Croatian Journal of Philosophy

Vol. XXIV · No. 71 · 2024

Croatian Journal of Philosophy

1333-1108 (Print)
1847-6139 (Online)

Editor:
Nenad Miščević (University of Rijeka)

Advisory Editors:
Dunja Jutronić (University of Rijeka)
Tvrtko Jolić (Institute of Philosophy, Zagreb)

Managing Editor:
Viktor Ivanković (Institute of Philosophy, Zagreb)

Editorial board:
Stipe Kutleša (Zagreb),
Davor Pećnjak (Institute of Philosophy, Zagreb)
Joško Žanić (Zagreb)

Advisory Board:
Elvio Baccarini (University of Rijeka), Carla Bagnoli (University of Bologna), Boran Berčić (University of Rijeka), István M. Bodnár (Central European University), Vanda Božičević (Bergen Community College), Sergio Cremaschi (Milano), Michael Devitt (The City University of New York), Peter Gärdenfors (Lund University), János Kis (Central European University), Friderik Klampfer (University of Maribor), Željko Loparić (Sao Paolo), Miomir Matulović (University of Rijeka), Snježana Prijić-Samaržija (University of Rijeka), Igor Primorac (Melbourne), Howard Robinson (Central Europena University), Nenad Smokrović (University of Rijeka), Danilo Šuster (University of Maribor)

Co-published by
"Kruzak d.o.o."
Naserov trg 6, 10020 Zagreb, Croatia
fax: + 385 1 65 90 416, e-mail: kruzak@kruzak.hr
www.kruzak.hr

and

Institute of Philosophy
Ulica grada Vukovara 54/IV, 10000 Zagreb, Croatia
fax: + 385 1 61 50 338, e-mail: filozof@ifzg.hr
www.ifzg.hr

Available online at http://www.pdcnet.org, http://www.ceeol.com

CROATIAN
JOURNAL
OF PHILOSOPHY

Vol. XXIV · No. 71 · 2024

In memoriam Nenad Miščević (1950 – 2024)

Book Symposium on Nenad Miščević,
Thought Experiments

Articles

Book Reviews

In memoriam
Nenad Miščević (1950–2024)

On May 11th 2024, we lost Nenad Miščević, a doyen and key figure in Croatian analytical philosophy, one of the most renowned Croatian philosophers in the world. This is what he was. But what essentially defined him was that he was a philosopher with his whole being, in every manifestation of his existence, not just by profession, not even just by vocation. From his high school days to his last moment, he directed and pursued all his activities—intellectual, social, emotional, and organizational—towards philosophy. I can personally testify to a good deal of those activities. Through his philosophical work, Nenad created his enviable philosophical reputation worldwide and, of course, in Croatia. However, what was more important for the environments in which he worked was that he inspired, if not founded, institutions where philosophy thrived everywhere he stayed. In personal contacts, he was an unsurpassed motivator, encouraging and motivating all those with interest and ability to engage in philosophical thinking and writing.

In the early days of his career, this encouragement was directed to postmodern philosophy. But for most of his life it was analytical philosophy. Starting with a circle of young philosophers and logicians, proudly referred to as the "Rijeka Circle," he created a pool for talented scientists who soon became the driving force of philosophical activities at various faculties. Nenad's professional path first led him to the Department of Philosophy at the Zadar Faculty of Philosophy. Soon, the small department became a significant center of analytical philosophy, where Nenad's personality attracted young promising philosophers, many from the mentioned "Rijeka Circle," as department employees and numerous rising stars of philosophy as guests, who are today leading world philosophers. I'll mention only Georges Rey and Michael Devitt, with whom Nenad remained friendly throughout his life. After being expelled from Zadar, he got a job at the Department of Philosophy at the University of Maribor. I believe Slovenian colleagues would agree that Nenad's role in constituting that department and its analytical orientation was enormous. Parallel to his work in Maribor, he became engaged at the Central European University in Budapest (CEU). Finally, and certainly not least importantly, Nenad played an enormous role in creating the Department of Philosophy at the Faculty of Humanities and Social Sciences in Rijeka. Although he was not employed there full-time, he remained an external associate until the end of his life.

Regarding Nenad's personal philosophical development, his period of learning began right after high school, when he spent the school year 1969/70 in Chicago. There he studied philosophy and classical philology (ancient Greek). From personal conversations, I got the impression that his stay in Chicago did not leave him with the fondest memories. After returning, he studied in Zagreb, where he experienced some more disappointments. His true philosophical ascent occurred during his postgraduate studies in Paris. At that time, postmodern and poststructuralist heroes were just articulating their philosophical positions. Foucault, Deleuze, and Derrida were real stars whose seminars he attended and privately discussed with them. He had especially close contact with Althusser. This engagement with postmodernism resulted in books Marxism and Post-Structuralism, The Speech of the Other, Essays in Philosophical Hermeneutics, *and* White Noise, *which had an exceptional impact on the domestic philosophical public. However, after this youthful intoxication, a sobering followed. In a brief autobiographical note Nenad provided a concise but perhaps the most illustrative sketch of the poststructuralist deconstruction technique: "You slip under Schelling's skin, poke at his metaphors, reverse and ironize his story, add a bit of psychoanalysis and attribute various things to him to show that he is actually, through no fault of his own, a victim of metaphysics. This amused me for seven or eight years, but after a while, it started to seem somewhat dishonest to me. [...] I didn't like it, it was mean, and great philosophers didn't do this."*[1]

A break with postmodernism and deconstruction followed, and he became acquainted with analytical philosophy. He familiarized himself with it gradually, reading and introducing himself to this different (compared to previous engagements) and difficult subject matter. At the same time, he continuously participated in significant analytical philosophy conferences, presenting and building his own philosophical name and position. Parallel to his own improvement, Nenad spared no time and energy to encourage others to engage in philosophical work and their personal advancement. The result was a series of collections co-authored with less experienced authors, who paved their own paths through these activities, following Nenad's guidance.

Nenad engaged in and left a significant mark on many disciplines characteristic of analytical philosophy. To a superficial connoisseur of Nenad Miščević's work, he is a philosopher of language. However, he was equally knowledgeable in almost all philosophical disciplines. This breadth of interest and activity can be followed through texts published in the domestic and international journals. From around 1985 onwards, he published a series of articles in Croatian Philosophical Investigations *(and* Synthesis Philosophica*), mainly in the fields of philosophy of lan-*

[1] S. Prijić-Samaržija and P. Bojanić (eds.). 2012. *Nenad Miščević – All Faces of Philosophy*. Belgrade: University of Belgrade, The Institute for Philosophy and Social Theory, p. 13.

guage and philosophy of psychology (i.e., philosophy of mind). After that, roughly from 1990 forward, an interest in epistemological topics is visible. This can be found in articles in Dometi *and again in* Philosophical Investigations. *A significant area of Nenad's work is epistemology and the philosophy of mathematics (in* Acta Analytica *and the Slovenian* Scientific Journal*) to which he contributed original theoretical theses. This line of interest, roughly speaking, can be followed from 1995 onwards. Simultaneously with his interest in the philosophy of mathematics, his published texts show contributions to the philosophy of science. What Nenad gained special recognition for globally is the philosophy of politics, particularly his contribution to understanding the phenomenon of nationalism. In 2001 he published a text on nationalism in the* Stanford Encyclopedia of Philosophy *and the book* Nationalism and Beyond *(2001). His other important books are* Rationality and Cognition *(2000),* Philosophy of Language *(2003),* Nationalism: The Ethical View *(2006),* Curiosity as an Epistemic Virtue *(2020) and* Thought Experiments *(2022).*

Apart from scientific achievements it is especially important to highlight Nenad's unparalleled impact on promoting the reputation of Croatian philosophy on the international philosophical scene. This reputation was created directly, through numerous connections that Nenad built with leading contemporary philosophers, through numerous conferences and symposia he participated in, and by organizing visits of many philosophers to domestic institutions. Indirectly, it was created through the work of people who gained philosophical recognition under Nenad's mentorship, as well as through the activities of institutions established due to Nenad's efforts.

Recognizing the full breadth and reach of Nenad Miščević's philosophical work, we are only gradually becoming aware of the void his departure has left, first in the lives of those of us who knew him well and socialized with him, and then in the entire cultural, especially philosophical, space to which we belong.

NENAD SMOKROVIĆ
University of Rijeka, Rijeka, Croatia

Farewell to Nenad

Dear Nenad,
There are many metaphors for Life: Life is a race, Life is school. The most common one is that Life is a struggle, but my favorite is that Life is a journey. For your and my forty or more years of life as colleagues and friends, I can say that our shared life was a beautiful journey. Over those long forty years, we traveled together, both literally and metaphorically. Literally, you from Rijeka and me from Split, when we were both hired at the Faculty of Philosophy in Zadar almost at the same time. That's where we first met... we were young! I remember the moment when you appeared at the door of my office and said: "I am Nenad... and I am Dunja..." and as usual (to which I later got used to), you asked for a book that I might have. I don't remember which one. And you immediately invited me to give a lecture at your department of Philosophy. At that time, you had already left behind your previous phases of philosophical life with Derrida, Foucault, as well as your professors in Zagreb, and you entered, as you said yourself, into the "labyrinth of analytic philosophy," and that's where we immediately bonded professionally because my interest has always been primarily language. Literal physical travel was complemented by conversations, reflections, discussions, debates, and trips to symposiums, mostly in Dubrovnik, at the IUC, and abroad.

Our journey and stay in Zadar, which I could call "romantic," were brought to an end by the fall of Yugoslavia, war and the war years. We were still traveling because we had to work even though we were all together in the shelter in the basement of the Faculty of Philosophy, under gunfire all around, without water and electricity. Many, including myself, feared for the lives of their children. But as if that wasn't enough! In 1991, the faculty leadership "wonderfully" decided that certain departments should definitely (!), I can freely say, be eradicated, the philosophy department being their primary target because there were so-called "political saboteurs" influenced by the Croatian Liberal Party. You and your assistants were their first targets as representatives of "false liberalism and Western democracy and whatever-kind-of analytic philosophy..." You were expelled, and I left shortly after.

But let's get back to our journey. It ended in Zadar, with a stormy northern wind, bura*, as we would say in Dalmatia. But the journey didn't stop or end. First, with the financial help of our dear colleagues, Michael Devitt and George Rey, we traveled to Aix-en-Provence in 1993 for the European meeting on Analytic Philosophy, the largest gathering of European philosophers of analytical orientation. You were elected*

*president on that occasion. I was happy and proud because it was a sign of respect for you as the main promoter of analytic philosophy in the Balkans, especially in Croatia. Our journey quieted down a bit. It leaned more towards the southern wind (*jugo*) when the sea threatened subtly and when there was a bit of a sharp* maestral. *We had to fight for establishing the Department of Philosophy in Rijeka against the stubborn and unfounded resistance of the Ministry of Science in Zagreb. Perhaps the metaphor "Life is a struggle" rather than a journey fitted better in this period. The students were the loudest. Let me remind them a little, although many present know. Students shouted: "I think, therefore I can't study." "Political censorship of science." "Another attempt to marginalize Rijeka." "We want our professors." "I want to study philosophy in Rijeka now." But as Bob Dylan sings,* Times are a-changing. *Victory was eventually achieved, and the best analytically oriented department of philosophy in Croatia exists in Rijeka.*

We found our new home together in Maribor, thanks to Bojan Bostner. Now we didn't commute to work anymore, I from the south, and you from the north like in Zadar, but our journey became completely mutual. We went to Maribor together in my already quite old Seat every week. And so, for 20 years, every week! In the car, we listened to music, mostly operas that you liked. In it, we devised all sorts of things: how to improve teaching, which symposiums to organize, planned trips, how to bring the greatest philosophical names like Davidson, Pietroski, Ludlow, Yagisawa, Jeff King, Stephen Neale, and many many others to Rijeka and Maribor. But it was also important how to spend evenings with our dear colleagues from Maribor.

Intellectually, for me (and I'm sure for others), you were a tour de force. *You taught me that in philosophy, as in life, clarity, honesty, rationality, and creativity go hand in hand. We discussed many topics, especially in the philosophy of language, always from a naturalistic standpoint. In that regard, you didn't change my mind. I followed your footsteps with great enthusiasm but also opposed them. We were a real pair! Just as we were in this journal – the* Croatian Journal of Philosophy *– for twenty years, you as the Editor and me as the Advisory Editor.*

And what were you like as I remember you: Honest to the end. As we would say in Dalmatia: Drito u sridu *(straight to the point). You said: "We don't have to hide behind metaphors, stylistic circuses, erudition if we have it (and yours was immeasurable). We can say what we think directly to someone's face."*

Yes, you were straightforward. You called the Ministress (I won't name her now) who forcefully wanted to abolish the Department of Philosophy "a puppet on the string" (completely deservedly). For her assistant, Mr. Z., you wrote that he was "one of the gravediggers of the Rijeka philosophy department." I didn't hold back either... I wrote: "Mr. Z., where were you when it was thundering?" You were extremely argumentative, with a mild irony. We all remember your columns in the newspaper Novi list

and lately in the Novi list *supplement* Vox Academiae. *There are countless examples, but here's one more philosophical: "In 2010, the question of* a posteriori *knowledge was coming into fashion, while* a priori *was going out of fashion. Luckily, in these parts, we're always late, so we're always just ready for the next, reverse, phase."*

You always loved to talk and discuss. In the car with you, I mostly listened and learned a lot. I recently found out that they sent you around the classrooms when you were in the first grade to tell the story of Ciplić Njuškalić (Snoopy Bream Fish). And that was the first story I've ever read! About that little fish that quietly and fearfully left the safe harbor and sailed out into the world. It seems that you and I started a journey that I try to describe long before it is documented here!

And now, at the end of our shared journey, there is no bura *or* jugo*, or* maestral *but only unreal calm sea, Kvarner* bonaca. *I ask myself and I ask you too, as our favorite Croatian poet Danijel Dragojević phrased it in one of his last poems:*

I won't ask if you're still alive somewhere out there
Maybe you are, maybe you're not,
Maybe yes, maybe no.
Let the unspoken question
swing in doubt...
Without it, you're at any station, on all sides,
in freedom for all or nothing...
("Question" – by Danijel Dragojević)

And I'd add:
You are here in our hearts forever!
Thank you, Nenad, for everything.
Let the heavenly birds travel with you now!

Read at the memorial in Rijeka, May 16th, 2024.

DUNJA JUTRONIĆ
University of Split, Split, Croatia

A Goodbye to Nenad

Farewells are always difficult. Especially when saying goodbye to someone with whom you have had a strong connection for almost half a century. Therefore, I begin in an unusual way, in a personal tone. My first memory of Nenad goes back to the final year of my studies in Ljubljana when I came across a brown softcover book titled Marxism and Post-structuralism. *What a discovery! A new insight into the current development of French philosophy. Masterfully written. Then seven years passed. Alive meeting in Ljubljana. A congress in memory of Locke, where I presented my first serious philosophical paper. And Nenad was already commenting on some of my claims. Then, in the IUC course* Philosophy of Science *in Dubrovnik, where Nenad introduced me to a circle of important analytical philosophers. The first conference in Zadar followed, opening doors for us to the world of analytical philosophy because Simon Blackburn, then editor of* Mind*, was there. Then Nenad's first arrival and lectures in Maribor (it was forty years ago in April). We organized two major conferences under the title* Science and Philosophy*, in which analytical philosophers from ex Yugoslavia (Nenad Miščević, Neven Sesardić, Miša Arsenijević, Saša Pavković, Andrej Ule, Matjaž Potrč) played a significant role. The first conference in Rijeka (Faculty of Economics and Nenad Smokrović), where I gave a co-lecture on Nenad's paper. Joint trips to the Wittgenstein Congress in Kirchberg. Congresses in Radgona and Bad Radkersburg. All these events deepened our connection. Then there was the breakup of Yugoslavia. In Maribor, we finally got the opportunity to open a philosophy study program. Nenad received a Slovenian half-year scholarship to contribute as a prominent expert to the development of philosophy and humanities at the University of Maribor. He got to know Maribor and made a decision. On October 1st, 1993, he began regularly teaching in Maribor. Together, we planned the future development of the Department of Philosophy.*

New colleagues arrived—today's professors. Nenad was not only a doctoral mentor to many but also an advisor and friend who, with his insight, erudition, and sound rationality, always found the right way out of puzzling situations in both philosophy and personal life. I must emphasize that the Department of Philosophy, with Nenad's immense help, was recognized on the world map of analytical philosophy by the end of the 1990s. During this time, Nenad also became the founder of philosophical studies at CEU and taught there until CEU moved to Vienna.

Nenad was always a visionary. However, everyday obligations often slipped out of his hands. Fortunately, Dunja (Jutronić) and I were always there to solve the arising problems. A kind of guardian angels, but unfortunately not omnipotent. Once Nenad comforted me when things went wrong: "You know, not everything is in our power. Don't worry."

I will conclude this moment with the words of two poets we both admired and, as Nenad described: "Tin Ujević is closer to you, Bojan, since you are a Bohemian at heart, but you don't want to show it, and for me, it's Dane Zajc, because I am a true avant-gardist."

In foreboding, in longing, distances, distances;
in the heart, in the breath, mountains, mountains.
There, there to travel,
there, there to grieve;
To no longer know myself,
nor the smoke of pain in the mists.
("Departure" – by Tin Ujević)

There comes a time when there is no more time.
A step stops and cannot move forward.
Time when you stop.
When you yourself are the ice.
Your time.
("Your Time" – by Dane Zajc)

Nenad, I hope you are now, in some other world, enthusiastically debating with the philosophers you have always loved.

To his wife Vera, daughter Heda and her family, I express my sincere condolence in my name and on behalf of my colleagues at the Department of Philosophy, the Dean of the Faculty of Humanities, and the Rector of the University of Maribor.

Read at the memorial in Rijeka, May 16th, 2024

BOJAN BORSTNER
University of Maribor, Maribor, Slovenia

Book Symposium on Nenad Miščević, Thought Experiments, *Springer 2022.*

Introduction

A book symposium was held on Nenad Miščević's important new book, Thought Experiments, in April 2022 at the annual Philosophy of Science meetings in Dubrovnik. The participants were James Robert Brown, David Davies, Marko Grba, and James McAllister. Nenad replied to these contributions.

The participants thought the topic and the discussions to be of sufficient interest that they agreed to write up their oral presentations for publication.

Since the symposium we learned the sad news of Nenad's death. Unfortunately, it seems that he was not able to write up his oral comments. Each of the participants has added some relevant remarks about Nenad more generally. He was someone we all liked and admired greatly. He had a tremendous influence on us all and will be missed.

JAMES ROBERT BROWN
Toronto, June 2024

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.1
Received: September 19, 2022
Accepted: June 26, 2024

Miščević: Mental Models and More

JAMES ROBERT BROWN
University of Toronto, Toronto, Canada

This is a review discussion of Nenad Miščević's stimulating new book, Thought Experiments (2022). His mental models account is of great importance in the various current debates about the nature of thought experiments. I discuss some of the pros and cons of his account.

Keywords: Thought experiments; mental models; intuitions.

Thought experiments (TEs) are remarkable devices for producing knowledge. Physics and philosophy are full of them, and it would be hard to imagine either discipline progressing as they have without a heavy dose of the kind of imaginative thinking produced by TEs. Galileo's ship, Einstein's elevator, Schrödinger's cat, and a great many more have played a central role in the development of physics. Plato's cave, Leibniz's mill, Putnam's twin earth, the trolley problem have similarly enriched and shaped the course of philosophy. In his new book Nenad Miščević offers a justification that I think we can all endorse. "Thought experiments are indispensable. Philosophy does not use a laboratory to test its theories; the only experiments available here are those in thought. TEs play in philosophy the crucial role that laboratory experiments play in science. Philosophers are vitally interested in *connections between our spontaneous understanding of important items, like meaning and content of our thoughts, and the results of science*" (2022: 87).

Many questions arise. How do TEs work? What are the different kinds? Why do some disciplines have a lot of TEs while others have few or none? The central question is this: How is it possible to learn something new about reality merely by thinking?

Nenad Miščević has an answer: *mental models*. His account can be found in various of his papers and now in his stimulating new book, *Thought Experiments* (Miščević 2022). Mental models, he claims, can address all (or most) issues concerning TEs. He introduced this ap-

proach in a talk about 30 years ago (subsequently published as Miščević 1992, simultaneously with, but independently from Nancy Nercessian (1993)). I fondly remember the occasion. It was in Dubrovnik in the siege during the violent breakup of Yugoslavia. One could come into the city or leave only by a daily ferry from Rijeka. The Inter-University Centre where the annual philosophy of science conference was held had been bombed and was largely destroyed. So, we met in temporary surroundings elsewhere in the old city. At night we heard machine gun fire. Snipers in the hills helped to focus the mind.

Nenad's mental models account is extremely plausible. As the term suggests, we form a model in our heads then read off the details that are consequences of the model. One of the strongest pieces of evidence for this account comes from our ability to make inference almost instantaneously. Imagine a turtle on a log. A fish swims under the log. Is the fish under the turtle? We immediately say yes, because we can see it in our mental model. A rival account of thinking would have us make inferences (deductive or inductive) from the given premisses. The trouble is that it takes several slow steps to get to the conclusion that the fish is under the turtle. This makes the mental model account much better at explaining how we actually reason in a wide variety of cases. And it makes a great deal of thought experimenting easy to understand as simply being instances of mental modelling.

At one point Nenad remarks, "It is Kant whose account of 'construction in intuition' comes closest to the mental model view" (2022: 61). This might need some explanation, since Nenad is a naturalist and a liberal empiricist, so there could be some tension. But this is a point I will not pursue. Instead, I will note the contrast with my own view. I think that (some) intuitions should be understood as producing genuine new knowledge. This is not a construction in imagination, nor an examination of our concepts, but rather a kind of perception of something existing independently from us. Such an account is anathema to empiricists and naturalists. Serious intuitions involve a kind of intellectual grasp, seeing with the mind's eye.

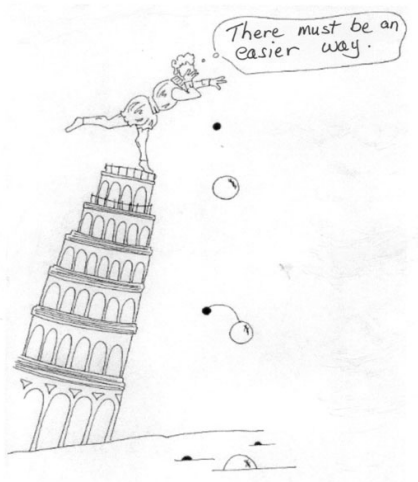
Nenad argues that: "the mental modelling theory and the 'voice-of-competence' proposal can account for most, perhaps even all, puzzling phenomena tied to thought experiments and intuitions" (2022: vii). Evolution comes to the rescue: "The evolutionary, adaptationist hypothesis offers a hope that a part of our primitive intuitional knowledge does reflect the deep make up of our environment, and thus, in spite of its fallibility, carry information about real and philosophically important properties of some states of affairs in the world" (2022: 68). Nenad also says, "Whereas Brown thinks that intuition capacity is a basic capacity, I prefer to think of it is a derived capacity that employs various basic capacities, prominently reasoning and quasi-perceptual imagination in the off-line fashion" (2022: 73). Moreover, he adds, "Intuitions should be studied as any other sources of cognition; one should search

for already known capacities and try to account for intuitions starting from them, instead of ad hoc postulating new capacities" (2022: 74).

Of course, it is difficult to disagree. As a general rule we should not introduce anything new, including new cognitive mechanisms, when we can account for everything with the equipment we already have. Here is a simple example. Those who fish sometimes marvel at the ability of some who seems to know where the fish are. We might say they have great fishing intuitions. I have no such intuitions, nor had my father. It turns out there is a simple empirical explanation for why some people do so well. They have tacit empirical knowledge of the situation. A fast flowing river will create eddies, pools of slow moving water, say, around a large rock. A trout will lurk in such a region because it requires relatively little energy to stay in place. If it stays near the seam of the two regions, the fast water will be a moving buffet, bring food for the hungry fish. None of this might be consciously noticed but it is all empirically absorbed by the alert fisher. Most of our unexplained intuitions will have an empirical source like this. Most – but not all.

I think there are cases, albeit quite rare, where we would be very hard pressed to give a naturalistic account of our intuitions. I will give two examples of this, one from physics and the other from mathematics. The first is obviously a thought experiment; the second is next of kin.

First, a word of explanation. Nenad has introduced useful terminology to cover this. An IET is an imaginative exercise in thought. It covers thought experiments and more, and would include the mathematics example I am about to present. I resist the desire to define thought experiment; I prefer a characterization that sets rough boundaries but does not try to make them precise. A definition can come at the end of inquiry. This is how we treat all sorts of important concepts. Religion and democracy, for instance, are not precisely defined, yet we can rationally discuss them. As for thought experiments, I only want to insist that they be performed in the mind and have an experiential character. We might ask about what the tides would be like, if there were 25 moons instead of one. We cannot "see" the answer; we would need to calculate. So, I would not call that a thought experiment, though others often do. On the other hand, some visual reasoning in mathematics might not be a thought experiment, but it is next of kin. Nenad's term, IET, captures this nicely. Now to my two examples of intuitions that produce genuinely new results.



Galileo first noted that Aristotle and common sense claim that a heavy object such as a cannon ball falls faster than a lighter object such as a musket ball (H > L). From this, it follows that combined cannon and musket balls would fall faster than the cannon ball alone. (H+L > H)

However, the lighter musket ball would tend to slow down the heavier cannon ball with the result that the cannon ball alone would fall faster than the combined object (H > H+L). Thus, we have a contradiction. Aristotle and common sense must be wrong. Galileo was able to resolve the situation by simply having all objects fall at the same speed (H = L = H+L). In other words, all bodies fall at the same rate, regardless of their weight.

This is a truly remarkable result. It is certainly a prime candidate for *a priori* knowledge. Why? There are unquestionably empirical concepts involved, such as weight and falling. But experience did not give us the result; that took the thought experiment. In fact, there was no new experience that moved us from Aristotle's to Galileo's view of falling bodies. The result is not derived from previous experience. Nor is it any kind of logical truth. After all, objects could fall at different rates based on their colour. Those who recall the rise and fall of the fifth force will remember the main claim that different rates of fall would depend on chemical composition. In any case, thanks to this example it can be plausibly claimed that we have *a priori* knowledge of nature. This is something no empiricist or naturalist can entertain.

My second example is from elementary number theory. What is the sum of the first n numbers? A theorem answers this question. The standard proof of this theorem is by mathematical induction, a technique that everyone takes to be a legitimate proof. A diagram is generally considered illegitimate as evidence. Of course, a picture can be pedagogically useful and perhaps helpful in suggesting a legitimate proof, but it is not thought to be acceptably rigorous.

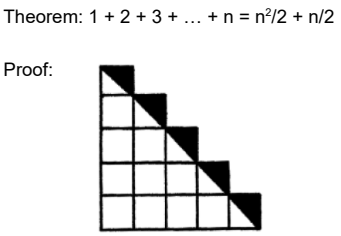


Figure 2. *Picture proof of a theorem.*

Spend a moment on the picture to see how it works. If you need a hint, here it is: Starting from the top add the squares, 1 + 2 + 3 + 4 + 5. Imagine this is a 5 x 5 square. Cut it in half with a diagonal. This represents $n^2/2$. Now restore the half squares (blacked out) that were removed by the diagonal. This represents $n/2$.

After studying the example, you should be persuaded of two things. First, the picture proof is just as rigorous as a proof by mathematical induction. And second, thanks to the first point, intuition is essential. This will be obvious when you realize that the proof holds for every number n, all infinitely many of them, even though the actual diagram is only for the number 5. Needless to say, there are different kinds of intuition, most are compatible with empiricism. Many of Nenad's uses of the term involve cases such as Putnam's twin earth. Here intuition means something like common sense judgement, which is based on empirical experience. As I mentioned earlier, I have no quarrel with these uses and quite agree on their empirical respectability. It is the rare kind that are not empirically respectable that I claim exist. The picture proof and the Galileo case are examples.

I take the Galileo thought experiment and the picture proof of the number theory theorem to demonstrate the existence of genuine knowledge-producing intuitions. I call them platonic intuitions. Such intuitions are not at work in every thought experiment, only a few. We reason about other cases in a variety of ways, as I acknowledged when asserting my pluralism about TEs. Some of these use mental models, just as Nenad claims. In fact, there are a large number of things on which we agree. I should mention some of these, as they are important. The first – and I want to stress this – is that I like Nenad's mental models account very much. It is probably the most popular account of TEs, and for good reason. My own view is often misunderstood, since I embrace intuitions and a generally platonistic outlook. In fact, to repeat again, I am a pluralist about thought experiments. I think Nenad is right about lots of them. I think John Norton, whose view is at the other end of the empiricist-rationalist spectrum from mine, is often right, too. Real experiments work in lots of very different ways. It should come as no surprise that the same is true of thought experiments.

One thing that Nenad took up that is otherwise little discussed is the difference between thought experiments in philosophy and in the sciences. We agree that in some broad sense they are the same kind of thing. Of course, philosophers of science talk about thought experiments in the sciences while regular philosophers focus on thought experiments in ethics, language, mind, and so on. But this is not what I have in mind. There is often a difference in methodological approach. Nenad put his finger on it: "One issue that has been prominent in the discussion is the contrast between 'extroversion' and 'introversion': are intuitions concerned with their external objects, the domain of items and facts, or with our concepts? Is Galileo investigating the falling bodies, or of the concept of the falling body? My sympathies are with external reference. Concepts often play a role in the process, but they are not the object of intuitions, and their role is subordinate to the role played by the external referential domain" (2022: 25). This is a hugely important point and I wholly agree with Nenad. Of course, it is important to know how language and our various concepts work, but ultimately, we are concerned with how the mind-independent world works.

Incidentally, I think every philosopher of science would also agree. This is one of the obstacles to overcome in finally getting something like a unified account of TEs in philosophy and the sciences. Like Nenad, I think that thought experiments are the same in both disciplines, but when some are focusing on *things* and others are talking about *concepts of things*, it can be difficult.

Nenad answers a question I have often raised hoping for an answer. Why are some disciplines more likely to use TEs than others? In particular, why does chemistry have so few, possibly none? Nenad puts it this way and provides an answer: "Why don't we normally have very reliable intuitions about chemistry? A natural answer is that chemical knowledge is not part of our folk theories, and that chemical reactions are not accessible to us to the degree physical reactions are. Therefore, there are no relevant assumptions that a thought experimenter might use. The [mental] models view offers a direction for explaining the phenomenon; I wonder whether the Platonist has anything comparable" (2022: 62).

No, probably not. Nenad's explanation sounds plausible to me initially. But I hesitate to embrace it because it skirts close to a view of TEs held by Daniel Dennett. Dennett has long been a critic of TEs for several reasons. One of these is his claim that TEs rest on folk science. We should expect the world to be very different from our folk conceptions, he says, and therefore, we should really give them no heed at all. We face a two-part problem: first, according to Nenad, we need folk concepts, which we don't have in chemistry, then, according to Dennett, we should reject folk concepts as fundamentally misguided. Consequently, if we need folk concepts but they are misguided in principle, then thought experimenters are right out of business. I think both of these claims are wrong, especially the later.

TEs use concepts at hand. Often these are folk concepts, but they needn't be. TEs are frequently constructed at a high level in physics with very sophisticated concepts. They would not be intelligible to the untrained (the folk) and might only be understood after years of study. Rather than thinking that folk concepts are required, it would be better to say that *familiar concepts* are required and that this can include highly sophisticated concepts that have been internalized by the thought experimenter so as to be second nature.

The second point, which is Dennett's, not Nenad's, is that folk concepts are misleading or useless. Not so. Some folk concepts might turn out to be of great scientific value. For instance, Galileo's ship example uses everyday concepts about the motion of a ship and our typical experience inside and outside the ship. It lead to the principle of relativity in both Newtonian physics and special relativity. For a second example, consider Turing's analysis of computability. His account of what is now known as Turing machines is often said to be a thought experiment. I won't describe it here except to say that a very simple, readily understood mechanism leads to some spectacular results. One of these is that most functions are not computable. In both of these cases, folk concepts have given us spectacular results that we now consider among our most sophisticated beliefs.

Nenad's new book is rich in detail and powerful in defence of mental models. His mental models account has become one of the most important and influential accounts of TEs, arguably more popular than any other. *Thought Experiments* will reinforce this opinion. It is a richly rewarding contribution to our better understanding of TEs in particular and how we learn about things in general.

Unfortunately, I must end on a sad note. Recently Nenad died. He was a wonderful friend, interested in everything and with an opinion about it no matter what it was. Every discussion was lively, funny, and included a touch of scurrilous gossip. We shared a seriously left-wing outlook and shared similar views on religion and current politics. Most of all I shall miss future discussions on thought experiments. As I already noted, he (and Nancy Nercessian) were the first to propose the very popular and plausible mental models view of thought experiment. He was particularly insightful on political thought experiments. Again and again I found myself persuaded and always looked forward to the next encounter. The loss is hard to fathom.

References

Miščević, N. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6: 215–226.

Miščević, N. 2022. *Thought Experiments*. Cham: Springer.

Nercessian, N. 1993. "In the Theoretician's Laboratory." In D. Hull, M. Forbes, and K. Okruhlik (eds.). *PSA 1992*. East Lansing: Philosophy of Science Association, 291–301.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.2
Received: September 19, 2022
Accepted: June 26, 2024

Thought Experiment as Bridge Between Science and Common Sense

JAMES W. McALLISTER
University of Leiden, Leiden, Netherlands

This reflection on the recent work of Nenad Miščević on thought experiment pursues two themes. One is the congruence between the historical development of the practice of thought experiment in science over the centuries and the development of philosophical accounts of thought experiment. The second is the idea that thought experiment provides a point of contact between common-sense and scientific conceptions of particular phenomena.

Keywords: Common sense; mental modelling; science; thought experiment.

1. *Twin histories*

There is not just a single history of thought experiment, but two. History 1 is the record of the rise and use of thought experiment in natural philosophy and science over the centuries. This history includes, among its high points, the classic thought experiments proposed by Galileo Galilei, Isaac Newton, and Albert Einstein. History 2, by contrast, is the succession of accounts thematising and analysing thought experiment as a distinctive device in scientific practice. This history consists of theories of the philosophy, epistemology, and methodology of thought experiment. It includes landmark contributions by such writers as Alexandre Koyré and Thomas S. Kuhn, as well as the revived debate among philosophers of science since the 1990s (Stuart and Fehige 2021).

The relation between these two histories presents an oddity. We expect the history of philosophy of science to mirror the history of science in various ways, of course: the former is, in part, a reflection on con-

ceptual changes and methodological innovations in the latter. In most cases, however, philosophical accounts of a facet of science do more than merely recapitulate that facet: they account for it at a higher conceptual level. In the case of thought experiment, by contrast, the relation appears more mechanical: history 2 simply reiterates history 1, it seems. Every conception of thought experiment put forward in history 2 is seemingly already present in history 1.

Here are some examples. Roy A. Sorensen (1992) in history 2 proposed a philosophical account of thought experiment as a species of concrete experiment: in history 1, natural scientists of the eighteenth and nineteenth centuries progressively incorporated a mature practice of thought experiment into a broader experimental methodology. John D. Norton (2004) in history 2 analysed thought experiments as arguments with suggestive premises: Aristotelian natural philosophers in history 1 constructed a variety of arguments *secundum imaginationem*, consisting of imaginative and counterfactual reasoning. James R. Brown (2011) in history 2 proposed a Platonist account, according to which some thought experiments allowed us to apprehend laws of nature: in history 1, Galileo used thought experiment to portray the laws of the new mechanics as evident and indubitable. The same holds, lastly, for my own contribution. I have argued that thought experiment acquires evidential significance only on certain metaphysical assumptions: where these assumptions are not accepted, thought experiment is evidentially inert. I have been able to find many examples in history 1 of researchers outside nomothetic domains who declined to lend evidential significance to thought experiment for this reason (McAllister 2018).

Why do accounts of thought experiment in history 2 seem fated to repeat what instances of the use of thought experiment in history 1 already offer? One possible explanation is that philosophers in history 2 have seen their task as clarifying, endorsing, and justifying examples of thought experiment that they found in history 1. That sounds unlikely, however: philosophers usually set themselves more ambitious goals.

A more intriguing hypothesis is that history 2 parallels history 1 on this topic because the two explore the same space of conceptual possibilities. There are only so many possible conceptual structures for thought experiment, and both histories exhaust them. This hypothesis gains plausibility in the light of the special role of thought experiment in theorising in philosophy. Philosophical analysis of other evidential devices in science—laboratory experiment, fieldwork or computer simulation, say—does not itself consist of laboratory experiment, fieldwork or computer simulation. Philosophical analysis of thought experiment, by contrast, consists largely in thought experiment—that is, in imaginative modelling of possible uses of the device in reaching conclusions. If thought experiment were restricted to a limited set of conceptually coherent options, then it would not be surprising if this framework constrained both history 1 and history 2.

This suggests that there are two ways of developing the philosophy of thought experiment, and thereby extending and enriching history 2. One way is to continue the project of creating accounts that explain and justify yet further individual examples of thought experiment found in history 1, clarifying their epistemology and methodology. Many writers have pursued this project, as we have already seen. The second way is to survey and elucidate the overarching space of conceptual possibilities that the device of thought experiment inhabits in both history 1 and 2.

Nenad Miščević in his book, *Thought Experiments*, makes a contribution to both these projects. Miščević's primary aim is to present a specific account of thought experiment, thus occupying a particular place in the conceptual space. In passing, however, he also offers a valuable suggestion about the space as a whole that instances of thought experiment inhabit.

The first project takes off in chapter 3: Miščević critically reviews some previous accounts of thought experiment, including inferentialist, Platonist, and Kantian theories. From chapter 4 onwards, Miščević develops his own alternative proposal in this repertory. This is a mental modelling account of thought experiment. In particular, Miščević argues that the function of thought experiment is to prompt a researcher to activate and draw upon unarticulated (and perhaps inarticulable) cognitive resources. Some of these resources may be innate, whereas others are the accumulation of our experiences of the world.

Miščević's thinking along these lines stretches back over thirty years, and his ideas have stimulated wide debate (Miščević 1992; Borstner and Gartner 2017). The new book adds much detail. For example, Miščević now suggests that the performance of a thought experiment traverses seven stages: these start with retrieving an unarticulated intuition, and conclude with identifying the significance of the thought experiment for our broader understanding of the world. This schema amounts to a practical guide for performing thought experiments (Miščević 2022: 17–22).

I suggested above that every conception of thought experiment that philosophers put forward in history 2 is already found in history 1. This holds for Miščević's conception too. Its counterpart in history 1 is an iconic thought experiment in mechanics, featuring a *clootcrans* or "wreath of balls," which Simon Stevin proposed in 1586. Stevin used this thought experiment to conclude that a chain draped over a frictionless prism would not slide off in either direction, and thence to derive the condition for the balance of forces on inclined planes (Dijksterhuis 1955: 176–179).

Miščević returns to Stevin's thought experiment several times in the course of the book. The example is particularly apposite for Miščević, for two reasons. In general terms, it is an instance of mental modelling: Stevin asks us to picture the dynamics of the chain in our mind. On a more specific level, Stevin's reasoning in the thought experiment

turns on the principle of impossibility of perpetual motion: this appears suddenly as a premise in the course of the argument, as if the thought experiment had prompted the natural philosopher to recall it at the appropriate step. This illustrates what Miščević describes as the tendency of thought experiment to activate implicit cognitive resources. In Miščević's words:

> Stevin's TE and the resulting intuition that the chain will not move, deploys some spatial-geometrical knowledge that might be innate and in this sense a priori, some naïve physics that is partly innate (a priori) and partly derived from and justified by experience (a posteriori), and we can trace each of the lines of justification to its distinctive source. (Miščević 2022: 25–26)

All this fits together well. In fact, however, Miščević has not only a counterpart in history 1, but also a precursor in history 2. Ernst Mach also propounded a mental modelling account of thought experiment. Mach hypothesised that a scientist had a store of intuitive knowledge laid down from previous experience:

> Everything which we observe in nature imprints itself uncomprehended and unanalysed in our percepts and ideas, which, then, in their turn, mimic the processes of nature in their most general and most striking features. In these accumulated experiences we possess a treasure-store which is ever close at hand and of which only the smallest portion is embodied in clear articulate thought. The circumstance that we are easier able to employ these experiences than we are nature itself, and that they are, notwithstanding this, free, in the sense indicated, from all subjectivity, invests them with high value. (Mach [1883] 2013: 28)

Thought experiment allowed the scientist to tap into this store and retrieve items of knowledge that were suited to tackling a particular problem, according to Mach. Furthermore, Mach too took Stevin's chain thought experiment to illustrate this conception, and presented a detailed analysis of it (Mach [1883] 2013: 24–31). Since both Miščević's theory and his understanding of Stevin's thought experiment recall Mach quite closely, it would have been interesting if he had contrasted his views in detail with those of Mach; in fact, he touches on the similarity only briefly (Miščević 2022: 31).

2. *Bridge function*

I see in Miščević's book also a contribution to the second project that I identified above, namely the investigation of the overarching conceptual space in which thought experiment operates. Rather than striving to add to our stock of individual models of thought experiment, this project attempts instead to identify the range of possibilities that accommodates all such models.

Miščević locates this conceptual space between the domains of science and common sense. Since antiquity, philosophers have been intrigued by the existence of two forms of knowledge: everyday, practical conceptions of the world that people share widely and apply in concrete

cases, and specialist, formal or technical conceptions that are the product of systematic investigation and reasoning within disciplinary settings. A particular question has concerned the relation between these two forms of knowledge. Should they be seen as separate domains, or is there some point of contact between them?

Miščević's intriguing proposal is that thought experiment acts as a link between everyday and technical modes of knowing:

> Why is […] a TE indispensable? Because philosophers are vitally interested in connections between our spontaneous understanding of important properties […] and the results of science. In order to answer the question about the relation between, say, scientific determinism and our belief in freedom, we need to confront them, and we cannot do it within science alone. We need the bridge, and TE is a perfect candidate. (Miščević 2022: 28)

The example of free will is well chosen. This concept features prominently in both domains: common sense includes well-entrenched assumptions about human freedom to make decisions and take actions, while physics and the neurosciences advance theories about the degree to which human acts can be explained by—and thus be reduced to—more basic causal factors. If we are to develop a unified view of this domain, then these two discourses must communicate: insights from science may refine and correct common sense, but it is also important that the view put forward by science speak to our everyday experience (Nahmias 2014). Thought experiments about free will are able, as Miščević suggests, to provide a bridge between these two discourses.

If this proposal is to contribute to the second project that I identified above, of systematising the overarching conceptual space of thought experiment, then it must provide a framework that is demonstrably more encompassing than individual models of thought experiment are, and sufficiently flexible to do justice to a wide variety of them. Miščević's proposal is capable of meeting this challenge. To appreciate this, we need only note that there are many different ways of—and purposes for—building a bridge between common sense and science: different examples and models of thought experiment correspond to these different possibilities.

Consider the following instances. We may wish to forge a link between science and common sense by spurring science to take up and resolve puzzles suggested by everyday intuition. This is the function carried out by Einstein's light beam thought experiment. Second, we may wish to test scientific theories against criteria of acceptability rooted in common sense. This is what Galileo's falling body thought experiment does. Third, we may wish to probe the implications of particular scientific theories for everyday conceptions of the world—Erwin Schrödinger's cat thought experiment in quantum theory does this. Many further alternatives can be devised.

Miščević's suggestion, that what is common to all instances and models of thought experiment is a capacity to bridge the gap between science and common sense, is an original and powerful contribution

to elucidating thought experiment in its variety. It is more than that, though. It is also a novel and convincing answer to the debate about the relation between science and common sense that has endured since Arthur S. Eddington's "two tables" parable (Eddington 1928).

Eddington intended his parable to highlight the incommensurability between the dominion of common sense, in which a table was solid, sharply bounded, and coloured, and that of science, in which a table was none of these things. Philosophers over the decades have been divided about the most convincing response to Eddington. Some have embraced eliminativism, holding that only one of the two worlds genuinely exists; others have postulated priority of one over the other. Miščević, by contrast, succeeds in placing the two domains on the same level by the simple and flexible notion of constructing a bridge between them.

Tamar Szabó Gendler (2007) has already gone some way in this direction, albeit for philosophical rather than scientific thought experiments. Gendler pointed out that discussion of a philosophical problem may take very different forms and elicit differing intuitions depending on whether it is based on a description of an abstract and general state of affairs, or on a portrayal of a concrete and particular scenario. An abstract and general description is the typical centrepiece of scientific conceptions of the world, whereas concrete particulars are more often the object of common-sense conceptions. Gendler ascribed to philosophical thought experiments the function of linking and comparing these two conceptions, somewhat similar to that which Miščević attributes to scientific thought experiments.

To my mind, the greatest value of Miščević's book is to be found in his contribution to this second project, even more than in that to the first. His arguments for the mental modelling account of thought experiment will be received with interest by philosophers inclined to a cognitive science approach to scientific methodology. However, I find Miščević's idea about the functions that thought experiments play regardless of the particular epistemology that we attribute to them, creating a link between the domains of science and common sense, to be of greater significance and originality. It will be a pleasure to see to what further insights and developments this intriguing suggestion gives rise in years to come.

Acknowledgements

I presented a previous version at the 47th Annual Philosophy of Science Conference, Inter-University Centre Dubrovnik, April 2022. I remember Nenad Miščević (1950–2024) with affection and gratitude for invariably interesting and friendly discussions of thought experiment and other topics in Dubrovnik over many years.

References

Borstner, B. and Gartner, S. (eds.). 2017. *Thought Experiments between Nature and Society: A Festschrift for Nenad Miščević*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Brown, J. R. 2011. *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. Second edition. Abingdon: Routledge.

Dijksterhuis, E. J. (ed.). 1955. *The Principal Works of Simon Stevin*. Vol. 1. Amsterdam: C. V. Swets and Zeitlinger.

Eddington, A. S. 1928. *The Nature of the Physical World*. Cambridge: Cambridge University Press.

Gendler, T. S. 2007. "Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium." *Midwest Studies in Philosophy* 31: 68–89.

Mach, E. [1883] 2013. *The Science of Mechanics*. Translated by T. J. McCormack. Second edition. Cambridge: Cambridge University Press.

McAllister, J. W. 2018. "Historicism and Cross-culture Comparison." In M. T. Stuart, Y. Fehige, and J. R. Brown (eds.). *The Routledge Companion to Thought Experiments*. London: Routledge, 425–438.

Miščević, N. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6: 215–226.

Miščević, N. 2022. *Thought Experiments*. Cham: Springer.

Nahmias, E. 2014. "Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences." In W. Sinnott-Armstrong (ed.). *Moral Psychology*. Vol. 4, Freedom and Responsibility. Cambridge, Mass.: MIT Press, 1–25.

Norton, J. D. 2004. "On Thought Experiments: Is There More to the Argument?" *Philosophy of Science* 71: 1139–1151.

Sorensen, R. A. 1992. *Thought Experiments*. New York: Oxford University Press.

Stuart, M. T. and Fehige, Y. 2021. "Motivating the History of the Philosophy of Thought Experiments." *HOPOS: The Journal of the International Society for the History of Philosophy of Science* 11: 212–221.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.3
Received: September 19, 2022
Accepted: June 4, 2024

Miščević On Thought Experiments

DAVID DAVIES
McGill University, Toronto, Canada

I address two claims that Miščević makes in his book Thought Experiments. The first claim is that literary fictions belong to the broader category of what he terms "Imaginative Enactments in Thought" (IET's), but are not TE's properly understood. The second claim is that TE's are indispensable to analytic philosophy. Both claims appeal to Miščević's discussion in the opening chapter of what it is for something to be a TE. I argue for the following conclusions: (1) If TE's are defined in the way that Miščević proposes, then there can in fact be (and indeed are!) works of literary fiction that qualify as TE's. (2) If TE's are defined in this way and are explained in terms of mental models, then whether there can in fact be analytic philosophy without TE's depends upon how we understand the relationship between TE's and counter-factual thinking more broadly construed.

Keywords: Thought experiments; fictional narratives; mental models; analytic philosophy.

Foreword

It is very sad that Nenad's untimely passing has deprived us of what would, I am sure, have been his very lively responses to these papers exploring themes in his wonderful book *Thought Experiments*. But I am very pleased to include, in this commemorative issue of the *Croatian Journal of Philosophy,* a brief paper that celebrates some of Nenad's insightful and valuable contributions to the literature on thought-experiments, contributions that I, like many others, have learned from and drawn upon in my own work.

1

The centrepiece of Nenad Miščević's very interesting book *Thought Experiments* is the further elaboration and defence of his 1992 account of how we are able to learn from thought experiments (TE's) in both science and philosophy. Carving out a middle ground between the pessimistic views of the empiricists—where the best we can hope to get from thought experiments is deductive arguments in sheep's clothing—and the heady views of the Platonists, Miščević has argued that, when we "run" a TE, we are able to activate and draw upon unarticulated and/or unarticulable cognitive resources, some of these innate and some the unarticulated residue of our experiential engagements with the world. He draws here upon Johnson-Laird's idea (1983) that the construction of "mental models" is a crucial part of our comprehension of narratives.

But Miščević's book advances his earlier thinking on these matters in at least two important ways. First, stressing the analogies between real experiments and TE's, he analyses the cognitive work of a TE into a number of distinct stages. The first five stages incorporate the conception and formulation of the TE, and its initial reception resulting in an "intuition" on the part of the receiver. The further stages incorporate processes of (a) "intuitive induction", where we gauge the more general import of the TE through comparison with other related TE's, and (b) seeking "reflective equilibrium", where the import of the TE is determined by locating it in the broader framework of our understandings of the world. Citing Stevin's famous "chain" TE, Miščević notes that "scientists, philosophers and teachers know that [engaging with the narrative] is not the end of the story: one can and should vary the story and generalize the result, and then test the intuition and generalization, comparing them to other spontaneous intuitions and generalizations, or even to information from psychology of belief-formation" (Miščević 2022: 9).

This analysis in terms of stages serves two roles in Miščević's response, in chapter 6, to the challenges to the cognitive status of TE's that have come from experimental philosophy. First, although Miščević does not stress this point, it seems to follow that the intuitions evoked by TE's have cognitive value only when the TE's are elements in the kind of broader investigative practice that the "stages" model describes. Second, analysing the workings of TE's in terms of the "stages" model allows us to identify different places where our intuitions might be untrustworthy and, thereby, to consider measures that might render TE's more epistemically reliable. Both of these points are of special importance for analytic philosophy, Miščević maintains, because TE's are indispensable for the latter. Finally, Miščević argues (chapter 5) that, to properly understand how TE's work in philosophy we need to view them diachronically, as the means whereby philosophical thinking in a given field may develop through engagement with and development of a powerful TE. He develops this point at some length, taking as his

principle example the manner in which thinking in the philosophy of language and of mind has developed in different ways in response to Putnam's original Twin Earth TE's.

2

I have always found the "mental model" view of TE's in its various incarnations an attractive one. It preserves, with philosophically modest resources, our sense that TE's can have genuine cognitive value. It also solves nicely Kuhn's puzzle (1964) as to how we can acquire new knowledge of the world without new empirical input. We can do so, it is claimed, because, in constructing a mental model in our comprehension of the narrative of a TE, we are able to draw on otherwise inaccessible understandings of the world that we already possess. I think Miščević does an excellent job of deepening and expanding his earlier published defence of the "mental model" account in this book. So I shall not be questioning Miščević's general positive account of TE's.

What I do want to address, however, are two further claims that Miščević makes, one in the opening chapter of the book and the other in his account of the role to be accorded to TE's in philosophy. The first claim is that literary fictions belong to the broader category of what he terms "Imaginative Enactments in Thought" (IET's), but are not TE's properly understood. The second claim is that TE's are indispensable to analytic philosophy. Both claims appeal, directly in the first case and indirectly in the second, to Miščević's discussion, in the opening chapter, of what it is for something to be a TE. My two critical reflections will take this discussion as premise and argue for the following conclusions:

(1) If TE's are defined in the way that Miščević proposes, then there can in fact be (and indeed are!) works of literary fiction that qualify as TE's.

(2) If TE's are defined in this way and are explained in terms of mental models, then whether there can in fact be (analytic) philosophy without TE's depends upon how we understand the relationship between TE's and counter-factual thinking more broadly construed, an aspect of Miščević's account of TE's that perhaps needs further clarification.

3

In specifying what he takes to be the constitutive features of a TE, Miščević contrasts his own view with Mach's somewhat expansive account. According to Mach, "the planner, the builder of castles in the air, the novelist, the author of social and technological utopias is experimenting with thoughts; so, too, is the hard-headed merchant, the serious inventor and the enquirer" (Mach 1976: 29; cited in Miščević 2022: 10). Miščević does not question the interest of this grouping, but

proposes that we view it as a broader genus—"Imaginative Enactments in Thought"—of which "strict TE's" of the sort that we find in science and philosophy are a species. The latter "have as their primary purpose increase of knowledge" whereas the other kinds of IET's listed by Mach have "a different primary motivation".

One kind of IET that Miščević wishes to exclude from the class of strict TE's is works of narrative fiction such as novels and films. He cites my piece on "Art and Thought Experiments" in the *Routledge Companion to Thought Experiments* as following Mach in using the term TE "in a very wide sense" so as to include such artistic fictions (Miščević 2022: 11). He then argues that, while the latter may have *some* cognitive function, their primary function will be either to achieve artistic ends of an expressive or formal nature or to induce enjoyment or other kinds of affect. In defence of his exclusion of artistic fictions from the realm of strict TE's, he further claims that, in such fictions, "the requirements of strictness are weaker than in TE's. In science and philosophy the TE should have a clear and univocal goal, and the proposal that is tested by it has to be decided in a non-ambiguous way. In a literary work ambiguity is often praised as a goal" (Miščević 2022: 11).

Let me note first that, in my piece in the *Companion*, far from following Mach's profligate employment of the term "thought experiment", my use of the term agrees in all essential respects with Miščević's. My aim in that piece was to assess the extent to which—as other authors such as Catherine Elgin (2007), Noel Carroll (2002), and James Young (2001) have claimed—at least some artistic fictions *meet* Miščević's requirements for being strict TE's. According to these authors, at least some literary or cinematic fictions are IET's whose primary intended purpose is to increase our knowledge or understanding. The authors in question further claim that, as a result, at least some works of artistic fiction have significant cognitive value. They thereby espouse some form of what is usually termed "literary cognitivism". In my piece, drawing on a couple of earlier articles (Davies 2007, Davies 2010), I argue that the first claim is correct but express significant reservations about the second claim.

These reservations obtain because a defender of literary cognitivism must meet certain empiricist challenges analogous to those that Miščević surveys in his overview of empiricist criticisms of the cognitive credentials of TE's in science. A representative sample of the kinds of challenges confronting the literary cognitivist can be found in Jerome Stolnitz's paper (1992) "On the cognitive triviality of art". Stolnitz begins by suggesting that we cannot learn anything interesting about the world through reading works of fiction because the supposed "truths" in such works are generally banal and imprecise. All we might hope to learn from reading Jane Austen's *Pride and Prejudice*, for example, is that "stubborn pride and ignorant prejudice keep attractive people apart," and even here it is unclear what the scope of this claim

is. To the response that this fails to do justice to the general truths about the world that may be gleaned from works of fiction, Stolnitz responds that, even if there *were* genuinely interesting truths about the world exemplified in the narratives of works of literary fiction, we couldn't *learn* those truths in our engagements with those works of fiction because the work provides no *empirical support* for such putative truths. All we are given in the fictional narrative is a single non-real example which has been gerrymandered to make those "truths" apparent. Echoing empiricist critics of TE's in science, Stolnitz maintains that the best we might get from reading literary fictions is interesting hypotheses that might then be subjected to independent empirical test.

The literary cognitivists cited above respond to this kind of challenge by arguing that at least some literary works function as extended TE's, and can therefore share in the kinds of cognitive value ascribable to TE's in science (Elgin) and philosophy (Carroll). In critically discussing this strategy on the part of literary cognitivists, I have pointed out (see especially Davies 2010) that the strategy can serve cognitivist aims only if we counter the empiricist criticisms of TE's in the latter domains. In fact, a model of TE's like the one defended by Miščević seems to be just what the literary cognitivist needs. If the running of a scientific or philosophical TE can yield genuine knowledge of the world—without the need for independent empirical testing—because the TE draws on genuine but unarticulated, or unarticulable, cognitive resources, then, if literary fictions are TE's, surely the same can apply to them, and Stolnitz's objections are answered.

Literary cognitivists have generally assumed that their case is made once it is granted that some literary fictions are TE's, but even if one supplements the cognitivist's case with something like a "mental model" account of TE's, there are still issues that need to be addressed (see again Davies 2010). Miščević's "stages" model, in fact, provides further reason to be sceptical about the literary cognitivist's claims, since the consumption of literary fictions does not seem to be part of a larger practice of consuming and testing TE's, and it is, according to this model, the location of our running of TE's within such a practice that confers cognitive credibility upon the intuitions they evoke.

But the issue of present concern is whether at least some works of literary fiction can meet Miščević's requirements to count as "strict TE's", and here I think the answer must be a positive one. The requirement, we may recall, is that the principal aim of the narrative be a cognitive one: the primary purpose should be to increase knowledge, and, with this in mind, the "lesson" of the TE should be clear and not trade in ambiguity. Perhaps fittingly, we can show that this requirement can be met by means of a (philosophical) TE! Let us imagine two literary authors—let us call them Edward and Graham. Suppose that Edward, in a literary essay, expresses the view that our moral duties to our friends should outweigh our moral duties to our country. When

Graham hears of this, he strongly disagrees and undertakes to demonstrate how duty to country can, at least on occasion, outweigh duty to friends. He does so by writing a literary fiction where, when the reader grasps the genuinely conflicting nature of the duties to friend and country confronting the main protagonist, her intuitions will accord with those of the protagonist when the latter decides to weight duty to country over duty to friend. The motivation for composing the fictional narrative in this case is clearly cognitive, and there is no attempt to make the situation ambiguous in any relevant respects. Thus, by Miščević's criteria, we have a work of literary fiction that is a strict TE.

In fact, we do not need to appeal to a TE to make this case. For, at least on some accounts, what we have described in hypothetical terms was what actually led Graham Greene to write his novel *The Third Man* (1950) to counter a claim about how to balance moral duties to friend and country voiced by E. M. Forster in his essay "What I believe" (Forster 1938/1951). And it is not difficult to find other examples of works of literary fiction whose primary aim is cognitive in this way. The original edition of Anthony Burgess's *A Clockwork Orange* (1962), for example, is an extended IET intended to explore the moral issues surrounding the treatment of social deviance. Here again the purposes motivating the construction of the narrative are clearly cognitive in the manner required by Miščević. But it is *not* sufficient to meet Miščević's criteria that the author of a literary fiction works with the elements that define a philosophical issue: Carl Reiner's film *All of Me* arguably takes as its basis the kinds of hypothetical cases that drive debates about the place of embodiment in our sense of personal identity, but the aim of the film is clearly to entertain rather than enlighten the viewer (for a discussion of this case, see Smith 2006).

We thus have examples of existing literary fictions that (1) have as their primary purpose the increase of knowledge or understanding, (2) are not intentionally ambiguous, and (3) are, if Johnson-Laird's "mental model" account of narrative comprehension is correct, comprehended through constructing a mental model. They thereby fit Miščević's description of a "strict" TE in chapter 2 of his book: "We have characterized a TE as a process that starts with a design, which involves the determination of the goal(s), in particular the thesis/theory to be tested, and the construction of a scenario to be considered. We noted that it then proceeds with the presentation of the scenario thus constructed to the experimental subjects. On the side of the subject, the experiment then continues with the typically imaginative contemplation of the scenario plus some piece of reasoning, culminating in the decision ("intuition") concerning the thesis/theory to be tested."

4

In the final section of this paper, I want to at least raise some questions about Miščević's claim that TE's are "indispensable" for analytic philosophy. We find an argument for this claim, at least with respect to practical philosophy, in the following passage: "The traditional sources of insight here are either facts (including presumed facts), principles or TE's. Facts are useful and indispensable, but taken alone they don't teach us about what is valuable, morally prohibited, morally indifferent and so on. We need at least principles. But how do we test principles? The only source here are intuitions and the indispensable testing grounds are TE's" (Miščević 2022: 26). As he later puts this, for philosophy "TEs are indispensable. Philosophy does not use [a] laboratory to test its theories; the only experiments available here are those in thought....Although life without TEs might be possible for science, it is practically impossible for philosophy" (Miščević 2022: 87, 98).

We might reformulate this argument as follows:

(1) The claims that philosophers seek to evaluate are *modal* in the sense that they are not just claims about how things actually are but about how things must be, or can't be, or ought to be.
(2) To evaluate a modal claim, we need to engage in counter-factual reasoning.
(3) To engage in such counterfactual reasoning is to entertain a thought experiment.
(4) So philosophy cannot do without TE's.

Points (1) and (2) seem valid if we restrict ourselves to attempts to *defend or establish* a modal claim. To defend a general modal claim is to maintain that it would lead to the right results in possible as well as actual cases, and to assess a possible case requires counterfactual reasoning. It might seem that the cases brought *against* such a claim could be actual cases and would therefore not call for counterfactual reasoning: in countering the claim "all A's must be B", we might point to an actual A that is not B. It might be responded that we will still need counterfactual reasoning to establish that we have a genuine counter-example to the universal claim. But rather than pursue this issue, I want to look at the move from (2) to (3) and (4).

As we saw, Mach understood the idea of a TE very broadly—it includes any process of working out in one's head how to proceed in a given instance, where this necessarily involves considering various options and thus counter-factual reasoning. On this account, the merchant in the market who deals with a customer trying to haggle for a cheaper price is engaged in a TE. Miščević is critical of this broad construal of TE's, but this is on the grounds that a TE must have a primarily cognitive purpose. But does Miščević hold that, *as long as this further condition is satisfied*, any instance of counter-factual reasoning is a TE? Suppose we term such a view the "cognitively motivated

counterfactual reasoning" (CMCR) view of TE's. While the CMCR view seems required if (4) is to follow from (1) and (2), it also raises a number of questions:

(i) The CMCR view will incorporate many examples of counterfactual reasoning in philosophical and other contexts that we would not normally think of as (philosophical) TE's of the sort discussed throughout Miščević's book. This *broad* conception of TE's would resemble the one that Miščević ascribes (2022: 43) to Buzzoni according to which "TEs are the condition of the possibility of REs because, without the a priori capacity of the mind to reason counterfactually, we could not devise any hypothesis and would be unable to plan the corresponding RE that should test it" But Miščević seems sceptical about Buzzoni's approach.

(ii) If Miščević is operating with the CMCR view of TE's, it is difficult to make sense of his sympathetic response to Williamson's account, which clearly rejects the CMCR view. Indeed, both Williamson (2016) and Miščević seem concerned to distinguish TE's from cognitively motivated counterfactual reasoning more generally. Miščević cites here Williamson's discussion of the hunter who deliberates about whether to attempt to ford a stream by jumping across it at its narrowest point. What distinguishes such a case from counter-factual reasoning more generally, for Williamson, is the hunter's use of imagination, something that cannot be replaced by more abstract reasoning. Miščević develops this idea by proposing that the imagination here serves a particular role, namely, the construction of a mental model of the counterfactual situation. On the mental-modelling approach, TEs are sophisticated "re-modellings in the head" whose most important feature "is precisely their concrete and quasi-spatial character" (Miščević 2022: 47). This strongly suggests that for Miščević only those cases of counter-factual reasoning that have these distinctive features of mental modelling count as TE's, contrary to the CMCR view. But in this case, it seems, we cannot derive (4) from (2).

(iii) However, certain other remarks by Miščević seem to place him closer to the CMCR view. For example, in discussing the distinctive features of mental models, he states that "'TE's might involve language-like representations and inference and computation on them, but *typically*, they involve more concrete representations, such as are used in imaginative operations" (Miščević 2022: 53, stress added). This seems to erase the distinction that Williamson is trying to draw in his appeal to the use of the imagination in TE's as contrasted with other more formal kinds of counter-factual reasoning. On the other hand, in another puzzling remark which seems to indicate a departure from the CMCR view, Miščević claims that

> a TE need not involve counter-factual reasoning because some cases considered in a TE can be real" (Miščević 2022: 46). One wonders here whether, for such cases to count as TE's, they must in fact involve counter-factual reasoning about the real case. If not, why think of them as TE's rather than imaginative engagements with an actual case, as occurs in the mental modelling of a non-fictional narrative. Also, this seems to conflict with the claim that "thought-experimenting involves proposing and considering counterfactual scenarios (Miščević 2022: 44).

These are issues upon which I am sure Miščević would have provided further clarification and enlightenment had he been able. But they are issues that only present themselves because of the intellectually engaging aspects, as described earlier, of Miščević's overall enterprise in this very interesting book.

References

Burgess, A. 1962. *A Clockwork Orange*. London: William Heinemann.

Carroll, N. 2002. "The Wheel of Virtue: Art, Literature, and Moral Knowledge." *Journal of Aesthetics and Art Criticism* 60 (1): 3–26.

Davies, D. 2007. "Thought Experiments and Fictional Narratives." *Croatian Journal of Philosophy* 19: 29–46.

Davies, D. 2010. "Learning through Fictional Narratives in Art and Science." In R. Frigg and M. Hunter (eds.). *Beyond Mimesis and Convention: Representation in Art and Science*. Boston Studies in the Philosophy of Science 262. Dordrecht: Springer, 51–70.

Davies, D. 2018. "Art and Thought-experiments." In M. T. Stuart, Y. Fehige, and J. R. Brown (eds.). *The Routledge Companion to Thought Experiments*. London: Routledge, 512–25.

Elgin, C. Z. 2007. "The Laboratory of the Mind." In W. Huerner, J. Gibson, and L. Pocci (eds.). *A Sense of the World: Essays on Fiction, Narrative, and Knowledge*. London: Routledge, 43–54.

Forster, E. M. 1951 [1938]. "What I believe." In *Two Cheers for Democracy*. New York: Harcourt Brace.

Greene, G. 1950. *The Third Man, and the Fallen Idol*. London: William Heinemann.

Johnson-Laird, P. N. 1983. *Mental Models*. Cambridge: Harvard University Press.

Kuhn, T. 1964. "A Function for Thought Experiments." Reprinted in *The Essential Tension*. Chicago: University of Chicago Press, 1977, 240–265.

Mach, E. 1976. *Knowledge and Error*. Dordrecht: Reidel.

Miščević, N. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6 (3): 215–226.

Miščević, N. 2022. *Thought Experiments*. Cham: Springer.

Smith, M. 2006. "Film Art, Argument, and Ambiguity." *Journal of Aesthetics and Art Criticism* 64: 33–42.

Stolnitz, J. 1992. "On the Cognitive Triviality of Art." *British Journal of Aesthetics* 32 (3): 191–200.

Williamson, T. 2016. "Knowing by Imagining." In G. Currie (ed.). *Knowing Through Imagination*. Oxford: Oxford University Press, 113–126.

Young, J. 2001. *Art and Knowledge*. London: Routledge.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.4
Received: September 19, 2022
Accepted: June 26, 2024

The Mystery of Intuition in Einstein's Thought Experiments

MARKO GRBA
University of Rijeka, Rijeka, Croatia

The role of intuition in understanding in general and in scientific understanding in particular is still very much a subject of a lively philosophical discussion. The role of intuition in thought experimenting is much disputed in its own right, and the arguments range from those that deny any substantial role of intuition in the final inference that the thought experiment is meant to illustrate (eg. Norton or Williamson) to the pivotal role some form of intuition might play (eg. Brown or Miščević). So far, mostly Platonists were defenders of intuition, but in his recent book, Miščević takes on a formidable task to mount a defense of intuition as seen from a naturalist-evolutionist point of view and within his mental-modelling approach to thought experiments. I will, while acclaiming certain – and considerable – merits of his approach, nevertheless, insist that certain aspects of intuitive comprehending as it is meant to be going on in the process of thought experimenting remains inexplicable in the naturalist scheme such as Miščević's. The more mysterious (not to say Platonist) aspects of intuition will, hopefully, be revealed through the analyses of the two very famous thought experiments of Einstein which also figure quite importantly in his scientific opus. I will also have something to say about a few related problems as addressed by Miščević in his book regarding the description of thought experiment and more general imaginative enactments in thought, as well as on whether there is an essential difference between scientific (primarily physical) and metaphysical thought experiments and other thought experiments or related modes of thinking.

Keywords: Einstein; thought experiments; intuition; Miščević.

1. *The merits of Miščević's approach*

Miščević's new book on *Thought Experiments* (2022) is a most welcome addition to the growing literature on an important aspect of thinking in the natural sciences but also, more broadly, in theoretical and practical philosophy. The book is unique in being intended as a broad as possible an account of different theories of thought experiments (TEs) on offer in philosophical literature as well as of other related modes of thinking, from metaphysical TEs (Descartes' demon) to literary fiction (SF-stories for example), political utopias or dystopias, or even religious meditations (for example Ignacio Loyola's). For the whole lot of these mental modelling schemes which he, following Ernst Mach, sees as congenial, Miščević proposes a most ingenious phrase of *imaginative enactment in thought* (IET) (2022: 11). Thought experiment is then seen more specifically in the following manner:

> A typical TE starts with a design, which involves the determination of the goal(s) in the thesis/theory to be tested, and the construction of a *scenario to be considered*. It then proceeds with the presentation of the scenario thus constructed to the experimental subject, either the author of the scenario, or an interlocutor. In the later situation, the testing is done independently of the author: she is supposed to sit and wait for the verdict from the interlocutors, i.e. experimental subject's '*laboratory of the mind*'. On the side of the subject, the experiment starts with understanding of the proposed scenario, and then continues with the typically imaginative contemplation of it. Some reasoning might intervene. If all goes well, the subject ends with a verdict concerning the thesis/theory to be tested. Usually, in her mind it is presented to her as an invitation to believe or directly as a belief, most often seeming obvious and compelling. Such states (invitations to believe, or immediate beliefs) have been traditionally described as '*intuitions*'; they are often likened to similar states concerning mathematical insights or obviously looking linguistic judgments on sentences in subject's native language. Once the verdict is achieved, it can be and often is compared with results of other scenarios in the vicinity, or other versions of roughly the same scenario. Finally, interesting and provocative verdicts are normally being brought to comparison with items of knowledge or widely accepted beliefs. If they clash, the arduous task of balancing is required, in which the particular verdict might win (as has historically been the case with Galileo's verdict on falling bodies), or, alternatively, the established knowledge might, or, thirdly, some compromise is made. *The result is usually described as* '*reflective equilibrium*' (2022: 14, my italics)

Miščević is tying in one finely knit fabric a vast body of views and analyses found in literature, such as James R. Brown's (1991/2005) idea of a *laboratory of the mind* as the scene of thought experimenting, or John Rawls' *reflective equilibrium* of judgments, and presenting to the reader a unified picture of the whole realm of modes of thinking which have been used by various authors and to various purposes for millennia under one name and one guiding principle. As I take it, this guiding principle is to see how scientists, philosophers and authors are generally arriving at their ideas, more or less revolutionary, relying on their

intuitions and employing their imagination, perhaps (at least at times) more than their logical reasoning. This fits very well with what Einstein said about the respective role of intuition/imagination and logic in the context of discovery of the fundamental laws of nature:

> The supreme task of the physicist is to arrive at those universal elementary laws from which the cosmos can be built up by pure deduction. There is no logical path to these laws; only intuition, resting on sympathetic understanding of experience, can reach them. (Einstein 1919: 226)

Connected to the idea of a unifying approach to studying different modes of thinking in as diverse fields as physics and political theory, we might speculate on why certain ideas were historically seen as (more) revolutionary than others, say, Copernican revolution in astronomy as more revolutionary than Plato's epistemology, or on a par with the ideas of the French revolution. Could it not be that many a time ideas and, indeed, the values of ideas were judged more on the merits of their practical application, or potential for such an application, rather than on their intrinsic (theoretical) value? Although, Miščević is not likely (based on what I know form our conversations) to agree with Einstein's deductivist position (as espoused in Brown 1991/2005: 112–121) as to the methodology of science, or embrace a Platonist epistemology, nevertheless, his account is potentially broad enough to accommodate even such widely differing positions on the epistemological spectrum.

Reading Miščević's book, one could gain an impression that his intention was to write a sort of a guidebook on how to conduct thought experiments, given the detailed analysis of their structure or the breadth of examples and references. In many respects, I would say, one would not be amiss to take advice from this book. However, one must always take it with a pinch of salt, especially when it comes to how to understand intuition as such and what exactly it takes to reach a conclusion from a thought experiment. These are the issues I will now take on in the next two sections basing the discussion on two most consequential thought experiments of Einstein.

2. *Two conundrums regarding physical TEs (applicable to other scientific TEs)*

Miščević's account of IETs (2022: 67–68), includes model building, thought-experimenting and intuition-producing. Regarding the TEs as a subspecies of IET he demands that they are scenario-based rather than inference-based, that they produce intuitions as their final products in a process of mental modelling where various highly particularised scenarios are played as in front of our eyes and in the creation of which imagination of the experimenter (speaker/interlocutor) has a central part to play. He insists that such scenarios have both cognitive and justificatory role in TEs, whereas inference plays a subordinate role. Furthermore, he gives a pivotal role to intuition as having to do

with the external referential domain, not merely concepts. For him this intuitioning is largely innate and related to a specific competence(s) of the brain along the lines of the standard Chomsky's proposal. Miščević, however, goes a step further and generalises the specific lingusitic competence to other crucial competences when it comes to understanding and dealing with the world (such as spatial, temporal, numerical etc.). He does not claim that what intuitions share is primarily the underlying structure(s) in as much as it is the manner of representation. These competencies are ultimately regarded in an evolutionary-adaptationist way. This approach to understadning TEs, and more generally IETs, he names the *Moderate Voice of Competence* proposal (MoVoC):

> So, we now have the minimal necessary elements to formulate a proposal concerning the nature of intuitions and TEs producing them. I have called it Moderate voice of competence view ('MoVoC' for short). It starts from the admission that there are intuitions-dispositions and judgments, which form a distinct group of phenomena, and there is the intuition-capacity, the capacity to use our imaginative and judgmental competencies in an off-line fashion. It is the voice of competence, most often discreet. Intuitional data are thus the minimal 'products' of tentative production – linguistic, philosophical, moral or mathematical – by naïve thinker (or speaker-listener) and not their opinions about the data. The data involve no theory and very little proto-theory. Although there might be admixtures of guesswork in the conscious production of data, these are routinely weaned out by linguists. As against predominantly conceptualist understanding of TEs and intuitions (Peacocke, Boghossian) it claims that intuitions are concerned with their external objects, the domain of items and facts, rather than with concepts. Concepts often play a role in the process, but they are not the object of intuitions, and their role is subordinate to the role played by the external referential domain. (2022: 67)

Although I agree with the general framework, especially with putting the stress on the key part the imagination plays in TEs, the view that imagined scenarios have both cognitive and justificatory role as well as with assigning the intuition an external referential domain, as all these features seem to me prominent in scientific TEs, especially Galilei's and Einstein's, I would be somewhat hesitant in committing to the very narrow evolutionist-adaptationist account of the origin of intuition-capacity or the thought process as such. Given, first, that what we know on these matters is still mostly informed by research from psychology rather than neuro-science which is both more "naturalistic" as well as more accurate in its measurements, and hence conclusions than psychology will ever be, and yet does not really give us much to muse about at the present state of development. (One may wonder whether it ever will, given that some of the problems are related to the problem of *qualia*, which is notoriuosly difficult to solve from any point of view). Furthermore, even if we assumed that valuable intuitions which will have some bearing for the understanding of the world might be arising in special mental capacities pertaining to specific brain region(s), we could ask why the more sophisticated intuitions do not arise much

more often, as I believe it could be agreed on that the insights of the kind Galilei or, even better, Einstein had, arise in others even in a span of a century. If this were the case, our science would have been on a much more developed stage by now?

In sum, the first conundrum I see unresolved in Miščević's work so far (as in most of other authors) and not much commented on either, would be the origin or, (I guess) in Miščević's case, the mechanism of generating the more sophisticated intuitions. If I understood him well, in case of scientific (physical) TEs, the proposed source of these ideas to be identified with some sort of folk science/physics (as modelled on folk psychology, which in my above described view already makes it a problematic idea), the concepts/intuitions which are then clashed with the *accumulated wisdom* of the ages (2022: 64) and tempered by new (real) experimental data, simply will not do. Even Miščević agrees that the ideas of the folk sciences are usually fallible as: "Our innate geometry might be false, our possibly innate folk-physics certainly is" (2022: 65). Not to mention that it is hard to sometimes even formulate what the folk-scientific ideas would even be, say in the case of chemistry (as most humans do not perform that many relevant real chemical experiments to acquire a significant body of observations which could then be conceptualised in any meaningful way). In the case of physics, the situation should by no means be underestimated, given that almost all fundamental physics concepts are to a high degree sophisticated. There is nothing obvious or simple in any of the concepts we use in, say, Newtonian mechanics: such concepts as speed or acceleration already have both a scalar and a vector representation (which obviously assumes the knowledge of a sort of vector algebra); the ideas of motion, continuity of space and time or matter are debated since the pre-Socratic philosophers and still mostly unresolved. Galilei and Newton came to their fundamental postulates of the science of mechanics by a combination of highly sophisticated abstraction and pure guesswork (with a little bit of experimentation where the limitations of air-resistance or friction allowed it). But the real challenge is to try to account for any of the sublime thought experiments of Einstein by way of relating his new ideas to some form of folk physics, or folk mathematics, especially for the more consequential of his TEs. Some of the challenges will be presented shortly.

The second conundrum I see unaddressed is how exactly does the inference come about from the thought experiment as this again is by no means obvious. Especially so in the case of the sophisticated TEs like Einstein's. Miščević, in my view, provides persuasive arguments in favour of an intuitionistic view of TEs (and IETs) as opposed to inferentialist or conceptualist views, but I would like to have seen this relation of inference to the scenario of the thought experiment described in more detail as this is where the real trouble begins when it comes to interpreting the TEs or ascribing any value to them in the context of,

say, theory building such as was Einstein's regular practice. Norton's famous account (1991) skillfully avoids asking part of the epistemological question about the origin of this relation. He is only interested in spelling out the logical part as he much later admitted in a response to a criticism (Norton 2021: 125–126). The origin of the relation for Norton is resolved by assumption of identity, where *thought experiments are simply picturesque arguments*. But what about Einstein's words as quoted above insisting that: "There is no logical path to these laws; only intuition, resting on sympathetic understanding of experience, can reach them?"

3. *Process of discovery and process of justification in the case of Einstein*

I would argue that the two most consequential questions to be answered when it comes to interpreting the results of a TE are: *Which idea(s) do(es) the explaining?* and *How does one arrive at the idea(s)?* The second question is not only relevant in the context of discovery but could also be in the context of justification, to use the famous distinction by Reichenbach. It could happen, namely, that the path to discovery (the heuristics if one prefers) might be of a significance also as steps of justification, which is how Einstein often argued when trying to give an account of justification of his theories (including both special and general theory of relativity), as Norton convincingly argued in (1995). Most often (definitely in the case of Einstein's TEs) the idea(s) that actually serve(s) as explanans is/are quite subtle and unexpected (so it would appear that there is not much there in terms of folk physics, as Miščević demands it), to the point of being of inexplicable origin, or at least origin hard to trace. Two very famous TEs will be used to illustrate: Einstein's elevator and his light momentum TE with the help of which he derived $E = mc^2$. But before those, a word or two on the comparison of Norton's views to Miščević's as I believe some interesting thoughts might emerge.

Even a rationalist and inferentialist, when it comes to analysing the origin or structure of a TE, like Norton, admits (1995: 63) that a rationalistic account of the discoveries (and thought experiments) of Einstein leaves room for *arational*, in Norton's own words, elements and, as he puts it, *perhaps even Einstein's "free inventions of a human mind*." But the key question here, surely, is how much exactly in genius's (like Eintein's) process of discovery is rationally accountable and how much remains perhaps forever inexplicable, at least by a rationalist analysis? Of course, this is very hard to establish. However, it does matter a great deal for the following reasons.

First, if key ideas came to Einstein in some sort of an epiphany (much like to the mathematician Ramanujan in a dream, according to his own recollection passed to G. H. Hardy. This, of course, annoyed rationalistic and logical mind like Hardy's, especially given Ramanu-

jan's insistence that he needed no proofs for the mathematical propositions thus revealed). They were presumably unique, or at least quite specific to one mind, that of Einstein's. This means that it would be a gross oversimplification to claim that the subsequent rational analysis of the origins of these ideas is possible or even useful. To the contrary, one could, after reading such analysis, acquire a completely distorted picture of the real process (if there was any) and assume that if only one would follow the steps of the rational analysis, one could repeat the same kind of discovery, or achieve the discovery of the same calibre as some of Einstein's discoveries. Now, I am not arguing that no rational analysis is ever possible – far from it – but simply that more space and a more of an open mind should be left to the possibility of the contrary. The contrary could then be seen as either a Platonic insight of a sort, or a naturalistically founded intuition as understood by Miščević and described above. This would then be an argument in favour of Miščević's conception of the process of discovery in TE, but also a defense of Miščević's account of nature and value of a TE as against Norton's. However, I have an issue with Miščević's account of the discovery process as too narrow in not allowing for anything but a naturalistically understood intuition. But how then to account for the rarity of such deep insights as Einstein's? Surely, if evolution has programmed us for such deep thinking, then it must have programmed more of us, proportionally many more then the history of science would allow for (by which I mean the history of those ideas in science that have proved fruitful especially when it comes to the predictive power of natural sciences)! On the contrary it would appear, that Einstein was quite unique in his way of thinking as well as discovering.

The second objection to a thoroughgoing rationalist analysis of the type of Norton's is to my mind even more serious, if not even deeper. Namely, the objection that follows from the point raised by Einstein as quoted above, that *only intuition, resting on a sympathetic understanding of experience,* can reach deep insight into the fundamental laws of nature, which, as I claimed at the beginning of this paper would go in favour of Miščević, but not necessarily of his naturalism. For Einstein surely knew what he was talking about and his *dictum* was inspired by his own experience of working for decades at the forefront of research in foundations of physics, from particle physics to cosmology, and so his emphasis on intuiton as *opposed* to logic must have had some grounding in observing his own process of discovery. This insistence would appear to agree well with Leibniz's view of reasons of the world of physical phenomena which never necessitate but only incline (Russell 1937/1992: ch. 3), meaning that the connection of no two ideas in physics is logically necessary, hence it is not possible to discern such a connection by applying pure logic. It would be interesting to know what Miščević's thoughts were on this aspect of the problem of acquiring knowledge about the physical world.

Finally, I arrive at my perhaps most controversial point, which I am inevitably led to, especially after having spent some time assessing the merits and demerits of various accounts of how Einstein came to discover his general theory of relativity and this associated elevator thought experiment. My point can again be well posed as against Norton's claim in the above referred paper of 1995 (62–63) to the effect that the better the rationalistic reconstruction of the process of discovery is, the less mystifying the process appears and, consequentially, the more likely the steps of the process of discovery are to be also seen as the steps in justification or explanatory process, if this can at all be achieved (as Norton, I believe, justifiably claims Einstein himself was in the habit of doing, at least when it came to the theories of principle, as he called them[1]). The point is that if Norton is right that sometimes (at least in the case of some of the steps along the path of discovery of Einstein's theories of relativity) the process (or parts of the process) of discovery can be supplanted for the process of justification, so heuristics could be supplanted for logic. If so we should be extremely observant as to the details involved as is best seen in the case of Einstein using the so-called *equivalence principle* in discovering the general relativity which will now be briefly described as some of the best available accounts in the literature. The claim I will be led to is that in the case of using the equivalence principle (or principles as Einstein actually changed the meaning of his postulate on several occasions), and as most famously exemplified in his elevator TE, Einstein was in the end making up a just-so-story rather than presenting a genuinely valid argument or operational TE to support his quest for general relativity, although not doing it consciously, at least not at all times.

4. *Einstein's elevator TE and the equivalence principle as idée fixe*

Soon after completion of his special theory of relativity, which was Einstein's response to the most pertinent issue of the day, namely, the conflict between Newtonian mechanics which embodied the principle of relativity of all motions and Maxwell-Lorentz electrodynamics which seemed to suggest the independence of the speed of light of any source or direction of motion, Einstein embarked on an even more ambitious

[1] The theories of principle, as opposed to constructive theories, according to Einstein, are those that are founded on a minimal number of preferably empirically suggested basic principles (axioms) which do not assume anything about the structure of the material world, only state some universal properties of natural processes or their theoretical representations which then have to be cast into mathematical form (Einstein 1919: 228; Brown 1991/2005: 103–105). An example of such a theory of principle, after which Einstein modelled his theories of relativity, is classical thermodynamics with its main principles being the laws of thermodynamics (viz. the impossibility of building the perepetuum mobile of either the first (1st law of thermodynamics!) or the second kind (2nd law!).

quest – to generalise his theory. Although the special theory of relativity was a remarkable achievement in its own right, especially given the minimalist nature of its structure and the scarcity of experimental evidence at the time (early 20th century), it was a theory of a limited domain of application, applying only to systems in uniform non-accelerated motion, to motion of the so-called *inertial reference frames*. But Einstein sensed, rightly as it turned out, that the basic structure of the theory, what in mathematical terms would amount to invariants of motion with respect to a certain group of symmetry transformations and in physical terms would have implications for the way we represent spatial and temporal relations between phenomena, held a much bigger promise. However, this was initially only a pretty vague impression, although strongly present in his mind. For Einstein in his twenties (when he was developing his special theory) was not yet a fully trained mathematical physicist as he was to become during the work on his generalised theory, for which he had to develop a mastery of the latest developments in then-contemporary mathematics (such as the absolute differential calculus, or tensor calculus, of Ricci, Levi-Civita and Cartan). Indeed, he had to learn to appreciate the fact that further advances in ever more abstract theories of fundamental physics come (perhaps) exclusively at a very high price in terms of the mathematical knowledge requisite in their development (see eg. Norton 1995: 61–62).

As late as 1914, Einstein still had doubts about whether he should follow the path of mathematical simplicity and elegance or carry on in his familiar way through direct physical insight. Different authors see these internal quibbles as a consequence of Einstein before 1915 still being naive to abstract mathematics of his day, but one could, at least with a hindsight, see in these an originality of approach to physics as Einstein's characteristics, as perhaps one of only very few physicists or scientists of his statue. Namely, Einstein was hesitant to adopt the predominantly mathematical approach to problem-solving in the realm of physics as he was genuinely baffled by the ever-increasing demands in terms of the level of mathematical sophistication, which is usually accompanied by an appropriate increase in the level of abstraction, with every new and more subtle problem in physics In effect, Einstein was overawed by the ramifications of the relation between mathematics and physics. Rightly so! As I would dare say, whoever takes the complexity of this relation lightly usually pays the price of losing the compass as to what could exist in reality but which is not revealed in mathematics alone. And there would appear to be such an element in at least every mature physics theory. So Einstein was not wrong in being *prima facie* suspicious towards giving mathematics the predominant role in guiding the research in physics/science, but only extremely cautious. At his expense, as it ultimately turned out and is well known, since by 1916 he was able to complete the general theory of relativity which was actually a new theory of gravity, only by fully adopting all the sophisticated

mathematics which he could learn from his mathematician friends, some of which were among the greatest mathematical geniuses of all times (like Tullio Levi-Civita, Felix Klein, David Hilbert, Emmy Noether, Hermann Weyl and Elie Cartan, more or less in historical order of appearance in Einstein's professional life). The question, however, remains, what was Einstein relying on, if not highly abstract mathematical techniques, in deriving his conclusions in physics? The answer is also well known: primarily thought experiments!

As early as 1907, Einstein thought about generalizing his theory of relativity since he was naturally dissatisfied with it being applicable only to a very narrow domain of uniform inertial motion and was looking to extend the domain of application of, first and foremost, the relativity principle to all the relative motions. Einstein's original train of thought might have looked like this (see eg. Janssen 2014 with my own insertions here and there): given that in special theory no one's frame of reference could be thought as absolute (as one cannot prove its existence by, say, observing motion relatively to this frame) and that all the inertial motions are hence only considered as relative, one should expect that all the reference frames are equivalent (including the accelerating ones) and so any one could be deemed as at rest for a given observer. In essence it is to find the most general form of the laws of nature, independent of the choice of the coordinate frame. The task seems meaningful enough, indeed, something to be desired, as if there is no favoured frame of reference, then surely the fundamental laws being universally valid entails their mathematical formulation being coordinate-independent. The trouble is that the effects of non-inertial motion (such as tidal forces due to gravitating masses) are discernible for all the observers alike, whether moving with the frame or apart from it. At the time the only candidates for fundamental forces (to which all others would reduce in final account) were electrical, magnetic and gravitational forces. Since Maxwell showed electrical and magnetic forces to be two sides of the same unified electromagnetic force (or rather field), and given that magnetic force was shown by Einstein himself to be eliminable by a change of a reference frame, Einstein might have been inspired (we do not know this for sure!) to ponder upon the idea of somehow eliminating gravity as a force, or, rather, transforming gravity away and therefore transforming between inertial and non-inertial reference frames. Thereby, in the long run, perhaps achieving the transformation between any and all the reference frames as if any one of them could be at any moment seen as at rest. While still at the patent office in Bern, Einstein had, in his own words, *the happiest thought of his life* (as quoted in Janssen 2014: 174 and note 30): what if a man was falling with the elevator, would he not have the same experience as if in a state of weightlessness in a space without gravitational fields? By extension, an observer who is performing various observations in a stationary elevator in a gravitational field would have the same experi-

ence as the one who is moving in an elevator (in space free of gravitational fields) which is acted upon by a force different from gravity but acting so that inside the elevator everything appears as if there is no external force on the elevator but gravity. Does that not suggest a way to eliminate gravitational force and still have the gravitational effects?

As mentioned, Einstein's efforts towards generalization of his special theory started as early as 1907; at first by him trying to develop a special-relativistic account of Newtonian force of gravity, but that, he soon realised, was impossible given that Newtonian force assumed instantaneous action at a distance and special relativity implied relativity of simultaneity. Einstein struggled for several years with different versions of a special-relativistic theory of gravity as did some of his contemporaries (like Max Abraham or Gunnar Nordström) and managed to conclude that no such theory (either a (3+1)- or 4-dimensional) is possible. The interesting point from those early attempts is that Einstein used what he is to call the equivalence principle only later. At the time, this meant that Galilei's law of free fall holds, namely that all objects falling from the same height in a homogenous gravitational field fall at the same rate and that the vertical velocity of fall is independent of the horizontal component of motion, if there is such. The special-relativistic theories of gravity did not fulfill the second part of the statement (as well as the law of conservtion of energy which was by then taken as "sacrosanct" in physics). Now, Einstein was to relate to the Galilean law of fall another implication, namely, that of the equivalence of inertial and gravitational mass, which is a fact tacitly assumed in deriving the law of motion of a (point) mass under the influence of Newtonian force of gravity, which Newton too was aware of. But none of these on its own, or standing together, would enable Einstein to make any progress from special theory of relativity to the generalised form of any kind, as is clear from his elevator TE, which actually assumes much more, albeit not clearly expressed. Einstein, indeed, was aware of the limitations of merely coupling special relativity with the law of free fall, the equivalence of inertial and gravitational mass and having Newton's law of gravity as a limiting case of his new theory. And, yet, what else could he demand for his general theory to fulfil?

He played with various ideas, having a variable speed of light (sacrificing the second of the two postulates of special theory of relativity all to generalise the first), but all in vain, as it turned out. He then envisaged another of his thought experiments (Janssen 2014: 178-181), the rotating disk as a frame of reference. Wondering how an observer at the circumference would see the passage of a light beam sent from the observer at the centre of the disk towards the circumference, he concluded that although the light would travel the straight line path, it would not be perceived as such by the observer at the circumference due to difference in linear velocity of the two observers. Given that, after the elevator example, the rotating disk frame (with centripetal

force) is equivalent to a disk at rest with a centrifugal gravitational field acting on it, we are justified in concluding that gravity bends light, as indeed the final form of the general theory of relativity accurately predicts regardless of how bizarre either the conclusion in the first instance of the rotating disk or the transference of it to the disk at rest may at first seem. Now, the geometry of the rotating disk is meant to be Minkowskian as in special relativity, which enabled Einstein to make deductions, given its familiarity. The principle of equivalence (proclaiming the equivalence of effects as seen from the appropriate system in accelerated motion or the one at rest in a corresponding gravitational field) enables the transfer of deduction to another type of reference frame, namely the one in a gravitational field, hence giving Einstein an essential insight of the outline of the sought after general theory. It also may inspire, as indeed it might have inspired Einstein, according to Stachel (1989), to consider alternative geometries to Euclidean for the space-time structure of general theory (here Einstein would have considered contraction to measuring rods along the radius and circumference and from these deduced ratios of circumference to the radius of a rotating disk which might differ from 2π).

Even if Einstein has managed by, more or less, following the path described above to reach correct conclusions, there are certain conceptual problems which cannot be ignored, and, indeed, Einstein could not ignore them when discovering general relativity. The alleged equivalence which Einstein crucially relied upon in making his deductions valid for the case, first, of a homogeneous gravitational field, and then any gravitational field turns out to be extremely difficult to articulate, so much so that Einstein changed the meaning given to his principle on several occasions, after more than ten years finally reaching the mature form which aims to make any gravitational field as having only relative existence as one side of the so called *inertio-gravitational field*: "There is only an inertio-gravitational field that breaks down differently into inertial and gravitational components depending on the state of motion of the person making the call" (Janssen 2014: 178).

Now, the problem with this formulation, although Einstein claimed it was the key for discovering general relativity, was that it retroactively(!) sanctioned the inference of gravitational effects from accelerated frames with Minkowski space-time given that the very metric of Minkowski space-time would, according to mature equivalence principle, be taken as a particular inertio-gravitational field (Janssen 2014: 179). This means Einstein would have been able to make the deduction as found within the Minkowski frame valid in a frame at rest with gravity acting only *since* the metric of Minkowski space-time represents a form of gravity! There was never actually an equivalence between an accelerated frame and a frame at rest with gravity, but only between two types of frames with gravity. This could further be read as a curious case of a let's-pretend game (or just-so-story) that Einstein

was playing on himself for several years. Of course, it could only have worked if the metric can be identified with gravity, but here again we have several major issues to consider which Einstein was to become gradually aware of either through his own efforts or through constructive criticism from colleagues.

Einstein claimed for many years in his papers and correspondence with many authors/critics that gravitational fields should only have relative existence, that equivalence principle helped in a crucial way on his path towards the covariant field equations of general theory and that the general theory was a generalization of the invariant special theory of relativity in the sense of relativity of all reference frames and all motions. We now know that none of these actually holds, and that even the extent to which the equivalence principle really helped or hindered his research is not quite clear as seen from the writings of different authors (for instance Janssen in (2014) argues for more of a hindrance case, whereas Norton in (1985: 40) praises Einstein's use of equivalence principle *as one of the most beautiful of Einstein's insights*).

With respect to the claim of relative existence of gravitational fields (Janssen 2014: 178–179), it follows from Einstein's insistence that the gravitational field is to be related to the metric tensor not the Riemann curvature tensor of Einstein's field equations. This follows from his reinterpretation of special relativity theory as a theory of special case of gravitational field, namely the one generated by the Minkowski metric, which means that now even non-accelerated frames could be associated with a field. Thus any field could be thought of as related to a reference frame and transformable, therefore of relative existence. At first, and for a long time, Einstein thought this idea, coupled with what he in 1918 dubbed as Mach's principle,[2] could finally remove any trace of absolute motion from physics. In this he turned out to be wrong as the Dutch astronomer and mathematician Willem De Sitter managed to show after which Einstein gave up attempts at a Machian account of inertia, but not before introducing his (in)famous cosmological constant to the equations of gravitational field, which he later denounced as his *biggest blunder*. By 1954, in the final year of his life, Einstein wrote:

[2] As Einstein wrote to De Sitter in 1917 (quoted in Janssen 2014: 202): "It would be unsatisfactory, in my opinion, if a world without matter were possible. Rather, it should be the case that the $g_{\mu\nu}$-field is *fully determined by matter and cannot exist without the latter*. This is the core of what I mean by the requirement of the relativity of inertia." Which means that there is no field without matter which generates it and given that all motion is with respect to metric ($g_{\mu\nu}$), it is in actuality with respect to some constellation of masses as Mach originally conceived in response to Netwon's famous bucket experiment which had as aim proving the existence of absolute motion by example of rotation of a water inside a bucket even when the bucket does not move relative to the water, after being set in motion from rest with the initially still water and then stopped at the point when the water reaches the highest point of ascent of concave surface as against the walls of the container.

> In my view one should no longer speak of Mach's principle at all. It dates back to the time in which one thought that the 'ponderable bodies' are the only physically real entities and that all elements of the theory which are not completely determined by them should be avoided. (I am well aware of the fact that I myself was long influenced by this *idée fixe*). (Einstein to Felix Pirani, February 2nd 1954)

I claim that similar judgement could be passed about the equivalence principle and hence about what is usually claimed to be the gist of the elevator TE. Indeed, this judgement was passed by the leading relativist of the later generation, John L. Synge, and is a predominant view among the physicists working in the field of general relativity and cosmology (Janssen 2014: 178–179) given that the modern day criterion for the presence of a gravitational field is whether the curvature (not metric) tensor has non-vanishing components: "The Principle of Equivalence performed the essential office of midwife at the birth of general relativity. […] I suggest that the midwife be now buried with appropriate honours and the facts of absolute space-time faced" (Synge 1960: ix–x).

As Janssen explains in conclusion of his überblick of the genesis of general relativity (2014: 208), Einstein could indeed be said to have developed a theory which can be interpreted as seeing gravitational fields as relative, but definitely not all motions as relative. Also, the invariance of the covariant equations of general relativity which Einstein was for a while conflating with invariance of special relativity (as pointed out by Erich Kretschmann already in 1917 and by several authors ever since (Janssen 2014: 186–187)) thinking that there is a principle of relativity of motion related to the general theory as well. Furthermore, Einstein's hopes to make general theory into a Machian theory of gravitational field and inertia failed and Einstein, as we have seen, in the end gave up Machian notions. What, then, remains, is a new theory of gravity with absolute pseudo-Riemannian space-time, still with some vestiges of absolute motion and hard to trace genesis.

Next to the ideas of non-Euclidean geometry of space-time and covariance of the field equations which Einstein picked up through studying abstract mathematical theories and musing on whether these could have any consequence for the physical reality—much like Gauss once wondered whether the sum of the angles in a physical triangle (made of, say, light beams from different lanterns sufficiently far apart) is 180° or more, or less—there is one more idea which makes for a constant in his thinking throughout the process of discovering general relativity. This is the idea I mentioned at the beginning as I believe it was one of the earliest thoughts in this process, namely, the intuition that gravity should be eliminable as a force. Now, at the end of the discussion of the elevator TE and related equivalence principle, let us examine how credible this idea is. I think there are reasons to believe this is the idea the elevator TE was supposed to illustrate all along and it is the one constant that crops up again and again in Einstein's thinking after 1905.

At the time of the development of general relativity this idea, however bizzare, could have appeared reasonable enough to push forward, given in particular the analogy with how magnetic field can be eliminated in special realtivity. But what about electric force/field, can it be eliminated in the same way? Or, even better, what about nuclear forces, the strong and the weak, which could not only be said not to be eliminable but also could not even be conceived as classical fields? So one wonders whether Einstein would have hit on his covariant field equations, at least if he would have discovered them starting from the same originating ideas, if he already in the 1910s could have known of the other two fundamental forces? To conclude, the mixed messages we get from the elevator TE are just a sign of the more general cacophony that still remains when it comes to disentangling all the subtleties involved in discovering and justifying the general theory of relativity.

I would now go on to analyse messages of another of Einstein's famous TEs, this time with a more positive conclusion, and return at the end of the next section to the problems surrounding the Einsteinian justification process which could also arise as problems for a non-Platonist account such as Miščević's.

5. *Einstein's light momentum TE: deducing $E = m c^2$*

I believe the less discussed of Einstein's thought experiments, the light momentum TE, deserves perhaps the highest status. It would appear that Einstein regarded it highly too, as he developed versions of it virtually throughout his working life, from 1905 to 1946. The version presented here is Norton's adaptation of Einstein's 1946 and final rendering (Norton 2014: 96–98). Why would Einstein return on multiple occasions in the span of more than four decades to try to demonstrate that $E = m c^2$, or that energy and mass are equivalent? Each time trying to render his "proof" simpler and using fewer elements from parts of physics different from special theory of relativity. And what can be said about the nature of his proofs? Are they to be taken as sufficient *per se* to establish the relation, so as *a priori* (or mathematical) proofs, or should we still require experimental evidence that the relation holds (which we have by now obtained on many occasions from different type experiments)?

The answer to the first query would appear to be that Einstein wanted at least one quantitative result of his two main contributions to theoretical physics and science, in general, to be fully within grasp of even a high school student of physics, and I believe with the final version of his derivation (as presented by Norton in any case) he indeed succeeded, given the minimal requirements of knowledge of either physics theories or its experimental results (which Einstein anyway lists and none of which is too difficult to understand or at least appreciate in its significance for the derivation) as well as of the level of mathematical skill (basic high school vector algebra will suffice). It would

make perfect sense for Einstein to try to achieve such a derivation/argument, as in spite of his theories of relativity being quite abstract and conceptually extremely demanding (especially the general theory, as we have even if only partly seen in the previous section), not to mention the mathematical requirements they impose on the student, Einstein fostered a firm belief that the fundamental ideas of his physics, as indeed of all physics, can be expressed in simple terms (at least some of those ideas). But I believe there is yet another reason which he expressed perhaps most clearly in his famous 1933 Oxford lecture on the methods of theoretical physics:

> Our experience up to date justifies us in feeling sure that in Nature is actualised the ideal of mathematical simplicity. It is my conviction that pure mathematical construction enables us to discover the concepts and the laws connecting them which give us the key to the understanding of the phenomena of Nature. Experience can, of course, guide us in our choice of serviceable mathematical concepts; it cannot possibly be the source from which they are derived; experience, of course, remains the sole criterion of the serviceability of a mathematical construction for physics, but the truly creative principle resides in mathematics. In a certain sense, therefore, I hold it to be true that pure thought is competent to comprehend the real, as the ancients dreamed. (Einstein 1934: 163–169)

I mean, in particular, his emphasis that *pure thought is competent to comprehend the real, as the ancients dreamed,* and the identification of this thought with the (predominantly) mathematical process of discovery, which of course is primarily aprioristic, as is its justificatory process. By referring to the *ancients* (not the modern day philosophers!) Einstein is further underlying to which genealogy he as a thinker belongs, to the genealogy of Platonic thinkers (at least partly, given that Einstein did use different epistemologies in an opportunistic way depending on the needs of his science), those who *dream* that reality can be comprehended by pure thinking (where this is clearly not meant in a pejorative sense). To the latter testifies the opening phrase of the quoted paragraph: *in Nature is actualised the ideal of mathematical simplicity*. The experience Einstein here refers to is his own experience of work (by the time of his speech measured in a few decades) at the forefront of research in theoretical physics. To me it is also significant that he uses another of Leibnizian expressions about actualization of principles. Einstein as an avid reader in philosophy and, if not consciously a follower of Leibniz but explicitly a follower of Spinoza, with whom, as is well known, Leibniz had many points in common (not to mention that he went to study with him as a young men). So Einstein, in polishing his "proof" which he reached by pure thought (as no experimental evidence was available for it around 1905 and for years to come), like Spinoza was polishing his lenses, could be seen as trying to present a perfect proof of his general approach to physics and partly of his worldview.

Before we endeavour to answer the second query, let us examine the derivation. The process of light emission is seeen from two frames of reference, one at rest (S', with mass of the emitter m' and emitted energy in both directions $E'/2$) and one moving with velocity v perpendicularly to the direction of emission as seen from S'. Momentum of light is given from Maxwell's theory by $p = E/c$, and thus from the viewpoint of S' the momentum of light in each direction is $E'/(2c)$, whereas the new momentum from viewpoint of S is

$$(E/2c)(v/c)=1/2(E/c^2)v,$$

taking into account only the vertical portions which are a v/c fraction of the total momentum in each direction. Hence, total change of momentum in the direction of motion is $(E/c^2)v$, and since the particle is losing the same momentum expressed as mv, we obtain:

$$m=E/c^2.$$

One cannot deny the simplicity and brevity of the derivation, but does it suffice as a "proof" (by pure thinking), and how general it actually is? In his derivation Einstein relies on:

- law of conservation of momentum – that it holds for light as well as material particles;
- formula for the momentum of light (waves or photons alike) from Maxwell's theory;
- the Lorentz contraction coefficient known from experiments of Fizeau (known to Einstein at the time of the first derivation) and Michelson-Morley experiments (which Einstein was adamant he did not know whilst in process of discoverig STR).

Surely, we could accept this derivation just as given in a thought experiment as a proof in the sense of an *a priori* proof that visible light carries inertia. However, the question remains: what does it take to generalise the deduction to all forms of, first, electromagnetic energy and, then, to all forms of energy.

To generalise the conclusion: that *all electromagnetic energy* (light) has inertia, he assumes that all the different EM-waves differ only by frequency/wavelength and that all the emission or absorption processes are equivalent, in the sense that all the above assumptions/facts hold for any of them. But what does it take to generalise the result to *all energy* has inertia, as he was later to do? General validity of the laws of conservation for all matter-energy, new concept of mass-energy, Lorentz transformations for momentum-energy, Noether's theorems…? These were not all spelled out in the original TE or its versions, as Einstein knew he could do only as much when it came to generalizing his results by using TE as the only tool. However, his equation does hold generally, for all forms of energy, known and yet perhaps to be discovered, and to motivate this we would need to expand our discussion much more to examine the general framework of the relativity theories and the physical and philosophical reasons as to why it should hold, or

why the theories to be discovered should also be expected to be relativistically invariant theories. If we had all these clearly spelled out, we could, I think, be excused, for believing that the argument derived from TE alone suffices to justify the belief in the correctness of the deduction. As Miščević puts it:

> First, note that the alleged minuses of TEs are not really minuses of thought experimenting as such, but rather *deficiencies of available wider frameworks!* Further, if an important thesis is scientifically testable in some reasonable time, then TEs teaching us about it can still be very useful. (2022: 118, my italics.)

So even though I agree with Miščević's overall analysis of generic TEs, I would still point to how truly surprising is not only the result of the light momentum TE, but also the fact that it illuminates aprioristic deductive thinking, albeit in a limited domain of application, something more akin to a Platonist account of TEs rather than a naturalist one along the lines of Miščević's proposal (cf. Brown 1991/2005: especially ch. 4).

6. *Einstein's cocksureness and the reason why scientific and metaphysical TEs could be regarded as special*

Einstein was famous for his open-mindedness when it comes to the potential revision of his, at times even most cherished, beliefs as well as for his honesty in admitting errors of judgment (cf. eg. Janssen 2014: 216). He was also pretty cocksure. When asked by biographer and philosopher Ilse Rosenthal-Schneider about how he received the news from Eddington's solar eclipse expedition of 1919, the results of which proved Einstein's calculations of the bending of light rays from a distant star passing the Sun as predicted by general relativity, he was not exhilarated as expected but laconically retorted:

> 'I knew that the theory is correct. Did you doubt it?' I answered, 'No, of course not. But what would you have said if there had been no confirmation like this?' He replied, 'I would have had to pity our dear God. The theory is correct all the same.' (Rosenthal-Schnieder 1980: 74)

This response could seem nothing short of blasphemy, not to say arrogance. But, of course, it wasn't Einstein's intention to be either. He liked to couch his musings on the nature of physical reality, his philosophical outlook, the meaning of life and other big questions in theological terms, not necessarily adopting any particular theology. He had no interest in being arrogant, especially late in life and after achieving not only the main results of his physics, but also worldwide fame reaching far beyond the community of physicists or scientists in general. So should we take it for granted that Einstein knew when he was right, even before having been given experimental data, that he had some special insight into the nature of things, a direct line to God? Although it is tempting, we should be reminded that Einstein did exclaim in the

past, and on more than one occasion, that he was certain he was in possession of the true theory when he in reality was not. For instance, in 1914, in a letter to Michele Besso, his lifelong friend from student days, Einstein wrote about his perfect satisfaction with the prototype of a general theory of relativity but which lacked the general covariance (the so called *"Entwurf" theory*):

> Now I am completely satisfied and no longer doubt the correctnes of the whole system, whether the observation of the solar eclipse works out or not. The sense [*vernunft*] of the matter is too evident. […] The general theory of invariants functioned only as a hindrance. The direct path proved itself to be the only passable one. (As quoted in Norton 1995: 61–62)

Notice that exactly the opposite was to ultimately show iteslf to be true, namely, that the final theory was to be a covariant theory with certain invariants seen as the crucial part of the whole system, and that the line of thought Einstein was, even against his own will, forced to follow the one of mathematical simplicity and elegance which he will much later praise in his Oxford lecture quoted above, not the direct path of physical insight. Yet, the phrasing is almost identical to the response he gave Rosenthal-Schneider, up to denying the relevance of the solar eclipse results (which are historically the *observatio*, if not *experimentum crucis* for general relativity!). Again, what are we to think of Einstein self-confidence, was it a mere joke? Obviously not to him, as the letter to Besso testifies, where he was in earnest about what he was saying, even if we could doubt the same being true in case of the conversation recorded by Rosenthal-Schneider. The plain matter of the fact is that Einstein was not always sure and could not always be sure about the correctness of his theoretical constructs, and that it would be a mistake to take for granted that he somehow always knew. However, it is also quite evident that Einstein did have a special insight into the nature of things as witnessed by his light momentum TE and so many other similar examples. Einstein was, as said at the beginning of this paper, a master of manipulating thought experiments so as to reveal Nature's secrets – this is what he presumably meant by the *direct path* – and one could not blame him that he preferred this direct, or at least more starightforward, pathway into Nature's hidden realm, not least as it usually served as a kind of shortcut. That sometimes he could not find appropriate shortcut, or that sometimes there was not one, but only the ardous path of abstract mathematics was available, surely cannot be taken against his general outlook. We could say of Einstein as it was said of Benjamin Franklin, and even more trully: *eripuit fulmen coelo sceptrumque tyrannis*. I would also maintain that his cocksuredness was not only a byproduct of his method of thought experimenting and coming to the far reaching conclusions about the nature of things, but that it was a prerequisite for it, as Einstein was first and foremost a theorist and the purest of the pure, but not a stranger to all experimenting (after all he spent some years in the Bern patent office).

It was essential to his method as a theorist to be able to not only do the so called *back-of-the-envelope calculations* but also to try to guess at the solutions before even attempting to solve (or put forward) an equation. In order to develop this kind of method you need cocksuredness as part of your character. I think much here is explicable rationally, but there is a residuum which escapes any rationalistic or naturalistic analysis and is best described by my deliberately chosen words *insight into the nature of things*.

Finally, let me remark on the claim put modestly by Miščević (2022: section 6.4) that he sees scientific and, broadly speaking, philosophical TEs as on a par, and although he can see that scientific TEs usually always have some effect on the discussion at hand (even if disproved by real experiments), he does not see any reason why we should be forced to decide on the comparative value of either based on usefulness only. My response would be, based on the studies of primarily scientific (in particular Einstein's) TEs, but also the metaphysical ones, and comparing them with related *genera* as Miščević espouses throughout his book, that one could potentially try to mount a serious objection to the claim that all the TEs, or rather IETs, stand equal in terms of epistemological value. Namely, behind every scientific (and I would also say metaphysical) TE there must be a general framework which Miščević also mentions in the passage quoted in the previous section of this paper, within which there is a hierarchy of statements, from axioms/hypotheses of highest degree of generality to more specific claims; there is, in Leibnizian jargon, a whole hierarchy of reasons which can justify this or that belief. I am quite doubtful as to the existence of such principles in such areas of philosophy as ethics, politics or philosophy of history. Let us take ethics as an example. If we take ethics heteronomously, then its foundations are outside it, so its grounding principles are not ethical. If, on the other hand, we take it autonomously, we end up with all the intricacies of the problem of the relationship of individual interests as against the interests of the group. And even a theonomously considered ethics has its problems as the main reasons for something happening to us, the grounding principles, are perhaps for ever to stay unclear to us humans (take the Biblical example of Job, who keeps suffering as a righteous man, something that should not be happening according to general morality that comes accross in the *Old Testament*). Does anyone believe, to borrow a phrase of Herman Melville from his *Moby Dick* (ch. 64), that *angels are nothing more than sharks well governed*? Are there universally knowable universal rules of ethics to be found by some thought experimenting such as John Rawls'? And one would be hard pressed indeed to try to find the laws of history or politics. As the course of fate of both individuals and societies is determined by so many factors, it is impossible to know its turns, and the so called *real politics* is usually just bestial, so the only rule is the rule of the jungle.

Acknowledgments

I would like to thank the late Professor Nenad Miščević, my mentor and friend, and Professor James Brown for inviting me to participate in the 47th Philosophy of Science conference in Dubrovnik in April 2022 and contribute to the discussion of Professor Miščević's valuable book. This paper is dedicated to the ever-lasting memory of my Professor.

References

Brown J. R. 1991/2005. *Laboratory of the Mind: Thought Experiments in the Natural Sciences*. London: Routledge.

Einstein A. 1919/1988. "What is the Theory of Relativity?" English translation in A. Einstein 1954. *Ideas and Opinions*. New York: Bonanza Books.

Einstein A. 1934. "On the Method of Theoretical Physics." *Philosophy of Science* 1: 163–169.

Einstein A. 1954/1988. "Letter to Felix Pirani February 2nd 1954." In A. Einstein. *Ideas and Opinions*. New York: Bonanza Books.

Janssen M. 2014. "'No Success Like Failure...': Einstein's Quest for General Relativity, 1907–1920." In M. Janssen and C. Lehner (eds.). *The Cambridge Companion to Einstein*. Cambridge: Cambridge University Press, 167–228.

Miščević N. 2022. *Thought Experiements*. Cham: Springer.

Norton J. D. 1985. "What Was Einstein's Principle of Equivalence?" *Studies in History and Philosophy of Science* 16: 203–246. Reprinted in D. Howard and J. Stachel (eds.). 1989. *Einstein and the History of General Relativity: Einstein Studies* Vol. I. Boston: Birkhäuser, 5–47.

Norton J. D. 1991. "Thought Experiments in Einstein's Work." In T. Horowitz and G. J. Massey (eds.). *Thought Experiments in Science and Philosophy*. Lanham: Rowman and Littlefield Publishers, 129–144.

Norton J. D. 1995. "Eliminative Induction as a Method of Discovery: How Einstein Discovered General Relativity." In J. Leplin (ed.). *The Creation of Ideas in Physics: Studies for a Methodology of Theory Construction*. Dordrecht: Kluwer, 29–69.

Norton J. D. 2014. "Einstein's Special Theory of Relativity and the Problems in the Electrodynamics of Moving Bodies That Led Him to It." In M. Janssen and C. Lehner (eds.). *The Cambridge Companion to Einstein*. Cambridge: Cambridge University Press, 72–103.

Norton J. D. 2021. "Author's Responses." *Studies in History and Philosophy of Science* 85: 114–126.

Rosenthal-Schneider I. 1980. *Reality and Scientific Truth*. Detroit: Wayne State University Press.

Russell B. 1937/1992. *A Critical Exposition of the Philosophy of Leibniz with an Appendix of Leading Passages*, 3rd ed. London and New York: Routledge.

Stachel J. 1989. "The Rigidly Rotating Disk as the 'Missing Link' in the History of General Relativity." In D. Howard and J. Stachel (eds.). *Einstein and the History of General Relativity: Einstein Studies* Vol. I. Boston: Birkhäuser, 48–62.

Synge J. L. 1960. *Relativity: The General Theory*. Amsterdam: North-Holland.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.5
Received: March 19, 2024
Accepted: May 13, 2024

Generative Linguistics and the Computational Level

FINTAN MALLORY
Durham University, Durham, UK

Generative linguistics is widely claimed to produce theories at the level of computation in the sense outlined by David Marr. Marr even used generative grammar as an example of a computational level theory. At this level, a theory specifies a function for mapping one kind of information into another. How this function is computed is then specified at the algorithmic level before an account of how this is algorithm is realised by some physical system is presented at the implementation level. This paper will argue that generative linguistics does not fit anywhere within this framework. We will then look at several ways researchers have attempted to modify either the framework of generative theory to reconcile the two approaches. Finally, it presents and discusses an alternative position, anti-realism about generative grammar. While this position has attracted some recent support, it also runs into some of the problems that earlier modifications faced.

Keywords: Generative grammar; Marr; computation; cognitive science; linguistics; Chomsky.

What is the relation between generative linguistics and the rest of the cognitive sciences? Despite the historical role played by generative grammar in the cognitive turn of the 1950s and 60s, linguists have complained for decades that the rest of the cognitive sciences largely ignore their findings.[1] This concern has only intensified with the revival of connectionism in cognitive science under the guise of "artificial intelligence" (i.e. deep neural networks). Consider the recent claim that

[1] For example, Ian Roberts asks "why is mainstream generative syntax overlooked in cognitive science as a whole?" (Roberts 2014: 22) or as Ray Jackendoff's phrased it in the title of a Topic-Comment piece from 1988, "Why are They Saying These Things about Us?" (Jackendoff 1988).

"After decades of privilege and prominence in linguistics, Noam Chomsky's approach to the science of language is experiencing a remarkable downfall" (Piantadosi 2023). Behind these concerns is a lack of clarity about the metatheory of generative linguistics, i.e. claims about what generative models—grammars—actually model, and specifically, in what sense they describe computations. For decades it has been customary to claim that generative grammar was a computational level theory akin to David Marr's program within the cognitive neuroscience of vision. If so, generative grammarians would be seeking the same kinds of explanations that have had success across the cognitive sciences.

A glance at the literature would suggest that this is the case. Marr explicitly invoked generative grammar as an example of a computational level theory when he introduced his levels of analysis writing that "Chomsky's (1965) theory of transformational grammar is a true computational theory in the sense defined earlier" (Marr 1982: 28). Chomsky, in turn, agreed claiming, "We may consider the study of grammar and UG [universal grammar] to be at the level of the theory of computation" (Chomsky 1982: 48). This idea remains the consensus position among linguists. As Klaus Abels puts it "theorising at the most abstract, the computational-level has remained the mainstay of work in theoretical linguistics."[2] In the last few years, appeal to Marr has been used to justify or explain theoretical disagreements between traditional generative grammarians and usage-based theorists (Yang 2017, Adger and Svenonius 2015) as well as model-theoretic syntax (Neeleman 2013; Graf 2017). It has appeared in debates concerning how the levels of description should interact in linguistics (Yang 2017; Hornstein 2013; Abels 2013; Hornstein and Pietroski 2009), has been invoked in debates about language evolution (Johnson 2015, 2016; Berwick and Chomsky 2016; Perfors 2017), as well as the independence of knowledge from production and comprehension (Neeleman and Van de Koot 2010). Appeal to Marr has also formed the basis of the interaction between theoretical linguistics and other subfields within cognitive science (Poeppel 2017; Kobele 2012; Embick and Poeppel 2014; Jackendoff 2012; Murphy 2015). The idea that generative grammar is a computational level theory in Marr's sense constitutes the most successful response to the "realism" debates which have followed generative linguistics since its inception (see Pylyshyn 1973 for a discussion of the "psychological reality" of generativist claims).

This paper will first argue that, despite widespread claims to the contrary, generative grammar is not a computational-level theory in

[2] See also, "What Chomsky (1965) calls a theory of 'competence' or 'knowledge of language' corresponds to Marr's computational theory" (Jackendoff 2012: 1133). "[t]he competence theories of linguistics correspond to Marr's (1980) topmost level of computational theory [...]" (Heinz 2011: 140). "A theory of grammar corresponds to Marr's abstract theory of a computation" (Berwick 1985: 9). Countless other examples of this claim can be found in the literature.

the sense articulated by Marr (sections 1-4). Simply put, Marr's computational level concerns a theory of performance, not competence whereas generative grammar purports to describe linguistic competence. I will then go on to consider the alternative interpretations of generative grammar that have been proposed; that generative grammar is a metamathematical theory of computation, a theory of parsing (i.e. performance), and a description of a separate data structure utilised by the parser. While each of these approaches has its merits, we will see that each requires rejecting core features of generative linguistics. Finally, I will articulate a position that is, at least, implicit in some core generative literature, that of *modal anti-realism* about computation. This position makes sense of some theoretical practice but ultimately denies that the structure-building operations posited by grammarians are realised as processes in the human brain.

2. *Marr's computational level*

According to David Marr, information processing systems are best described at three different levels: computational, algorithmic, and implementation. At the computational level, "the performance of the device is characterised as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated" (Marr 1982: 24). The mapping is stated as a function: $f: I \rightarrow O$ (hence the alternative name "function-theoretic explanation"). At the algorithmic level, a representation of the input and output of this mapping is provided, and an algorithm that produces the output from the input is proposed. At the implementation level, an account is given of how this algorithmic process is physically realised by some physical system, e.g. the activations of neurones, oscillatory dynamics, transistors, etc. While these levels are conceptually distinct, the development of a theory at one level may inform theory construction at others (for a brilliant demonstration of how this works, see Jonas and Kording (2017).

Standard examples of computational-level theories are the analysis of the auditory system in terms of Fourier transforms (e.g. Schneider and Mores 2013), hand-eye coordination using vector subtraction (Perrone and Krauzlis 2008), Marr's model of edge detection, and the use of path integration by various animals (Eteinne and Jeffery 2004). So that we have a concrete example to work with, we'll consider Marr's own example of a computational level analysis of a cash register. While this is rough and simple, it shall serve our purposes going forward.

Imagine we want to understand a cash register. At the computational level, we might note that a cash register computes addition. This computational-level characterisation is blind to questions of representation, e.g. whether the cash register uses binary or Roman or Arabic numeral systems. At the algorithmic level, a particular algorithm that

computes this function is posited. The algorithm makes claims about the representational format of the information being processed. At the implementational level, it is explained how this algorithm may be realised by circuitry and micro-transistors. Returning to the computational level, the function characterised is:

(1) Addition: $f(x, y) = x + y$

The computational theory does not only specify which function is implemented by the system but also justifies *why* the function is the appropriate one for our theory. What is at issue here is not *why the system* implements the function but *why our theory* says that this function is the right one (the focus is not teleological but methodological). It is often the case that several different functions could produce the mapping identified and a why-story connects these mathematical properties with relevant features of the world or the task under consideration.[3] In the cash register example, Marr connects the algebraic properties of addition with commercial practices. There is a zero element because buying nothing costs the same as not buying anything. The order in which goods are purchased shouldn't affect the total price (commutativity), nor does it matter if they are paid for separately (associativity) and the register can also handle the existence of a refund policy (inverses).[4] What allows us to speak of the different properties of these functions is the intensionality of our characterisations, i.e. the function is described independently of its inputs and outputs. As Egan observes, one might use an algorithm to specify a function at the computational level without making a commitment to the algorithm which implements the function (Egan 2017). For now, we note two important things that follow from the intensionality of the computational level description.

First, having an intensional characterisation of the function allows us to discuss the function independently from the environment in which it is embedded—where "environment" may be understood as

[3] Example 1: In Marr's account of stereopsis (i.e. binocular vision), he takes into account where dots may actually appear on physical surfaces in the world in order to select between functions: "We have to examine the basis in the physical world for making a correspondence between the two images" (Marr 1980: 112). Example 2: Perrone and Krauzlis' (2008) account of eye rotation. The task modelled is the subtraction of image movement as a result of eye-rotation from image movement that occurs as a result of an agent's traversal through space, i.e. vector subtraction. One way of answering the question of how vector subtraction occurs in the brain involves the use of arctan (the inverse tangent function) while another involves treating the vectors as cosine curves in which their length and direction correspond to the amplitude and phases of the curves. The authors observe that "the problem of singularities associated with the inverse tangent also seems to preclude any simple biological implementation" and so choose the second approach.

[4] While Marr claims that these properties uniquely individuate the operation of addition, they will actually hold of any operation on an abelian group and while Marr's initial claims that the "why"-part of computational-level theorising can individuate a unique function is clearly too strong, it can be amended (see Anderson on Rational Analysis, Anderson 1990).

either the physical environment in which the agent is located or the wider properties of the information-processing system (or cognome). For example, cash registers don't compute the addition function for values less than some bound and "error" for values above it. Adding or removing memory from our cash register without altering the addition algorithm will alter the function's domain and range but intensionally it won't change the function computed. Likewise, the visual system presumably computes the same function for edge-detection whether one has glaucoma or not.

Second, the fact that the function can be characterised independently of the actual activity of the system serves a normative purpose. Once it has been determined what function the system computes, the theorist is in the position to assess if the system is functioning normally (Egan 2017). Supporting this is the fact that the mathematical theory which characterises the function is independent of the psychological theory which describes its implementation. In the example above, the theory of arithmetic is not grounded by the theory of cash registers. There is no suggestion that 1+1=2 because that's how cash registers see the world. Rather, the cash registers are designed (though they could be naturally occurring, Darwinian-evolved cash registers) to track this independent mathematical fact. The fact that the truths of mathematics are independent of the existence of cash registers doesn't entail that cognitive systems implementing functions can only be individuated by reference to extra-mental mathematical reality. It simply acknowledges that we don't expect facts about cash registers to ground facts about numbers.

These two points will be important when we consider whether or not generative grammar provides computational level theories. First, though, we must turn to generative grammar.

3. *Generative grammar*

Generative grammar is the branch of cognitive science aimed at characterising the state of the human mind corresponding to an individual's knowledge of a language. A generative grammar is a function-theoretic characterisation of this knowledge. A quick glance at the literature would suggest that the function corresponding to a grammar characterises a mapping from sounds (or visual signs) to meanings.[5] This suggests a function along the lines of:

[5] This function is often spoken of in Marrian terms: "Generative accounts of linguistic phenomena are couched at a level of analysis that is close to Marr's (1982) computational-level. That is, the theory specifies a system that guarantees a particular pairing of sounds and meanings across a potentially unbounded domain" (Adger and Svenonius 2015: 6). "A computational account of language has two parts. First is a specification of which sounds (or more generally signals) convey which meanings" (Kobele 2012: 411).

(2) Grammar: f(sound) = meaning

However, matters are not so simple. Standard function-theoretic characterisations are methodologically possible because researchers have a pre-existing mathematical account of what the functions are; they are typically number-theoretic. Number theoretic functions are used because these functions are often defined over quantifiable inputs, e.g. Marr's appeal to the Laplacian of a Gaussian ($\nabla^2 G$) of the retinal array is possible because it is defined over (numerical) intensity values. In contrast, we do not have a pre-syntactic grasp of the sets of possible phonological and semantic structures between which our grammar characterises a mapping. Furthermore, the sets of sounds and meanings we are interested in are unboundedly large (we want to know how unboundedly many strings can map to unboundedly many meanings). As a result, generative linguists don't attempt to characterise this mapping directly. Instead, they describe an operation for building syntactic structures from lexical inputs. More accurately, they describe a function that recursively enumerates the set of ordered ⟨sound, meaning⟩ pairs where each element of this set is individuated by its syntactic structure. This is still a characterisation of the sound-meaning function but "from below."

The input to this function is typically presented as either a finite set of lexical items, either the whole lexicon as in Collins and Stabler (2016) or as a numeration. I'll represent it here as the power set of the lexicon (strictly speaking, this should include multisets, see Adger 2021), while the output will be the set of structural descriptions (SDs), or syntactic structures of the language.

(3) Grammar: g(lexicon) → Syntactic Structures

The grammar recursively enumerates the set of sound-meaning mappings as structured by the syntax of the language. In effect, it decomposes the function *f*, by revealing the structure of each sound-meaning pair.[6]

To know a language, according to generative linguistics is, in part, to realise a function that outputs the set of syntactic structures for that language. This function is, in turn, defined by describing an operation for combining items in the lexicon into more complex structures. This operation can apply iteratively to the structures it has already generated, as we see in the following toy example where we imagine that ⊕ is the structure-building operation:

[6] In other words, function *g* recursively enumerates the extension of function *f*. Instead defining a mapping from a set of sounds to a set of meanings (sounds as inputs, meanings as outputs), function g takes ordered sound-meaning pairs (i.e. lexical items) as the primitives and enumerates the set of possible combinations of sound-meaning pairs. It should be noted that this is also the case in systems like HPSG in which phonological and semantic information are combined in the same feature structure.

(4) "the" ⊕ "dog" → [the dog]
(5) "likes" ⊕ "the dog" → [likes[the dog]]
(6) "the cat"⊕ "likes the dog" → [[the cat] [likes [the dog]]]

Along with a simple operation for merging syntactic objects, grammars are also capable of moving these items to different locations in the syntactic structure, thereby building more complex structures. For example:

(7) "the dog$_1$" + "the cat likes the dog$_1$" → [[the dog][the cat] [likes [~~the dog~~]]]

In (7), the noun phrase "the dog" which had its grammatical case determined by the verb "likes," has been raised to form the relative clause "the dog the cat likes." As a result of movement, even though "the dog" was the first element in the phrase to be constructed, it finds itself at the (linear) beginning of the phrase. Within minimalist syntax, it is possible for the most embedded element to be the first merged. Furthermore, which structures can be built by the structure-building operations will be determined by *syntactic features* of the lexical items they take as their input. Just as the combinatorial capacities of legos and atoms are determined by the intrinsic properties of those entities, the combinatorial properties of lexical items, and thus what the structure-building operation can do with them, are determined by their syntactic features (e.g. an item with the features V, =N is labelled as V and can combine with an item labelled as N).

Within Minimalism, the core structure-building operation is called "merge," within Head Driven Phrase Structure Grammar and Generalised Phrase Structure Grammar, it is called "unification," within Tree-Adjoining Grammars, it is "tree adjunction," within categorial grammars, it is "function composition."[7] The exact details of these frameworks aren't relevant to this discussion; instead, what matters is that they all characterise the knowledge of language in terms of an operation for building syntactic structures. I will be using "merge" as a cover-all term for *the* core syntactic operation in what follows. By characterising a function that outputs unboundedly many different syntactic structures, the linguist, in theory, gives an account of what is involved in knowing a language. The question is whether describing

[7] Within minimalism, merge and movement are constrained by a series of further feature-checking operations (agree, probe, labelling etc.) which determine whether two lexical items can be merged, but ultimately, it is merge that builds the syntactic structures. Not every framework has "movement" as illustrated in our toy case but all describe a basic operation for constructing complex structures. While unification is also used in some varieties of Construction Grammar (e.g. Kay and Fillmore 1999) it is unclear how much of the discussion in this paper would apply to CxG theories which draw directly on usage. I suspect that CxG approaches will relate to both performance and the Marrian framework quite differently to more generative frameworks. For a recent discussion of the connections between CxG and the Predictive Processing model of cognition, see Michel (2023).

this function amounts to giving a computational-level theory for our knowledge of language.

Superficially, this appears to be the case. Just as with computational level theories, generative grammars provide intensional characterisations of a function: "I-languages are functions regarded in intension" (Chomsky 1995: 26). Grammars describe linguistic competence independently of the social and cognitive environments in which they are embedded including any constraints on memory facing parsers.[8] The theories are computational, in that, they specify a general method by which an output can be generated in a finite number of steps. Furthermore, generative linguistics is often involved in comparing extensionally equivalent systems of grammar and presenting arguments for why one is superior to another. For example, while multiple context-free grammars and Minimalist grammars can generate the same sets of syntactic structures, linguists have given compelling reasons to believe that the latter is the more cognitively plausible, in effect, presenting the *why* component of a computational level theory. The problem is that a grammar is not a computational level explanation.

4. *Differences between generative grammar and computational level explanations*

We'll consider here two differences between the function characterised in (3) and Marrian computational level theories.

1. Performance/Competence: The first and most obvious reason that a generative grammar is not a computation-level theory in Marr's sense is that it does not describe a process and so, by extension, does not characterise information processing. A syntactic derivation is not a real-time event. Actual linguistic processing, or at least our parsing models of it, must incrementally construct syntactic representations from unlabelled inputs, starting with the words at the start of the sentence. In contrast, the syntactic derivations posited in generative grammar build structures from the most embedded constituents outwards, applying movement operations when necessary.

Furthermore, the inputs to syntactic derivations are not unlabelled strings as the inputs to parsing are but highly specified feature structures that represent information about the elements combined (this

[8] It is worth noting that, if we embrace the Marrian interpretation of generative grammar, then the Minimalist Program would appear to be a perfectly reasonable application of Anderson's "Principle of Rationality" for computational-level theorising. This is the idea that we take the function computed by some cognitive system to be the optimal function for the task and incrementally and iteratively modify our proposal on the basis of how the system's behaviour diverges from the function proposed. This has proven to be methodologically well-motivated for narrowing down which functions should be posited at the computational level (for a recent defence of the method see Van Rooij et. al. 2018). It is also worth noting that Anderson is one of the few theorists who doesn't equate competence theories with computational-level theories (Anderson 1990: 8–9).

was left out of the example above). This disconnect between the two approaches is quite explicit. While on the computational level, "[t]he performance of the device is characterised as a mapping from one kind of information to another" (Marr 1982: 24), "[a] generative grammar [...] in no sense purports to be a description of his actual performance, either as a speaker or as a listener" (Chomsky 1965: 3). It is instead a theory of competence; an account of what an agent must *know* (or "cognise") in order to perform some task, not a computational characterisation of the task itself. It is presumably this issue that Chomsky is raising when he writes: "David Marr's influential ideas about levels of analysis do not apply here at all, contrary to much discussion, because he too is considering input-output systems [...]" (Chomsky 1995: 12).

2. Grounding: A second difference is that, unlike the mathematical functions typical of computational-level theorising, generative grammars ground the structures they output. According to standard generative assumptions, a sentence has the syntactic structure linguists ascribe to it because that is the structure assigned to it by a grammar.[9] A cognitively realised grammar makes it the case that a sentence has its particular syntactic structure and not some other one, and the structures which a grammar generates are characterised solely with reference to that grammar; there is no independent method for characterising the grammatical structures of language (at least over an infinite set). Contrast this with the cash register example above, it is clear that one needn't be a Platonist to think that the values the cash register computes don't depend on the cash register for their existence. 1 + 1 does not equal 2 because that is how cash registers see the world. Rather it's the fact that 1 + 1 = 2 is true independently of the cash register that allows us to determine whether or not the cash register is functioning properly since we can contrast the results of addition with the results of a broken cash register. The natural numbers have their structure independent of cash registers. In contrast, the syntactic structures characterised in generative grammar are function-dependent. We cannot characterise the full set of syntactic structures output by a grammar except with reference to the grammar itself.

Since syntactic structures can only be ascribed to a sentence with reference to a particular grammar, it isn't possible to distinguish between the function a grammar is supposed to compute and the function that it actually does compute (though we can still say that a sentence is ungrammatical relative to a grammar). As a result, it simply doesn't make sense to say that the grammar is computing the *wrong* function (generative grammarians are descriptivists, not prescriptivists about grammar).[10] This problem does not arise for other cognitive systems

[9] There are some realists (or Platonists) about linguistic structure who claim it exists independently of the human mind (Devitt 2006, Katz 1981, Katz and Postal 1991) but this remains a fringe position.

[10] This point shouldn't be mistaken for the familiar bugbear of functional indeterminacy. The problem of functional indeterminacy concerns whether or not

studied within the Marrian framework. We do not assume that the world is 3D because our visual system creates a 3D representation from its input whereas we do assume that a sentence has the particular structure it possesses in virtue of speakers' particular grammars. Similarly, the retinal array exists independently of the function which uses it to construct 3D objects, whereas linguistic structure does not exist independently of a grammar.

While Chomsky (1995) appears to reject the idea that generative linguistics produces computational level theories in Marr's sense, more recently Berwick and Chomsky explicitly endorse a Marrian interpretation of generative grammar and imply that the operation merge is implemented on a lower algorithmic level (Berwick and Chomsky 2016: 132-139). However, *even more recently*, Chomsky et al., (2023) cites a 2012 interview in which he does suggest that the Marrian framework is ill-suited for understanding "internal capacities." The quote cited doesn't appear in the printed version of the interview but it is worth examining:

> As discussed in Marr (1982), complex biological systems must be understood at different levels of analysis (computational, algorithmic, implementational). Here we discuss internal language, a system of knowledge, which we understand at a computational level. Since such a system is intensional, therefore not a process, there's no algorithm. In contrast, externalization, a process of using the internal system, may find an algorithmic characterization. (Chomsky et al. 2023: 8)[11]

This is by far the most explicit statement of how Chomsky regards generative grammar to relate to the Marrian framework. I suspect

it is possible to identify a correct function in cases where the function's domain of application is infinite or even just very large. The problem here, however, has nothing to do with the size of the function's input or output. It arises as a result of the widespread commitment to the mind-dependence of linguistic structure within generative grammar. If the function implemented by the language faculty is what makes it true that a sentence has the structure it possesses, that function cannot be assessed with regards accuracy.

[11] In the same interview, Chomsky considers how one would characterise knowledge of mathematics and the case is very similar to language. "If you try to find out what that internal system is of yours, the Marr hierarchy doesn't really work very well. You can talk about the computational level—maybe the rules I have are Peano's axioms that describes a core set of basic rules of arithmetic and natural numbers, from which many useful facts about arithmetic can be deduced, or something, whatever they are—that's the computational level. In theory, though we don't know how, you can talk about the neurophysiological level, nobody knows how, but there's no real algorithmic level. Because there's no calculation of knowledge, it's just a system of knowledge. To find out the nature of the system of knowledge, there is no algorithm, because there is no process. Using the system of knowledge, that'll have a process, but that's something different" (Chomsky 2012). While Chomsky is only one generative linguist among many, his ideas have been uniquely influential in the field and if he has a non-standard interpretation of what it is involved in developing a computational level theory, it may be useful for researchers working the same tradition to be aware of this, if only to reflect on what they mean when they say generative grammar is computational.

many would agree that, if there is no algorithm computing the function, it is not a computational description as Marr presents it and so Chomsky's continued use of "computational level" to describe generative grammar is non-standard. In any case, these are not matters to be settled by appeals to authority and whether any of the proponents of the Marrian interpretation of generative grammar mentioned above share this interpretation is unclear.

To summarise these claims; unlike computational level theories, generative grammars don't describe processes, are strongly intensional, and are structure grounding. Generative theories give a procedural characterisation of a state, knowledge of language, while computational theories give a static characterisation of a process. This much shouldn't be controversial, though it is seldom acknowledged in print and it certainly isn't itself an objection to the generative method.[12] It's not surprising that language, which must be acquired and varies at least in its surface manifestations, would require a different approach from other cognitive capacities. What I will examine in the rest of this paper is whether we can give an account of generative grammar that captures the methodological virtues of the Marrian method while adhering to the constraints placed on it by the properties listed above. In the next section, I will look at several ways in which theorists have attempted to reinterpret generative grammar as a computational level theory and discuss the challenges they face before considering an alternative approach that may have a better chance (but nonetheless has problems of its own).

5. *Alternative interpretations*

5.1. *Grammars as metamathematical descriptions*

One option is to treat generative grammars as highly abstract description at the computational level. Manfred Krifka responds directly to Chomsky's claim that generative grammar does not concern input-output systems writing: "I do not see why representation levels should only be applicable to computation arising in input/output systems. In particular, one could see level 3 descriptions as idealisations, for example, the Peano axioms for our integrated arithmetic abilities, or the rules postulated by a generative grammar for our linguistic abilities" (Krifka 2011: 55). Analogies between generative grammar and Peano arithmetic are quite common and so this proposal deserves consideration.[13] The core idea seems to be that, by providing metamathematical

[12] Poeppel (2012) and Poeppel and Embick (2015) raise a range of challenges faced by attempts to connect linguistic theory and the neurobiology of language. However, they do appear to accept that traditional generative linguistics has been targeted as computational-level theorising (Poeppel 2012: 50; Poeppel and Embick 2015: 359).

[13] Chomsky (1999: 41–42), Adger and Svenonius (2015: 1422), and Boeckx (2010) all speak of the computational level as the study of the "logical properties" of the language faculty.

characterisations of the function implemented we can learn about the language faculty. Chomsky himself has made similar claims:

> One of the properties of Peano's Axioms PA is that PA generates the proof P of '2 + 2 = 4' but not the proof P of '2 + 2 = 7' (in suitable notation). We can speak freely of the property 'generable by PA' f holding of P but not P¢, and derivatively of lines of generable proofs (theorems) and the set of theorems without postulating any entities beyond PA and its properties. (Chomsky 2001: 41-42)

This excerpt has been a constant source of debate over the last two decades and I won't recapitulate the controversy here.[14] The axiom analogy does help clarify matters to an extent. Steps in a proof or derivation can be ordered and when viewed as mathematical objects there is no need to regard that order as temporal rather than structural. The connections between computation and deduction are relatively well understood and whether a set of axioms and associated rules generate a proof does not depend upon their actually being used in real-time to generate that proof.

The problem with the analogy is that it raises as many questions as it answers. The axiomatic view of grammar, while aligning with the parsing-as-deduction approach to grammar (e.g. Johnson 1989), commits us to some internal system of representation, i.e a formal language. If this is the case, then it merely pushes any question about our grasp of language back to another level (a "language of thought" needs a grammar too). The concern is that any formalism with the expressive power to represent the full range of syntactic structures found across natural languages (e.g. weak monadic second order logic) would itself require a grammar in which its combinatorial possibilities are specified. While the strength of this criticism depends on how robustly the notion of axiom is taken, most axiomatic systems require their syntax to be specified in some metalanguage.[15] Furthermore, metamathematical statements such as the Peano axioms don't tell us anything without either a set of inference rules, e.g. modus ponens, or a model. Lacking either a proof theory or a semantics, they are merely marks. It would presumably be these "logical capacities" in which we are interested when appealing to such an axiomatisation. Yet if this is the proposal, the theory doesn't support any further claims. For example, it is unwarranted to claim that a cash register is capable of inferring according to modus ponens on the grounds that it computes addition. To say that the cash register has in any sense the capacity to make logical inferences according to the rules of a classical proof system or that it can map variables to models of Peano arithmetic is simply false.

This problem is not solved by weakening our proof system (e.g. using Heyting arithmetic or intuitionist arithmetic) to get us any closer to

[14] Paul Postal described it as "the most irresponsible passage written by a professional linguist in the entire history of linguistics" (Postal 2004: 296).

[15] A discussion of the role that such concerns played in the early development of generative grammar can be found in Mallory (2023).

the "psychological reality" of the machine. A cash register is neither capable of inferring "p" from "~~p" nor incapable of it. The trivial reason is that computing the sum of x and y is not the same as deriving a proof that x + y = z. The tasks described at the metamathematical level by our formal system and at the computational-level by our function are different. Nevertheless according to Pylyshyn, "[t]his is exactly the goal Chomsky declared many years ago for linguistics: find the least powerful formalism for expressing the range of human languages and you will have a formalism that you can view as intensionally (as opposed to merely extensionally) significant" (Pylyshyn 1991: 14).

The moral here is that, when making inferences from the existence of one capacity to the existence of another we must be careful not to conflate metamathematical and algorithmic levels of description. For example, if the system in question performs multiplication using the Karatsuba algorithm, we can infer that it is capable of addition as well since the algorithm requires this. We can't however, make this kind of conclusion based on metamathematical ideas alone, e.g. the fact that recursive definitions of multiplication tend to utilise addition doesn't tell us much. In Skolem arithmetic (a complete and decidable subsystem of Peano arithmetic) multiplication is defined independently of addition. If we want to preserve the function-theoretic outlook by making the theory more abstract, it becomes less clear what the theory is actually telling us about the mind.

5.2. *Grammars as theories of performance*

5.2.1. *Grammars as parsers*

The next option is to construe grammars as computational-level descriptions of parsers. Parsing is the process of incrementally building representations of the syntactic structure of input sentences, in other words, mapping strings (one kind of information) to hierarchical structures (another kind of information). It can therefore be understood as an input-output system. Several researchers have explicitly attempted to bridge the gap between generative practice and Marrian metatheory this way. One of the most sophisticated developments of the grammar-as-parser view is provided by Neeleman and Van de Koot (2010).[16] Neeleman and Van de Koot present the minimalist operation merge as an abstract characterisation of the actual operation which a parser uses to build structural representations of sentences.

Just as Marr decomposed the algebraic properties of addition in the cash register example, Neeleman and Van de Koot discuss the abstract properties that structure-building operations might possess. Constraints on the structure-building process such as whether it is binary-

[16] A similar "one system" view is defended by Lewis and Phillips (2015). Ruth Kempson's program of Dynamic Syntax may also be seen as a higher-level description of the parser but one which takes account of the linear order of sentences (Kempson et al. 2001).

branching, inclusive, whether labels are assigned, and so forth, can be either built into the parser's structure-building process or left as filters on outputs. When these constraints are understood as properties of the parser's structure-building operation, then the computational burden of parsing is lightened significantly—fewer candidate structures are constructed and filtered. Furthermore, the properties can be characterised and discussed independently of any algorithm which implements them.

The decision to interpret merge as a parsing operation is to treat it as an aspect of performance rather than competence. Parsing a sentence is a real-time, memory-bounded cognitive process, whereas generative linguistics was initially developed as a theory of competence, a description of the state of mind corresponding to knowledge of a language, not its actual use (Chomsky 1965).[17] Whether or not you consider this reinterpretation to be a bad thing will depend upon your prior metatheoretical commitments. Nonetheless, it is worth noting some issues with this approach.

First, the operation "merge" for example, applies first to the most embedded constituent in a sentence (e.g. the "$\_\_\_\__x$" in "what$_x$ did the kitten swallow $\_\_\_\__x$?") and builds a syntactic structure outwards from this, moving constituents to higher (and more fronted) positions in the process. In English, more embedded constituents tend to appear closer the end of a clause. Parsing, in contrast, begins with the most leftward constituent in a structure and builds structure from there. As a higher-level theory of parsing, then, generative grammars start their derivation at the wrong end of the sentence. Second, the information a grammar has available to it is much richer than the information a parser has access to. Formal models of grammar assume that the inputs to the merge operation contain explicit, structured sets features which are checked off in the process of structure-building. The inputs to a grammar wear their syntactic properties on their sleeve. In contrast, the inputs to the process of human parsing are underspecified chunked phonological units or bare strings (as garden-path sentences show). This capacity for ambiguity makes parsing a difficult challenge. So if we are to view a generative theory of structure-building as a model of how the parser builds structures, it's reasonable to ask if it's likely to be a good one. Parsing is a cognitive process that lends itself well to computational-level theorising but whether the operations described by competence grammars can be translated neatly into a such a theory is open for debate (much of which is outlined in Pereplyotchik (2017)).

5.2.2. *Implementation level concerns*

[17] This is not to say that researchers weren't almost immediately attempting to connect generative claims to models of performance in work culminating in the derivational theory of complexity (Fodor and Garrett 1966; Garrett, Beaver and Fodor 1966).

Some of the most exciting contemporary research is coming from psycholinguistics. Before continuing, we should consider what the foregoing discussion means for attempts to identify the neurological correlates of the operations described by generative grammarians. In an influential body of research, Angela Friederici has assembled considerable evidence that merge occurs in the ventral part of BA44 (the posterior inferior frontal gyrus) (Friederici 2017; Liu et al. 2023). Similarly, Eliot Murphy has developed a sophisticated account of the implementation of merge, according to which the operation is realised by cross-frequency interactions between θ and ɣ frequencies where lexical items indexed by ɣ cycles are embedded within slower θ and **δ** oscillations (Murphy 2020). Others have sought the neural correlates of other operations posited by generative grammarians such as *search* (Ohta, Fukui and Sakai 2013), *label* (Murphy 2015), and *scrambling* (Makuuchi et al. 2013).

However, if we accept the claim that generative grammar does not describe performance, then there is no way to reconcile the psycholinguistic claims that the *inferior-frontal cortex,* (IFG) (Caplan et al. 1998) putatively affords core-syntactic operations such as "merge" (Zaccarella et al. 2017) and "movement" (Grodzinsky and Santi 2008; Makuuchi et al. 2013) with claims by Chomsky like the following:

> [A] generative system involves no temporal dimension. In this respect, generation of expressions is similar to other recursive processes such as construction of formal proofs. Intuitively, the proof 'begins' with axioms and each line is added to earlier lines by rules of inference or additional axioms. But this implies no temporal ordering. It is simply a description of the structural properties of the geometrical object proof. (Chomsky 2007: 6)

None of these positions are compatible with the view of merge or any other syntactic operation as an atemporal "logical" operation—logical abstractions don't show up in fMRI scans (which makes it striking that Chomsky (2017) agrees with Friederici's findings). If merge is a "logical operation" not taking place in space or time, it doesn't take place in the inferior frontal gyrus.[18] It seems reasonable to conclude that, while much of this research is exciting and important, there is little reason to believe it is tracking what generative grammarians are describing when they develop theories of syntactic structure-building *unless we reinterpret those theories as theories of parsing or some other linguistic performance.* This position has been advocated by some psycholinguists for independent reasons (Phillips 2013; Embick and Poeppel 2005).

[18] This is known as the problem of ontological commensurability (Poeppel 2017). "The tendency in generative syntax, for example, is to speak as if the computations proposed in syntactic analyses need not be regarded as computations that are performed in real-time […] This assumption simply makes the link between linguistics and neuroscience harder to bridge, for reasons that are ultimately historical, and not necessarily principled" (Poeppel and Embick 2005: 114).

5.3. *Grammars as data-structures*

The final account we'll look at claims that generative grammars specify the data structures that are accessed by the parsing mechanism. This would place grammars at what Christopher Peacocke calls "level 1.5," somewhere between computational and algorithmic level theories. According to Peacocke, a theory at this level "identifies the information drawn upon by an algorithm" (Peacocke 1989).[19] This approach aligns with Bresnan and Kaplan's *Strong Competence Hypothesis* (the idea that a competence grammar is used by performance systems).[20] If correct, parsers are algorithms that utilise grammars to produce appropriate syntactic structures for an input string. This gives substance to the idea that a grammar "underlies and accounts for," "determines" or "is put to use by" performance (each of these phrases occur without further elaboration in Chomsky (1964)). The question then is, what is the role of a structure-building "computational" operation like merge in the theory of parsing? To be clear, we are not now considering the operations that might combine and label input constituents during parsing, but instead, we are looking at the role of an operation like merge in the grammar accessed by the parser. Parsing algorithms have their own range of real-time computational operations, e.g. pushing a unit of information to a stack, adding information to a table, searching for a representation of a rule in the grammar (see Jurafsky and Martin 2008 for introductions to basic parsing algorithms). These operations are described at the algorithmic level although one can perhaps have a computational level theory of the parser as Neeleman and van de Koot demonstrate. Typically, as algorithmic level theories, they involve representational commitments, e.g. a grammar is in Chomsky Normal Form, and so the question is how the computational operations described by a generative grammar are represented within the grammar.[21] What we are concerned with is the role of structure-building operations posited by grammarians in these accounts.

[19] Momma and Phillips also suggest that Marr's hierarchy may be best viewed as a continuum in order to accommodate the anomalous position of linguistics and neurolinguistics (Momma and Phillips 2018).

[20] Whether grammars are causally implicated in performance has been a matter of debate for decades. While many argue that grammars are causally implicated in performance (Fodor 1985; Peacocke 1986, 1989; Rey 2003; Hornstein 2009), John Collins argues that they aren't causally implicated (Collins 2017, 2023). Higginbotham claims that whether or not grammars play a causal role in production is to be determined by empirical enquiry (Higginbotham 1982). Generally, formulations of how a grammar relates to performance haven't added much detail to Chomsky's original brushstrokes, e.g. "[k]nowledge of language guides/provides the basis for actual use, but does not completely determine use" (Boeckx 2009: 134). Rey (2020) claims that a grammar makes claims about cognitive processes and architecture "at some level of abstraction," a position I think many would get behind (Rey 2020: 112).

[21] This is a considerable simplification. Designing a parser often involves deciding which rules should be represented in the grammar (i.e. a data structure separate

For example, the operations of phrase rewriting in a phrase structure grammar manifest as relations between categories when that grammar is being consulted by a parser. In the simplest example, a rule of grammar, NP → Det N, is not to be viewed as a rewrite rule for deriving some structure for another but as a statement in a database which can be accessed by the parsing algorithm. While it is sometimes suggested that grammatical operations have to be applied to generate structures so that the structure is, in effect, built twice during parsing, this isn't the role that grammar operations play in contemporary, highly-lexicalised parsing models where the structure generating information is built in to the lexical item rather than into rules relating grammatical categories. Consider Stabler's influential Minimalist parser (Stabler 2014; Berwick and Stabler 2019; Hunter 2019). While Minimalist grammars are typically "bottom-up"—derivations are formed by merging lexical items into more complex units and then merging those units until the derivation is complete, Stabler's minimalist parser is top-down, in the sense, that it starts with the highest category of a phrase along with a queue of predictions for what lies below it before applying rules operating over the input and the queue of predictions. What we need to examine is the role of merge in the grammar of this model. When we do, we see that the role of merge in this model is to specify the properties of the grammar, i.e. the sets of features associated with each lexical item. The features that lexical items are taken to have are just those that they *would* need *if* merge *were* the actual operation by which syntactic structures were built. Whether or not two items can be merged by the grammar is determined by those items syntactic features. The conceptual function of merge in the grammar is simply to induce upon the lexicon, the set of features they would require if syntactic structures were to be constructed by means of merge. But once we have the grammar on the table—the set of lexical items with their rich array of syntactic features, then we can, in effect, ignore merge when talking about how the grammar interacts with the parser. The grammar is simply a structure of the lexicon upon which parsing operations can then be defined. At no point in the model is it assumed that merge is actually implemented in the brain to generate syntactic structures. Merge is not really a computational operation at all. It doesn't describe a lower-level algorithmic process but it does give us the means to identify a set of features which such a (parsing) process may access.

Isn't this just an algorithmic level theory? Yes and no. A parser model is an algorithmic-level theory. It specifies an algorithm for computing an output for a given input in a finite series of steps. The algorithmic level theory tells us what those steps are, it specifies the algorithm,

from the parser which can be altered while keeping the parsing algorithm the same) and which rules are to be built into the operation of the parser itself. Pereplyotchik (2017) gives a helpful overview of these issues.

and in the process it makes claims about how the information must be represented. In the case of minimalist parsing models, information is represented as a set of features of lexical items. In Stabler's parsing model, this information can be understood as a structure within the lexicon. However, merge is not the computation that searches through the lexicon to incrementally build syntactic structures. This brings us to a final approach to understanding generative grammar as a computational level theory.

6. Anti-realist computational level theories

This interpretation of computation in generative grammar is both *modal* and *anti-realist* (or perhaps instrumentalist). It is modal in that it treats merge not as an operation that does apply to build syntactic structures but a computational operation that *could* apply. While it is *anti-realist* because it regards the merge-story of how structure could possibly be built as a useful theoretical device for describing how syntactic information is organised in some cognitive structure rather than a representation of an actual cognitive operation (in the style of Marr). The core idea is that, one way to describe what syntactic features lexical items need to have to be effectively learned and parsed is to describe a simple device for building syntactic structures and ask what information it would require to do its job. Then, once we know what this information is, we discard the structure-building operation. It is not posited as "cognitively real," in the sense that it doesn't pick out any real-time process. What is genuinely represented in the brain are syntactic features which are accessed by the parser. An operation like merge is a notational device for figuring out what those features might be.

If this is actually the idea that has been implicit within the literature, it would make sense of some of the stranger claims one finds in Chomsky. For example, consider the following:

> We can discuss the set of expressions or derivations generated by a grammar but in doing so no new entities are postulated in these usages beyond FL [the faculty of language], its states L [some language], and their properties. Similarly, a study of the solar system could introduce the notion HT = {possible trajectories of Halley's Comet within the solar system}, and the studies of motor organization or visual system could introduce the notions plans for moving the arm or visual images for cats (vs. bees). But these studies do not postulate weird entities apart from planets, comets, neurons, cats, and the like. (Chomsky 2001: 41-41)

Read descriptively, there are obvious problems with this analogy. Firstly, a comet does actually follow one of these trajectories—the set of trajectories is a description of paths the comet *might* take. They constitute a modal claim about the possible behaviour of the comet. In contrast, the set of derivations a grammar generates is not a description of a grammar's *possible* performance. Secondly, we do not characterise a comet as a device recursively enumerating its possible trajectories.

The trajectories of a comet are determined by things like mass and velocity, properties of the comet which constrain its possible behaviour and apply to other objects as well. Finally, it is often claimed that the outputs of grammars are involved in mediating the interface between systems of phonology and semantics. One might reasonably think that such entities would have to exist in order to do this. However, if these outputs are not generated, it is not clear how they could serve this role in transduction.

However, these objections arise only if we understand the modal component of the interpretation as a *de re* claim about an actual computational operation. If we read this section as a discussion of what the merge operation *could* do, then it seems like Chomsky is making a claim about an actual, cognitively implemented computational operation (that can be functionally localised etc.). However, if we understand it as a claim about the kind of theory that can help us uncover the structure of the lexicon, which features lexical items possess, what unpronounced items such as functional heads there might be, then it becomes a more plausible, instrumental claim about a useful kind of theory building. It is not just that the syntactic structures enumerated by a grammar are an idealisation of some cognitive structure, but the operation involved in their generation, merge, is an idealisation of this structure as well.[22] This kind of anti-realism about computation in generative linguistics has recently been advocated by John Collins: "what makes a system a computer is that only a computational theory is adequate for its explanation, independently of whether or not any physical states are discriminable as realising the computation" (Collins 2023). If this interpretation is correct, then merge is not an operation in the brain. It is merely a way of describing a data structure. It isn't just that merge doesn't occur in real time, it isn't supposed to characterise a process that does. Its function within the generative theory is to help linguists to identify syntactic features and phenomena that emerge from how syntactic information (which is actually drawn upon by the parser) is organised in the brain. Accordingly, when humans evolved the capacity for merge, they developed the capacity to organise information in their brains in such a way that, *were* merge to occur using information organised this way, it would generate the structures we find in a languages.[23]

[22] It is easy to be misled by reference to "idealization" here. In this case, merge would not be an idealisation of some actual structure building operation (considered independently of memory limitations, for example), but a theoretical tool which gives us the formal resources required to describe the functional properties of the lexical items which any such structure-building operation would have to have access to. Insomuch as it idealises actual performance processes, it does so obliquely, by enabling researchers to better describe the information that such processes have access to (i.e. as a competence theory).

[23] This is similar in a respect to Adger (2022). Adger argues that the representations posited in generative grammar are structured abstractions of brain states.

This still leaves us with a range of questions for the antirealist: why think that the relevant syntactic features are those that facilitate the operation of a structure-building device that isn't actually responsible for building syntactic structures in real time? Why aren't the features we posit the ones that are directly posited to facilitate efficient parsing (as in Ruth Kempson's Dynamic Syntax program)? If reference to merge is an expression that one is adopting a distinct framework of idealisation, how should we think of the empirical content of generative theories as well as theoretical debates about the exact nature of merge (e.g. binary merge, parallel merge, workspace models)? These are important questions for anyone who adopts the generative framework as it has been described here and it is far from certain that they have easy answers. For some, I suspect that this position will be too great a concession to abstraction.

The present paper has merely sought to illustrate that generative grammar is not a computational level theory in the traditional Marrian sense assumed throughout much of computational cognitive neuroscience and suggest that researchers in psycholinguistics are unlikely to find the structure-building operations discussed by linguists in their labs. I have also suggested that a more instrumentalist interpretation may make sense of how generative "computations" are appealed to within parsing theory. Throughout this, I have tried to balance both empirical, theoretical, and to some extent, hermeneutic evidence. The practitioners of a scientific discipline are by no means obligated to interpret their own research in accordance with the ideas and images of prominent figures within their field. The fact that Chomsky might interpret the subject matter of generative linguistics in a certain way should not bind others in the field. I do, however, think there is value in being explicit about the promises and challenges of different metatheoretical commitments, in particular due to the interdisciplinary nature of current research.

Acknowledgement

The bulk of this paper was written long ago and benefitted from discussion with David Ager and Mark Textor. More recent additions have benefited from comments at the OMLET in Oslo, in particular from Drew Johnson and Nicholas Allot, and in Durham from James Miller, Aadil Kurji and Keith Begley.

References

Abels, K. 2013. "Comments on Hornstein." *Mind & Language* 28 (4): 421–429.

Adger, D. and Svenonius, P. 2015. "Linguistic Explanation and Domain Specialization: A Case Study in Bound Variable Anaphora." *Frontiers in Psychology* 6: 421.

Adger, D. 2021. "The Architecture of Computation" In N. Allot, T. Lohndal and G. Rey (eds.). *A Companion to Chomsky*. Hoboken: John Wiley And Sons, 123–139

Adger, D. 2022. "What are Linguistic Representations?" *Mind & Language* 37 (2): 248–260.

Anderson, J. R. 1990. *The Adaptive Character of Thought*. Hillsdale: Lawrence Erlbaum Associates.

Berwick, R. C. 1985. *The Acquisition of Syntactic Knowledge*. Cambridge: MIT Press.

Berwick, R. C. and Chomsky, N. 2016. *Why Only Us?* Cambridge: MIT Press.

Berwick, R. C., Pietroski, P., Yankama, B. and Chomsky, N. 2011. "Poverty of the Stimulus Revisited." *Cognitive Science* 35 (7): 1207–1242.

Berwick, R. C. and Stabler, E. P. (eds.). 2019. *Minimalist Parsing*. Oxford: Oxford University Press.

Boeckx, C. 2010. *Language in Cognition*. Oxford: Wiley-Blackwell.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

———. 1980. "Rules and Representations." *The Behavioural and Brain Sciences* 3 (127): 1–61.

———. 1982. *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge: MIT Press.

———. 1995. "Language and Nature." *Mind* 104 (413): 1–61.

———. 2001. "Derivation by Phase." In M. Kenstowicz (ed.). *Ken Hale: A Life in Language*. Cambridge: MIT Press, 1–52.

———. 2007. "Approaching UG from Below." In U. Sauerland and H. Gärtner (eds.). *Interfaces + Recursion = Language?: Chomsky's Minimalism and the View from Syntax-Semantics*. Berlin: Mouton de Gruyter, 1–29.

———. 2013. "Problems of Projection." *Lingua* 130: 33–49.

Chomsky, N., Seely, T. D., Berwick, R. C., Fong, S., Huybregts, M. A. C., Kitahara, H. and Sugimoto, Y. 2023. *Merge and the Strong Minimalist Thesis*. Cambridge: Cambridge University Press.

Collins, J. 2017. "Faculties and Modules: Chomsky on Cognitive Architecture." In J. McGilvray (ed.). *The Cambridge Companion to Chomsky*. Cambridge: Cambridge University Press, 217–234.

———. 2023. "Generative Linguistics: 'Galilean Style'." *Language Sciences* 100: 101585.

Egan, F. 2017. "Function-Theoretic Explanation and the Search for Neural Mechanisms." In D. M. Kaplan (ed.). *Integrating Mind and Brain Science: Mechanistic Perspectives and Beyond*. Oxford: Oxford University Press, 145–163.

Embick, D. and Poeppel, D. 2005. "Defining the Relation Between Linguistics and Neuroscience." In A. Cutler (ed.). *Twenty-first Century Psycholinguistics: Four Cornerstones*. Mahwah, NJ: Lawrence Erlbaum Asso-

ciates, 103–118.

———. 2015. "Towards a Computational(ist) Neurobiology of Language: Correlational, Integrated and Explanatory Neurolinguistics." *Language, Cognition and Neuroscience* 30 (4): 357–366.

Etienne, A. S. and Jeffery K. J. 2004. "Path Integration in Mammals." *Hippocampus* 14 (2): 180–192.

Fodor, J. A. and Garrett, M. 1967. "Some Syntactic Determinants of Sentential Complexity." *Perception & Psychophysics* 2 (7): 289–296.

Fodor, J. 1985. "Some Notes on What Linguistics is About." In J. J. Katz (ed.). *The Philosophy of Linguistics*. Oxford: Oxford University Press, 146–160.

Friederici, A. D. and Kotz. S. A. 2003. "The Brain Basis of Syntactic Processes: Functional Imaging and Lesion Studies." *Neuroimage* 20 (1): S8–S17.

Friederici, A. D. 2017. *Language in Our Brain: The Origins of a Uniquely Human Capacity*. Cambridge: MIT Press.

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A. and Bolhuis, J. J. 2017. "Language, Mind and Brain." *Nature Human Behaviour* 1 (10): 713–722.

Garrett, M., Bever, T. and Fodor, J. 1966. "The Active Use of Grammar in Speech Perception." *Perception & Psychophysics* 1 (1): 30–32.

Grodzinsky, Y. and Santi, A. 2008. "The Battle for Broca's Region." *Trends in Cognitive Sciences* 12 (12): 474–480.

Graf, T. 2017. "A Computational Guide to the Dichotomy of Features and Constraints." *Glossa: a Journal of General Linguistics* 2 (1): 18.

Heinz, J. 2011. "Computational Phonology Part I: Foundations." *Language and Linguistics Compass* 5 (4): 140–152.

Higginbotham, J. 1982. "Noam Chomsky's Linguistic Theory." *Social Research* 49 (1): 143–157.

———. 1991. "Remarks on the Metaphysics of Linguistics." *Linguistics and Philosophy* 14 (5): 555–66.

Hornstein, N. 2009. *A Theory of Syntax*. Cambridge: Cambridge University Press.

———. 2013. "Three Grades of Grammatical Involvement: Syntax from a Minimalist Perspective." *Mind & Language* 28 (4): 392–420.

Hornstein, N. and Pietroski, P. 2009. "Basic Operations: Minimal Syntax-Semantics." *Catalan Journal of Linguistics* 8 (1): 113–139.

Hunter, T. 2019. "Left-corner Parsing of Minimalist Grammars." In R. C. Berwick and E. P. Stabler (eds.). *Minimalist Parsing*. Oxford: Oxford University Press, 125–158.

Jackendoff, R. 1988. "Topic…Comment: Why are They Saying These Things about Us?" *Natural Language & Linguistic Theory* 6 (3): 435–442.

———. 2012. "Language as a Source of Evidence for Theories of Spatial Representation." *Perception* 41 (9): 1128–1152.

Johnson, M. 1989. "Parsing as Deduction: The Use of Knowledge of Language." *Journal of Psycholinguistic Research* 18 (1): 105–128.

Jonas, E. and Kording, K. 2017. "Could a Neuroscientist Understand a Microprocessor?" *PLoS Computational Biology* 13 (1): e1005268.

Johnson, K. 2015. "Notational Variants and Invariance in Linguistics." *Mind & Language* 30 (2): 162–186.

Johnson, M. 2016. "Marr's Levels and the Minimalist Program." *Psycho-*

nomic Bulletin & Review 24 (1): 171–174.
Kay. P. and Fillmore, C. J. 1999. "Grammatical Constructions and Linguistic Generalizations: The What's X doing Y? Construction." *Language* 75 (1): 1–33.
Katz, J. 1981. *Language and Other Abstract Objects*. Oxford: Basil Blackwell.
Katz, J. and Postal, P. 1991. "Realism vs. Conceptualism in Linguistics." *Linguistics and Philosophy* 14 (5): 515–554.
Kempson, R. Meyer-Viol, W. and Gabbay, D. 2001. *Dynamic Syntax: The Flow of Understanding*. Oxford: Blackwell.
Kobele, G. 2012. "Ellipsis: Computation of." *Wiley Interdisciplinary Reviews* (3): 411–418.
Krifka, M. 2011. "In Defence of Idealizations: A Comment on Stokhof and van Lambalgen." *Theoretical Linguistics* 37: 51–62.
Lewis, S. and Philips, C. 2015. "Aligning Grammatical Theories and Language Processing Models." *Journal of Psycholinguistic Research* 44: 27–46.
Liu, Y., Gao, C., Friederici, A. D., Zaccarella, E. and Chen, L. 2023. "Exploring the Neurobiology of Merge at a Basic Level: Insights from a Novel Artificial Grammar Paradigm." *Frontiers in Psychology* 14: 1151518.
Michiru, M. and Friederici, A. D. 2013. "Hierarchical Functional Connectivity between the Core Language System and the Working Memory System." *Cortex* 49 (9): 2416–2423.
Mallory, F. 2023. "Why is Generative Grammar Recursive?" *Erkenntnis* 88: 3097–3111.
Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.
Michel, C. 2023. "Scaling up Predictive Processing to Language with Construction Grammar." *Philosophical Psychology* 36 (3): 553–579.
Momma, S. and Phillips, C. 2018. "The Relationship between Parsing and Generation." *Annual Review of Linguistics* 4: 233–254.
Murphy, E. 2015. "Labels, Cognomes, and Cyclic Computation: an Ethological Perspective." *Frontiers in Psychology* 6: 715.
Murphy, E. 2020. *The Oscillatory Nature of Language*. Cambridge: Cambridge University Press.
Neeleman, A. 2013. "Comments on Pullum." *Mind & Language* 28 (4): 522–531.
Neeleman, A. and van de Koot, H. 2010. "Theoretical Validity and Psychological Reality of the Grammatical Code." In M. Everaert, T. Lentz, H. De Mulder, H. Nilsen and A. Zondervan (eds.). *The Linguistics Enterprise*. Amsterdam: John Benjamins, 183–212.
Ohta, S. and Sakai, K. L. 2017. "Computational Principles of Syntax in the Regions Specialized for Language: Integrating Theoretical Linguistics and Functional Neuroimaging." In N. Fukui (ed.). *Merge in the Mind-Brain*. New York: Routledge, 237–264.
Peacocke, C. 1986. "Explanation in Computational Psychology: Language, Perception and Level 1.5." *Mind and Language* 1 (2): 101–123.
———. 1990. "When is a Grammar Psychologically Real." In A. George and N. Chomsky (eds.). *Reflections on Chomsky*. Oxford: Basil Blackwell, 111–130.
Pereplyotchik, D. 2017. *Psychosyntax: The Nature of Grammar and its*

Place in the Mind. Springer Cham.

Perfors, A. 2017. "On Simplicity and Emergence." *Psychonomic Bulletin & Review* 24 (1): 175–176.

Perrone, J. A. and Krauzlis, R. J. 2008. "Vector Subtraction Using Visual and Extraretinal Motion Signals: A New Look at Efference Copy and Corollary Discharge Theories." *Journal of Vision* 8 (14): 1–14.

Philips, C. 2013. "Parser Grammar Relations: We don't understand everything twice." In M. Sanz, I. Laka and M. K. Tanenhaus (eds.). *Language Down the Garden Path*: *The Cognitive and Biological Basis for Linguistic Structure*. Oxford: Oxford University Press, 1–23.

Piantadosi, S. 2023. "Modern Language Models Refute Chomsky's Approach to Language." *Lingbuzz Preprint* url: https://lingbuzz.net/lingbuzz/007180.

Poeppel, D. 2017. "The Influence of Chomsky on the Neuroscience of Language." In J. McGilvray (ed.). *The Cambridge Companion to Chomsky*. Cambridge: Cambridge University Press, 155–174.

Postal, P. 2004. *Skeptical Linguistic Essays*. Oxford: Oxford University Press.

Pullum, G. 2009. "Computational Linguistics and Generative Linguistics: The Triumph of Hope over Experience." *ILCL 09 Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* 12–21.

Pylyshyn, Z. W. 1973. "The Role of Competence Theories in Cognitive Psychology." *Journal of Psycholinguistic Research* 2 (1): 21–50.

Pylyshyn, Z. 1991. "Rules and Representations: Chomsky and Representational Realism." In A. Kashir (ed.). *The Chomskian Turn*. Cambridge: Blackwell, 231–251.

Rey, G. 2003. "Chomsky, Intentionality, and a CRRT." In L. M. Anthony and N. Hornstein (eds.). *Chomsky and his Critics*. Oxford: Blackwell, 105–139.

———. 2020. *Representation of Language: Philosophical Issues in a Chomskyan* Linguistics. Oxford: Oxford University Press.

Roberts, I. 2014. "The Mystery of the Overlooked Discipline: Modern Syntactic Theory and Cognitive Science." *Revue Roumaine de Linguistique* 58: 151–178.

Schneider, A. and Mores, R. 2013. "Fourier-Time-Transformation (FTT), Analysis of Sound and Auditory Perception." In R. Bader (ed.). *Sound - Perception - Performance. Current Research in Systematic Musicology*. Vol 1. Heidelberg: Springer, 299–329.

Stabler, E. P. 2013. "Two Models of Minimalist, Incremental Syntactic Analysis." *Topics in Cognitive Science* 5 (3): 611–633.

van Rooij, I., Wright, C., Kwisthout, J. and Wareham, T. 2018. "Rational Analysis, Intractability, and the Prospects of 'As if'-explanations." *Synthese* 195 (2): 491–510.

Yang, C. 2017. "Rage Against the Machine: Evaluation Metrics in the 21st Century." *Language Acquisition: A Journal of Developmental Linguistics* 24 (2): 100–125.

Zaccarella, E., Schell, M. and Friederici, A. D. 2017. "Reviewing the Functional Basis of the Syntactic Merge Mechanism for Language: A Coordinate-based Activation Likelihood Estimation Meta-analysis." *Neuroscience & Biobehavioral Reviews* 80: 646–656.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.6
Received: February 9, 2023
Accepted: January 26, 2024

Propositions, Concepts, and the Fregean/Russellian Distinction

DUŠAN DOŽUDIĆ
Institute of Philosophy, Zagreb, Croatia

In this paper, I deal with recognising an appropriate criterion for distinguishing two competing conceptions of the propositional content among the content realists—the Fregean and the Russellian—especially in connection to some classical proponents of the realist view (Frege, Moore, and Russell). My starting point is a survey characterisation of the two conceptions and the accompanying classification of Russell's and Moore's conceptions of the propositional content, which I find problematic on several accounts. I set up a context for my consideration and elaborate on why I find it problematic. My central point is that, given how the classical proponents of propositions understood their respective conceptions, as well as how more recent proponents of propositions (for example, David Kaplan) understood them, one should draw the distinction between the Fregean and the Russellian conception on the grounds of what propositional components do rather than the nature of propositional components (unless, of course, one ultimately reduces the latter to the former).

Keywords: Concepts; Frege Gottlob; Fregean; Moore Georg E.; propositions; Russell Bertrand; Russellian.

1. *The unfitting demarcation*

Two disagreements prevail in the debate over propositions. One disagreement is whether such entities exist at all; the other is a disagreement among proponents of propositions themselves, and it concerns the nature and function of such entities. Two competing conceptions prevailed among the authors involved in the latter disagreement for the last hundred and fifty years. One of the conceptions started with

Frege and the other one with Russell. In their entry on propositions, McGrath and Frank (2023: sect. 1) consider several classical proponents of propositions and propose the following characterisation:

> In their early writings, Russell and Moore endorse propositionalism. In his 1903 book *The Principles of Mathematics*, Russell affirms the existence of propositions, taking them to be complexes of ordinary concrete objects (the referents of words) rather than of Fregean senses (p. 47). Propositions so conceived are now standardly called *Russellian*, and propositions conceived as complexes of senses or abstract entities are called *Fregean*. In his 1899 paper, "The Nature of Judgment," Moore affirms the existence of propositions, taking them to be broadly Fregean in nature (in particular as being complexes of mind-independent Platonic universals which he calls concepts).

According to this passage, although Russell and Moore agreed at one point that one needs propositions to explain the relevant phenomena, they disagreed about the nature of such entities. For Moore, they were *Fregean* propositions "conceived as complexes of senses or abstract entities"; for Russell, *Russellian*, namely, "complexes of ordinary concrete objects […] rather than of Fregean senses." According to this characterisation, the disagreement between the two primarily comes down to the disagreement about the nature of proper constituents of propositions—whether their constituents are ordinary concrete objects or senses (i.e., abstract entities). In what follows, I will use McGrath and Frank's characterisation of Russell's and Moore's conceptions (as well as the Russellian and the Fregean conceptions) to point out what I consider the key feature that separates the Fregean and the Russellian conceptions. I will first show why the above-quoted characterisation of the Fregean and Russellian propositions is inadequate and why the accompanying characterisation of Russell's propositions is essentially wrong (sect. 2). Then, I will show why the characterisation of Moore's propositions suggested in the same passage is incorrect (not necessarily for the same reason the first two characterisations are wrong) (sect. 3). Finally, I will use the mischaracterisations detected in the quoted passage to point out what I take to be the key distinguishing feature of the competing conceptions of propositions (sect. 4). In the rest of this section, I briefly characterise Fregean and Russellian propositions using the apparatus Frege has provided.[1]

If one draws parameters for characterising *Fregean* propositions from Frege (1984a), propositions turn out to be entities that stem from the fundamental division of objects and concepts on the one hand and their modes of presentation (senses) on the other.[2] For the sake of ter-

[1] A reviewer objected that throughout the paper I uncritically follow McGrath and Frank in attributing to Frege the view that propositions (i.e., Frege's thoughts) are *structured* entities, thus neglecting the alternative view that for Frege propositions were not structured. In the paper, however, I mainly talk about Frege*ans*, not Frege, and where I talk about Frege, I remain neutral about the matter.

[2] A reviewer suggested I should explicitly state my assumption that for any object or any concept, there is a mode of presentation that (re)presents it uniquely.

minological consistency, neutrality, and brevity, I will call all the examples of objects and concepts Frege had in mind "items." Thus, objects such as Socrates and Aristotle, and properties and relations, such as wisdom, death, cat, older than, or son of, will be "items."[3] Given the characterisation, only modes of presentation of items are constituents of propositions sentences express, never items themselves. Thus, on the one hand, there are complexes, such as *Socrates being older than Aristotle*, which consist of various items (here at least: Socrates, Aristotle, and the *older than* relation) arranged in a particular manner. On the other hand, there are modes of presentation of items arranged in a propositional complex and expressed by the corresponding sentential complex. The expressed proposition, in turn, relates the sentential complex with the complex of items; it is the mediator between the two complexes.

There are apparent exceptions, one being the attitude and indirect speech sentences (reports). In such cases, modes of presentation become items that enter complexes about which one talks using an attitude or indirect speech sentence.[4] In such cases, however, it is not the mode of presentation about which one says something that enters the proposition but *its* mode of presentation, namely, the mode of presentation of that mode of presentation. With the hierarchy of modes of presentation in mind, the fundamental distinction between items (that enter complexes about which one talks using the sentence) and their modes of presentation (which make it possible to talk about complexes in the first place) is preserved. The direct speech sentences that target linguistic expressions as items make another exception.

Accordingly, the point of Fregean propositions is this: When one refers to and says something about items, whatever they may be, these items, relative to the context, are never regarded as senses of given expressions or sentences. Whenever items are referents, they never function as constituents of the expressed proposition. On the other hand, Russellian propositions do not presuppose the Frege-like division, namely, concepts and objects on the one side and senses on the other. The idea of Russellian propositions is that items to which one refers

In fact, here, the assumption is not mine but Frege's, and it would be curious to adopt Frege's apparatus yet deny the assumption. I do not think that Frege ever questioned it.

[3] Thus, items would be similar to what Russell (1992: 43–44) called "terms" (cf. Cartwright 2003: 115–116). Of course, the convention is tentative, and one should bear in mind the potential threat of Frege's "concept horse" problem and his dispute with Russell over it (Frege 1984b; 1980a).

[4] See Frege (1984a: 159, 166–167; 1980b: 164). One should note that Frege, unlike many later Fregeans, strongly opposed any commitment to complexes consisting of objects and concepts, not only as candidates for propositions, but also as candidates for their truthmakers. Instead, he eventually adopted the view that all true sentences refer to the True and all false ones to the False. After he introduced the sense/meaning (reference) distinction, Frege nowhere considered in an approving way any complexes in addition to sentences, thoughts, complex concepts, and complex physical objects (see Frege 1980b: 163–164; 1984a: 161–165).

and about which one says something are precisely entities that function as constituents of the expressed proposition. Indeed, all items that enter complexes and about which one says something function as constituents of the expressed proposition (Russell 1992: 42–52; 1980: 169).

In summary, both Fregeans and Russellians acknowledge the level of items one can refer to and which, if their respective metaphysics allow them, enter complexes about which one talks using the sentences. The disagreement comes at the point of deciding what are the constituents of the proposition and how they come to function that way. One typically considers that point of disagreement in semantic terms of how one succeeds in referring to something and expressing propositions that enable a sentence to hook onto a segment of the reality—a complex. Given the characterisation in the opening quote, what seems to be of interest here is the what-enters-the-proposition disagreement. However, having the Fregean/Russellian distinction and disagreement between Fregeans and Russellians in mind, as already indicated, the question is not merely what enters the proposition, i.e., what are its constituents, but also what the constituents of the proposition do. By acknowledging this what-enters/what-it-does distinction, consider next the characterisation of Russellian and Fregean propositions suggested in the opening quote.

2. *The Fregean and the Russellian*

The opening quote contains the characterisation of Russellian propositions as "complexes of ordinary concrete objects." McGrath and Frank do not specify what ordinary concrete objects would be (except that they are "the referents of words") nor provide examples. I suppose they primarily have well-familiar particulars in mind, such as the pen I am currently writing with, my present computer, the book I am reading right now, or my gluttonous dog lying next to the table. Russell's (1992: 53) neat example that fits here is "an actual man with a tailor and a bank-account or a public-house and a drunken wife." Suppose such candidates exhaust the list of ordinary concrete objects (and I do not see what else would appropriately be described as *ordinary* and *concrete* that would significantly differ from the listed entities).[5] In that case, the characterisation of Russellian propositions proposed in the quoted passage appears inadequate in several respects.[6]

[5] In fact, earlier in the section, McGrath and Frank remark something that supports the proposed reading of "ordinary concrete objects." They briefly consider Plato's view and conclude that "it is far from clear that he takes the objects of belief to be statements rather than simply the ordinary concrete objects (e.g., Theaetetus) and forms (e.g., flying), which the statement is about" (McGrath and Frank 2023: sect. 1). Here, forms (attributes, universals) are clearly excluded from the list of the ordinary concrete objects.

[6] McGrath and Frank explicitly attribute their explanation of Russellian propositions to Russell (1992). But Russell in his 1903 *Principles of Mathematics* (or anywhere else, for that matter) gives no such characterisation of propositions. In

For one thing, those Russellians whose underlying metaphysics comprises more than ordinary concrete objects would not accept it. And I suspect that would be a majority of Russellians (past and present), if not all of them (see Caplan 2007 and Schiffer 2007: 270–271). It is even hard to conceive the possibility of Russellian propositions consisting only of ordinary concrete objects. Proposals of the trope theory could hardly come to the rescue here (as one of the reviewers suggested) since tropes are far from ordinary and are certainly not concrete (cf. Loux and Crisp 2017: 70–75). It is equally challenging to imagine a sentence expressing such a proposition. What would make a complex consisting exclusively of concrete parts *proposition,* and what would make the sentence that expresses it *declarative*? What would bind its ordinary concrete constituents to match the structure of a declarative sentence? Or, as Russell (1992: 35, 39) puts it, what would enable a proposition to *assert* anything of its subject?

Accordingly, some Russellian propositions would not be Russellian on the characterisation proposed in the opening quote. And these would be all the propositions that have abstract in addition to concrete constituents. An example would be the proposition *that Socrates was stubborn*, in which *Socrates* is an ordinary concrete particular and *stubbornness* an abstract entity. Suppose the sentence "Socrates was stubborn" expresses a proposition. That proposition cannot consist only of Socrates as an ordinary concrete object, and there is nothing ordinary and concrete in other candidates for constituents suggested by the sentence. Russell's (1992: 45) more illustrative example involves the proposition *that humanity belongs to Socrates*. Here, within the corresponding sentence, one refers to *humanity* and *Socrates* (using the expressions "humanity" and "Socrates") and indicates they stand in a particular relation to each other. As Russell puts it, a concept "does not walk the street, but lives in the shadowy limbo of the logic-books" (Russell 1992: 53, 64). Particular humans, dogs, books, etc., being ordinary and concrete, indeed do not inhabit such a limbo. But it is not only that the propositions of the kind would not be Russellian by the proposed characterisation. Such propositions would neither be Fregean. As clearly stated in the quote, Fregean propositions are "complexes of senses or abstract entities," and propositions mentioned so far all have concrete in addition to abstract entities: *Socrates* in the proposition *that Socrates was stubborn* in neither a sense nor an abstract object. It is as if the quoted passage presupposes a metaphysical clear-cut between Fregean and Russellian constituents of propositions; that for the former, they are supposed to be abstract, for the latter, concrete. But there is no such clear-cut overlap.

One should thus modify the initially proposed characterisation of Russellian propositions by saying that such propositions are complexes

fact, he insists that in every proposition there must be at least one constituent that is not a term but a concept (1992: 212).

of items. And items would include more than ordinary concrete objects; they would also include properties and relations (and abstract objects, too, if one's metaphysics allows them). Alternatively, they would include no ordinary concrete (or abstract) objects but only properties and relations (I will return to that in the next section).

What holds for Russellian propositions also holds for Russell's (1992) propositions.[7] For him, at the turn of the twentieth century, a proposition is a structured entity—a unity consisting of at least two constituents (1992: 44, 508). One can distinguish constituents of propositions in several ways. Still, the fundamental distinction is to things and concepts (1992: 44). And one can further distinguish concepts into class concepts (e.g., *smart* or *dog*) and relations (e.g., *mother of* or *older than*) (1992: 44–45); class concepts are universals that can have instances, whereas relations are universals without instances (1992: 51–52). Things are terms of a proposition that can occur only as its subjects. In contrast to things, concepts can occur within propositions as subjects (terms) or irreplaceable parts of assertion (in Russell's sense) (1992: 39, 44–45). Every proposition must contain at least one concept that is not a term in it (1992: 212). As far as things are concerned, Russell distinguishes several kinds: ordinary concrete objects (such as the computer on which I am currently typing this or one of my particular mental states (1992: 45)), but also "many other entities not commonly called things" (1992: 44), namely, abstract entities, such as classes or geometrical points (1992: 45–46), but also propositions themselves (1992: 35, 48–49). Russell conveniently summarises his position:

> Whatever may be an object of thought, or may occur in any true or false proposition, or can be counted as *one*, I call a *term*. […] A man, a moment, a number, a class, a relation, a chimaera, or anything else that can be mentioned, is sure to be a term; and to deny that such and such a thing is a term must always be false. (1992: 43)

Characterised in that way, Russell's propositions obviously do not fit McGrath and Frank's characterisation of the proposition class to which such propositions indisputably belong, namely, the class of Russellian propositions.

Once it is granted that Russell's propositions, and Russellian propositions in general, would have to consist of at least one constituent that is not an ordinary concrete object, one can note another problem with the opening characterisation of Russellian propositions. For Russellians who are realists about abstract entities and hold that such entities can be named and not merely described, there would surely be Russellian propositions that consist only of abstract entities. Examples of these might be the proposition *that redness is relational*; *that redness is not greenness*; *that five is greater than three*; *that two is not seven*; etc. Perhaps even the proposition *that five's being greater than two implies*

[7] See Cartwright (2003: 113ff.) and Hylton (2003: 207ff.) for further discussion about Russell's early conception of propositions and their constituents.

two's being less than five, and the like, would belong here, provided one considers propositions themselves to be abstract entities and "that"-clauses referential devices, as some Russellians consider them to be (cf. Schiffer 2006: 268–271).

Thus, it is not only that the above-mentioned Russellian propositions do not satisfy the opening characterisation of Russellian propositions. These propositions satisfy the opening characterisation of Fregean propositions as complexes of senses *or abstract entities*. Take the proposition *that two is not seven*; every constituent of that proposition is indeed abstract, if it is anything at all. Thus, by the proposed characterisation, some Russellian propositions would be Fregean. In fact, given the characterisation, one could hardly find any candidate for Russellian propositions and, accordingly, any actual proponent of Russellian propositions that would fit the characterisation.

For the same reason, the opening characterisation of Fregean propositions fails, too: Some of the propositions that one would, guided by the characterisation, identify as Fregean would, in fact, be Russellian. The problem with the opening characterisation of propositions is not that it is too sketchy and thus allows different interpretations. Its problem is that it emphasises a less important metaphysical aspect, which shifts the focus from what is more important for drawing the Fregean/Russellian distinction.

3. *Moore's concept(ion)*

Mislabelling Russellian propositions as "Fregean," licenced by the opening characterisation, does not stop at the so-far considered cases of Russellian propositions that consist only of abstract entities. There is an interesting kind of Russellian proposition, which McGrath and Frank also labelled "Fregean", namely, Moore's (1993a) propositions. In the opening quote, one reads that "Moore affirms the existence of propositions, taking them to be broadly Fregean in nature." The statement follows up with a brief explanation: "in particular as being complexes of mind-independent Platonic universals which he calls concepts." It is hard to figure out what "broadly Fregean" means here. I take it to mean something like: *not typical Fregean, but definitely not Russellian*.

Why is the just quoted characterisation of Moore's propositions incorrect? Are mind-independent Platonic universals (concepts) not abstract? The provided characterisation is incorrect because Moore's propositions are Russellian, not Fregean. Indeed, one should say that Moore's propositions are broadly Russellian in nature and thereby mean that such propositions are *not typical Russellian but definitely not Fregean*. The reason Moore's propositions are not typical Russellian, one should note, has nothing to do with the key feature of Russellian propositions but rather with Moore's unusual metaphysical conception of their constituents adopted as a reaction to the British idealist tradition (I will return to that shortly) (cf. Hylton 2003: 207–208). As for the

question of what such constituents of propositions do, the answer is the same as in cases of more typical candidates for Russellian propositions. Constituents of Moore's propositions do the same thing that, e.g., *Socrates* or *Aristotle* would do in Russell's propositions. And they occur in propositions in the same way and for the same reason *Socrates* or *Aristotle,* in Russell's case, do. Russell's or Moore's propositions are a means to challenge idealism, and to be able to do so on the ground of propositions, they thought, propositions should not involve any mediation; otherwise, one might end up where idealists are. I will elaborate on that because to see why Moore's propositions are Russellian, not Fregean, it is important to understand what concepts for Moore are and how they relate to propositions on the one hand and the world on the other.

Moore (1993a) starts as a reflection on some of the doctrines proposed in Francis Bradley's *Logic*. He puts the matter as follows: Although in his *Logic*, Bradley attempted to preserve the objective reality independent of one's ideas, he, nevertheless, ended up with ideas alone, fuzzily separating them as something that designat*es* and as something designat*ed* (cf. Russell 1992: 47). In his reaction to Bradley's idealist conception, Moore took the radical realist stance on the issues Bradley dealt with in *Logic*. Accordingly, he substituted "Bradley's ideas" with objectively existing "logical ideas." Bradley called such entities "universal meanings," and Moore decided to call them "concepts." For him (1993a: 4), concepts (including both properties and relations) are not psychological (subjective) or linguistic entities. They exist objectively and are related to language and thought only as their objects (in the way a ball is the object of someone's kicking) but ontologically independent of such a relation. Concepts are not created. They are causally inert, incapable of change (1993a: 4–5), and something immediately known (1993a: 6), be they empirical or a priori (1993a: 14).[8] As it turns out, Moore's proposal here seems to be a peculiar realist version of Berkeley's (1998) bundle theory (minus the God). He writes:

> All that exists is thus composed of concepts necessarily related to one another in specific manners, and likewise to the concept of existence. I am fully aware how paradoxical this theory must appear, and even how contemptible. But it seems to me to follow from premisses generally admitted, and to have been avoided only by lack of logical consistency. (Moore 1993a: 6)

And continues afterwards along Berkeleyan lines:

> It seems necessary, then, to regard the world as formed of concepts. These are the only objects of knowledge. They cannot be regarded fundamentally as abstractions either from things or from ideas; since both alike can, if

[8] Following Moore, Russell adopted the outlined metaphysical characterisation of concepts and applied it to all *terms*: "[E]very term is immutable and indestructible […] no change can be conceived in it which would not destroy its identity and make it another term" (1992: 44). For a further discussion about Russell's *terms*, see Cartwright (2003: 115ff.). I return to Moore's impact on Russell in the following section.

> anything is to be true of them, be composed of nothing but concepts. A thing becomes intelligible first when it is analysed into its constituent concepts. The material diversity of things, which is generally taken as starting point, is only derived […]. (Moore 1993a: 8)[9]

If Moore intends concepts to supplant Bradley's ideas, one might think that concepts accordingly have the same function Bradley's ideas do, with the sole difference of being external and objective rather than mental and subjective. In one sense, that is true. If Bradley's ideas are something with the help of which one comes to know the world (whatever it may be), concepts coincide with ideas. If Bradley's ideas make words and sentences meaningful, concepts also coincide with ideas by that feature. And if Bradley's ideas are all one ultimately needs, and thus all that ultimately exists, as Moore (1993a: 1–3) suggests it holds in Bradley's case, then by that feature, concepts coincide with Bradley's ideas, since for Moore concepts are all there is (this last thesis is particularly important for understanding and classifying Moore's conception of propositions.). But if Bradley's ideas *represent* something other than ideas or even other ideas, then concepts do not coincide with Bradley's ideas, not by that feature.

Precisely here lies the crucial point for understanding Moore's conception of propositions, without which one could hardly assign it the proper label, "Fregean" or "Russellian." For Moore (1993a: 4–6), propositions are entities composed of at least two concepts that stand in a specific relation to one another. The truth or falsity of a proposition does not depend on what exists in the world independently of the proposition and the correspondence between the proposition and the existent. Instead, it depends on the nature of the relation between concepts within the proposition. Indeed, since concepts are all that exists, the notions of correspondence and representation become utterly redundant. All one could ultimately have are simple concepts (such as *red*), complex concepts formed out of the simple(r) ones (such as *rose*), and propositions composed of simple or complex concepts connected by a specific relation (for example, the proposition *that this rose is red*). By this characterisation of the constituents of propositions, concepts for Moore in no way coincide with Frege's senses or alternative constituents of Fregean propositions besides being abstract. Moore's concepts are not representational; Frege's senses are, and other sense-like entities within the later Fregean semantic tradition are also supposed to be of the kind.

[9] See also Moore (1993a: 18; 1993b: 21). Just as for Berkeley (1998), the exception here would be the particular knowing subjects. It should be noted that, although Russell (1992) diverged from Moore's conception in allowing *things* beside and independent of *concepts* (as I already mentioned), in his later writings, he apparently returned precisely to Moore's outlined conception. Thus, one finds Russell later writing: "I wish to suggest that 'this is red' is not a subject-predicate proposition, but is of the form 'redness is here'; that 'red' is a name, not a predicate; and that *what would commonly be called a 'thing' is nothing but a bundle of coexisting qualities* such as redness, hardness, etc." (1961: 97, emphasis added; unlike in his early writing, Russell here uses "proposition" for sentences, not their contents).

Another thing that separates Moore from Frege (and later Fregeans who agreed with Frege on that point) is that Frege presupposes a hierarchy of senses (Frege 1984a; cf. Carnap: 1956: 129). For any particular item, there is the item, senses of that item, senses of senses of that item, senses of senses of senses of that item, etc. That feature of senses allowed Frege to explain the peculiarities of indirect speech or attitude sentences without abandoning the familiar pattern of the explanation he introduced for the "customary" sentences, such as "Socrates is stubborn." One might think that, at least in that respect, Moore's concepts do not differ from Frege's senses. One should, however, bear in mind that even if Moore would allow for such a hierarchy (namely, concepts of concepts, concepts of concepts of concepts, etc.), and I can see no reason why he would not, that would still not justify a Fregean interpretation of the higher-level concepts. They would not be something that uniquely picks out the lower-level concepts and constitutes the meaning of the words in question. But, to my knowledge, Moore never suggested something along these lines, and the concepts he considers are not singular (individual). Singularity, if any, could come only from a specific combination of concepts into a single complex concept. Thus, it is reasonable to suppose that the only hierarchy Moore would have allowed in the case of concepts would resemble the classical realist hierarchy of universals. The realists typically hold that universals directly related to particulars are also directly related to "higher" properties and relations, "higher" properties and relations to "still higher" properties and relations, etc. The particular apple, for example, is directly related to *redness*, *redness* to *colour*, *colour* to *monadic*, etc. Such a sequence would then constitute a hierarchy, but not the one resembling Frege's sense hierarchy.

Now consider the following case: Imagine a reformed Fregean whose metaphysical investigations lead him to conclude, plausibly or not, that senses and complexes of senses are all there is, that senses constitute the world the way Moore's concepts do, and that they are of the single level. No hierarchy of senses thus exists by that metaphysical account. But, for whatever reason, the reformed Fregean still holds that senses are constituents of propositions, just as an ordinary Fregean would. According to that "reformed" conception, propositions would actually be Russellian, not Fregean, even though they would have senses as constituents. Of course, one might protest at this point that such "reformed" senses would not really be *Fregean* because they would not do what Fregean senses are supposed to do, namely, (re)present items (including lower-level senses) in a unique manner. That is true (although Frege allowed senses that present nothing), but it does not undermine the point here: One might have entities that resemble Fregean senses in other respects save their function. For that reason alone, one would not have Fregean but Russellian propositions if such senses were their constituents.

4. *Mediation and the puzzling "concept"*

The existence of senses does not make a conception of propositions Fregean, but the particular assumption about what senses within propositions do. In the Fregean case, the assumption is that senses are *identifying mediators* between items and bits of language (or certain psychological states), which, *as mediators*, enter propositions that are themselves identifying mediators. If senses were not mediators but would still, for whatever reason, enter propositions as their constituents, such propositions would not be Fregean. Russell provides an example.

Up to now, Russell was identified as a classical proponent of—not surprisingly—Russellian conception of propositions. But there is a point at which Russell's (1992) conception turns roughly Fregean and where his position, unlike Moore's, might be classified as "broadly Fregean." This is precisely the point that nicely illustrates the proposed demarcation criterion governed by the question of what constituents of propositions do. Namely, early Russell seems to be on the same track as Frege when it comes to denoting phrases like "a man," "the present queen of England," "any number," or "all dogs" (cf. Hylton 2003: 214). Here is how Russell (1992: 53) puts it:

> A concept *denotes* when, if it occurs in a proposition, the proposition is not *about* the concept, but about a term connected in a certain peculiar way with the concept. If I say 'I met a man', the proposition is not about *a man*: this is a concept which does not walk the streets, but lives in the shadowy limbo of the logic-books. What I met was a thing, not a concept, an actual man with a tailor and a bank-account or a public-house and a drunken wife.

What makes this a Fregean addendum to Russell's otherwise Russellian conception of propositions is not the kind of entity that could now occur within some propositions—the concept—since entities of the same kind occur in cases of Russell's previously considered propositions, too (Russell 1992: 48). But in Russell's propositions considered so far, concepts as their constituents did nothing logically in addition to occurring within them. Indeed, in the case of the proposition *that Manhood belongs to Socrates*, the proposition is about the concept *man(hood)*. In denoting cases, however, even when concepts occur as subjects of propositions—as in the proposition *that a man walked into the bar*—the proposition is not about the concept *a man* but about what that concept denotes instead. And this case significantly differs from the proposition *that 'a man' is a denoting concept*. The latter proposition is about the denoting concept *a man* and, at the same time, contains another denoting concept, namely, the concept *a denoting concept*, which functions differently from the first one within this proposition. And one could go further along the same lines. For example, one could say (adopting Russell's italic letters convention), "*A denoting concept* does not denote in the proposition expressed by this very sentence, but *the proposition expressed by this very sentence* does."

Denoting concepts thus do more than merely occur within propositions as inactive constituents on par with *Socrates* or (the number) *nine*; they denote. As constituents of propositions in which they actually denote, denoting concepts are about something other than themselves, something which is typically not a constituent of these same propositions.[10] As Russell puts it, denoting concepts "are symbolic in their own logical nature" (1992: 47; see Hylton 2003: 207ff. for a more detailed overview). Russell soon became discontented even with this restricted Fregean burden of his theory. A year later, he writes to Frege: "In the case of a simple proper name like 'Socrates', I cannot distinguish between sense [*Sinn*] and meaning [*Bedeutung*]; […] I see the difference between sense and meaning only in the case of complexes whose meaning is an object, […] But I admit that there are certain difficulties in this view" (Russell 1980: 169; cf. Russell 1992: 47). And a year after that letter, he completely eliminated the notion of denoting concepts as Fregean constituents of propositions from his explanation, supplanting it now with the well-familiar apparatus of contextual definitions accompanied by a cryptic criticism of the meaning/denotation distinction (Russell 1968; cf. Hylton 2003: 219–222).

In a sense, Frege's analysis of the attitude or indirect speech sentences supports that, too (e.g., Frege 1980b: 163–165). When a sense becomes the object of discourse or thinking, it no longer performs its function relative to that context. The sense to which a word refers within an indirect construction (e.g., "Plato" in "Aristotle claimed that Plato was on the wrong track," or the whole "that"-clause "that Plato was on the wrong track"), although by its nature still a mediator, that is, a (re)presentation of something, its representational character is irrelevant relative to that particular case. Therefore, the customary sense of "Plato" does not enter the proposition expressed by the whole sentence "Aristotle claimed that Plato was on the wrong track." And one can quickly think of a sentence in which tokens of the same name within the same sentence are not coreferential, say, "Aristotle believed that Plato was dead, although at that time Plato was still alive."

Therefore, drawing the distinction between Fregeans and Russellians is not primarily about what the constituents of propositions *are* but what such constituents within propositions *do*. This is where Russellians and Fregeans primarily disagree. I will point out another example to support the claim further.

The point about what-constituents-do-rather-than-what-they-are also gets supported if one considers Kaplan's characterisations of singular (i.e., Russellian) propositions (which are opposed to Fregean propo-

[10] Again, one could think of examples where precisely the untypical happens (adopting Russell's italic letters convention): The sentence "A man is not a denoting concept, but *a men* is" expresses the proposition containing the concept *a man* both as a denoting concept *and* as an inactive item (analogously to sentences such as "A man is not a denoting phrase, but 'a man' is"). That, however, in no way goes against what I have said here. For a related discussion, see Russell (1968: 45–51).

sitions). One can detect a number of places in Kaplan's writings where he expresses it clearly. Here are several examples: "[…] certain singular terms refer directly without the mediation of a Fregean *Sinn* as meaning. […] the proposition expressed by a sentence containing such a term would involve individuals directly rather than by way of the 'individual concepts' or 'manners of presentation'" (Kaplan 1989a: 483). Or: "Directly referential expressions are said to refer directly without the mediation of a Fregean *Sinn*. […] the relation between the linguistic expression and the referent is not mediated by the corresponding propositional component, the content or what-is-said" (Kaplan 1989b: 568). Or: "The 'direct' of 'direct reference' means unmediated by any propositional component, not unmediated *simpliciter*. The directly referential term goes directly to its referent, *directly* in the sense that it does not first pass through the proposition" (Kaplan 1989b: 569). Notice that Kaplan mentions no metaphysical feature (such as *abstract* or *concrete*) in his characterisations of singular propositions.

Then, according to Fregean conception, constituents of propositions are *mediators* between referents and expressions; it is what they do. They (re)present referents in a certain way. For the Russellian conception of propositions, there is no *such* mediation. Rather, Russellians will typically hold that the directly referential terms within a sentence refer to objects (referents) via causal or historical chains which do not enter the propositions expressed by the sentence. Constituents of propositions are referents themselves and do nothing in addition to that. In the three quoted passages, Kaplan does not mention other features of propositional components for a good reason.

Regarding the Fregean/Russellian distinction, all other features of such components are irrelevant unless they have direct bearings on the question of what propositional components do. And whether propositional components are abstract or concrete certainly has no such bearings. Moore, for example, thought that the world and propositions consist of concepts; concepts are abstract, yet Moore's propositions are not Fregean but Russellian.

However, I do not think the mischaracterisation of Fregean and Russellian propositions is the only reason why McGrath and Frank characterise Moore's propositions as "broadly Fregean." I believe they would characterise them in the same way even if their characterisation of Fregean and Russellian propositions would be entirely in order and in complete accordance with Kaplan's characterisation. The main reason they characterise Moore's propositions the way they do, I suspect, is that Moore's characterisation of propositions—especially his "concept" talk—*sounds* much like something some later Fregeans would say. It was already mentioned that Kaplan (1989a: 483) characterised Russellian propositions as entities that "involve individuals directly rather than by way of the 'individual concepts' or 'manners of presentation.'" And "individual concept" is the term taken from Carnap and Church, not Frege.

Carnap (1956), for example, distinguishes the extension of an expression from the expression's intension (the former is an entity to which the expression refers, the latter the concept of that entity expressed by the expression). Thus, the distinction is intended to be an adaptation of Frege's sense/meaning (reference) distinction. Then he introduces the term "concept" "as a common designation for properties, relations, and similar entities," which are intensions of expressions (Carnap 1956: 21). And then he writes things such as: "let us look for entities which we might regard as intensions of individual expressions. […] Now it seems to me a natural procedure, in the case of individual expressions […] to speak of concepts, but of concepts of a particular type, namely, the individual type" (Carnap 1956: 40–41). And Church (1964: 438–439) writes along similar Fregean lines:

> A name is said to *denote* its denotation and to *express* its sense, and the sense is said to be *a concept of* the denotation. The abstract entities which serve as senses of names let us call *concepts* […] Thus anything which is or is capable of being the sense of some name in some language, actual or possible, is a concept. The terms *individual concept*, *function concept*, and the like are then to mean a concept which is a concept of an individual, of a function, etc. A *class concept* may be identified with a *property*, and a *truth-value concept* (as already indicated) with a proposition.

Thus, both Carnap and Church take *concepts* to be precisely what Frege called "senses" or "modes of presentation."[11] Reading Moore not too carefully with the intensional semantics tradition in mind easily leads to interpreting his position along these lines. All one needs to do is to combine McGrath and Frank's characterisation of Fregean propositions "as complexes of senses or abstract entities" with Church's stipulation that concepts will be "abstract entities which serve as senses of names" and then note that Moore treated propositions as entities composed of concepts. But, as I have argued, such identification is licenced by nothing Moore says about concepts in his paper. Indeed, as one can notice, what is indicative in Church's quote is the repeating phrase "concept of," which displays the representational nature of Fregean senses. No phrase of this kind (in this context, at least) ever occurs in Moore's paper.

In addition, one should consider Russell's (1992: xxiii, 24, 44) repeating acknowledgements to Moore's (1993a) conception as the source of influence.[12] These acknowledgements, too, support the claim that

[11] One should note, however, that, from Frege's perspective, both Carnap and Church would make sort of a category mistake here since they identify some of the concepts with properties and relations. But, strictly speaking, properties and relations are at the same level as objects (*Socrates*, for example). They are all *items* (as previously defined). Frege avoids this by distinguishing concepts from *senses* of concepts, and only the latter ones enter propositions according to him. For related point, see Gabriel (2004: 2, 12).

[12] For example: "On fundamental questions of philosophy, my position, in all its chief features, is derived from Mr G. E. Moore. I have accepted from him the non-existential nature of propositions (except such as happen to assert existence) and

Russell is Moore's heir on this point, not a proponent of the rival conception. Russell also remarks that "[t]he notion of a term here set forth is a modification of Mr G. E. Moore's notion of a concept in his article "On the Nature of Judgment, […], from which notion, however, it differs in some important respects" (1992: 44, footnote). The modification, i.e., difference, Russell had in mind here concerns the nature of constituents of propositions, not what these constituents do and how they figure into propositions.[13] In particular, Russell allows constituents of propositions, which are neither concepts nor bundles of concepts, namely, *things*. But he also allows propositions consisting only of concepts, and all such propositions are on par with Moore's propositions (unless, of course, a denoting concept occurs in them). A previously considered example was the proposition *that redness is not relational*. Thus, the denoting cases aside, Russell's departure from Moore has nothing to do with the issue concerning the nature of the propositions from the perspective of the Russellian/Fregean distinction. Both Moore's and Russell's propositions are Russellian. Not that it matters much now, but given the characterisation of Moore's and Russell's propositions, as well as the fact that they were first proposed by Moore (1993a) in 1899 and only then adopted by Russell in the short period to come, *Russellian* propositions would be more appropriately labelled "Moorean."[14]

References

Berkeley, G. 1998. *A Treatise Concerning the Principles of Human Knowledge*. J. Dancy (ed.). Oxford and New York: Oxford University Press.

Caplan, B. 2007. "On Sense and Direct Reference." In M. Davidson (ed.). *On Sense and Direct Reference: Readings on the Philosophy of Language*. Boston and Toronto: McGraw Hill, 2–16.

Carnap, R. 1956. *Meaning and Necessity: A Study in Semantics and Modal Logic*. 2nd ed. Chicago and London: The University of Chicago Press.

Cartwright, R. L. 2003. "Russell and Moore, 1898–1905." In N. Griffin (ed.). *The Cambridge Companion to Bertrand Russell*. Cambridge: Cambridge University Press, 108–127.

Church, A. 1964. "The Need for Abstract Entities in Semantic Analysis." In J. A. Fodor and J. J. Katz (eds.). *The Structure of Language: Read-*

their independence of any knowing mind" (Russell 1992: xxiii; for a more detailed discussion, see Cartwright 2003).

[13] It should be noted that, as one of the reviewers remarked, Russell's biggest departure from Moore is his theory of *denoting* concepts—and denoting concepts differ from other concepts precisely on the account of what they do, namely, denote. Cartwright (2003: 120) notes Moore was dissatisfied with Russell's theory of denoting concepts from the start.

[14] I thank to the anonymous reviewers for the helpful comments about the earlier version of the paper. The paper was produced within the project *Antipsychologistic conceptions of logic and their reception in Croatian philosophy* (APsiH) at the Institute of Philosophy, Zagreb, reviewed by the Ministry of Science and Education of the Republic of Croatia and financed through the National Recovery and Resilience Plan by the European Union—NextGenerationEU.

ings in the Philosophy of Language. Englewood Cliffs, NJ: Prentice-Hall, 437–445.

Frege, G. 1980a. "Frege to Russell, 28.7.1902." In G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, and A. Veraart (eds.). *Gottlob Frege: Philosophical and Mathematical Correspondence*. Chicago: The University of Chicago Press, 139–142.

———. 1980b. "Frege to Russell, 13.11.1904." In G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, and A. Veraart (eds.). *Gottlob Frege: Philosophical and Mathematical Correspondence*. Chicago: The University of Chicago Press, 160–166.

———. 1984a. "On Sense and Meaning." In B. McGuinness (ed.). *Gottlob Frege: Collected Papers on Mathematics, Logic, and Philosophy*. Oxford and New York: Basil Blackwell, 157–177.

———. 1984b. "On Concept and Object." In B. McGuinness (ed.). *Gottlob Frege: Collected Papers on Mathematics, Logic, and Philosophy*. Oxford and New York: Basil Blackwell, 182–194.

Gabriel, G. 2004. "Introduction: Frege's Lectures on *Begriffsschrift*." In E. H. Reck and S. Awodey (eds.). *Frege's Lectures on Logic: Carnap's Student Notes 1910–1914*. Chicago and LaSalle, IL: Open Court, 1–15.

Hylton, P. 2003. "The Theory of Descriptions." In N. Griffin (ed.). *The Cambridge Companion to Bertrand Russell*. Cambridge: Cambridge University Press, 202–240.

Kaplan, D. 1989a. "Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals." In J. Almog, J. Perry, and H. Wettstein (eds.). *Themes from Kaplan*. New York and Oxford: Oxford University Press, 481–563.

———. 1989b. "Afterthoughts." In J. Almog, J. Perry, and H. Wettstein (eds.). *Themes from Kaplan*. New York and Oxford: Oxford University Press, 565–614.

Loux, M. J. and Crisp, T. M. 2017. *Metaphysics: A Contemporary Introduction*, 4th ed. New York and London: Routledge.

McGrath, M. and Frank, D. 2023. "Propositions." In E. N. Zalta and U. Nodelman (eds.). *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.

Moore, G. E. 1993a. "The Nature of Judgement." In T. Baldwin (ed.). *G. E. Moore: Selected Writings*. London and New York: Routledge, 1–19.

———. 1993b. "Truth and Falsity." In T. Baldwin (ed.). *G. E. Moore: Selected Writings*. London and New York: Routledge, 20–22.

Russell, B. 1961. *An Inquiry into Meaning and Truth*. London: George Allen and Unwin LTD.

———. 1968. "On Denoting." In R. C. Marsh (ed.). *Logic and Knowledge*. London: George Allen and Unwin LTD, 41–56.

———. 1980. "Russell to Frege 12.12.1904." In G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, and A. Veraart (eds.). *Gottlob Frege: Philosophical and Mathematical Correspondence*. Chicago: The University of Chicago Press, 166–170.

———. 1992. *The Principles of Mathematics*. London: Routledge.

Schiffer, S. 2006. "Propositional Content." In E. Lepore and B. C. Smith (eds.). *The Oxford Handbook of Philosophy of Language*. Oxford and New York: Clarendon Press, 267–294.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.7
Received: January 20, 2023
Accepted: November 24, 2023

Reclaiming Russellian Singular Thoughts

HEIMIR GEIRSSON
Iowa State University, Iowa, USA

There is an important difference between a thought that is directed towards a particular object and a thought that is not so directed. For example, there is a difference in my thoughts about my brother, and my thoughts about brothers, more generally. The first has the earmarks of singular thought, while the latter does not. After showing that there is no agreement about the nature of singular thought, I revisit early Russell to find greater clarity. I then advance a version of Millianism that builds on early Russell's view of singular thought. I argue that the advocates of the direct reference view who argue that being on the receiving end of a name is sufficient for having singular thoughts about the object named have not provided good reasons for their views. Passing on a name can provide the recipient with a general understanding of the name, but not specific understanding. That is, when acquiring the name, the recipient may not learn the identity of the object named as this very object which, I argue, is required for one having singular thoughts of that object.

Keywords: Singular thought; direct reference; acquaintance.

1. *Singular and general thoughts*

There is an important cognitive difference between a thought that is directed towards a particular object and a thought that is not so directed but is instead about a certain kind of object. For example, there is a difference between my thoughts about my brother and my thoughts about brothers more generally. The former is an example of a singular thought (often called de re thought) while the latter exemplifies general thought. Similarly, when I come upon a particularly grizzly murder then there is an important cognitive difference between my thought

that Smith's murderer is insane when I know who the murderer is and when I do not have such knowledge. In the former case my thought is about a particular person while in the latter case it is not about a particular person but rather about the murderer *whoever it might be*. Again, the former exemplifies singular thought and the latter general thought.

While there is a general agreement about there being singular thoughts and general thoughts there is little agreement on what exactly constitutes a singular thought.[1] Similarly, there is not much agreement on the conditions for one acquiring singular thoughts. In particular, the disagreement about the latter focuses on the one hand on the epistemic requirement of acquaintance and, on the other hand, on the metaphysical requirement of existence. Even those who insist on an acquaintance requirement for singular thoughts do not agree on the strength of the acquaintance relation.

Direct reference theorists who identify as Millians generally accept a very weak requirement when it comes to acquaintance, namely the view that one can be sufficiently acquainted with an object and so obtain a singular thought about it in virtue of being on the receiving end of a use of a name of the object that stretches back to an initial baptism of it. I will argue that we have good reasons to doubt that singular thought comes so cheaply. Instead, I will argue that we, including the Millians, should stay close to Bertrand Russell's 1903 view, when he held that we could be acquainted with ordinary objects. At that time Russell introduced a distinction that is sensible and important, but one that has been lost by relaxing the acquaintance constraint too much.

The paper will proceed as follows. I will first present three general constraints on singular thoughts as well as examples of prominent views that relax and/or reject some of these. The discussion will show that there are very significant disagreements about even foundational issues of singular thought. The disagreement will lead me to search for a fresh start, and to that effect I will provide an overview of Russell's 1903 view of singular thought, a view that I find sensible and valuable and a view that differs significantly from the 1912 view that he when he held that we can only be acquainted with sense data, universals, and oneself. The main sections of the paper will develop a Millian version of singular thought in the tradition of Russell's 1903 view. I will argue that while being on the receiving end of a causal chain linking a name to an object is sufficient to secure the reference of the name, having singular thought of the object requires acquaintance with the object and having paid conscious attention to it.

[1] Admittedly, Hawthorne and Manley (2012) have used the lack of agreement on what constitutes singular thought to argue that there are no such thoughts.

2. *(Severe) Lack of agreement on singular thoughts*

Sarah Sawyer has listed three constraints that guide the various views on singular thoughts:

> *Content constraint*: the object is thought about directly and not descriptively.
> *Metaphysical constraint*: there is an object thought about.
> *Epistemic constraint*: the subject is acquainted with the object thought about. (Sawyer 2012: 270)

A quick review reveals that there is little agreement about Sawyer's constraints.

Semantic instrumentalists about singular thought maintain that it is sufficient for having a singular thought that one introduces a name and so they only accept the content constraint.[2] Accordingly, instrumentalists maintain that we can have singular thoughts after introducing directly referring terms by means of Kaplan's "dthat," and by doing so converting an arbitrary singular term into a directly referring term, thus enabling singular thought about the term's referent.[3]

Robin Jeshion rejects both the epistemic and the metaphysical constraints, as she claims that one can have singular thoughts about something that does not exist. When discussing the case where Leverrier introduces the name "Vulcan" and then entertains a thought such as "Vulcan is a planet" she writes "Intuitively, it seems to me [...] plausible to hold that [the Vulcan case is an instance] in which an agent has a singular, non-descriptive belief [...] I wish to carve out a theory that respects these intuitions" (Jeshion 2010: 117–118).[4]

While Jeshion only accepts the content constraint above, she does add a condition that distinguishes her view from that of the instrumentalists. For Jeshion, one thinks a singular thought by thinking *through* or *via* a mental file that one has of the relevant object. And one only forms a mental file of an object if Jeshion's significance condition is satisfied.

> *Significance condition*: a mental file is initiated on an individual only if that individual is significant to the agent with respect to her plans, projects, affective states, motivations. (Jeshion 2010: 136; 2014: 83)

Both Jeshion and Francois Recanati make singular thoughts dependent on mental files.[5] Recanati, however, denies that we can have a successful empty singular thought, while claiming that we can have

[2] Both Kaplan and Harman have advocated instrumentalism (Harman 1977; Kaplan 1989b).

[3] For any definite description f, dthat(f) refers directly to the object that satisfies the description. Accordingly, dthat(the largest whale in the ocean) refers directly to the largest whale in the ocean. See, for example, Kaplan (1989b).

[4] I discuss Jeshion's view in Geirsson (2018).

[5] See for example Recanati (2012; 2021).

the vehicle for singular thought (a mental file) in an empty case. Recanati accepts the content constraint, the metaphysical constraint, as well as a modified version of the epistemic constraint, namely

> *Epistemic constraint*$_{FR}$: the subject is acquainted with or correctly anticipates becoming acquainted with the object thought about.

The modified epistemic constraint allows one to be acquainted with an object by being on the receiving end of a causal chain of a name that stretches back to that object. Additionally, it allows one to have singular thoughts about objects one is not acquainted with provided that one will be appropriately acquainted with them at some point. For example, when Leverrier hypothesized that Neptune exists then Recanati claims that he has a singular thought when thinking, for example, "Neptune is a planet." The reason Recanati gives is that "Neptune" is a referring name and so the resulting thought is truth-evaluative and as such qualifies as a singular thought and, furthermore, Leverrier did perceive the planet during his days.

While rejecting the epistemic requirement is the norm for semantic instrumentalists, advocates of the direct reference view have a history of accepting a very weak version of the requirement. Consider the following example. I am looking over the list of students in my class. I have not met any of the students and have never had any contact with them. As I look at the list, I think to myself, "Jessica Alba is taking my class," "Jessica Alba" being the first name on the roster. It is commonly accepted by direct reference theorists that I have a singular thought about Jessica, the reason being that the name is passed on to me via a causal chain following a baptism, and that I intend to use the name with the same reference as those I acquired it from. Nathan Salmon (Salmon 2004) points out that the *de re* connection need not be direct and intimate. Instead it may be remote and indirect, perhaps consisting of a network of causal intermediaries interposed between the cognizer and the object. Advocates of this view include Nathan Salmon, Scott Soames, and Robin Jeshion, to name a few.[6] The view, generally, assumes that names bring objects into thought, resulting in singular thought or de re thought of that object.

For direct reference theorists accepting or sincerely assenting to a sentence that expresses a singular proposition is generally deemed sufficient for having a singular thought. Accordingly, when I hear from what I take to be a reliable source that Thales was a philosopher, then I come to believe a singular proposition containing Thales as a constituent and so come to have a singular thought about Thales. The direct reference theorists are therefore likely to accept all three of Sawyer's constraints. The metaphysical constraint, which appears to cause problems with cases of seemingly empty names such as "Vulcan" and "Santa Claus," is often limited in scope by introducing literary, theoretical,

[6] See for example Salmon (1986), Jeshion (2002) and Soames (1995).

and/or imaginary objects that, in some sense, exist.[7] We therefore have philosophers who argue that, e.g., "Santa Claus" refers to an object,[8] albeit not an ordinary object, and that one can therefore have singular thoughts about Santa Claus. Singular thoughts remain object dependent on these views.

While the direct reference theorists tend to make singular thoughts object dependent, Tim Crane, Mark Sainsbury, and Jody Azzouni, in addition to Jeshion, want to allow singular thoughts about objects that do not exist (Azzouni 2011; Crane 2013; Sainsbury 2005). It seems that any characterization one gives of singular thoughts should, at least initially, be open to the possibility of such thoughts not being object dependent. However, the main issue facing us is how to account for singular thoughts of ordinary objects.

When moving on we need an account of singular vs general thought that is useful and at the same time does not come with too much theoretical baggage. For example, it is preferable that such an account does not saddle one with a commitment to mental files and/or metaphysical presuppositions about the objects of thought. I will suggest below that we can find such an account in early Russell.

3. *Russell and singular thoughts*

The discussion of singular thought can be traced back to Bertrand Russell. While the discussion during recent decades has been driven primarily by semantic concerns having mostly to do with direct reference, Russell's reasons for introducing singular thoughts focused more on epistemology and philosophy of mind, i.e., representation. In *Points About Denoting*, dating from 1903, Russell writes:

> [...] if I ask: Is Smith married? And the answer is affirmative, I then know that 'Smith's wife' is a denoting phrase, although I don't know who Smith's wife is. We may distinguish the terms [objects, individuals] with which we are *acquainted* from others which are merely denoted. E.g. in the above case, I am supposed to be acquainted with the term [object, individual] *Smith* and the relation *marriage*, and thence to be able to conceive a term [object, individual] having this relation to Smith, although I am not acquainted with any such term [object, individual].
>
> [...] we know that every human being now living has one and only one father [...] This shows that to be known by description is not the same thing as to be known by acquaintance, for 'the father of x' is an adequate description in the same that, as a matter of fact, there is only one person to whom it is applicable. (Russell 1994: 306)

[7] Admittedly, most of the advocates of this approach accept some, but not all of literary, mythical, or fictional objects, thus still leaving a problem of empty names. See for example Salmon (1998, 2002). Also Braun (2005). Even Braun, who is existentially most generous of the above, claims that there are still some empty names. See Braun (2021).

[8] A clear example here is Azzouni (2021).

So, one has *direct knowledge*, knowledge by acquaintance, of those objects that one is acquainted with. One can have knowledge by description of those objects with which one is not acquainted. The latter enables us to think about objects with which we are not acquainted.

It is interesting that at this time, when Russell first introduced his distinction, he uses knowledge of an individual, namely Smith, as a paradigm example of knowledge by acquaintance. Clearly, he thought that one could be acquainted with individuals, and presumably other ordinary objects, via perception. On the other hand, we can extend our knowledge beyond that with which we are acquainted via knowledge by descriptions. Descriptions are denoting phrases that denote the objects that uniquely satisfies them and so we can have knowledge of and talk about, for example, Smith's wife, Triphena.

On Russell's view some propositions contain objects. For example, the proposition expressed by *Smith is married* contains Smith. If I am to be able to believe the proposition expressed by *Smith is married* then, somehow, I need to turn Smith into a cognitive object. Acquaintance allows for that to happen. If I am acquainted with Smith, he is a constituent of the proposition expressed when I think or say *Smith is married*. But since I am not acquainted with Triphena, she is not a constituent of the proposition expressed when I think or say *Triphena is married*. Instead, the proposition contains a denoting complex.

The introduction of sense data changed the picture outlined above, but one can view the change as resulting from tightening up the acquaintance requirement while leaving the other aspects of the picture as they were. In 1912 Russell writes:

> We shall say that we have *acquaintance* with anything of which we are directly aware, without the intermediary of any process of inference or any knowledge of truths. Thus, in the presence of my table I am acquainted with the sense-data that make up the appearance of my table [...]. (Russell 1961: 191)

At this point Russell would not say that I am acquainted with Smith. Instead, I am acquainted with my sense-data that make up the appearance of Smith. My knowledge of Smith is knowledge by description. He is the physical object that causes such-and-such sense data (Russell 1961: 192).

In 1903 Russell allows that I am acquainted with Smith. In 1912 he does not allow that and instead argues that I am only acquainted with the sense data that make up the appearance of Smith. Acquaintance, if you will, comes on a sliding scale, and Russell has moved the scale so that we no longer can be acquainted with ordinary objects and so we cannot have singular thoughts about them. In 1912 Russell writes:

> Common words, even proper names, are usually really descriptions. That is to say, the thought in the mind of a person using a proper name correctly can generally only be expressed explicitly if we replace the proper name by a description. (Russell 1961: 195)

So, proper acquaintance allows for a type of thought, and it is the type of thought that determines whether a thought is singular or not. A singular thought is not descriptive in nature.

Decades later, most direct reference theorists moved the acquaintance scale in the other direction, claiming that we can have singular thoughts about an object provided that it is at the end of a causal chain of a name that we have acquired. When doing so they gave up Russell's initial requirement that acquaintance requires, at minimum, that one perceive the relevant object. Others allow for singular thoughts of objects when we are only familiar with causal traces of it, such as a footprint, provided that some additional constraints are met. And some, the semantic instrumentalists, allow that we can introduce a name with a uniquely identifying description and thereby come to have a singular thought about the object so named. The question then remains, can such relaxed requirements result in a thought of an object that is not descriptive?

4. *Minimal criteria for singular thought*

There is no uncontroversial account of singular thoughts to be found in recent literature. The direct reference theorists tend to account for singular thoughts in terms of content, where singular thoughts are mental states with singular as opposed to general content. The singular content is presented to us with singular propositions, which have objects and properties as constituents. General content, on the other hand, is presented to us with general, or descriptive propositions. Thus, the sentence

1. Obama is a former president of the United States

is understood as expressing a proposition that can be represented as an ordered couple consisting of Obama and the property of being a former president of the United States.

The direct reference account is not without problems. It allows for a very weak acquaintance relation, namely a name passed on with the intent that it continues to refer to the same object providing sufficiently strong acquaintance relation for one having singular thought. One has to wonder how such weak relation can provide one who so acquires a name with non-descriptive content instead of, e.g., metalinguistic content such as "the person I heard about from so-and-so" or "the person named so-and-so."

I suggest that a minimal criterion be based on the general ideas captured by Russell's initial criteria when he explained that when the thought in the mind of a person using a proper name correctly can

generally only be expressed explicitly if we replace the proper name by a description, then the thought is not a singular thought. We can then think of a general thought about an object as one where the object is thought about in terms of being a possessor of a certain set of properties that it then satisfies. The referent of a general thought is thought about by means of descriptions. An object of singular thought, in contrast, is not thought of in such a way. In that sense singular thought is not satisfactional.[9]

5. *Reference and thought according to the Millian*

A Millian accepts the following: i) beliefs are binary relations, ii) names refer via causal chains, iii) simple sentences express singular propositions, iv) the name or indexical that occurs in a simple sentence contributes its referent to the proposition, and the predicate contributes the property it designates.[10] According to the Millian version of the direct reference theory a reference of a name is not determined by how an object fits a given set of descriptions. Instead, reference is secured via a causal chain, where one user of a name passes it on to another who then intends to use it with the same reference. The reference of the name is therefore not satisfactional. Because the reference of the name is not satisfactional Millians have often been quick to conclude that the thought that results from sincerely assenting to a simple sentence that contains the name of an object, and the thought that one reports with a simple sentence containing the name, is a singular thought. Consequently, since "Obama" is a referring proper name, if John sincerely assents to (1) then the claim is that John has a singular thought about Obama. Similarly, were John to utter

2. I believe that Obama is a former president of the United States

then the claim is that it clearly indicates that John has a singular thought about Obama. The claim is the same if the name is of someone that John is only acquainted with via the name. If John were to sincerely assent to

3. Thales was a philosopher

then the Millians would generally take that as a clear indication that he has a singular thought about Thales. Being on the receiving end of the causal chain of a name, according to the view, is acquaintance enough for having a singular thought about the relevant object. The point is succinctly made by Marleen Rozemond in the following quote:

> (Kripke) points out that many people who use the name 'Feynman' only know that Feynman is an important physicist. Yet they manage to refer to him by using the name [...] It seems clear [...] that they can have *de re* (sin-

[9] Goodman (2018) provides a similar account.

[10] The Millian also accepts semantic innocence, namely that a simple sentence expresses the same content when embedded in belief context.

> gular) thoughts about Feynman by virtue of a causal chain going from their use of a name to a famous physicist. (Rozemond 1993: 278)

Robin Jeshion shares Rozemond's understanding. She puts the point as follows: "If you finally met me, would you thereby better understand the term 'Robin Jeshion?' Surely this is something that the Millian denies" (Jeshion 2001: 130). The point is that according to the advocates of the direct reference view there is no "additional meaning" beyond what is referred to found in names and so there is no "additional understanding" to be had once one has acquired the name. Acquiring a name of an object enables one to have singular thoughts about it.

Rozemond and Jeshion seem to be echoing a point made earlier by Kent Bach when he argued that when a name is passed on "a speaker cannot just express but can actually *display* his *de re* way of thinking of the object and thereby enable the hearer to think of it in the same way" (Bach 1987: 32).[11] However, when Bach explains what he means by someone *displaying his way of thinking* it is clear that he is assuming that the preservation of reference of a name passes on a de re, or a non-descriptive way of thinking as well. He writes "Since the hearer's mental token of the name 'inherits' the same object as the speaker's, the object of the hearer's thought is determined relationally, not satisfactionally" (Bach 1987: 32). The underlying assumption that Bach, who is not a direct reference theorist, appears to be working with is the following:

> *The testimony requirement*: A sufficient condition for one having a singular, or a de re thought of an object is that one acquires a name of the object, the name having been initially introduced with an acquaintance relation.[12]

Note that the requirement is shared by the direct reference theorists, and it allows one to be on the receiving end of a long chain of use, stretching back to an initial baptism, and still have singular thought about the object named. I believe that we should not accept the requirement.

Suppose that Bach tells me about his new neighbor, Travis, and informs me that Travis is newly retired. I pick up the name and form the appropriate belief that I can express by saying that Travis is retired. According to Bach he has displayed his way of thinking about Travis to me and so I now have singular thoughts about Travis. But let us look again at Russell's basic criteria for one having singular thought. Can I express my thoughts about Travis properly without resorting to

[11] Bach (2010) emphasizes his relaxed conditions for singular thoughts. There he writes "[...] even hearing about or reading about [an] individual from someone else who has perceived that individual or who at least has heard or read about that individual from someone who has heard or read about that individual [...] from someone who has perceived that individual" (2010: 57–58).

[12] Granted, Jeshion would not agree with the sufficiency claim, as she would insist on at least the significance condition being satisfied as well.

descriptions? The answer is no. The only thoughts I have about Travis are descriptive, including thoughts such as "Bach's new and newly retired neighbor." Bach, having interacted with Travis, presumably has a wealth of non-descriptive thoughts about him, but none of them are displayed to me or passed on to me with the simple passing on of a name.

Both Keith Donnellan and David Kaplan agree that singular (de re) thought does not come as easy as Rozemond, Jeshion, and Bach assume. In "The Contingent *A Priori* and Rigid Designators," Donnellan presents a skeptical view of anyone being able to acquire a priori de re (singular) knowledge with stipulative descriptive reference fixing (Donnellan 1981). The proposition that Donnellan is primarily concerned with is expressed by the following sentence, presumably uttered by Leverrier when fixing the reference of 'Neptune:'

> If the planet that caused such and such discrepancies in the orbit of Uranus exists, then Neptune is the planet which caused such and such discrepancies in the orbit of Uranus.

Towards the end of the paper Donnellan characterizes the requirement for de re (singular) knowledge by adopting Kaplan's view that one must be *en rapport* with the object. He then emphasizes that one is not *en rapport* with an object if one has to resort to using stipulative reference fixing. The argument can be presented as follows:

1. In order to have de re (singular) knowledge of an object, one must be *en rapport* with it.
2. When having to use stipulative descriptive reference fixing, one is not *en rapport* with the object being named.
3. So, stipulative descriptive reference fixing does not provide one with de re (singular) knowledge.

Of course, the notion of *en rapport* is not spelled out, but the lesson learned from the argument is still rather clear. It is precisely when one does not have direct contact with objects, when one does not perceive the object being named, as Leverrier with regard to Neptune, that one resorts to stipulative descriptive reference fixing. And that seems to be the core of Donnellan's rejection of stipulative descriptive reference fixing enabling one to acquire de re (singular) knowledge about the object named. Because the stipulator is not in the right relationship with the object being named, she cannot acquire de re (singular) knowledge about the object. Instead, the resulting knowledge is de dicto, or descriptive.

In "Afterthoughts," David Kaplan states that a name does not put us *en rapport* with an object and so does not provide one with de re (singular) beliefs about it.

> On my view, acquisition of a name does not, in general, put us *en rapport* (in the language of 'Quantifying In') with the referent. But this is not required for us to use the name in the standard way as a device of direct reference. Nor is it required for us to apprehend, to believe, to doubt, to assert, or to

> hold other *de dicto* attitudes toward the proposition we express using the name. (Kaplan 1989a: 605)

So, acquiring a name is not sufficient for one to have a singular or a de re thought about the object named. A stronger connection is required.

Gareth Evans provides an example that clearly questions Bach's account of displaying or inheriting a way of thinking via the use of names as well as Jeshion's claim about understanding names. Suppose that person X joins a group that is talking about a certain Louis. X listens in for a while and then joins in the conversation with appropriate uses of the name "Louis." It certainly seems that he is, when doing so, successful in referring to the same Louis that his friends are talking about. The discussion is about King Louis XIII. If that is so then Jeshion, as well as most Millians, are committed to attributing to X singular thoughts about Louis XIII. Suppose now that due to some massive errors X comes away from the discussion believing that Louis is a basketball player, Evans comments on this:

> [N]otice how little *point* there is in saying that he (entertains a singular thought about) one French king rather than another, or any other person named by the name. There is now nothing the speaker is prepared to say or do which relates him differentially to the one King. This is why it is so outrageous to say that he believes that Louis XIII is a basketball player. The notion of (singular thought) has simply been severed from all the connections that made it of interest. (Evans 1973: 274)

It appears to me that Evans is pulling on the right intuitions here. Even though the subject in the story comes away using the name "Louis," and even though the subject can use that name to refer to Louis, he did not come away with singular thoughts about Louis. Even though the name is passed on, singular thought is not. This is very much in line with Donnellan and Kaplan. As Kaplan might point out, X has acquired a name that he can use as a referring device, but that does not enable X to have singular (de re) thoughts about the object named. And Donnellan might point out that the discussion can only provide X with content akin to one acquired from descriptive reference fixing; Louis is whoever my friends are talking about. And that is not sufficient for singular (de re) thought.

We now have on the one hand the direct reference view that requires a very weak acquaintance relation for singular thoughts, namely one that can be satisfied by the proper acquisition of a name, and on the other hand we have Donnellan, Kaplan, Evans' example and the example of Bach's new neighbor, all of which suggest that the direct reference view is too permissive. Most Millians, as well as Bach, accept The Testimony Requirement while Evans and Kaplan clearly reject it. The Millians offer as a support for their view the direct reference claim that all there is to the meaning of a name is its referent, while Donnellan, Kaplan, Evans' example and the example of Bach's new neighbor provide support for one acquiring a name not being sufficient for having a singular thought about the object named. There is a way to explain

the intuitions that drive Donnellan's example, Kaplan's view, Evans' example and the example of Bach's neighbor while accommodating the main tenets of the direct reference view and Millianism, but The Testimony Requirement falls by the wayside as a result.

Jeshion claimed that a Millian should maintain that he doesn't understand her name any better after meeting her than he did before doing so. That is right when we are talking about the typical Millian who accepts a very weak acquaintance requirement. But the typical Millian, I believe, is not right. Someone who has never met Jeshion should argue that he does understand the name "Robin Jeshion" better after meeting her than he did before doing so. Before meeting her he had a *general understanding* of the name, that is, he knew the semantic role the name plays as a proper name which suffices to enable him to use the name competently as a referring device. This agrees with Kaplan's view that one can acquire a name and use it in a standard way as a device of reference without being able to have de re (singular) thoughts about the object named. But since he did not know who the referent was, he did not have a *specific understanding* of the name, that is, he did not know that it was *this very individual* who was the semantic value of the name. Since he has specific understanding *and* general understanding of the name "Robin Jeshion" after meeting her, he now has a better understanding of it than he did before meeting her.

Nathan Salmon makes a similar point in a footnote. He writes:

> There may be a weaker sense of 'understand' in which the reference-fixer 'understands' the word 'metre' simply by knowing that it was introduced in such a way that 'one metre' refers to whatever length S has at t_0, if S exists. But understanding 'metre' in this weak sense does not give one the basic semantic knowledge that 'one metre' refers, if S exists, specifically to one metre. (Salmon 1987: 200, n. 210)

Salmon thus allows that one can use a name as a semantic device without having full semantic knowledge of its reference. But while Salmon discusses the possibility of weak and strong understanding in connection with naming, I intend, capturing Kaplan's observation above, specific and general understanding to apply to established names.

When a name is passed on without the hearer being otherwise acquainted with the named object then the hearer can only have a *general understanding* of the name. While having general understanding is sufficient to successfully use the name in a public language it does not provide one with singular thoughts. It is not until one is in a position to have *specific understanding* of the name of the object it refers to, i.e., in a position to have thoughts that are not descriptive in nature, that one can acquire singular thoughts. The understanding that makes singular thoughts interesting and relevant is to be found in the specific understanding of names; the knowledge that the name is of *this very individual*. This is the insight that is reflected in Russell's 1903 account of singular thought.

When I have a general understanding of a name, then I can use it competently and appropriately to refer to its bearer. I then typically have some descriptive beliefs about the object named filed away. The descriptions might not reveal much about the object. Instead, they might be very general in nature, such as descriptions to the effect that I acquired the name in a recent conversation with my friends and, as in the case of Evans' X, many of the beliefs might be false. However, the competent use of a name does not entail that one has a non-satisfactional representation of the bearer of the name in any interesting way. Such representation typically requires one perceiving or having perceived the object. And while a proper name refers to its bearer in a non-satisfactional way, a speaker does not display (in the sense of showing or passing on non-satisfactional ways of thinking about an object) how she represents an object when using that name. A simple example should suffice. When I utter "Arya is fast," speaking to a person who is hearing the name "Arya" for the first time and who knows nothing about Arya, then I have not displayed or shown or indicated how I think about Arya, and I have not displayed whether I am acquainted with Arya. When uttering the sentence, I have not even indicated to my listener that Arya is a dog. And were I to indicate that she is a dog, uttering for example "Arya is fast for a dog," then I have not displayed or revealed when saying so what kind of a dog she is, nor have I indicated what she is fast at doing. In fact, my use of a proper name when passing it on to a new user generally does not display or indicate or show how I think about its bearer. Here the predicates and context are more helpful for a listener. Even so, the resulting thought will not be a non-satisfactional thought about Arya. Instead, the listener will have descriptive thoughts about her, such as "the dog I talked about with so-and-so," or "the fast dog," or "the dog named Arya."[13]

The Testimony Requirement assumes that it is sufficient for one to have a general understanding of a name in order to have a singular thought about the object named. But the distinction between general and specific understanding of names explains the appeal of Evans' example as well as the example of Bach's neighbor. While the subjects in the examples have general understanding of the names "Louis" and "Travis," they do not have specific understanding of the names. Having general understanding of a name is not sufficient for one having singular thought about the object named as such understanding only provides general thoughts. While the causal connection between an object named and the use of the relevant name secures non-satisfactional reference, it does not provide the information needed for one to have

[13] It is fairly evident that acquiring names is not a necessary condition for singular thoughts as one can have singular thoughts about something without having a name for that object. I can, for example, have singular thoughts about a soccer ball that I am trying to juggle without me having a name for the ball and even without formulating any thoughts that explicitly use names or indexicals to refer to it.

non-satisfactional thoughts, singular thoughts, of the object named. Something more is required for that.

Several philosophers have suggested that causal connections other than the one required by testimony are sufficient for one having singular thoughts about objects. For example, Jeshion and Recanati allow that Leverrier had singular thoughts about Neptune without ever perceiving the planet. It suffices, on their account, that he has seen the appropriate causal traces of Neptune, namely the perturbations in the orbit of nearby known planets and that, of course, he satisfies Jeshion's significance condition and Recanati's requirement that he later become more directly acquainted with it. Similarly, one can, as Jeshion suggests, have a singular thought about a nearby bear even though one has only encountered the bear's scat. Upon seeing the fresh scat, one might think "he is close to us," thus entertaining a singular thought about the bear.

Relying on causal connections of the kind described above is not likely to be helpful in clarifying the nature of singular thought, as these connections are too permissive. I am causally connected to the person who finalized the online purchase of the endnote program that I am currently using, I am causally connected with the person who drove my car off the assembly lot wherever it was assembled, I am causally connected with the person I never see who assembled my hamburger at a drive-through, and I am causally connected with the person who made the final inspection of the shirt that I am wearing. Such connections do not enable me to have singular thoughts about the relevant people, regardless of how much I otherwise care about them and regardless of whether I at some point in the future I will meet these people.[14]

6. *Strong acquaintance and conscious attention*

Testimony and causal connections are too weak to provide one with singular thoughts. While passing on a name secures reference and provides a general understanding of a name, it does not provide a non-descriptive representation of the object named.

Someone might suggest at this point that we might resort to referential use of descriptions and when doing so allow singular thoughts to be descriptive. The idea would then be that I can employ the distinction between referential and attributive uses of descriptions to appropriately connect with the object of thought. For example, while it appears that I cannot have a singular thought about the person who assembled my burger, I might use the description "the person who assembled this burger" referentially to pick out that very person. But this approach will not work. When Keith Donnellan introduced the referential/attributive distinction then one of the important differences between the

[14] Similar points have been made by Jody Azzouni (2011) and Filepe Martone (2016).

examples of the two uses was that one could identify the referent when one used a description referentially as *this very person/object*. In the case of attributive uses, on the other hand, the referent could not be so identified. Instead, one referred to the object or person who fit the description *whatever or whoever it is*.[15] My use of the description "the person who assembled this burger," in this light, has to be attributive. It is no different from my use of "the person who drove my car off the assembly lot" in the regard that I cannot identify the person beyond that. It is the person who fits the description, *whoever it is*.

In the non-controversial cases of singular thoughts of ordinary objects, the one having the thought has perceived the object the thought is of. It is not controversial that I have singular thoughts about my spouse, my parents who raised me, my children whom I helped raise, and the soccer ball that I regularly try to juggle with less than stellar results. In each of these cases I am directly acquainted with the relevant objects. That is, I have perceived them. But the non-controversial examples are also examples of objects that I have paid conscious attention to, that is, the kind of attention that allows me to indicate that it is directed at *this very object*, and the examples thus satisfy what I think is a second necessary condition for one having singular thoughts of ordinary objects. The examples below show why perceiving an object and paying attention to it is not sufficient for one having singular thoughts about it and why *conscious attention* is needed as well.

Consider first an example that most have encountered in some form, where I drive or walk some distance towards my destination. Once I safely arrive at my destination, I realize that I cannot recall what I encountered on my way there. Clearly, I was paying some kind of attention to my environment and there is a clear sense in which I perceived various obstacles as I managed not to run into them as I navigated towards my destination. But this kind of a focused attention is not the kind of attention that allows one to acquire singular thoughts about various objects that one encounters.[16] While I clearly perceived various objects on my way and paid enough attention to them not to run into them or stumble over them, I cannot recall any of them once I reach my destination. I have no current representation of these objects and no beliefs about them.[17]

[15] See Donnellan (1966). Anne Bezuidenhout (2021) argues that the referential/attributive distinction is in fact an epistemic distinction with different uses representing differences in the epistemic access to the entity denoted by the description.

[16] For more on the various kinds of attention see Montemayor and Haladjian (2015).

[17] Someone might suggest here that Pylyshin's fingers of instantiation, FINSTs, provide unconscious content to mental files. However, FINSTs lock onto objects and so allow us to track them in a way that is independent of our representation the objects. FINSTs provide links to object files without endowing them with content and so without providing any representation of the object being tracked, nonconceptual (in a philosophically relevant way) or otherwise. What FINSTs do

Or consider an example of a face in the crowd. When I encounter a crowd of people, I might scan the crowd and take in its size and diversity. When doing so I might not pay attention to any particular individual. While I might have singular thoughts about the crowd at this point, I do not have singular thoughts about any of its members. That changes when I, for some reason, focus on one particular face in the crowd. At that point I am paying conscious attention to that very person and so I am able to have singular thoughts about that person. When paying such attention to the face in the crowd I satisfy how Monemayor and Haladjian characterize conscious attention; namely as one that "requires a demonstrative awareness of attending to a specific object (e.g., "that" or "this" object). Such attention also entails voluntarily maintaining attention to an external object that has been perceptually selected" (Montemayor and Haladjian 2015: 229). On their account conscious attention must include contents that are available for thought and report (2015: 143). The kind of attention that I paid to my environment when driving to my destination did now provide me with content that was available for thought and report and so I was not paying conscious attention to my environment at the time. While scanning the crowd does provide me with content that is available for thought and report, it is only when I focus my attention on a specific person that I can attend to *that* person specifically. Conscious attention paid to that person enables me to have singular thoughts, non-satisfactional thoughts, about the person.

Consider again the example of Arya. Can I perhaps show you a picture of Arya and in doing so enable you to have a singular thought about her? While Russell did not discuss that possibility, perhaps we should accept *some* intermediaries as sufficient for one acquiring a non-satisfactional representation that we can say is of *this very object*. While I have never met Obama, I have seen photographs of him as well as TV footages and interviews that feature him prominently. Given the faithful representation that the technology gives us, it is clearly capable of providing us with non-descriptive representations. It is not unreasonable to accept that such viewing counts as perceiving Obama and thus resulting in singular thoughts about him.[18]

is open up information channels; they provide access to information. They do not provide information in the sense of providing representations of what is being tracked. Instead, they make it possible to receive information as representations. See Pylyshyn (2004; 2007). For a detailed discussion of FINSTs relevance, or lack thereof, to mental files as philosophers use that concept, see Geirsson (2018).

[18] That is not to say that all representations can provide non-descriptive representations. Clearly, some of the representative works of Pablo Picasso and Paul Klee, to name two examples, are too abstract or too stylistic to provide an accurate representation of a subject that it is of. When one views some of their portraits, it is not likely that one can recognize them of portraying one particular person rather than another.

If perceiving an object and paying conscious attention to it is required for singular thought, then that entails that me seeing bear scat does not enable me to have singular thoughts about the bear who left it there. While I have perceived the scat, I have not perceived the bear and not paid conscious attention to the bear itself. All I have experienced are some causal traces left by the bear. Until I perceive the bear my thoughts of it are general (e.g., "the bear that left the scat"), not singular, and my attempted references to it are attributive in nature. I refer to the bear that left the scat, whatever bear that is.

Those who have claimed that one can have singular thoughts about an object by being on the receiving end of a causal chain of names advocate a view that admits of very weak causal traces being sufficient to acquire singular thoughts. But, as we have seen, the main reason given for accepting that view is that *reference* is secured via the causal chain. As I have argued we can accommodate that view by acknowledging that one can acquire a *general understanding* of the relevant name that way, but not *specific understanding*. General understanding gives us general thoughts. More is needed for one to acquire singular thought.[19]

7. *Taking stock*

Someone might object at this point that I have restricted singular thoughts too much; that it is too hard to acquire singular thoughts. And it is true that the view presented here is more restrictive than those of Jeshion, who gives up the acquaintance and the metaphysical constraints,[20] Crane, who accepts Jeshion's intuitions regarding acquaintance (Crane 2013: 152), and Recanati, who advocates very weak epistemic relations, to name a few. But the view I have presented restricts the scope of singular thoughts in a very similar way to Russell's 1903 view. That is, singular thoughts are non-descriptive thoughts, one can have singular thoughts of ordinary objects, and one needs to be acquainted with (having perceived) such objects in order to have singular thoughts about them. Any other relationship results in descriptive thoughts. Can one be acquainted with an object without perceiving it directly? Perhaps, yes, provided that one's experience of the object is of the kind that enables one to form a non-descriptive thought of the object. This might allow for one being acquainted with an object after seeing it on TV, for example.

Finally, the view presented here has consequences for one believing singular propositions. Singular propositions contain the object referred to. However, the view I have advocated entails that quite frequently

[19] Some might wonder how statements containing different but codesignative names can resist substitution on the account that I am providing. I discuss that in Geirsson (2013; 2021).

[20] See for example her Dessert Sensations example, where her father thinks singular thoughts about a cake-delivering business yet to exist. (Jeshion 2010: 117–118).

we don't grasp a non-descriptive mode of presentation of the relevant object. Instead, we likely replace the name with a description, thus coming to believe a general proposition. If, after meeting Smith, someone tells me that he is married to Triphena then, as Russell observed in 1903, my thought that Triphena is not present is a general thought more appropriately expressed as "Smith's wife is not present."

8. *Concluding remarks*

After showing that there is no agreement about the nature of singular thought, I revisited early Russell to find greater clarity. I then advanced an account in the spirit of early Russell. I have argued that the advocates of the direct reference view who argue that being on the receiving end of a name is sufficient for having singular thoughts about the object named have not provided good reasons for their view. Passing on a name can provide the recipient with a general understanding of the name, but not specific understanding. That is, when acquiring the name, the recipient may not learn the identity of the object named as *this very object*. For that we need strong acquaintance. While names do play an important role in communication when passing on information the explanation is not, as Bach would have it, that the name displays how the object is thought about.

References

Azzouni, J. 2011. "Singular Thoughts (Object-Directed Thoughts)." *Proceedings of the Aristotelian Society Supplementary* LXXXV: 45–61.

Azzouni, J. 2021. "Singular Thoughts, Sentences and Propositions of That Which Does Not Exist." In S. Biggs and H. Geirsson (eds.). *The Routledge Handbook of Linguistic Reference*. New York and Oxford: Routledge, 409–420.

Bach, K. 1987. *Thought and Reference*. Oxford: Clarendon Press.

Bach, K. 2010. "Getting a Thing into a Thought." In R. Jeshion (ed.). *New Essays on Singular Thought*. Oxford: Oxford University Press, 39–63.

Bezuidenhout, A. 2021. "The Referential-Attributive Distinction." In S. Biggs and H. Geirsson (eds.). *The Routledge Handbook of Linguistic Reference*. New York and Oxford: Routledge, 53–70.

Braun, D. 2005. "Empty Names, Fictional Names, Mythical Names." *Nous* 39 (4): 596–631.

Braun, D. 2021. "Mill and the Missing Referents." In S. Biggs and H. Geirsson (eds.). *The Routledge Handbook of Linguistic Reference*. New York and Oxford: Routledge, 373–383.

Crane, T. 2013. *The Objects of Thought*. Oxford: Oxford University Press.

Donnellan, K. 1966. "Reference and Definite Descriptions." *The Philosophical Review* 75 (3): 281–304.

Donnellan, K. 1981. "The Contingent A Priori and Rigid Designators." In P. A. French, T. E. Uehling and H. K. Wettstein (eds.). *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press, 45–60.

Evans, G. 1973. "The Causal Theory of Names." *Aristotelian Society Supplementary* 47 (1): 187–208.

Geirsson, H. 2013. *Philosophy of Language and Webs of Information*. New York: Routledge.

Geirsson, H. 2018. "Singular Thought, Cognitivism, and Conscious Attention." *Erkenntnis* 83 (3): 613–626.

Geirsson, H. 2021. "Eliciting and Conveying Information." In S. Biggs and H. Geirsson (eds.). *The Routledge Handbook of Linguistic Reference*. New York and Oxford: Routledge, 153–166.

Goodman, R. 2018. "On the Supposed Connection Between Proper Names and Singular Thought." *Synthese* 195 (1): 197–223.

Harman, G. 1977. "How to Use Propositions." *American Philosophical Quarterly* 14: 173–176.

Jeshion, R. 2001. "Donnellan on Neptune." *Philosophy and Phenomenological Research* LXIII (1): 111–135.

Jeshion, R. 2002. "The Epistemological Argument against Descriptivism." *Philosophy and Phenomenological Research* 64 (2): 325–345.

Jeshion, R. 2010. "Singular Thought: Acquiantance, Semantic Instrumentalism, and Cognitivism." In R. Jeshion (ed.). *New Essays on Singular Thought*. Oxford: Oxford University Press, 105–140.

Jeshion, R. 2014. "Two Dogmas of Russellianism." In M. Garcia-Carpintero and G. Marti (eds.). *Empty Representations: Reference and Non-Existence*. Oxford: Oxford University Press, 67–90.

Kaplan, D. 1989a. "Afterthoughts." In J. Almog, H. Wettstein and J. Perry (eds.). *Themes From Kaplan*. New York: Oxford University Press, 565–614.

Kaplan, D. 1989b. "Demonstratives." In J. Almog, H. Wettstein and J. Perry (eds.). *Themes From Kaplan*. Oxford: Oxford University Press, 481–564.

Martone, F. 2016. "Singular Reference Without Singular Thought." *Manuscrito* 39 (1): 33–59.

Montemayor, C. and Haladjian, H. H. 2015. *Consciousness, Attention, and Conscious Attention*. Cambridge: The MIT Press.

Pylyshyn, Z. W. 2004. "Visual Indexes, Objects, and Nonconceptual Reference." *Notes for Isle d'Oleron Summer Workshop on Objects*.

Pylyshyn, Z. W. 2007. *Things and Places: How the Mind Connects with the World*. Cambridge: The MIT Press.

Recanati, F. 2012. *Mental Files*. Oxford: Oxford University Press.

Recanati, F. 2021. "Reference and Singular Thought." In S. Biggs and H. Geirsson (eds.). *The Routledge Handbook of Linguistic Reference*. New York and Oxford: Routledge, 399–408.

Rozemond, M. 1993. "Evans on *De Re* Thought." *Philosophia* 22 (3–4): 275–298.

Russell, B. 1961. "Knowledge by Acquaintance and Knowledge by Description." In R. E. Egner and L. E. Denonn (eds.). *The Basic Writings of Bertrand Russell*. London: Routledge, 191–198.

Russell, B. 1994. "Points About Denoting." In A. Urquhart and A. C. Lewis (eds.). *The Collected Papers of Bertrand Russell*, Vol. 4: Foundations of Logic, 1903–05. London: Routledge, 305–313.

Sainsbury, R. M. 2005. *Reference without Referents*. Oxford: Oxford University Press.

Salmon, N. 1986. *Frege's Puzzle*. Cambridge: The MIT Press.

Salmon, N. 1987. "How to Measure the Standard Metre." *Proceedings of the Aristotelian Society* 88 (1987–1988): 193–217.

Salmon, N. 1998. "Nonexistence." *Nous* 32 (3): 277–319.

Salmon, N. 2002. "Mythical Objects." In J. K. Campbell, M. O'Rourke and D. Shier (eds.). *Topics in Contemporary Philosophy 1*. New York: Seven Bridges, 105–123.

Salmon, N. 2004. "The Good, the Bad, and the Ugly." In M. Reimer and A. Bezuidenhout (eds.). *Descriptions and Beyond*. Oxford: Clarendon Press, 230–260.

Sawyer, S. 2012. "Cognitivism: A New Theory of Singular Thought?" *Mind and Language* 27 (3): 264–283.

Soames, S. 1995. "Beyond Singular Propositions." *Canadian Journal of Philosophy* 25 (4): 515–550.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.8
Received: May 7, 2023
Accepted: April 25, 2024

Embedded Metaphor and Perspective Shifting

GONG CHEN
Guangzhou Medical University, Guangdong, China

GRAHAM STEVENS
University of Manchester, Manchester, UK

Non-cognitivism is an approach to metaphor that denies the existence of any metaphorical meanings. A metaphor's only meaning is its literal meaning. The interpretation of metaphor, on this approach, does not consist in metaphorical contents being communicated by being either semantically encoded or pragmatically communicated. Rather, metaphor operates in an entirely non-linguistic way that does not require the postulation of such meanings. Metaphors cause people to see connections, even to grasp new thoughts, but they do not do so by meaning those thoughts or connections. Non-cognitivism faces a stern challenge from the problem of embedding: metaphors embedded in propositional attitude reports seem to require metaphorical meanings in their truth-conditions. In this paper, we argue that existing attempts to solve this problem for non-cognitivism have been unsuccessful. We then offer a new solution that differentiates two scope readings of embedded metaphors and explains each in turn. The paper thus suggests that non-cognitivism has enough rescores to account for embedded metaphors.

Keywords: Metaphor; non-cognitivism; perspective shifting.

1. *The problem of embedded metaphors*

Do metaphors mean something more than their literal contents? Most philosophers of language think that they do. They divide, roughly into those who think that metaphorical meanings are the result of pragmatic processes applied to literal contents to generate metaphorical meanings which are conveyed through either implicature or explica-

ture (Grice 1975; Wilson and Carston 2006; Récanati 2004) and, more rarely, those who postulate metaphorical meanings as semantic values (the most prominent contemporary proponent of this view is Stern 2000). A more radical alternative view, originating with Davidson (1978), is that there is no such thing as metaphorical meaning: "metaphors mean what the words, in their most literal interpretation, mean, and nothing more" (1978: 32). This view is known as non-cognitivism. Metaphors may *cause* people to grasp intended thoughts, or to see connections between things, and so on, but they do so by other means than by encoding those things as contents (or, indeed, implicating or explicating them). Metaphors are understood in the way that paintings or pieces of music are understood, not in the way that sentences are.

Embedded metaphors, for example metaphors embedded under attitude verbs such *believes, hopes, knows*, etc., pose an immediate problem for non-cognitivism. According to non-cognitivism, metaphors have no meaning beyond their literal meaning. The metaphor "hope is the thing with feathers that perches in the soul" is simply false, because hope is a feeling of expectation. Whatever explanation the non-cognitivist offers of how this metaphor is employed in human communication cannot appeal to some metaphorical content encoded by this sentence. Rather the explanation will have to appeal to a story about how an utterance of a straightforward falsehood stands in a causal relation towards its hearer such that this relation results in the speaker achieving something by that utterance. The above non-cognitivists offer a range of detailed accounts of how this can be elucidated. But what about an utterance of the following?

(1) James believes that hope is the thing with feathers that perches in the soul.

The attitude report in (1) does not say something literally false about hope, it reports James' state of mind. And, like all propositional attitude reports of the form S Vs that P, a plausible semantic theory would predict that it is true just in case A stands in the correct attitude relation to P: that James stands in the belief relation to the proposition expressed by the sentence "hope is the thing with feathers that perches in the soul." But, intuitively, (1) does not report that he believes a literal falsehood. It might report that James believes that hope allows us to rise above or overcome adversity. So the proposition that (1) reports James as believing is the metaphorical content of the embedded sentence, not its literal meaning. In short, the truth-conditions of (1) require that the embedded proposition is the very same thing that non-cognitivism denies the existence of, namely the metaphorical content of the sentence. If we accept the relatively uncontroversial premises that truth-conditions supply the meanings that speakers understand and that (1) is a perfectly meaningful construction that ordinary English

speakers can understand with ease, then we seem to have a powerful counterexample to non-cognitivism.[1]

In addition, it is worth noticing that this type of belief report may either report that S represents some content to themselves metaphorically and believes it, or it can be a metaphorical representation of a content that S believes without representing it to themselves metaphorically. This distinction seems to support cognitivism. This is because if a metaphor M encodes a non-literal meaning M* when uttered by S, cognitivists will distinguish cases where M contributes its literal meaning or M* compositionally to the content of the construction *S believes that M*.

In this paper, we will argue that non-cognitivism can account for metaphors in the above belief report cases. We will begin Sect. 2 with an introduction explaining the distinction between these two readings. We name them "de re readings" and "de dicto readings." After that, Sect. 3 will consider two unsuccessful non-cognitivist solutions dealing with the problem, both of which refuse to accept the legitimacy of de re readings and insist that de dicto readings are the only admissible readings of metaphorical belief reports. In Sect. 4, we aim to propose a non-cognitivist account of de re readings of embedded metaphors. In Sect. 5, we offer an account of de dicto readings of embedded metaphors.

2. *Two readings for belief report cases*

These two readings correspond to the common distinction between de re and de dicto attitude reports. For example, the literal belief report underlined in 2 has both a *de re* (2a) and a *de dicto* (2b) reading:[2]

(2) Having tried them both in the guitar store, Amy believes that the 1972 SG sounds better than the 1989 SG.

(2a) [The x: x is a 1972 SG][The y: y is a 1989 SG](Amy believes that x sounds better than y).

(2b) Amy believes that ([The x: x is a 1972 SG][The y: y is a 1989 SG] (x sounds better than y)).

Whereas 2b requires Amy to conceptualise the two guitars under the concepts provided by the definite descriptions, 2a is true simply if she believes that the objects in question stand in the right to relation to

[1] Gricean implicature accounts of metaphor also owe us an explanation of (1) just as much as non-cognitivists do.

[2] For ease of exposition we have treated the definite descriptions as quantifier phrases containing bound variables along the lines developed by those who endorse a Russellian theory of definite descriptions. Alternative accounts of definites can be offered and those accounts can also recognise the distinction between de re and de dicto attitude reports. The Russellian analysis of the distinction as a matter of the relative scope of the attitude verb and a quantifier is a simple way of making the distinction apparent, however, hence our choice to draw on it in this example.

one another, regardless of how she herself conceptualises those objects (perhaps she cannot distinguish one from the other). The same holds true of metaphorical belief ascriptions: when we report that James believes that hope is the thing with feathers that perches in the soul this can be true just in case it employs a metaphor in order to communicate that the metaphor can convey what James believes (*de re*) or it can also be read as reporting that James entertains that very metaphor himself (*de dicto*).

Some might object to our recognition of de re readings of embedded metaphors. For example, if one holds to a view along the lines of Camp (2006), according to which metaphors are *characterizations* of objects, the de re reading may seem less plausible. For Camp, roughly, a characterization gives us a set of salient properties that the speaker of the metaphor is communicating by their choice of that metaphor. One persuaded that this is the right way to think about metaphor may well take this to support a rejection of the plausibility of de re readings. If we are reporting James' attitudes when we state that James believes that hope is the thing with feathers that perches in the soul, we might have grounds here for insisting that the choice of metaphor will only be apt if it characterizes hope in the way that James does. Accordingly, this may count against recognising the de re reading as plausible.[3] However, we do not think this is the case. Camp's notion of a characterization is an insightful one and is particularly useful for understanding how to think of metaphors on de dicto readings of embedded cases. But it can also admit de re readings. It is important to note that, on the de re reading, the characterization would effectively take wide scope over the propositional attitude verb. In other words, it is a way of characterizing the belief from the perspective of the reporter, not of the attitude holder. So a de re report "James believes that hope is the thing with feathers that perches in the soul" characterizes James' attitude in accordance with the perspective of the reporter, rather than characterizing hope from James' perspective. These two can be hard to disentangle as they are obviously closely aligned. Characterising James' attitude in this way obviously will be a very similar enterprise to reporting his characterisation of hope as the thing with feathers that perches in the soul. But they are not the same enterprise. Suppose, for example, that James lacks the imaginative resources to understand that labelling hope as he thing with feathers that perches in the soul can be an effective metaphor to communicate its function. But he does nonetheless think that hope is a thing with feathers. Then the de dicto reading is false, but the de re one is true. Why? Because James does not characterize hope under the representation *a thing with feathers that perches in the soul.* But our use of that representation to characterize his highly negative and distrustful attitude towards hope is apt

[3] We are not attributing this rejection of the de re reading to Camp herself, it should be noted.

all the same. Thus we maintain that both de re and de dicto readings are plausible.

3. *Some non-cognitivist proposals*

Some prominent non-cognitivists have offered responses to the problem of embedding. Here we explain why we find those responses inadequate.

The first proposal is offered by Davies (1984). Davies denies that metaphors should be understood in accordance with the same framework as we apply to ordinary descriptive contents. The function of a metaphor, he insists, is not to encode a propositional content that is true or false. Metaphors function by helping those who appreciate them to recognise certain truths, but those truths themselves are not part of the content of the metaphors. Understanding metaphor, on this view, is not a matter of linguistic competence but of something more akin to aesthetic appreciation. Echoing ideas in Davidson, it seems that understanding metaphor for Davies is of the same kind as appreciating a painting or a work of music. Whatever content is arrived at in this process, it is not linguistically encoded. Davies does not directly address the problem of embedding as we have presented it here (namely, in terms of the truth-conditions of propositional attitude reports that embed metaphors). He does however briefly consider cases that raise the spectre of metaphors being the objects of belief. He takes the fact that the following cases sound infelicitous to support his claim that metaphors are not believed:

(3) I believe this: you are a rose.
(4) Of course this is true: you are a rose.

If Davies is correct that metaphors cannot be the objects of belief, then it would presumably follow that belief reports apparently employing them should not be interpreted at face value. Davies' examples strike us as puzzling, however. They do indeed sound odd, but it is not really clear that this has anything to do with the fact that they contain metaphors. They are just peculiar ways of speaking. More natural constructions like the following sound quite acceptable:

(5) I believe that Juliet is a rose.
(6) Truly, Juliet is a rose.

Given that this is so, Davies owes us an explanation of how his account is to be extended to constructions like these. On Davies's view, when James says "hope is the thing with feathers that perches in the soul" and Benvolio says "yes, I agree," James is intending Benvolio to see something, and Benvolio's utterance "I agree" signifies that he grasps and endorses it. So no metaphorical meaning as such is involved.

Could this explanation be extended to account for "James believes that hope is the thing with feathers that perches in the soul?" Well presumably it would then have to be a report that James had also grasped

the thing this metaphor is supposed to get us to appreciate. But then Davies' proposal would have to argue that James did so by understanding *this very sentence*. But this is not the only situation where we would be warranted in asserting that James believes that hope is the thing with feathers that perches in the soul. Perhaps this particular metaphor has never actually occurred to James but he does have a different description for hope. This would warrant an assertion of (1) on its de re reading. But this eludes explanation on Davies's account as that would clearly lead us straight back into the very problem that as a non-cognitivist we are trying to avoid—the idea that the metaphor *means* something that S is being reported to believe. In other words, Davies's account seems to work only for cases where (1) reports a situation where the metaphor "hope is the thing with feathers that perches in the soul" was a "live" metaphor[4] in James thought, but cannot accommodate a use of (1) to metaphorically describe James' belief that the metaphor "hope is the thing with feathers that perches in the soul" can convey what he believes that does not attribute to him the conscious apprehension of that metaphor.

In other words, Davies only offers an explanation of the de dicto reading of (1) and denies that there is a de re reading. But, as we have argued above, this is contrary to the evidence. Accordingly, his defence of non-cognitivism is incomplete unless he can offer compelling reasons why (1) should only be read as a de dicto report not as a de re one.

Another recent defence of non-cognitivism that does attempt to pursue just this kind of response to the embedding problem, is offered by Lepore and Stone (2010). On their account, as we also saw with Davies, metaphors serve a non-linguistic purpose: they are used to influence hearers, to make them see things in a certain sort of a way, even to get them to believe certain things—but they do not achieve these ends by *meaning* these contents that hearers arrive at. But the point of a metaphor, the thing it is used to get hearers to arrive at, seems to be playing a semantic role when the metaphor is embedded.[5] This is clearly contrary to their approach. They respond by denying that there is some content which can be isolated from the metaphorical vehicle as its content even in these situations. The response rests on their insistence that an embedded metaphor can only be truthfully reported by an attitude report if the attitude holder actively accepts the metaphor

[4] When we say it is "live" metaphor, it means that the speaker entertains the very metaphor himself.

[5] Keating (2015) objects to Lepore and Stone's account of how metaphors function on the grounds that it is not clear why the propositions that speakers of metaphors intend to cause their hearers to grasp are not thereby counted as speaker meanings. This objection seems more pressing for them than other non-cognitivists as they seek to ground metaphorical communication within a co-operative process between speakers and hearers. Keating's call for greater justification in construing this process as somehow fundamentally different to the co-operative processes familiar in pragmatics seems reasonable to us.

(i.e., if the belief report is understood as reporting that the embedded metaphor is apprehended "live" by the attitude holder). For example, if Chris is reported as believing that *No man is an island*, this report is not veridical simply if Chris believes that humans are socially interconnected beings. It:

> [...] also requires that Chris accepts the metaphor as apt, and moreover that Chris is drawn from there by metaphorical thinking to appreciate that people are all inter-connected by social relationships. The metaphor must be active in Chris's thought, and so it must somehow also be active in the truth conditions of [*No man is an island*]. (Lepore and Stone 2010: 175)

By taking this line, Lepore and Stone think that they can sidestep the embedding problem by effectively insisting that rather than being a full-blown belief report, a metaphorical belief report is in fact a report of the metaphorical thinking that the subject underwent. In other words, "Chris believes that no man is an island" does not report the content of Chris's belief, it reports that Chris was in the situation where he believed something that was connected, in whatever way the non-cognitivist recognises as generally explaining how metaphor works, to the literal sentence "no man is an island." For example, if we understand what speakers intend us to grasp by an utterance of this sentence by recognizing relevant similarities, then the belief report simply reports that Chris was in the position of recognizing those similarities in response to entertaining that literal content. In short, like Davies, Lepore and Stone respond to the embedding problem by first insisting that embeddings under attitude verbs are de dicto by default. However, they do offer some justification for this exclusion of the de re reading by seeking to reduce the de dicto reading to a report that the attitude holder stands in the relevant non-cognitive relation to the very sentence displayed in the report. But the only thing grounding this reduction, it seems, is their intuition that this is the only way to interpret the report.

Unfortunately, no evidence in support of Lepore and Stone's intuition that the metaphor must be "live" in the thoughts of the attitude holder is provided. Our view is that the intuition is incorrect. If we report that James believes that hope is the thing with feathers that perches in the soul, it seems to us that our report is true if (though not *only if*, as we shall explain below) James thinks hope allows us to rise above or overcome adversity. We can indeed report attitudes using metaphors that the reported attitude holder is simply not in a position to understand, let alone to have as active in their thoughts. For example, I can describe a six year-old child as thinking *that the entire universe revolves around them*. But the six year old child does not have this metaphor active in their thought—they may simply lack the cognitive resources to have that kind of metaphor active in their thought—but they can have a self-centred attitude towards themselves and lack of consideration towards others. This is all that is needed to license the

metaphorical description of their cognitive state and this is all that is meant by the belief report in this particular instance.[6]

So far we have argued that both Davies and Lepore and Stone are mistaken in seeking to deny the plausibility of a de re reading. We will argue below, in fact, that de re readings can be fully explained by a non-cognitivist account of metaphor. Of course, one might think that we therefore have little grounds for complaint against these two competitors: if we can explain de re readings, then surely all we need to do to secure a robust non-cognitivist analysis of metaphors embedded under attitude reporting verbs is to add our account of de re readings to one of these accounts which explain the de dicto readings. However, while we are indeed in general sympathy with the non-cognitivist project of these authors, we do think that their particular accounts of the de dicto readings are problematic. Thus, while we hope that our account can strengthen the non-cognitivist's case, and so is intended to be offered as a contribution to that project, we also think there are significant explanatory gaps in the accounts that we have considered. In the remainder of this section, we aim to identify those explanatory gaps. We will then go on to propose an account of how de dicto readings function to fill those gaps, alongside our account of the de re readings.

Not only is the position Lepore and Stone defend at fault in its failure to recognise de re readings of embedded metaphors, it also leaves the interpretation of the verb under which the metaphorical material is embedded shrouded in secrecy. To put this point very simply: what triggers a hearer to recognize when an attitude verb is a genuine propositional attitude reporting relation, and when it is reporting the non-cognitive relation that is taken to underlie the speakers interaction with the metaphor? Take, for example, a case of the sort that Cohen (1978) labels "twice true" metaphors—namely those metaphors which are intuitively understood as communicating a metaphorical truth while also being literally true:[7]

(7) Trump is an animal.

This can be embedded under a belief attribution:

(8) Biden believes that Trump is an animal.

[6] It could well be that this apparent report is not really describing a belief of the child's at all, but simply giving a metaphorical description of their general character, selfish attitudes, or lack of concern for others. That might look like a de re belief report but in fact it would not be because it wouldn't be a *belief* report at all. We agree that such cases are tricky. But it is not unrealistic that some such utterances are genuine de re reports of a child's belief that they are *entitled* to X, without attributing to them the de re belief that takes them to entertain the very sentence "the world revolves around me!" Other examples, such as James' belief that hope allows us to rise above or overcome adversity as grounding a de re belief attribution using (1) are less controversial—see our discussion of Camp on characterizations in section 2 for further defence of the reading there.

[7] See also Keating (2015) for discussion of such cases.

But there is a difference in what belief is being reported, depending on whether the embedded sentence is literally or figuratively interpreted. Now of course it is not particularly problematic for belief sentences to be ambiguous: we have already noted that attitude reporting sentences are ambiguous between de re and de dicto readings; and, more mundanely, placing any lexical or syntactic ambiguity under the scope of a belief report will usually preserve that ambiguity:

(9) Mary believes that Suzy likes to play a little guitar to relax.
(10) Jane believes that Mary met Suzy when she was living in London.

In the case of (9) the word *little* may denote the size of the guitar, or the amount of playing that Suzy likes to do. Hence (9) is ambiguous between at least two belief reports. Similarly, the surface grammar of (10) is ambiguous between reporting Jane's belief that Mary met Suzy when *Mary* was living in London, or Jane's belief that Mary met Suzy when *Suzy* was living in London, as well as Jane's belief that Mary met Suzy when *Jane* was living in London. However, the situation Lepore and Stone envisage is more problematic. On their view, we do not really have an ambiguity in the object of Biden's belief in (8). There is no metaphorical meaning of the embedded sentence for Biden to believe on their view. Hence, this reading is not really a belief report at all. The ambiguity, if there is one, does not reside in the embedded sentence as this sentence only has its one, literal, meaning. It must reside, then, in different senses of the verb "believes." Hence, Lepore and Stone seem committed to the view that the "metaphorical" interpretation (whereby we take Biden to see the same connections that we do if we interact with the embedded sentence in the right sort of way) does not attribute a belief to Biden at all. But, in that case, what does the word "believes" mean in (8)? It now looks dangerously close to itself encoding a metaphorical meaning here—Biden does not *literally* believe any metaphorical meanings according to the non-cognitivist, because there are no metaphorical meanings for him to believe. In which case, presumably, a non-cognitivist analysis of this seemingly metaphorical use of "believes" must be provided. At best, we now have an undesirable and, we submit, implausible level of complexity at work in the account. For we will now have to say that we have two levels of connection-seeing at play: the "metaphorical" use of "believes" triggers some process in us which allows us to see certain kinds of connections which in turn lead us to recognize another episode of connection-seeing which Biden is being reported as having taken part in, as triggered by the embedded sentence. But surely part of the appeal of non-cognitivism is that it avoids unnecessary interpretive complexities of this sort. This proposal seems to us no simpler than simply postulating metaphorical meanings in the first place. Indeed we are not convinced that this double layering of non-cognitivism to explain away an apparently metaphorical sense of belief that arises out of a prior attempt to explain away an apparently metaphorical object of this "belief" is even coherent.

What we think the above discussion shows is that Lepore and Stone (and, indeed, Davies) leave at least two unacceptable explanatory gaps in their accounts of embedded cases. On the one hand, they fail to explain what guarantees that embedding metaphors under propositional attitude verbs has a different result to embedding literal sentences under such verbs. But, if they are going to deny that the resulting constructions can be read in a de re as well as de dicto fashion, this needs explaining. After all, literal belief reports are seemingly ambiguous between the two, so what makes the embedding of metaphor special? On the other hand, no account is given of precisely how we effect the shift in perspective whereby we understand (11) to be reporting a connection-seeing event by Trump that does not routinely happen for belief reports. After all, we do not need to understand (11) as directing us to consider an episode of connection-seeing from Trump's perspective:

(11) Trump believes that Biden lives in the White House.

In the next section, we will elaborate further on the kind of perspective shift that is at work in the de dicto reading of metaphors embedded under attitude verbs. In the following section, we will propose a solution to fill those gaps.

4. *De re and de dicto embeddings of metaphor under believes*

Recall our original belief report, (1). We can intuitively recognise an ambiguity in this report that is best explained by appeal to the de re/de dicto ambiguity of the report, as outlined above:[8]

De Re:
[Hope is such that x] James believes that x is the thing with feathers that perches in the soul.

De Dicto:
James believes that [hope is such that x] x is the thing with feathers that perches in the soul.

In the de re reading, the metaphor "hope is the thing with feathers that perches in the soul" conveys the thing that James believes without committing to the claim that James has that metaphor in mind. In the de dicto reading, James believes that hope is the thing with feathers that perches in the soul by virtue of having that very metaphor before

[8] An alternative reason one might take for the scope behavior of the embedded metaphor here may be that the metaphorical part "hope is the thing with feathers that perches in the soul" is perhaps ambiguous between a descriptive and a pejorative or insulting sense. On this view, it will take wide scope if occurring in the latter sense by virtue of the semantic properties it has as an expressive (see Potts 2007 for extensive discussion of expressives). Nonetheless, we take this to be inessential to the issues as hand, as there are plenty of non-insulting metaphors that display the same behavior with respect to scope. For example, "James believes that hope is the thing with feathers that perches in the soul."

his mind and assenting to its truth. The strategy we have observed consistently emerging among those who face difficulties in accounting for the two readings has been to deny the reality of the de re reading. For our non-cognitivists this was because they were able to offer some account of the de dicto reading by insisting that it is not really a report of a belief, rather it is a report of the sort of situation that James found himself in when confronted by the metaphor *hope is the thing with feathers that perches in the soul*. Rather than grasping a metaphorical meaning, he engaged with the sentence in some non-cognitive fashion (perhaps he saw some connections, or was caused to entertain some thought that was not linguistically encoded or implicated by the original sentence). This is what is really being reported, according to the non-cognitivist.

From the perspective of non-cognitivism, this strategy strikes us as ill-advised and unnecessary. It is ill-advised because the de re reading is just as plausible as the de dicto, so it puts non-cognitivism in the dialectically weak position of having to argue that people are wrong to think metaphors can be employed to describe the beliefs we report others as having. It is unnecessary because there is a non-cognitivist explanation of the de re reading. The easy explanation is to insist that the de re reading is not a description of a belief at all. Why? Because it characterises the belief metaphorically and, according to non-cognitivism, metaphors do not describe things. They function by causing hearers to see connection between things in a way that does not require any semantic content above the literal meaning of the metaphor. The non-cognitivist interpretation of the de re reading should be no different: on the de re reading, (2) is an attempt on the part of the reporter to cause their audience to see something by saying something literally false about James' state of mind. It is a metaphor apparently about James' belief, in the same way that *hope is the thing with feathers that perches in the soul* is a metaphor apparently about hope. Or if one prefers to find a metaphor which employs a verb phrase to make the similarity clearer, *James believes that hope is the thing with feathers that perches in the soul*, characterizes James' belief metaphorically in the same way that *Cobain sings with the voice of the dispossessed* characterizes Cobain's singing metaphorically. No special explanation is required for the former that was not already needed for the latter.

The non-cognitivist explanation of de re cases then, insists that they are not reports of a belief in a metaphor, they are metaphors themselves. The explanation of how a metaphor communicates de re information about what James believes should therefore take the same form as the explanation of how a metaphor inspires people to see the similarities between hope and the thing with feathers that perches in the soul. The metaphor *hope is the thing with feathers that perches in the soul* functions by saying something about hope that leads us to arrive at information concerning hope that is not linguistically encoded in the

original sentence. The metaphor *James believes that hope is the thing with feathers that perches in the soul* functions by saying something about *James*' state of mind which leads us to arrive at information concerning *James*' state of mind that is not linguistically encoded in the original sentence. If the non-cognitivist can explain the metaphor *hope is the thing with feathers that perches in the soul* by appeal to connections it leads its hearer to recognize, which lead that hearer in turn to the thought that hope allows us to rise above or overcome adversity, then a precisely similar explanation can be provided of the connections recognized in arriving at the thought that James' state of mind is distrustful towards hope.

On this view, the de re cases should not only be recognised by the non-cognitivist, they can be easily explained by them. It is the de dicto cases that are hard. Here, the usual non-cognitivist explanations run into a new obstacle because we now have the metaphorical content seemingly playing an essential role in the truth-conditions of (1): it is the metaphor itself that James is being reported to believe here—hence a metaphorical content is demanded as the object of his belief in order to explain what would need to be the case for (1) to be true on the de dicto reading. We are not simply trying to get our audience to see connections in the de dicto case: we are reporting that James sees those connections.

We have argued that a range of theorists including non-cognitivists like Davies, Lepore and Stone, share two difficulties in the face of the problem of embedded metaphors. Firstly, they fail to accommodate de re readings of embedded metaphors. This, we have argued, is implausible as de re readings seem clearly available. Presumably, what we have called the non-cognitivist explanation of de re cases would be consistent with Davies, and Lepore and Stone's non-cognitivist ambitions and, therefore, a welcome additional resource for them. Secondly, these theorists all lack an account of what ensures that attitude reports take obligatory wide scope when the attitude verb operates on an (apparent) metaphorical content but not when it operates on a literal content. Related to this complaint, we have argued that they lack any explanation of the mechanism which explains how the de dicto reading is achieved. If the embedded metaphor has no metaphorical meaning, and simply serves to place us in the position where we obtain a clear picture of the non-cognitive relation that the reported attitude holder was in, there should be some kind of explanation of how this is achieved. Otherwise, we have argued, we will be in danger of treating believes as itself having a metaphorical function in such roles, and this is a dangerous route for the non-cognitivist. We have offered a proposal to avoid the first difficulty. Having recognized de re readings as well as de dicto readings, we do not face the challenge of needing to account for attitude verbs taking wide scope over what they report, as we are not taking such readings to be obligatory: the reports are simply ambiguous between

de re and de dicto readings. However, we do still need an explanation of the de dicto ones. We now proceed to provide an explanation of these and the mechanism which facilitates them. This, again, is one which will draw only on the resources of non-cognitivism.

5. *A new non-cognitivist solution to the problem of embedded metaphor*

We explained above that non-cognitivism does not lack the resources to account for de re readings of embedded metaphors if we grant them the resources to explain ordinary, non-embedded, metaphors. We will now propose an explanation of how de dicto readings of embedded metaphors work in a way that is consistent with non-cognitivism. Returning to our preferred example:

(1) James believes that hope is the thing with feathers that perches in the soul.

What we seek is an account of what this metaphor means when it reports that James himself entertained that very metaphor and took its content to be true.

Our proposal is that embedded metaphors understood on a de dicto reading are quotational constructions. In particular, we suggest that they should be read as implicit examples of what we will call *echoic quotation*. Echoic quotation, as we will see, can be produced in many quotational contexts but it is particularly evident in cases of open quotation. "Open quotation" is a term coined by Récanati (2010)[9] to describe a distinct form of quotation that does not recruit the quoted material to occupy the grammatical role of a singular term. Rather than referring to the material that is quoted, it acts as a context-shifting device to enable speakers to mimic or echo the thoughts and words of others in order to express the mimicked speaker's perspective.[10] Closed quotation, by contrast, is quotation which does recruit the material as a singular term. Both closed and open quotation can generate the echoic uses of quotation we are interested in, as we will see in several examples below. Open quotation is particularly useful for illustrating the echoic role, however, as this seems to be the primary role of open quotation.

Consider a simple form of quotation like we have in the following example:

[9] Récanati would not, of course, endorse our desire to defend non-cognitivism, as he has his own account of metaphor as resulting from pragmatic explicatures. See references in the introduction above.

[10] Note that Récanati was not directly motivated by a desire to explain context shifting but motivated by the idea that (what he takes to be the core cases of) quotations are demonstrations in Clark's (1996) sense and that they are not singular terms. However, open quotation includes cases where "the very words which are used to express the content of the reported attitude (or speech act) are at the same time displayed for demonstrative purposes" (Récanati 2010: 240).

(12) The current prime minister of the UK is called "Boris Johnson."

In (12) the function of the quotation operation is extremely simple—it just acts as a nominalizing operation to convert an expression into a name of that expression. But quotation is also used of course to report exact speech, as in this example:

(13) Bertrand Russell said, "I am not a Christian."

This form of speech report can be understood in much the same way as the first kind of quotation: the quotation names the thing that Russell said. But now consider the following examples:

(14) Oxford vaccine shows "encouraging" immune response in older adults.

(15) There are things of which we cannot speak, and I agree with Wittgenstein that, on these, "we must remain silent."

In these examples, we have a form of *echoic* quotation in which the quoted material is both used and mentioned at the same time. The quotation makes clear that this is a word for word transcription of the quoted material, but that material is put to use by the person who is doing the quoting in their own assertion. We might say that, in these examples, the material is both cited and endorsed.

Quoted material in echoic quotational contexts does not have to be endorsed however. It can also be a way of presenting the perspective of another without endorsement. Consider this news headline, which employs echoic quotation but (unlike 13) does so in a way that makes it clear that the speaker does not share the perspective introduced by the quoted material:

(16) "Human foot" spotted in Gateshead turns out to be potato.

Clearly, the quotation in (16) is not a mere mention of the material it quotes, but is a way of using it to portray a perspective without sharing or endorsing it. On the contrary, it introduces the perspective identifying the mistake of the speaker who mistook a potato for a human foot. These sorts of instances of echoic quotation are common. Notice that we have examples of both open (14, 15) and closed (16) quotation performing this echoic function. Open quotation, however, gives a particularly vivid example, as it often quotes material that can only be naturally understood in this echoic manner. Consider this pair of examples (adapted from the examples used below by Récanati):

(17) Come on now, Donald! "This election was rigged," "they stole my Presidency," [...] when are you going to face up to the truth?

(18) Donald keeps getting upset and saying "this election was rigged."

Both are reports of Donald's speech and attitudes, but they report in very different ways. Whereas (18) merely reports the words that Donald said, (17) employs those words to occupy his perspective in recounting the episode. It echoes, or mimics, his speech so as to imitate him in representing his view. Récanati helpfully characterises the difference, as follows:

> The contrast between open and closed quotation is illustrated by the following pair of sentences:
> (7) Stop that John! 'Nobody likes me,' 'I am miserable' … Don't you think you exaggerate a bit?
> (8) John keeps crying and saying 'Nobody likes me.'
> In (7) a token of 'Nobody likes me' and 'I am miserable' is displayed for demonstrative purposes, but is not used as a singular term, in contrast to what happens in (8), where the quotation serves as a singular term to complete the sentence 'John keeps crying and saying ___.' Sentence (7), therefore, is an instance of open quotation, while (8) is an instance of closed quotation. (Récanati 2010: 231)

Open quotation, then, provides a clear illustration of the echoic reading of quotation. But once we recognise it in the open cases, we can also identify it in the closed cases, as discussed above. It is this echoic reading, we suggest, that is perfect for capturing the de dicto readings of embedded metaphors.

Echoic quotation in the cases considered thus far introduces the quoted material as demonstrating what a speaker said in order to mimic that speaker. This mimicry is not restricted to contexts involving verbs of saying.[11] We can just as readily report a range of propositional attitudes in ways that make it plain that we are adopting the perspective of the attitude holder under a form of pretence. Consider this example:

(19) Trump believes that us whining liberals are undermining his authority.

In the example, it is natural to interpret the pejorative phrase "whining liberals" as mimicking the attitude that the speaker attributes to Trump, and not at all natural to interpret it as expressing the speaker's own attitude. It would, in fact, be reasonable to reconstruct the sentence by adding quotation marks around the phrase to make this clear.

A similar story holds from embedded metaphors, which, we have seen, have two readings: a de re one which does not present the metaphor as being itself before the mind of the reported attitude holder, and a de dicto one which does. This de dicto one simply presents the metaphor as it occurs from the perspective of the subject. Such cases, we submit, are best understood as instances of echoic quotation of the sort we have just outlined. The de dicto reading is thus a mimicry of the agent of the reported attitude which is effected by an implicit echoic quotation operation.[12] The operation can be made explicit to illustrate this:

(20) James believes that hope is the "thing with feathers that perches in the soul."

With de dicto readings secured by a context-shifting echoic quotation operator, the non-cognitivist has a complete account of embedded

[11] See Récanati (2010) in 226–228.

[12] See more about implicit echoic quotation examples in 20–21.

metaphors in attitude reports. The quotation "thing with feathers that perches in the soul" mimics the attitude attributed to James. There is no special problem of explaining metaphors embedded under attitude verbs for the non-cognitivist. The context-shifting nature of echoic quotation allows us to present the attitude-holder's perspective in such a way that the same mechanism involved in making sense of James' own utterance of the metaphor can extend to the mimicry of his utterance in the embedded case.

6. *Conclusion*

Our solution to the problem of embedded metaphors demonstrates that the non-cognitivist faces no special problem in explaining how metaphors can be embedded under attitude reports. The solution rests on a strategy of "divide and conquer:" first, we divide attitude reports into de re and de dicto readings, and then proceed to explain each differently. A de dicto reading employs an implicit echoic quotation operator to shift the embedded material to the reported context. Accordingly, the explanation of how we understand "S believes that M," where M is a metaphor, on this reading is the same as that which explains how we understand S when she herself utters M. The reported context is accessed by the context-shifting quotation operator, meaning that any non-cognitivist account of how S's utterance of M is to be understood will transpose to the de dicto reading of "S believes that M." A de re reading does not require the same context shift as it gives a metaphorical description of the attitude holders state of mind, rather than a description of a metaphor that the attitude holder has in mind. Thus whatever explanation the non-cognitivist avails herself of in explaining a metaphorical description of an object *o* as *F*, will transpose to the case where *o* is the state of mind of the individual so described. Of course, one who is unconvinced by non-cognitivism in general will not be likely to find anything in this explanation to change their mind. But they should be willing to concede that the non-cognitivist faces no *additional* challenge when it comes to explaining metaphors embedded under attitude reports. Non-cognitivism does not stand or fall on this issue, we conclude.

Despite our solution to the problem of explaining metaphors when embedded under attitude reporting verbs, there is a remaining puzzle about embedding that seems especially problematic for non-cognitivism. The following example takes the same form as a problematic case noted by Wilson and Carston (2019):

> Tim: Robert is a bulldozer.
> Bob: Robert is better to be a bulldozer than a Robin Reliant.

Although Wilson and Carston label such cases as "embedded metaphors," they seem rather different to the cases of embedding under propositional attitude reporting verbs. The puzzle for the non-cogni-

tivist, however, is very similar.[13] Bob's reply takes for granted, and indeed develops, the *metaphorical* meaning of Tim's assertion. Taken literally, the predicate "is a bulldozer" does not support the inference that anything in its extension is being a better bulldozer than being any Robin Reliant (where "being a Robin Reliant" is taken literally). The conversation appears to presuppose, and exploit, the metaphorical content that non-cognitivists refuse to recognize. It might be more helpful to call this the problem of *extended* metaphor, rather than embedded metaphor, given that (a) there does not seem to be any obvious lexical item that the metaphor is embedded under, and (b) the same puzzle might be thought to arise even if Tim alone were to extend the metaphor without assistance from Bob (hence it need not be conversationally embedded either). Whatever we call the problem, it requires a solution, although such considerations make us hesitant to subsume a solution to it under any general solution to the problem of embedding.

There is more to be said about extended metaphor than we can offer here, so our suggestions on this are tentative, but it does seem plausible that the mechanism we have employed to analyse de dicto attitude ascriptions can be utilised by the non-cognitivist to make sense of what is happening in these cases. Bob's reply adopts Tim's perspective, hence it is naturally interpreted in the same quotational manner that we have provided for de dicto belief ascriptions involving metaphor. Effectively, what Bob is doing in the example is occupying Tim's perspective and then building on the same metaphorical narrative that Tim develops in the first utterance. If this is right then, again, the non-cognitivist can appeal to a context-shifting operation to put Bob's audience in the same situation as Tim's—if the non-cognitivist can explain how Tim's metaphor impacts on his audience, they will be able to inherit the same explanation when it comes to explaining Bob's extension of it. If there is a good non-cognitivist explanation of non-embedded, non-extended

[13] A related "problem of embedding" for the non-cognitivist is the problem of embedding under logical operators. In sentences like "If hope is the thing with feathers that perches in the soul, then we had better keep a watchful eye on it," or "unless our intelligence agents are mistaken, hope is the thing with feathers that perches in the soul," most will have the intuition that it is the metaphorical content of "hope is the thing with feathers that perches in the soul" that is contributed to the truth-conditions of the complex sentence. But the non-cognitivist cannot offer any such metaphorical content to play that role. Alas, we do not see a way to extend our solution to the problem of embedding under attitude verbs to these cases. Here, it seems to us that the non-cognitivist simply has no choice but to "bite the bullet" and deny that any metaphorical content is contributed to the conditional. Just as with atomic sentences, the non-cognitivist has to insist that no content beyond the literal meaning is at work in these cases. The non-cognitivist, in our view, should construe these sentences as literal conditionals that perform a function of causing hearers to see things in a certain kind of way. They should not see them as conditionals which assert that *if* one views things a certain kind of way, *then* some literally described content follows. That would be asking metaphors to contribute a content to a conditional that the non-cognitivist has no right to recognize.

metaphors, there should be no special problem of either embedding or extending them.

References

Camp, E. 2006. "Metaphor and That Certain 'Je Ne Sais Quoi'." *Philosophical Studies* 129 (1): 1–25.

Clark, H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Cohen, T. 1978. "Metaphor and the Cultivation of Intimacy." *Critical Inquiry* 5 (1): 3–12.

Davidson, D. 1978. "What Metaphors Mean." *Critical Inquiry* 5 (1): 31–47.

Davies, S. 1984. "Truth-Values and Metaphors." *The Journal of Aesthetics and Art Criticism* 42 (3): 291–302.

Grice, H. P. 1975. "Logic and Conversation." In P. Cole and J. Morgan (eds.). *Syntax and semantics*. New York: Academic Press, 41–58.

Keating, M. 2015. "Thinking about Embedded Metaphors." *Journal of Pragmatics* 88 (2): 19–26.

Lepore, E. and Stone, M. 2010. "Against Metaphorical Meaning." *Topoi* 29 (2): 165–180.

Potts, C. 2007. "The Expressive Dimension." *Theoretical Linguistics,* 33 (2): 165–198.

Récanati, F. 2004. *Literal meaning*. Cambridge: Cambridge University Press.

——. 2010. *Truth-Conditional Pragmatics*. Oxford: Oxford University Press.

Stern, J. 2000. *Metaphor in Context*. Cambridge: MIT Press.

Wilson, D. and Carston, R. 2006. "Metaphor, Relevance and the 'Emergent Property' Issue." *Mind and Language* 21 (3): 404–433.

——. 2019. "Pragmatics and the Challenge of 'Non-propositional' Effects." *Journal of Pragmatics* 145 (2): 31–38.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.9
Received: March 2, 2023
Accepted: September 14, 2023

Rationality and Intransitivity

WALTER VEIT
University of Reading, Reading, UK
Ludwig-Maximilians-Universität, München, Germany

The axiom of transitivity has been challenged in economic theorizing for over seventy years. Yet, there does not seem to be any movement in economics towards removing classical rational choice models from introductory microeconomics books. The concept of rationality has similarly been employed in the cognitive sciences and biology, and yet, transitivity has here not only been shown to be violated, but also rationally so. Some economists have thus responded with attempts to develop alternative theories that give up on the axiom of transitivity. In this paper, I argue that there is a conceptual confusion in this debate that rests on the mistaken idea that there is something like the "one true theory of rationality" that can determine axioms like transitivity to be true or false. Instead, I defend a shift towards a pluralism of concepts of rationality as well as models in which transitivity should play a role depending on the purposes of the model at hand.

Keywords: Idealization; rationality; transitivity; preference; choice; evolution; models.

> "Shall I say, 'a rational animal'? No, for then I should have to examine what exactly an animal is, and what 'rational' is, and hence, starting with one question, I should stumble into more and more difficult ones."
>
> Meditation II of *Meditations on First Philosophy*
> – René Descartes (2008: 25)

1. *Introduction*

When Descartes set out to provide a new metaphysical system for philosophy, he rejected the Aristotelian answer or rather definition of man as the "rational animal" as methodologically flawed. While I share little agreement with Descartes' metaphilosophy, he rightfully recognized that the question of what it means to be rational is a highly complex one. Aristotle's motivation behind classifying humans as the "rational animal" was to distinguish humans from other animals. This definition, of course, runs into a number of conceptual and empirical problems—even being mocked by Bertrand Russell:

> Man is a rational animal—so at least I have been told. Throughout a long life I have looked diligently for evidence in favour of this statement, but so far I have not had the good fortune to come across it. (Russell (2009: 45)

Naturally, the concept of rationality has been the subject of one of the longest conceptual debates in the history of philosophy. When is an agent rational? Is there a difference between the rationality of human and non-human animals (henceforth animals)? Do rational agent models accurately represent these targets in the real world? If not, can they nevertheless be explanatory? Despite the attention "rationality" has received, only little consensus has emerged. The debate is so vast indeed that no single *Stanford Encyclopedia of Philosophy* article on rationality has even been attempted. There is, however, a large number of articles on preferences, decision-making, utility, practical reason, and instrumental rationality.[1]

In this paper, I argue that this scattered picture should be taken serious as a reflection of the *disunified* nature of the cluster of ideas relating to rationality, rather than a mere reflection of the philosophical complexity of the term "rationality." I will argue that a lot of confusion in this debate rests on the mistaken idea common among philosophers (though also economists, psychologists, and biologists) that there is something like the *one true theory* of rationality that we only have to uncover and formalize. Instead, I defend a pluralist view of the concepts of rationality, as well as a pluralist view of rational choice models, where different assumptions can be more or less appropriate depending on the purpose of the model at hand. I will do so by focusing on one of the most controversial subjects in debates on rationality, i.e. whether our choices must be transitive to be rational, i.e. the *axiom of transitivity*. But before I explain this notion in more detail and outline the structure of this paper, let me briefly introduce a distinction due to Alex Kacelnik (2006) that will be useful throughout the rest of this article.

While philosophers qua philosophers can often be overly ambitious in trying to offer accounts that are as general as possible, scientists routinely lament that such attempts can often neither be successful nor

[1] See Rysiew (2015) for an elegant and brief overview of the conceptual debate.

useful, due to the particular conceptual and methodological challenges of their disciplines. So perhaps it shouldn't be surprising it was a behavioural ecologist, who has been incredibly influential for his interdisciplinary work on rational choice in animals combining methods from economics, biology, and psychology, to cast significant doubts on the idea that we can have a single cross-discipline definition of rationality. In an inter-disciplinary edited volume on the question whether animals can be rational, Kacelnik (2006) lamented that there could not be a definite answer to this question because different fields use the term rationality in very distinct ways. To make this clear, he introduced a distinction between what he called PP-Rationality, E-Rationality, and B-Rationality.

Beginning with the first, the PP in PP-Rationality stands for the concept of rationality as used in philosophy and psychology. Here, Kacelnik (2006) argues that philosophers, psychologists, and cognitive scientists are largely interested in the process of reasoning, and whether beliefs are formed in response to appropriate reasons.[2] In opposition, Kacelnik calls E-Rationality the concept of rationality employed in economics. The target here are actions rather than beliefs, and the outcome, rather than the process of deliberation. For economists, actions are rational if the maximize expected utility. Furthermore, Kacelnik argues that economists not only emphasize—but built their theory of rationality—on the consistency of choice. While this is perhaps an unfairly simple picture of economic concepts of rationality it will serve us well for the purposes of the present paper. As I mentioned above and indicated with the title of this paper, my concern is the axiom of transitivity, which we can simply define as follows: If a rational agent prefers A over B and B over C, they should prefer A over C. To put it more formally, while making room for indifference:

(Weak) Transitivity: If $A \succsim B \,\&\, B \succsim C \rightarrow A \succsim C$

Intuitively, this perhaps most fundamental idealizations in economic theorizing might seem like a common-sense criterion for rationality—not only in economics, but also in psychology, philosophy, and biology.[3] Yet, this seemingly innocent assumption has caused a lot of controversy. Many psychologists and behavioural economists have rejected it as an accurate idealization to describe human behaviour. But there has also been opposition to transitivity as a normative standard for behaviour to meet to be considered rational. Indeed, one immediate objection one could raise to Kacelnik's PP-Rationality, is that philosophers as well as psychologists are very much interested in the rationality of actions, rather than just beliefs. Nevertheless, we could simply expand this concept here to include the process of rational belief

[2] This concept may require introspective capacities, and may thus surprisingly be applied to non-human animals and AIs (Browning and Veit 2023).

[3] Unsurprisingly, philosophers have been among those who have criticized the rational choice axiom of transitive preferences early on (Schumm 1987).

formation as well as decision-making. This, however, is already quite the substantial commitment about the nature of rationality and does not reflect the entire spectrum of philosophers. Let me therefore follow Okasha (2018) and abbreviate PP-Rationality as P-Rationality. Unlike Okasha, however, I do not intend this merely as an abbreviation, but a reflection of the narrower conception of rationality within the psychological sciences to focus on a descriptive rather than normative account of rational belief formation and decision-making. Economists, as we shall see, are much closer philosophers than psychologists in their motivation to offer a concept of rationality that is also normative. Finally, B-Rationality describes the rationality concept used in biology as a place-holder for fitness-maximization. Just like for E-Rationality, behaviour is considered rational if it maximizes a quantity, but instead of utility it is fitness (i.e. reproductive value). Indeed, fields like evolutionary game theory make clear how these conceptions can influence each other (Veit 2023c).

As I shall argue in this paper, the conflicts about the status of transitivity for rationality not only reflect different disciplinary goals, but also within-discipline disagreements about the goals of our concepts and models. There is no one correct way of evaluating intransitive preferences and choices. There are parts of economic, and other sciences, where the assumption of transitivity is unproblematic and yields both predictive and explanatory insights, while there are others in which it is misleading. There is no a-priori answer that could help us determine in advance whether this idealization is a good or bad one. Sometimes, the use of this idealization functions as a deliberate misrepresentation of reality for some other purpose, explanatory or otherwise, such as the need to assign utilities to alternative options or to explain an agents choices across a narrow set of options. Worse, economists, cognitive scientists, biologists, and philosophers differ substantially in the reasons and goals for "rationality-talk" even within their own disciplines. I will thus argue here, that we should surrender the idea that a term as polysemous as "rationality" has anything like a one true account that could unify all its different usages. With this throat-clearing out of the way, let me provide a brief outline of the structure of this paper.

1.1. *Outline*

In Section 2, I offer a brief history behind the adoption of transitivity as an axiom of rationality in economics and discuss why transitivity has been so controversial. In Section 3, I will discuss intransitivity observed in animal experiments and debates on the evolution of rational behaviour that cast doubt on the idea that there is a simple answer to the question of whether transitivity should be part of our concept of rationality or not. In Section 4, I draw on the philosophy of science literature on modeling and idealization to argue that the transitivity axiom of rationality cannot simply be assessed as being either correct

or false. Rather, we should adopt a pragmatic and pluralist stance in which we employ different concepts and models of rationality depending on the goal we are using them for. Lastly, Section 5 summarizes and concludes the discussion.

2. *Transitive preferences and rationality*

Leaving aside the question of group-rationality and how intransitive group choices can emerge from individually "rational" behavior or vice-versa, I shall offer here a brief overview of the roles *transitivity* plays in economic theorizing and how it has been defended. I should note, however, that collective entities such as companies can are often usefully treated as individuals that conform to a rational agent model. A similar point applies to much work in contemporary political science that treats nations as individual rational agents, an assumption that has not gone without criticism (Green and Shapiro 1996).[4] What began with Adam Smith (2010) as the study of wealth, quickly became the science of rational choice theory. Many decision and game theorists, especially those working in philosophy, and arguably even the founders of decision theory itself, von Neumann and Morgenstern (1944), argued that it is a normative, rather than descriptive theory of how humans should act.

In one of the most influential monographs on economic methodology, Lionel Robbins (1935) detached economic thinking from psychological welfare considerations and material exchanges. He redefined the discipline more abstractly as "the science which studies human behavior as a relationship between ends and scarce means which have alternative uses" (Robbins 1935:16). This could be considered the birth of microeconomics in its modern sense, i.e. the study of individual choice behavior of economic agents. Others, i.e. many behavioral economists (Camerer 1999; Ashraf et al. 2005; Thaler 2016) and philosophers (Rosenberg 1992, 1994, 1995, 2009; Angner and Loewenstein 2007), see this as an unfortunate mathematization and loss of realism of the discipline. But as economists following Robbins argued: economics is not necessarily about humans or the human domain traditionally seen as markets[5]—it is about the optimization of choice behaviour.

Naturally, this conception of economics has led to an expansion of the proper domain of economics and invited the charge of economics imperialism, i.e. the extension and application of economic methods and models to explain and predict phenomena traditionally viewed beyond

[4] These models, after all, are fundamentally based on the original one of individual human agents in economics. There are, however, interesting parallels here between such collective human organizations and collective multi-cellular organisms (Okasha 2018; Veit 2019a, 2021a).

[5] In addition, biologists have extended market thinking to develop what they call *biological market theory*. See Noë and Hammerstein (1994, 1995) and Noë et al. (2001).

the scope of economics (Becker 1976; Stigler 1984; Tullock 1972; Levitt and Dubner 2005; Mäki 2009a). Rational agent models have been used to explain criminal behaviour (Becker 1973, 1974), marriage (Becker 1968), politics (Tullock 1972), and science itself (Diamond 2008). For his work on expanding the bounds of economics and rational choice theory, Chicago economist Becker was eventually awarded the Nobel Memorial Prize in Economic Sciences. In his Nobel lecture, he stated:

> I have intentionally chosen certain topics for my research—such as addiction—to probe the boundaries of rational choice theory. [...] My work may have sometimes assumed too much rationality, but I believe it has been an antidote to the extensive research that does not credit people with enough rationality. (Becker 1993: 402)

The charge of economics imperialism against the likes of Becker can be seen in two ways, one of which is to be condemned, the other appreciated. When Becker (1993) argues that social scientists have not taken rationality of humans seriously enough, it would be a stretch to defend the thesis that all human choice behavior corresponds to a demanding set of axioms satisfying both *completeness* and *transitivity*. Behavioral economics is an antidote to this way of doing economics, not as a grand unifying theory of human rationality, but as an alternative methodology that provides a variety of models that explain the anomalies of rational choice theory. If economics is conceived of as a more pluralist discipline with a variety of alternative and complementary models for the same phenomena, there wouldn't be a problem of economic imperialism, since all that is imported is a variety of new tools to formerly distinct disciplines.[6] Perhaps though, the label imperialism is misplaced for the latter approach. Instead, one should see the application of economic theories and models to phenomena in other fields as *economics borrowing*, and only the additional goal of replacing theories of "irrationality" with rational choice models as economics imperialism. With this lesson in mind, let us turn to actual economic modeling practice and how the *axiom of transitivity* is defended.

For the purposes of this paper, Kacelnik's definition of economic rationality as consistency will do well enough. Here, he is not so much drawing his own distinction, but rather using the notion of rationality that rational-choice theorists have defended for decades. This way of thinking about rationality goes beyond Robbins' definition of economics as the study of the optimal achievement of goals under scarcity, i.e. instrumental rationality. With the introduction of expected utility theory (von Neumann and Morgenstern 1944), consistency as transitive orderings among preferences became a necessary axiom to calculate utility. Von Neumann and Morgenstern's theorem assumes probability distributions to be given over the outcomes of actions. Their theorem shows that we can only assign utilities if an agent's preferences conform to the axioms of rational choice theory. Because we often do not know the

[6] Thaler (2016) and Rodrik (2015) offer similar conciliatory words.

objective probabilities over outcomes, Savage (1954) developed a highly influential theory of "subjective" probability that was subsequently adopted and used to calculate subjective expected utility. The axiom of transitivity plays therefore a necessary role in much of economic theorizing and has been defended as a necessary idealization. Critics on the economics side have attempted to develop more realistic alternatives such as bounded rationality (see Herbert Simon 1955, 1972, 1991, 1997) that is in line with research in behavioral economics. Despite the development of alternatives, however, most of contemporary rational-choice models, whether normative or descriptive continue to rely on the transitivity of preferences. But as already pointed out, it is not my goal here to defend one account over another. Indeed, as the following discussion will illustrate—I will argue these methodological discussions to rest on outdated views in the philosophy of science.

Transitivity of preferences is at the very center of methodological debates about rational choice theory. Much empirical evidence, however, has accumulated showing that the assumption of transitive preference orderings lacks real-world evidence.[7] Economic models that make use of transitive preference orderings frequently fail to make accurate predictions about the choice behavior of humans. Unfortunately, however, many of these economic models are reliant on this assumption, without which it would not be possible to move from preferences to utility. Due to considerations of space, I leave the question open here of *what* preferences are. It would be a mistake, however, to think that psychological approaches to economics are all in support of a mentalistic interpretation. The phenomenon of rationalization in psychology, i.e. the retrospective attribution of hidden beliefs and desires to oneself, could support a behaviorist interpretation of preferences (see Veit et al. 2020). If the "behaviorist" interpretation of preferences is correct, E-Rationality and B-Rationality would move closer together. If unification is the goal, however, there is strong case to be made for a preference account based on Daniel Dennett's (1989) *intentional stance*, which attributes beliefs and desires to systems to predict and explain their behaviour as those of a rational agent. This idea has subsequently been developed by Don Ross (2005, 2014) for the purposes of economics. I have sympathies for this ambitious account, as unlike anything offered in the literature so far, it has at least some potential to unify all three accounts of rationality. In a recent work with others, Don Ross has attempted to develop the idea of a "quantitative intentional stance," as a truly economic, rather than merely philosophical, account of preferences as constructions (see Alekseev et al. 2019). Intransitive preferences could then (at least to some extent) be explained away as mere "noise."

Some economists have proposed alternatives that seek to maintain

[7] See Sen 1969, 1970, 1971, 1977; Grether and Plott 1979; Suzumura 1983; Korhonen et al. 1990; Bradbury and Ross 1990; Fishburn 1991 for a number for important criticisms and proposed alternatives.

something close to "quasi-transitivity" (Sen 1969; Panda 2018) in order to improve the realism in their models. Others have defended the transitivity assumption as a normative principle, rather than an empirical one—but even this assumption has been challenged by many philosophers and economists. These debates are notably absent from most economic textbooks (with the exception of behavioral economics). Anand (1993) while very critical of transitivity assumptions in economics, considers the basic idea of "considerable pedagogical value" (1993: 345). This is an idea that has been picked up by several economists and philosophers to argue that introductory books and lectures to economics give a misleadingly narrow picture of the field at large.[8] This, however, need not be a problem. The subject matter of economics is complex and it might be best to start with highly idealized models that include the axiom of transitivity, even when its role is merely heuristic.

Nevertheless, the literature has provided three primary arguments for transitive preference-orderings that Anand (1993) in his influential essay sought to dispel. Firstly, Anand argues that transitivity has been defended as logical consistency. Here, intransitivity is simply a logical mistake—analogous to a mistake in logical reasoning—defended for instance by Broome (1991). This, Anand argues does not work, for it locates the mistake not in the logical preference relation, but the assumption that preferences cannot change if options are added or removed (an assumption that has been challenged in the literature, see Sugden 1985).

Secondly, Anand points to the defense of transitivity as something embedded in the concept of rationality itself. Here Anand (1993: 340) quotes a passage Davidson (1980),[9] who argues that:

> theory [...] is so powerful and simple, and so constitutive of concepts assumed by further satisfactory theory [...] that we must strain to fit our findings, or interpretations, to fit the theory. If length is not transitive, what does it mean to use a number of measure length at all? We could find or invent an answer, but unless or until we do, we must strive to interpret 'longer than' so that it comes out transitive. Similarly 'for preferred to'. (Davidson 1980: 273)

Anand argues that we should not overestimate this metaphor. In order to do so, he introduces an alternative metaphor, i.e. idea of pair-wise competitions of sport teams. While the highest ranked team frequently beats the second ranked team, a lower-ranked team might have the perfect composition to beat the first ranked team. There is nothing surprising about such reversals in sports, indeed, it would be ludicrous and boring if the highest ranked team beats all others, the second highest ranked team beats all except for the first – and so on for the entire ranking list.

[8] See Rodrik (2015); Ylikoski and Aydinonat (2014); Aydinonat (2018); Veit (2019b, 2021b).

[9] Anand (1993: 340) accidentally cites page 237 of Davidson. The actual page number is 273.

Anand (1993) does not so much as argue that this is the right interpretation of preferences, but rather to make the point that these are mere metaphors and there is no a-priori reason or empirical evidence as of yet to think that one of them is *the way* of seeing preferences. Instead, we might be well-advised to see these different suggestions as mere metaphors. Interestingly, Nancy Cartwright (2019) makes a similar argumentative move when she criticizes the metaphorical idea of "laws of nature" and "nature doing it by the book," instead introducing her own metaphor of "nature as an artful modeler." While I find the metaphor misplaced, one can see how easy it is to be tempted by metaphors. If one disagrees with the metaphors of a particular theory, whether in philosophy or science, it will often be necessary to come up with alternative metaphors. Dennett vaguely alludes to this possibility as "war of metaphors" (1991: 455), when he defends the use of metaphors as tools of thought. When there are two sides of a debate, and one has metaphors in their arsenal while the other doesn't, the latter will be put into a disadvantaged position. Defenders of the transitivity axiom unfortunately had this unrecognized advantage for the majority of the debate.

In addition to Anand's criticism, it is important to note that Davidson's defense of transitive preference orderings is based on outdated views in the philosophy of science. Davidson states that "Hempel set out to show that reason explanations do not differ in their general logical character from explanation in physics or elsewhere" and that his own "reflections reinforce this view" (1980: 274). While he avoids the conclusion that we can extrapolate to general laws about human behavior—he argues that we can find general laws about individual humans such as *Gerald Ford* that would apply under certain conditions. This idea is deflating the idea of laws to such narrow domains, that it is hardly even worth speaking of laws, and even in such a narrow domain they are unlikely to be exceptionless. More commonly, philosophers of science are now following Nancy Cartwright's (1983) suggestion to see such generalizations as useful idealizations in models. The discovery of general laws is no longer seen as a necessary condition for successful explanation.

Lastly, and perhaps most importantly, Anand (1993) discusses a popular *reductio ad absurdum* argument against critics of the transitivity axiom, i.e. the money pump. The argument goes as follows. Suppose we have an agent who prefers A over B, B over C, and C over A. Suppose now that this agent is in possession of B. Because of the cyclical preference structure of this agent, a merchant who is in possession of A and C should be able to swap his own A for the agent's B in addition to a tiny amount of money such that the preference relation between A and B remains intact. Since the merchant is also in possession of C he will be able to expose the agent to a continuous set of exchanges with a minor additional cost that he would be "rational" to agree to given

his cyclical preference ordering. These repeated exchanges, however, would eventually lead to the bankruptcy of our agent holding cyclical preferences. Hence, they are being money-pumped.

This argument is a strong and intuitive one, for it seems to suggest that unless we accept the transitivity of preferences as a necessary requirement of rationality—it would be rationally required to give away all of one's money. The assumption has been criticized on the grounds that it seems to assume a stable preference set over an entire life, but this does not seem to be a requirement of rationality. There is a stronger counterpoint against the money pump argument, however, that draws on literature in evolutionary biology and behavioral ecology. But before we turn to the literature on intransitive choice in animals, let me briefly summarize this section.

As this section hoped to make clear, the axiom of transitivity has long played a central role in economics in order to enable meaningful attributions of utilities to alternative choices. This instrumentalist defense of transitivity, however, has been criticized by economists and psychologists who were interested in actual choice behaviour. One might describe this conflict thus as one between the normative-idealist stance of mainstream economics and the descriptive-realist stance of behavioural economists and psychologists. Some economists may object to being described as "normativists," but arguments like the money pump rely upon the normative assumption that it is bad to be exploited. Nevertheless, economists have tried to justify the normativity of the transitivity axiom through recourse on a purely descriptive kind of normativity in biology to which we shall now turn, i.e. the maximization of fitness.

3. *Intransitivity and evolution*

Unlike the "Rational Animal," non-human animals are often taken to be irrational. This philosophical conception of rationality goes back to Aristotle and was intended to distinguish man from animal. For the purposes of this paper, we will discard this a priori distinction between humans and animals and show that there is much to learn from the debate on intransitive preferences in non-human animals.

While the P- and E-concepts of rationality seemed incompatible, economists frequently suggest that there is a more important form of rationality economists can rely on, even if the E-concept fails to represent and accurately explain actual human thought processes in markets, i.e. B-Rationality. This Biological Rationality concept is simply the maximization of fitness—and, hence, was often used as an analogue to justify models that assume the maximization of utility (see Okasha 2018; Okasha and Binmore 2012). E-Rationality, however, is frequently violated by both humans and animals. So it is worth exploring whether the connection to B-Rationality can actually help economists to justify their highly idealized form of E-Rationality.

In a biological context, "optimal" often replaces talk of "rational" (see Smith and Harper 2003; Okasha 2018). The optimal choice, in terms of maximization of fitness, then becomes the parallel to the rational choice, i.e. the choice that maximizes utility. The parallel is obvious, but it is not clear how far the analogy stretches and whether it is, indeed, a useful one.

When it comes to E-Rationality there is now an extensive literature on rational choice behavior in animals. McGonigle and Chalmers (1992) for instance argue that squirrel monkeys are capable of transitive choice behavior. For non-human animals, it is sometimes assumed that optimal behavior, i.e. fitness-maximizing behavior, would always correspond to the transitivity axiom, but as Okasha (2018) points out this need not be the case. He discusses a biological optimality model of Houston et al. (2007) in which transitivity is violated—and yet fitness maximized. The Houston et al. (2007) paper is thus aptly titled "Violations of transitivity under fitness maximization." In their model, animals have to choose between three different foraging options. Each option is associated with a different predation risk and an associated chance of success. The nutritional value itself is equal for all. Whether a particular option is preferred to another depends on the state the animal is in. The "goal" for the animal, however, as Okasha (2018) notes is to survive the winter and avoid starvation. Houston et al. (2007) show that the best strategy (to maximize fitness) involves intransitive choices for a range of intermediate energy reserves, i.e. neither full nor starved.[10]

The moral here, as Okasha points out, is a similar one to an important result in the behavioral economics literature. When we analyse choices in isolation, they may violate transitivity and appear irrational. The actual strategies that underlie the choice behavior, however, might be rational because they are about repeated actions. What should be rationally evaluated then is not the individual choice but the strategy itself.

Consider the simple thought experiment of a hypothetical conference meeting with a long queue in front of the food-stand. Our human agent, let us call him Bob, is given the option between eating a salad, a plate with sliced peaches, or a steak. Bob picks the steak. However, it turns out there is more food than participants so everyone is allowed to choose again. After Bob has enjoyed his steak, he proceeds to join the queue again. This time, however, he chooses the salad. How odd you say? Let us make matters worse. Once again, there are food leftovers. Bob joins is faced with the three items once more. This time, however, he chooses the sliced peaches. Now our straw-man economists might yell: "How irrational!" Psychologists, of course, have no problem with explaining such choice behavior. But neither do contemporary economists.

[10] Okasha (2018) discusses this example in more detail.

Clearly, it need not be irrational if Bob chooses the steak, and is subsequently allowed to once again choose between the two after he has devoured the steak, other people have made their choices, and there are leftovers. As Okasha (2018) nicely illustrates, behavioral economists have here responded in a similar way to biologists such as Houston et al. (2007); McNamara et al. (2014) who note that the irrationality disappears once we change our perspective to look at the level of strategies, rather than just the individual choices, a view that is gaining support through recent work in the neurosciences (see Kalenscher et al. 2010). Thus, the evolutionary most "rational" strategy can lead to intransitivity among individual choices.

This explanation is also able to explain the tendency of children and infants to exhibit intransitive preferences that seems to stem from a preference over novelty that is lost over time (Bradbury and Ross 1990). We could rationalize this as the progressive development of "rationality" into adulthood—or a beneficial exploratory phase during early years. Curiosity could be a useful exploratory strategy in rapidly changing environments, for instance. Similar patterns can be found in the foraging behavior of bees (Shafir 1994). This is a better response to the money pump argument: we often need to take the context, time, and number of repeated choices into account. This has led Gigerenzer and Todd (1999) and Smith (2003) to develop, what they call *Ecological Rationality*, as an alternative to standard Rational Choice Theory. Again, it is not my goal here to defend one "Rationality" account over another, but rather to highlight the importance of idealization when the concept is used in practice.[11]

Having addressed the major opposition to the abolition of the transitivity axiom we shall now turn to the much more interesting philosophical questions concerning idealization and representation by drawing on the philosophy of models literature.

4. *Rationality Redux*

As the previous sections should make clear, the disagreements about how we should conceptualize rationality do not just reflect the complexity of the concept. Rather, the disagreements are indicative of deeper differences in regard to why we use the models, concepts, and other clusters of ideas related to rationality at all. Thus, my goal in this section will be to draw on the philosophy of science literature on modeling and idealization to argue that the transitivity axiom of rationality cannot simply assessed as being either correct or false. Instead, I will

[11] I will note, however, that this doesn't mean that there can be useful connection between these concepts. As I've argued in a recent book, the demands on animals to engage in optimizing behaviour could explain the evolution of Benthamite creatures with economic agency that have a common currency to rank/evaluate alternative actions, thus perhaps providing an evolutionary bridge between these concepts (Veit 2022, 2023a).

defend a pragmatic and pluralist stance in which we employ different concepts and models of rationality depending on the goal we are using them for.

As is indicative of the rational choice axiom of transitivity that I have focused on in this article, the last 70 years appear to show no success in removing classical rational choice models from introductory microeconomics books despite many criticisms. Indeed, in these 70 years a huge variety of elegant alternatives have been developed that do not rely on the axiom of transitive preference ordering, or least only a weaker version. To some extent, this literature may appear an endeavor in futility. None of the successor models have achieved sufficient prominence to replace the original status of the transitivity axiom. Here, both economists and philosophers have been misguided. It is a mistake the following quote from Fishburn's (1991) review of the literature elegantly illustrates:

> If the variety of representations is more confusing than illuminating, one would hope that further research during the next few decades will help to identify the most viable models on the basis of philosophical arguments, empirical robustness, and applications potential. General but elegant models that are capable of representing what most researchers agree are reasonable patterns of preference will likely prevail. Some of these surely await discovery. (Fishburn 1991: 131)

Almost 30 years later, we must recognize that Fishburn's prediction failed. No general model has been "discovered" that is able to represent all reasonable patterns of preference.[12] Is this a failure of economics? I suggest not. Indeed, we should see the extreme proliferation of rational choice models as an utter success. But we need to change our understanding of what economists have achieved. Even though many of the economists engaged in this debate had the goal of developing a general model that is able to cover a broader range of phenomena, almost all of them failed. But this does not mean that there was no progress in the last 70 years in our understanding of rational choice behavior. A consensus has emerged that there are certain circumstances under which the transitivity axiom is unproblematic, elegant, and predictively powerful.

Reasonable economists have given up on the idea that transitivity of preferences is a general feature of *all* rational choice behavior. To this end, a large number of theoretical and empirical contributions from psychology, economics, philosophy, and biology have added to our understanding of "rationality" as a cluster of concepts, rather than a single one. There is no single phenomena of rationality in nature that could unify these different concepts and models. To recognize this, however, we must shift our understanding of models away from what Veit (2019b, 2023b) has called "model monism" or "model essentialism," and towards are more pluralist position he has dubbed "model pluralism:"

[12] Let alone elegant.

> (i) any successful analysis of models must target sets of models, their multiplicity of functions within science, and their scientific context and history and (ii) for almost any aspect x of phenomenon y, scientists require multiple models to achieve scientific goal z. (Veit (2019b: 92–93)

While unification is certainly a worthwhile goal, there is a misguided tendency within economics to seek *the one perfect and general model*. This tendency should be avoided. But in practice, not much will have to change for economists—they can and should continue to build new models and expand our toolkit of possible explanations. Articles, such as Regenwetter et al. (2011), attempt to rationalize many of the empirical studies on intransitive choice as *actually* consistent with transitive preferences. I see this as a double-edged sword. On the one hand, I am reluctant to accept the calls to abolish traditional rational choice theory by some of its critics. On the other, I am not willing to grant that the conclusion, that because many of these studies are somewhat consistent with axiom of transitive preference orderings, we do not need alternative models. The debate, however, is often put in a very monist and competitive way. This, I hope to have succesfully illustrated, is a mistake. Instead, we need to embrace a pluralism of alternative models.

Granted, for my proposed changes to succeed, there will have to be a major change in the public understanding of the core role of idealizations in economics. Philosophers are well-advised to promote this change, rather than argue against the viability of idealizations in science. Idealizations are everywhere. It is important to see them as tools for our models to perform their intended roles. Whether it is explanation, prediction, or even unification—idealizations are a must.

The topic of idealization, however, has been one of the most longstanding debates in the philosophy of science literature, much of which we consider too critical (e.g. Cartwright 1983, 2009; Hausman 1992; de Donato Rodriguez and Bonilla 2009; Knuuttila 2009; Mäki 2009b; Reiss 2012; Northcott and Alexandrova 2015; Fumagalli 2015, 2016). Idealizations as distortions, misrepresentations, and falsehoods, have often been viewed with suspicion, if not contempt, by more traditionally inclined philosophers. These views are indicative of a more general tendency among philosophers of science to come up with sweeping generalizations about science—a dangerous tendency that has contributed to a sometimes quite dismissive picture of philosophy of science by scientists.[13]

This way of thinking, however, is beginning to change. Thanks to philosophers such as Michael Weisberg,[14] Ronald Giere (1999, 2006), Peter Godfrey-Smith (2006), Angela Potochnik (2017), N. Emrah Aydinonat (2018), and hopefully myself (Veit 2019b), there is now a growing un-

[13] See Maynard Smith (1997); Godfrey-Smith (2003); Veit (2019b, 2023b).

[14] Weisberg has published a number of highly influential articles on models that I deem to be of special importance for the shift towards a more pluralist understanding of models in the literature: see Weisberg (2003, 2006b,a, 2007b,a, 2012), Weisberg and Reisman (2008), Matthewson and Weisberg (2009), Weisberg et al. (2011), Elliott-Graves and Weisberg (2014)

derstanding of the necessary and diverse roles idealizations play within science. It is into this tradition that the present article squarely falls.

As I have illustrated above, the debate about rationality in economics has unfortunately suffered from a lot of bad methodological and conceptual confusions regarding the need for consensus on a single definition of rationality. Akin to debates between political parties a rift has opened between critics and proponents of economics, with both sides seeing the other as political partisans and holders of naive views about science. Economists have responded to challenges of the transitivity axiom in variety of ways. Critics, however, especially from the psychology-friendly side of economics, i.e. behavioral economics, remain unconvinced. Subsequently, economists have developed a number of alternatives for traditional expected utility maximization that do not rely on transitive preference orderings and that are more or less in line with the idea of bounded rationality (see Morrison 1962; Tversky 1969; Fishburn 1982; Bell 1982; Loomes and Sugden 1982).

How should one interpret these alternative models of rational choice? It was my goal here to dispel the perceived need for a unified account that covers all of economic (and possibly biological) choice behavior. Economic imperialism has led to the application of rational choice theory to a variety of phenomena, formerly seen as outside the domain of economics. The problem here is not the application of the models itself. We should treat them as idealized tools that can at best only partially represent the world. Yet, the use of diverse tools enables us to discover new explanatory insights, a point that has recently gained prominence through a position that has come to be named "Perspectivism" or "Perspectival Realism."[15] This does not entail that we should become anti-realists about "Rationality," yet it does require changes in how we perceive it.

Should we, for instance, consider failure to exhibit transitive choice behavior in other animals, such as hoarding gray jays (Waite 2001), as a depiction of their "irrational" behavior? I think not. The question is ill-posed and presumes that there is a general answer to questions involving the concept of "Rationality," which Kacelnik (2006) early on tried to warn us off. As I hope to have convinced the reader, rationality might not be the unified phenomenon philosophers have taken it to be. Rather, it is a lose collection of metaphors, models, and idealizations—epistemic tools that help us to explain and make sense of the world. The different concepts we may associate with rationality, such as E-, P-, and B-Rationality reflect genuinely different phenomena that may have similarities, but shouldn't be grouped together. Indeed, we should move away from attempts to provide the one true account of rationality. This is, as has hopefully become clear now, a hopeless endeavor. A more subtle and pragmatic way forwards for economics (and other

[15] See Giere (2006) for the first articulation of the view, and Massimi (2017) for a recent overview.

disciplines making use of the concept of rationality), would be to embrace a pluralist perspective, and defend models that are not intended to replace all others but instead illuminate a novel aspect or provide a new perspective of a phenomena.

5. *Conclusion*

In this article, I have criticized the common attempts to find something like the one true theory of rationality or for that matter truth or falsity of the axiom of transitivity. One immediate response to such criticisms will naturally be what we should be doing instead. Drawing on the philosophy of modeling literature, I have therefore argued that we should reconceptualize these debates in terms of determining useful models for different purposes. This will help us to recognize the different conceptualizations of rationality in (evolutionary) biology, economics, and psychology as reflecting different interests. We should see the concepts of rationality and its axioms such as transitivity as idealized conceptual tools, rather than accurate explications of "the one true" concept of rationality.

There is a special explanatory force that comes from explanations invoking "Rationality" and "Reason" to us as cognitively limited agents that evolved to talk and think in normative terms—but it is a tempting force that might lead us into the wrong conclusions if we mistake what are useful tools for representations of reality.[16] The final conclusion for economists (and for that matter biologists and cognitive scientists) is a simple, but philosophically less interesting one: there is set of cases where it is reasonable and/or useful to accept the axiom of transitive preference orderings, while it is not for others. No generalized defense or rejection of this idealization can be offered. The real insight and philosophically much more interesting one is this: we may have to give up on the idea of rationality as a unified concept or phenomena, and instead think of it as a useful set of metaphors, models, and idealizations.[17]

References

Alekseev, A., Harrison, G. W., Lau, M. and Ross, D. 2018. "Deciphering the Noise: The Welfare Costs of Noisy Behavior." *Working paper / Center for Economic Analysis of Risk (CEAR)* 2018-01: 1–45.

Anand, P. 1993. "The Philosophy of Intransitive Preference." *The Economic Journal* 103 (417): 337–346.

Angner, E. and Loewenstein, G. 2007. "Behavioral Economics." In U. Mäki (ed.). *Philosophy of Economics*. Amsterdam: Elsevier, 641–690.

Ashraf, N., Camerer, C. F. and Loewenstein G. 2005. "Adam Smith, Behavioral Economist." *Journal of Economic Perspectives* 19 (3): 131–145.

[16] See Godfrey-Smith (2013) and Okasha (2018).

[17] I would like to thank Heather Browning and two anonymous reviewers for their feedback on this manuscript.

Aydinonat, N.E. 2018. "The Diversity of Models as a Means to Better Explanations in Economics." *Journal of Economic Methodology* 25 (3): 237–251.

Becker, G. S. 1968. "Crime and Punishment: An Economic Approach." In N. G. Fielding, A. Clarke and R. Witt (eds.). *The Economic Dimensions of Crime*. New York: St. Martin's Press, 13–68.

_____. 1973. "A Theory of Marriage: Part I." *Journal of Political Economy* 81 (4): 813–846.

_____. 1974. "A Theory of Marriage: Part II." *Journal of Political Economy* 82 (2): S11–S26.

_____. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

_____. 1993. "Nobel Lecture: The Economic Way of Looking at Behavior." *Journal of Political Economy* 101 (3): 385–409.

Bell, D. E. 1982. "Regret in Decision Making Under Uncertainty." *Operations Research* 30 (5): 961–981.

Bradbury, H. and Ross, K. 1990. "The Effects of Novelty and Choice Materials on the Intransitivity of Preferences of Children and Adults." *Annals of Operations Research* 23 (1): 141–159.

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty and Time*. Oxford: Basil Blackwell Press.

Browning, H. and Veit, W. 2023. "Studying Introspection in Animals and AIs." *Journal of Consciousness Studies* 30 (9): 63–74.

Camerer, C. 1999. "Behavioral Economics: Reunifying Psychology and Economics." *Proceedings of the National Academy of Sciences* 96 (19): 10575–10577.

Cartwright, N. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

_____. 2009. "If No Capacities Then No Credible Worlds. But Can Models Reveal Capacities? *Erkenntnis* 70 (1): 45–58.

_____. 2019. *Nature, the Artful Modeler: Lectures on Laws, Science, How Nature Arranges the World and How We Can Arrange It Better*. Open Court Publishing.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

de Donato Rodriguez, X. and Bonilla, J. Z. 2009. "Credibility, Idealisation, and Model Building: An Inferential Approach." *Erkenntnis* 70 (1): 101–118.

Dennett, D. 1991. *Consciousness Explained*. Boston: Little, Brown and Co.

_____. 1989. *The Intentional Stance*. Cambridge: MIT press.

Descartes, R. 2008. *Meditations on First Philosophy: With Selections From the Objections and Replies*. Oxford: Oxford University Press.

Diamond, A. M. 2008. "Science, Economics of." In M. Vernengo, E. Perez and C. J. Ghosh (eds.). *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, 1–9.

Elliott-Graves, A. and Weisberg M. 2014. "Idealization." *Philosophy Compass* 9 (3): 176–185.

Fishburn, P. C. 1982. "Nontransitive Measurable Utility." *Journal of Mathematical Psychology* 26 (1): 31–67.

_____. 1991. "Nontransitive Preferences in Decision Theory." *Journal of Risk and Uncertainty* 4 (2): 113–134.

Fumagalli, R. 2015. "No Learning From Minimal Models." *Philosophy of Science* 82 (5): 798–809.

_____. 2016. "Why We Cannot Learn From Minimal Models." *Erkenntnis* 81(3): 433–455.

Giere, R. 1999. *Science Without Laws*. Chicago: University of Chicago Press.

_____. 2006. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Gigerenzer, G. And Todd P. M. 1999. „Ecological Rationality: The Normative Study of Heuristics." In G. Gigerenzer and P. M. Todd (eds.). *Ecological Rationality: Intelligence in the World*. Oxford University Press, 487—497.

Godfrey-Smith, P. 2003. *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago: University of Chicago Press.

_____. 2006. "The Strategy of Model-based Science." *Biology and Philosophy* 21 (5): 725–740.

_____. 2013. "Darwinian Individuals." In F. Bouchard and P. Huneman (eds.). *From Groups to Individuals: Evolution and Emerging Individuality*. Cambridge: MIT Press, 17–36.

Green, D. and Shapiro I. 1996. *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. Yale University Press.

Grether, D. M. and Plott, C. R. 1979. "Economic Theory of Choice and the Preference Reversal Phenomenon." *The American Economic Review* 69 (4): 623–638.

Hausman, D. M. 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Houston, A. I., McNamara, J. M. and Steer M. D. 2007. "Violations of Transitivity Under Fitness Maximization." *Biology Letters* 3 (4): 365–367.

Kacelnik, A. 2006. "Meanings of Rationality." In S. Hurley and M. Nudds (eds.). *Rational Animals*. Oxford: Oxford University Press, 87–106.

Kalenscher, T., Tobler, P. N., Huijbers, W., Daselaar, S. M. and Pennartz, C. 2010. "Neural Signatures of Intransitive Preferences." *Frontiers in Human Neuroscience* 4: 49.

Knuuttila, T. 2009. "Isolating Representations Versus Credible Constructions? Economic Modelling in Theory and Practice." *Erkenntnis* 70 (1): 59–80.

Korhonen, P., Moskowitz, H. and Wallenius, J. 1990. "Choice Behavior in Interactive Multiple-criteria Decision Making." *Annals of Operations Research* 23 (1): 161–179.

Levitt, S. D. and Dubner, S. J. 2005. *Freakonomics*. New York: William Morrow. Harper Collins.

Loomes, G. and Sugden, R. 1982. "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty." *The Economic Journal* 92 (368): 805–824.

Mäki, U. 2009a. "Economics Imperialism: Concept and Constraints." *Philosophy of the Social Sciences* 39 (3): 351–380.

_____. 2009b. "MISSing the World. Models as Isolations and Credible Surrogate Systems." *Erkenntnis* 70 (1): 29–43.

Massimi, M. 2017. "Perspectivism." In J. Saatsi (ed.). *The Routledge Handbook of Scientific Realism*. Routledge, 164–175.

Matthewson, J. and Weisberg, M. 2009. "The Structure of Tradeoffs in Model Building." *Synthese* 170 (1): 169–190.

Maynard Smith, J. 1997. "Web of Stories." Interview by Richard Dawkins. https://www.webofstories.com/play/john.maynard.smith/1 [Online; accessed 13-September-2023].

McGonigle, B. and Chalmers M. 1992. "Monkeys are Rational!" *The Quarterly Journal of Experimental Psychology* 45 (3): 189—228.

McNamara, J. M., Trimmer, P. C. and Houston, A. 2014. "Natural Selection Can Favour 'Irrational' Behaviour." *Biology Letters* 10 (1): 20130935.

Morrison, H. 1962. *Intransitivity of Paired Comparison Choices*. PhD thesis. University of Michigan.

Noë, R. and Hammerstein, P. 1994. "Biological Markets: Supply and Demand Determine the Effect of Partner Choice in Cooperation, Mutualism and Mating." *Behavioral Ecology and Sociobiology* 35 (1): 1—11.

_____. 1995. "Biological Markets." *Trends in Ecology and Evolution* 10 (8): 336—339.

Northcott, R. and Alexandrova, A. 2015. "Prisoner's Dilemma Doesn't Explain Much." In M. Peterson (ed.). *The Prisoner's Dilemma*. Cambridge: Cambridge University Press, 64—84.

Noë, R., Van Hooff, J. A. R. A. M. and Hammerstein, P. 2001. *Economics in Nature: Social Dilemmas, Mate Choice and Biological Markets*. Cambridge: Cambridge University Press.

Okasha, S. 2018. *Agents and Goals in Evolution*. Oxford: Oxford University Press.

Okasha, S. and Binmore, K. (eds.). 2012. *Evolution and Rationality: Decisions, Co-operation and Strategic Behaviour*. Cambridge: Cambridge University Press.

Panda, S. C. 2018. "Rational Choice with Intransitive Preferences." *Studies in Microeconomics* 6 (1—2): 66—83.

Potochnik, A. 2017. *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Regenwetter, M., Dana, J. and Davis-Stober, C. P. 2011. "Transitivity of Preferences." *Psychological Review* 118 (1): 42—56.

Reiss, J. 2012. "Idealization and the Aims of Economics: Three Cheers for Instrumentalism." *Economics and Philosophy* 28 (3): 363—383.

Robbins, L. 1935. *An Essay on the Nature and Significance of Economic Science*. London: MacMillan and Co.

Rodrik, D. 2015. *Economics Rules: Why Economics Works, When it Fails, and How to Tell the Difference*. Oxford: Oxford University Press.

Rosenberg, A. 1992. *Economics–Mathematical Politics or Science of Diminishing Returns?* Chicago: University of Chicago Press.

_____. 1994. "If Economics isn't Science, What is it?" In D. M. Hausman (ed.). *The Philosophy of Economics: An Anthology*. Cambridge: Cambridge University Press, 376–394.

_____. 1995. "The Metaphysics of Microeconomics." *The Monist* 78 (3): 352–367.

_____. 2009. "If Economics is a Science, What Kind of a Science is it? In D. Ross and H. Kincaid (eds.). *The Oxford Handbook of Philosophy of Economics*. Oxford: Oxford University Press, 55–67.

Ross, D. 2005. *Economic Theory and Cognitive Science: Microexplanation*. Cambridge: MIT Press.

_____. 2014. *Philosophy of Economics*. London: Palgrave Macmillan.

Russell, B. 2009. *The Basic Writings of Bertrand Russell*. London: Routledge.

Rysiew, P. 2015. *Rationality*. Oxford: Oxford Bibliographies. https://doi.org/10.1093/OBO/9780195396577-0175.

Savage, L. J. 1954. *The Foundations of Statistics*. New York: Wiley.

Schumm, G. F. 1987. "Transitivity, Preference and Indifference." *Philosophical Studies* 52: 435–437.

Sen, A. 1969. "Quasi-transitivity, Rational Choice and Collective Decisions." *The Review of Economic Studies* 36 (3): 381–393.

_____. 1970. *Collective Choice and Social Welfare*. San Francisco: Holden Day.

_____. 1971. "Choice Functions and Revealed Preference." *The Review of Economic Studies* 38 (3): 307–317.

_____. 1977. "Social Choice Theory: A Re-examination." *Econometrica: Journal of the Econometric Society* 45 (1): 53–89.

Shafir, S. 1994. "Intransitivity of Preferences in Honey Bees: Support for 'Comparative' Evaluation of Foraging Options." *Animal Behaviour* 48 (1): 55–67.

Simon, H. A. 1955. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics* 69 (1): 99–118.

_____. 1972. "Theories of Bounded Rationality." *Decision and Organization* 1 (1): 161–176.

_____. 1991. "Bounded Rationality and Organizational Learning." *Organization Science* 2 (1): 125–134.

_____. 1997. *Models of Bounded Rationality: Empirically Grounded Economic Reason*. Vol. 3. Cambridge: MIT Press.

Smith, A. 2010. *The Wealth of Nations: An Inquiry Into the Nature and Causes of the Wealth of Nations*. Harriman House Limited.

Smith, J. M. and Harper, D. 2003. *Animal Signals*. Oxford: Oxford University Press.

Smith, V. L. 2003. "Constructivist and Ecological Rationality in Economics." *American Economic Review* 93 (3): 465–508.

Stigler, G. J. 1984. "Economics: The Imperial Science?" *The Scandinavian Journal of Economics* 86 (3): 301–313.

Sugden, R. 1985. "Why be Consistent? A Critical Analysis of Consistency Requirements in Choice Theory." *Economica* 52 (206): 167–183.

Suzumura, K. 1983. *Rational Choice, Collective Decisions, and Social Welfare*. Cambridge: Cambridge University Press.

Thaler, R. H. 2016. "Behavioral Economics: Past, Present, and Future." *American Economic Review* 106 (7): 1577–1600.

Tullock, G. 1972. "Economic Imperialism." In J. M. Buchanan and R. D. Tollison (eds.). *Theory of Public Choice*. Ann Arbor: University of Michigan Press, 317–329.

Tversky, A. 1969. "Intransitivity of Preferences." *Psychological Review* 76 (1): 31—48.

Veit, W. 2023a. *A Philosophy for the Science of Animal Consciousness*. New York: Routledge.

_____. 2023b. "Model Anarchism." *THEORIA. An International Journal for Theory, History and Foundations of Science* 38 (2): 225–245.

_____. 2023c. "Evolutionary Game Theory and Interdisciplinary Integration." *Croatian Journal of Philosophy* 23 (67): 33–50.

_____. 2022. "Complexity and the Evolution of Consciousness." *Biological Theory* 18 (3): 175–190.

_____. 2021a. "Agential Thinking." *Synthese* 199 (5): 13393–13419.

_____. 2021b. "Model Diversity and the Embarrassment of Riches." *Journal of Economic Methodology* 25 (3): 237–251.

_____. 2019a. "Evolution of Multicellularity: Cheating Done Right." *Biology and Philosophy* 34 (3): 34.

_____. 2019b. "Model Pluralism." *Philosophy of the Social Sciences* 50 (2): 91–114.

Veit, W., Dewhurst, J., Dolega, K., Jones, M., Stanley, S., Frankish, K. and Dennett, D. 2020. "The Rationale of Rationalization." *Behavioral and Brain Sciences* 43: e53.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Waite, T. A. 2001. "Intransitive Preferences in Hoarding Gray Jays (Perisoreus canadensis)." *Behavioral Ecology and Sociobiology* 50 (2): 116–121.

Weisberg, M. 2003. *When Less is More: Tradeoffs and Idealization in Model Building*. PhD thesis. Stanford University.

_____. 2006a. "Forty Years of 'The Strategy': Levins on Model Building and Idealization." *Biology and Philosophy* 21 (5): 623–645.

_____. 2006b. "Robustness Analysis." *Philosophy of Science* 73 (5): 730–742.

_____. 2007a. "Three Kinds of Idealization." *The Journal of Philosophy* 104 (12): 639–659.

_____. 2007b. "Who is a Modeler?" *The British Journal for the Philosophy of Science* 58 (2): 207–233.

_____. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Weisberg, M., Okasha, S. and Mäki, U. 2011. "Modeling in Biology and Economics." *Biology and Philosophy* 26 (5): 613–615.

Weisberg, M. and Reisman, K. 2008. "The Robust Volterra Principle." *Philosophy of Science* 75 (1): 106–131.

Ylikoski, P. and Aydinonat, N. E. 2014. "Understanding with Theoretical Models." *Journal of Economic Methodology* 21 (1): 19–36.

Croatian Journal of Philosophy
Vol. XXIV, No. 71, 2024
https://doi.org/10.52685/cjp.24.71.10
Received: November 16, 2022
Accepted: July 27, 2023

A Sufficientarian Proposal for Discharging Our Moral Duties Towards Emigrants[*]

ADELIN-COSTIN DUMITRU
National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania
National University of Political Studies and Public Administration, Bucharest, Romania

In this article I investigate the nature of the moral duties that citizens of a legitimate state have towards emigrants. A large part of the literature dedicated to the normative study of the migration phenomenon focuses on two major topics: the brain drain phenomenon and the legitimacy of restricting immigrations. If the first of these concerns the moral obligations that individuals have towards a state and their co-nationals, the second regards the policies that a state can justifiably adopt in order to manage migration flows. With the exception of temporary labor migration, less discussed in the literature are the moral duties that we have towards those citizens who chose to emigrate. My answer is that a state has neither more, nor less responsibilities towards its emigrants than it has towards the other citizens. However, the particular way that it can discharge those duties have to pay attention to each citizen's particular situation, so that public policies dealing specifically with the emigrants are required. If we embrace a sufficientarian position, we could see how public policies have to be forged in order to be morally justifiable. I compare in the article 2 potential ways in which a state could try to discharge its moral duties towards emigrants. The first consists in promoting policies that focus on reverse migration. The second is based on cooperating with host societies and ensuring that emigrants' rights and well-being are protected to the fullest degree. I argue that the second

[*] This work was supported by a grant of the Romanian Ministry of Research and Innovation, CNCS—UEFISCDI, project number PN-III-P1-1.1-BSH-2-2016-0005, within PNCDI III, as part of the "Spiru Haret" Scholarship that the author had between June 2020–June 2021 from the aforementioned SNSPA Program.

proposal is the one that can be morally defended, and is in line with moral defenses of reformed temporary labor migration programs which would take into account the rights and legitimate interests of migrants (Baubock and Ruhs: 2022).

Keywords: Brain drain; migrants; reverse migration; sufficientarianism.

1. *Introduction*

The normative study of emigration has focused in the last couple of years on two major topics: the *brain drain* phenomenon (Blake and Brock 2015; Owen 2016; Ypi 2016; Pevnick 2016; Okeja 2017; Yuksekdag 2018, 2019; Niimi, Ozden and Schiff 2010; Glytsos 2010; Beine, Docquier and Oden-Defoort 2011; Ferracioli and De Lora 2015; Kaplan and Hoppli 2017) and the legitimacy of imposing restrictions on immigration (among the defenders of such restrictions are Walzer 1983; Kymlicka 2001; Miller 2005; Pevnick 2009, 2011; Wellmann 2008, 2011; while among the proponents of relaxing them are Carens 1992; Kukathas 2005; and Cole 2011). Whereas the first subject concerns the duties that citizens have towards a state in which they had been educated and towards the citizens of that state, the second tries to shed light on what measures states can justifiably take when it comes to the admission of potential immigrants. What seems to be undertheorized, however, is the subject of the duties that we have towards our compatriots who chose to emigrate. What are those duties and how can we justify them? Furthermore, given that most our duties are usually discharged through institutions, what are the public policies that can be taken by the state towards emigrants? One important exception is the literature on temporary labor migration programs (Carens 2008; Lister 2014; Barry and Ferracioli 2018), which sometimes explicitly deals with what is owed to migrants by both the destination and the origin countries (Baubock and Ruhs 2022).

My position in this article is that there is nothing *sui generis* about the duties that we have towards emigrants. Nonetheless, we must take into account the fact that the particular way in which we discharge those duties might have to be sensitive to them living in another country. For instance, if we embrace a sufficientarian view, according to which social justice is realized when people have secured enough resources, capabilities, or welfare, one must account for the different strategies that can be employed in order to achieve this ideal for the residents of a state and for its citizens living abroad. Starting from such a sufficientarian position, I investigate two potential ways in which a state can fulfill its moral obligations towards emigrants. The first consists in creating some conditions that are good enough at a national

level so that any emigrant who so desires could return. This would be founded on a supposed right to stay (Oberman 2011). The second, which I favor, entails carefully drafted policies that ensure that the host country guarantees the emigrants' level of well-being. One way of achieving this is through joint programs involving the country of origin and the host country (Delano 2010). This approached could be called the dual responsibility model and will be further developed towards the end of the paper.

The proposal that I put forward is meant to satisfy a feasibility criterion, and as such it belongs to the realm of non-ideal political theory, in that it issues achievable and desirable recommendations (Stemplowksa 2008: 324). Non-ideal theory is important because it helps us rank options in circumstances that are far from perfect: real-world individuals do not comply with the principles of justice, our resources are limited, it is difficult to judge whether or not the implemented measures will reach their purpose (Swift 2008). Thus, one of the assumptions that I make is that the global political order is unchangeable for the foreseeable future, and that states and borders are here to stay. Feasibility considerations are an important reason why I argue that we should opt for the second solution, in that a right to stay would be too onerous on many on the existing states. Furthermore, assuming that decision-makers are not fully compliant with what justice requires of them means that in real-world scenario such a right to stay would become associated with a deeply ethnical nationalist rhetoric. An advantage of the second proposal is that it fits our intuition that there is something fundamentally problematic in neglecting the responsibility that developed states have towards citizens of less developed states (Blake 2015: 223). Regarding state responsibilities, this is a formulation that I employ in order to avoid wordiness. My approach is individualistic, and it is individuals who are the ultimate bearers of moral duties. However, there are numerous empirical reasons which encourage us to use the institutional framework in order to discharge our duties (Nussbam 2005: 213; Dumitru 2017: 142). According to North (1991), institutions reduce uncertainty and transactional costs, and thus oftentimes are moral duties will have to be discharged through the institutions of the state. Another important point (which is going to be developed further on in the article) is that the proposal is going to be focused on legitimate states, where legitimacy is understood in a minimal sense that hinges on a state respecting human rights.

In order to advance my proposal I proceed as follows. In the first section I present the asymmetry extant in the literature between emigrants' moral duties and their entitlements. In the second section I attempt to explain why this asymmetry exists. The third section tries to answer the question of what duties we might have towards emigrants, employing sufficientarianism as the distributive pattern which might offer an answer to this inquiry. It is in the fourth section that I analyze

two potential ways of discharging those duties, opting for what I labeled the dual-responsibility model. In this forth section I also present how my proposal relates to previous literature, especially the one on temporary labor migration programs.

2. *The asymmetry between emigrants' moral duties and entitlements*

Much of the normative literature on emigrants focuses on the duties that they have towards their countries of origin, while their entitlements are largely a matter analyzed in reference solely to the country in which they immigrated. This is what I call the asymmetry. For instance, much has been written lately about brain drain, "the phenomenon by which the most skilled agents from one economy migrate to live and work in another, where their own personal prospects are enhanced" (Brassington 2012: 113). Brain drain is conceived as "a sort of moral tragedy" (Brock 2015: 272), in that it entails a value conflict between the freedom of the would-be emigrants to pursue a career and a life of their choice and the achievement of justice at the level of their national states, which are going to suffer economically if doctors or other vital workers leave their borders. Many consider that the moral dilemmas associated with this phenomenon stem from the fact that "there are no permissible paths to directly and fully address the brain drain in our current inegalitarian world" (Hobden 2017: 33).

There are several ways in which the brain drain phenomenon challenges our morality. On the one hand, "skilled workers should have the right to exit countries in which they no longer wish to live;" on the other, "there are normative questions about citizens' responsibilities, fair terms of exit, and whether migration should be managed to ensure the burden of migration does not fall disproportionately on the world's worst off" (Brock 2015: 12). The brain drain is considered a problem because it leads to a loss of human capital that in some situations could be extremely detrimental to the development of a country. In order to limit the impact of potential emigration, solutions such as mandatory national employment periods or taxing imposed upon exit have been proposed (Brock 2015: 49–51).

However, there are those, like Blake, who consider that there is a human right to exit, and that "any attempt by a state to forcibly prevent people from leaving that state—to coercively insist upon allegiance and obligation, against the wishes of the would-be emigrant—is fundamentally unjust, and [represents] a violation of the most basic norms of human rights" (2015: 111). Others, like Brock, consider that under special circumstances limiting the right to exit is justifiable. Such conditions include aspects such as thwarting the governments' attempts to discharge their duties by leaving, and having "received important benefits during their residence in the state of origin and

failure to reciprocate for those past benefits involves taking advantage of others or free-riding unfairly" (Brock 2015: 251). Thus, she considers that "programs aimed at combatting the burdens associated with brain drain, such as compulsory service or taxation arrangements, are a helpful set of remedies that can aid the transition to a more just state of affairs" (Brock 2015: 272).

However, no matter how important the emigration of skilled workers is, these are not the only citizens of a country who might choose to emigrate. Low skilled workers are also emigrating in large numbers. Brassington argued that a potential explanation of why brain drain is morally problematic in the context of a migratory route from South to North is that, "by employing Southern experts, North is effectively taking life-sustaining resources from South, thereby wrongfully depriving the Southern population of the means necessary to lead a minimally tolerable life" (2012: 116). However, he also states that this argument is vulnerable to a Kantian objection, in that "it seems to require that the Southern government adopts quite a questionable attitude to its stock of experts, along the lines that they are merely a resource that can be put to better or worse use" (Brassington 2012: 117). Focusing only on containing the emigration of skilled workers and ignoring the emigration of low-skilled workers could reflect a tendency to treat them not by taking into account their rights and entitlements, but rather the ones of the whole society. Here one could advance an objection similar to the one addressed by Rawls to classical utilitarianism, that it "fails to take seriously the distinction between persons" (1971: 163). Blake makes this argument in his defense of the right of skilled workers to exit, mentioning that "the idea is that justification of a sort of coercive policy would have to be made to the person, considered as an individual" (2015: 203–4).

In reply, someone who wants to limit the brain drain phenomenon could make the counterpoint that skilled emigrants who leave their country have not fulfilled yet their duties towards their conationals, and that rather than framing the discussion in terms of the benefits that they bring to source societies, we could rephrase it as involving their duties to host societies. This counterargument only works in non-ideal circumstances and if we assume that the only way potential high-skilled emigrants could discharge their duties would be to remain and work in their source societies. Brock (2015: 88), for instance, considers that actually being in the country of origin is sometimes necessary, giving the example of "a severe shortage of skilled personnel who can assist with particular needs such as administering vaccines or dispensing appropriate drugs." Oberman also develops an argument that includes the following conditions for justifying emigration restrictions on brain drain grounds: 1) a skilled worker owes assistance to her poor compatriots and 2) a skilled worker's duty to assist is enforceable if she stays in her country of origin (2013: 452). However, although Brock takes into account

the unskilled citizens, she does not consider their presence necessary in order that they discharge their duties: "unskilled workers who leave might assist best by working in foreign countries and having a portion of their wages taxed, thereby providing an important revenue stream for source country governments" (Brock 2015: 93). It is unclear why taxing the income earned abroad by high skilled workers, a venue which might generate a greater revenue stream, is not sufficient to reach the conclusion that that they discharged their duties to the citizens remaining in their country of origin. Leaving this aside, although Brock does have something to say about the situation of low-skilled or unskilled citizens who emigrate, she only refers to their obligations. What are these citizens entitled to? Of course, the same question could be asked of the high-skilled citizens, who might end up being discriminated in the host society, being treated disrespectfully or having lower wages than their peers born there. It is more probable, however, that the situation of the unskilled citizens who emigrate would require attention.

There's an important literature that has recently regained ground which takes into account the situation of unskilled or low-skilled emigrants. Baubock and Ruhs, for instance, argue that "temporary migrants" should be "included as local citizens in destination countries and as national citizens in their countries of origin," as "they are still citizenship stakeholders," and both countries "have duties to help them realise their life projects and to involve them in shaping the future of these societies" (2022: 531–2). Furthermore, given that they remain citizens of their countries of origin, it is those that have "special duties to assist them in realizing their life plans through facilitating remittances, return migration and reintegration after return" (2022: 543). Baubock and Ruhs' approach, however, seems to differ from the way other authors discuss temporary labor migration programs, which see persons taking part in such programs *qua* immigrants rather than as emigrants. The difference is a subtle one, but it stems from the fact that most discussions center around the fact that, initially, "worries about temporary labor migration […] stem from an image of the programs that existed in Germany. Foreign workers, most famously from Turkey, worked for extended periods, eventually bringing in family members, but were never allowed access to full societal membership" (Lister 2014: 97). As such, the main focus is on whether or not it is justifiable for temporary migrants not to have a clear path to citizenship (Lister 2014) or on what conditions have to be fulfilled on the labor market in order to avoid the potential exploitation of temporary migrants (Carens 2008; Barry and Ferracioli 2018).[1]

Thus, even with this important exception, there seems to be a noticeable asymmetry between the postulated duties of emigrants and their entitlements *qua* emigrants and members of countries of origin.

[1] I thank an anonymous reviewer for asking that I take into account the literature on temporary labor migration programs.

Too much attention is paid to what they have to do for their countries of origin, and too little to what their countries of origin ought to do in order to help them. In the following section I explore potential reasons for this asymmetry, and I argue that someone who considers brain drain morally problematic should also consider the rights and entitlements of emigrants as morally pressing.

3. *Making sense of the asymmetry*

How can we account for the asymmetry? There are two plausible explanations why there is so much emphasis placed upon the duties of the skilled migrants and so little on the entitlements of emigrants, be they skilled or unskilled. In this section I intend to show why these explanations are not convincing, and ultimately the asymmetry is not morally justifiable.

The first—and more unconvincing one—is that taking care of the migrants falls under the jurisdiction of the country of destination. With few exceptions (Baubock and Ruhs 2022; Lenard 2022), this also seems to be the norm when it comes to moral discussions of temporary labor migration programs. Nonetheless, even in developed and democratic countries there are serious shortcomings regarding the integration of the migrants. In October 2020, *The Guardian* published an expose in which it was shown that migrants in England had been denied treatment by the NHS for an average of 37 weeks, a consequence of the fact that "the NHS deems them not ordinarily resident in the UK."[2] In the context of the global COVID-19 pandemic, the situation of many migrants has been worsened. The most affected have been the refugees and asylum seekers: "depending on the informal economy, they were among the first to suffer the economic impacts of lockdown, losing their jobs and being evicted from their homes."[3] However, the well-being of regular immigrants has also been negatively impacted: "due to a range of vulnerabilities such as a higher incidence of poverty, overcrowded housing conditions, and high concentration in jobs where physical distancing is difficult, immigrants are at a much higher risk of COVID-19 than the native born. Studies in a number of OECD countries found an infection risk that is at least twice as high as that of the native-born."[4]

[2] The Guardian, "Migrants in England denied NHS care for average of 37 weeks, research finds," 14 October 2020, https://www.theguardian.com/society/2020/oct/14/migrants-denied-nhs-care-for-average-of-37-weeks-research-finds, last accessed on 20 October 2020.

[3] UNHCR–United Nations Refuge Agency, COVID-19 crisis underlines need for refugee solidarity and inclusion, 7 October 2020, https://www.unhcr.org/news/latest/2020/10/5f7dfbc24/covid-19-crisis-underlines-need-refugee-solidarity-inclusion.html, last accessed on 20 October 2020.

[4] OECD, What is the impact of the COVID-19 pandemic on immigrants and their children? 19 October 2020, http://www.oecd.org/coronavirus/policy-responses/what-is-the-impact-of-the-covid-19-pandemic-on-immigrants-and-their-children-e7cbb7de/, last accessed on 20 October 2020.

Thus, the challenges faced by immigrants in host societies are sometimes highly specific and often more pressing than the problems faced by the citizens of those countries. Governments focus first and foremost on their citizens, and only then extend aid to immigrants, many of whom are only residents in the countries of destination. One could make the case that the governments ought to treat everyone in the society the same. But it is highly probable that most of the real-world states would try to shirk from their responsibilities concerning a new category of beneficiaries of distributive and welfare policies and would add more immigration restrictions, should their duties to immigrants become more onerous. Thus, assuming that host governments are the main or only duty-bearers in the case of the immigrants' rights will probably not lead to the intended result of improving the well-being of migrants. This would be especially true for more vulnerable categories of migrants—such as temporary migrants.

Of course, there are important exceptions here. On the one hand, we have refugees and asylum seekers, as their countries of destination are ones that fall short of any definition of legitimacy. For them, we'd have to rely on the international protection system, as well as the country in which they receive asylum or other forms of protection. The other exception would be of those individuals who permanently relocate to another country. In their case, it seems that asking the country of origin to continue to discharge its duties towards such individuals would be supererogatory in the case of developed countries and too burdensome in the case of developing or underdeveloped countries. In their case indeed, the intuition that the host society government is first and foremost responsible for their well-being might turn out to be correct to a certain degree.

A second explanation for the asymmetry is the assumption that there is no such thing as an (unqualified) right to leave. Pevnick, for instance, holds that a right to exit one's country can only be defended instrumentally. In his view, "neither rights of emigration nor rights of immigration are basic moral rights, but are instead of instrumental value, because they have the ability to sometimes protect interests that do rise to the level of moral rights" (2011: 98 –99). Stilz starts from the Universal Declaration of Human Rights, which stipulates such a right to exit. However, she argues that this does not imply that the right to leave should be unqualified: "a legitimate state would be within its rights to tax and regulate those who seek residence or citizenship elsewhere [although] such a state should still permit its citizens to travel and relocate to other countries, though it may enforce their citizenship obligations at the point of exit or during their stay abroad" (2011: 60). Not only that, but she considers that all legitimate states can require individuals to work for a time in their country of origin, or apply taxes on the income that they earn abroad, if these taxes are deemed "essential to sustaining a just distributive scheme for their compatriots" and

are not forcing the emigrant to pursue "an obligation he loathes" (Stilz 2011: 74). On the other hand, Blake considers that a right to leave is based not only on the international legal practice, but also on the fact that, "while we certainly have duties of justice to other members of our society while we are residents within that society, we cannot be thought to have any obligation of justice to continue to be part of that society;" in other words, "what we owe, morally speaking, might be distinct from what we can be morally forced to provide" (2015: 120).

The purpose of this article is not to settle whether leaving one's country should really be classified as a right or not. The discussions surrounding the right to leave, however, serve an important aim: they show that what interests many of the authors who endorse limiting the emigration of skilled citizens in order to mitigate the effects of the brain drain phenomenon is that those skilled workers discharge their duties towards their compatriots. Sometimes, the freedom of emigrating from a country can be defended in order to ensure that the potential emigrants discharge said duties. But then it seems difficult to understand why low skilled citizens should have an unqualified right to leave, especially when the benefits of emigration (such as remittances) are only amplified when the income of the emigrant is higher (and the more skilled she is, the more probable it is that she will have a higher income). Brock does mention that "actually being here is indispensable," like in the example of having skilled personnel conducting a surgery or undertaking other medical acts (2015: 89). Does this argument really hold, however? Would it not be the same if a country could afford to pay a foreign doctor to operate on a patient? If the matter of a lack of resources is brought into consideration, why not require that developed states help more? Perhaps a solution to the negative consequences reached because of the brain drain consists in relying more on international fora and on developed states discharging their own duties than on qualifying the right to exit. Certainly this would seem to be a better option than asking developed states to tighten their immigration policies so that they refuse doctors from underdeveloped countries (a measure endorsed by Ferracioli and De Lora 2015).

Once again, the purpose of this article is neither to elucidate the status of leaving one's country as a moral right or as a weaker claim, nor to decide how to tackle the brain drain phenomenon. The discussions extant in the literature do have to be mentioned, nonetheless, in order to highlight the asymmetry between focusing so much on what is required of some individuals who intend to emigrate and so little on what is due to some individuals who intend to emigrate. If citizens who temporarily emigrate are tied with obligations of justice with the country of origin, then they should also have some entitlements with correlative obligations of their compatriots who chose to remain in a country. How are we to interpret our duties to emigrants? What could be the basis for such duties, besides an attempt to mitigate the asym-

metry? And what exactly are our duties to emigrants? In the remainder of this article I try to offer some provisional answers to these questions. "Emigrants" will be considered all individuals who leave their country of origin for a prolonged period of time, whether they have the intent of returning home or not. For my purposes in this article, "emigrants" can be considered an umbrella-term which can also include temporary migrants. It excludes individuals naturalized in the country of destination. The completion of the naturalization process thus marks a transfer of responsibilities to the country of destination.[5]

4. *Emigrants, duties of justice, and the sufficiency view*

Do we have special obligations to our compatriots (Mason 1997)? Some, like Richard Dagger, consider that we do. Since compatriots take part in "a cooperative enterprise for mutual advantage," they are obligate do to their fair share (Dagger 1985). Others, like Goodin, consider that sometimes we are permitted to treat our countrymen with partiality, whereas at other times those who should benefit from our actions are foreigners. This is because we should not consider "special duties" to be "magnifiers and multipliers;" instead, we should regard such special duties as "merely distributed general duties; merely devices whereby the moral community's general duties get assigned to particular agents," following a model that he deems "the assigned responsibility model." Thus, the so-called duties that stem from sharing citizenship are not intrinsically special, but are general duties discharged for administrative ease in the form of special responsibility. Goodin reaches the conclusion that in an ideal world, where each state would have all it needs to discharge its duties, there would be no requirement of redistribution across borders: each state would just know better how to discharge its general duty through special concern for the ones that happen to live on their territory. Since we are living in a non-ideal world, says Goodin, states cannot claim that they are fulfilling their general duty when they give priority to their citizens (1988: 678–686). Finally, we have cosmopolitan views which state that each human being has equal moral worth and that we have certain responsibilities towards all human beings *qua* human beings (Beitz 2005). However, Beitz' own theory of global justice states that we are "concerned with the moral relations of members of a universal community," but in which "state boundaries have a mere derivative significance" (Beitz 1999).

The answers to the above question thus range from a loud and clear "yes" to a qualified "no." Irrespective of what the answer is, however, we do have some duties to our compatriots—whether these are in virtue of them being our compatriots or in virtue of them being human beings. Alternatively, we could have "localized duties," which are part

[5] I thank an anonymous reviewer for inviting me to better define what categories of emigrants I focus on.

of the more fundamental duty to eradicate poverty, which is nonlocal. This is what Estlund calls the "think globally, act locally model" (2008: 148–150). For the purposes of this discussion, I will hold that we have some obligations of justice to people which are grounded in some features of the individuals themselves. This represents a conception of subject-centered justice (Buchanan 1990). Such a conception is compatible with accepting that under non-ideal circumstances sometimes it is easier to discharge your duties to other members of the same political community, mediated by a well-established institutional framework. As such, although we have duties of justice to all the individuals on this planet, it might be easier to fulfill our duties to our compatriots. How about the emigrants? Would such a model be compatible with stating that we have duties to emigrants, or would they fall under the jurisdiction of the country of destination? In the previous section I stated some reasons why it is difficult to believe under the same non-ideal circumstances that relying on the countries of origin only represents a viable strategy. If we want to maximize the probability that the rights of emigrants are respected, then we ought to consider that countries of origin serve an important function in protecting the emigrants' entitlements.

A subject-centered conception of justice which could account for our obligations to emigrants is sufficientarianism. Different versions of the sufficiency view have been endorsed as global principles of distribution (Miller 2007; Laborde 2010; Kuo 2014), or defended as a solution for selecting refugees (Gerver 2020). What is lacking from the sufficientarian literature, however, is a clarification of what happens to the persons who emigrate from a community. Who is responsible for their well-being? The arguments above emphasized the role played by the country of origin, but it remains to be seen whether other relevant agents have correlative duties, and what these duties actually are. Thus, in a sense, it could be said that the present paper also contributes to the refinement of sufficientarianism as a distributive pattern.

Sufficientarianism holds that social justice is accomplished when each individual has a certain amounts of a preferred currency of justice—be these resources, capabilities, rights or welfare. Sufficientarians hold that the real distributive problem is not that there are inequalities among individuals, but rather that some individuals are in a state of absolute deficiency and cannot lead a decent life (Frankfurt 1987; Crisp 2003). Thus, sufficientarianism is a non-comparative view of justice, holding that we should judge each case separately, and that we can assess an individuals' well-being without relying on interpersonal comparisons with other individuals' well-being levels. Furthermore, a sufficientarian conception considers that, above a certain threshold, our moral concern for other individuals should either dwindle (Shields 2012, 2016) or disappear completely (Casal 2007). In the former case, above the threshold we can apply other principles of justice, but first and foremost we have to ensure that all individuals reach the thresh-

old. Such sufficientarians adopt what Fourie (2017) calls a weak positioning claim, which simply states that we are agnostic regarding the distributive principles that should apply above the superior threshold. In the latter, it is considered that if an individual has enough resources/capabilities/welfare/rights, what happens to her above that threshold of interest ceases to be a question of social justice and thus she should not be the focus of distributive policies anymore. Such sufficientarians embrace a strong positioning claim (Fourie 2017). Irrespective of their stance concerning what happens above the threshold, all sufficientarians accept what Casal (2007) calls the "positive thesis" and Benbaji (2005) labels "the basic intuition," which states that it is bad in itself if someone is badly off and that such persons should be helped with priority. The argument of this paper is unaffected by additional details, so this brief sketch should suffice.

What duties do we have towards the emigrants? My position is that there are no special obligations that we have towards emigrants—they are due the same things as the rest of the citizens. However, the way that we discharge our duties towards them has to be sensitive to their particular situation, i.e. the fact that they reside in another country. Thus, each state will need a specific set of public policies that concern its diaspora. These public policies do not concern any kind of special entitlements that the emigrants might have, but are the consequence of us paying attention to their special circumstances (the most important of which being, as mentioned, the fact that they do not live within the borders of that country anymore). I shall only refer to legitimate state, where legitimacy should be interpreted in a minimal way. Following Buchanan, "an entity has political legitimacy if and only if it is morally justified in wielding political power" (2002: 689). Legitimate here should be understood in such a way as to exclude states that persecute their own citizens, or allow armed groups to persecute its citizens, or are unable to fulfill even the basic needs of their citizens. For example, even if brain drain occurs in such states, the fact that they do not satisfy minimal legitimacy criteria excludes them from consideration, as they're unable or unwilling to fulfill their duties to most of their citizens (be they remaining in the country or emigrating). This corresponds to the view put forward by Brock, who places at the heart of legitimacy the ability of states to respect their own citizens' human rights (2020: 38). For her, full legitimacy (in contrast to "interim legitimacy") requires the simultaneous satisfaction of additional criteria, such as "participation in the cooperative project needed to create or sustain a justified state system" (2020: 56).[6] Bringing legitimacy into

[6] The legitimacy of the state system is too large a topic to be tackled in this article. However, I believe that it is in the spirit of Brock's argument to hold that full legitimacy would also encourage states to become involved in bilateral projects which aim at improving the prospects of emigrants, and it is for this reason that I brought into discussion the difference between interim and full legitimacy. I thank an anonymous reviewer for raising this point.

discussion also serves an important purpose, as it entails that a state that intends to be perceived as legitimate has to do whatever it can reasonably do in order to safeguard the rights of its citizens, whether they are living within their territory or have chosen to temporarily live abroad. An account of legitimacy inspired by Brock's approach can thus explain why the duties of the sending countries do not wither away once someone emigrates to another country, up until the point where those citizens acquire a different citizenship. An important question that remains at the moment unanswered is whether Brock's account of legitimacy and the dual responsibility model that I endorse below would promote dual-nationality universalization, as a practice meant to better protect the rights of individuals. Although a definitive answer to this inquiry will not be offered in this article, I'm inclining towards a provisional "yes," as dual-nationality would multiply the number of agents of protection that could extend aid to individuals in need.

5. *How should we discharge our duties to emigrants?*

There are two potential ways in which states could fulfill their moral obligations towards emigrants. The first consists in establishing good enough conditions at the national level so that an emigrant who so desires could return. This could be founded, for instance, on a supposed right to stay (Oberman 2011). It can involve obligations of developed states to send financial aid to developing countries. The second is based on policies that involve a cooperation between the host country and the country of origin. One way of achieving this is through joint programs involving both countries (Delano 2010). The purpose of such joint programs would be to ensure that emigrants have good enough conditions in the host society, where good enough should be interpreted in a sufficientarian way. In what follows I want to dismiss the first model and defend the second.

5.1 *The encouragement of reverse migration model*

Oberman (2011) sets out to criticize what he calls the choice view, which states that rich states "can either admit poor foreigners as immigrants or they can provide alternative means of assistance, such as development aid, to poor people in their home states" (2011: 253). The reason for doing so is that "to pursue an immigration-based solution to poverty when alternative means of assistance can be implemented without severe cost is to perform an injustice, for it violates the human right people have to stay in their own state" (2011: 253). The strength of his argument is dependent on the extent to which such a right can be justified. Oberman mentions that such a right intends to protect individuals from three distinct sorts of threats: against expulsion, against persecution and against desperate poverty (2011: 257). He seems to follow an interest theory of rights, as he mentions an "interest that people have

in freely being able to make personal decisions without restrictions on their range of options" (2011: 258). Oberman provides three potential justifications for a right to stay: the freedom justification, the cultural membership justification, the territorial attachment (2011: 258). Since individuals have "important personal, cultural and territorial ties that connect them to their home state, they should not be expected to migrate to a foreign state if they are willing to enjoy a level of well-being to which they are entitled" (2011: 265). In order to help individuals realize this interest, rich states ought to assist individuals from poorer countries "in their home state rather than having to migrate abroad" (2011: 264). Furthermore, the stipulation of such a right could even entail the natural duty to establish just institutions, such as a global institution which would "assign which states have responsibility for assisting which poor people rather than [letting those states] try to fulfil their duties in an uncoordinated fashion" (2011: 262). Presumably, the necessity of such an institution would derive from the possibility that some poor societies will not be helped due to collective action problems.

The encouragement of reverse migration model starts from such a right to stay and states that the duties towards emigrants are best fulfilled by creating favorable conditions for their return, so that they would be able to reach a sufficiency threshold at home. A potential question that might arise concerns whether a postulated right to stay is not one applicable mostly to individuals who are living in a given country—that is, not to individuals who have already emigrated.[7] I believe that Oberman's position could be interpreted as being applicable to both categories of individuals. In his words, "a person has a particularly strong interest in being with her family, pursuing her career, practicing her religion, and taking an active part in her community. So more can be expected of governments to enable people to honor their attachments than to enable people to pursue possibilities." Furthermore, this interest that people have in maintaining attachments is one that grounds the already mentioned right to stay in one's own state: "for most people, the options that represent their most important attachments are situated within their own state. Thus, for most people, the human right to stay is a particularly important right, more important than the human right to immigrate" (2015: 246). In the scenario in which a person has already emigrated—and thus probably formed attachments in the host society as well—the right to stay might still be used to promote reverse migration if not sufficient time has passed for those attachments to be meaningful ones.

What is problematic with this model? I believe that it is vulnerable to both feasibility and desirability objections. Regarding the feasibility issues, it seems rather complicated to replicate those favorable condition in the home country. Brock, for instance, believes that "there is more that developing countries can do to make practicing medicine at

[7] I extend my gratitude to an anonymous reviewer for raising this question.

home more attractive […] Often, this is more of a resourcing issues than a lack of will on the part of governments" (2015: 277). In fact, the governments who could create better conditions for their citizens but refuse to do so would not fulfill the criterion of legitimacy mentioned above. Furthermore, given that migrants often choose a *much* richer country as their destination, the costs entailed by such an approach could be tremendous. Furthermore, even if we assumed that all countries were to benefit from a manna from heaven type scenario, there are other consideration that prevent us from endorsing this model. Safran (1991) mentions that not all host countries are willing to take their diasporas back, "as they might unsettle its political, social or economic equilibrium" (1991: 94). Tsuda (2010) mentions for instance that a couple of countries have encouraged ethnic return migration policies which "encourage a country's diasporic descendants born abroad to return home;" nonetheless, such states have mostly embraced "an ethnic conception of the nation state and therefore face stronger ethnonationalist pressures compared to civic nation states" (2010: 619). Such states also have in place "restrictive and exclusionary immigration policies" (Tsuda: 2010: 621, quoting observations made by Brubaker 1992; Castles and Miller 2003). If we consider that a civic conception of the nation-state is the only one compatible with cosmopolitan principles, then we have additional reasons to reject policies that only serve at encouraging ethno-nationalistic tendencies. As Tsuda mentions, "although some type of ethnic protection rationale can be invoked, the underlying justification is based on a sense of state responsibility/obligation toward their diasporic descendants abroad" (2010: 623). Joppke (2005) reaches a similar conclusion, stating that sometimes ethnic preference in immigration selection procedures is based on protection against foreign persecution. To the extent that is true, however, it is difficult to pinpoint exactly why those emigrants have to return to the home state. Furthermore, even if they were not aiming to return for the foreseeable future, this does not mean that their rights should not be protected (until such a moment where the host society would bear increasingly more of the correlative duties that it has to such individuals, which can be identified as the moment when they are naturalized/obtain citizenship in the country of destination). Under these circumstances, perhaps our duties to the emigrants can better be discharged if we resort to a model that does not insist that they have to return to the country of origin. I defend such a model in the next sub-section.

5.2 *The dual responsibility model*

What I hold to be more promising than the encouragement of reverse migration model is discharging our duties as part of a shared project in which, to varying degrees, both the host country and the country of origin play an important part. This represents the essence of what I call the dual responsibility model. Kapur and McHale mention that

"an emigrant diaspora can be a source of trade, investments, remittances, taxes, knowledge, and, eventually, capital-enhanced returnees. A policy approach is to look for ways to strengthen positive connections so that those remaining behind are less adversely affected by the absence of talented compatriots," which could be accomplished by "compensating the poorest countries for losses they bear, and efforts to ensure that emigrants remain as connected as possible—financially and otherwise—to their former homes" (2006: 319). Delano argues that "programs promoting education are based on the idea that the improvement of the lives of the Mexican-origin population in the US should be addresses through collaboration between both countries" (2010: 253). Leblang notices how "home countries have deployed a number of strategies to engage their diasporas and entice them to remit their human physical capital. These range from the creation of government agencies focusing on their citizens abroad to the establishment of hometown associations, which engage expatriates in their new communities" (2016: 76). One of these strategies also involves the adoption of dual citizenship (2016: 80).

The dual responsibility model is based on acknowledging that both countries have a role in ensuring that the emigrant reaches a certain sufficiency threshold. The appropriate way of discharging our duties to emigrants is by carefully drafting policies that ensure that the host country guarantees the emigrants' level of well-being. This could be achieved by joint programs that involve both countries (Delano 2010). Espindola and Jacobo-Suarez (2018) endorse a similar model, in the specific context of the normative obligations to children of immigrants. They mention that "when any two countries are immersed in [circular] migratory flows, they have a shared duty of justice toward the children of returned migrants" (2018: 55). More specifically, they mention that children of immigrant families should "have the skills and knowledge to adapt to their parents' homeland, should they be expelled from the host society or leave voluntarily" (2018: 66). Additionally, they state that this "is a responsibility of all societies involved in a specific migratory flow," which entails "bilingual and bicultural curricula and pedagogy, as well as a system of equivalencies and certifications that allow children of immigrants to transition between both education systems" (2018: 67). The dual responsibility model that I endorse generalizes this consideration: both the host and the origin country owe duties of justice to emigrants. One of the specific ways in which our duties of justice could be discharged is, of course, through educational policies, which might take the form advocated by Espindola and Jacobo-Suarez. However, our duties are not confined to the children of migrants, but to all migrants.

The dual responsibility model has more going on for it. It is in line with Ypi's observation that "the burdens between migrants, citizens of host states and citizens of source states should be distributed fairly"

and that "it is wrong to prioritize past-oriented relations between migrants and their source states at the expense of present ones between migrants and their host states" (2016: 43). It is also in line with the consensus reached by Blake and Broke regarding the fact that "both developing and developed states might work to make the world within which employment decisions are made a less thoroughly unjust one" (Blake 2015: 294). It corresponds to the requirements for legitimacy mentioned by Brock, who argues that "states have obligations to cooperate in a host of trans-border activities, programs, agreements, institutions that aim to secure arrangements capable of effective human rights protection" (2020: 193).

It also takes into account the fact that the developed states have often become developed due to their colonial past or to other historical injustices that they had committed, sometimes against countries that nowadays are struggling financially. It does not let such states off the hook, or simply expects them to pay more to international organizations, but asks them to carefully be involved in remedying past wrongs by accommodating the needs of emigrants from countries which suffered in the past or are still suffering the effects of an unjust institutional framework. Finally, it fits the commitments of cosmopolitans regarding international migration that "each individual person's well-being is of moral concern regardless of where he lives" and that "the place where a person can be best off is not necessarily the place where he was born and has lived" (Kapur and McHale 2006: 305).

Another advantage of the dual responsibility model is that it can distinguish between different categories of emigrants. For instance, high skilled workers do not have to be supported financially—one must rather ensure that their rights are protected, that they can be involved in the host community's life (even if they are not eligible to vote there or to hold an office), that they are not discriminated on the labor market, at the work place or in society in general. On the other hand, low skilled workers should benefit from redistributive policies, besides being guaranteed what has been mentioned above for high skilled workers. The dual-responsibility model can also provide additional normative justifications for several of the recommendations that have been made in the literature on temporary labor migration programs. For instance, Barry and Ferracioli mention that "a[nother] key threat to migrant workers is that employers may take unfair advantage of their vulnerability. They may misrepresent or make fraudulent claims regarding the nature of the work and the benefits the migrants will receive," amounting to "practices […] not consistent with treating temporary migrants as having equal moral status" (2018: S162). In order to reduce the potential impact of such practices, Barry and Ferracioli hold that "problems of this sort can and must be addressed through intelligent institutional design," giving the examples of Canada, which "enforces work agreements in the native language of temporary workers" and of Mauritius,

which "has a special migrant workers' unit, which has both the mandate and resources to investigate abuse against temporary workers [by] making use of translators, hotline for complaints, workplace inspections" (2018: S162). Similar practices could be employed by several other states, and sending countries would have the duty to encourage their adoption. At first glance, this might seem as a way of discharging their duties in a rather indirect way. However, it could also be understood as a way of discharging what Gilabert and Lawford-Smith call "dynamic duties," i.e. "duties that do not focus merely on what can be done in given circumstances, but also on how to change circumstances so that new things can be done" (2012: 812). The concept of dynamic duties can help us understand why the dual-responsibility model is not as limited by feasibility considerations as it might seem. A poorer state, for instance, holds less negotiating power in comparison to the richer and better positioned states to which its citizens might emigrate. However, the dynamic duties notion urges that the sending state engage in diplomatic procedures—not only bilateral, but also multilateral—to the best extent it can. This might entail drawing attention to the international community of potential human rights violations occurring against its citizens, contesting the legitimacy of certain policies and practices that affect its citizens, and so on. All of these help bring about better circumstances for the future safeguarding of its citizens' rights, and a state is not exempt from attempting to do those things just by feasibility considerations.[8] The concept of dynamic duties thus serves as an important guarantee that considerations of justice are not set aside for the sake of feasibility, as the sending states have a no less important duty of expanding the frontiers of what is feasible.

The dual-responsibility model does not ask that sending states directly provide a range of membership-specific rights (Carens 2013)[9]. Instead, it is compatible and endorses several proposals that have already been made in the literature regarding temporary labor migration programs, for instance, be they ways of ensuring that the period of time that temporary workers is taken into account for their pensions (Carens 2008: 247), guaranteeing freedom of movement (Lister 2014: 114), or even precluding the possibility that they pay rates for temporary workers fall under the threshold of protecting their basic rights (Barry and Ferracioli 2018: S162). A comprehensive list of such measures is outside the scope of this article, and it would be impossible to offer a one-fits-all checklist. The dual-responsibility model refers first and foremost to the idea that Baubock and Ruhs summarize as conceiving temporary migrants as citizenship stakeholders, who "must be included as local citizens in destination countries and as national citizens in

[8] I thank an anonymous reviewer for addressing the question of whether feasibility considerations might not be used by sending states to avoid discharging their duties towards their emigrants.

[9] I thank an anonymous reviewer for bringing membership-specific rights up and inquiring how the model relates to them.

their countries of origin," as both countries "have duties to help them realize their life projects and to involve them in shaping the future of these societies" (2022: 531—2). This principle not only applies to temporary migrants, but also to other categories of emigrants, as defined above (and provides further grounds for embracing a universalization of dual-citizenship). After all, it is not only temporary migrants who face the challenge mentioned by Baubock and Ruhs of "find[ing] that their absentee status diminishes their political clout or that home country governments use them only instrumentally for their own economic or political purposes" (2020: 541). A sufficientarian conception of justice would help individuals realize their life plans no matter where they are situated, and—depending on what currency of justice we employ—would also have something to say about the political standing of emigrants. For instance, Nussbaum's capabilities list includes control over one's environment, which entails "being able to participate effectively in political choices that govern one's life; having the right of political participation, protections of free speech and association" (1997: 288). The dual-responsibility model cannot offer a definitive answer to the question of whether this capability would imply that emigrants have voting rights in the countries of destination, but it would probably push for sending states to advocate the political inclusion of emigrants at least at a local level. Once again, this correspond to Baubock and Ruhs' position that it is important to "take sufficient account of the interests and fair representation of migrants" (2020: 546), which also implies bestowing upon them various forms of local citizenship, which "provides them with additional protection—symbolically through a status of temporary membership and practically through the attention that candidates have to pay to their interests of potential voters" (2020: 543). The dual-responsibility model embraces the idea that the passage of time has normative implications, contributing towards long-term emigrants having "located life plans" in their countries of destination (Stilz 2013). Thus, it would urge sending states that they push for the inclusion in what form of another of their emigrants in the sending state's demos the more time has passed since they have lived there.

Such a model is also not incompatible with taxing the high skilled workers as proponents of limiting the right to exit hold; it only holds that their entitlements are not ignored, and that their country of origin discharges its duties towards them. Furthermore, such a model could even lead to redistributions from high-skilled emigrants to low-skilled emigrants, up to a certain threshold of sufficiency.

Thus, unlike the return of reverse migration model, the dual responsibility model better fulfills the desirability and feasibility criteria. It is desirable for several reasons, two important ones that also distinguish it from the other model being that it takes into account the historical injustices caused by the countries which today represent main destinations for emigrants and that it embraces the aforemen-

tioned cosmopolitan position that "the place where a person can be best off is not necessarily the place where he was born and has lived" (Kapur and McHale 2006: 305). It is also feasible because it is based on already-existing examples of cooperation between host and destination countries which have functioned well. The dual responsibility model provides a normative justification for universalizing such practices. It is also bound to be acceptable by large parts of the destination countries' citizens as it highlights the fact that the sending country also has a role to play in helping its emigrants reach a threshold of sufficiency (thus making it more publicly acceptable than a potential third model which would hold that a state is responsible for all the residents on its territory). The sufficientarian pattern itself has an important function in ensuring the feasibility of this model, as it is less demanding than alternative conceptions (such as an egalitarian one). The dual responsibility model thus also fills a previously existing gap within sufficientarianism regarding what happens to citizens who emigrate to another society. Finally, the model that I endorsed aims to reduce the asymmetry between the postulated duties of emigrants and their entitlements *qua* emigrants by emphasizing what emigrants are owed—to reach a sufficient level of well-being, with both the sending and the destination countries playing a part in helping them reach the threshold.

6. *Conclusions*

In this article I endorsed a particular conception of the duties that we have towards emigrants, the dual responsibility model. This holds that the best way to ensure that the emigrants have a sufficient level of well-being (measured in whatever we agree to be the most appropriate currency of justice) is by establishing programs together with the country of destination that are aimed at helping emigrants integrate in the host society, at ensuring that their rights are protected, at preventing discrimination at the workplace, in educational programs, and elsewhere. I compared and defended this model against an alternative one, that I called the encouragement of reverse migration model, which is based on a supposed right to stay. My main concern was with defending the dual responsibility model—the task that lies ahead is to develop specific policy proposals that could help implement this model. Whatever form these policies do end up taking, however, it is my contention that they will contribute to a more just world.

References

Barry, C. and Ferracioli, L. 2018. "On the Rights of Temporary Migrants." *The Journal of Legal Studies* 47 (S1): S149–S168.

Baubock, R. and Ruhs, M. 2022. "The Elusive Triple Win: Addressing Temporary Migration Dilemmas Through Fair Representation." *Migration Studies* 10 (3): 528–552.

Beine, M., Docquier, F. and Oden-Defoort, C. 2011. "A Panel Data Analysis of the Brain Gain." *World Development* 39 (4): 523–532.

Beitz, C. 1999. *Political Theory and International Relations*. Second Edition. New Jersey: Princeton University Press.

———. 2005. "Cosmopolitanism and Global Justice." *The Journal of Ethics* 9 (1–2): 11–27.

Benbaji, Y. 2005. "The Doctrine of Sufficiency: A Defence." *Utilitas* 17 (3): 310–332.

Blake, M. and Brock, G. 2015. *Debating Brain Drain: May Governments Restrict Emigration?* New York: Oxford University Press.

Brassington, I. 2012. "What's Wrong with the Brain Drain?" *Developing World Bioethics* 12 (3): 113–120.

Brock, G. 2020. *Justice for People on the Move: Migration in Challenging Times*. New York: Cambridge University Press.

Brubaker, R. 1992. *Citizenship and Nationhood in France and Germany*. Cambridge: Harvard University Press.

Buchanan, A. 1990. "Justice as Reciprocity Versus Subject-centered Justice." *Philosophy and Public Affairs* 19 (3): 227–252.

———. 2002. "Political Legitimacy and Democracy." *Ethics* 112 (4): 689–719.

Carens, J. 1992. "Migration and Morality: A Liberal Egalitarian Perspective." In B. Barry and R. Goodin (eds.). *Free Movement: Ethical Issues in the Transnational Migration of People and Money*. Pennsylvania: Pennsylvania State University Press, 25–47.

———. 2008. "Live-in Domestics, Seasonal Workers, and Others Hard to Locate on the Map of Democracy." *The Journal of Political Philosophy* 16 (4): 419–445.

———. 2013. *The Ethics of Immigration*. Oxford: Oxford University Press.

Casal, P. 2007. "Why Sufficiency is Not Enough." *Ethics* 117 (2): 296–326.

Castles, S. and Miller, M. 2003. *The Age of Migration: International Population Movements in the Modern World*. London: Macmillan Press.

Crisp, R. 2003. "Equality, Priority and Compassion." *Ethics* 113 (4): 745–763.

Dagger, R. 1985. "Rights, Boundaries and the Bond of Community: A Qualified Defense of Moral Parochialism." *American Political Science Review* 79 (2): 436–447.

Delano, A. 2010. "Immigrant Integration Versus Transnational Ties? The Role of the Sending Ties." *Social Research* 77 (1): 237–268.

Dumitru, A. 2017. "On the Moral Irrelevance of a Global Basic Structure: Prospects for a Satisficing Sufficientarian Theory of Global Justice." *Croatian Journal of Philosophy* 17 (50): 233–264.

Espindola, J. and Jacobo-Suarez, M. 2018. "The Ethics of Return Migration and Education: Transnational Duties in Migratory Processes." *Journal of Global Ethics* 14 (1): 54–70.

Estlund, D. 2008. *Democratic Authority: A Philosophical Framework*. Princeton: Princeton University Press.

Ferracioli, L. and De Lora, P. 2015. "*Primum Nocere:* Medical Brain Drain and the Duty to Stay." *Journal of Medicine and Philosophy* 40 (5): 601–619.

Fourie, C. 2017. "The Sufficiency View: A Primer." In C. Fourie and A. Rid (eds.). *What is Enough? Sufficiency, Justice and Health*. Oxford: Oxford University Press, 11–29.

Frankfurt, H. 1987. "Equality as a Moral Ideal." *Ethics* 98 (1): 21–43.

Gerver, M. 2020. "Sufficiency, Priority, and Selecting Refugees." *Journal of Applied Philosophy* 37 (5): 713–730.

Gilabert, P. and Lawford-Smith, H. 2012. "Political Feasibility: A Conceptual Exploration." *Political Studies* 60 (4): 809–825.

Glytsos, N. 2010. "Theoretical Considerations and Empirical Evidence on Brain Drain Grounding the Review of Albania's and Bulgaria's experience." *International Migration* 48 (3): 107–130.

Goodin, R. 1988. "What is So Special about Our Fellow Countrymen?" *Ethics* 98 (4): 663–686.

Hobden, C. 2017. "Taking Up the Slack: The Duties of Source State Citizens in the Brain Drain Crisis." *South African Journal of Philosophy* 36 (1): 33–34.

Joppke, C. 2005. *Selecting by Origin: Ethnic Migration in the Liberal State.* Cambridge: Harvard University Press.

Kaplan, D. and Hoppli, T. 2017. "The South African Brain Drain: An Empirical Assessment." *Development Southern Africa* 34 (5): 497–514.

Kapur, D. and McHale, J. 2006. "Should a Cosmopolitan Worry About the Brain Drain?" *Ethics and International Affairs* 20 (3): 305–320.

Kukathas, C. 2005. "The Case for Open Immigration." In A. Cohen and C. H. Wellman (eds.). *Contemporary Debates in Applied Ethics*. Oxford: Blackwell University Press, 208–220.

Kuo, Y. 2014. "Global Sufficientarianism Reconsidered." *The Taiwanese Political Science Review* 18 (1): 181–225.

Kymlicka, W. 2001. "Territorial Boundaries: A Liberal Egalitarian Perspective." In D. Miller and S. H. Hashmi (eds.). *Boundaries and Justice: Diverse Ethical Perspectives*. Princeton: Princeton University Press, 249–276.

Laborde, C. 2010. "Republicanism and Global Justice." *European Journal of Political Theory* 9 (1): 48–69.

Leblang, D. 2016. "Harnessing the Diaspora: Dual Citizenship, Migrant Return Remittances." *Comparative Political Studies* 50 (1): 75–101.

Lenard, P. 2022. "Restricting Emigration for Their Protection? Exit Controls and the Protection of (Women) Migrant Workers." *Migration Studies* 10 (3): 510–527.

Lister, M. 2014. "Justice and Temporary Labor Migration." *Georgetown Immigration Law Journal* 29 (1): 95–123.

Mason, A. 1997. "Special Obligations to Compatriots." *Ethics* 107 (3): 427–447.

Miller, D. 2005. "Immigration: The Case for Limits." In A. Cohen and C. H. Wellman (eds.). *Contemporary Debates in Applied Ethics.* Oxford: Blackwell University Press, 193–220.

———. 2007. *National Responsibility and Global Justice.* Oxford: Oxford University Press.

Niimi, Y., Ozden, C. and Schiff, M. 2010. "Remittances and the Brain Drain: Skilled Migrants Do Remit Less." *Annals of Economics and Statistics* 97–98: 123–141.

North, D. 1991. "Institutions." *Journal of Economic Perspectives* 5 (1): 97–112.

Nussbaum, M. 1997. "Capabilities and Human Rights." *Fordham Law Review* 66 (2): 273–300.

———. 2005. "Beyond the Social Contract: Capabilities and Global Justice." In G. Broke and H. Brighouse (eds.). *Political Philosophy of Cosmopolitanism*. New York: Cambridge University Press, 196–218.

Oberman, K. 2011. "Immigration, Global Poverty and the Right to Stay." *Political Studies* 59 (2): 253–268.

———. 2013. "Can Brain Drain Justify Immigration Restrictions?" *Ethics* 123 (3): 427–455.

———. 2015. "Poverty and Immigration Policy." *American Political Science Review* 109 (2): 239–251.

Okeja, U. 2017. "Reverse Migration, Brain Drain and Global Justice." *South African Journal of Philosophy* 36 (1): 133–143.

Owen, D. 2016. "Compulsory Public Service and the Right to Exit." *Moral Philosophy and Politics* 3 (1): 55–65.

Pevnick, R. 2009. "Social Trust and the Ethics of Immigration Policy." *Journal of Political Philosophy* 17 (2): 146–167.

———. 2011. *Immigration and the Constraints of Justice: Between Open Borders and Absolute Sovereignty*. New York: Cambridge University Press.

———. 2016. "Brain Drain and Compulsory Service Programs." *Ethics and Global Politics* 9 (1): 1–7.

Rawls, J. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Safran, W. 1991. "Diasporas in Modern Societies: Myths of Homeland and Return." *Diaspora: A Journal of Transnational Studies* 1 (1): 83–99.

Shields, L. 2012. "The Prospects for Sufficientarianism." *Utilitas* 24 (1): 101–117.

———. 2016. *Just Enough: Sufficiency as a Demand of Justice*. Edinburgh: Edinburgh University Press.

Stemplowska, Z. 2008. "What's Ideal About Ideal Theory?" *Social Theory and Practice* 34 (3): 319–340.

Stilz, A. 2011. "Is There an Unqualified Right to Leave?" In S. Fine and L. Ypi (eds.). *Migration in Political Theory: The Ethics of Movement and Membership*. Oxford: Oxford University Press, 57–79.

———. 2013. "Occupancy Rights and the Wrong of Removal." *Philosophy and Public Affairs* 41 (3): 324–356.

Swift, A. 2008. "The Value of Philosophy in Non-ideal Circumstances." *Social Theory and Practice* 34 (3): 363–387.

Tsuda, T. G. 2010. "Ethnic Return Migration and the Nation-state: Encouraging the Diaspora to Return *Home*." *Nations and Nationalism* 16 (4): 616–636.

Wellman, C. H. 2008. "Immigration and Freedom of Association." *Ethics* 119 (1): 109–141.

Wellman, C. H. and Cole, P. 2011. *Debating the Ethics of Immigration: Is There a Right to Exclude?* New York: Oxford University Press.

Ypi, L. 2016. "Sharing the Burdens of The Brain Drain." *Moral Philosophy and Politics* 3 (1): 37–43.

Yuksekdag, Y. 2018. "Health Without Care? Vulnerability, Medical Brain Drain, and Health Worker Responsibilities in Underserved Contexts." *Health Care Analysis* 26 (1): 17–32.

———. 2019. "The Right to Exit and Skilled Labour Emigration: Ethical Considerations for Compulsory Health Service Programs." *Developing World Bioethics* 19 (3): 169–179.

Book Reviews

Owen Flanagan, How to do Things with Emotions: The Morality of Anger and Shame across Cultures. *Princeton and Oxford: Princeton University Press, 2021, ix + 309 pp.*

In his recent book, Owen Flanagan discusses the so-called disciplinary emotions: anger, shame, and guilt. These emotions are called disciplinary due to their punitive character. However, they are not only punitive; they have a higher goal. Flanagan describes them as emotions that "are more sticks than carrots, [and] the goal of using them must be to reap the rewards of a shared, harmonious, mutually beneficial common life" (9).

So, the question is, how exactly emotions conceived as "bad emotions" make us do good things? Flanagan provides an answer to that question. In short, these emotions have a bad reputation that needs to be rebuilt. He proposes working towards reconstruction of emotions as well as rehabilitation of their reputation. Flanagan's method for the reconstruction of emotions is set from the perspective of cultural psychology, anthropology and cross-cultural philosophy. It aims at using "the evidence of variation as an invitation to think about how we do these emotions, to think of how we do these emotions as something we are in charge of and that we can change if we have reason to" (42). The overall idea based on such method is to critically think about "how we do emotions, and how we might do them better" (42).

The book is organized in three parts and eight chapters. The first part, "Anger," is divided into three chapters ("Anger and Morals," "Anger across Cultures," "Anger and Flourishing"). As Flanagan thinks, anger mistakenly has a good reputation because we are taught to think that daily display of a minimal amount of anger is good, healthy, permissible, and sometimes necessary since it shows that we care about something. There is a problem with the moral categorization of anger due to the fact that many people, as well as many moral philosophers, think that some forms of anger are virtuous (see 49). Hence, anger also needs to be rebuilt and rehabilitated. Rehabilitation consists of teaching that anger is bad, yet not every form of anger is a vice. Anger should not be a part of a healthy moral community, although there are some varieties of anger (e.g., anger against structural sexism) that help to increase awareness of things that we need to overcome. Forms of anger that we should get rid of are payback and pain-passing anger. Both are common and similar insofar as they aim to hurt and humiliate others.

Specifically, payback anger, which includes revenge, is intentionally cruel; it is set on the intention "to cause another physical or mental pain and suffering, and/or status harm, typically because they caused me pain" (67). Pain-passing anger is a kind of anger where one intends to cause pain to another because one is in pain, but that pain is not caused by the person who is the subject of inflicted pain at the moment (see 67). Pain-passing anger is "thoughtless and self-indulgent" (68). Both payback and pain-passing anger "hurt others for no greater good or higher purpose, such as improving the other, balancing a relationship, or changing harmful practices or institutions. The arguments against them apply to the other kinds of anger insofar as they embed, enact, and encourage payback or pain-passing" (68).

Thus, the rehabilitation consists of rethinking what we are being taught. For instance, Flanagan reassesses contemporary American attitude (more specifically, the attitude of the American Psychological Association and the dominant American view among psychologists and psychiatrists) towards anger that considers such emotion as a healthy and normal human emotion which needs to be expressed, externalized or released "otherwise there will be addiction, eating disorders, skin disorders, migraines, divorce, and general mayhem" (56). He challenges such an attitude: "Except when one examines the evidence, it is all bullshit in the technical, philosophical sense [referring to Harry Frankfurt's *On Bullshit*]. The message is designed to persuade, but with complete disregard for the truth and evidence" (56).

The truth that Flanagan has in mind includes, on the one hand, accepting that "[t]he world I live in partakes in an orgy of anger but doesn't see or acknowledge it" (57), meaning that expressing or releasing anger produces more anger (which he is trying to emphasize but which the world around him, by getting more and more angrier, does not realize). On the other hand, we need to include evidence about other cultures that may help us examine how *others* do anger (with the possibility to learn something from them and do *our* emotions better).[1]

According to the evidence, the Japanese—as Flanagan informs us—leave the room when they are angry, and the Ifaluk people stop eating. Americans associate anger with yelling and hitting, and Belgians with withdrawal and ignoring (80). Regarding the Ifaluk people, it is interesting to point out that they disapprove of most kinds of anger, especially about personal hurt feelings or personal misfortune. The only kind of anger considered justified among them is "primarily in response to selfishness and stinginess" (83). Among Utku Inuits, anger towards their sled dogs is justified, although all forms of interpersonal anger are considered vicious (see 82).

The Minangkabau, a numerous ethnic group of people in Indonesia, believe that anger is a vice. It is harmful to socialization because it goes against respecting others. Admittedly, shaming children for the Minangkabau is useful and beneficial. Respecting others is an important value of

[1] "Our" or "We" refers to contemporary Americans and/or some groups of people who are connected regionally, politically, socioeconomically, religiously, educationally, ethically, by age etc. and/or the WEIRD cultures (Western, Educated, Industrialized, Rich, and Democratic) since "most psychology is based on experiments with North American college students, and this is one of the most unrepresentative populations in history" (110). North-American students are WEIRD.

the Minangkabau people, and shame cultivated at an early age ensures this common and social emotion. The Bara, an ethnic group in Madagascar, believes that anger is necessary for teaching children the norms of good behaviour. Their life credo is to live well. Moreover, if that good life is somehow disturbed, anger is necessary. Anger, in the mentioned monocultures, is a social and moral emotion—it has a moral feature, for it represents what one ought and ought not to do. In that sense, "[t]hese emotions [namely, anger and shame] are used to inform others that they are out of normative conformity or, at minimum, that they are doing something we don't like or approve of" (7).

Both Minangkabau and Bara people agree that anger is bad in interpersonal adult relationships, but Bara people consider it useful in upbringing and socialization. Both the Utku Inuits and the aforementioned ethnic groups are examples of monocultures in which it is possible to live in an "unambiguous" collective in which a norm violation is experienced as a collective violation of the norm. In a multicultural society, the matter of emotions is not that simple.

Since doing emotions in a multicultural society and in general is not that simple, Flanagan calls attention to several things about anger that are worth mentioning and that are emphasized throughout the book:

(1) there is no universal agreement on what anger is (see 126) since it is "a cultural matter, the result of cultural learning, including, especially, how elders model it for the young" (xiii);
(2) "[w]hat is universal is that anger is unpleasant; it has negative valence for the person who experiences it, and it is unpleasant for the recipient, producing pain, fear, anxiety, and sadness" (126);
(3) "[t]he best world is one in which when anger is necessary, it is motivated by love and compassion for the person or community of persons that one is angry at or with and does not aim at revenge or harm but only to make the person or persons, at the limit the world, better. This is loving anger" (59);
(4) "[a]nger and shame are generally even more implicated in normative life than emotions like sadness, fear, and happiness" (34);
(5) what we could do is examine the culturally scripted emotions and borrow emotional patterns in the same way we borrow "a cuisine or fashion or practice from an alien tradition because they like it or it looks good on them or it improves mental or moral health" (120).

The conclusion regarding anger is that as a moral emotion, it is, like all emotions, culturally scripted. By getting informed on different ways of living a human life, we can rethink how, when and why we get angry and think of ways to improve that.

The second part of the book, titled "Shame," is divided into four chapters ("Generic Shame," "The Science of Shame," "Shame across Cultures," and "The Mature Sense of Shame"). According to Flanagan, we lack shame when we ignore or violate values—what is good, true and beautiful. He puts it as follows: "Shamelessness is common, and it reflects a situation in which many values are weakly held, and in which norms suited for a common life that aims at the common good yield to precepts for winning friends and influencing people, gaming, and getting ahead. In a world in which it is every

ego for itself, it is better to seem honest than to be honest, and acquisitiveness of the "greed is good" sort—once a deadly sin—has various honorific disguises" (xi).

That is why Flanagan proposes to upgrade shame to a level of mature sense of shame. He cheers for the positive acceptance of shame or good shame "as an ideal protector of deep value commitments [...] [as] an emotional instrument that can be used to teach and protect values" (134). As Flanagan sees it, shame is an emotion that "starts out feeling bad but is eventually autonomously endorsed as a positive self-monitoring emotion" (134). So, the crucial part of upgrading shame is considering it as a shield for values. Currently, there are two dogmas about shame:

(1) "Shame is an essentially social emotion, ultimately a response to the disapproving eyes of others" (181);
(2) "Shame is directly morally bad" (181).

Flanagan discusses both dogmas. He believes shame is a complex social emotion whose moral categorization, like anger, depends on how each culture defines it. Despite that, there are two very widespread dogmas about shame.

According to the first dogma, shame is a social emotion arising from disapproval or non-compliance with norms. Flanagan does not deny this but adds that it does not necessarily have to be an emotion that entails the gaze of others.

According to the second dogma, shame is directly a morally bad emotion because we associate it with "one kind of bad feeling" (134) that an individual has when another judges him for violating a norm. Furthermore, shame occurs in combination with feelings of embarrassment, fear, anxiety and sadness (some consider that this mixture of emotions is shame itself), and it is a "social emotion" (135), not an individual one, which means that the individual does not, in principle, feel it self-initiated. The initiator of shame is always the other. As a collective emotion, shame opens up the possibility of exclusion from that collective. In this sense, shame is a painful and humiliating emotion; shame is public, comes from outside and is not an emotion that an individual chooses independently.

Flanagan sees shame in another manner. The idea of a mature sense of shame or good shame is that such an emotion is autonomously endorsed and serves as a positive self-monitoring emotion. Thus, shame results from setting boundaries one does not want to cross because otherwise, he would do something wrong. This does not mean that with this kind of shame, we would have a perfect or sinless individual. It only means that the individual who endorses shame can relate to the sociomoral order and is open to feedback from others. In other words, shame is related to social relationships but also to personal values and ideals, so in that sense, it is based on personal choice, not on criticism from others.

Shame is thus separated from humiliation and embarrassment, and it is far from a bad emotion. It is elevated from a bad and unnecessary emotion that depends on another's judgment (as an emotion that, e.g., "attacks a person") to an emotion that protects values.

On the other hand, guilt is an emotion that is conceptualized throughout the book in the same way shame is. Flanagan considers shame and guilt "different to some extent," although he "often use[s] the terms interchange-

ably" (192). The extent he has in mind is that shame, in contrast to guilt, is focused on character traits, more precisely on weaknesses or shortcomings of an individual, while, for example, guilt is linked to an action or an act.

The third part of the book is "Conclusion" and has one chapter, "Emotions for Multicultures." In that part, Flanagan summarizes what he wanted to achieve with the book, namely, to offer assistance for moral imagination about various moral possibilities and, thus, a mature attitude towards emotions.

In a gist, Flanagan's idea is simple: we need to do emotions better because we can be better at feeling shame and anger, as well as many other emotions. There are possibilities for changing how we do emotions (5) and by recognizing them, we can experience emotions differently and live a better life. The basis of this is the understanding that emotions are the things we do (xiv). Emotions are under our control. Moral or disciplinary emotions are designed to produce bad feelings because the idea is to stop doing what we should not—that is their intention. The ultimate idea of rehabilitation regarding moral emotions is to achieve self-regulation or self-observation in terms of norms, values and ideals.

This book is a work of philosophical art, and this review cannot do justice to how engaging and valuable it is. It was so refreshing to read about emotion from a philosophical point of view and, at the same time, get such a dense and insightful look on moral emotions. Reading an author who can deliver a fascinating philosophical book written in plain language is always a privilege.*

ANA GRGIĆ
Institute of Philosophy, Zagreb, Croatia

Frauke Albersmeier, The Concept of Moral Progress. *Berlin: De Gruyter, 2022, 248 pp.*

The phenomenon of moral progress has been attracting increasing interest in philosophy in recent years. Ever since the publication of Peter Singer's book *Expanding Circle* in 1981, numerous authors have attempted to grasp the concept of moral progress and to answer the question of whether there is indeed progress in morality and how we should understand it. It is not surprising that, like many other philosophical concepts, there is not much consensus on the concept of moral progress. What is specific to this concept is that the attempt to understand it delves into the very heart of the question of how to understand morality itself. In order to arrive at a plausible concept of moral progress, it seems that we must address, if not resolve, a whole range of contentious questions that accompany ethical thinking. Frauke Albersmeier has embarked on such an attempt in her book *The Concept of Moral Progress*.

The book is a revised doctoral thesis the author defended in 2020 at the University of Düsseldorf. It consists of five main chapters in which the

* This review is an output of the project "Moral Progress: Individual and Collective" supported by the Croatian Science Foundation (Grant No. IP-2022-10-5341).

author provides an explication of the concept of moral progress. In the first chapter, which is dedicated to methodological explanations of the procedures she will apply in the rest of the book, the author rejects the method of conceptual analysis of moral progress. The conceptual analysis aims to identify a set of necessary and sufficient conditions for the application of a particular concept, with success being achieved if the proposed definition of the concept aligns with our intuitions about individual cases to which the concept should apply. However, as with the analysis of other concepts, our intuitions about what changes should be considered "clear instances" of moral progress vary greatly from person to person. By analyzing concepts, we can gain useful clarifications of the concept itself, but mere conceptual analysis will not take us far in understanding the concept of moral progress (12). As a better approach to exploring the concept of moral progress, the author chooses the method of explication, characterized by Carnap as "the process of replacing an inexact (pretheoretical) concept (or term) with a more exact one for the purposes of scientific theory-building" (14).

Explaining the concept of moral progress and establishing its meaning is the first step in its explication. Albersmeier undertakes it in the second chapter titled 'Moral Progress: Conceptual Commitments, Pragmatic Expectations.' Breaking down the various meanings of the term progress, the author categorizes moral progress as a form of improvement whereby it is "a *process* of *change* undergone by something that persists through this change and it is *directed*" (28, emphasis in original). Explaining the "moral" component of moral progress is much more challenging. Various ethical theories explain morality in very different ways, emphasizing different essential aspects of the phenomenon of morality. In an attempt to offer a portrayal of morality that would enable the explication of the phenomenon of moral progress, Albersmeier starts from the understanding of morality as a practice of making judgments. In our moral discourse, moral judgments seem to express certain beliefs and can be true or false (32). Setting aside some controversial characteristics of morality, such as categoricity, universality, intersubjectivity, or impartiality, the author singles out the connection to actions as another key characteristic of moral judgments. Additional insights into the phenomenon of morality are gained when we observe it in the light of moral agents, i.e., individuals who are sensitive to moral reasons even though they may not always act in accordance with them, and the recognition that the capacity for moral progress is often considered a condition for morality (37). In order for the explication of the concept of moral progress to be as widely acceptable as possible, the mentioned characteristics are selected to clarify the phenomenon of morality as precisely as possible without (excessively) relying on specific normative and metaethical theories.

The exploration of how normative and metaethical theories influence the concept of moral progress is presented in the third chapter titled "Ethics and the Idea of Moral Progress." In this section, Albersmeier compares the attempt to define moral progress to the challenge of addressing moral problems in the domain of applied ethics, where solutions must be found without relying too heavily on normative theories. Assuming such a pluralism of normative and metaethical viewpoints, the search is for a solution that

would explain the phenomenon of moral progress in a manner acceptable to different theories (45-46). Each normative ethical theory naturally has its own vision of what moral progress should be, but it is understandable that the targeted explication of the concept of moral progress cannot benefit from such "narrowly" defined understandings. The reason for addressing different normative viewpoints is that each of them emphasizes different elements in our understanding of moral progress. An adequate concept of moral progress can benefit from considering various theoretical perspectives on the discussed phenomenon.

Although consequentialism, due to its emphasis on inclusivity (e.g., Singer's "expanding circle"), is considered an ethical theory closely associated with the idea of moral progress, the author believes that the idea of inclusivity, despite being widely accepted, cannot be used in explicating the concept of moral progress due to its normative charge. Namely, one can imagine theories that see moral progress in the exclusivity of taking into moral consideration. Additionally, the problem lies in the consequentialist focus on outcomes, which, in one sense, sidelines moral agents in the process of moral judgment. In global consequentialism, what is morally valued is not only actions, rules, and motives but also everything else that influences the outcomes. However, it seems problematic to consider an improvement in the state of the world that is not linked to an improvement in the moral practices of agents as an example of moral progress. Acknowledging the fact that improvements in the state of affairs are an integral part of moral progress, the author concludes that such improvements require the constant involvement of moral agents (57). Many authors writing about moral progress believe that people are somehow capable of improving their practices. Ethical reflections inspired by Kant warn us, however, that this does not necessarily mean it is moral progress. Starting from the premise that we can consider morally valuable only those actions done from right motives, philosophers inspired by Kantian ethics believe that an increase in the number of morally good actions and the resulting morally good effects says nothing about their moral worth. This is the main lesson from this tradition of ethical reflection that the author adopts for her explication of the concept of moral progress (62-63). When we talk about moral progress, we are not only discussing the state of affairs and the type of actions but also the moral agents themselves. We expect them to be morally better. This is precisely the area where virtue ethics has something to say. Like in the case of other normative theories, the author points out why appealing to some of the substantive ethical doctrines of this ethical theory would hinder a widely acceptable concept of moral progress. However, as a significant contribution from this theory, she adopts the perspective that what matters for a moral agent is the disposition to act well (66). From the domain of political philosophy, inspired by Mill's thinking, Albersmeier draws a warning that with the proliferation of moral beliefs comes the threat of loss of ethical understanding and consequently the threat of moral regression (78). Nevertheless, metaethics is the key challenge for any theory of moral progress. Since moral progress is often portrayed as a process of approaching moral truth, it seems as if moral progress presupposes the truth of moral realism, the claims that there exists an order of moral facts independent of us. In this segment of her research, the author demon-

strates that this connection between moral progress and moral realism is not necessary, given the weakness of the arguments put forward in favor of the claim that moral progress proves the truth of moral realism (transcendental argument from progress and abductive argument from progress).

After positioning the concept of moral progress in relation to normative and meta-ethical theories, the fourth chapter, "The Phenomenon of Moral Progress," presents "a proposal of how we should come to think of different types of moral progress, based on considerations that go beyond our initial conceptual intuitions" (99). When discussing dimensions of moral progress, it is common to talk about differences between individual versus collective and local versus global progress. Albersmeier, in her discussion, does not exclude the possibility of collective moral progress but considers that the clearest examples of moral progress can still be found at the individual level, with progress at the collective level being explained by progress at the individual level. Regarding the temporal dimension of moral progress, the author believes that moral progress does not necessarily have to represent an epochal and permanent change but still needs to demonstrate a certain durability that does not dissipate as soon as it appears. Moreover, it can be said that moral progress does not have to be global but may occur only in one domain of morality. Therefore, special attention is devoted to the possibility of moral progress in our beliefs (in theory) and moral progress in our practices.

Determining whether moral progress requires progress in both of these domains proves to be a key task of this chapter. If someone has achieved moral progress in theory (Albersmeier in this case uses the term ethical progress), it means that they have advanced their beliefs, desires, or judgments. Ethical progress does not necessarily have to be accompanied by progress in our moral behavior. Of course, such a situation is deeply problematic, and we could not consider it an example of moral progress. Albersmeier argues that moral progress must manifest itself in the practical domain of morality. By using examples in which a person changes their beliefs and/or behaviors in different circumstances, the author demonstrates that we can indeed speak of moral progress even in situations where there is no outwardly observable action in line with improved moral beliefs. What is crucial for us to consider it as a case of moral progress is that the person changes their dispositions for acting in a moral way. Changing dispositions is moral progress because, under favorable circumstances, it gives us confidence that the person will act in a morally correct manner. Albersmeier refers to this type of moral progress as dispositional moral progress. For the author, this is a genuine type of moral progress because its practical relevance lies in the fact that "theoretical change is required to impact moral performance as the occasion for the relevant type of action arises" (174).

In contrast to dispositional, real moral progress is "the improvement in the moral agent's moral performance over a certain period of time" (146). It is worth noting that real moral progress cannot happen "by fluke." For the progress to be considered real moral progress, the author believes there must be a moral agent involved who possesses at least a minimal moral consciousness that their actions are morally correct. Although she argues that there is no moral progress without ethical progress, the author acknowl-

edges that in some cases, it is difficult to distinguish between examples where behavioral change occurred with moral awareness and those where it did not (but happened for reasons unrelated to morality). However, she believes it is important to establish this conceptual distinction because unlike cases of moral progress, these other cases resemble "morally desirable non-agential changes" (161). In the case of dispositional moral progress, however, it is still considered moral progress because dispositions do not entirely fit into the standard division into the theoretical and practical parts of morality.

Considering that Albersmeier starts from the premise that it concerns individual moral progress, changes in moral behavior, even when they reach the point where they can be qualified as real moral progress, do not necessarily reflect broader societal moral character changes, which are usually considered examples of moral progress. The term encompassing this dimension of moral progress phenomenon is *impactful moral progress*. It is "actual moral progress that brings about an improvement in states of affairs" (175). Summarizing her explication of the concept, the author concludes the chapter with the assertion that "*moral progress is (a) durable change for the better in moral performance, (b) on given occasions, (c) that is sufficiently suited to effect change for the better in states of affairs*" (177, emphasis in the original).

In the final chapter entitled "Moral Progress and Moral Motivation: Improvement as a Fetish?" the author explores whether moral progress can motivate our actions. There seems to be something suspect in the idea that someone would act based on an abstract ideal simply because it is the right thing to do (*de dicto*), rather than wanting to perform a particular act that they consider right in a given situation (*de re*). The objection here is that acting on very general moral principles turns morality into a fetish. In rejecting this objection, Albersmeier points out that moral progress should only serve as a motivation for our actions in cases where we have reasonable belief that improvement is necessary (which includes it being possible and appropriate), effective, and optimal.

While most contemporary discussions on moral progress primarily consider this phenomenon at the level of broader social processes, the virtue of this book lies in its focus on moral progress at the individual level. Frauke Albersmeier provides a detailed insight into the various ways we can observe moral changes in individuals—in their desires, the content of their beliefs, the development of dispositions, and their actions—while also pointing out the ways these changes have broader social impacts. Therefore, we can conclude that this book makes a significant contribution to understanding the complex dynamics of the process of moral progress, especially regarding the relationship between moral progress at the individual and collective levels.*

TVRTKO JOLIĆ
Institute of Philosophy, Zagreb, Croatia

* This review is an output of the project "Moral Progress: Individual and Collective" supported by the Croatian Science Foundation (Grant No. IP-2022-10-5341).

Croatian Journal of Philosophy is published three times a year.
It publishes original scientific papers in the field of philosophy.

Croatian Journal of Philosophy is indexed in
The Philosopher's Index, PhilPapers, Scopus, ERIH PLUS
and in *Arts & Humanities Citation Index* (*Web of Science*).

Payment may be made by bank transfer

SWIFT PBZGHR2X
IBAN HR4723400091100096268

Croatian Journal of Philosophy is published with the support
of the Ministry of Science, Education and Youth
of the Republic of Croatia.

Instructions for Contributors

All submissions should be sent to the e-mail: cjp@ifzg.hr. Submissions must be in English and formatted to be double-spaced with suitably wide margins, an A4 page size, and automatic page numbering.

Articles are normally no longer than 8,000 words of main text including bibliography. The Journal will consider longer papers, but, once these limits are exceeded, authors should bear in mind the editorial policy that the acceptance bar raises with increasing length.

Manuscripts should be compiled in the following order: cover page; title; abstract (not exceeding 200 words); keywords (3 to 6); main text; appendices (as appropriate); references.

All the authors of a paper should include their full names, affiliations, postal addresses, telephone and fax numbers and email addresses on the cover page of the manuscript. If a paper is co-written, one author should be identified as the Corresponding Author. The cover page must be submitted as a separate document. All submitted manuscripts must be prepared for blind review, with revealing acknowledgements and self-identifying references removed.

Sources are cited in the text by the author's last name, the publication date of the work cited, and a page number if needed, e.g. (Barber 2007: 324). Full details appear in the reference list in which the year of publication appears immediately after the author's name:

Barber, A. 2007. "Linguistic Structure and the Brain." *Croatian Journal of Philosophy* 21 (7): 317–341.

Williamson, T. 2013. *Identity and Discrimination*. Oxford: Wiley-Blackwell.

The publication of a manuscript in the Croatian Journal of Philosophy is expected to follow standards of ethical behavior for all parties involved in the publishing process: authors, editors, and reviewers. The journal follows the principles of the Committee on Publication Ethics (https://publicationethics.org/resources/flowcharts).

ISSN 1333-1108
9 771333 110001