CROATIAN
JOURNAL
OF PHILOSOPHY

Vol. XXIV · No. 72 · 2024

Explorations in Human and Artificial Cognition

Introduction
ANITA AVRAMIDES 329

Concepts are Containers
ROBERT O'SHAUGHNESSY and MARK SPREVAK 333

How is Content Externalism Characterized by Vehicle Externalists
DUNJA JUTRONIĆ 351

Intentions and Representations
SHAUN GALLAGHER 367

Minds, Machines and Gödel
ZVONIMIR ŠIKIĆ 381

Human and Artificial Decision Making: A Unified View
KONSTANTINOS V. KATSIKOPOULOS 387

Table of Contents of Vol. XXIV 397

Introduction

Kathleen Vaughn Wilkes (known to all as Kathy) was a Tutor and Fellow of Philosophy at St. Hilda's College and Lecturer in the Faculty of Philosophy at Oxford University from 1973 until her untimely death in 2003. In April 2018—the year of St Hilda's 125th Anniversary—the College celebrated her life and work by holding a two-day conference in her honour. Close to one hundred people gathered from all around Britain and Europe to share memories of Kathy both as a philosopher and as a political activist.

Kathy's political activities are well known to many. Her political work began in 1979 when she became involved with the dissident political community in Prague, Czechoslovakia (today the Czech Republic). She worked alongside political dissidents in that country, helping to bring about what has come to be known as the Velvet Revolution, bringing the then Czechoslovakia out from under 40 years of Communist rule. The work Kathy did in Czechoslovakia she did in the company of Sir Anthony Kenny, Sir Roger Scruton, Professor Denis Noble, and Professor Bill Newton Smith. These luminaries were among the academics who were invited by Kathy to give seminars in Prague—only to have their seminars raided by the authorities and their persons escorted to the border. This was duly reported in the Times of London, which, in turn, helped to raise awareness of what was happening to our colleagues in Czechoslovakia. For her work in support of this community, President Václav Havel awarded Kathy the Commemorative Medal of the President of the Czech Republic in October 1998. Her work in Eastern and Central Europe was not restricted to Czechoslovakia/The Czech Republic. In 1981 she began her association with the Inter-University Centre (the IUC) in Dubrovnik, an association that was to last until her death in 2003. Over the many years of her association with the IUC Kathy organized courses in the philosophy of science together with Professor Bill Newton-Smith and others. The year 1991 saw the beginning of yet another chapter in Kathy's political career. This time it was to defend the Croatian cause in their war of Independence. Her efforts on behalf of this cause are numerous (for those who want to learn more about Kathy and the city of Dubrovnik I refer you to the paper by Nada Bruer Ljubišić, the Croatian Journal of Philosophy*, Vol. XXII, No. 66, 2022). What I will record here is the fact that Kathy was awarded honorary citizenship for her efforts on behalf of the city of Dubrovnik, and her portrait has been permanently placed in the City's Council Hall.*

At the end of the 2018 Conference in her honour at St Hilda's the then Croatian Ambassador to the United Kingdom, Igor Pokaz, approached the then Principal of St Hilda's, Sir Gordon Duff, urging that we re-establish the link that Kathy had built with the IUC in the decades before her death. As the Southover Manor Trust Fellow and Tutor in Philosophy at St. Hilda's College at that time it was my job to make this happen. After much discussion, I managed to put together a Memorandum of Agreement (for three years in the first instance) between the IUC and St Hilda's to hold a yearly Kathy Wilkes Conference in Cognitive and Social Science. The Chair of the Herbert Simon Society in Turin, Italy, Professor Riccardo Viale, a one-time academic collaborator with Kathy, asked if that organization might join in. It was agreed that these three Institutions would take it in turns to organize a conference each Spring on a topic that had links with some aspect of Kathy's work.

Kathy was trained as a philosopher of Ancient Greek and, indeed, her Fellowship at St Hilda's was to teach the Classics part of the course. While Kathy taught and published papers in ancient philosophy, she also did ground-breaking work in the philosophy of mind. She was extremely interested in keeping up with developments in psychology and the nascent neurosciences, and she was as knowledgeable about certain issues in these disciplines as she was about a large range of topics in philosophy. Much of her published work centred around issues to do with personal identity and the self. Her books, Physicalism *(published by Routledge in 1978), and* Real People *(published by Oxford University Press in 1988) are still read by students today. She also published* Modelling the Mind *in 1990, a volume edited along with K. A. Mohyeldin Said, W. H. Newton-Smith, and R. Viale and also published by Oxford University Press.* Modelling the Mind *came out of inter-disciplinary work that Kathy was engaged in with colleagues in the social sciences. Kathy was ahead of her time when she engaged in this interdisciplinary work. She was also very involved in interdisciplinary work with psychologists and physiologists (many of whom we would now classify as neuroscientists)—work which at that time was brought together under the heading of 'the cognitive sciences.' While we aim to celebrate all of Kathy's work, we have chosen to highlight her interdisciplinary work in the cognitive and social sciences in our annual conference in her honour.*

The first of our memorial conferences took place in Dubrovnik in the Spring of 2021. The topic for the conference was "(Re) Assessing Goal Directed Activity," and we were delighted to have as our Inaugural speaker a one-time collaborator and great friend of Kathy's Professor Denis Noble, currently Emeritus Fellow at Balliol College Oxford. The proceedings of that conference were published in the issue of the Croatian Journal of Philosophy *mentioned above.*

In the Spring of 2022, the second of the Kathy Wilkes Conferences was held at St Hilda's College, Oxford. The topic for discussion was Explorations in Human and Artificial Cognition. *The speakers were scheduled*

as follows: Professor Mark Sprevak, Edinburgh University, Professor Dunja Jutronić, University of Split, Croatia, Professor Shaun Gallagher, University of Memphis, Professor Konstantinos Katsikopoulos, University of Southampton, Professor Philipp Koralaus, St Catherine's College Oxford, Professor Zvonimir Šikić, University of Rijeka, Center for Logic and Decision Theory, Croatia. At the last minute, Professor Sprevak was unable to join us, and Professor Peter Milllican stepped in to replace him. While Peter gave us a terrific paper on Alan Turing and the use of his work in connection with human-like intelligence, the pressures of time mean that we are unable to publish his paper in this volume. However, Mark Sprevak (together with Robert O'Shaughnessy from the University of Edinburgh) has given us a paper to include in this volume. Finally, we are not able to publish the paper by Professor Koralus, although we can refer readers to his book, Reason & Inquiry: the erotetic theory*, published by Oxford University Press in 2022. The paper he gave at the conference was based on this book which had been published only a few months earlier.*

We are delighted that the Croatian Journal of Philosophy *has once again chosen to publish the proceedings of the second of our Kathy Wilkes Memorial Conference in Social and Cognitive Science.*

ANITA AVRAMIDES
Emeritus Fellow in Philosophy
St Hilda's College, Oxford

Croatian Journal of Philosophy
Vol. XXIV, No. 72, 2024
https://doi.org/10.52685/cjp.24.72.1
Received: October 18, 2024
Accepted: November 19, 2024

Concepts are Containers

ROBERT O'SHAUGHNESSY and MARK SPREVAK
University of Edinburgh, Edinburgh, UK

In this paper, we propose and defend a theory of concepts. According to Machery (2009), psychologists and philosophers mean different things by 'concept'. Psychologists mean bodies of knowledge used to categorise and infer; philosophers mean constituent of propositional thought. Machery's conclusion would drive a wedge between contributions by psychologists and philosophers on concepts. Theories about the former would have no clear role to play in, and cast no light on, the latter, and vice versa. We argue that, on the contrary, 'concept' has a single core meaning: a container of stored knowledge pertaining to a single category. This single meaning satisfies both the theories of psychologists and philosophers. The divergence in use of the term 'concept' on which Machery focuses arises because words for containers are often used to refer to (a) what is contained by the container and (b) the label of a container. Our account explains what a concept is, and how one might be misled by Machery's challenge.

Keywords: Concepts; mental files; pointers; language of thought; eliminativism; Machery; Fodor.

1. *Introduction*

Machery (2009) claims that philosophers and psychologists mean different things when they use the term 'concept'. Machery identifies two desiderata for something to be a concept:

1. *Judgement desideratum*: A concept must permit us to make categorisation, typicality, and inferential judgements.
2. *Propositional desideratum*: A concept must be capable of being used as a constituent in propositional thought.

Machery claims that the term 'concept' means something different when it is used to satisfy the first desideratum than it does when it is used to satisfy the second. Our claim is that 'concept' has a single core meaning that satisfies both desiderata: *a container of stored knowledge pertain-*

ing to a single category.[1] If we are right, then philosophers and psychologists are talking about the same thing when they talk about concepts.

To explain the divergence in use of the term 'concept' that Machery describes, we claim that psychologists and philosophers extend a single core meaning in different ways. Unlike Machery, we do not conclude that there are two entirely distinct entities that stand behind our concept talk. Rather, we argue that we have a case of polysemy patterned on forms of polysemy familiar when containers are in play. 'Concept', in this respect, is like 'DVD': the word for the container may be used to refer to the contents of the container (for example, "that DVD was boring" meaning *the movie on that DVD was boring*) or the word for the container ('DVD') may be used to refer to the title or label of the container (for example, handing someone a list of movie titles and asking her to "pick one of those DVDs" meaning *pick one of those titles of, or labels for, DVDs*). 'DVD' may be used to refer to the *movie* or its *associated title*, but no one would think that a DVD *is* a movie or that a DVD *is* a title. What a DVD *is* is a container of stored information.

Our claim is that a concept is a container of stored knowledge that pertains to a single category. Psychologists tend to refer to the *contents* of the container. Philosophers tend to refer to the *label* of the container. No one should conclude from this that a concept *is* the contents or that it *is* the label, any more than they would for 'DVD'.[2]

2. *Machery's claim*

Machery (2009) claims that philosophers and psychologists refer to two entirely distinct entities by 'concept'. Philosophers are disposed to favour models in which, for example, RED is an atomic concept linked to other related concepts (such as COLOUR) which do not leave a role for the exemplars and prototypes that psychologists think of as concepts and which facilitate categorization and inference. Psychologists favour

[1] We use the term 'knowledge' the way Machery and psychologists do: as a term for "any contentful state that can be used in cognitive processes" (Machery 2009: 8). States of knowledge in this sense do not have to be true or justified, nor do they have to be explicit or propositional: they may be imagistic (perceptual) or procedural (sensorimotor).

[2] What we are proposing could be described as an elaborated version of the "different aspects" response to Machery's claim (Machery, personal communication). That is to say, philosophers and psychologists are talking about different aspects of the same thing, not about different things (for responses along this line, see Margolis and Laurence 2010; Piccinini 2011). Machery replies to these objections that the aspects that satisfy the propositional desideratum play no role in theories about the aspects that satisfy the judgement desideratum, and vice versa. This suggests that there are two different entities, not one, involved in concepts. Our account differs in that we take there to be a *third* entity—a container of stored knowledge—and we hold that psychologists are interested in one aspect of this entity (its contents) and philosophers in another aspect (its label). This single third entity unifies the (otherwise puzzlingly different) multiple aspects of concepts and satisfies both Machery's desiderata.

models in which concepts are pieces of information (exemplars, prototypes and theories) with rules of engagement. But there do not seem to be good rules of engagement that allow such information to be constituents of propositional thought. If correct, Machery's claim would create a puzzle. Philosophers' concepts and psychologists' concepts would, in principle, have nothing in common. When you form the propositional thought THE DOG CHASES THE CAT, your understanding of the proposition could be entirely unrelated to your ability to make the kinds of judgements in which psychologists are interested about the constituent concepts DOG, CAT, CHASE, and vice versa. But divorcing the interests of philosophers and psychologists here seems too strong. Rich connections exist between philosophers' concepts and psychologists' concepts. These connections are obscured and relegated to purely contingent happenstance on Machery's account.

Chalmers (2011) makes a distinction between verbal disputes and substantive disputes. A verbal dispute arises when parties take themselves to be using an expression for which they both have the same proposition in mind when, in fact, they have different propositions in mind. The parties disagree, failing to realise that they are talking about different things. Once it is discovered that they are using the same term in different ways the dispute may vanish. For a substantive dispute, the parties use the expression in the same way but they disagree about the underlying facts.

Machery (2009) is, in effect, claiming that disputes between philosophers and psychologists over the nature of concepts are verbal disputes. Philosophers and psychologists use the term 'concept' and they might assume that they are talking about the same thing. But, according to Machery, they are talking about different things. For example, when Fodor says, "most of what contemporary cognitive science believes about concepts is radically, and practically *demonstrably*, untrue" (Fodor 1998: viii), Fodor (according to Machery) is engaged in a verbal, rather than a substantial, attack on cognitive science. Our claim, *contra* Machery, is that philosophers and psychologists mean the same thing by 'concept'. There is scope for substantive, not merely verbal, dispute between philosophers and psychologists about concepts.

3. *Polysemy*

Polysemy arises when a term has multiple meanings that are semantically related. For example, the term 'bank' is polysemous: 'bank' means both *financial institution* and *physical building* where this institution offers services—the words have different meanings but they are semantically related. Polysemy differs from mere homonymy. Homonymy arises when a term has multiple meanings that may be semantically unrelated. The term 'bank' also functions as a mere homonym: it means both *financial institution* and *side of a river*—different meanings that are semantically unrelated.

A polysemous term that will prove instructive to us later is 'table'. Murphy (2002) lists fourteen semantically related meanings of 'table' including: what furniture makers mean by 'table' (*four-legged piece of furniture*), what geographers mean by 'table' (*a flat or level area of land*), and what jewellers mean by 'table' (*a facet of a cut precious stone*). Murphy claims that these meanings are related because they share a common etymological origin. Some original meaning of 'table' (according to Murphy, *four-legged piece of furniture*) was extended to generate the rich palette of meanings now associated with 'table'.

Machery observes that philosophers and psychologists use the term 'concept' in different ways. Our claim, and Machery's view, is not that 'concept' is a mere homonym. Machery's proposal is that 'concept' is polysemous: it has multiple, semantically related meanings that pertain separately to philosophy and psychology. What is distinctive about Machery's proposal is that 'concept' is polysemous *in a particular way*: 'concept' is polysemous in the mouths of philosophers and psychologists in roughly the same way that 'table' is polysemous in the mouths of furniture makers and geographers. In both cases, *entirely distinct and independent entities* are denoted by the same linguistic expression. 'Table' as used by furniture makers refers to a *piece of furniture*. 'Table' as used by geographers refers to a *flat piece of land*. Machery claims that 'concept' as used by philosophers refers to a *constituent of propositional thought*; 'concept' as used by psychologists refers to a *body of knowledge used for classification*. There would be no point in a furniture maker and geographer entering into a dispute about whether "tables have four legs" or whether "tables in South Africa are sedimentary deposits". Similarly, there would be no point in a philosopher and psychologist entering into a dispute about whether "concepts govern typicality judgements" or whether "concepts are involved in linguistic thought".

We agree that 'concept' is polysemous, but we disagree about the kind of polysemy involved. The polysemy involved in 'concept' is not the same as that exhibited by 'table' as used by furniture makers and geographers. A referring term, such as 'table' or 'concept', may be polysemous without denoting entirely distinct and independent entities. If we are correct, then despite the divergence in use of 'concept' that Machery describes, both philosophers and psychologists can and should agree about what a concept is.

4. *Polysemy without proliferation*

Consider two other ways in which 'table' is polysemous (Murphy 2002: 404):

a. *the company of people eating at a table*: "The entire table shared the plate."
b. *a painting, sculpture or photograph of a table*: "The table is painted very soulfully."

Our claim is that 'concept', as used by psychologists and philosophers, is polysemous in roughly the same way that 'table' is polysemous in the preceding sentences. Psychologists follow the pattern exemplified by (a). Philosophers follow the pattern exemplified by (b).

Consider (a). Here, 'table' is used to refer to a *person or group of people*. For example, the waiter might say, "Table six wants a coffee refill". Observe that although 'table' means *person* or *people*, a second table (a piece of furniture) is also involved. One cannot use 'table' in isolation to mean *person* or *people*. One cannot point to a person walking down the street and say, "Look at that table". In contexts where 'table' means *person* or *people*, there must be also a table (a *piece of furniture*) present. This is because what is usually meant by 'table' is *person or people seated at the table*. The listener must be able to identify a table (a piece of furniture) in order to know which person or people are intended.

Compare this to a geographer's 'table'. In that case, no table (piece of furniture) is involved. Except from at the origin of the geographer's term, tables as pieces of furniture are irrelevant to the intended meaning. In contrast, for the waiter's 'table' (person or people), a table (piece of furniture) is essential to our present use and understanding. If one did not know what a table (piece of furniture) is, one would not know which person or people are being referred to. And it is because the table (piece of furniture) serves the function of grouping people together that the term can be used to refer to those people.

For psychologists, the polysemy involved in 'concept' is not that of the geographer's 'table', for which the term refers to two entirely distinct and otherwise unrelated entities (*piece of furniture* and *flat piece of land*). It is instead like that of the waiter's 'table', for which the original referent is used to identify something linked to, contained by, or grouped together by, that item. When psychologists use 'concept' what they mean, we claim, is not *body of knowledge* but *container of body of knowledge (pertaining to a single category)*.[3]

[3] Note that Machery cannot make the same response to our claim as his (2010) reply to Margolis and Laurence (2010). Margolis and Laurence suggest that we should think that concepts are mental symbols akin to words and that psychologists are interested in one aspect of these concepts (the exemplars, prototypes and theories linked to such concepts). Machery responds along the following lines. Mental symbols that are constituents of thought play no role in, and cast no light on, the theories of concepts that occupy psychologists. Indeed, Margolis and Laurence's characterization makes most of what psychologists say about concepts literally false. On the other hand, if 'concept' has the meaning set out by Machery then most of what psychologists say about concepts comes out true, and everything else being equal, this is to be preferred. Machery cannot make a similar response to our proposal. First, concepts conceived as containers do play the required role in psychological theories (see Section 7). Second, our proposal does not require what psychologists say about concepts to be false, provided it is understood as we claim: as a linguistic shorthand for *bodies of knowledge contained in (or stored in) a concept* where 'concept' means *container of stored knowledge pertaining to a single category*.

Now consider (b). The polysemy arises here because a single term is used to refer to both the object being represented and the representation of the object. We might say of a painting of a table, "That table is painted very soulfully". It may even take a moment's reflection to realise that this is a case in which 'table' deviates from the furniture makers' intended meaning. But what we mean is not that a table (a piece of furniture) is painted soulfully (perhaps the table has not been painted at all). We mean that *the image of the table* or *the representation of the table* has been painted soulfully.

Imagine that a restaurant owner has a graphical computer program that allows her to arrange representations of tables on screen to match the bookings for that evening. When she moves the outlines of tables on her screen, perhaps composing outlines into greater wholes, it is natural for her to speak of the representations on the screen as 'tables', even though they are merely proxies for tables. In the same way, we suggest, when philosophers use 'concept' what they mean is *label of a concept* or *proxy for a concept* (where 'concept' means the same as it did for psychologists: *a container of stored knowledge pertaining to a single category*).

What is distinctive about both kinds of polysemy is that both involve reference to entities that are *related to single common entity*. 'Table' refers to *person/people seated at a table* or to a *representation of a table*. Just as one cannot use 'table' to mean *person/people* but only *person/people at a table*, so one cannot use 'table' to mean *representation* but only *representation of a table*. This differs from the polysemy of the geographer's 'table' and the kind that Machery claims is involved in 'concept'. The waiter and the restaurant owner use 'table' to mean different things (*person/people seated at a table* and *representation of a table*) but they can do this only because a third entity, a table (piece of furniture), is related to both. So, we claim, psychologists and philosophers use 'concept' to mean different things (*body of knowledge for categorisation* and *constituent of thought)* but only because a third entity, a concept (*container of stored knowledge pertaining to a single category*), is related to both. A concept, we propose, is not a generic container but one which has the purpose of containing information pertaining to a single category.

5. *Container talk*

The term 'table' behaves in this way because tables are containers: tables group people together.[4] The linguistic patterns described above

[4] Note we are not claiming that the term 'table' is an ideal analogy for the term 'concept' in all respects ('table' has many more meanings than 'concept'). Rather we make use of Murphy's (2002) discussion to distinguish between what we call polysemy with proliferation (which Machery is arguing for) and polysemy without proliferation (which we are arguing for).

are characteristic of container talk. If one has a term for a container, or a thing that groups other things together, that term can also be used to refer to the *contents* grouped together or contained. 'Table' can be used to mean *person/people sitting at the table* because the table groups the people into a single category. Consider the expression, "the *x* is boring" (cf. Murphy 2002: 438). For *x* we may substitute words for *containers* when we mean *the contents of the container*. We may substitute any of the following: 'DVD', 'newspaper', 'CD', 'video', 'TV', 'file', 'website', 'room', 'bottle of wine', 'Christmas stocking', 'book', and so on. In each case, we mean not *x* itself is boring but the contents of *x* are boring. We submit that this is how psychologists use 'concept': they mean the *contents* of the concept—the stored bodies of knowledge.

A similar pattern holds for *x* meaning *representation of x*. We can use 'DVD' to refer to *the title associated with the DVD*. Suppose someone were to hand you a list of titles and say, "Pick one of those DVDs; we'll watch it tonight". No one in their right mind would reply, "What you should have said is 'pick one of those *titles* standing for a DVD containing the movie of that title'". Nobody thinks that a printed title *is* a DVD. Similarly, if given a catalogue and asked to "underline your favourite *x*s", where *x* could be 'newspaper', 'CD', 'video', 'TV', 'file', 'website', 'room', 'bottle of wine', 'Christmas stocking', 'book', and so on, what is meant is not *underline x itself* (how could one underline a website?) but *underline the label of, or proxy for, x*. We submit that this is how philosophers use 'concept': they mean the *label of* or *proxy for* the container of stored information. A DVD is a container of stored knowledge that may have an associated label. A concept is a container of stored knowledge that may have an associated label.

Container talk can rapidly switch between referring to the *containers*, referring to the *contents of the containers*, and referring to the *labels of the containers*. In normal communication these switches rarely cause a problem, but pitfalls lurk if one attempts to read off from this practice which entities stand behind the talk. Thus far, our intention has been to observe that container talk sanctions two forms of linguistic shorthand: container talk may refer to the contents of containers or to the labels of containers. With this point in mind, one should not be surprised that, if concepts *were* containers, philosophers and psychologists would use 'concept' in these two different ways even if they could agree that a concept really *is* just one thing: *a container of stored information pertaining to a single category*.

6. *A concept is a container*

But why think that a concept is such a container? We believe that the view has much in its favour. We introduce the view in this section, and in Sections 7 and 8 we argue that under this view concepts satisfy both of Machery's desiderata.

Concepts are often described as the entities that furnish the mind.[5] Our proposal is that concepts are not the furnishings but the rooms: the containers of the furnishings. What furnishes the mind are pieces of information. Concepts contain and group together those pieces of information in pertinent ways. Rooms contain furnishings. Concepts contain pieces of information or knowledge. Rooms do not contain types of room. Concepts do not contain types of concepts. Concepts contain pieces of information (encoded in the form of exemplars, prototypes, theories), but they are not identical to those pieces of information. Rooms contain pieces of furniture but they are not identical to those pieces of furniture. Inside rooms, we gather together pieces of furniture that belong together. Inside concepts, we gather together pieces of information that belong together. Information, such as "cows have udders" is not anyone's COW concept, but it could be contained within someone's COW concept. The same goes for an exemplar (for example, a visual image) of a cow, or any other type of information one might associate with cows.

Notice that the furniture inside a room is important for determining the kind of room it is. The furniture in a bedroom determines the kind of room it is. Similarly, the information inside a concept determines the kind of thing that the concept refers to. Imagine that one labels the doors of one's rooms so that someone can tell which room they are without having to look inside. The label might be a useful proxy, but it does not determine the kind of room it is. If we label the dining room 'bedroom', that does not turn it into a bedroom. However, if we were to swap all the furniture from the dining room with the furniture in the bedroom, even if we leave their labels intact, the dining room would be the bedroom. Rooms are individuated functionally by their contents, not by their labels. As we will see, concepts are individuated functionally by the information that they contain.

A drawback to the room analogy is that buildings do not contain upwards of ten thousand rooms each dedicated to a different purpose. Files are a better model for concepts.[6] Files are containers of informa-

[5] Locke famously says that the mind is furnished with ideas: "Let us then suppose the Mind to be, as we say, white Paper, void of all Characters, without any *Ideas*; How comes it to be furnished? […] Whence has it all the materials of Reason and Knowledge?" (Locke 1975: 104). Ideas, for Locke, are the constituents of propositional thought and permit us to make categorisation, typicality, and inferential judgements. Later, Hume and Reid also describe ideas as the "furniture of human understanding" (Hume 1976: 180; Reid 1983: 116). For Locke and his contemporaries, 'furniture' meant *that which furnishes* in the sense of stocking or equipping some container; it was not restricted to tables and chairs (Lewis 1967; Pasanek 2015). The furniture in the quotation, for example, are written characters. Our point is that a concept should always be identified with the container (the room, the page) not with its contents (tables, chairs, written marks). The mind contains rooms (concepts) which contain furniture (ideas).

[6] Margolis (1998); Prinz (2005); Papineau (2006); Fodor (2008); Recanati (2013) suggest mental files are a helpful model when discussing concepts.

tion. A file is suited to gathering together information pertaining to a single topic. One might imagine having thousands of files that contain pieces of information on specific topics. Files sometimes have labels that help them to be retrieved or referenced more easily. So, we argue, do concepts: concepts gather together pieces of information that pertain to a single category and they may have associated labels that provide an easy way for the rest of the cognitive system to get hold of them.

Unlike both rooms and files, concepts are functional, not spatial, containers. The pieces of information inside a concept are not located within some specific spatial boundary. The pieces of information inside a concept are grouped together, and distinguished from other information in the cognitive system, by a functional relation. To see the contrast between a spatial and a functional container, compare how a movie is stored on a DVD with how it is stored over the internet using BitTorrent. In the case of a DVD, the information that comprises the movie is spatially contained inside (it is "on") a physical container: a discrete, spatially bounded, storage disc. In the case of BitTorrent, the container is a functional, not a spatial, container of the same information. Torrents divide up the information of the movie into small chunks and each chunk is stored on a different computer or host. These host computers may be spatially scattered around the world and they may change rapidly over time. A tiny '.torrent' file contains data that indicates how to retrieve all the various chunks of information so that they can be viewed as a coherent movie. The '.torrent' file specifies a functional relationship between the pieces of information that groups them together and distinguishes them from other pieces of information on the internet. A torrent contains a movie, but it does not spatially contain it, or at least not in a sense that has any specific spatial boundary. In the same way, a concept contains pieces of information but it does not spatially contain them. A concept contains information because that information satisfies some functional relationship that allows it to be found and coherently deployed by the brain under the right circumstances. We will see some proposals for this functional containment relation in the next section.

An issue we can discuss here is that the idea of an empty concept or container of stored information might seem to be difficult to make sense of. There are two issues here. First, one might argue that a DVD can exist with nothing stored on it, but it seems odd to say that one has a concept that does not pertain to any category—that does not contain any inferential or recognitional information. We would argue that this oddness is because concepts are functional containers like torrent files. It makes no sense to open a torrent file without information to store in it and similarly it makes no sense to "open" a new concept until there is a category it is targeting (e.g. some entity in the world that the creature has encountered). However, we submit, that if concepts were spatial containers like little DVDs in the brain it would not be

odd for the brain to pre-fabricate empty ones awaiting allocation to a particular category. It does not affect our argument if one prefers to call such empty containers, for example, 'proto-concepts' and reserve the term 'concept' for when they have a certain amount of information contained in them. The second issue is related. When do we possess a concept or grasp a concept? Can we have a concept when there is only a small amount of information in the container or must we be able to fully recognize, infer, and form associated propositions? We prefer to distinguish between fledgling and fully-fledged concepts (they are both types of concept) but as with the first issue it does not affect our argument that concepts are containers. One may, if one prefers, distinguish between proto-concepts and concepts depending on how much information is in the container. To sum up, on both issues, there can be differences of opinion about when a concept (truly called) comes into existence but this does not harm our proposal that concepts are containers.

We now turn to Machery's two desiderata: the judgement desideratum and the propositional desideratum. Machery claims that no single entity, a concept, meets both the judgement desideratum and the propositional desideratum. We disagree: concepts as containers meet both. Let us consider each desideratum in turn.

7. *Meeting the judgement desideratum*

The judgement desideratum says that a concept should permit us to make categorisation, typicality, and inferential judgements. If you have a (fully-fledged) concept that satisfies the judgement desideratum, all else being equal, you should be able to do two things: (T1) identify to which category a relevant object belongs or is typical; (T2) apply information you have stored about that category to the current instance. To do this, you need to draw on two sorts of information from somewhere inside your cognitive system: (I1) information pertaining to recognition/identification of instances as belonging to, or typical of, that category; (I2) information pertaining to that category that is relevant for making inferences about and interacting with that instance. Either I1 or I2 in the absence of the other would be insufficient for you to have a concept that meets the judgement desideratum. It would be of no use to you to be able to identify a cow if you could not bring to bear information when you have identified a cow. And it would be of no use to you to draw an inference about cows if you do not know how to identify a cow. To satisfy the judgement desideratum, you need both I1 and I2.

Consider now that you likely have *many* pieces of (sometimes incompatible) information about cows under the headings I1 and I2 inside your cognitive system. You are likely to have many pieces of information that are relevant to identifying a cow. You are also likely to have many pieces of information that are relevant to drawing inferences about cows. What distinguishes the pieces of information that fall in-

side your COW concept from those that fall outside is some functional condition. Machery proposes that this is something like the default, or preferentially available, information you use for solving tasks T1 and T2. Which pieces of information get recruited, normally, rapidly, by default, to solve T1 and T2? This condition draws a functional boundary around certain pieces of information inside your cognitive system. That functional boundary unites certain pieces of information and distinguishes those pieces from others in your cognitive system.

Whether this is the right functional containment relation for concepts may be questioned. Machery proposes that the functional containment relation draws a boundary based around the pieces of information in the cognitive system that the agent uses "by default" in solving T1 and T2 (Machery 2009: 11). This condition is explained in terms of the idea of default inference in artificial intelligence: an inference that is normally drawn by the agent, except when some specific additional information is provided that defeats it. Machery equates this with information that, for the agent, is preferentially available, presumptively taken to be relevant, and spontaneously comes to mind (Machery 2009: 11–12). Prinz suggests that the relevant functional containment relation for concepts is that of information being "under organismic control": pieces of information inside a concept should be capable of being retrieved and manipulated intentionally (Prinz 2004: 45). Dennett suggests that the functional containment relation is that of being available to "call to mind": the relevant pieces of information should be capable of being objects of the agent's second-order, personal-level thoughts (Dennett 1996: 157). We do not wish to argue for the advantages of one specific functional containment relation for concepts. Our claim is merely that any theory of concepts will invoke some functional containment relation or another: it must distinguish between those pieces of information under I1 and I2 in your cognitive system that fall inside your concept from those that fall outside. Concepts are by their nature in the business of containment.

Notice that your COW concept is not, and cannot be, *identical to* the pieces of information that we claim are inside your COW concept. Those pieces of information need to be grouped together to distinguish them from other pieces of information about cows in your cognitive system. Merely enumerating *those specific pieces of information* would not suffice to specify your COW concept. And nor would it be necessary: your COW concept could involve any number of different specific pieces of information. Indeed, the specific pieces of information associated with your COW concept are likely to change over time as you learn more about cows. The pieces of information inside a COW concept are neither sufficient nor necessary for having the concept. Why those pieces of information are important is that they are grouped together by a functional relation that separates them from other pieces of information and hooks them up to behaviour in the right way (for example,

to drive your response in solving T1 and T2). The information inside your COW concept is not your COW concept. Your COW concept is a container that holds specific pieces of information about cows. A more basic concept such as RED will contain pieces of information such as exemplars and prototypes which permit inference and categorization. These pieces of information are not concepts and do not need to be relationships to anything else. Of course, as a creature becomes more sophisticated concepts may contain relationships to other concepts (e.g. RED is a COLOUR).

Interestingly, support for this view comes from Machery himself. Initially, Machery says that for a psychologist, "a concept of *x is* a body of information about *x* that is stored in long-term memory" (Machery 2009: 4, emphasis ours). Later, however, he says revealingly:

> [...] the knowledge that is *stored in a concept x* is preferentially available when we think reason and so on about *x*. So to speak it spontaneously comes to mind. By contrast, the knowledge about *x* that is *not stored in a concept of x* is less available—it does not spontaneously come to mind. (Machery 2009: 11–12, emphasis ours)

Note a shift from saying that a concept *is* a body of knowledge to saying that a concept *contains* a body of knowledge. This is precisely the kind of shift we would expect with container talk. Someone might use 'concept' to refer to the container or to its contents. This may not cause confusion in some quarters, but the difference between the two matters. Merely *having a body of information* does not suffice for having that concept. This motivated us, and is motivating Machery here, to switch to a container view about concepts. A concept is not, and cannot be, *a body of information pertaining to a single category*. That information must in addition satisfy functional constraints on use that distinguish it from other information in the cognitive system that pertains to the same category. Only information that satisfies those functional criteria is *stored inside* the concept. The concept is a container and the bodies of information are stored inside it. The container permits the categorisation, typicality, and inferential judgements that interest psychologists. Those mere bodies of information do not. Concepts as containers satisfy the judgement desideratum. Concepts as bodies of information do not.

One might have the concern that when we specify that concepts are containers that pertain to single categories all of the work is being done by the term 'category', which we have left undefined. What determines what a single category is? We believe that one of the major advantages of container theory is that it allows us to say very simply what a category is: where there is a single container, there is a single category. It is not the case that the system must first decide what constitutes a single category and then open a container for it. Instead, because containers have boundaries (information is either in the container or not), the intentional content is fixed by the contained information. The general idea is that a new container is "opened" when the system encounters some information that does not fit (according to some similar-

ity metric e.g. in exemplar theory or prototype theory) into any existing container. Information accumulates inside the container when further information that passes the similarity metric is encountered. The fact that our concepts generally accord with what we intuitively consider to be good and useful categories is explained by I1 and I2 above: things you identify as the same are the things that drive successful inferences and action and vice versa.

8. *Meeting the propositional desideratum*

The propositional desideratum is that a concept should be capable of being used as a constituent of propositional thought. Concepts allow us to contemplate and make judgements about complex states of affairs and events rather than about single categories. We claim that the same entity that satisfies the judgement desideratum also satisfies the propositional desideratum.

As Machery observes, psychologists tend to focus on categorisation, typicality and inferential judgements, whereas philosophers tend to focus on how concepts are combined to make (a potentially unbounded number of) complex propositional thoughts. Philosophers have a model of what concepts might be like to enable complex propositional thought: word-like symbols. If concepts are word-like symbols with a fixed meaning, then those symbols could be strung together, with a recursive syntax and compositional semantics, to express an unbounded number of complex propositional thoughts. This is Fodor (1975)'s language of thought hypothesis: we form complex propositional thoughts by combining atomic symbols inside our heads using a recursive syntax and compositional semantics. Fodor claims that the word-like symbols simply *are* concepts and that they directly stand for categories in the world (Fodor 1998; Fodor 2008). Your COW concept is a word-like atom inside your cognitive system that refers to *cows*. In contrast, the view we put forward is that these word-like atoms are not concepts but *proxies for* concepts (i.e. proxies for containers of stored knowledge which refer to categories in the world).

Before getting to this, we first need to show something simpler: that concepts as containers satisfy the propositional desideratum. In other words, concepts as containers *could* be the constituents of our propositional thought. In the course of establishing this, it will become clear that there is no logical *necessity* for labels or word-like atoms to be involved in propositional thought at all.

We can make progress with a simple example. Let us imagine for the moment that the atoms in question are the building blocks that young children play with. Children's building blocks sometimes have letters on them and they can be combined to form words. Suppose that our building blocks have words on them and that they can be combined to form sentences. The crucial idea here is that each building block is an atom in the sense that each block is the smallest unit that can par-

take in a construction. Within a construction, a block retains its identity and it can be manipulated or exchanged as a single item. So, for example, one could remove a block with the label 'cow' and substitute it with a block with the label 'dog' without having to do anything to the rest of the blocks.

Now imagine that our building blocks are containers (for example, crates with lids). For each concept, there is a separate building block. Inside the COW block is stored all of our (preferentially available) pieces of information pertaining to cows: all the information that gives us the ability to recognise instances and make inferences about cows needed in solving T1 and T2. Each block is a container that contains pieces of stored knowledge pertaining to a single category. Observe that each block now satisfies *both* the judgement and propositional desiderata. One can use the stored knowledge in the blocks to make categorisation, typicality, and inferential judgements. One can also use the atomic blocks as the constituents (smallest units) of propositional thoughts: just line them up so that they compose a sentence.

The fact that our blocks contain pieces of information does not affect their ability to be atoms that compose into larger wholes, but it does mean that the blocks' labels are no longer necessary. As we first considered them, each building block is labelled in a way that distinguishes it from the other blocks. It would be pointless to form a composition with indistinguishable building blocks—how would you know which complex thought was expressed? But if the blocks contain pieces of information pertaining to a single category, this removes the need for labels. The contents can be used to tell the blocks apart. You can look inside a block to see exemplars (along with other pieces of information used for categorical judgement) of the kind of thing that a particular block refers to. Note that when you form a complex of blocks, you are not composing each individual piece of information inside each block (you do not remove and compose every piece of information inside the COW block with every piece of information inside the RUN block when you form COWS RUN). Rather, you compose the entire containers (the discrete blocks), each of which pertains to a category. You may use the contents of the blocks to justify or interpret or illustrate the resulting statements. A container satisfies both Machery's desiderata: a concept contains information that can be used to identify instances of that category and a concept can be used as a constituent in complex propositional thought.

In our example, the containers were imagined to be spatial containers: they contain by having items placed inside a crate. The composition relation was also imagined to be spatial: place the blocks one after another—line them up in a row—to express a complex proposition. But concepts are not spatial containers. And as Fodor (1975) observes, the composition relation in propositional thought is unlikely to be spatial. Brains do not place information inside little crates in the head and

they do not move those crates around to make a propositional thought. Brains use functional properties for both concept containment and concept composition. We have seen a number of proposals for the brain's functional containment relation. The concept composition relation is also unknown. Current thinking is that complex propositional thought involves individual concepts being tokened in working memory or some similar central workspace.[7] The neurocomputational properties involved are unclear.[8] But the fact that there is more work to be done here does not affect our specific point. No matter how one composes those containers—be it arranging them in a spatial row or via some functional relation—those containers can also be used to express a structured proposition in which the containers are constituents. Moreover, the containers do not need to be individuated by labels: they are already individuated by, and have their intentional content fixed by, what is inside them. As discussed above, concepts/containers will only come into being when there is some information to put in them but just how precisely the intentional content is fixed may depend on whether the concept is a fledgling one or a fully-fledged one.

Containers (with or without associated labels) can be the constituents of propositional thought. But some reflection shows that using concepts as containers to express propositional thought would probably not be an efficient way for a cognitive system to operate. Let us return to the example of the building blocks. To form our building blocks into complex wholes requires a space in which to order them. If we want to form a thought with the concept COW in it, we need to transport the COW block or a copy of it (with all of its stored contents, the many pieces of information pertaining to the category COW) into that workspace. Similarly in the case of the brain, to form a propositional thought would require bringing each concept with all its associated contents into working memory or some similar central workspace. One thing we know about working memory or central workspaces in human thought is that it has limited capacity.[9] Copying or transporting an entire concept with all its associated contents would likely be inefficient as an information processing strategy. A more efficient solution would be to token in the workspace *labels of* (or, to borrow a notion from computer science, *pointers to*) the container. One could compose the labels of, or pointers to, concepts as proxies for the real concepts. That would do just as well for the purpose of forming complex propositional thoughts. Note that these labels or pointers stand proxy for *concepts* (containers

[7] Baars (1988); Baars (1997); Carruthers (2014); Carruthers (2015); Dehaene and Changeux (2011); Dehaene and Naccache (2001); D'Esposito and Postle (2015); Fodor (2008); Oberauer and Hein (2012); Penn, Holyoak, and Povinelli (2008); Shanahan and Baars (2005).

[8] Although see Piantadosi, Tenenbaum, and Goodman (2016).

[9] Baddeley (2010); Baars (1997); Cowan (2000); Ma, Husain, and Bays (2014); Miller (1956).

of stored knowledge) not for *categories in the world* as Fodor and other LOT theorists propose.

There is no logical necessity that the constituents of complex propositional thoughts be labels or word-like atoms. Containers can play the role of conceptual atoms in propositional thought. However, containers are bulky: they do not have the desirable features of being easily transportable or easily copyable. For that reason, it is likely that labels of, or pointers to, concepts are composed to form propositional thoughts.[10] The labels or pointers are proxies for the container of stored information pertaining to a single category. As we saw above, the labels or proxies by themselves give no understanding. Understanding comes from what is contained in the container. Mental words or conceptual labels should be seen as cues for accessing those contents. Using labels or proxies, rather than containers packed with information, makes composing concepts into complex constructions, and taking those complexes as the subject matter of further thought, easier.

Concepts as containers satisfy the propositional desideratum. They can either satisfy the desideratum directly by being the entity composed in propositional thought. Or they can satisfy it by having associated labels or pointers which are composed in place of the concepts themselves. Notably, the labels or pointers are not, as Fodor has it, concepts. They are proxies for concepts as containers. It is concepts as containers (which may or may not have associated labels) that have the content-fixing properties and that satisfy the propositional desideratum.

8. *Conclusion*

According to Machery, two distinct and independent types of entity stand behind concept talk in philosophy and psychology. We have argued that this is not the case. A single entity stands behind this talk: *a container of stored knowledge pertaining to a single category*. This entity (which has associated content and may have an associated label) satisfies both the judgement desideratum and the propositional desideratum for concepts. The linguistic divergence between philosophers and psychologists in their use of 'concept' that motivates Machery's view is

[10] There is much empirical evidence that monkeys use pointers in working memory areas to keep track of conceptual information in associative areas in delayed match to sample tasks (Miller et al. 1996; Fuster 1995; Goldman-Rakic 1995 amongst others). In these tasks the subject is shown a number of colours on a screen with one colour being indicated as being the "reward" colour. The colours then disappear from the screen during a delay period. They reappear this time without an indicator of the "reward" colour. The monkey must retain information about which colour to select during the delay period in order to select the right one. The consensus amongst the researchers is that the monkeys use loops between cells in working memory and cells in associative areas (that previously had been found to activate in the presence of e.g. the colour red) to track which colour it needed to indicate after the delay. In other words, the conceptual information was not imported into working memory; instead, cells in working memory pointed to where the information was.

explained by a general property of container talk: container talk can rapidly switch between referring to the *contents* of the container and to the *label* of the container. Despite the pattern of linguistic use that Machery describes, philosophers and psychologists should agree that a concept is just one thing: a container of stored information. They might disagree about specific features of that thing: about the functional containment relation (is "preferential availability" the right relation?) or about the functional composition relation and its implementation (how is composition done in the brain?). But these disagreements are substantive disagreements, not verbal disagreements. Treating concepts as containers untangles Machery's bind. We arrive at a desirable outcome: philosophers and psychologists share a common, rationally explicable, interest in concepts.

Acknowledgements

We are grateful for helpful comments on earlier versions of this paper from Edouard Machery, Jesse Prinz, and Andy Clark.

References

Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. 1997. *In the Theater of Consciousness*. Oxford: Oxford University Press.

Baddeley, A. 2010. "Working Memory." *Current Biology* 20: R136–R140.

Carruthers, P. 2014. "On Central Cognition." *Philosophical Studies* 170: 143–62.

Carruthers, P. 2015. *The Centered Mind*. Oxford: Oxford University Press.

Chalmers, D. J. 2011. "Verbal Disputes." *Philosophical Review* 120: 515–566.

Cowan, N. 2000. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24: 87–185.

Dehaene, S., and J.-P. Changeux. 2011. "Experimental and Theoretical Approaches to Conscious Processing." *Neuron* 70: 200–227.

Dehaene, S., and L. Naccache. 2001. "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition* 79: 1–37.

Dennett, D. C. 1996. *Kinds of Minds*. New York: Basic Books.

D'Esposito, M., and B. R. Postle. 2015. "The Cognitive Neuroscience of Working Memory." *Annual Review of Psychology* 66: 115–142.

Fodor, J. A. 1975. *The Language of Thought*. Sussex: The Harvester Press.

Fodor, J. A. 1998. *Concepts*. Oxford: Blackwell.

Fodor, J. A. 2008. *LOT2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Hume, D. 1976. *The Natural History of Religion and Dialogues Concerning Natural Religion*. Edited by A. W. Colver and J. V. Price. Oxford: Oxford University Press.

Lewis, C. S. 1967. *Studies in Words*. 2nd ed. Cambridge: Cambridge University Press.

Locke, J. 1975. *An Essay Concerning Human Understanding*. Edited by P. H. Nidditch. Oxford: Oxford University Press.

Ma, W. J., Husain, M. and Bays, P. M. 2014. "Changing Concepts of Working Memory." *Nature Reviews Neuroscience* 17: 347–356.

Machery, E. 2009. *Doing Without Concepts*. Oxford: Oxford University Press.

Machery, E. 2010. "The Heterogeneity of Knowledge Representation and the Elimination of Concept." *Behavioral and Brain Sciences* 33: 231–244.

Margolis, E. 1998. "How to Acquire a Concept." *Mind and Language* 13: 347–369.

Margolis, E., and Laurence, S. 2010. "Concepts and Theoretical Unification." *Behavioral and Brain Sciences* 33: 219–220.

Miller, G. A. 1956. "The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63: 81–97.

Murphy, G. L. 2002. *The Big Book of Concepts*. Cambridge: MIT Press.

Oberauer, K., and Hein, L. 2012. "Attention to Information in Working Memory." *Current Directions in Psychological Science* 21: 164–69.

Papineau, D. 2006. "Phenomenal and Perceptual Concepts." In T. Alter and S. Walter (eds.). *Phenomenal Concepts and Phenomenal Knowledge*, Oxford University Press, 111–144.

Pasanek, B. 2015. *Metaphors of Mind*. Baltimore: Johns Hopkins University Press.

Penn, D. C., Holyoak, K. J. and Povinelli, D. J. 2008. "Darwin's Mistake: Explaining the Discontinuity Between Human and Nonhuman Minds." *Behavioral and Brain Sciences* 31: 109–178.

Piantadosi, S. T., Tenenbaum, J. B. and Goodman, N. D. 2016. "The Logical Primitives of Thought: Empirical Foundations for Compositional Cognitive Models." *Psychological Review* 123(4): 392–424

Piccinini, G. 2011. "Two Kinds of Concept: Implicit and Explicit." *Dialogue* 50: 179–193.

Prinz, J. 2004. *Gut Reactions*. Oxford: Oxford University Press.

Prinz, J. 2005. "The Return of Concept Empiricism." In H. Cohen and C. Lefebvre (eds.). *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier, 679–699.

Recanati, F. 2013. *Mental Files*. Oxford: Oxford University Press.

Reid, T. 1983. *Inquiry and Essays*. Edited by R. E. Beanblossom and K. Lehrer. Indianapolis, IN: Hackett.

Shanahan, M., and Baars, B. 2005. "Applying Global Workspace Theory to the Frame Problem." *Cognition* 98: 157–76.

Croatian Journal of Philosophy
Vol. XXIV, No. 72, 2024
https://doi.org/10.52685/cjp.24.72.2
Received: February 4, 2024
Accepted: July 4, 2024

How is Content Externalism Characterized by Vehicle Externalists

DUNJA JUTRONIĆ
University of Split, Split, Croatia

Content externalism and vehicle externalism (what-externalism and how—externalism) or more commonly known as the thesis of extended mind, are said to be two totally independent views that "diverge sharply" (Stanford encyclopedia). There are advocates, adversaries but also agnostics about the extended mind thesis. The approach has been much debated and the controversies about vehicle externalism are importantly manifold. I am not going into any of them. My aim is different and focused on why and how content externalism is characterized by vehicle externalists. Content externalism is labelled by extended mind theorists as: merely causal, taxonomic (Wilson), reactionary (Rowlands), passive (Clark), while vehicle externalism is: constitutive, radical and active. Since content externalists (to my knowledge) have not reacted to a rather negative presentation of their ideas, I restrict myself to showing that many of vehicle externalist (VE) presented views about content externalism (CE) are partly unjustified, not definitive and even wrong. I zoom on the following: 1. CE being 'merely' causal. 2. Active vs. Passive distinction, 3. CE being behaviourally inert.

Keywords: Vehicle externalism; content externalism; causal vs. constitutive; passive vs. active; non-intentional vs. intentional.

1. *Introduction*

In 1998 Andy Clark and David Chalmers published an essay in *Analysis* which started an exciting debate about the nature and study of mind and cognition. Their thesis begins with the question "where does the mind stop and the rest of the world begin?" (Menary 2010: 1) and the claim is that the mind does not stop with the head but spreads into the world. Thus, Clark and Chalmers (1998) in their *extended mind thesis* hold that the mind and the cognitive processes that constitute it extend beyond the boundary of the skin of the individual agent (Menary 2010).

This radical thesis about the mind is usually called *the extended mind thesis* by its proponents and creators (Clark and Chalmers 1998) but it is given a number of other names: *locational externalism, enabling externalism* (Wilson 2000, 2004), Rowlands calls it *environmentalism*, *vehicle externalism* (1999, 2003)*, wide computationalism* and it is named *how-externalism* by Susan Hurley (1998, 2010) and Wilson (2010), *transcranialism* by Adams and Aizawa (2010), sometimes *process externalism* (Keijzer and Schouten 2007).

The other kind of externalism, externalism about mental *content* has been around for a long time. This externalism was the reaction against what Jerry Fodor (following Hilary Putnam) called "methodological solipsism," i.e. against the belief that meanings/contents takes place solely inside the head. Philosophical doubts against "methodological solipsism" or individualism were first raised in the now classical arguments of Hilary Putnam (1975) and Tyler Burge (1979). Content externalism also goes under a number of other names: *semantic externalism, traditional externalism, philosophical externalism, meaning externalism, what-externalism* (Wilson 2000, 2004).

Content (semantic) externalism and vehicle externalism (what-externalism vs. how-externalism) are said to be two totally independent views that "diverge sharply" (Stanford encyclopedia). "We conflate vehicles and contents, as Dennett (1991) and Hurley (1998) stress, at our philosophical and scientific peril" (Clark 2005: fn.1). One is about mental content and the other about vehicles, i.e., about cognitive processes.[1]

There are staunch advocates, but also many adversaries and of course some agnostics about the extended mind thesis.[2] The approach has been much debated and the controversies about vehicle externalism are importantly manifold and often very argumentative and heated. There are also attempts to show that the two externalisms do not "diverge sharply and in a radical way" (Sprevak and Kallestrup 2014; Lyre 2016; Vosgerau 2018). I am not going into any of the above mentioned controversies.

My aim is different and focused on *how content externalism is characterized by vehicle externalists.* I try to show that many of vehicle externalists' (EV) presented views about content externalism (CE) are partly unjustified, not definitive and even wrong. I zoom on the following: 1. content externalism (CE) being "merely" causal. 2. active vs. passive distinction, i.e., distal, historical vs. proximal, "here-and-now." 3. CE being behaviourly inert.

[1] A vehicle need not necessarily be a process. In the previous sentence the inscription 'peril' is a vehicle of the meaning/concept PERIL, but it is not a process. Of course, one might say that the complex *sound* /peril/ is a process. Vehicle is often used to mean a state and/or a process. I look at it as a process.

[2] The literature on radically extended cognition has burgeoned. For a good review, see Shapiro (2011).

Vehicle externalists are keen on stressing the difference between two externalisms.[3] The way that the difference is stressed seems to me to downplay the role/importance of content externalism. Or at least I shall try to show that. The very names/labels and qualification given to semantic externalism by vehicle externalists in their discussion about the differences between the two indicate that they think that content externalism is: *merely causal* (Wilson), *reactionary* (Rowlands), *passive* (Clark). These labels sound rather negative especially in contrasts to the positive qualifications given to vehicle externalism as being: *constitutive, radical* and *active.*

The following is one of overt (and there are many more covert) quotations that point to the *generally* negative view of content externalism.

Wilson and Clark say: "If the extended mind thesis is true, it is true of something implementationally *deep* about cognition, rather than some *debatable view of mental content* […] the extended mind thesis is *not simply a view of how we 'talk about' or view cognition and the mind*—about the epistemology of the mind, one might say—but about what cognition and the mind are—about the ontology of the mind" (Wilson and Clark 2009: 4, italics mine).

When talking about the difference between content externalism and vehicle externalism Mark Rowlands says:

> [W]e might distinguish between what we can call reactionary and radical forms of content externalism. Reactionary content externalism is the view that some propositionally individuated mental properties are externally individuated. Radical content externalism is the view that tokens or instances of some propositionally individuated mental properties are externally located. Reactionary content externalism is a thesis about mental properties and entails rejection of the internalist Possession Claim […] What makes it reactionary is its preservation of at least one core aspect of the Cartesian conception of the mind: the idea that the mental is, ontologically speaking, an internal entity, one located, in one way or another, inside the skins of mental subjects. (Rowlands 2003: 137)[4]

2. *"Merely causal"*

I first take a look at vehicle externalists' claim that content (semantic/meaning) externalism is "merely causal."

One of the big, if not even the most important, issues in the extended mind proposal involves the relation between causality and constitu-

[3] Sprevak says: "HEC has more distant relationship to other kinds of philosophical externalisms such as content externalism […] content externalism says that the representational content of our cognitive states does not supervene on the internal physical state of our brains. HEC has almost nothing to say about this" (2019: 10). However, VE are referring and criticizing CE all along.

[4] The above account of content externalism is basically right but why would the lack of talk about the location of processes make it reactionary? The term "reactionary" is surely offensive and one of many inflammatory rhetoric that VE use about CE.

tion. Adams and Aizawa (2008) and Aizawa (2010) argue that the extended mind hypothesis makes an unjustifiable inference from causal dependence (where bodily and environmental factors play a causal role in support of cognitive processes) to constitutive dependence (where the claim is that such factors actually are part of the cognitive processes). The theory is said to confuse causality with constitution. This is the so-called causal-constitution (C-C) fallacy.

I cannot go into the intricate and much discussed issue whether vehicle externalism is causal or constitutive thesis. Let us, for the present, accept that the vehicle externalism is constitutive thesis as extended mind theorists try to show. What I want to challenge is the extended mind overt and covert statements that content externalism is merely causal or causally weak in supposedly big contrast to vehicle externalism which is constitutive. Here are some chosen passages where it is rather clear that vehicle externalists think that content externalism is "merely causal." Some are more covert and others more overt.

Robert Rupert in distinguishing content and vehicle externalism says:

> Here is a final reason to reject the close association of content externalism and HEC (hypothesis of extended cognition). Recall the sorts of examples externalists typically give in support of their views, examples where content-reference, most clearly is determined by *causal interaction* between the subject and that to which the mental representation in question refers: the subject's 'water' concept refers to H_20 because she has had the right sort of *causal intercourse* with samples of H_2O (Rupert 2024: 401).

Robert Wilson in talking about the chief difference between the two says: "The first (CE) involves the *causal integration* of explicit symbols located in an organism's environment […]" (2010: 181). Richard Menary characterizes content externalism as asymmetric vs. vehicle externalism as symmetric. He says that "vehicle externalism is symmetric form of externalism while content externalism is asymmetric because active externalism (i.e. vehicle externalism) is a constitutive thesis, *it is not a matter of asymmetric causal influence of the environment on internal processes"* (2007: 49, italics mine). The implication surely being that content externalism is just causal and not constitutive.

All the authors mention causal connection, causal integration or causal influence and we know, however, that "causal dependencies are relatively cheap, metaphysically speaking" (Robbins and Aydede 2009: 6).

Vehicle externalists' statements about content externalism most of the time claim that content externalism is a causal thesis and nothing else. It is never mentioned that content externalists claim and show that content itself is not only caused but is *constituted* by certain links to the world. I try to point this in the following discussion.

3. *Causation and constitution*

Daniel Harris (2018) in *Convention, Causation, and Grounding* (on the web) states the difference between causal explanation versus grounding (constitutive) explanation as follows:

1. Roughly speaking, a causal explanation accounts for a phenomenon by spelling out the events that led to it and saying how they brought it about.
2. To give a grounding (constitutive) explanation of a fact is to spell out the more fundamental facts in virtue of which it obtains—i.e. the facts that ground it, that *make it the case* or in *virtue of which* it obtains.

When discussing content externalism, the proponents of VE always mention just Putnam-Burge externalists' claim. In what follows I will, however, help myself with the externalist causal-historical theory of content (or "picture" as Kripke called it), as further importantly developed by Michael Devitt (1981, 2001, 2015), a leading content externalist, to show that externalist theories of content are far from being merely causal.[5]

As early as 1974, in the presentation of the causal theory of proper names in Devitt (1974) the opening sentence reads: "The main problem in giving the semantics of proper names is that of explaining the *nature of the link between name and object in virtue of which the former designates the latter*" (1974: 183, italics mine). In 1981, in his book *Designation* Devitt, said:

> It is important to distinguish our main problem from another. Our problem is to explain *the nature* of the link that certain kinds of words have to the world. The other problem is to explain *how* words *come to be so linked to the world*: what is the historical or causal explanation? *Causal theories of reference are sometimes seen simply as solutions to this other problem.* As such they may seem true enough but trivial. However, they are offered primarily as solutions to the main problem: *they claim that the nature of the link is to be found by looking to the historical explanation.* (Devitt 1981: 8, italics mine)

Here is another relevant passage: "I emphasize that we look to d-chains not merely to discover how a word *came* to designate an object but to discover *the nature of designation.* Understanding designation is understanding groundings, thoughts (of a certain sort), and reference borrowings" (Devitt 1981: 138, italics mine). Obviously, the talk of nature runs right through these passages. To ask about the nature of X is not to ask about the cause of X. It is to ask about the constitution or grounding of X.

[5] Panu Raatikainen says: "Now the critical literature on externalism has a regrettable tendency to focus solely on the earliest statements of semantic externalism and the causal theory of reference, and totally ignore its later developments […] Critics of externalism tend to ignore important improvements" (2020: 80).

Let us look as some concrete examples: What is the meaning of the term 'horse'? The answer is: The meaning of 'horse' is a *property*, the property of referring to horses by a certain causal mode. That's what *constitutes* the meaning. So, horses partly constitute the meaning property. We can then ask: How much of the horse itself goes into the meaning of 'horse'. The answer is: The horse gets into the meaning (so "direct reference" got that right). But more gets in: the mode of referring to the horse. Let's take another example: Dunja has the property of being Croatian. That is the property of being appropriately related to Croatia. So Croatia partly *constitutes* the property Dunja has. How much of Dunja herself goes into the meaning of 'Dunja'. Dunja gets into the meaning (so "direct reference" got that right). But more gets in: the mode of referring to Dunja.

It seems obvious that the above externalist story is far from showing merely causal dependence. On the contrary, the causal story of 'horse', etc. is partly constitutive of its meaning.[6]

Here is another example about the distinction between causation and constitution that often gets blurred: 1. What caused gold is one thing (some dramatic developments in the conditions of the Earth's surface). 2. What constitutes gold is another thing (having atomic number 79). The meaning has to be a property that at least determines that 'gold' refers to gold, i.e. to anything that has the essence/nature of gold. This is the answer to statement 2. The Kripke-Putnam view is that the latter is atomic number 79, and what does the determining is a causal network of reference borrowing back to those that fixed the reference in that essence (more about it in the next section). In sum, the meaning is the property of referring to stuff (gold) with that essence by that causal mode. The point is that the meaning and reference of the name are *constituted* by these causal links.[7] Thus historical-causal theory isn't merely causal: it is the thesis that meanings are constituted by causal link to reality.[8] Descriptions theories of reference are theories of what *constitutively* determines reference (not of what *causally* determines reference). Causal-historical theories (or "pictures", as Kripke would say) are explicitly presented as *rivals* to description theories. So how could they be simply causal? The warning is/was, not to confuse the two theories; how-externalism is different from what-externalism.

[6] Devitt in correspondence: "Right from the beginning in 1970 I had to deal with the objection to the causal theory of names that '*of course*, a name gets its reference at a dubbing,' so what's new?"

[7] "[I]t is not a consequence [of Putnam's slogan] that no aspect of meaning is in the head. The point of the slogan is simply to deny that meanings are entirely in the head. In my view, the meaning of a term is likely to involve many psychological states [...] the slogan emphasizes that extra-cranial links to reality are also necessary to meaning" (Devitt 1990: 83).

[8] Katalin Farkas says: "We already know that meaning is outside the head: so the content of beliefs is also outside the head. Similar considerations will apply to other instances of intentional directedness. Hence some mental features are *constitutively* determined by things outside a thinking subject" (2019: 261).

Nevertheless, in both theories the boundaries of cognition extend beyond the boundaries of individual organisms, beyond the boundary of the skin. Extended mind (interesting or controversial) bold thesis is that their externalism is a *constitutive* thesis as rehearsed by the slogan "cognitive processes ain't (all) in the head," while, they say, content externalism with the slogan "meanings just ain't in the head" is *merely causal* one. If the above discussion is true, that cannot be right since content externalism described above is the thesis that the meaning *properties* of mental states (particularly thoughts) are partly *constituted* by external (causal) relations. So their thesis is not just a causal one.

In sum, we can concede that vehicle externalism is a bolder thesis but it certainly is not bolder because it is constitutive while content externalism is supposedly merely causal or weakly causal. Although the Kripke-Putnam-Devitt thesis is about mental properties and not mental processes, the former is a constitutive thesis, not a merely causal one. "Meanings 'just ain't in the head" means that meanings are partly constituted by the external (horses, Croatia, etc.). Andy Clark's words "cognitive processes ain't in the brain" means that cognitive processes are partly constituted by the external. The main *controversial* part is that processes occurring outside of the brain can be partial constituents of cognitive processes. Whether they are constituents is much discussed and many think that they are not. The issue is undecided so far.[9] Whichever way this interesting thesis turns out, the matter of constitutivity itself is not the main bone of contention between the two (rival) theories.

4. *Active versus passive externalism*

Vehicle externalism also goes under the name of active externalism. Clark and Chalmers (and others) pay great attention to show how the *active externalism* can be distinguished from the more traditional content externalism, familiar from the writings of Putnam (1975) and Burge (1986), which they label *passive* externalism. What I am concerned with in this section is why content externalism is seen and defined by vehicle externalists as passive. I try to point out what is wrong with this characterization.

Here is one of the most important (relevant) quotes from Clark and Chalmers:

> This externalism [radically extended cognition] differs from the standard variety advocated by Putnam (1975) and Burge (1979). When I believe that water is wet, and my twin believes that twin water is wet, the external features responsible for the difference in our beliefs are *distal and historical*, at the other end of *a lengthy causal chain.* Features of the *present are*

[9] Daniel Dennett asked whether the enactive program was really revolutionary or rather a welcome shift in emphasis (1993: 122). He thought it was too soon to answer the question in 1993, and it is not obvious that the matter has been settled since then.

> *not relevant*: if I happen to be surrounded by XYZ right now (maybe I have teleported to Twin Earth), my beliefs still concern standard water, because of my history. In these cases, the *relevant external features are passive.* Because of their distal nature, they play no role in driving the cognitive process *in the here-and-now* […]
>
> In the cases we describe, by contrast, the relevant external features are *active*, playing a crucial role in the *here-and-now.* Because they are coupled with the human organism, they have a direct impact on the organism and on its behavior. In these cases, the relevant parts of the world are in the loop, *not dangling at the other end of a long causal chain*. Concentrating on this sort of coupling leads us to an *active externalism*, as opposed to the *passive externalism* of Putnam and Burge. (Clark and Chalmers 1998: 9, italics mine)

Why is content externalism not active?

1. Because (when I believe that water is wet, and my twin believes that twin water is wet), the external features responsible for the difference in our beliefs are *distal and historical, at the other end of a lengthy causal chain.*
2. Also features of the *present* are not relevant. Because of their *distal nature*, they play no role in driving the cognitive process in the *here-and-now.*

By contrast in vehicle externalism

1. There is no lengthy causal chain. (The relevant parts of the world are in the loop, *not dangling at the other end of a long causal chain*).
2. Vehicle externalism *is active, playing a crucial role in the here-and-now*. "Features of the present are relevant." They "play a role in driving the cognitive process in the here-and-now."

In a nutshell, the claims are that contents of beliefs depend on *my history and that because of that they are distal and thus they do not play an active role in here-and-now.* The two claims about CE—that it is historical and distal and thus not relevant for here-and-now—are related so I shall look at them together.

Whether the above assessments are true/correct depends in large part on the characterization of CE.[10] As stressed before, content externalism is defined and identified only with Putnam and Burge's claims and no other elaborations of the CE are mentioned in the extended mind discussions. However, we should look more closely at content externalism where the theory is elaborated in much more details than what we find in Putnam and Burge. Here again I take the theory of content externalism as developed by Devitt which is a relatively straight-

[10] Let me stress once more that what Clark and Chalmers are after is quite different from Putnam-style semantic externalism: their focus is on the locus of cognitive processes, whereas Putnam, Burge and others are concerned with the external conditions that ground the content of mental or linguistic tokens. However, my concern is not the difference between the two but VE's characterization about content externalism.

forward development of Kripke's (1980) revolutionary idea/picture known as "the causal theory of reference."[11] Devitt's development of content/meaning externalism is within a naturalistic and anti-Cartesian framework.[12] The theory has two parts: a theory of initial fixing of reference, and a theory of reference borrowing. First, a referring expression is typically introduced in a "baptism" or a dubbing event, in the perceptual contact with the referent or a sample of the kind. Second, other language users not present at the name-giving occasion acquire the word from those present at the dubbing, still others from the former, and so on. This is the idea of reference borrowing.

When VE say that all beliefs are historical and thus distal they do not take into account (or ever mention) the first part of the theory, that is, reference/content fixing, i.e., they do not mention grounding.

Let us take the name 'Elvis' for Elvis Presley. In the grounding or reference fixing scenario the name is introduced at a dubbing (formal or informal). The dubbing is in the presence of the object (baby Elvis) that will from then on be the bearer of the name.[13] The grounder (Elvis's mother) has a dispositional property that *caused* a certain thought and *the nature* of that thought is partly explained by its causal connection to the object (baby Elvis). What is crucial for the present discussion is that *it is not the causal history* that grounded the representational or "aboutness" relation to Elvis, or Elvis's name. It was the present Elvis's mother thought that played a role in direct causal connection to the object (baby Elvis). More generally, the grounders of the term 'F' are the people who fix the reference of 'F' that others then borrow. So a key thing for the reference of 'F' is what as a matter of fact goes on in the groundings by those people, whatever anyone's opinion about Fs is. This is Kripke important "ignorance and error" claim/discovery. There is no "lengthy causal chain." There is nothing *historical or distal* in the grounding scenario and nothing "dangling at the other end of a long causal chain." It is not the causal history that plays a role in this interaction. The represented entity (baby Elvis) in the environment is represented precisely because it (he) has a direct impact on the cognizing organism (Elvis's mother) and its (her) behavior. There is nothing distal and historical about such scenario.[14]

[11] Other names are "the historical theory of reference," "the causal-historical theory of reference," or simply "the new theory of reference." See Raatikainen (2020).

[12] Devitt says: "This is not to say, of course, that the theory is complete. I have emphasized that any theory of reference at this time must look to future psycholinguistics for more details. And it is not to say either that the details already provided are certainly right. The point is simply that we have good reason now to think that this theory is more or less right, so far as it goes, and it goes as far as it is reasonable to expect at this time. And we can see that such adjustments as may be necessary will not be large and will be in terms of the same reality of designating or denoting-chains" (2015: 128).

[13] Devitt and Sterelny (1999: 67). The example with Elvis is mine.

[14] If you consider demonstratives rather than names, then the causal link to the referent is typically immediate.

One may wonder: Why would content externalism be passive when the grounder is in the direct contact with the thing grounded in the dubbing scenario? The situation cannot be more direct than it is. And it does not seem to be passive. Why not? In the naming ceremony, the entity (baby Elvis) that individuates the contents of mental states (his mom's), has an impact on the organism (his mom) and there does not seem to be any legitimate reason not to say that the relation is active and that it is "driving the cognitive process in the here-and-now." Both sides, the grounder and the object grounded, are influencing one another. There does not seem to be any passivity in the naming ceremony and its completion. On the contrary the entity in the world (baby Elvis) plays an active role in cognition because the result of the interaction is that the environment (baby Elvis) partly constitutes grounder's (his mom's) cognitive states.

Hajo Greif in defining the active externalism says: "The activity of interest is in the environment and the organism at the same instance, and it is that *concurrent activity* which serves to make both cognition extended and externalism active" (Greif 2017: 4313). Content externalist can say that this is exactly what we have in the grounding scenario. The grounder and the thing grounded are in "concurrent relation," and the interaction is active. In this relation there are without doubt relational processes that make the interaction dynamic and thus makes referring an activity in which the relational bond (between Elvis's mom and baby Elvis) is established (more on this issue follows). So much for grounding.

When vehicle externalists say that CE is distal and historical, they are obviously referring only to reference borrowing (indirectly, since they do not mention it under this name) which supposedly make content externalist's approach *historical and distal and thus* (*consequently*) *passive.*

What is happening in reference borrowing? Language users who were not at the grounding gain the ability to refer with the expression in virtue of an appropriate causal-historical chain going back to the introduction of the expression. The borrower may borrow its reference from that of others, whether she knows anything about this borrowing or not,[15] and she can be totally ignorant about this chain or the referent. Nevertheless, she can successfully refer with the expression. Going back to our example. My (or present) term 'Elvis' is about Elvis Presley in virtue of a designating-chain going back to him involving people participating in the convention of designating him by 'Elvis.' That *underlying d-chain* is a causal relation that *constitutes* the content of 'Elvis' (see part one of this paper) and it is true that the d-chain is historical. It can go a bit into the past or centuries into the past. Because of this the content ELVIS, or the term 'Elvis' then, according to vehicle exter-

[15] She presumably knows about her own borrowing at the time of her borrowing but that is a minor point.

nalists, has a "passive" role since it is "removed from the cognitive processes of the individual (it is distal)" through a long chain. VE are thus denying that represented entities could play an active role in cognition just because they are distal (i.e. located at a distance to cognitive activity) and thus their impact on cognition, supposedly cannot be direct.

The question is: Does the fact that there is a long d-chain going back to Elvis make the process passive in the way that vehicle externalists assume? I think it does not. The reference/content borrower is connected to the thing through a historical chain. That is distal and historical for sure. But why is this passive when the content of the belief is *constituted* by the d-chains? It is in *virtue of* that content that the belief plays its role in cognition and also in causing behavior here and now. In other words, how could the fact that represented entities are historical and distal (i.e. located at a distance to cognitive activity), be the ground for denying that represented entities play an active role in cognition? There is no reason to assume the past to be in any normative sense irrelevant. Hajo Greif, in presenting kind of defense of CE, says: "On the other hand it is this *history of interactions* that explains any possible difference between the contents of two *prima facie* identical mental or linguistic tokens, no matter what the current interactions may look like to participants and observers" (Greif 2017: 4313). The implication being, I think, that history of interactions is even more important that the current interaction (here-and-now). The reference borrowing with its d-chains is just such a scenario. The first interaction, namely the grounding and then the past interactions had been relevant to shaping (constituting) the content of some present linguistic or mental token, and there is no reason to assume the past is "remote" from the present.

However, VE argue that because these entities are distal and historical, their impact on cognition cannot be direct. This is surely misleading. The constancy and past-endorsement criteria show that the causal history is constitutive of belief. So, the fact that contents are distal and historical does not matter since the representer/speaker is *ipso facto* appropriately receptive here-and-now by the constitutive features of the content. *The historically determined content plays a role here-and-now. It is in virtue of that content that the belief plays its role in cognition and in causing behavior.* Vehicle externalists seem to be committed to thinking that historically represented environmental entities—those entities that individuate the representational contents of mental states as content externalism suggests—are not represented in virtue of "driving the cognitive process in the here-and-now." Take Burge's example about the arthritis in my hip. My belief should be established (or is grounded) in an existing active relationship with the doctor and then it would presumably "drive the cognitive process in the here-and-now," but my causal-historical relationship to a language community (reference borrowing) would not.

In sum, the statement that the relevant external features "because of their distal nature, do not play a role in driving the cognitive process

in the here-and-now […]"overlooks (1) the fact that the content of my belief plays a causal role here-and-now even though it is partly constituted by historical causal links, and moreover it overlooks the fact (2) that terms can be, and typically are, *multiply* grounded in their referent. As a result, words can change their reference; 'Madagascar' used to refer to a part of the African mainland but now refers to an island.

5. *Possible objections*

One might say that the causal theory being relational is therefore static. The term 'Elvis' (in our example) has the content in virtue of standing in the relation to a famous American rock star. But relational does not necessarily mean static and meaning constitution as primarily the expression of thought surely includes some process. The content states are formed in interaction between the environment and some inner processes going on in the grounder. The result of the interaction is the belief that ends up being in person's mind or it can be outside the mind because produced in speech and writing. If one still insists that the relationship is static this is not surely the same as passive.

However, vehicle externalists argue that more is included in the characterization of active externalism. It is insisted that external features are "in the loop," where this indicates more than "merely playing a crucial role in the here-and-now." It is a "two-way interaction" between the human organism and external entities by which externalism is distinctively "active" and supposed to be part of what it is to be "in the loop" (Greif 2017: 4313). Whether external features being in the loop is more appropriate characterization of active externalism is a question that is not a part of this discussion. There are convincing arguments given by Sprevak and Kallestrup in showing that "many external resources […] do not satisfy ACTIVE's conditions" (2014: 87). They conclude that "it is rather misleading to say that what distinguishes radically extended cognition from Putnam-Burge anti-individualism is that the former is distinctively active and the latter is passive" (2014: 83–84). One cannot but agree. But that is not the concern here. What was important to point out (and hopefully show) is that VE's claim that CE is passive is at least questionable or maybe downright wrong. Whichever way it turns out for the active externalism to be, it is still simply false to say that content externalism plays a passive role in cognition. The fact that (mental) representation is importantly relational does not show that it is passive.

Furthermore, content externalists, although not primarily interested in cognitive processes, are not immune to this particular issue. The question arises in the so-called *qua* problem, the name coined by Kim Sterelny. Continuing with our example, the question is: why 'Elvis' refers to the whole individual and not to his face or his lips. By virtue of what is the grounding term grounded in the object *qua*-Elvis and not in some of his parts. There have been a number of attempts to solve the

qua problem.[16] The most recent statement is found in Devitt who says: "I have struggled mightily with this problem (1981a: 61–4; Devitt and Sterelny 1999: 79–80), but I now wonder whether this was a mistake: perhaps the problem is more for psychology than philosophy" (2002: 115, footnote 15). Why it is more a problem for psychology than philosophy? Because it is concerned with mental processes of the grounder. In virtue of what has the grounder grounded the term 'Elvis' and not Elvis's lips? In order to find the answer one has to go beyond looking at the mental processes of the grounder to the mechanisms/processes of perceptual experiences which will tell us if applied to the whole object and not just parts of it. In order to complete the causal theory of content Devitt and Sterelny's suggestion is to add the teleological elements to the causal story, to appeal to the biological function in the explanation of the mechanisms/processes of referential relation. Recently their suggestion seems more plausible with the fine elaboration of the teleosemantic explanation of the preconceptual/nonconceptual level of sensory perceptual representations found in Neander (2017). Needless to say, we cannot go into any details of such suggestion or claim that it is true. The main point here is that content externalism is not immune to the problems (and possible solutions) to the workings and structure of cognitive processes. Giving a detailed account of the actual mechanisms might not be, pace Devitt, a philosopher's task but the concern again points to the fact that CE worry about processes which VE do not mention at all.

6. *Content externalism is behaviorally inert (irrelevant)*

The third point to look into is vehicle externalists' claim that content externalism is behaviorally inert. It is inert because it does not affect the results of behavior and it does not generate action. Here are two quotes:

> Many have complained that even if Putnam and Burge are right about the externality of content, it is not clear that these external aspects play a causal or explanatory role in *the generation of action*. In counterfactual cases where internal structure is held constant but these external features are changed, *behavior looks just the same*; so internal structure seems to be doing the crucial work (Clark and Chalmers 1998, in Menary 2010: 29, italics mine)
>
> […] the relevant external features are *passive*. Because of their distal nature, they play no role in driving the cognitive process in the here-and-now […] This is reflected by the fact that the *actions performed by me and my twin are physically indistinguishable, despite our external differences*. (Clark and Chalmers 1998, in Menary 2010: 29, italics mine)

In sum the claim is that content externalism is not action-guiding, it does not explain behavior because external changes do not cause internal changes. External differences leave the Twins physically indistin-

[16] For a rather comprehensive review see Jutronić (2019: 449–477).

guishable, their behavior is physically the same. Thus, external component is behaviorally irrelevant and inert.[17]

What kind of action or behaviour vehicle externalists have in mind? "V(ehicle)E(xternalism) requires that the external resource guide the agent's action in the here and now. The relevant sense of 'action' is non-intentional; 'action' means something like *bodily movement*" (Clark and Chalmers 1998: 8–9)."[18] Thus Otto walking to 53rd street and Otto walking to 51st street involve different bodily movement, different neural activity which explains the difference in their non-intentional walking behavior. But if the relevant sense of 'action' *is non-intentional*, i.e. 'action' meaning something like *bodily movement* then vehicle externalists' claim that content externalism cannot explain behavior is misplaced since what Twin-Earth example tries to explain is not non-intentional action in the guise of some bodily movement or neural activity but it tries to explain *intentional behavior*. If that is the case then how can we explain different intentional behaviors of two Otto's going to two different streets? Otto's and Twin Otto's intentional behavior cannot be explained with the difference in the neural activity. When VE say that CE is behaviouraly inert because it does not affect the results of behavior and it does not generate action they are talking at cross purposes. Content externalists are concerned to explain intentional behaviour and not some bodily movement. Moreover, vehicle externalists' opinion about the generation of action is based on Narrow dogma. i.e. the belief that only narrow meanings are needed for the scientifically proper explanation of behavior.[19] What narrow psychologists have in mind is fairly brute-physical, neural impulses or mere bodily movements to explain behaviour. Tyler Burge (1986) has shown that narrow dogma is wrong and that "many relevant specifications of behavior in psychology are intentional, or relational, or both" (Burge 1986: 11). There is nothing to show how narrow meanings of a sentence—as a functional role involving other sentences, proximal sensory input, and proximal behavioral output—might explain intentional behaviors. Devitt brings out the crux of the problem: "In brief, Narrow psychology committed to functional-role meanings faces a dilemma. Either it claims that psychology should explain only brute-physical behaviors, or it accepts that psychology explains intentional behaviors. If the former, then Narrow psychology is committed, implausibly, to denying intentional behaviors. If the latter, then Narrow psychology is committed, implausibly, to narrow meanings explaining intentional behaviors" (Devitt 1990: 298). Consequently, the intentionally described

[17] Lyre says: "Clark and Chalmers endorse Putnam's and Burge's externalism as a thesis about content individuation, although they find it insufficient to account for all aspects of cognitions (in particular, the current causal contribution made by the environment), and therefore ultimately reject it" (2015: 2).

[18] In Sprevak and Kallestrup (2014: 83–84).

[19] On the terminology "narrow" and "wide," see Putnam (1975: 220–2). Also Devitt (1990, 2001).

behavior of Otto who walks to 53rd Street and Twin Otto who walks to 51st cannot be explained with their neural differences or their bodily movements. The intentionally described behavior of Otto is not the same as that of his Twin because it involves 53rd street, not 51st street.

7. *Conclusion*

I tried to show that CE is not merely causal, that it is active and behaviourly relevant.

Content externalism entails that (1) some entities, that are biologically external to an organism, are theoretically important for understanding organism's cognitive psychology and that (2) these entities play an active cognitive role in having a direct impact on the cognizing organism and its intentional (not non-intentional) behavior. Content externalism is neither merely causal, or simply passive and behaviorally inert. If true, then the content externalism's arguments showing that meanings of our words "aren't in the head" is not a totaly independent view that "diverge sharply" and is in opposition to the arguments of the extended mind claim that cognitive processes just "ain't in the head." In other words, the externalism about content carries over into the externalism about mind. However, that does not show that the vehicle externalism is true.

In her article "Modelling the Mind" K.V. Wilkes said: "A danger as far as psychology is concerned, comes when we switch from indefinite to definite article, when we stop talking of 'a' model, metaphor [...] and sneak in the term 'the'" (1990: 63–64). Her suggestion and belief was: "Let a hundred models bloom" (1990: 82).

References

Burge, T. 1986. "Individualism and Psychology." *The Philosophical Review* 14 (1): 777–780.

Clark A. and Chalmers, D. 2010. "The extended mind." In Menary, R. (ed.). *The Extended Mind*. Cambridge: MIT Press, 27–43. /Originally in *Analysis* in 1998/.

Devitt, M. 1974. "Singular Terms." *Journal of Philosophy* 71 (7): 183–205.

Devitt, M. 1981. *Designation*. New York: Columbia University Press.

Devitt, M. 1996. *Coming to our senses: A naturalistic program for semantic localism*. Cambridge: Cambridge University Press.

Devitt, M. 2001. "A shocking idea about meaning." *Revue Internationale de Philosophie* 208: 449–472.

Devitt, M. 2015. "Should proper names still seem so problematic?" In A. Bianchi (ed.). *On reference*. Oxford: Oxford University Press, 108–143.

Devitt, M. 2020. "Stirring the Possum: Responses to the Bianchi Papers." In A. Bianchi (ed.). *Language and Reality from a Naturalistic Perspective Themes from Michael Devitt*. Springer, 371–457.

Devitt, M. and Sterelny, K. 1999. *Language and reality: An introduction to the philosophy of language* 2nd ed. Cambridge, MA: MIT Press. 1st edition 1987.

Farkas, K. 2019. "The Boundaries of the Mind." In A. Kind (ed.). *Philosophy of Mind in the Twentieth and Twenty-First Centuries. The History and the Philosophy of Mind* Vol. 6. Routledge, 256–279.

Greif, H. 2017. "What is the extension of the extended mind?" *Synthese* 194: 4311–4336.

Harris, D. 2018. *Convention, Causation, and Grounding*. URL: https://danielwharris.com/papers/DanielWHarris-ConventionCausationGrounding.pdf

Jutronić, D. 2019. "The *Qua* Problem and the Proposed Solutions." *Croatian Journal of Philosophy* 19 (57): 449–477.

Lyre, H. 2016. "Active Content Externalism." *Review of Philosophy and Psychology* 7: 17–33.

Lyre, H. 2018. "Socially Extended Cognition and Shared Intentionality." *Frontiers in Psychology* 9: 831.

Menary, R. 2007. *Cognitive Integration Mind and Cognition Unbounded*. New York: Palgrave Macmillan.

Menary, R. (ed.). 2010. *The Extended Mind*. Cambridge: MIT Press.

Raatikainen, P. 2020. "Theories of reference. What was the question?" In A. Bianchi (ed.). *Language and Reality from a Naturalistic Perspective Themes from Michael Devitt.* Springer, 69–105.

Robbins, P. and Murat, A. (eds.). 2009. "A Short Primer on Situated Cognition." *The Cambridge Handbook of Situated Cognition*, 3–11.

Rowlands, M. 2003. *Externalism, Putting Mind and World back together again*. Chesham: Acumen.

Rupert R. D. 2004. "Challenges to the Hypothesis of Extended Cognition." *The Journal of Philosophy* 101 (8): 389–428.

Shapiro, L. 2011. *Embodied Cognition.* Routledge/Taylor & Francis Group.

Sprevak, M. 2019. "Extended cognition" In T. Crane (ed.). *The Routledge Encyclopedia of Philosophy Online*. London: Routledge.

Sprevak, M. and Kallestrup, J. 2014. "Entangled Externalism." In M. Sprevak and J. Kallestrup (eds.). *New Waves in Philosophy of Mind*. Palgrave: Macmillan, 74–98.

Vosgerau, G. 2018. "Vehicles, Contents and Supervenience." *Philosophy and Society* 29 (4): 471–646.

Wilkes, K. 1990. "Modelling the Mind." In K. A. Mohyeldin Said, W. H. Newton-Smith, R. Viale and K. V. Wilkes (eds.) *Modelling the Mind.* Oxford: Clarendon Press, 63–82.

Wilson, R. A. 2004. *Boundaries of the Mind: The Individual in the Fragile Sciences: Cognition*. New York: Cambridge University Press.

Wilson R. A. and Clark, A. 2009. "How to situate cognition: Letting nature take its course." In M. Aydede and P. Robbins (eds.). *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press, 55–77.

Wilson, R. A. 2010. "Meaning Making and the Mind of the Externalist." In R. Menary (ed.). *The Extended Mind*. Cambridge: Bradford Book, 167–189.

Croatian Journal of Philosophy
Vol. XXIV, No. 72, 2024
https://doi.org/10.52685/cjp.24.72.3
Received: May 6, 2024
Accepted: July 4, 2024

Intentions and Representations

SHAUN GALLAGHER
University of Memphis, Memphis, USA

Kathy Wilkes's essays on explanations and representations, and especially her interaction with Daniel Dennett, raise questions about whether some notion of representation can explain action intention. Wilkes is not sure whether subpersonal representations are real, but she thinks that the most pragmatic strategy is to take the intentional stance and accept the usefulness of personal level intentions, even if we have to worry that this does not give us a scientific explanation. Wilkes's skepticism about subpersonal representations, and even about the appropriateness of the notion of subpersonal levels of explanation, seems to fit with more recent embodied-enactive approaches to cognition. Considerations about the nature of cognitive mechanisms and animal intelligence prevent her from moving in that direction, however. These insights suggest that Wilkes' analysis continues to be directly relevant to contemporary discussions.

Keywords: Action intention; representation; subpersonal levels of explanation; intentional stance; enactivism.

1. *Introduction*

My aim in this paper is to look at some things that are missing from Kathy Wilkes' essay "Representations and explanations" (1989a) and to see if by bringing those missing items into the discussion it could clarify some of the problems she is considering, and also give us some idea of how Wilkes would fit into some current debates about representation. Some of the missing things are missing simply because Wilkes ignored them; other things were not yet available when she wrote her essay. I'll also make reference to a second essay she published in the same volume, "Explanations—How not to miss the point" (1989b). Both of these essays are part of a volume, *Goals, No-goals and Own Goals. A Debate on Goal-directed and Intentional Behaviour,* based on a set of seminars that took place in Oxford in the 1970s and 1980s, as Monte-

fiore and Noble indicate in their editorial Introduction to the volume. Due to the nature of the volume, she makes reference to other essays in the volume (by Dennett, Montefiore, McFarland, and Noble), and she is explicitly in dialogue with Dennett on some specific points. It's notable, I think, that neither Wilkes nor any other contributor to this volume, which explores issues having to do with intention, makes mention of well-known work by Elizabeth Anscombe (1957) or John Searle (1983) on these issues—even to disagree with them.

2. *Some stage setting*

Wilkes, in her essay "Representations and explanations" (1989a), is concerned with the notion of intention, in the sense of having or forming an action intention, understood as a representation of a goal to be attained. Her question is whether one needs these concepts (intention, representation) to explain behavior. Her answer ultimately is yes with some important qualifications. To work out this answer she explores the notion of explanation and along the way discusses, and endorses, Dennett's intentional stance (without naming it as such).

One issue that we might be tempted to set aside, not only because it seems to be a terminological issue, but because Wilkes herself sets it aside, concerns the distinction between intentionality (with a 't') and intensionality (with an 's'). For her, intentionality with a 't' signifies "the forming or having of intentions, representations of goals to be achieved" (Wilkes 1989a: 159), and this is taken to be a sub-category of intensionality (with an 's'). Wilkes (1989a) equates intensionality (with an 's') with what most would consider Brentano's concept of the aboutness or object-directedness of mental states, which is usually spelled 'intentionality' (with a 't'). Wilkes, then, does not use the standard or orthodox understanding of this terminological distinction.[1] According to Searle, for example, this is something of a mortal sin:

> One of the most pervasive confusions in contemporary philosophy is the mistaken belief that there is some close connection, perhaps even an identity, between intensionality-with-an-s and Intentionality-with-a-t. Nothing could be further from the truth. They are not even remotely similar. Intentionality-with-a-t is that property of the mind (brain) by which it is able to represent other things; intensionality-with-an-s is the failure of certain sentences, statements, etc., to satisfy certain logical tests for extensionality. The only connection between them is that some sentences about Intentionality-with-a-t are intensional-with-an-s […]. (Searle 1983: 24)

Whether we accept Searle's characterization or not, it does seem that in contemporary debates about representation, the distinction holds some significance; in some broad sense it relates to an issue discussed by Wilkes, about whether having a representation with semantic con-

[1] Despite the fact that she cites Chisholm (1957), the likely source for the orthodox view. "Any reader who wants a short and clear description of what 'intensionality' is should consult chapter 11 of [Chisholm 1957)]" (Wilkes 1989a: 182).

tent depends on having linguistic ability. We'll return to that issue. For now I'll adopt the more standard distinctions between action-intention, mental intentionality (aboutness), and language-based semantic intensionality.

At this point, however, Wilkes is more concerned to distinguish action-intention, from intentionality in the broad sense of aboutness. Action-intention is a special form of the more general concept, of course, since action-intention also has the character of being about something or being directed towards something (e.g., a goal).

3. *Levels of explanation*

One issue addressed by Wilkes concerns causal explanation, and whether that would be the right kind of explanation for explaining action intentions, or for explaining behavior using intentions as part of the *explanans*. In this respect, there is also some question about levels of explanation. An action intention seems to be a personal-level phenomenon; but a representation is sometimes considered a sub-personal thing. Consider that neuroscientists keep telling us that they can identify what someone is thinking, or perceiving, or intending to do by looking in that person's brain using e.g., fMRI (e.g., Coles 1989; Cox and Savoy 2003; Frith and Gallagher 2002; Haynes et al. 2007). Chris Frith, for example, thinks that an intention is part of the neural mechanism involved in motor control for intentional action, specifically a representational part of the comparator process which, to keep an action on track checks the match between intention and efferent copy (generating a sense of agency) (Frith 1992). One can find similar ideas in discussions of predictive processing (Friston and Frith 2015; Gallagher and Allen 2018). As I understand her, Wilkes wants to rule out such explanations, i.e., any explanation that would treat intentions as subpersonal representations with causal power, and she wants to limit the notion of intention to the personal level.

With respect to the issue of explanation, Wilkes cites an example from Hilary Putnam (1980): why the square peg won't fit into the round hole cannot be explained by a molecular-level explanation. That would be the wrong sort of explanation. Just as molecular processes do not cause the mismatch between square peg and round hole, neural processes do not cause the intention—"*even if* intentions, or representations, can ultimately be described as 'no more than' sets of processes in amongst nerve cells" (Wilkes 1989a: 169). This would be a statement of composition rather than causality. Neural processes do not cause intentions, and do not causally interact with intentions; even if they constitute intentions (on some kind of identity theory).

I think we get closer to her reasoning by considering her own example: a neural explanation won't explain why Flora flounced out of the party—one needs to explain that behavior in terms of intentions, beliefs, desires, etc. in what Sellars would call the space of reasons (folk

psychology) rather than the space of causes. No one in this collection of essays mentions Sellars either. No need to, since we have Dennett and the intentional stance with its distinction from design and physical stances (Dennett 1989: 237). These stances reference different levels of explanation: the personal- or intentional level versus the subpersonal, distinguished into functional (design) and physical levels. Things get complicated, however, since explanatory levels can be defined in different ways (Wilkes 1989b: 198–199); mereological/constitutional *versus* functional/causal for example.

This kind of complication is discussed in more recent philosophy of science. Phillip Gerrans itemizes a number of level types.

> The notion of levels is ubiquitous [in scientific explanation], but not everyone uses it in the same way. It can refer to ordering relationships between theories…; the objects of theories ordered by size or complexity, e.g., cells are smaller and less complex than the organs they make up; functional analyses, e.g., vision is a higher-level property than edge detection; or levels of mereological containment, e.g., parts are at a lower level than wholes. (Gerrans 2014: 229–230)

Kenneth Schaffner (2020: 384) provides a more exhaustive list of level types: levels of "abstraction, analysis, aggregation, behavior, complexity, function, perspective, organization, generality, and processes—as well as causation and control—as well as description and explanation, and more."

Given this complex multiplicity of levels, James Woodward (2020: 428) expresses an attraction to "levels eliminativism […] the thought that we would be better off avoiding level talk entirely." He nonetheless introduces what he calls the 'interaction level'. He takes the interaction level to include any factor, regardless of size or composition, that has a causal relation to the system that needs to be explained. Such factors are testable by his notion of interventionist causality. This puts neural processes, psychological processes, social processes, etc. all on the same level, so that the explanation doesn't have to reference any other level (defined by different criteria). In this sense, if one's explanation is confined to one level, so defined, then one doesn't have to talk about levels at all.

This approach might have appealed to Wilkes since she had a skeptical view of levels and a quite pluralistic or "tolerant" (1989b: 209) view of causes ("We describe things as causes when they interest us, when they seem important to us, when we can juggle and manipulate them" [1989b: 205]).

> In psychology and neuroscience …we have practically no idea what, and where, the relevant 'levels' between (at one extreme) the macro-states postulated by the behavioural sciences, and (at the other extreme) the individual synaptic connections described by neurophysiology, are. We lack an agreed neuropsychological taxonomy 'in the middle'; and, as noted already, psychology at the 'macro' level still has little consensus about its taxonomy of explananda. The top 'level' is very loosely characterised as yet; and the levels beneath that are still largely matters of mystery. Thus we do not

> know whether 'intentions', or 'goal representations', have a suitably systematic relation to anything on the next level down (whatever that might be). (Wilkes 1989a: 174)

I think this is still the case, but also I think that Wilkes gets tripped up about levels by her examples, which are examples of physical level mereological relations rather than functional-causal relations. That's clear in her appeal to Putnam's example of the square peg; and also her example of the ornament. An intention is like an ornament, but "the way in which atoms and molecules constitute ornaments is not something amenable to scientific investigation; and for that very reason 'ornament' is not an explanandum for physics" (Wilkes 1989a: 171)—at least in regard to its cultural significance. If the complications about levels of explanations lead her to endorse the intentional stance, this is not the only reason. Another reason is her uncertainty about the reality or status of representations. In this regard Wilkes argues that "it is vastly unclear what it means to say that 'there are such things as' intentions, or goal representations. Yet if they are to be worth citing as 'causes' in the explanation of behaviour, then, evidently, they must exist. If they have a role in explanation, but not as causes of behaviour, then the matter is less clear" (Wilkes 1989a: 170).

4. *Are representations real?*

Here, Wilkes engages with Dennett and the idea that representations must be "physically structured objects" that play a causal role in cognition. Quoting Dennett: "... information is represented explicitly in a system if and only if there actually exists in the functionally relevant system a physically structured object ..." (Dennett 1982-3: 216; Wilkes 1989a: 161). Specifically, for Dennett, a representation is a physically structured object plus some kind of interpretation or interpretive mechanism. The two together realize a representation. But, Wilkes is hesitant: "what it means to say that representations or intentions exist—is a highly vexed business."

One interpretation of the Dennettian view is that a representation is a kind of subpersonal entity, explicable from a design stance. This would be distinguished from whatever might (or might not) be on the personal level, grasped via the intentional stance. The scientific explanation is focused on the design level. Wilkes, however, is drawn to the personal level, where an intention is equated with what she calls an 'explicit' representation. For her there is something like a "sliding scale" that descends from an explicit representation (intention) to lower-level operations. Here Anscombe is not mentioned, but an Anscombian analysis seems to be implied:[2]

[2] I have in mind Anscombe's example of the man working a pump. What counts as the action can be described in many ways, including just the physical use of muscles to pump the water. But the circumstances will say what the most appropriate description is. For example, if the water is poisoned and the occupants

> For instance, if one intends some end—killing Lincoln, say—then in a sense one intends the various means to that end; one may be said to intend whatever the guiding intention implies. Thus, perhaps parasitically, Wilkes Booth ('also'?) intended to shoot Lincoln; to fire a gun; to pull the trigger; to crook his right index finger. We are, I think, much less certain about whether these are 'explicit' or not. We might call them 'implicit', in that they are implied by something that is (perhaps) explicit. But we have no clear intuitions about whether, or when, they are 'worth individuating', or what it means to say they 'exist'. The problems of individuating intentions are, unsurprisingly, exactly the same as those of individuating actions […] evidently [we] have a hefty theoretical problem in the specification of 'the' intention, or representation, that guides, governs, explains, modifies, or perhaps causes, purposive behaviour. (Wilkes 1989a: 162)

It's not clear that in this listing of implicit intentions/representations Wilkes (Cathy, not Booth) goes far enough down to get to anything subpersonal. Nothing here resembles a physically structured object of the sort that would count as a subpersonal representation.

Even when she considers what she calls 'tacit representations' they are not necessarily subpersonal, although subpersonal processes are clearly involved—"Tacit representations seem to be dispositions, abilities, know-how: where and what we can do depends upon the way we are, or—sometimes—on what we have learned" (Wilkes 1989a: 163). It's not necessarily the case that individuating actions or intentions remains theoretically problematic once one introduces some of these Anscombian descriptions, but they lead directly to pragmatic considerations about circumstances, embodied degrees of freedom and ecological affordances (Gallagher 2020).

It's still an open question (at least for some) whether one can lower the analysis into the subpersonal scale, to find Dennett's 'physically structured objects.' In the contemporary discussion (unavailable to Wilkes) such objects are called 'structural representations.' Structural representations are described precisely as mechanisms on the subpersonal, neural level. Gualtiero Piccinini (2022: 6), for example, suggests a way to think of such representations as physically structured objects. For him, representational content is just the information contained in the occurrent physical structure of neurons or neuronal networks (which can be understood following Gabor [1946] and Miłkowski [2023] as Shannon information instantiated in the quantifiable independent degrees of freedom of such physical entities):

of the house die, then pumping the water could be a case of murder, depending upon the agent's knowledge and intention. "[A] single action can have many different descriptions […]. Are we to say that the man who (intentionally) moves his arm, operates the pump, replenishes the water supply, poisons the inhabitants, is performing four actions? Or only one? […]" (Anscombe 1957: 11). In short, the only distinct action of his that is in question is this one, A. For moving his arm up and down with his fingers round the pump handle is, in these circumstances, operating the pump; and, in these circumstances, it is replenishing the house water-supply; and, in these circumstances, it is poisoning the household. So there is one action with four descriptions" (Anscombe 1957: 45–46).

> A specific content may be distributed over a relatively large ensemble of neurons. Yet content is relatively localized in the sense that it is carried by a specific vehicle born by a specific bearer (neuron/ensemble/circuit) and not diffused through the whole neurocognitive system, or even a large part thereof. (Piccinini 2022: 6)

If one stays with the concept of Shannon information, then this means that content just is the physical configuration that defines the neuron's function in the neuronal ensemble, the degrees of freedom of a neuronal network, etc. If this does not solve all problems, it nonetheless is a good candidate for Dennett's physically structured object. Questions still remain whether this isn't just a neural structure that covaries with environmental stimuli, and in that sense why we should consider it a system-relative representation rather than an observer-relative interpretation, or deflationary gloss (Egan 2014). Moreover, if the physical pattern is indirectly, yet still physically, coupled to the environment or object correlated with that co-varying pattern, it can be explained in terms of dynamical causality rather than anything resembling good-old-fashioned semantic content. Although Piccinini thinks such a neuronal structure can be decoupled from its target, he also contends that we don't even get this far without the system being embodied, embedded, enactive and affective. That's where the structure comes from. In which case one can ask about the embodied origins of so-called 'non-derived' original content. At the very least, these are questions that we can raise about why we would want to call this a representation in the first place.

A couple of years after the publication of Wilkes' essay, Dennett published his essay "Real patterns" (1991) which suggests an answer to the question, are representations (of this subpersonal type) real? (This is another thing that was not available to Wilkes, although I wonder whether the dialogue here between Dennett and Wilkes didn't directly motivate Dennett's thinking about patterns). Whereas Wilkes has to say she just doesn't know, Dennett would contend that representations are real in a scientific pragmatic sense. That is, they are real enough if science finds them useful components of an explanation—if they serve some pragmatic purpose in empirical explanation. This is not an ontologically heavy conception of reality; it seems to go along with the notion of content that is not of the heavy semantic intensionality (with an s) type, as well as with the notion of intention in the intentional stance, which seemingly does not come along with ontological commitments. Wilkes embraces this latter kind of explanation—the intentional stance—which is just what allows her to remain uncommitted about a subpersonal explanation (Wilkes 1989b: 195):

> [I]t is almost entirely irrelevant what (if any) neurophysiological processes underlie the psychological dispositions or processes which we cite in such explanations—these have no bearing on what interests us […] there may be no systematic correlations between descriptions of intentions and of cerebral processes. Objects picked out by common sense, since they are not

> necessarily (indeed not often) natural kinds, won't usually have any systematic reductive correlations with any microstructural descriptions. (Wilkes 1989a: 168-169)

Even if a representation were radically token-token correlated, perceived object to neural processes, something that could still be simple co-variation, Wilkes doesn't think this is explanatory:

> But more than that: it might put into question the viability of [person-level intentions or] representations as appropriate scientific explananda or explanantia. If representations cannot be explained systematically by states of the brain, what is the scientific justification for postulating them in the first place? One reply, of course, is to say that psychology is 'autonomous.' (Wilkes 1989a: 171)

If person-level intentions can be explained by a physically structured object (a structural neural representation), then person-level intentions would play no part in (or be redundant in) explanation of behavior, and psychology would not be autonomous; if they can't be so explained, then they seem less real. Wilkes is here anticipating and opting for Dennettian pragmatism:

> Nonetheless it is hard to avoid the conclusion that explanations lacking 'intentions,' or 'goal representations' will, by and large, come out as superior to those that possess them […]. To defend such woolly postulates as 'intentions' or 'representations' we'd need to establish that there were instances of behaviour which could not be explained without them; or which could only be explained in a highly unwieldy way without them. This is absolutely crucial, for, if this could be established, then other deficiencies of 'intentionalist' theorising can perhaps be overlooked; I have already argued that it's better to have some explanation than no explanation. (Wilkes 1989a: 176)

5. *Extended mind but not enactivism*

For Wilkes, the concept of intention is clearly the idea of a prior intention, formed in deliberation. Wilkes discusses a representation of a goal that a person forms as the result of some deliberation (going to the bank tomorrow), which action the person then sets aside until tomorrow. There is no mention of intention-in-action when tomorrow comes, something one would find in Searle. Rather, Wilkes compares the representation held in memory to an example that has since become a central example in the idea of the extended mind—setting this decision down in a notebook:

> These [notebook and natural] representations help guide our behaviour. They seem, too, to be phenomena that we want to construe realistically: phenomena needing to be individuated, and which 'really exist' […]. What is special about the 'guiding' of diary entries, or sudden recollections of an earlier decision? Simply that they cannot guide us unless their semantic content is understood. The marks in a diary must have meaning for the user […] (Wilkes 1989a: 181)

Indeed, she seems to anticipate and endorse the idea of extended mind: "Some people do not need diaries, and keep their decisions 'recorded' in

short-term memory—but the difference between written and remembered intentions seems insignificant" (Wilkes 1989a: 181).

This example points to a significant distinction that remains implicit in Wilkes's account—and it may clarify things to make it explicit. Both forms of representations (one the result of biological memory; the other an external representation in a notebook) are *products* of some cognitive doings. They are not representations with original content, or representations found in the mechanistic or physiological processes that may be doing our cognitive work. Representations-as-products may have an influence on such mechanistic or physiological processes in a derived or secondary way, if they loop back into subsequent cognition, or inform our intentions-in-action as we set out to do our action.

Accordingly, there is an important distinction between:

- *Representations as products* of cognitive processes, operating at the personal level—e.g., memory is the representation of a past event—this may involve language—and may be part of a folk psychological explanation.
- *Representations as components of (or processes in) the mechanisms* that explain cognition, operating on a subpersonal level, providing a functional or physical explanation.

For Wilkes, "The earlier deliberation, *resulting* in some intention [representation] or other, seems required by any adequate explanation of the behaviours in question […]" (Wilkes 1989a: 181), at least in a folk psychological explanation.

What's been confusing throughout this discussion is the difference between a representation that is posited as part of the mechanism that produces cognition (these are the putative subpersonal structural representations characterized in causal terms) and a representation as a personal-level intention that is the product of a deliberative cognition. Wilkes is not sure whether either form of representation is real, but thinks, along with Dennett, that the most pragmatic strategy is to take the intentional stance and accept the usefulness of personal level intentions, even if we have to worry that this does not give us a scientific explanation:

> That the explanation of much behaviour can only be given in teleological, intensional idioms, and even idioms that cite 'intentions,' I accept. That behaviour *as classified in ways appropriate to a science of behaviour* can only be handled by reference to […] internal representation [framed in terms of] the neurophysiological ('hardware') nitty-gritty" […] I doubt (with a few qualifications). (Wilkes 1989b: 204–205)

If Wilkes thinks that this is in some agreement with Dennett, Dennett, in a critical response to Wilkes, disagrees:

> I certainly agree that explanations are not all of the same type. I distinguish physical stance explanations, design stance explanations and intentional stance explanations. There are finer distinctions that also seem well-motivated to me, but I don't yet see why we can't use them all in science—and in everyday 'commonsense' explanations. (Dennett 1989: 237)

In regard to current debates about representation, let me note that enactivists who tend to be the strongest anti-representationalists in the embodied cognition camp, would not necessarily object to Wilkes' or Dennett's pragmatic defense of action-intentions, that is as products of deliberative processes, especially if such deliberations involve language. They may add intentions-in-action (P-intentions) and motor intentions (M-intentions, processed at the subpersonal level) to get a fuller story (see Pacherie 2008); but they would object, as Wilkes does, to positing subpersonal representations as part of the explanatory mechanism for any of these intentions.

There are, however, two things that suggest that Wilkes would not be happy moving in the direction of enactivism. The first is her story about the fuel-saving device; the second is her citation of Macphail about animal intelligence.

First, Wilkes repeats a story she learned from Naomi Sheman:

> An advertisement claims that unbeknownst to most drivers (perhaps because the automobile manufacturers are in cahoots with the oil companies), there is a fuel-saving device in all cars, and for a mere $29.95 we will send you what you need to know in order to activate it. When you send in your money what you receive is a set of tips such as: avoid jack rabbit starts, use the highest possible gear, do not overuse the choke, disconnect your air conditioner, and so on. Now, it's true that if you follow such tips you and your car will be performing the function of conserving fuel, but it is worse than misleading - it is simply false - to claim that there is in the car a fuel-saving device. That is, there is no physical token—however complex—which corresponds to the functional description 'fuel-saver.' (Wilkes 1989a: 163, quoting Sheman)

Wilkes would want her money back whereas enactivists would not, although they might also lodge a complaint about the misleading term 'device.' The enactivist solution is not to look for or expect to find a device or physically structured object. The enactivist would be satisfied with what Wilkes had called tacit representations (at the personal level)—"dispositions, abilities, know-how: where and what we can do depends upon the way we are, or—sometimes—on what we have learned," to which we can add habits of avoiding jack rabbit starts, using the highest possible gear, not overusing the choke, disconnecting your air conditioner, and so on. Now Wilkes might say of the enactivist, 'a sucker is born every minute;' and the enactivist might reply, the sucker is the one who expected to find a fuel-saving device in the box. Although that's not Wilkes, she still seems to worry that there is no such device.

Second, Wilkes considers a thesis by Evan Macphail (1982, 1986). She summarizes:

> [His] hypothesis proposes that there is no quantitative or qualitative difference in intelligence among the vertebrate species, excluding man. He claims that there is no solid evidence that any of the cognitive feats ascribed to allegedly more intelligent species, like chimpanzees, cannot be rivalled by any other vertebrate—once one has taken into account and allowed for differ-

> ences in perceptual capacity, motivation, physical capacity, and other such contextual variables. (Wilkes 1989a: 178)

Wilkes suggests this is a radical claim that is difficult to access—she neither accepts nor rejects it, but finds it useful to make a point, which is about the importance of language.

Enactivists would reject Macphail's claim outright, not because they would reject the importance of language, but because it presumes to define intelligence as if intelligence were constitutionally independent of all the differences listed, which are primarily differences in embodiment, which also means differences in brain structure, motility, skills, etc.—all of which adds up to what we consider to be intelligence. Macphail's hypothesis would deny that intentions are embodied, embedded (in physical and social environments), enactive, affective, and perhaps extended. Rather, on Macphail's hypothesis, intentions are exclusively the product of human linguistic ability. Wilkes too points to the idea that the addition of language is what allows for the addition of intention formation:

> [I]f language is, as I believe, crucial for consideration of the need to postulate intentions, and if chimpanzees have some capacity for linguistic communication, then maybe some chimpanzee behaviour might require us to talk in terms of goal-representations. If not, not. (Wilkes 1989a: 180)

This would be to ignore P-intentions and M-intentions, but also the embodied and situated (wider) features of deliberation and D(istal)-intention formation, which are, at least in the human, always (explicitly or implicitly) socially embedded. Wilkes instead opts for a narrow conception of, if not in-the-head, then in-the-sentence form of intentionality:

> This is a very modest conclusion, though. It restricts 'goal representations' to language-using creatures, and even there argues for their utility only in cases where the deliberation, and framing of intentions, is explicit, prior to the action taken, and linguistic. (Wilkes 1989a: 182)

Consider, however, the rat. Wilkes considers a suggestion made by David McFarland, that "rats are capable of some cognitive evaluation" (McFarland 1989: 223; Wilkes 1989b: 209). McFarland, appealing to experiments by Adams and Dickinson (1981), offers what I think is a curious claim, that rats can cognitively evaluate in a way that involves a practical inference operating on a proposition-like form, which means that "the animal makes use of declarative representations in evaluating the likely consequences of its behaviour" (McFarland 1989: 223), but that this does not involve goal-directedness.

Wilkes clearly rejects McFarland's rejection of the goal-directed nature of such representations. I think that for Wilkes, *if* rats cognitively evaluate, then that involves intentionality and goal-directedness. Still, she does not necessarily accept that rats can cognitively evaluate (or deliberate). "I am left agnostic about 'representations of goals' in non-human animals, or human behaviours that do not obviously require linguistic talents" (Wilkes 1989b: 209).

More recent empirical evidence suggests that rats do deliberate about goals, although whether this involves propositional declarative representations is, to say the least, an open question. Martin Miłkowski summarizes some recent research:

> [R]odents plan future paths, which is reflected in the future-oriented navigational activity of place cells in the hippocampus in the brains of rats. This activity was directly observed in an elegant experiment (Pfeiffer and Foster [2013]). As it turns out, rats pause before taking a journey. During that pause, place cells emit sharp-wave-ripple events: irregular bursts of brief (100– 200 ms) large-amplitude and high-frequency (140–200 Hz) activity. These are distinct from regular spikes in place cells. Using an algorithm proposed earlier for decoding similar neural events […] Pfeiffer and Foster were able to show that place cells are used to represent the future journey of the rat to the location of a previously observed reward. (Miłkowski 2023: §5.2).

Milkowski is making a claim about subpersonal neural (structural rather than declarative) representations. Dennett might accept this, but enactivists and Wilkes would likely reject it. In any case this puts us right back into the problematic that Wilkes was wrestling with, and it suggests that we still have to work out some unresolved issues. In effect, Wilkes' analysis continues to be directly relevant to contemporary discussions.

References

Adams, C.D. and Dickinson, A. 1981. "Instrumental responding following reinforcer devaluation." *Quarterly Journal of Experimental Psychology* 33 (2b): 109–112.

Anscombe, G. E. M. 1957. *Intention*. Oxford: Blackwell.

Chisholm, R. M. 1957. *Perceiving: A Philosophical Study*. Ithaca and London: Cornell University Press.

Coles, M. G. 1989. "Modern mind-brain reading: psychophysiology, physiology, and cognition." *Psychophysiology* 26 (3): 251–269.

Cox, D. D., and Savoy, R. L. 2003. "Functional magnetic resonance imaging (fMRI)'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex." *Neuroimage* 19 (2): 261–270.

Dennett, D. C. 1989. "Comments." In Montefiore, A. and Noble, D. (eds.). *Goals, No-goals and Own Goals. A Debate on Goal-directed and Intentional Behaviour*. London: Unwin Hyman, 229–237.

Dennett, D. C. 1982-3. "Styles of mental representation." *Proceedings of Aristotelian Society* LXXXIII: 213–226.

Egan, F. 2014. "Explaining representation: a reply to Matthen." *Philosophical Studies* 170: 137–142.

Friston, K. J., and Frith, C. D. 2015. "Active inference, communication and hermeneutics." *Cortex* 68: 129–143.

Frith, C. D. 1992. *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale: Lawrence Erlbaum Associates.

Frith, C. D. and Gallagher, S. 2002. "Models of the pathological mind: An interview with Christopher Frith." *Journal of Consciousness Studies* 9 (4): 57–80.

Gabor, D. 1946. "Theory of communication. Part 1: The analysis of information." *Journal of the Institution of Electrical Engineers-part III: Radio and Communication Engineering* 93 (26): 429–441.

Gallagher, S. and Allen, M. 2018. "Active inference, enactivism and the hermeneutics of social cognition." *Synthese* 195 (6): 2627–2648.

Gerrans P. 2014. *The Measure of Madness: Philosophy of Mind, Cognitive Neuroscience, and Delusional Thought*. Cambridge: MIT Press.

Haynes J-D, et al. 2007. "Reading hidden intentions in the human brain." *Current Biology* 17 (4): 323–328.

MacPhail, E. M. 1982. *Brain and Intelligence in Vertebrates*. Oxford: Clarendon Press.

MacPhail, E. M. 1986. "Vertebrate intelligence: the null hypothesis." In L. Weiskrantz (ed.). *Animal Intelligence*. Oxford: The Clarendon Press, 37–50.

McFarland, D. J. 1989. *Problems of Animal Behaviour*. London: Longmans.

Miłkowski, M. 2023. "Correspondence theory of semantic information." *The British Journal for the Philosophy of Science* 74 (2): 485–510.

Pacherie, E. 2008. "The phenomenology of action: A conceptual framework." *Cognition* 107 (1): 179–217.

Pfeiffer, B. E. and Foster, D. J. 2013. "Hippocampal place-cell sequences depict future paths to remembered goals." *Nature* 497: 74–79.

Piccinini, G. 2022. "Situated neural representations: Solving the problems of content." *Frontiers in Neurorobotics* 16: 846979.

Putnam, H. 1980. "Philosophy and our mental life." In N. Block (ed.). *Readings in Philosophy of Psychology*, Volume I. Cambridge: Harvard University Press, 134–143.

Schaffner, K. F. 2020. "Approaches to multilevel models of fear: The what, where, why, how, and how much?". In K. S. Kendler, J. Parnas, and P. Zachar (eds.). *Levels of analysis in psychopathology: Cross-disciplinary perspectives*. Cambridge University Press, 384–409.

Wilkes, K. 1989a. "Representations and explanations." In Montefiore, A. and Noble, D. (eds.). *Goals, No-goals and Own Goals. A Debate on Goal-directed and Intentional Behaviour*. London: Unwin Hyman, 159–184.

Wilkes, K. 1989b. "Explanations—How not to miss the point." In Montefiore, A. and Noble, D. (eds.). *Goals, No-goals and Own Goals. A Debate on Goal-directed and Intentional Behaviour*. London: Unwin Hyman, 194–210.

Croatian Journal of Philosophy
Vol. XXIV, No. 72, 2024
https://doi.org/10.52685/cjp.24.72.4
Received: November 20, 2023
Accepted: January 15, 2024

Minds, Machines and Gödel

ZVONIMIR ŠIKIĆ
University of Rijeka, Rijeka, Croatia

A very popular argument for the difference between mind and machine are Gödel's incompleteness theorems. Here we present some of the most famous such arguments, as well as their most famous criticisms. Finally, we offer our own reconstruction of the argument and show why it is not valid.

Keywords: Gödel's incompleteness theorems; mind vs. machine; consistency; ω-consistency.

1. *Gödel's theorems*

The vast majority of those who use Gödel's theorems of incompleteness to argue for mind-machine non-equivalence do not fully understand what Gödel's theorems are claiming. So we will begin by presenting the theorems. Gödel's first incompleteness theorem reads as follows.

If formal mathematical theory M includes an appropriate amount of arithmetic it contains an explicitly definable sentence G which asserts its own unprovability and is such that, if M is consistent then $\nvdash_M$ G and if M is ω-consistent then $\nvdash_M$ –G.

In what follows $\vdash$ is $\vdash_M$ and M is a formal mathematical theory which includes an appropriate amount of arithmetic and we think of it as a machine.

Gödel's second incompleteness theorem reads as follows.

If formal theory M is consistent it cannot prove its consistency, Con(M), which is expressed by –Pr ('⊥'), because $\vdash$ Con(M) ↔ G. (About provability predicate Pr (x) see in the appendix.)

Concerning formal unprovability of G and –G, it can be proved that

$\vdash$ –Pr ('G') ↔ Con(M) and $\vdash$ –Pr ('–G') ↔ Con (M + Con(M)).

Notice that Con (M + Con(M)) is stronger than Con(M), by the second incompleteness theorem. On the other hand, it can be proved that Con (M + Con(M)) is a weaker requirement than ω-consistency (even

weaker than 1-consistency which is a weakening of ω-consistency). Ideas of the proofs of some of these results are given in the appendix.

Let us now turn to "Gödelian dualist" arguments and their refutations.

2. *Gödel*

We will start with Gödel. In (Gödel 1951) he admits the possibility that human mind is a machine unable to understand completely its own functioning. By the end of the article I will explain that there are very good reasons for such a point of view.

Gödel even says it is conceivable that it would be known with empirical certainty that the brain suffices for the explanation of all mental phenomena and is a machine in the sense of Turing.

Hence, "Gödelian dualist" would have a hard time convincing Gödel himself.

3. *Penrose, Boolos and Good*

Penrose claims that we can see that G is true as follows (Penrose 1999). If G is provable in Peano arithmetic PA then it is false (because it asserts that it is not provable). But that is impossible *"because our formal system should not be so badly constructed that it actually allows false propositions to be proved* [in other words the system should be correct]*."* So, G is unprovable and therefore true.

Boolos asks what about ZFC (Boolos 1990). If ZFC is correct its Gödel sentence G is also unprovable in ZFC and therefore true. But we don't know if ZFC is correct; "*we could be in the same situation regarding ZFC that Frege was before receiving the letter from Russell."*

Anyway, the argument could be much simpler. If we know that M is correct and therefore consistent then $\vdash \mathrm{Con(M)} \leftrightarrow \mathrm{G}$ implies that we know that G is true. And that's it.

Of course, M also "knows" that, because $\vdash \mathrm{Con(M)} \leftrightarrow \mathrm{G}$.

It could be that we know that Con(M) is true and that we therefore know more than M. But then, we can extend M to M + Con(M) and our knowledge of the truth of Con(M) is then successfully formalized. Of course, now the question is do we know that Con (M + Con (M)) is true etc. The "Gödelian dualist" must verify that the Con sentences of all these extensions are true. But Good successfully argued that no such proof is possible (since it would imply that the smallest non-constructible ordinal is constructible) (Good 1969).

4. *Lucas and Lewis*

Lucas bypasses this hierarchy of extensions (Lucas 1961). Introducing Gödel's theorems, we already said that there is a function Con that assigns a sentence Con(M) to each theory M in such a way that:

C1. Con(M) is true if and only if M is consistent.
C2. If M is correct then Con(M) is true.
C3. Con(M) is provable if and only if M is inconsistent.

Call C a *consistency sentence* for set of sentences S iff there is M such that S is the set of its provable sentences and C = Con(M). Lucas introduced the following rule of inference which is valid, by C2:

L. If C is a consistency sentence for S, infer C from S.

Lucas extended PA to LA, with the rule L. If PA is correct then LA is correct, because L is a valid rule of inference. Furthermore, if LA is a formal theory, its consistency sentence C = Con (LA) would be its theorem, by L, and LA would be inconsistent, by C3. Hence, by C1, the falsehoods would follow from PA. Therefore, if PA is correct we know that Lucas arithmetic is not the output of any formal theory.

So if Lucas can verify all the theorems of Lucas arithmetic then Lucas is no machine.

But we are given no reason to believe that he can. As Lewis warned, in order to check whether Lucas's rule L has been used correctly, a checking procedure would have to decide whether a given set S of sentences is the output of a formal theory and that, we know, is an undecidable problem (Lewis 1989). So we do not know how many theorems of LA Lucas can produce. He can certainly go beyond PA, but he can go beyond it and still be a machine, because limitations on his ability to verify theoremhood in LA may leave him unable to recognize a lot of theorems of LA.

5. *McCall not understanding Godel's theorem*

McCall's reasoning differs from the earlier "Gödelian dualist's" arguments in his admission that the recognition of truth of G, assigned to a formal theory M, depends essentially on the unproved assumption that M is consistent (McCall 1999). That is why McCall refers to the distinction between following formal and informal claims:

A. If M is consistent then G is not provable
   $\vdash$ Con(M) $\rightarrow$ –Pr ('G')
B. If M is consistent then –G is not provable
   $\vdash$ Con(M) $\rightarrow$ –Pr ('–G')).

He claims that both informal sentences are true. He also claims that the formal version of A. is a theorem, whereas the formal version of B. *"to the best of* [*his*] *knowledge"* is not. Hence, McCall concludes that B. yields the informally true but formally unprovable sentence.

But, informal sentence in B. is not true! Unprovability of –G depends on ω-consistency. We can recognize that –G is not provable, if we assume not only the consistency of M, but the ω-consistency of M. And M can do even better, because

$\vdash$ Con (M + Con(M)) $\leftrightarrow$ –Pr ('–G')

and ω-consistency implies Con (M + Con(M)) but is not implied by it (of course, when M proves something then we know it too).

6. *My account*

My own account of dualists' argument is as follows (see Šikić 2005). "Gödelian dualist" argue that no machine M can be identical to a human mathematician H, in the following way. Let M_p be the set of arithmetical sentences provable by M and H_k is to be the set of arithmetical sentences knowable by H (the only property of the notion of knowledge we will need is that knowledge entails truth and that truth does not entail knowledge).

It must be that $M_p \subseteq H_k$ or $M_p \nsubseteq H_k$.

In the second case $M_p \neq H_k$, hence $M \neq H$.

In the first case whatever is provable by M is knowable by H and that means that all sentences in M_p are true. Therefore H knows that M is a correct system. But then H knows that it is a consistent system, i.e. $Con(M) \in H_k$. But $Con(M) \notin M_p$, by second Gödel's theorem, hence $M_p \neq H_k$ and therefore $M \neq H$.

Hence, $M \neq H$ in every case.

But the above conclusion "Therefore H knows that M is a correct system" is not justified. From the truth that every sentence provable by M is knowable by H it follows that every sentence provable by M is true (i.e. that M is correct) but it does not follow that H knows that, because truth does not entail knowledge. It is possible that $M_p \subseteq H_k$ and that H does not know that.

In some specific cases we may know just enough to conclude that M is a correct system. On the other hand, it remains possible that there may exist mathematical machines which in fact are equivalent to our mathematical intuitions. For example, we could be such machines.

What follows from Gödel's incompleteness theorem is that:

There is no machine which could capture all our mathematical intuitions *and which we could understand well enough to know that it is consistent (i.e. that G is true).*

It does not follow that:

There is no machine which could capture all our mathematical intuitions.

We may conclude. As far as Gödel's incompleteness theorems are concerned we could well be machines. But if we are then we are definitely not capable of the complete knowledge of the machines, i.e. of the complete knowledge of ourselves.

That explains Gödel's understanding of the problem in Gödel (1995).

7. *Appendix*

If formal mathematical theory M includes an appropriate amount of arithmetic it can refer to its expression F with its Gödel's number 'F'.

Gödel defined arithmetical predicate Prv (x, y) which represents "x is proved by y" (within M itself) and proved that:

1) n is Gödel's number of a provable formula $\Rightarrow \vdash$ Prv (n, m) for some m
2) n is not Gödel's number of a provable formula $\Rightarrow \vdash$ –Prv (n, m) for every m

Gödel then defined Pr (x), which represents "x is provable", as $\exists y$ Prv (x, y).

Furthermore, (B1) easily follows from 1) and (B1') easily follows from 2) and ω-consistency. It is also easy to prove (B2) and somewhat more difficult (B3).

(B1) $\vdash X \Rightarrow \vdash$ Pr ('X'),
(B1') $\vdash$ Pr ('X') $\Rightarrow \vdash X$ if M is ω-consistent
(B2) $\vdash$ Pr ('X $\rightarrow$ Y') $\rightarrow$ (Pr ('X') $\rightarrow$ Pr ('Y')),
(B3) $\vdash$ Pr ('X') $\rightarrow$ (Pr ('Pr ('X')').

For any predicate P (x), substitution of 'P (d (x))' for x in P (d (x)) gives P (d ('P (d (x))')), or D for short. It immediately follows that D $\Leftrightarrow$ P ('D'). Hence, there is a sentence G such that

(DL) $\vdash G \leftrightarrow$ –Pr ('G')

From (DL), (B1) and (B1') we can deduce the first incompleteness theorem. Namely,

$\vdash G \Rightarrow \vdash$ Pr ('G') $\Leftrightarrow \vdash$–G
$\vdash$ –G $\Leftrightarrow \vdash$ Pr ('G') $\Rightarrow \vdash G$

Both implications contradict the consistency of M. Hence, $\nvdash$G and $\nvdash$–G. Note that we used (B1'), i.e. ω-consistency, to prove the unprovability of –G.

From (DL), (B1), (B2) and (B3) we can deduce the second incompleteness theorem:

$\vdash G \rightarrow$ (Pr ('G') $\rightarrow \perp$)
$\vdash$ Pr ('G') $\rightarrow$ (Pr ('Pr ('G')') $\rightarrow$ Pr ('$\perp$'))
$\vdash$ Pr ('G') $\rightarrow$ Pr ('$\perp$')
$\vdash$ –Pr ('$\perp$') $\rightarrow$ –Pr ('G') i.e. $\vdash$ Con(M) $\rightarrow$ G
$\vdash \perp \rightarrow$ G
$\vdash$ Pr ('$\perp$') $\rightarrow$ Pr ('G')
$\vdash$ –Pr ('G') $\rightarrow$ –Pr ('$\perp$') i.e. $\vdash$ G $\rightarrow$ Con(M)

Now, from $\nvdash$ G and $\vdash$ Con(M) $\leftrightarrow$ G it immediately follows that $\nvdash$ Con(M).

So, by (DL) and $\vdash$ Con(M) $\leftrightarrow$ G, unprovability of G is provably equivalent to the consistency of M:

$\vdash$ –Pr ('G') $\leftrightarrow$ Con(M)

What do we know about the unprovability of –G, which is the other part of the first incompleteness theorem? From $\vdash$ –G $\leftrightarrow$ Pr (), by (B1) and (B2), we get

$\vdash$ –Pr ('–G') $\leftrightarrow$ –Pr ('Pr ('$\perp$')').

But –Pr ('Pr ('⊥')') expresses the consistency of M + Con(M). Namely, if $Pr_{M+Con(M)}$ is the provability predicate of M + Con(M), then the consistency of M + Con(M) is expressed by $–Pr_{M+Con(M)}$ ('⊥'). But,

$$–Pr\ (\text{'Pr}\ (\text{'}\bot\text{'})\text{'}) \Leftrightarrow –Pr_{M}\ (\text{'}–Con(M)\text{'}) \Leftrightarrow –Pr_{M+Con(M)}\ (\text{'}\bot\text{'})$$

Hence

$$\vdash –Pr\ (\text{'}–G\text{'}) \leftrightarrow Con\ (M + Con(M)).$$

References

Boolos, G. 1990. "On seeing the truth of the Gödel sentence." *Behavioural and Brain Sciences* 13: 655–656.

Gödel, K. 1931. "Über formal unentscheidbäre Sätze I." *Monatshefte für Mathematik und Physik* 38: 173–198.

Gödel, K. 1995. "Gibbs Lecture, 1951." In S. Feferman (ed.). *Collected Works, Vol. 3: Unpublished Essays and Lectures*. Oxford: Oxford University Press, 290–323.

Good, I. J. 1969. "Gödel's theorem is a red herring." *The British Journal for the Philosophy of Science* 18: 359–373.

Lewis, D. 1989. "Lucas against mechanism II." *Canadian Journal of Philosophy* 9: 373–376.

Lucas, J. R. 1961. "Minds, machines and Gödel." *Philosophy* 36: 112–137.

McCall, S. 1999. "Can a Turing machine know that the Gödel sentence is true?" *Journal of Philosophy* 96: 525–532.

Penrose, R. 1999. *The Emperor's New Mind*. Oxford: Oxford University Press.

Penrose, R. 1994. *Shadows of the Mind*: *A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Šikić, Z. 2005. "Gödel's Incompleteness and Man-Machine Non-Equivalence." *Grazer Mathematische Berichte* 304: 75–78.

Croatian Journal of Philosophy
Vol. XXIV, No. 72, 2024
https://doi.org/10.52685/cjp.24.72.5
Received: April 3, 2024
Accepted: June 3, 2024

Human and Artificial Decision Making: A Unified View

KONSTANTINOS V. KATSIKOPOULOS
University of Southampton, Southampton, England

Machines can now match, or outperform, human performance in several reasoning and decision tasks. Some say that all that intelligence amounts to is smart computation. This is not a new thesis, dating back to Leibniz as well as Simon and Newell, but what is new is what smart means. Today it is identified with complex statistics and optimisation. Simon's meaning, however, of smart rested on bounded rationality, a unified view of human and artificial decision making. This view was fleshed out by Gigerenzer as fast-and-frugal heuristics. Interestingly, such heuristics are typically sparse, as some machine learning models are optimised to be. So, one might hope that we can make sense of artificial intelligence in human terms after all, and face the upcoming challenges with open-mindedness and courage, just like Simon, and of course Wilkes, would have done.

Keywords: Human decision making; artificial intelligence; bounded rationality; heuristics; smart computation; sparsity.

1. *Overview*

Colin Cherry was what we would today call a cognitive scientist… Or a communications engineer, or a researcher of artificial intelligence... Well, anyway, Cherry was all of the above. His "cocktail party effect" can serve to illustrate my contribution to the 2023 Kathy Wilkes Memorial Conference, also weaving in the ideas of Herb Simon. Cherry, Simon, and Wilkes align in all being open-minded, courageous researchers, who asked the tough questions of what effective human and machine communication, reasoning and decision making is, and provided answers that invite deep thought. My contribution connects to the conference contributions by Philipp Koralus (human reasoning and decision making) and Peter Millican (artificial intelligence).

First, on the cocktail party effect. Cherry (1953) run laboratory experiments where participants listened to two different messages from the same speaker and were instructed to separate them. Whether and the extent to which people can perform such tasks accurately depends on factors such as the direction from which the messages are coming, the pitch of the messages and the rate of speech. In a classic demonstration, it was found that people can detect message segments they are not actively attending to only if these segments are important to them, as when their name is spoken (Moray 1959). For a long time, human performance on cocktail party settings could not be matched by machines. But Xie et al. (2015) combined acoustic metamaterials with computational sensing to achieve competitive performance, wherein three overlapping and independent auditory sources were separated with 97% accuracy.

This machine success raised some eyebrows then, but it might have done much less so today. In the last decade, we have had the reinforcement learning algorithm Alpha Zero mastering simultaneously the games of Go, chess, and shogi at world-class level (Silver et al. 2018). And you have probably all heard enough about large language models such as Chat GPT in the past couple of years. It is definitely on the table now that machine behavior can be as intelligent as human behavior, or even more so. Furthermore, some forcefully say that intelligent behavior *only* requires smart computation.

This is not a new thesis. In the seventeenth and eighteenth centuries Gottfried Leibniz had the dream of a *Characteristica universalis*, a universal language of formal computation. Herb Simon and Allan Newell received the 1975 Turing award for their *physical symbol systems hypothesis*, stating that symbol manipulation is a necessary and sufficient condition for intelligence, be it human or artificial. What *is* new is that smart computation has recently been identified with complex statistics and optimization. But this was not Simon's meaning of "smart." The objective of this article is to present, drawing from my presentation at St Hilda's College conference at Oxford University, research that fleshes out Simon's unified view of an integral part of human and artificial behavior, decision making, as it has been developing from the seventies until today.

2. *Simon's meaning of "smart"*

Herb Simon has been the only person to date who received both the Nobel prize in 1978 and the Turing award in 1975. He received the Nobel prize for economics, the Turing award is for computer science. Economics is a social science and computer science is—in Simon's own words (1968)—a science of the artificial. Both economics and computing study behavior, of humans and computers respectively. In Simon's view, the analogy can be pushed further because he saw both humans and computers fundamentally as systems that exhibit intelligent be-

havior because they process information; and more specifically, as per the physical symbol systems hypothesis, they process symbols. At the highest level of abstraction, this is what being smart meant for Simon.

Now, Herb Simon was a polymath if there ever was one. He was so prolific and wrote over seven decades, that is not always easy to follow his various intellectual threads and see them fleshed out (Petracca 2021). Yes, intelligence for Simon was information processing, but is that all? Can one flesh Simon's vision out? Yes, one can and some did. Before presenting Gerd Gigerenzer's implementation, analysis, and testing of Simon's vision in the next section, at the research center he directed at the Max Planck Institute for Human Development in Berlin, let us discuss Simon's vision a little more.

Simon was gifted in formal modeling, mathematical and computational. He has produced multiple articles in areas such as statistics, decision theory and operations research. For instance, he developed fundamental methods for distinguishing spurious from genuine statistical correlations (Simon 1954) and for deriving optimal policies for stochastic dynamic programming (Simon 1956). Even though he was accused of unnecessarily, even according to some damagingly, "hardening" the social sciences, he was a dedicated scientist who understood what formal models can and cannot do for theoretical development and empirical testing (Katsikopoulos, Marewski and Hoffrage 2024).

In other words, Simon can be said to have acted respectfully to Einstein's maxim "everything should be made as simple as possible, but not more so." As such, the formal expression of smart computation Simon endorsed is considerably simpler than today's reliance on complex statistics and optimization, evident in areas such as big data analytics and statistical machine learning (Katsikopoulos and Canellas 2012). An obvious reaction to this observation is that such areas have become much more technically sophisticated since Simon's time—he passed away in 2001 and produced most of his more technical work in the fifties, sixties, seventies, and eighties—and in the most recent couple of decades. This is true of course but I do not believe that it accounts for the whole difference. For example, Simon resisted vehemently the then state-of-the-art statistical ritual of null hypothesis significance testing. And he was also pushing the envelope for developing new methods that can address stubborn problems, rather than changing problems so that the known methods can address them; hence his crusade on developing artificial intelligence. The exchange with leading applied mathematician Richard Bellman (Simon and Newell 1958; Bellman 1958), where he forcefully argued that current decision methods could not handle ill-structured problems, is particularly telling in this regard. (It would have been intriguing to hear Simon's take on today's explosion of data science and machine learning, but, alas, we do not have this privilege).

In line with his overall attitude to decision modelling, Simon (1956, 1968) voiced consistent concerns about the effectiveness of mathematical optimization outside toy problems, outside the lab, or how one might

call it, "in the wild." It is a truism that something optimal according to a model of the world might be not only suboptimal, but even poor performing, in the wild. But it is a truism that modelers all too often do not heed, preferring to go about the usual business of optimizing, without even testing how benchmarks, such as non-optimizing models, perform in the wild. The fields of "soft" and "behavioral" operations research have been more sensitive than "hard" operations research, considering and acting on such points (Ackoff 1979; Rosenhead and Mingers 2001; for a historical and conceptual perspective on all these types of operations research, see Katsikopoulos 2023). But it is interesting that Simon first made such suggestions decades before (for another aligned contemporary perspective, see Kimball 1958).

What neither Simon, nor the fields of soft and behavioral operations research, did, however, is to develop *formal decision models*, computational and mathematical, that are applicable to the wild. Simon (1956) sketched the idea of *satisficing*—this word is a portmanteau of "satisfying" and "suffice"—models, which, contra optimization, do not search and settle only on the theoretically best option, but may choose another option based on criteria other than optimality, such as adaptiveness and robustness. Simon did not empirically test how far this idea can go but conjectured that it can: "The presence of uncertainty places a premium on robust adaptive procedures instead of optimization strategies that work well only when finely tuned to precisely known environments" (Simon 1968: 35). Was he right? The next section provides some answers.

3. *Gigerenzer's analysis, implementation, and testing of Simon's "smart"*

A broad and deep investigation of Simon's conjecture has been ongoing since the mid-nineties. Gerd Gigerenzer, a psychologist with a philosopher's inclination to analyze conceptually and a scientist's skill to implement and test empirically, did exactly that with Simon's concept of "smart." Whereas many have claimed to stand on Simon's shoulders—including no less the founders of the modern field of heuristics, psychologists Amos Tversky and Daniel Kahneman—scrutiny reveals that perhaps it was Gigerenzer who most closely did so. Historian Enrico Petracca has even, fittingly I believe, suggested that the work of Gerd Gigerenzer, Peter Todd and colleagues "could appear 'more Simonian than Simon'" (2021: 710). In a nutshell, Gigerenzer, Todd and the ABC research group (1999) can take credit for making clear and distinguishing three possible interpretations of Herb Simon's key idea of *bounded rationality*, as explained below.

The dominant interpretation of bounded rationality in economics (Sargent 1993) is that it is still optimization, but under constraints (e.g., cognitive, systemic). The leading interpretation of bounded ra-

tionality in psychology (Kahneman, Slovic and Tversky 1982) is consistent with the economics one, but takes the form of attributing the (supposedly regrettable) lack of (full) optimization to people's heuristics. This is the well-known *heuristics-and-biases* research program. The third interpretation of bounded rationality is more interdisciplinary, aligned with the concerns of disciplines that assess decision performance in the wild, such as human factors, operations research, and artificial intelligence (Katsikopoulos 2023). According to Gigerenzer et al. (1999), bounded rationality is not the study of what theoretically can be called lack of rationality, but the study of what practically is a real rationality that real organisms can and aim to exhibit. Bounded rationality is implemented by heuristics, some of which are of the satisficing variety originally proposed by Simon and others of which are *fast-and-frugal heuristics*, which constituted a new major research program, and according to some such as Kelman (2011), the main antipode to the heuristics-and-biases program (see also the paper by Koralus in the conference and in this collection). A main thesis of fast-and-frugal heuristics is that intelligent people, and machines, use few, informative, pieces of information, and combine those pieces in mathematically simple ways.

Fast-and-frugal heuristics are, just like most research developments, not entirely new. Early demonstrations of the concept can be found in the seventies in the work of Robyn Dawes (Dawes and Corrigan 1974) and Robin Hogarth (Einhorn and Hogarth 1975), who empirically showed that just tallying variables, without weighting them differentially as in least-squares regressions, could lead to equally, and sometimes even more, accurate predictions. Such results were, however, not typically taken that seriously. Gigerenzer's greater volume of empirical results, and of supporting theoretical analyses, with the help of some dozens of researchers at the Max Planck Institute for Human Development and in a world-wide network (Gigerenzer, Hertwig and Pachur 2012) eventually attracted more attention, and served to establish the success of fast-and-frugal heuristics across many domains in the wild (Katsikopoulos, Şimşek, Buckmann and Gigerenzer 2020). The remaining of this section samples two empirical demonstrations, both from the geopolitical sphere trying to connect to Kathy Wilkes' activism and gives a glimpse of the analytical theory.

Predicting election outcomes. Ahead of the 2016 U.S. presidential election, big data algorithms predicted a 71.4% chance of Hilary Clinton winning (Katsikopoulos et al. 2020). Furthermore, polls and prediction markets made the same prediction. Historian Allan Lichtman, on the other hand, predicted that Donald Trump would win. Lichtman (1980) *13 keys to the White House* is a tallying heuristic he derived based on domain knowledge, blending theories of politics, economics, sociology, and psychology. The keys—also called attributes or features—were fixed once and for all before the 1984 election and have been used to

correctly predict all U.S. elections since. Each key is an issue that matters to voters; they are stated so that each is either true or false ahead of an election. Some of the keys are facts, others require judgment. It is of course key (pun intended) that all keys are judged and scored before the election.

Key 1: Incumbent party holds more seats in the House of Representatives after this midterm election than the previous one.
Key 2: No serious contest for incumbent-party nomination.
Key 3: Incumbent-party candidate is the sitting president.
Key 4: No significant third-party or independent campaign.
Key 5: Economy not in recession during campaign.
Key 6: Real annual per capita economic growth during the term equals or exceeds mean growth during two previous terms.
Key 7: Incumbent administration effects major changes in national policy.
Key 8: No sustained social unrest during the term.
Key 9: Incumbent administration untainted by major scandal.
Key 10: Incumbent administration suffers no major failure in foreign or military affairs.
Key 11: Incumbent administration achieves a major success in foreign or military affairs.
Key 12: Incumbent-party candidate is charismatic or national hero.
Key 13: The challenging-party candidate is not charismatic or national hero.

Lichtman proposed the following heuristic:

> *Score all keys and tally the number of false keys. If this tally is six or more, the challenger will win.*

For instance, in the 2012 election, Mitt Romney challenged Barack Obama. Lichtman counted only three keys as false (1, 6 and 12), and correctly predicted that Obama would win. In late September 2016, Lichtman found six false keys (1, 3, 4, 7, 11 and 12) and predicted, again correctly, that Trump would win (for further discussion, including some subtleties, see Katsikopoulos et al. 2020).

Unlike big data analytics, the 13-key rule is transparent. The rule contradicts campaign wisdom: All keys—except key 13—refer to the incumbent party, i.e., the party holding the White House and its candidate. That is, incumbents tend to lose, rather than challengers tend to win. The heuristic delivers a simple theory, a process-based explanation for behavior, and creates a platform for discussion, qualities important in a healthy democracy (Katsikopoulos and Canellas 2022).

Understanding and improving the operation of checkpoints. In checkpoints set up by NATO in Afghanistan between 2004 and 2009, soldiers had to classify approaching cars as a friend or a foe, and decide how to make the car stop, so that it can be respectively searched or neutralized. How did soldiers make these decisions? Did it work? Can research help do better?

The Wikileaks reports mined by Keller and Katsikopoulos (2016) referred to 1160 incidents, of which 7 were suicide attacks and 1053 civilian. Suicide attacks resulted to all car occupants and soldiers dying. The civilian incidents resulted to 204 people injured or killed. Applying standard operations research or artificial intelligence techniques is tempting but does not work because the empirical estimate of the probability of a hit (i.e., soldiers classify a car with suicide attackers as a foe) is zero. This cannot be right and would lead to extreme classifiers such as classifying all cars as civilian, which is also not right.

It seems that soldiers relied almost exclusively, for 1020 out of 1053 civilian incidents, on a heuristic that uses one attribute, *compliance*:

> *If an approaching car appears to comply with soldier instructions (e.g., slows down), then classify it as a friend and ensure peacefully that it stops and is searched. If the car does not appear to comply (e.g., speeds up), then classify it as a foe and ensure that it is neutralized, which might require shooting at it, etc.*

Can this reasonable, but to a good extent ineffective, heuristic be enhanced so that it leads to improved decision making that would have resulted to fewer than 204 civilian casualties? The authors consulted with experts, such as military personnel and teachers in military academies, about how to classify an approaching car as friend or foe and sought to combine their insights with the compliance attribute. The resulting method is more complex than a single-attribute heuristic, but still a fast-and-frugal heuristic, of the type called a *fast-and-frugal (decision) tree* (Martignon, Katsikopoulos and Woike 2008). This tree is shown in Figure 1 below (in the tree, threat cues refer to any information that makes a car seem suspicious, e.g., intelligence information).

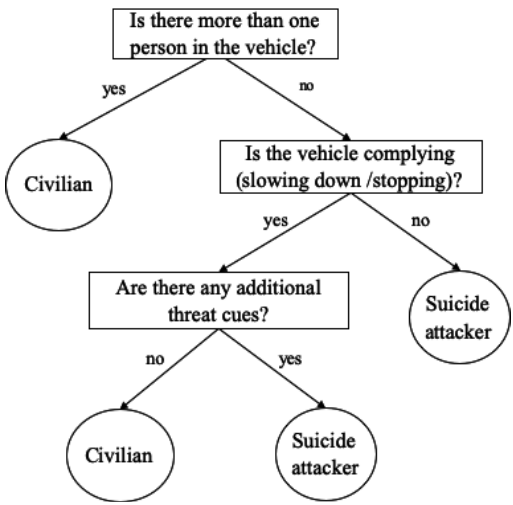

Figure 1. A fast-and-frugal decision tree for classifying cars approaching a checkpoint as friend or foe (adapted from Keller and Katsikopoulos 2016).

Decision trees are a popular type of transparent model in machine learning (Breiman et al. 1984; Bertsimas, Dunn and Mundru 2019). Fast-and-frugal trees are designed to be even more transparent. For example, the Figure 1 tree asks only a few questions, it asks those questions one at a time, and each time it asks a question it is possible that a decision is made immediately. Is the tree only transparent, or is it also accurate? Had it been applied to the *Wikileaks* dataset, it would have led to 78 casualties, that is a 60% decrease from what transpired.

Is he cherry picking? No. Katsikopoulos et al. (2020) compared the classification error of Breiman's (1984) classification and regression trees (CART) to tallying and fast-and-frugal trees in 64 classification tasks, containing 95 to 32,561 instances (median 904) and three to 1,418 cues (median 19). Across the 64 tasks, each fast-and-frugal heuristic predicted nearly as well as CART, falling behind by only half a percentage point. There is an advantage for CART in problems where the error is small, that is, in easy tasks, and an advantage for fast-and-frugal heuristic when the error is larger, that is, in more difficult tasks. Furthermore, decades of competitions among such simple heuristics and more complex optimization models such as linear regressions (including regularized versions), Bayesian networks, decision trees, random forests, and support vector machines, spread across disciplines such as psychology, economics, engineering design, statistics, and machine learning, have shown that the differences in predictive accuracy between these two model families are not that large, and that each family enjoys a region of superior performance. In other words, heuristics can be *robust*. Why? Are there explanations for these results?

Theory: The role of sparsity. Yes. For reviews, see Martignon and Hoffrage (2002), Katsikopoulos et al. (2018), and Katsikopoulos (2023: Chapter 6). A good, short answer is *sparsity*. A model is sparse if only a small proportion of its parameters are different from zero. For example, regularization techniques push regressions towards sparsity. Sparsity can make a model more predictively accurate because it does not overfit in the training set (Geman, Bienenstock and Doursat 1992; Rudin 2019).

The checkpoint fast-and-frugal tree of Figure 1 is a sparse version of full-blown CART decision trees. Tallying is a sparse version of linear regression with potentially differentially weighted attributes (Lichtenberg and Şimşek 2019). Whereas it is overall appreciated that such transparent models can be accurate as well, it should be noted that there are multiple approaches to deriving those. Bertsimas et al. (2019) and Rudin (2019) suggest that sparse models may be derived as solutions to optimization problems, while fast-and-frugal heuristics researchers may additionally generate such models by deeply observing human decision making (Gigerenzer et al. 2011; Katsikopoulos et al. 2020).

4. *Epilogue*

Cherry, Simon, and Wilkes all sought to understand human and artificial intelligence, by taking open-minded, penetrating, and courageous approaches. Humanity always faces challenges, and perhaps the artificial intelligence one will prove to be a very tough, even existential, one. May a unified view of human and artificial decision making, as the one championed by Gigerenzer and presented here, act as a resource to keep such issues at bay.[1]

References

Ackoff, R. L. 1979. "The future of operational research is past." *Journal of the Operational Research Society* 30: 93–104.

Bellman, R. 1958. "On 'Heuristic Problem Solving' by Simon and Newell." *Operations Research* 6 (3): 448–449.

Bertsimas, D., Dunn, J. and Mundru, N. 2019. "Optimal prescriptive trees." *INFORMS Journal on Optimization* 1 (2): 164–183.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. 1984. *Classification and Regression Trees*. London: Chapman and Hall.

Cherry, E. C. 1953. "Some experiments on the recognition of speech, with one and with two Ears." *The Journal of the Acoustical Society of America* 25 (5): 975–979.

Dawes, R. M. and Corrigan, B. 1974. "Linear models in decision making. " *Psychological Bulletin* 81 (2): 95–106.

Einhorn, H. J. and Hogarth, R. M. 1975. "Unit weighting schemes for decision making." *Organizational Behavior and Human Performance* 13 (2): 171–192.

Geman, S., Bienenstock, E. and Doursat, R. 1992. "Neural networks and the bias/variance dilemma." *Neural Computation* 4 (1): 1–58.

Gigerenzer, G., Hertwig, R. and Pachur, T. (eds.). 2011. *Heuristics*: *The foundations of adaptive behavior*. New York, NY: Oxford University Press.

Gigerenzer, G., Todd, P. M. and the ABC Research Group 1999. *Simple heuristics that make us smart*. New York: Oxford University Press.

Kahneman, D., Slovic, P. and Tversky, A. (eds.). 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Katsikopoulos, K. V. 2023. *Cognitive operations: Models that open the black box and predict our decisions*. London, UK: Palgrave Macmillan.

Katsikopoulos, K. V. and Canellas, M. C. 2022. "Decoding human behavior with big data? Critical, constructive input from the decision sciences." *AI Magazine* 43 (1): 1–13.

[1] Since the time this article was written and accepted for publication, two events occurred that necessitate the following updates: First, Herb Simon is no longer the only person who has received both the Nobel prize and the Turing award as Geoff Hinton has now also received those (Nobel prize in physics in 2024 and Turing award in 2019). Second, Allan Lichtman's 13 keys to the White House has no longer predicted all U.S. elections since 1984 because the model failed to predict the outcome of the 2024 election.

Katsikopoulos, K. V., Durbach, I. N. and Stewart, T. J. 2018. "When should we use simple decision models? A synthesis of various research strands." *Omega*: *The International Journal of Management Science* 81: 17–25.

Katsikopoulos, K. V., Marewski, J. N. and Hoffrage, U. 2024. "Heuristics for metascience: Simon and Popper." In G. Gigerenzer, S. Mousavi and R. Viale (eds.). *The Herbert Simon Companion*: *Standing the Test of Time*. Northampton: Edward Elgar, 300–311.

Katsikopoulos, K. V., Şimşek, Ö. Buckmann, M. and Gigerenzer, G. 2020. *Classification in the wild: The science and art of transparent decision making*. Cambridge: MIT Press.

Keller, N. and Katsikopoulos, K. V. 2016. "On the role of psychological heuristics in operational research; and a demonstration in military stability operations." *European Journal of Operational Research* 249 (3): 1063–1073.

Kelman, M. G. 2011. *The heuristics debate*. New York, NY: Oxford University Press.

Kimball, G. E. 1958. "A critique of operations research." *Journal of the Washington Academy of Sciences* 48 (2).

Lichtenberg, J. M. and Şimşek, Ö. 2019. "Regularization in directable environments with application to Tetris." *International Conference on Machine Learning*: 3953–3962.

Martignon, L. and Hoffrage, U. 2002. "Fast, frugal, and fit: Simple heuristics for paired comparison." *Theory and Decision* 52 (1): 29–71.

Martignon, L., Katsikopoulos, K. V. and Woike, J. K. 2008. "Categorization with limited resources: A family of simple heuristics." *Journal of Mathematical Psychology* 52 (6): 352–361.

Moray, N. 1959. "Attention in dichotic listening: Affective cues and the influence of instructions." *Quarterly Journal of Experimental Psychology* 11 (1): 56–60.

Petracca, E. 2021. "On the origins and consequences of Simon's modular approach to bounded rationality in economics." *The European Journal of the History of Economic Thought*: 1–24.

Rosenhead, J. and Mingers, J. (eds.). 2001. *Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict*. New York: John Wiley and Sons.

Rudin, C. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1 (5): 206–215.

Sargent, T. J. 1993. *Bounded rationality in macroeconomics: The Arne Ryde memorial lectures*. New York: Oxford University Press.

Silver, D., et al. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play." *Science* 362 (6419): 1140–1144.

Simon, H. A. 1954. "Spurious correlation: A causal interpretation." *Journal of the American Statistical Association* 49(267): 467–479.

Simon, H. A. 1956. "Rational choice and the structure of the environment." *Psychological Review* 63 (2): 129–138.

Simon, H. A. 1968. *The sciences of the artificial*. Cambridge: MIT Press.

Simon, H. A. and Newell, A. 1958. "Heuristic problem solving: The next advance in operations research." *Operations Research* 6: 1–10.

Xie, Y., Tsai, T. H., Konneker, A., Popa, B. I., Brady, D. J. and Cummer, S. A. 2015. "Single-sensor multispeaker listening with acoustic metamaterials." *Proceedings of the National Academy of Sciences* 112 (34): 10595–10598.

Table of Contents of Vol. XXIV

AVRAMIDES, ANITA
Introduction 329

BARCHANA-LORAND, DORIT
The Dark Side of Cultural Sensitivity:
Right-Wing Anxiety
and Institutional Literary Censorship in Israel 113

BORSTNER, BOJAN
A Goodbye to Nenad 143

BROWN, JAMES ROBERT
Introduction 145

BROWN, JAMES ROBERT
Miščević: Mental Models and More 147

CHEN, GONG and STEVENS, GRAHAM
Embedded Metaphor and Perspective Shifting 255

DAVIES, DAVID
Miščević On Thought Experiments 163

DOŽUDIĆ, DUŠAN
Propositions, Concepts,
and the Fregean/Russellian Distinction 219

DUMITRU, ADELIN-COSTIN
A Sufficientarian Proposal for Discharging
Our Moral Duties Towards Emigrants 295

GALLAGHER, SHAUN
Intentions and Representations 367

GEIRSSON, HEIMIR
Reclaiming Russellian Singular Thoughts 235

GRBA, MARKO
The Mystery of Intuition
in Einstein's Thought Experiments 173

HEIGL, ANTONIA
Beyond Reading: What it Means
to Encounter a Literary Work of Art 19

JUTRONIĆ, DUNJA
Farewell to Nenad 139

JUTRONIĆ, DUNJA
How is Content Externalism
Characterized by Vehicle Externalists 351

KATSIKOPOULOS, KONSTANTINOS V.
Human and Artificial Decision Making: A Unified View 387

KOSENA. ANTONIA and VIRVIDAKIS, STELIOS
Drawing Reflections: What Kind of Knowledge
Does Self-referential Literature Yield? 65

LAMARQUE, PETER
Literary Interpretation is Not Just About Meaning 3

MALLORY, FINTAN
Generative Linguistics and the Computational Level 195

MCALLISTER, JAMES W.
Thought Experiment as Bridge
Between Science and Common Sense 155

MCGREGOR, RAFE and BURNS, REECE
Social Science as a Kind of Writing 97

O'SHAUGHNESSY, ROBERT and SPREVAK, MARK
Concepts are Containers 333

OBRIGEWITSCH, ALEX
Intimations of a Lyricism sans Subject:
On the Poetics of Philippe Lacoue-Labarthe 35

PAGANINI, ELISA
Aesthetic Value of Immoral Fictions 53

ŠIKIĆ, ZVONIMIR
Minds, Machines and Gödel 381

SMOKROVIĆ, NENAD
Nenad Miščević 135

VEIT, WALTER
Rationality and Intransitivity 273

VIDMAR JOVANOVIĆ. IRIS, SLUGAN, MARIO and GRČKI, DAVID
Introduction 1

WIELAND, NELLIE
Escaping Fiction 81

Book Reviews

GRGIĆ, ANA
Owen Flanagan, *How to do Things with Emotions:*
The Morality of Anger and Shame across Cultures 319

JOLIĆ, TVRTKO
Frauke Albersmeier, *The Concept of Moral Progress* 323

LALIĆ, EMA LUNA AND VIDMAR JOVANOVIĆ, IRIS
Patrik Engisch and Julia Langkau (eds.),
The Philosophy of Fiction: Imagination and Cognition 131