon and Noam Chomsky have made significant contributions to its development. Since 1975, cognitive science has been recognised as an independent scientific discipline in the USA (Lakomy, 2023). Despite centuries of research and advances across various scientific domains, the mysteries of the human brain and mind remain only partially understood. Humans strive to understand themselves, yet the human mind, with its psychological and conscious layers, and the human brain, with its biological layer, remain unparalleled phenomena, continually inspiring further exploration and discovery of their secrets.

Technological progress has not only posed theoretical challenges but also created practical and social problems for everyday users. As science and technology evolve rapidly, individuals often find themselves struggling to cope with their complexities and unintended consequences.

The pace of technological development can leave people feeling helpless, unable to anticipate or mitigate the negative impacts of these innovations. Modern technologies are addictive and can potentially harm people's psychological well-being, leading to issues such as emotional numbness, boredom, and a sense of aimlessness of human life.

Developments in cognitive technologies and artificial intelligence currently include efforts to create artificial minds for humanoid robots. A notable example is the Phoenix humanoid robot, created by the Canadian Tire Corporation (CTC). This robot, equipped with legs, arms, and a head similar in size to that of an average human, weighs 70 kg (155 pounds), is 170 cm (5 feet 7 inches) tall, and exhibits human-like full body mobility. It is capable of moving at speeds of up to three miles per hour (five km/h) and carrying loads up to 55 pounds (25 kg). However, what distinguishes Phoenix is its sophisticated artificial 'mind', modelled to replicate the subsystems found in the human brain. This artificial mind is equipped with memory and sensory capabilities such as sight, hearing, and touch. It is designed to imitate the human mind, which requires implementing mental processes related to thinking and consciousness, particularly the modelling of cognitive and creative functions through networked means of logical and associative information processing, as well as neuromorphic approaches to processing information (Stankevich, 2022).

Interest in humanoid robots, particularly social robots, has surged dramatically in recent years, growing at what can only be described as an exponential rate. International scientific journals devoted to this field include *International Journal of Social Robotics*, *Computers in Human Behavior*, *International Journal of Human-Computer Interaction*, and *International Journal of Humanoid Robotics*. Numerous humanoid robots are currently available on the market, effectively replacing humans in various traditionally human activities such as customer service, scientific research, and conferences. Some of the most well-known humanoid robots include Sophia, Pepper, Optimus, Mika, Apollo, Valkyrie, and ASIMO, among others. Ongoing research focuses not only on computational systems but also on integrating cognitive and artificial intelligence technologies with other innovations, such as blockchain.

A significant aspect of this research is understanding public perceptions of human-robot interactions and evaluating the capability of intelligent machines to fulfil human needs. A study involving 700 participants

revealed that people are willing to benefit from the support offered by intelligent robots in several areas, including: reading human mood (56.7%); reading human thoughts (27.6%); reading human non-verbal cues (57%); reading and analysing human biomedical parameters (53.7%); reading and analysing human biomedical parameters and calling for medical assistance if necessary (46.9%); responding to and following human verbal commands (53.6%); recognising and meeting human emotional needs (47%); recognising and meeting human intimate (sexual) needs (35.9%); balancing and supervising human diet (36.4%) (Kuzior & Kwilinski, 2022).

The development of intelligent humanoid robots raises significant ethical questions. As artificial intelligence and cognitive technologies advance, modern robots are increasingly capable of performing tasks and intellectual activities traditionally reserved for humans. This prompts questions about whether ethical oversight should extend to robots. Traditionally, ethics have been concerned with humans as moral subjects and their moral actions and moral responsibilities. The rapid progress of cognitive technologies has sparked debate on whether subjectivity can be attributed to intelligent robots and whether they can be granted the status of ethical subjects. Ethical considerations in this area typically involve concepts of causation and intentionality. While robots can certainly be considered causal subjects (if this term can be used here) in a mechanical sense, the scope of their actions is ultimately determined by humans. Passive causation might be considered if robots were to refrain from certain actions. When it comes to intentionality, we enter the realm of normativity, a concept traditionally ascribed exclusively to humans in both its logical and conceptual dimensions. For instance, we do not punish a swarm of bees for stinging someone. Although it is conceivable that robots could be equipped with a certain scope of normativity, currently even the best-implemented rules of ethical conduct will not make a machine a subject of moral responsibility because responsibility requires freedom. If a machine can only act according to predefined rules, it can neither be praised nor reprimanded for its actions.

However, programmers continuously work in this area. If it becomes feasible to endow robots with ethical values and link causality with intentionality, it might be possible to attribute a form of subjectivity and, consequently, responsibility to these machines. In other words,

transforming passive subjectivity into active subjectivity would require meeting several additional conditions. These issues remain under active discussion (Kuzior, 2021).

Cognitive technologies are increasingly built upon generative artificial intelligence solutions. The year 2022 marked a significant breakthrough with the release of ChatGPT by OpenAI. This technology enables real-time human-to-machine and machine-to-human communication in natural language. ChatGPT represents a major innovation, facilitating dialogue between a human and an artificial neural network through deep learning techniques. Its applications span a wide range of uses, from generating poetry in specific literary styles to providing recipes for yeast dough. However, its potential extends far beyond these examples. On the other hand, there are also malicious variants like WormGPT, which lack ethical constraints and are exploited by cybercriminals for hacking and phishing activities (Lakomy, 2023).

Systematic reflection with conclusions and recommendations

Generative artificial intelligence (Gen-AI) has become pervasive in various fields, including education. The educational "Shifting Horizons: Transformative Trends Reshaping the Landscape of Higher Education" report (2023) highlights the urgent need to regulate the use of emerging technologies based on generative artificial intelligence. Students are increasingly seeking institutional approval to integrate these tools into their educational processes. Currently, over 52% of Polish students use tools like ChatGPT or Bard, which are powered by large language models (LLMs). Nearly 63% of these students believe that ChatGPT aids them in quickly finding information and solving academic problems (Raport edukacyjny, 2023). Of course, declaring that AI helps solve problems does not imply that a student-user believes it unreservedly and is unaware of its limitations. Below, two critical issues are addressed: the responsible use of cognitive technologies and the ethical concerns associated with legal regulations.

Modern technologies (including cognitive technologies) are shaping a new type of society built on the previous types: the information society

and the knowledge society. When it comes to the responsible use of cognitive technologies in this new society, two scenarios are possible: a positive one and a negative one. The positive scenario envisions using cognitive technologies to enhance and improve the quality of human life. The negative scenario involves a potential societal divide between the creators and users of IT and ICT technologies. This divide could lead to polarisation, where the creators and users dominate all economic sectors and aspects of human life, while those who are technologically illiterate are marginalised. However, this 'social margin' might not face the same risks of intellectual decline as unreflective users, who could lose their ability to think independently, reflect on reality, and maintain natural creativity. Intellectual complacency among the majority of users could lead to biological and psychological changes that result in a regression of human qualities such as consciousness, free will, morality, creativity, abstract thinking, symbolic reasoning, and cooperative skills rooted in natural language. Human thinking, reasoning, and decision-making are based on personal experiences, motives, and emotions that distinguish us from machines. While computers and robots today can execute tasks based on pre-programmed instructions, they lack the capacity for independent thought or decision-making.

A further consequence of the increasing blurring of the differences between humans and machine-robots equipped with ever-advancing artificial intelligence, combined with human intellectual complacency, could be a growing dependence of natural intelligence on artificial intelligence, potentially leading to learned helplessness. The risks of information manipulation, data distortion, and the erosion of boundaries between the real and virtual worlds will also intensify (Kuzior, 2017). Elon Musk's vision, widely promoted on social media, of a society coexisting with intelligent machines might seem somewhat dystopian. However, some signs of this development are already evident, such as driverless cars navigating traffic. But who bears responsibility when such a car causes an accident?

In response to these challenges, interdisciplinary Cognitive Technologies degree programs are being introduced at universities in the Netherlands (Eindhoven University of Technology), the UK (University of Hertfordshire), and Germany (Universität Osnabrück). Additionally, the Silesian University of Technology in Gliwice offers a second-cycle

programme “Cognitive technologies”, which blends traditional cognitive science with modern cognitive technologies. This programme addresses the needs of a rapidly evolving market that demands specialists capable of leveraging advanced technologies, particularly within modern business services focused on sustainable development.

Ultimately, it is students, not AI, who should leverage their cognitive abilities to solve increasingly complex problems. AI should serve only as a supportive tool. Ready-made hints and solutions provided by AI do not train the human brain; rather, they accustom it to reproducing ready-made patterns that may contain errors, distortions and entrenched stereotypes. For example, ChatGPT can sometimes generate misleading or incorrect information due to its ‘hallucinatory’ tendencies when processing vast amounts of data inaccurately. However, this does not mean we should reject such cognitive tools. Instead, it is crucial to learn how to use them effectively and adhere to best practices.

Regarding ethical and regulatory concerns, ChatGPT stands out as a prominent example of cognitive technology with broad applications (Lakomy, 2023). Yet, its use can pose several risks, including unauthorised sharing of personal data, vulnerability to cyberattacks, unclear policies on data storage and privacy, and potential legal issues related to copyright and intellectual property in training generative AI models. Additionally, ChatGPT may inadvertently generate disinformation or manipulate content, which could influence election results, tarnish individuals’ reputations (either randomly or deliberately), or impact financial markets. Users may fall prey to disinformation and unknowingly spread false content with good intentions. Furthermore, there is a risk that GenAI could be misused to fabricate data to support pre-existing scientific hypotheses (Raport edukacyjny, 2023).

The responsible development and application of artificial intelligence and cognitive technologies hinge on robust legal and ethical regulations. In the European Union, efforts to establish a framework for ‘ethical AI’ began with an expert group in 2018. The primary aim is to make the EU “a global leader in innovation in the data economy and its applications”. The “White Paper on Artificial Intelligence: A European approach to excellence and trust” specifies its aim as the creation of two key ecosystems: an ‘ecosystem of excellence’, which focuses on policy frameworks and partnership between the public and the private sectors, and

an ‘ecosystem of trust’, which emphasises a regulatory framework that protects fundamental rights and consumers’ rights, while also encouraging citizens to take up AI applications. The White Paper identifies seven essential principles for developing trustworthy AI:

1. Human agency and oversight,
2. Technical robustness and safety,
3. Privacy and data governance,
4. Transparency,
5. Diversity, non-discrimination and fairness,
6. Societal and environmental wellbeing,
7. Accountability.

The document also outlines both the opportunities and risks associated with the development of artificial intelligence. Opportunities include “improving healthcare (e.g., making diagnosis more precise, enabling better prevention of diseases), increasing the efficiency of farming, contributing to climate change mitigation and adaptation, improving the efficiency of production systems through predictive maintenance, increasing the security of Europeans” and many others, while potential risks cover “opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes”. Ethical regulations for AI development must be founded on respect for fundamental rights, which include:

1. Respect for human dignity (ensuring individuals are treated as moral agents rather than objects to be manipulated or categorised).
2. Protection of individual freedoms (safeguarding personal sovereignty, ensuring equal access to opportunities, and protecting privacy and security).
3. Respect for democracy, justice and the rule of law (using AI to support democratic values rather than undermine them or facilitate manipulation).
4. Ensuring equality, non-discrimination and solidarity (preventing biased outcomes and ensuring transparency and data quality).
5. Protection of citizens’ rights (guaranteeing that all individuals have equal rights).

Based on fundamental rights the following four ethical principles have been defined:

1. Respect for human autonomy.
2. Prevention of harm.
3. Integrity.
4. Credibility (Kuzior, 2021).

In the European Union, extensive work and negotiations have culminated in regulations governing the use of artificial intelligence. These regulations are particularly pertinent to cognitive technologies reliant on generative AI models (such as GPT-4 or Gemini), deep machine learning, and artificial neural networks. The AI Act (2024), recently passed by the European Parliament, introduces a risk-based approach to regulation, grouping AI systems into four risk categories:

1. Minimal risk – applies to all AI models that do not fall into the other three risk categories.
2. High risk – a limited number of AI systems potentially creating an adverse impact on people's safety or their fundamental rights (as protected by the EU Charter of Fundamental Rights).
3. Product safety risks covered by sectoral EU legislation.
4. Unacceptable risk – a very limited set of particularly harmful uses of AI that contravene EU values because they violate fundamental rights and will therefore be banned:
    a. Exploitation of vulnerabilities of persons, manipulation and use of subliminal techniques;
    b. Social scoring for public and private purposes;
    c. Individual predictive policing based solely on profiling people;
    d. Untargeted scraping of internet or CCTV for facial images to build-up or expand databases;
    e. Emotion recognition in the workplace and education institutions, unless for medical or safety reasons (i.e. monitoring the tiredness levels of a pilot);
    f. Biometric categorisation of natural persons to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs or sexual orientation. Labelling or filtering of datasets and categorising data in the field of law enforcement will still be possible;
    g. Real-time remote biometric identification in publicly accessible spaces by law enforcement, subject to narrow exceptions including 16 specified very serious crimes: terrorism;

> trafficking in human beings; sexual exploitation of children and creation of child pornography; illicit trafficking in narcotic drugs and psychoactive substances; illicit trafficking in weapons, munitions and explosives; deliberate homicide; serious causing of bodily harm; illicit trade in human organs and tissue; illicit trafficking in nuclear or radioactive materials; kidnapping, illegal restraint and hostage-taking; offences within the jurisdiction of the International Criminal Court; unlawful seizure of aircraft and ships; rape; environmental crime; organised and armed robbery; sabotage, participation in a criminal organisation (AI Act, 2024).

Given the rapid expansion of cognitive technologies and their pervasive applications, there is an urgent need for specialists who integrate technical expertise with humanistic and ethical considerations. Such professionals will be essential in shaping the ethical landscape of cognitive technologies and ensuring responsible development and implementation of cognitive architectures.

References

Act AI (2024), https://odo24.pl/blog-post.co-zmieni-unijny-akt-w-sprawie-sztucznej-inteligencji. Accessed 28.08.2024.

Bainbridge, W.S. (2006). Cognitive technologies. In: *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society* (pp. 203–226). Dordrecht: Springer.

Bremer, J. (ed.) (2016). *Przewodnik po kognitywistyce*. Kraków: Wydawnictwo WAM.

Duch, W. (1998). Czym jest kognitywistyka. *Kognitywistyka i Media w Edukacji*, *1*, 9–50.

Duch, W. (2009). Architektury kognitywne, czyli jak zbudować sztuczny umysł. *Neurocybernetyka teoretyczna*, *14*, 329–361.

Elia, G., & Margherita, A. (2021). A conceptual framework for the cognitive enterprise: pillars, maturity, value drivers. *Technology Analysis and Strategic Management*. DOI: 10.1080/09537325.2021.1901874.

Ernst, H.A. (1962). *MH-1, a computer-operated mechanical hand*. AFIPS Proceedings – 1962 Spring Joint Computer Conference, AFIPS, 142886, 39–50. DOI: 10.1145/1460833.1460839.

Kuzior, A. (2021). *Applied ethics*. Lublin: Wydawnictwo Naukowe Tygiel.

Kuzior, A. (2017). Problem bezrobocia technologicznego w perspektywie rozwoju Przemysłu 4.0. *Etyka Biznesu i Zrównoważony Rozwój*, *4*, 31–38.

Kuzior, A., & Kwilinski, A. (2022). Cognitive Technologies and Artificial Intelligence in Social Perception. *Management Systems in Production Engineering*, *2*, 109–115. DOI: 10.2478/mspe-2022-0014.

Lakomy, M. (2023). Technologie kognitywne a demokracja. *BELLONA QUART.*, *3*, 31–42. DOI: 10.5604/01.3001.0054.1831.

Minsky, M. (1961). Steps Toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1). DOI: 10.1109/JRPROC.1961.287775.

Newell, A. (1963). Documentation of IPL-V. *Communications of the ACM*, *6*(3), 86–89. DOI: 10.1145/366274.366296.

Rouse, M. (2018). Cognitive technology. In: *Techopedia*. https://www.techopedia.com/definition/32482/cognitive-technology. Accessed 12.12.2023.

Shifting Horizons – Transformacyjne trendy zmieniające krajobraz szkolnictwa wyższego. (2023). Coopernicus by Our Future Fundation. https://coopernicus.pl/wp-content/uploads/2024/01/Raport-Edukacyjny-PL.pdf

Stankevich, L.A. (2022). Cognitive Technologies and Artificial Mind for Humanoid Robots. Studies in Computational Intelligence, 1008 SCI, 3–8. DOI: 10.1007/978-3-030-91581-0_1.

Tadeusiewicz, R. (1993). *Sieci neuronowe*. Warszawa: Akademicka Oficyna Wydawnicza.

Woods, D.D. (1985). Cognitive Technologies: The Design of Joint Human-Machine Cognitive Systems. *AI Magazine*, *6*(4), 86–92. DOI: 10.1609/aimag.v6i4.511.

List of authors

Dariusz Adamek, Jagiellonian University in Krakow
https://orcid.org/0000-0002-5570-2259
Nina Bażela, Jagiellonian University in Krakow
https://orcid.org/0009-0003-1147-1086
Jacek Bielas, Ignatianum University in Cracow
https://orcid.org/0000-0001-9771-4162
Andrzej Bielecki, AGH University of Krakow
https://orcid.org/0000-0002-0192-3785
Józef Bremer, Ignatianum University in Cracow
https://orcid.org/0000-0001-9664-8896
Andrzej Dąbrowski, University of the National Education Commission in Krakow
https://orcid.org/0000-0002-3998-9608
Mariusz Flasiński, Jagiellonian University in Krakow
https://orcid.org/0000-0002-0910-4928
Paweł Fortuna, The John Paul II Catholic University of Lublin
https://orcid.org/0000-0002-0633-4453
Mateusz Hohol, Jagiellonian University in Krakow
https://orcid.org/0000-0003-0422-5488
Ewa Jaglarz, Ignatianum University in Cracow
https://orcid.org/0000-0001-8754-8770
Mateusz Jarmużewski, Windesheim University of Applied Sciences
https://orcid.org/0000-0002-8398-1398

Piotr Konderak, Maria Curie-Skłodowska University in Lublin
https://orcid.org/0000-0002-9161-4870

Aleksandra Kuzior, Silesian University of Technology
https://orcid.org/0000-0001-9764-5320

Marek Lechniak, The John Paul II Catholic University of Lublin
https://orcid.org/0000-0002-0768-7963

Ewa Odoj, Ignatianum University in Cracow
https://orcid.org/0000-0003-0821-9317

Anita Pacholik-Żuromska, Nicolaus Copernicus University in Toruń
https://orcid.org/0000-0002-8674-9412

Marek Pokropski, University of Warsaw
https://orcid.org/0000-0002-7423-6951

Sławomir Sztajer, Adam Mickiewicz University in Poznań
https://orcid.org/0000-0001-8975-2994

Michał Wierzchoń, Jagiellonian University in Krakow
https://orcid.org/0000-0002-7347-2696

Łukasz Wiraszka, Jagiellonian University in Krakow
https://orcid.org/0000-0002-3653-5926

Zbigniew Wróblewski, The John Paul II Catholic University of Lublin
https://orcid.org/0000-0003-4477-6903