Zbornik 25. mednarodne multikonference

INFORMACIJSKA DRUŽBA

Zvezek B

Proceedings of the 25th International Multiconference

INFORMATION SOCIETY

Volume B

2022

Kognitivna znanost

Cognitive Science

Uredniki • Editors:
Toma Strle, Borut Trpin, Olga Markič

Ljubljana, Slovenija
13. oktober
13 October
Ljubljana, Slovenia

http://is.ijs.si

Zbornik 25. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2022

Zvezek B

Proceedings of the 25th International Multiconference

INFORMATION SOCIETY – IS 2022

Volume B

Kognitivna znanost
Cognitive Science

Uredniki / Editors

Toma Strle, Borut Trpin, Olga Markič

http://is.ijs.si

13. oktober 2022 / 13 October 2022
Ljubljana, Slovenija

Uredniki:

Toma Strle
Center za Kognitivno znanost, Pedagoška fakulteta, Univerza v Ljubljani

Borut Trpin
Filozofska fakulteta, Univerza v Ljubljani

Olga Markič
Filozofska fakulteta, Univerza v Ljubljani

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2022

Informacijska družba
ISSN 2630-371X

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 127437571
ISBN 978-961-264-242-6 (PDF)

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2022

Petindvajseta multikonferenca *Informacijska družba* je preživela probleme zaradi korone. Zahvala za skoraj normalno delovanje konference gre predvsem tistim predsednikom konferenc, ki so kljub prvi pandemiji modernega sveta pogumno obdržali visok strokovni nivo.

Pandemija v letih 2020 do danes skoraj v ničemer ni omejila neverjetne rasti IKTja, informacijske družbe, umetne inteligence in znanosti nasploh, ampak nasprotno – rast znanja, računalništva in umetne inteligence se nadaljuje z že kar običajno nesluteno hitrostjo. Po drugi strani se nadaljuje razpadanje družbenih vrednot ter tragična vojna v Ukrajini, ki lahko pljuskne v Evropo. Se pa zavedanje večine ljudi, da je potrebno podpreti stroko, krepi. Konec koncev je v 2022 v veljavo stopil not raziskovalni zakon, ki bo izboljšal razmere, predvsem leto za letom povečeval sredstva za znanost.

Letos smo v multikonferenco povezali enajst odličnih neodvisnih konferenc, med njimi »Legende računalništva«, s katero postavljamo nov mehanizem promocije informacijske družbe. IS 2022 zajema okoli 200 predstavitev, povzetkov in referatov v okviru samostojnih konferenc in delavnic ter 400 obiskovalcev. Prireditev so spremljale okrogle mize in razprave ter posebni dogodki, kot je svečana podelitev nagrad. Izbrani prispevki bodo izšli tudi v posebni številki revije Informatica (http://www.informatica.si/), ki se ponaša s 46-letno tradicijo odlične znanstvene revije. Multikonferenco Informacijska družba 2022 sestavljajo naslednje samostojne konference:

- Slovenska konferenca o umetni inteligenci
- Izkopavanje znanja in podatkovna skladišča
- Demografske in družinske analize
- Kognitivna znanost
- Kognitonika
- Legende računalništva
- Vseprisotne zdravstvene storitve in pametni senzorji
- Mednarodna konferenca o prenosu tehnologij
- Vzgoja in izobraževanje v informacijski družbi
- Študentska konferenca o računalniškem raziskovanju
- Matcos 2022

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi ACM Slovenija, SLAIS, DKZ in druga slovenska nacionalna akademija, Inženirska akademija Slovenije (IAS). V imenu organizatorjev konference se zahvaljujemo združenjem in institucijam, še posebej pa udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

S podelitvijo nagrad, še posebej z nagrado Michie-Turing, se avtonomna stroka s področja opredeli do najbolj izstopajočih dosežkov. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel prof. dr. Jadran Lenarčič. Priznanje za dosežek leta pripada ekipi NIJZ za portal zVEM. »Informacijsko limono« za najmanj primerno informacijsko potezo je prejela cenzura na socialnih omrežjih, »informacijsko jagodo« kot najboljšo potezo pa nova elektronska osebna izkaznica. Čestitke nagrajencem!

Mojca Ciglarič, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2022

The 25th *Information Society Multiconference* (http://is.ijs.si) survived the COVID-19 problems. The multiconference survived due to the conference chairs who bravely decided to continue with their conferences despite the first pandemics in the modern era.

The COVID-19 pandemic from 2020 till now did not decrease the growth of ICT, information society, artificial intelligence and science overall, quite on the contrary – the progress of computers, knowledge and artificial intelligence continued with the fascinating growth rate. However, the downfall of societal norms and progress seems to slowly but surely continue along with the tragical war in Ukraine. On the other hand, the awareness of the majority, that science and development are the only perspective for prosperous future, substantially grows. In 2020, a new law regulating Slovenian research was accepted promoting increase of funding year by year.

The Multiconference is running parallel sessions with 200 presentations of scientific papers at eleven conferences, many round tables, workshops and award ceremonies, and 400 attendees. Among the conferences, "Legends of computing" introduce the "Hall of fame" concept for computer science and informatics. Selected papers will be published in the Informatica journal with its 46-years tradition of excellent research publishing.

The Information Society 2022 Multiconference consists of the following conferences:

- Slovenian Conference on Artificial Intelligence
- Data Mining and Data Warehouses
- Cognitive Science
- Demographic and family analyses
- Cognitonics
- Legends of computing
- Pervasive health and smart sensing
- International technology transfer conference
- Education in information society
- Student computer science research conference 2022
- Matcos 2022

The multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, i.e. the Slovenian chapter of the ACM, SLAIS, DKZ and the second national academy, the Slovenian Engineering Academy. In the name of the conference organizers, we thank all the societies and institutions, and particularly all the participants for their valuable contribution and their interest in this event, and the reviewers for their thorough reviews.

The award for life-long outstanding contributions is presented in memory of Donald Michie and Alan Turing. The Michie-Turing award was given to Prof. Dr. Jadran Lenarčič for his life-long outstanding contribution to the development and promotion of information society in our country. In addition, the yearly recognition for current achievements was awarded to NIJZ for the zVEM platform. The information lemon goes to the censorship on social networks. The information strawberry as the best information service last year went to the electronic identity card. Congratulations!

Mojca Ciglarič, Programme Committee Chair
Matjaž Gams, Organizing Committee Chair

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia
Sergio Campos-Cordobes, Spain
Shabnam Farahmand, Finland
Sergio Crovella, Italy

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič

Programme Committee

Mojca Ciglarič, chair
Bojan Orel,
Franc Solina,
Viljan Mahnič,
Cene Bavec,
Tomaž Kalin,
Jozsef Györkös,
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Niko Zimic
Rok Piltaver
Toma Strle
Tine Kolenik
Franci Pivec
Uroš Rajkovič
Borut Batagelj
Tomaž Ogrin
Aleš Ude
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek

KAZALO / TABLE OF CONTENTS

Zbornik 25. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2022

Zvezek B

Proceedings of the 25th International Multiconference

INFORMATION SOCIETY – IS 2022

Volume B

Kognitivna znanost
Cognitive Science

Uredniki / Editors

Toma Strle, Borut Trpin, Olga Markič

http://is.ijs.si

13. oktober 2022 / 13 October 2022
Ljubljana, Slovenija

PREDGOVOR

Na tokratni konferenci Kognitivna znanost sodelujejo avtorice in avtorji, ki se raziskovalno ukvarjajo s kognitivno znanostjo, in predstavljajo tako empirične rezultate svojih raziskav kot tudi teoretska raziskovanja z najrazličnejših področij – od psihologije in nevroznanosti do filozofije in umetne inteligence. Poseben poudarek na letošnji konferenci posvečamo kognitivnim vidikom zaupanja v znanost, kar avtorice in avtorji naslavljajo tako z družbenega, političnega, psihološkega in filozofskega vidika.

Upamo, da bo letošnja disciplinarno in metodološko bogata konferenca odprla prostor za povezovanje pronicljivih idej ter povezala domače in tuje, mlade in izkušene znanstvenice in znanstvenike, ki se ukvarjajo z vprašanji kognicije.

Borut Trpin
Toma Strle
Olga Markič

FOREWORD

At this year's Cognitive Science conference, the authors who actively research in scope of cognitive science present their empirical studies as well as theoretical research from a diverse range of disciplinary backgrounds – from psychology and neuroscience to philosophy and artificial intelligence. A special focus of this year's conference is on cognitive aspects of trust in science. The authors address this topic from a social, political, psychological, and philosophical viewpoint.

We hope that this year's cognitive science conference – rich in disciplinary approaches and methodologies – will open space for exchanging intriguing research ideas and will bring together local and international, junior and senior scientists from a diverse range of areas related to the exploration of the human mind.

Borut Trpin
Toma Strle
Olga Markič

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Toma Strle, Center za Kognitivno znanost, Pedagoška fakulteta, Univerza v Ljubljani

Borut Trpin, Filozofska fakulteta, Univerza v Ljubljani

Olga Markič, Filozofska fakulteta, Univerza v Ljubljani

Urban Kordeš, Center za kognitivno znanost, Pedagoška fakulteta, Univerza v Ljubljani

Matjaž Gams, Odsek za inteligentne sisteme, Institut »Jožef Stefan«, Ljubljana

Into the Constant Now—Comparing DES and micro-phenomenology, two methods for exploring consciousness

Julian Bass-Krueger
MSc University of Vienna
Vienna, Austria
julianbassk@gmail.com

Elisa Wiedemann
Department of Cognitive Science
Central European University PU
Vienna, Austria
wiedemann_elisa@phd.ceu.edu

Ema Demšar
Centre for Consciousness and Contemplative Studies
Monash University
Melbourne, Australia
ema.demsar@monash.edu

ABSTRACT

Here we compare two methods of examining conscious experience—Descriptive Experience Sampling (DES) and micro-phenomenology. Both look at short episodes of experience. Both have safeguards to limit biases and distortions from first-person reporting. But these methods are still different in terms of how they deal with memory, questioning, and analysis. In this pilot study (n=4), we use both methods in the context of a common task. Participants were interviewed about their experience of a mental imagery task using both methods. DES results focused more on fine-grained details of visual experiences. Micro-phenomenology results focused more on how experience extended over time, and how participants engaged with the task. These differences in results show that the investigated methods differ in scope. To further address this, we encourage a critical methodological pluralism where methods can continue to be improved and tested for validity.

KEYWORDS

Consciousness, inner experience, empirical phenomenology, DES, micro-phenomenology

1 BACKGROUND

The study of first-person experience has had a difficult time. In the early 20th century a prolonged disagreement between two rival introspectionist camps led to the field's essential banishing from psychology [1, 2]. A later influential study by Nisbett and Wilson [3] further solidified the notion that first-person data is flawed and distorted by heuristics, overgeneralizations, and memory problems. People simply don't know what's in their consciousness. To give a pragmatic definition, for a conscious person, there is *something that it is like* to be that person [4]. A conscious person might be sipping coffee noticing the rich smell and hearing birds chirp. An unconscious person could for example be in dreamless sleep.

Conscious experience takes up most of our day (presumably) and influences our identity and understanding of the world. It lies behind our sensations, emotions, and thoughts. It is important. And yet it's often either assumed as trivial, approachable through naive methods, or else unattainable, not worth even seeking to understand. New methods reject both premises–consciousness is neither trivial nor unattainable. These methods attempt to systematize consciousness research, in a field that has been dubbed "empirical phenomenology" [5]. They deal with past critiques and attempt to limit biases. Although validity cannot yet be proven, here we test the limits and constraints of these methods. Specifically, we look at Descriptive Experience Sampling, founded by Russell Hurlburt and refined with the aid of fellow researchers [1]. And we'll look at micro-phenomenology—adapted by Claire Petimengin from Pierre Vermersch's explicitation interview [2].

Descriptive Experience Sampling uses random beeps to direct participants towards specific, concrete episodes of experience. Micro-phenomenology guides participants to a state in which memory becomes immediate and lived.

Both methods then use different means to aim for a common goal, of revealing short episodes of experience. Experience described in the abstract is an amalgamation of warped memory, self-perception, conceptual frames, and fleeting impressions. 'This morning I had breakfast and felt sleepy.' In the concrete, however, experience manifests as a flow of vivid nows. '*Now* I'm watching the cream dissolve in my coffee. *Now* I'm picturing what would happen if gravity reversed overnight and I had to rearrange my furniture on the ceiling.' These *nows*, so vivid when lived, can dissolve in memory like cream in coffee, so that we might forget their original color. Methods of empirical phenomenology aim for that color.

Despite similar intentions, there has been some contention between methods. Akhter and Hurlburt have questioned the validity of micro-phenomenology [6]. Petitmengin has argued about DES that "the beeper is not suitable for observing very brief or very fine subjective events" [7]. Is this disagreement warranted? Do methods really reveal different aspects of experience when used with a common task? And if so, does this call into question the validity of one method or the other? Methods might just have different scopes, yielding different results [8]. To address these questions, we compared methods with a shared task.

2 METHODS

2.1 DES

DES uses random beeps through the day to help participants better grasp their own experience. This can involve a specialized beeper or a smartphone. The participant must have an earpiece directly in their ear throughout the procedure. The beeps are delivered at randomized intervals ranging between five minutes to one hour [9]. Six beeps are delivered a day. This usually takes around three or four hours. In most studies, they occur during the

participant's daily life, not in a lab, to increase ecological validity.

After each beep, the participant jots down notes on their inner experience right before the beep. So not inner experience during the beep (e.g., darn that's annoying!) but right before. The goal is to describe that last uninterrupted moment before the beep. Usually this moment is much shorter than what participants first have in mind, and can last a fraction of a second.

Questioning and training is needed in order to apprehend this moment. At the end of each day of sampling, participants are interviewed about the six beeps they collected. The interviews last an hour and any samples not discussed within that time are discarded. There are always multiple days of sampling, usually around 5 or 6, but occasionally many more. The first day of sampling is always discarded and considered training. Subsequent days are often discarded as well, if they don't hew to validity criteria.

Validity depends primarily on participants' ability to clearly describe specific moments of experience with little hesitation and equivocating language. Questioning aims to lead participants away from generalizations. For example, a participant might first say, "I was driving and kinda nervous I think. I'm always nervous when I drive." The use of the term 'always' may indicate that the participant was generalizing. The use of terms 'kinda' and 'I think' could indicate uncertainty stemming from lack of contact with direct experience. Further questioning may reveal that experience before the beep was something completely different—perhaps a mental image of a fat squirrel with the inner speaking "munchy munch." It is common in DES for results to go against participants' initial expectations [9, 10].

2.2 Micro-phenomenology

Micro-phenomenology aims to guide the participants towards vividly reliving and precisely describing a past conscious episode [7]. This episode is of underdetermined length, ranging from a few minutes to a few seconds. The episode can be in the recent past or have occurred many years ago. For the sake of bringing our methods as close as possible to compare them, here we'll apply micro-phenomenology to the recent past and to short episodes (10 seconds).

Memories can be indistinct, so micro-phenomenology aims to guide the participant to an "evocation state" where past experience is 're-lived' [7]. Participants have direct contact with what they saw, heard, or felt at the time of the target experience. Questions aim to 'stabilize' this evocation state and maintain the participant's contact with their experience. For example, participants are periodically asked to return to the beginning of the episode. If the participant digresses, the interviewer can repeat the participant's earlier descriptions.

As in DES, participants are asked for greater specificity about the elements they reveal. For example, if a participant has a mental image, an interviewer might ask, "Is it in colour or in black and white? Is it detailed or fuzzy? Is it dark or light?" [7].

Micro-phenomenology begins by eliciting the context and sensory modalities of past experience—what participants saw, heard, felt, etc. This helps the participant enter the evocation state. Once in this state, questions can be more open ended. Interviewers can ask about the sequence of experience and how different elements change over time. They then focus on specific elements in turn and ask questions to elicit greater specificity. Micro-phenomenology aims for nuance. Questioning can often focus on subtle emotional shifts of even shifts in body or posture that contribute to experience.

There are no firm guidelines for how long a micro-phenomenology interview lasts. However, it is not uncommon for short segments of experience to elicit hour-long interviews. The aim of micro-phenomenology is to uncover the complexity and nuance of the experiential episode both at a particular moment (synchronic dimension) and its development over time (diachronic dimension),with the focus of the interview depending on the research question of the particular study.

2.3 Main differences

Time - Micro-phenomenology typically deals with longer sections of time. Researchers can observe how elements change. Petitmengin writes, "To enter into contact with one's experience, it is necessary to respect its fluid and dynamic character" [11]. DES does also incorporate time though. Experience is not frozen into a static snapshot. For example, if a person is innerly speaking "I need to call mom" this might extend over time. And a fuzzy feeling in their chest might increase in strength over the moment. The difference here is thus of degree, not of type.

Retrospection - Micro-phenomenology, in general, involves substantially more retrospection. The target experience could be years before the interview [11]. In DES, the target experience is a few seconds before the notetaking and less than 24 hours before the interview. There are still memory demands but they are fewer. However, as mentioned, micro-phenomenology can also be done with the target experience shortly before the interview [12]. This is the case for our comparison study.

Directing attention - Micro-phenomenology aims for an evocation state in which participants re-live the original experience. DES takes a more skeptical approach. DES questions encourage the participant to doubt if reported elements were really part of their direct experience. DES acknowledges that this skepticism might lead it to miss out on elements of experience. But Hurlburt sees this as preferable to reporting elements that weren't there [9]. Micro-phenomenology prefers having as full an impression of experience as possible. It offers participants opportunities to revise and clarify their reports, but in service of maintaining an evocation state, doesn't 'grill' participants to the extent that DES does.

Questioning – Micro-phenomenology questioning is "non-inducive but directive" [7]. DES questioning is non-inducive and non-directive. For example, micro-phenomenology asks about specific sensory modalities in turn, i.e. 'Do you hear anything?' It holds that this is necessary to elicit greater detail since participants may not know where to direct their attention. DES would instead ask, 'Was there anything else in your experience?'.

In general, micro-phenomenology is more trusting of participant reports. DES places a greater emphasis on skepticism, training participants in order to get greater fidelity. For example, the first day of training is always discarded with DES. This is not the case with micro-phenomenology. Training interviews are occasionally used but optional.

Validity - There is agreement between methods about how to judge validity. Both acknowledge that rules and explanations of

the method make their own case for validity. A successful sample/interview then depends on these guidelines being followed, and questions being suitably content-neutral and non-leading. Other points of agreement include situating methods in a net of third-person observables—for example, can first-person data link with behavioral data? Can correlations be found with neuroimaging? No one correlation can address validity but networks of connections can help lead to first- and third- person methods informing each other through "mutual constraints" [13].

Differences include differing methods for judging veracity. Both methods rely on both verbal and non-verbal cues. But DES leans more heavily on verbal cues, like subjunctification [9]. Is the participant saying umm, I think, kindof, maybe, sorta, I guess? Then it's likely they're not describing direct experience. Micro-phenomenology relies more on visual cues—for example a participant's eyes pointing upwards indicating that they're in an evocation state.

Petitmengin also advocates checking a participant's reported experience against the researcher's own experience, calling this the "kingpin of all validation" [7]. Is it similar or at least plausible? Hurlburt and Akhter [6] see this as harmful—a participants' experience may be radically different from the researcher's, and so should be 'bracketed' as much as possible.

3 PROCEDURE

This study involved four participants—a small sample size aimed at highlighting certain method contours rather than generalizing or making claims of statistical significance. All four were female students residing in Slovenia, aged 23 to 26. They are referred to here using pseudonyms. Each participant underwent both the DES and micro-phenomenology procedure. However, two started with micro-phenomenology and two started with DES (to limit biasing). There was a break (at least six days) before switching methods.

To facilitate comparison, the interviews concerned participants' experience of a task. We used a mental imagery elicitation task, in which participants were given descriptive prompts and 10 seconds to form mental images. Examples of prompts include: "A child holds an ice cream cone with three scoops. The ice cream falls onto the hot pavement." "A storm cloud gathers over a city. A lightning bolt strikes."

Before the task came training. For DES, this involved three days of DES sampling during the participant's everyday life—going to class, cafés, etc. Participants received six beeps a day, jotted down their consciousness experience in the moment before the beep, and received hour-long interviews within 24 hours of sample collection [Fig. 1].

For micro-phenomenology, training was much shorter. Participants were given a task shortly before the main task—to spell the word octopus. Participants were then interviewed to give them some practice and familiarity with micro-phenomenology and the interview procedure [Fig. 2].

For the task, the DES portion involved 32 pre-recorded prompts. 10 seconds followed each prompt, allowing for mental imagery formation. Five random beeps were interspersed throughout the task, ranging from 1-10 seconds after the prompt concluded. There was a DES interview after each beep. J.B.-K. conducted these interviews [Fig. 1].

The micro-phenomenology task involved 2 prompts. These were on separate days. Participants again had 10 seconds after the prompt to form mental images. They were interviewed following the guidelines for micro-phenomenological interviews [7] after each prompt. E.W. conducted these interviews [Fig. 2].

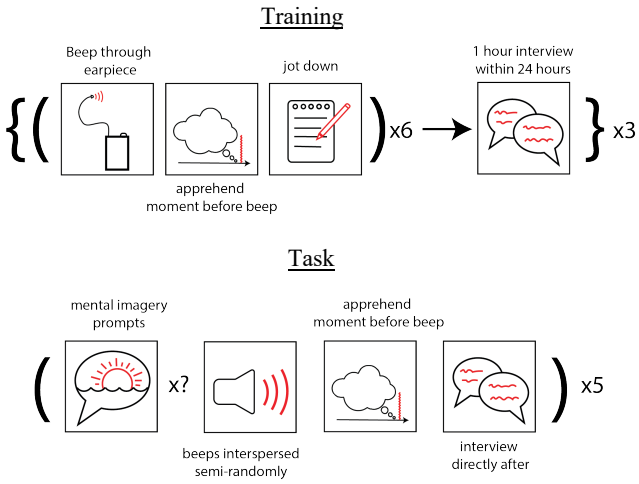


Figure 1: DES training and task

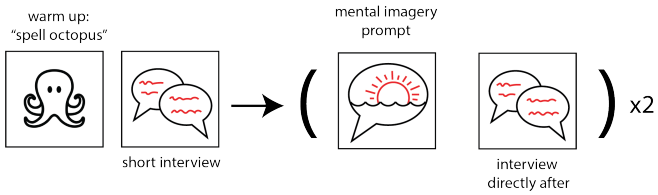


Figure 2: Micro-phenomenology training and task

4 RESULTS

4.1 Similarities

Image characteristics - Both methods uncovered common visual phenomena. One example of this is with GIF-like repetition. This may be something specific to our current digital age. These short, repeating moving images are common on social media. Many older people in DES sampling have mental images in black and white [14]. The technologies of our age may shape our perception and perceptual cognition.

Other commonalities include elements changing over time. For both methods, images didn't always emerge fully formed. And micro-phenomenology further shows how images morphed or how new elements entered. For both methods, images could either be moving or static.

With both methods, mental images had differing levels of detail—inter- and intrasubject. Images were sometimes clear. Sometimes they were fuzzy, indistinct, ghostly, or blurry. Visual elements were sometimes realistic and sometimes cartoonish.

Interactions with other modalities - Inner images could interact with other types of experience. Both methods revealed words and images interacting. Micro-phenomenology revealed participants sometimes innerly repeating words from the prompt. In one case, these words were a distraction from forming images. In another case, they spurred on a new visual perspective.

For DES one participant misheard the word 'chirp' as 'gerb'. At the moment of the beep, she was innerly repeating it, wondering what it meant, and had a visual impression without any visual elements explicitly present—just a large mass.

Images could also interact with feelings. DES found that 1/5 of samples involved feelings. These were sometimes positive in valence ('calm') or sometimes negative ('dislike'). Certain prompts correlated with negative feelings—like the prompt "A family gathers around the dinner table. The father starts serving food."

Micro-phenomenology also found feelings. For example, for a prompt about two children skating on a pond, Jelka added a mother to the scene and projected her own worry onto the mother.

4.2 Differences

Visual differences - While there were similarities concerning mental imagery formation, there were differences as well. With DES, for Jelka, all 5 prompt samples involved imagery with a dual vantage point. She was both looking at the image from a distance but at the same time had another vantage point of being surrounded by the scene. Think of simultaneously watching a movie on a screen and being in the movie as the main character. Since this dual vantage point was found in all of her samples, one might expect it to be a generalizable feature of her mental images. But micro-phenomenology didn't find it. It found instances of 3rd and 1st person inner images for Jelka, but never both at the same time. Perhaps the dual vantage point was present but not apprehended.

DES findings focused more on characteristics of mental images.

—Images can have borders, no borders, or focus can be on the center so the participant is unsure of whether or not the image has edges.

—Images can be in a separate mental space or positioned over the real world, for instance on a "3D screen."

—Mental images can involve aspects that would be impossible in real physical space.

—Two simultaneous visual spaces can be present at the same time. For example, Anna had one visual space of children skating on a frozen pond, and a separate space where she was creating a face to add to the children.

Time – Micro-phenomenology focused more on experience evolving over time. We can see how imagery changes. We can see how participants interact with prompts, referring back to them, and playing with them. We see the broader experience of the task.

—Some elements came naturally, others required concentration.

—Elements could be disproportionate and not fit with the scene. For example, Jelka imagined a tree with birds that didn't fit with the rest of the scene. It was too big, and a different color. We see how new elements enter and how they relate to previous elements.

—The task could involve constrained freedom or constraint. Jelka felt constrained at times. She had to imagine things she wasn't interested in. Anna, especially, felt freedom. She could imagine whatever she wanted. Anna also played with the prompts. For example, given a prompt about a boy with three scoops of ice-cream, Anna imagined three ice-cream scoop tools. We can see how she engaged with the task, lightheartedly testing how far she could push the prompts. DES could not have revealed this entire sequence of trying out different visual components.

5 DISCUSSION

Despite similarities, these methods have different scopes and reveal different results. Micro-phenomenology revealed more temporal dynamics. We saw how images evolved over time, and how participants interacted with the prompts. DES revealed more visual characteristics of images. This is contrary to Petitmengin's comment concerning DES's limited experiential detail: "I doubt whether the beep enables the interviewee to direct his attention from 'what' to 'how', unless by chance" [7]. It also goes against claims from Froese, Seth, and Gould that DES adheres only to a "shallow conception of consciousness" [15].

Note that methods differed in the treatment of fine-grained details. DES revealed dual aspect imagery and micro-phenomenology did not. This could have been the result of differing experience or a product of the research design where training with one method alters reporting with the other method.[1] It could also be a result of one or another method hewing more closely to experience. If this is the case, we need to make sure our methods are faithful. Methods that distort experience may lead to disagreements and stall progression of the study of consciousness. For this reason, issues with retrospection, memory distortion, presuppositions, and biases need to be handled carefully. Practitioners of any method need to question what its intent is, whether its guidelines are coherent, and what research questions it can and can't answer.

Horizons are open for refinement of methods and experimentation. Emerging research is even combining elements from micro-phenomenology and DES [16-19]. Oblak, for example, combined influences from both methods for interviews investigating experience during a visual-spatial memory task [16]. Springinsfeld conducted micro-phenomenology inspired interviews shortly after targeted experience–aiming for interviews on the same day as a bulimic individual's vomiting episodes, to minimize retrospection demands [17]. Caporusso used DES-style beeps with an interview method hewing more closely to micro-phenomenology in order to better understand sense of self and boundaries in daily life and compare this to experiences of boundary dissolution [18]. And Bass-Krueger adapted DES to a slightly wider temporal scope to investigate what is really meant by a 'moment' of experience [19]. Critical methodological pluralism is important going forward. We must acknowledge differing avenues of exploring lived experience, while questioning where exactly these avenues lead us.

[1] Procedurally, there seemed to have been an effect of experience with one method on participants' approach to the other (new) method of investigating experience. For instance, participants who started with DES and then moved on to the micro-phenomenological interviews approached their experience with more skepticism and caution than those who started with micro-phenomenology. Conversely, one participant who started with the micro-phenomenological interviews and then moved on to DES at first found the latter method 'too skeptical' and both required at least as much training as participants with no prior experience with first-person reporting. However, with such a small sample, it is hard to disentangle how experience with one method or the other may have influenced our final results.

ACKNOWLEDGEMENTS

Special thanks to Urban Kordeš and Toma Strle for the resources and guidance to carry out this project. And to our participants for the trust granted in our research.

REFERENCES

[1] Russell Hurlburt, 1993. Sampling Inner Experience in Disturbed Affect. Plenum Press.

[2] Pierre Vermersch, 1999. Introspection as practice. *Journal of Consciousness Studies 6,* 2–3, 17–42.

[3] Richard E. Nisbett and Timothy D. Wilson, 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84 3, 231–259. https://doi.org/10.1037/0033-295x.84.3.231*

[4] Thomas Nagel, 1974. What is it like to be a bat? *The Philosophical Review 84*, 4, 435-450.

[5] Urban Kordeš, 2016. Going beyond theory: Constructivism and empirical phenomenology. *Constructivist Foundations 11*, 2, 375–385.

[6] Russell Hurlburt and Sarah Akhter, 2006. The Descriptive Experience Sampling method. *Phenomenology and the Cognitive Sciences 5*, 271-301. https://doi.org/10.1007/s11097-006-9024-0

[7] Claire Petitmengin, 2006. Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive Sciences 5*, 229-269. https://doi.org/10.1007/s11097-006-9022-2

[8] Urban Kordeš and Ema Demšar, 2021. Horizons of becoming aware: Constructing a pragmatic-epistemological framework for empirical first-person research. *Phenomenology and the Cognitive Sciences*, 1-29. https://doi.org/10.1007/s11097-021-09767-6

[9] Russell Hurlburt and Christopher Heavey, 2006. *Exploring Inner Experience: The Descriptive Experience Sampling Method*. John Benjamins. https://doi.org/10.1075/aicr.64

[10] Russell Hurlburt and Christopher Heavey, 2015. Investigating pristine inner experience: Implications for experience sampling and questionnaires. *Consciousness and Cognition 31*, 148–159. http://doi.org/10.1016/j.concog.2014.11.002

[11] Claire Petitmengin, 2011. Describing the experience of describing? The blindspot of introspection. *Journal of Consciousness Studies 18*, 1, 44-62.

[12] Natalie Depraz, Maria Gyemant, and Thomas Desmidt, 2017. A first-person analysis using third-person data as a generative method: A case study of surprise in depression. *Constructivist Foundations* 12, 2, 191-203.

[13] Francisco Varela, 1996. Neurophenomenology: A Methodological Remedy for the Hard Problem. *Journal of Consciousness Studies* 3, 4, 330-49.

[14] Russell Hurlburt, 2011. Investigating Pristine Inner Experience: Moments of Truth. Cambridge University Press.

[15] Tom Froese, Anil Seth, and Cassandra Gould, 2011. Validating and calibrating first- and second-person methods in the science of consciousness. *Journal of Consciousness Studies 18*, 2, 38-64.

[16] Aleš Oblak, Anka Slana Ozimič, Grega Repovš, and Urban Kordeš, 2022. What Individuals Experience During Visuo-Spatial Working Memory Task Performance: An Exploratory Phenomenological Study. *Frontiers in Psychology* 13. https://doi.org/10.3389/fpsyg.2022.811712

[17] Constanze Springinsfeld [In Review]. Attention, dissociation, and feelings in bulimia nervosa [prospective title]. Mei:CogSci Master's Thesis

[18] Jaya Caporusso, 2022. Dissolution experiences and the experience of the self: An empirical phenomenological investigation. *Mei:CogSci Master's Thesis*. 10.25365/thesis.71694

[19] Julian Bass-Krueger, 2021. Consciousness and Time. *Mei:CogSci Master's Thesis* 10.25365/thesis.70232

LTP and LTD dependence on spontaneous activity in hippocampal and cortical glutamate synapses and the role of anaesthetics in the study of plasticity and learning

Maša Bratuša†
MEi:CogSci
University of Ljubljana
Ljubljana Slovenia
bratusa.masa@gmail.com

ABSTRACT

The following article is a condensed version of a review paper which was motivated by the hypothesis put forward by Benuskova and her colleagues that an ongoing pre- and postsynaptic spontaneous activity (SA) determines not only the degree of input-specific LTP elicited by various plasticity-inducing protocols, but also the degree of associated LTD in neighbouring non-tetanized inputs. It appears that understanding regularities of spontaneous activity can help us define boundary conditions for both LTP/LTD induction and maintenance. We look into LTP and LTD induction in excitatory glutamate synapses, their interrelatedness and connected non-glutamate plasticity. We then assess the role of SA in plasticity and consider what it means for in vitro studies where SA is limited. We inquire how anaesthetics affect the general capacity for LTP and LTD induction and maintenance and we join this with results on their effects on SA. All of this is taken together in order to suggest protocols of notable ecological validity and to provide an argument in favour of procedure standardization in the field.

KEYWORDS

Hippocampus, Cerebral Cortex, Anaesthesia, Sleep, Spontaneous Activity, Synaptic Long-Term Potentiation (LTP), Synaptic Long-Term Depression (LTD)

INTRODUCTION

In the process of learning, there is both an increase of electrochemical signalling in some synapses and a decrease thereof in others. Potentiation and depression include many physiological changes and are therefore more stable over time in comparison to facilitation and inhibition [1]. The general understanding of NMDA-dependent LTP is as follows: presynaptic stimulation opens postsynaptic NMDA channels which cause a rise in postsynaptic Ca2+. Strong depolarizations displace magnesium ions, which open more NMDA channels in a positive feedback loop manner. The postsynaptic neuron accepts even more Ca2+ ions, and this superfluous concentration of Ca2+ then activates CAMKII, increases cAMP and PKAII concentrations. Activated CAMKII is known to increase the volume of the dendritic spines [2] and stimulate new AMPAr integration [3], with both of these processes being key criteria for successful LTP.

In their seminal work, Abraham and Goddard [4] showed that there is otherwise a notable difference between homosynaptic and heterosynaptic plasticity: *“Homosynaptic plasticity occurs at synapses that were active during the induction. It is also called input-specific or associative, governed by Hebbian-type learning rules. Heterosynaptic plasticity can be induced by episodes of strong postsynaptic activity also at synapses that were not active during the induction, thus making any synapse at a cell a target to heterosynaptic changes. Both forms can be induced by typical protocols and operate on the same time scales but have differential computational properties and play different roles in learning systems. Homosynaptic plasticity mediates associative modifications of synaptic weights. Heterosynaptic plasticity counteracts runaway dynamics introduced by Hebbian-type rules and balances synaptic changes."* [5].

A conceptual shift in our understanding of “activity dependence” in heterosynaptic plasticity occured after the following experiment: Prior to stimulation the medial perforant pathway (MPP) and the lateral perforant path (LPP) were equally weighted. With low-frequency stimulation spontaneous input activity was largely correlated and only simultaneous or closely successive spikes at these two inputs could fire the postsynaptic granule cell. Meanwhile high-frequency stimulation of the MPP decorrelated the activity between LPP and MPP, which lead to lower postsynaptic activity. Notably, there was no heterosynaptic LTD when the presynaptic spontaneous activity was blocked [6]. This became known as the Benuskova-Abraham model which explains "heterosynaptic" LTD as a homosynaptic phenomenon due to presynaptic activity.

Meanwhile, the baseline difference between LTP- and LTD-inducing protocols can most simply be illustrated with a difference in stimulation protocols: "*900 pulses of stimuli induced LTD when applied at lower frequencies (1–3 Hz), and induced LTP when applied at a higher frequency (30 Hz).“* [7]

All of the aforementioned considerations led researchers [8, 9] to investigate the role of background SA in memory formation. It should be noted that any activity which is not

evoked by immediate sensory processing can be considered as spontaneous [10, 11, 12]. The goal of our review was to integrate evaluations of all known processes that affect the animals ability to “create a memory trace”, whether it is the physiological condition of the animal or how the inquiry into physiological change is performed.

METHODS

Data was collected from 232 peer-reviewed studies on excitatory glutamate synapses of granule cells in the dentate gyrus, CA1 neurons of the hippocampus (HPC), and cortical (CTX) networks, including those that dealt with developmental, pathophysiological and behavioural data. We also included computational studies of synaptic plasticity. In the process of integration, various types of methodological differences had to be kept in mind.

RESULTS and DISCUSSION

At the onset of writing we wanted to achieve a sound, precise and conclusive multivariate analysis. Yet this numerical approach proved to be impossible due to overarching disparities in experimental protocols. The differences in methods and materials make these experiments dissimilar to the point of barely studying the same phenomenon at all, not to mention the consideration that plasticity phenomena are not a uniform class to begin with [13]. In the following sections, we are nevertheless able to provide some conclusions about which variables ought to be controlled for so that the experimental work is ecologically valid while also giving results that are available for inferences on subsequent, more complex paradigms within the study of memory.

Firstly, the evidence that SA plays a key role in induction and maintenance of proper strength of LTP and concurring, homeostatic LTD is overwhelming [14, 15, 16]. In order to provide a realistic picture of synaptic plasticity (in which SA is as natural as possible), experiments on intact tissues should be given preference [17], since all nerve ablation limits physiological SA input. For example, when studying the CA1 region, its connections to CA3 [18], the dentate gyrus [19], the entorhinal CTX [20] and the medial prefrontal CTX [21] ought to be maintained. Considering norepinephrine [7] and dopamine [22] modifications on glutamate-synapse plasticity, there is good reason to believe that both the amygdala and nucleus accumbens should remain connected to the HPC area under study. But when it comes to the CTX, the scope of kept projections largely depends on the cortical region in question. Unsurprisingly, and in accordance with many authors referenced in the full paper, a preference for in vivo recordings is advised [17, 23, 24]. Nevertheless, many authors agree that thoughtful attention to in vitro conditions could still prove fruitful.

Secondly, no matter the nature of the preparation, we would do best to also keep track of what is happening on other, non-glutamate synapses – since these signalling chains are extensively interdependent. Along with the previously mentioned norepinephrine and dopamine receptors, endocannabinoid, GABA and various acetylcholine receptors should be accounted for in order for us to be able to interpret and generalize our findings [25]. Surveillance of tyrosine [25], serine [26], adenosine/ATP [27] and Ca2+ secretion [28] whether it be from neighbouring neurons or glial cells also appears to play a vital role in outcomes of synaptic plasticity. Especially in the case of astrocytes, close monitoring of glutamate secretion should not be neglected. As far as the author is aware, all of these recordings are not possible simultaneously - so a full analysis would require iterations of the same paradigm with different permutations of controlled variables. Although genetic similarity of laboratory animals is regular practice, we have found evidence that conditions regarding nutrition, activity, sleep and stress should be matched as closely as possible, as they all play a role in establishing baseline stress levels and ionic/aminoacid signalling [29, 30]. Stress/norepinephrine [31] minimization through ensuring environments that best resemble the ecological niche and allow for natural behaviours is crucial both in terms of deriving inferences on physiological plasticity in humans and ethical concerns. Due to dendrite [32] and button [33] restructuring that occurs in synapses after the process of learning, it would be advised to scan for their baseline structure since an intricate confluence of signalling chains appears to take place at that scale.

Thirdly, we have taken a stance that if we are to study memory itself, we should focus on studies where it is represented as a "fully learned association with practical effects" which can be doubtlessly confirmed only with experiments within behavioural paradigms [34, 35]. This functionalist approach requires multiple-synapse learning with behavioural timescales (seconds-to-minutes). Not only that, but it is also unquestionably dependent on replay during sleep [36], which means that an understanding of phosphorylations [37] and gene expression [38] is an indispensable part of the puzzle. If we are to understand memory, we ought to control for post-learning sleep duration and composition, but also for the quantity of operative gap junction [39] channels that extensively contribute to the plasticity-related signalling in sleep, both through slow oscillations and sharp-wave ripples [40].

In short, there is overwhelming evidence that SA within or outside the region of interest is crucial to synaptic plasticity in a myriad of forms (post-tetanic spiking [41], bursting [42], theta oscillations [43], slow oscillations [44] and sharp wave ripples [45, 46]) and that all of them should be taken into consideration. The more complex the type of learning (declarative vs. nondeclarative, behavioural sequences vs. single behaviours, simple classical conditioning vs. nonassociative learning), the larger the region of interest and the more notable the effect of these sleep phenomena. This compounding of

complexity also applies to most previously mentioned signalling, as the area of messenger perfusion also grows.

Lastly, both in vivo conditions and tissue extraction demand the use of anaesthetics. Due to its equal effect on inhibitory and excitatory receptors, which results in successful plasticity induction while also providing sufficient insentience, application of urethane seems to be the best option for studying plasticity, at least in adult subjects [47]. According to previous research, isoflurane appears to be the second best choice. There is some evidence to believe that sevoflurane is a good option for experiments in the neonatal period [48]. There might be some alternatives to anaesthetic predicaments, e.g. severing some sensory projections, usage of neurotransmitter perfusions that would correct for their effects, such as norepinephrine [36], or a combination of both measures. Nevertheless, a routine use of these remains in the realm of the hypothetical since the bioethical committees might remain sceptical about what lowering the anaesthetic dose would mean in terms of sentience and anguish [49, 50].

At this point in time, we are far from being in possession of any sort of statistical law that could be considered ecologically valid even in simpler types of learning/plasticity. The author is aware that the variables listed in the previous sections taken together are essentially calling for an "ideal experiment" which is entirely unattainable within one laboratory. Yet it appears that a combined effort of multiple institutions could overcome these constraints of time and funding and make greater strides in the integration of experimental results into a cohesive body of knowledge. A collaborative search for a law that could easily generalize across experimental conditions should most likely start with a standardisation of materials and methods and careful coordination of experimental tasks within the in vitro domain of plasticity in order to gradually build up towards the end goal of understanding declarative memory formation.

In conclusion, our research could only show that the spectrum of phenomena contributing to various levels of plasticity is strikingly wide and heavily interconnected - to the point that a comprehensive understanding of learning is apparently not achievable through inherently untransferable results of nonpartisan research.

ACKNOWLEDGMENTS

The author would like to thank prof. RNDr. Ľubica Beňušková, PhD. from Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University, Slovakia for her indispensable mentorship in the writing of the original paper [51].

REFERENCES

[1] Antonov, I., E. R. Kandel, and R. D. Hawkins. 2010. “Presynaptic and Postsynaptic Mechanisms of Synaptic Plasticity and Metaplasticity during Intermediate-Term Memory Formation in Aplysia.” Journal of Neuroscience 30 (16): 5781–91. https://doi.org/10.1523/jneurosci.4947-09.2010.

[2] Pi, Hyun Jae, Nikolai Otmakhov, Farida El Gaamouch, David Lemelin, Paul De Koninck, and John Lisman. 2010. “CaMKII Control of Spine Size and Synaptic Strength: Role of Phosphorylation States and Nonenzymatic Action.” Proceedings of the National Academy of Sciences 107 (32): 14437–42. https://doi.org/10.1073/pnas.1009268107.

[3] Lu, W., K. Isozaki, K. W. Roche, and R. A. Nicoll. 2010. “Synaptic Targeting of AMPA Receptors Is Regulated by a CaMKII Site in the First Intracellular Loop of GluA1.” Proceedings of the National Academy of Sciences 107 (51): 22266–71. https://doi.org/10.1073/pnas.1016289107.

[4] Abraham, W. C., and G. V. Goddard. 1983. “Asymmetric Relationships between Homosynaptic Long-Term Potentiation and Heterosynaptic Long-Term Depression.” Nature 305 (5936): 717–19. https://doi.org/10.1038/305717a0

[5] Chistiakova, Marina, Nicholas M. Bannon, Maxim Bazhenov, and Maxim Volgushev. 2014. “Heterosynaptic Plasticity: Multiple Mechanisms and Multiple Roles.” The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry 20 (5): 483–98. https://doi.org/10.1177/1073858414529829.

[6] Benuskova, Lubica, and Wickliffe C. Abraham. 2007. “STDP Rule Endowed with the BCM Sliding Threshold Accounts for Hippocampal Heterosynaptic Plasticity.” Journal of Computational Neuroscience 22 (2): 129–33. https://doi.org/10.1007/s10827-006-0002-x.

[7] Katsuki, Hiroshi, Yukitoshi Izumi, and Charles F. Zorumski. 1997. “Noradrenergic Regulation of Synaptic Plasticity in the Hippocampal CA1 Region.” Journal of Neurophysiology 77 (6): 3013–20. https://doi.org/10.1152/jn.1997.77.6.3013

[8] Beňušková, Ľubica, and Peter Jedlička. 2012. “COMPUTATIONAL MODELING of LONG-TERM DEPRESSION of SYNAPTIC WEIGHTS: INSIGHTS from STDP, METAPLASTICITY and SPONTANEOUS ACTIVITY.” Neural Network World 22 (2): 161–80. https://doi.org/10.14311/nnw.2012.22.010.

[9] Shirrafiardekani, Azam, Lubica Benuskova, and Jörg Frauendiener. 2019. “A Voltage-Based Metaplasticity Rule Applied to the Model Hippocampal Granule Cell Accounts for Homeostatic Heterosynaptic Plasticity,” February. https://doi.org/10.1101/557173.

[10] Vidaurre, Diego, Laurence T. Hunt, Andrew J. Quinn, Benjamin A. E. Hunt, Matthew J. Brookes, Anna C. Nobre, and Mark W. Woolrich. 2018. “Spontaneous Cortical Activity Transiently Organises into Frequency Specific Phase-Coupling Networks.” Nature Communications 9 (1). https://doi.org/10.1038/s41467-018-05316-z.

[11] Sun, W., and Y. Dan. 2009. “Layer-Specific Network Oscillation and Spatiotemporal Receptive Field in the Visual Cortex.” Proceedings of the National Academy of Sciences 106 (42): 17986–91. https://doi.org/10.1073/pnas.0903962106.

[12] Florin, Esther, and Sylvain Baillet. 2015. “The Brain’s Resting-State Activity Is Shaped by Synchronized Cross-Frequency Coupling of Neural Oscillations.” NeuroImage 111 (May): 26–35. https://doi.org/10.1016/j.neuroimage.2015.01.054.

[13] Rc, Malenka, and Bear Mf. 2004. “LTP and LTD: An Embarrassment of Riches.” Neuron. September 30, 2004. https://pubmed.ncbi.nlm.nih.gov/15450156/.

[14] Griffiths, Benjamin J., George Parish, Frederic Roux, Sebastian Michelmann, Mircea van der Plas, Luca D. Kolibius, Ramesh Chelvarajah, et al. 2019. “Directional Coupling of Slow and Fast Hippocampal Gamma with Neocortical Alpha/Beta Oscillations in Human Episodic Memory.” Proceedings of the National Academy of Sciences 116 (43): 21834–42. https://doi.org/10.1073/pnas.1914180116

[15] Umbach, Gray, Pranish Kantak, Joshua Jacobs, Michael Kahana, Brad E. Pfeiffer, Michael Sperling, and Bradley Lega. 2020. “Time Cells in the Human Hippocampus and Entorhinal Cortex Support Episodic Memory.” Proceedings of the National Academy of Sciences 117 (45): 28463–74. https://doi.org/10.1073/pnas.2013250117.

[16] Dehaene, Stanislas, and Jean-Pierre Changeux. 2005. “Ongoing Spontaneous Activity Controls Access to Consciousness: A Neuronal Model for Inattentional Blindness.” Edited by Larry Abbott. PLoS Biology 3 (5): e141. https://doi.org/10.1371/journal.pbio.0030141.

[17] Antkowiak, B. 2002. “In Vitro Networks: Cortical Mechanisms of Anaesthetic Action.” British Journal of Anaesthesia 89 (1): 102–11. https://doi.org/10.1093/bja/aef154.

[18] Lin, Xiaoxiao, Michelle Amalraj, Crisylle Blanton, Brenda Avila, Todd C. Holmes, Douglas A. Nitz, and Xiangmin Xu. 2021. “Noncanonical Projections to the Hippocampal CA3 Regulate Spatial Learning and Memory by Modulating the Feedforward Hippocampal Trisynaptic Pathway.” Edited by Thomas Klausberger. PLOS Biology 19 (12): e3001127. https://doi.org/10.1371/journal.pbio.3001127.

[19] Isomura, Yoshikazu, Anton Sirota, Simal Özen, Sean Montgomery, Kenji Mizuseki, Darrell A. Henze, and György Buzsáki. 2006. “Integration and Segregation of Activity in Entorhinal-Hippocampal Subregions by Neocortical Slow Oscillations.” Neuron 52 (5): 871–82. https://doi.org/10.1016/j.neuron.2006.10.023.

[20] Hahn, Thomas T G, James M McFarland, Sven Berberich, Bert Sakmann, and Mayank R Mehta. 2012. “Spontaneous Persistent Activity in Entorhinal Cortex Modulates Cortico-Hippocampal Interaction in Vivo.” Nature Neuroscience 15 (11): 1531–38. https://doi.org/10.1038/nn.3236.

[21] Lundqvist, Mikael, Jonas Rose, Pawel Herman, Scott L Brincat, Timothy J Buschman, and Earl K Miller. 2016. “Gamma and Beta Bursts Underlie Working Memory.” Neuron 90 (1): 152–64. https://doi.org/10.1016/j.neuron.2016.02.028.

[22] Atherton, Laura A., David Dupret, and Jack R. Mellor. 2015. “Memory Trace Replay: The Shaping of Memory Consolidation by Neuromodulation.” Trends in Neurosciences 38 (9): 560–70. https://doi.org/10.1016/j.tins.2015.07.004.

[23] Whittington, Miles A., Mark O. Cunningham, Fiona E.N. LeBeau, Claudia Racca, and Roger D. Traub. 2010. “Multiple Origins of the Cortical Gamma Rhythm.” Developmental Neurobiology 71 (1): 92–106. https://doi.org/10.1002/dneu.20814.

[24] Pollard, Christopher E., and Anthony Angel. 1990. “Spontaneous Single Cell Discharge in Rat Somatosensory Cortical Slices and Its Relationship to Discharge in the Urethane-Anaesthetized Rat.” Brain Research 518 (1-2): 120–26. https://doi.org/10.1016/0006-8993(90)90962-b.

[25] Bashir, Zafar I. 2003. “On Long-Term Depression Induced by Activation of G-Protein Coupled Receptors.” Neuroscience Research 45 (4): 363–67. https://doi.org/10.1016/s0168-0102(03)00002-6.

[26] Nishiyama, Makoto, Kyonsoo Hong, Katsuhiko Mikoshiba, Mu-ming Poo, and Kunio Kato. 2000. “Calcium Stores Regulate the Polarity and Input Specificity of Synaptic Modification.” Nature 408 (6812): 584–88. https://doi.org/10.1038/35046067.

[27] Chen, Jiadong, Zhibing Tan, Li Zeng, Xiaoxing Zhang, You He, Wei Gao, Xiumei Wu, et al. 2012. “Heterosynaptic Long-Term Depression Mediated by ATP Released from Astrocytes.” Glia 61 (2): 178–91. https://doi.org/10.1002/glia.22425.

[28] SUL, JAI-YOON, GEORGE OROSZ, RICHARD S. GIVENS, and PHILIP G. HAYDON. 2004. “Astrocytic Connectivity in the Hippocampus.” Neuron Glia Biology 1 (1): 3–11. https://doi.org/10.1017/s1740925x04000031.

[29] Roth, Richard H., Robert H. Cudmore, Han L. Tan, Ingie Hong, Yong Zhang, and Richard L. Huganir. 2020. “Cortical Synaptic AMPA Receptor Plasticity during Motor Learning.” Neuron 105 (5): 895-908.e5. https://doi.org/10.1016/j.neuron.2019.12.005.

[30] Dittman, Jeremy S., Anatol C. Kreitzer, and Wade G. Regehr. 2000. “Interplay between Facilitation, Depression, and Residual Calcium at Three Presynaptic Terminals.” The Journal of Neuroscience 20 (4): 1374–85. https://doi.org/10.1523/jneurosci.20-04-01374.200

[31] Tomar, Anupratap, Denis Polygalov, Sumantra Chattarji, and Thomas J. McHugh. 2021. “Stress Enhances Hippocampal Neuronal Synchrony and Alters Ripple-Spike Interaction.” Neurobiology of Stress 14 (May): 100327. https://doi.org/10.1016/j.ynstr.2021.100327.

[32] Ohtsuki, Gen. 2019. “Modification of Synaptic-Input Clustering by Intrinsic Excitability Plasticity on Cerebellar Purkinje Cell Dendrites.” The Journal of Neuroscience 40 (2): 267–82. https://doi.org/10.1523/jneurosci.3211-18.2019.

[33] Emptage, Nigel J., Christopher A. Reid, and Alan Fine. 2001. “Calcium Stores in Hippocampal Synaptic Boutons Mediate Short-Term Plasticity, Store-Operated Ca2+ Entry, and Spontaneous Transmitter Release.” Neuron 29 (1): 197–208. https://doi.org/10.1016/s0896-6273(01)00190-8.

[34] Cone, Ian, and Harel Z. Shouval. 2021. “Behavioral Time Scale Plasticity of Place Fields: Mathematical Analysis.” Frontiers in Computational Neuroscience 15 (March). https://doi.org/10.3389/fncom.2021.640235.

[35] Cao, Jun, Nanhui Chen, Tianle Xu, and Lin Xu. 2004. “Stress-Facilitated LTD Induces Output Plasticity through Synchronized-Spikes and Spontaneous Unitary Discharges in the CA1 Region of the Hippocampus.” Neuroscience Research 49 (2): 229–39. https://doi.org/10.1016/j.neures.2004.03.001.

[36] Reasor, Jonathan D., and Gina R. Poe. 2008. “Learning and Memory during Sleep and Anesthesia.” International Anesthesiology Clinics 46 (3): 105–29. https://doi.org/10.1097/aia.0b013e318181e513.

[37] Brüning, Franziska, Sara B. Noya, Tanja Bange, Stella Koutsouli, Jan D. Rudolph, Shiva K. Tyagarajan, Jürgen Cox, Matthias Mann, Steven A. Brown, and Maria S. Robles. 2019. “Sleep-Wake Cycles Drive Daily Dynamics of Synaptic Phosphorylation.” Science 366 (6462). https://doi.org/10.1126/science.aav3617.

[38] PEKNY, T., D. ANDERSSON, U. WILHELMSSON, M. PEKNA, and M. PEKNY. 2014. “Short General Anaesthesia Induces Prolonged Changes in Gene Expression in the Mouse Hippocampus.” Acta Anaesthesiologica Scandinavica 58 (9): 1127–33. https://doi.org/10.1111/aas.12369.

[39] Ross, F.M, P Gwyn, D Spanswick, and S.N Davies. 2000. “Carbenoxolone Depresses Spontaneous Epileptiform Activity in the CA1 Region of Rat Hippocampal Slices.” Neuroscience 100 (4): 789–96.https://doi.org/10.1016/s0306-4522(00)00346-8.

[40] Yuste, Rafael, Alejandro Peinado, and Lawrence C. Katz. 1992. “Neuronal Domains in Developing Neocortex.” Science 257 (5070): 665–69. https://doi.org/10.1126/science.1496379.

[41] Vandael, David, Yuji Okamoto, and Peter Jonas. 2021. “Transsynaptic Modulation of Presynaptic Short-Term Plasticity in Hippocampal Mossy Fiber Synapses.” Nature Communications 12 (1). https://doi.org/10.1038/s41467-021-23153-5.

[42] Goodman, Corey S., and Carla J. Shatz. 1993. “Developmental Mechanisms That Generate Precise Patterns of Neuronal Connectivity.” Cell 72 (January): 77–98. https://doi.org/10.1016/s0092-8674(05)80030-3.

[43] Wang, Mengni, David J. Foster, and Brad E. Pfeiffer. 2020. “Alternating Sequences of Future and Past Behavior Encoded within Hippocampal Theta Oscillations.” Science 370 (6513): 247–50. https://doi.org/10.1126/science.abb4151.

[44] Kim, Jaekyung, Tanuj Gulati, and Karunesh Ganguly. 2019. “Competing Roles of Slow Oscillations and Delta Waves in Memory Consolidation versus Forgetting.” Cell 179 (2): 514-526.e13. https://doi.org/10.1016/j.cell.2019.08.040.

[45] Farooq, U., and G. Dragoi. 2019. “Emergence of Preconfigured and Plastic Time-Compressed Sequences in Early Postnatal Development.” Science 363 (6423): 168–73. https://doi.org/10.1126/science.aav0502.

[46] Fernández-Ruiz, Antonio, Azahara Oliva, Eliezyer Fermino de Oliveira, Florbela Rocha-Almeida, David Tingley, and György Buzsáki. 2019. “Long-Duration Hippocampal Sharp Wave Ripples Improve Memory.” Science 364 (6445): 1082–86. https://doi.org/10.1126/science.aax0758.

[47] Hara, Koji, and R. Adron Harris. 2002. “The Anesthetic Mechanism of Urethane: The Effects on Neurotransmitter-Gated Ion Channels.” Anesthesia & Analgesia 94 (2): 313–18. https://doi.org/10.1097/00000539-200202000-00015.

[48] Feng, X., J.J. Liu, X. Zhou, F.H. Song, X.Y. Yang, X.S. Chen, W.Q. Huang, L.H. Zhou, and J.H. Ye. 2012. “Single Sevoflurane Exposure Decreases Neuronal Nitric Oxide Synthase Levels in the Hippocampus of Developing Rats.” British Journal of Anaesthesia 109 (2): 225–33. https://doi.org/10.1093/bja/aes121.

[49] Vahle-Hinz, C, and O Detsch. 2002. “What Can in Vivo Electrophysiology in Animal Models Tell Us about Mechanisms of Anaesthesia?” British Journal of Anaesthesia 89 (1): 123–42. https://doi.org/10.1093/bja/aef166.

[50] Hentschke, H., A. Raz, B.M. Krause, C.A. Murphy, and M.I. Banks. 2017. “Disruption of Cortical Network Activity by the General Anaesthetic Isoflurane.” British Journal of Anaesthesia 119 (4): 685–96. https://doi.org/10.1093/bja/aex199.

[51] Bratuša, M. 2022. “LTP and LTD dependence on spontaneous activity in glutamate synapses of the HPC and CTX and the role of anaesthetics.” exchange semester project MEi:CogSci. https://www.academia.edu/80928790/LTP_and_LTD_dependence_on_spontaneous_activity_in_glutamate_synapses_of_the_HPC_and_CTX_and_the_role_of_anaesthetics

Trusted sources and disinformation: studying the limits of science*

Rita Gsenger
Weizenbaum Institute,
Humboldt University Berlin
Berlin, Germany
rita.gsenger@hu-berlin.de

ABSTRACT

During the Covid-19 pandemic, the spread of disinformation became more apparent. Much of that disinformation focused on health-related topics and the current health crisis, often claiming to be scientific information. Trusting scientists became crucial to counter the pandemic effectively as a society; however, science-related disinformation and so-called pseudoscience provided new challenges for societies. These beliefs often overlap with other types of disinformation and conspirational thinking, making them very attractive to human cognition. Twenty semi-structured interviews were done in 2020 to investigate individuals' trust in science, governments, and media. The interviews focused on information sources and the conclusions drawn from the situation to determine how individuals estimate information sources' trustworthiness.

KEYWORDS

Pseudoscience, disinformation, Covid-19, trust

1 DISINFORMATION AND PSEUDOSCIENCE

1.1 Dimensions of disinformation

Disinformation is most commonly defined as false information that is deliberately propagated and distributed [1, 2, 3, 4]. The concept of disinformation includes various dimensions and aspects, which often overlap and influence each other [1, 3]. Kapantai et al. (2020) developed a taxonomical framework to include important types of disinformation, including the motive (profit, ideological, psychological, and unclear), facticity, and verifiability as dimensions. That resulted in eleven kinds of disinformation, including, for instance, conspiracy theories, pseudoscience, hoaxes, trolling, or clickbait. Disinformation can also be partly true to make it more credible [5].

*This abstract is partly based on the author's Master thesis: Digital Literacy and Pseudoscience in Crisis Response. The Case of COVID-19 in Austria (University of Vienna 2021).

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

Disinformation often speaks to human emotions and touches upon controversial or ideologically charged topics [5] as "people have a taste, a predisposition even, for it" [5, p. 57], and virality helps content to be distributed widely [6]. As soon as such topics are concerned, the content is often not that important to be empirically correct and reasonable. However, a social identity is afforded by believing that content is valuable [7]. Making certain beliefs their identity does not only lead to an ignorance of facts (ibid.), but it also enables people to think along ideologically polarised lines with affective disdain for outgroup beliefs [8]. Therefore, believing, for instance, that the earth is flat goes beyond holding a belief but is used to form identities. Around these identities, movements are formed, and, in the case of flat earthers, for instance, people are willing to lose their jobs, friends, and family relations to be part of the group [7].
A kind of disinformation that has become crucial, especially during the Covid-19 pandemic, has been dubbed *pseudoscience*.

1.2 What is pseudoscience?

Distinguishing science from pseudoscience is not a simple endeavour. Some paradigmatic cases might exist where philosophers and scientists agree, but other examples remain undecided or on the fringes of science. Ultimately, the question of defining science and delimitating it from non-science comes down to a fundamental question of philosophy, mainly what knowledge is and how we attain it [9]. Pseudoscience can be understood as a discourse about a specific subject matter, and what is considered pseudoscience, like science, changes [10]. Defining pseudoscience often "involves subjects that are either on the margins or borderlands of science and are not yet proven, or have been disproved, or make claims that sound scientific but in fact have no relationship to science" [11, p. 203]. Several characteristics can be identified to designate the likelihood of adherence to pseudoscientific or non-scientific claims. For example, the language used to describe the phenomenon or research results often indicates the credibility of the reported results or evidence. The excessive use of technical terms and scientifically sounding language, for instance, in press releases, might lead to trust and acceptance of the presented results due to the impression of smart people doing important work. These are, however, not doing a great job in communicating science, and

more importantly, some elements of good scientific practice can be commonly understood [12]. Pseudoscientific theories often use language full of epithets and refer to emotions and religion or use ideological markers.The presentations often include theses and evaluations presented as unequivocal [13]. The method used might not be scientifically sound. For instance, anecdotal evidence and not controlling for other variables, very small or unrepresentative sample sizes in establishing causal relationships, lack of control groups, or blind testing might indicate unsound methods. Moreover, many pseudoscientific studies tend to select parts of their evidence, which allows for a very charitable interpretation of studies to support a predefined conclusion [12]. As there is not one comprehensive definition of pseudoscience, issues fall more or less under its spectrum. I will consider the above-outlined characteristics during my empirical investigation. Pseudoscience and science are historical phenomena that inform the decisions societies make about what is considered the truth. Attempting to define pseudoscience involves making claims about the nature of science. Overall, no methodology has been developed that allows for a general and comprehensive distinction [10], with Popper's principle of falsifiability [14] not solving the problem satisfactorily [15]. In a culture that highly values science, other domains such as religion, politics, or literature are often closely associated with science and seem to borrow scientific language, theories, or methods [11]. Later, theories might be reevaluated and reclassified as science or pseudoscience [10]. Pseudoscientific beliefs are not a marginal phenomenon and influence public policies [9, 16]. For example, during the Covid- 19 pandemic, pseudoscientific explanations for the causes and cures of the virus surged [17]. Therefore, these beliefs, especially in crises, when public policies might be more crucial to follow, and such beliefs could be detrimental to society. However, belief acquisition is not always easy, as human beings are prone to biases and faulty conclusions.

2 STUDYING THE LIMITS OF SCIENCE

2.1 Method and participants

Twenty semi-structured [18] and problem-centred qualitative interviews [19] were conducted in November and December 2020 with Austrian volunteers (N=20, 16 female, age 19-65, SD = 13.8). Interviews were led in German, and the author translated quotes. Interview participants were volunteers. Therefore, the researcher did not have much influence on their gender. However, gender was determined not to be a crucial influence on the study. The discussions included several topics. However, only one part focusing on trust and attitude towards the government, scientists and media is contained in this paper. Some limitations must be outlined when doing qualitative interviews online. Conducting interviews online limits the information transmission compared to real life interviewing face-to-face. The qualitative study was done at a specific moment of the pandemic and thus only reflects participants' attitudes during that time. Furthermore, participants might be hesitant to share pseudoscientific beliefs or denial of science with a researcher. Therefore, no outright questions about such ideas were asked.

2.2 Results: "I don't know what to believe anymore": doubt and trust in times of crises

Participants used various sources of information about the Covid-19 pandemic, including online sources, TV, radio and conversations with friends and family. When asked about the sources participants considered trustworthy or not to provide information about the Covid-19 pandemic, various categories were mentioned, including the media, social media, social contacts, and the government. However, the trust did not seem to be easily acquired or granted among participants. Social aspects were considered influential in attributing trustworthiness (P3, P6). Therefore, a reason to trust a source would be that people from an individual's social circle would also trust it (P3, P6, P19). Furthermore, authenticity and "thinking outside the box" (P13) were considered trustworthy traits of people. Some said they would trust family members, doctors or journalists they knew personally (P2, P11).

Participants trust media if they provide sources with more information about the topic in question (P3), including links to other trusted websites (P4). Furthermore, if the information could be cross-referenced with scientific sources (P14), and if scientists, experts, or studies are included (P20), the trust in media sources is increased. Furthermore, including various opinions was considered a sign of trustworthiness (P6, P1). These opinions permit looking at a subject from multiple viewpoints (P3) and discussions by different people (P3, P14). The content would not be considered trustworthy if a personal opinion were presented as objective truth (P3). Furthermore, the presentation of information in the media and on the internet influences the attribution of trustworthiness. Accordingly, the way people post something, specifically the language (P14), if they write whole sentences and if they explain the context of an article (P4) or if something is not formulated blatantly (P18) and frequently based on emotion (P5) it is considered more trustworthy. However, trust was not attributed without reservation for many participants as they perceived the media as having their agenda (P8, P13) and being prejudiced (P14), but still more trustworthy than social media (P8). On the other hand, some did not consider the media "a source to find out what is really happening" (P13), and one participant mentioned that they "don't believe anything anymore" because "[...] it is not explained what the numbers mean at all or put into a context from which area the numbers come from and how they were created at all" (P16). Social media was not considered trustworthy because a lot of information originates from private individuals (P19). Moreover, assessing the trustworthiness of information on social media is challenging (P1), even though most participants considered some people they were friends with on Facebook trustworthy (P3, P4, P15). Some participants based their trust on intuition and how they felt regarding the media and online information. One participant described it as follows: "When I open that, how does it 'feel' if I move towards a platform, then I read how the information is structured, and I read the first paragraphs, and when something is in there that seems a bit strange to me, then I would get out of there and look it up somewhere else. So, it depends on how it is in a textual sense and how the information lies in front of me"

(P15). More specific descriptions of that feeling included if something seemed "out of touch with reality" (P5), what sounds reasonable (P3), to use one's common sense or if it appears strange (P18). Participants furthermore attributed trustworthiness to sources or information that would confirm a worldview. Accordingly, a participant described that other people with differing worldviews would find different information trustworthy and objective (P6). Additionally, reputation was a source of trustworthiness, especially in the media (P10, P19). Some participants mentioned the government and ministries as trustworthy sources of information. One participant said, "in the last months, I have experienced things where I was not sure in the moment can I trust anyone, and this is now a purely emotional thing because you cannot know anything anyway" (P8). Furthermore, a participant claimed that "somebody is telling me, I cannot go to university anymore, that I cannot see people anymore, who is that somebody who would permit that, who decides about me, that I cannot do that anymore" (P5). Some participants showed understanding of the difficult decisions the government needs to take right now, claiming as they would not want to be in their position or get involved, they would need to comply with measures (P11).

Furthermore, some participants claimed that everybody would need to find their way of dealing with the situation and meet as many people as they would think appropriate (P14), emphasizing the responsibility of individuals (P5, P14). Many participants mentioned the adverse consequences of the measures. Some agreed that these consequences, including the dangers of a lockdown (P1), were not discussed enough (P1, P2, P16, P17). Some were worried about restricting civil rights during the lockdown and possible dangers to democracy (P9, P1), claiming that the government could not implement a curfew as it violated human rights (P1). Participants wished that people were given more credit (P20, P13), which included telling them to take care of their immune system and take vitamins (P13). Another participant would have wanted different perspectives on the transformation happening in 2020, as communication is changing and more telepathy will be possible due to that change (P5). According to participant 13, not discussing alternative ways of handling the pandemic can be attributed to international pressure (P13).

The plurality of opinions is generally valued highly among participants as it is essential that everybody can share their standpoint and how they arrive at their conclusions because everybody has a good reason to think as they do (P2, P1). However, according to some participants, not all opinions and standpoints were listened to somewhat during the pandemic. For instance, the questions "masks yes or no these questions are not allowed to be asked because we are being beaten down by all these numbers" (P13), and they should listen to people who have other methods (P14). Some observed that a division between opinions and people was taking place in the general society. In that regard, only two contrary camps seemed to exist, and only to "be for Corona or against, a middle course or a differentiated account was not possible" (P17). That means, participants were worried that a nuanced debate about issues regarding the pandemic was more difficult.

Moreover, participants observed how people changed and suddenly believed entirely different things (P16, P2, P8). According to participants, everybody should state their opinion but has the responsibility to do it respectfully (P15). An individual's history is crucial to consider to make respectful interaction easier (P2). Participants elaborated more in detail on how they formed their own opinions about the measures, the communication, and the pandemic in general. Some attributed the decision to believe the information from a source to intuition (P12) or if it seems strange (P20), as highlighted previously in section 6.4.1. Furthermore, they highlighted the influence of social factors, such as the influence of people they would talk to (P5), for instance, in their workplace or people who had the illness (P14), even if they disagreed with them (P5). They would like to discuss these issues among their circle of friends as some would be more active and critical and might introduce other perspectives (P14). If some- thing seems strange, however, they would try to find other opinions (P20), and online they would follow links from friends (P9) or try and consume contrary opinions (P5). Overall, participants would form their opinions by combining various other opinions (P8), questioning their worldview, and staying open for new information (P14), and reflect on it (P5). Participants highlighted difficulties with opinion formation about the Covid-19 pandemic, as one participant summarized: "I believe a big problem is that there are so many people, where it is claimed, ok, I am a doctor in that area, and I say this and that. And the doctor then says that and you don't know, is that person really a doctor, do they really know about that. I mean, probably they are doctors but did they actually engage with that issue, or are they just saying anything? There are so many doctors that have different areas of expertise and, of course, various experiences and a different level of knowledge, so you don't know where the information is coming from" (P16).

3 DISCUSSION AND CONCLUSION

Participants highlighted some specific topics as instilling the most significant doubts about trustworthy sources considering the Covid-19 pandemic. Science was considered a trusted source, but various indications showed that participants had significant doubts regarding the scientific consensus about Covid-19, for instance, that they would not know what it really was (P1) and that it was the flu, which is unpleasant but not particularly dangerous (P13). Tests to determine infections were doubted in their validity and efficacy (P16) and are considered inaccurate (P13). Even though the interviews were led before Covid-19 vaccinations were widely available, mandatory vaccinations were already a big concern for some participants, which are thought to change society (P11) and should be well prepared to take people's fears about the vaccinations (P20) as chaos might ensue if vaccinations become mandatory (P12). Furthermore, the topic of not being told everything was present regarding the issue of vaccinations. Participants worried about what would happen to the Austrian culture and the country if vaccinations were mandatory (P11). An electronic compulsory vaccination certificate was mentioned as a source of worry for a functioning peaceful democracy (P9). Various conspirational elements seemed to be present during interviews. For example, some participants were worried about democracy and the rule of law in

Austria. One mentioned that a friend who is a doctor told them that now with the pandemic, the government can achieve things they could never have done without the pandemic (P13), which happens behind the scenes and might endanger our democracy (P9). In that regard, with the climate of fear, the government "is trying out how far it can tighten the thumbscrews" (P9), and the government lies, meaning something changed in a significant and sustainable way without people knowing (P9). In that regard, a participant stated the government was "catholic, dishonest and tendentious" (P11) and that it "actually does not have a plan [...] only false numbers, false facts, false something" (P17). Some doubted the democratic nature of the situation, as not everything is communicated (P18) in our "so-called democracy" (P12). One participant summarised the situation as follows:

"Honestly, it is a bit authoritarian because the information comes from above. Kurz [the Austrian chancellor] speaks from the microphone, and everyone listens, sits in front of the TV or channel, listens to him, and then it is done. I don't think that's democracy. How that has now developed individually that certain events were then possible, these self-initiatives that have then taken place in conformity with measures I find again thanks to people with whom I live together that we are democratic. These two perspectives in my social environment where you get together and ask if it's okay if you can hug someone or sing with each other even though choirs are not allowed to practice so that in agreement with the others, of course, works because we are not in the snitch system and the Biedermeier maybe it seems so but not quite." (P5).

All participants seemed to struggle with doubts regarding handling the pandemic in 2020. These doubts focused either on the government or on science. The media seemed to be the most trusted. However, some would argue that they would only report uncritically. If doubts focused on the government, they seemed to lean more towards a conspirational mindset. Doubts regarding scientific consensus about Covid-19 are mostly deemed to adhere to pseudoscientific beliefs such as Covid-19 is the same as the flu or tests/masks do not work. However, a mix of pseudoscience (vaccines do not work) and conspirational tendencies (there is some more extensive agenda) could be observed when discussing vaccinations. In conclusion, the frequency of social media use and the content consumed should not be overestimated, as an individuals' immediate social environment (i.e. friends and family) seems to have a more significant influence on their beliefs.

REFERENCES

[1] J. Möller, M. Hameleers and F. Ferreau, "Verschiedene Formen von Desinformation und ihre Verbreitung aus kommunikationswissenschaftlicher und rechtswissenschaftlicher Perpektive. Ein Gutachten im Auftrag der Gremienvorsitzendenkonferenz der Landesmedienanstalten (GVK).," die medeinanstalten - ALM GbR, Berlin, 2020.

[2] E. Humprecht, F. Esser and P. Van Aelst, "Resilience to Online Disinformation: A Framework for Cross-National Comparative Research," *The INternational Journal of Press/Politics,* vol. 25, no. 3, pp. 493-516, 2020.

[3] E. Kapantai, A. Christopoulou, C. Berberidis and V. Peristeras, "A Systematic Literature Review on Disinformation: Toward a Unified Taxonomical Framework," *New Media & Society,* vol. 23, no. 5, pp. 1301-1326, 2021.

[4] K. Shu, S. Wang, D. Lee and H. Liu, "Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements," in *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, Cham, Springer International Publishing, 2020, pp. 1-20.

[5] N. O'Shaughnessy, "From Disinformation to Fake News: Forwards into the Past," in *The Sage Handbook of Propaganda*, Thousand Oaks (CA), Sage, 2019, pp. 55-71.

[6] S. Vosoughi, D. Roy and S. Aral, "The Spread of True and False News Online," *Science,* vol. 359, no. 6380, pp. 1146-1151, 2018.

[7] L. McIntyre, How to Talk to a Science Denier. Conversations with Flat Earthers, Climate Deniers, and Others Who Defy Reason, Cambridge (MA): MIT Press, 2021.

[8] L. Mason, "Ideologues without Issues: Polarizing Consequences of Ideological Identities," *Public Opinion Quarterly,* vol. 82, no. S1, pp. 866-887, 2018.

[9] M. Pigliucci and M. Boudry, "Introduction: Why the Demarcation Problem Matters," in *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*, London, The University of Chicao Press, 2013, pp. 1-9.

[10] D. K. Hecht, "Pseudoscience and the Pursuit of Truth," in *Pseudoscience. The Conspiracy Against Science*, Cambridge (MA), MIT Press, 2018, pp. 3-20.

[11] M. Shermer, "Science and Pseudoscience. The Difference in Practice and the Difference It Makes," in *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem.*, London, The University of Chicago Press, 2013, pp. 203-223.

[12] C. W. Lack and J. Rousseau, Critical Thinking, Science, and Pseudoscience, New York: Springer, 2016.

[13] M. Szynkiewicz, "May you live in interesting times. Science vs. pseudoscience in the era of the internet," *Ethics in Progress,* vol. 11, no. 1, pp. 85-98, 2020.

[14] K. Popper, The logic of scientific discovery, London and New York: Routledge, 2002.

[15] M. Pigliucci, "The Demarcation Problem. A (Belated) Response to Laudan.," in *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*, London, The University of Chicago Press, 2013, pp. 9-29.

[16] J. H. Taylor, R. A. Eve and F. B. Harrold, "Why creationists don't go to psychic fairs: differential sources of pseudoscientific beliefs," *Skeptical Inquirer,* vol. 19, no. 6, pp. 23-28, 1995.

[17] T. T. Desta and T. Mulugeta, "Living with COVID-19 triggered pseudoscience and conspiracies," *International journal of Public Health,* vol. 65, no. 6, pp. 713-714, 2020.

[18] G. D. McCracken, The long interview, Newbury Park (CA): Sage, 1988.

[19] U. Flick, An introduction to qualitative research, London: Sage, 2009.

[20] R. Armitage and C. Vaccari, "Misinformation and Disinfomration," in *The Routledge Companion to Media Disinformation and Populism*, London and New York, Routledge, 2021, pp. 38-49.

[21] P. Carrillo-Santisteve and P. Lopalco, "Measles still spread in Europe: Who is responsible for the failure to vaccinate?," *Clinical Microbiology and Infection,* vol. 18, no. 5, pp. 50-56, 2012.

[22] J. M. Berman, Anti vaxxers: How to Challenge a Misinformed Movement, Cambridge (MA): Teh MIT Press, 2020.

[23] D. Kahneman, "A perspective on judgement and choice: Mapping bounded rationality," *American Psychologist,* vol. 58, no. 9, pp. 697-720, 2003.

[24] H. A. Simon, "Rational Decision Making in Business Organizations," *The American Economic Review,* vol. 69, no. 4, pp. 493-513, 1979.

[25] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," in *Judgment under uncertainty: Heuritsics and biases*, Cambridge (MA), Cambridge Unviersity Press, 1982, pp. 3-23.

[26] A. Tversky, "Elimination by Aspects: A Theory of Choice," *Psychological Review,,* vol. 79, no. 4, pp. 281-299, 1972.

Opacity and understanding in artificial neural networks: a philosophical perspective

Martin Justin
martin1123581321@gmail.com
Faculty of Arts, University of Ljubljana
Ljubljana, Slovenia

ABSTRACT

In the paper, I review some of the emerging philosophical literature on the problem of using artificial neural networks (ANNs) and deep learning in science. Specifically, I focus on the problem of opacity in such systems and argue that although using deep neural networks in cognitive science can produce better results, it can also acts as a barrier to gaining new understanding of cognitive processes.

KEYWORDS

explanation, understanding, scientific discovery, artificial neural networks, black-box problem

1 INTRODUCTION

Early on in their inception, connectionist approaches in cognitive science faced challenges from proponents of competing approaches. One of the leading theorists of the classical symbolic approach, J. Fodor and Z. Pylyshyn [7], for example, argued that connectionism could not account for four essential properties of cognition – i.e., productivity, systematicity, compositionality, and coherence – and thus was not a sufficient theory of the mind. We now have good reasons to believe that their argument does not hold [11]. Indeed, in their demonstration of the supposed inadequacy of connectionist models, Fodor and Pylyshyn only considered very simple models with local representations. But it turns out that more complex models with distributed representations can satisfactorily solve the explanatory task. Contrary to what Fodor and Pylyshyn claimed, we can therefore show that even connectionist cognitive models are powerful enough to exhibit the required properties.

Fast forward forty or so years in the future, scientists using artificial neural networks (ANNs) and deep learning to study cognitive functions now face a different problem. One of the key advantages of present day ANNs that use deep learning is their increased complexity and depth [4]. But because of their increased complexity, such systems can become opaque in a way that even the researchers developing them do not understand some key aspects of how they work [10]. Present day ANNs can thus be used to model cognitive functions much more successfully than before, but because of their opaqueness, it is unclear what new insights such successes are generating [5].[1] If researchers in the 1980s talked about an explanatory task, the problem scientists' face today could be called an explanatory barrier.

In this paper, I will review some of the emerging philosophical literature on the problem of using ANNs in science. First, I will briefly introduce the problem of opaqueness or the so-called black-box problem. Then, I will present a paper by Erasmus et al. [6] which provides a detailed analysis of the notions of explanation and understanding that are central to thinking about the problem. After that, I will present Florian J. Boge's [3] argument that we can talk about two distinct dimensions of opacity in ANNs. In the last section, Mazviita Chimirmuuta's [5] argument about the implications of the trade-off between predictive accuracy and opacity for research in computational neuroscience will be presented.

2 BLACK BOX PROBLEM

Let us first turn to the problem of opaqueness. Authors of one of the review papers [10] from the field of explainable AI (XAI) note that the "predictive accuracy [of machine learning systems] has often been achieved through increased model complexity." This increased complexity, "combined with the fact that vast amounts of data are used to train and develops such complex systems" has inherently reduced researches' ability to "explain the inner workings and mechanisms" of these systems. As a result, "the rationale behind decisions [of these systems] becomes quite hard to understand and, therefore, their predictions hard to interpret." Therefore, they say that "there is clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions." The authors of anther review paper [1] reached a similar conclusion: "Indeed, there are algorithms that are more interpretable than others are, and there is often a tradeoff between accuracy and interpretability: the most accurate AI/ML models usually are not very explainable (for example, deep neural nets, boosted trees, random forests, and support vector machines), and the most interpretable models usually are less accurate (for example, linear or logistic regression)."

Authors of [10] thus distinguish between "black-box" models, which have state-of-the-art performance but are opaque, and "white-box" or "glass-box" models, which are more easily interpretable, but not as powerful. In her paper, Chirmuuta [5] also specifies which aspects of deep neural networks suffer from opaqueness. She argues that scientists have a good understanding of "internal architecture and workings" of the systems, i.e., they know the activation values of the units, the learning rule, the depth of the network and the connectivity between the layers. But they do not know exactly how an already trained network arrives at a prediction or classification.

[1] In contrast to this, Sullivan [14] argues that the problem of contemporary ANNs is not their opacity, but "a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon." But see Räz and Breisbart [13] for an argument that her point rests on a weak and thus undesirable notion of understanding.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

3 EXPLANATION AND UNDERSTANDING

The black box problem or the problem of opaqueness has resulted in increased attention to research in explainable AI. But one salient feature of the literature on explainable AI is the imprecise or even interchangeable use of the concepts of explainability, intelligibility and interpretability. This is also recognized by the researchers themselves. For example, the authors [10] observe that there is "no concrete mathematical definition of the concepts of explainability and interpretability." Nevertheless, they make a conceptual distinction between these two terms. Interpretability, on the one hand, is understood in connection to the ability of researchers to intuitively understand the relationship between inputs and outputs of a system. Explainability, on the other hand, is understood in relation to the ability to understand the inner workings of a system. In contrast, authors of a different similar study [9] define explainability as possibility to provide a satisfactory answer to the "why" question regarding the functioning of a system. They also make a difference between two levels of explanation, connected to two different questions scientists can ask about a system. Namely, "why does this particular input lead to that particular output?" i.e., a question about the relationship between inputs and outputs, and "what information does the network contain?" i.e., a question about the internal workings of a system.

In their paper, Erasmus et al. [6] point to this shortcoming of the literature on explainable AI and argue that this imprecise use of the terms leads to a misunderstanding of the trade-off between performance and explainability of AI systems. Their argument proceeds in three steps. First, they offer a more precise analysis of the notions of explanation and understanding. Second, they show that the increased complexity of systems affects their undersandability rather than their explainability.[2] And third, they offer a typology of possible explanatory methods that could also increase the intelligibility of systems. Here, I will be interested mainly in the first and the second step. Therefore, in the remainder of this section I will first present (a) their definition of explanation, (b) their arguments that the possibility of explanation is independent of the complexity of the phenomenon itself, and (c) the argument that ANNs can be explained. Then I will present (d) their definition of understanding and (e) their argument that complexity affects the ability to understand.

Let us start with (a). In defining the notion of explanation, Erasmus et al. [6] draw on a longer tradition in philosophy of science which holds that explanation consists of three elements: (1) the *explanandum*, i.e., what we want to explain, (2) the *explanans*, i.e., with what we are explaining, and (3) *the process of explanation.* Different models of explanations differ in one or more of these elements. Four such models feature prominently in the literature. (I) Deductive Nomological model, in which the explanans includes empirical content plus a law-like preposition, and the process of explanation takes the form of deductive reasoning. (II) Inductive Statistical model, in which the explanans includes a statistical law about behavior of the variables, and the process of explanation takes the form of inductive or probabilistic reasoning. (III) Causal Mechanical model which aims to show "how the explanandum fits into the causal structure of the world", and thus involves giving information about the causal process and the causal interaction that leads to the emergence of the explanandum. (IV) New Mechanist model which takes as explanans the entities and their activities that are responsible for the emergence of the explanandum.[3]

Erasmus et al. [6] then argue (b) that the increased complexity of the phenomenon we are trying to explain (or of the concepts and data we use to explain it) does not affect our ability to offer an explanation for the phenomenon. And (c) that deep neural networks can be explained in all four of the ways described above. The argument for (b) is quite simple: Deductive Nomological explanation, for example, requires only that the explanans contains a law, and that the process of explanation takes the form of deductive reasoning. It does not matter how complex the two elements are. Thus, an explanation that contains a more complex explanans and requires more complex reasoning may be less desirable, but it is no less an explanation.

The argument for (c) is a bit more technical. To demonstrate this point, the authors provide an example of an explanation of how an ANN, trained to identify dense breast tissue on X-ray images, classify these images [6]. Let us see how a Deductive Nomological explanations of such ANN could work. As the empirical content of the explanans, we could use all the information about the activation values of the individual units in the network and about the weights between them, as well as the numerical values of the input data. We could also form a law-like proposition of the form "outputs with such and such numerical value are classified as such and such." In this way, the explanandum, i.e., the classification of the photograph F into the class r, would be explained using an explanans consisting of a law-like preposition and empirical content. In other words, we would have a Deductive Nomological explanation. Although the arguments for (b) and (c) were presented only for the case of Deductive Nomological explanation, authors argue they apply *mutatis mutandis* to other models of explanation as well.

Let us now turn to (d), the definition of understanding. As Erasmus et al. [6] point out, authors who study understanding do not, of course, entirely agree on its exact definition, but they commonly observe that, while explanation is necessary for understanding, it is not sufficient for it. So to gain understanding of a phenomenon, some other conditions besides having an explanation must be met. There are several candidates for these additional conditions in the literature, but, as Erasmus et al. argue, they all have in common that they are "psychological traits of the user of the explanation." One such condition is the criterion of intelligibility. It states that a theory T is intelligible to a scientist in a context C if the scientist is able to recognize the qualitatively distinct consequences of T without doing the exact calculations [5, 6].[4] Given this, it is obvious that increased complexity of an explanation or a phenomenon makes it less intelligible and thus less understandable. Thus, it can be concluded that (e) complexity affects the ability to understand.

4 TWO DIMENSIONS OF OPACITY

Erasmus et al. [6] argue that while the workings of deep neural networks are explainable, they are often not understandable for

[2] See Beisbart and Räz [2] for a critique of this point. They say that "the distinction that Erasmus et al. draw between interpretability and explainability in this way seems rather stipulative." In contrast, they argue that we should use these terms as synonyms. Nevertheless, I think that Erasmus et al. [6] point to an important and well established concetpual distinction between these two terms which should not be so easily dismissed.

[3] Woodword and Ross [17] present a slightly different typology. In particular, they add Salmon's statistical relevance model and pragmatic models of explanation.

[4] Chirimuuta [5] also lists four properties of a theory (or an explanation) that affect its intelligibility. Those are: (1) the possibility of visualization, (2) the simplicity of included theoretical assumptions, (3) the linearity of mathematical operations, and (4) functional transparency.

human users. In other words, they conclude that we should talk about a trade-off between the performance of AI systems and their understandability or intelligibility, not their explainability. Nevertheless, they seem to overlook another important aspect of the trade-off. As it is apparent from the definitions of explainability and understandability in Gilpin et al. [9] and Linardatos et al. [10], there seem to be different ways in which ANNs can be opaque to humans.

This point is explicated and extended upon by Boge [3]. In his paper, he presents the following three theses: (1) deep neural networks are instrumental, and their instrumentality is distinct from that of other mathematical models; (2) deep neural networks are opaque in two different ways; and (3) the combination of (1) and (2) means that in the future, we may not be able to understand potential new discoveries made by deep neural networks. In the rest of this section, I will be primarily interested in (2).

Boge [3] begins his exposition of the two aspects of opacity by defining opacity. He defines it as follows: "a process P is epistemologically opaque to a subject X at time t if and only if X does not know all the epistemically relevant elements of the process P at time t." He then distinguishes between two aspects of the opacity of deep neural networks. First, he describes h-opacity. It concerns the operation of a system: a system is h-opaque if it is the process of its operation that is not not intelligible to its human users. This is the opacity that results from the complexity of deep neural networks and hiders the understanding of the connection between input and output data. But as Boge notes, this type of opacity is not qualitatively different from, say, the opacity of other complex computational simulations, e.g., climate simulations. He therefore identifies another aspect of opacity that is specific to deep neural networks. This is w-opacity, which concerns the representational content of the system (what was learned). According to Boge, in deep neural networks, not just the process that takes a neural network from an input to an output, but also the properties of the input data that guide this process are opaque.

This difference is important as it points to a specific problem that the use of deep neural networks introduces to scientific research. H-opacity only hinders the understanding of the computational model itself, as it prevents researchers from seeing how it gets from input to output data. Such opacity can thus be problematic form an ethical point of view, as it makes it harder to justify the decisions made on the basis of a recommendation by an AI system. In contrast to this, w-opacity reduces the potential of deep neural networks to bring new understanding to the processes studied by the scientists. Even in the case where promising results would suggest that an ANN represents a given problem space in a better way than existing theories, w-opacity would leave this representation incomprehensible to scientists. Thus, w-opacity has important implications for the use of neural networks in scientific research.

5 PREDICTION VERSUS UNDERSTANDING

The implications of w-opacity for research in computational neuroscience are convincingly presented by Chirimuuta [5]. In this section, I will summarize her findings. I will do this in the following steps: (a) first, I will briefly outline the research program of computational neuroscience; (b) then, I will present examples of two studies from the field, one in which scientists approached their problem using a transparent mathematical model, and another in which they approached a very similar problem using a w-opaque deep neural network; (c) finally, I will present Chirimuuta's version of the trade-off between performance and understanding that arises when using ANNs in science.

Let us start with (a). Chirimuuta [5] defines computational neuroscience as "a tradition of research that builds mathematical models of neurons' response profiles, aiming both at predictive accuracy and at theoretical understanding of the computations performed by classes of neurons." It is based on the assumption that information about the external world is 'encoded' in the electrical and chemical signals of the neurons. It attempts to solve the so-called 'decoding problem', i.e., it tries to find a mathematical function that could successfully link neuron spikes to outside information. Specifically, according to Chirimuuta, scientists try to devise a theory of how neurons encode information about the outside world and then write a program, called an encoder, that performs the translation operation between the stimuli and the neural activity.

Thus, as Chirimuuta [5] points out, computational neuroscience pursues two separate epistemic goals. On the one hand, it aims at accurately predicting the relations between neural activity and external stimuli (e.g., to predict how neurons will fire if we show a picture of a square to a primate). On the other hand, it tries to understand how this translation takes place. Chirimuuta thus argues that in the past, when even very simple linear models have proved surprisingly accurate in certain contexts, there has been a convergence between these two goals. However, with the development of deep neural networks, which are much more accurate but w-opaque, these two goals started to diverge.

Chirimuuta [5] presents two examples of such divergence, one from modeling the functioning of the motor cortex and another from modeling the visual perception system. I will limit my presentation to the former, i.e., to her comparison between two studies that tried to model motor cortex activity, Georgopoulos et al. [8] and Sussillo et al. [15]. In both of these two experiments, researchers measured the activity of individual neurons in non-human primates while the primates were performing given tasks. Georgopoulos et al. [8] present an experiment in which a monkey was surrounded by eight buttons, with another button straight ahead. In the experiment, first the button in front of the monkey lit up. After the monkey held it for one second, one of the other eight buttons lit up, and the monkey had to press it with the same hand. Meanwhile, the scientists measured the activity of a population of neurons in her motor cortex, and tried to establish a correlation between this activity and the direction of her arm movement. They did this by simply converting the activity of a neuron into a vector in three-dimensional space according to a formula they devised, and then summing the vectors of the individual neuronal cells to obtain one vector that represented the whole neuron population. They found out that the direction of this vector quite closely matched the direction of arm movement. Because of the fairly simple math they used, their model was completely transparent. In addition, the researchers themselves determined which information about the neural activity is important and should be used to calculate the movement vector. The accuracy achieved by the model can thus be seen as a partial confirmation that these features of neural activity are indeed important for directing arm movement.

The experiment reported by Sussillo et al. [15] is a bit different. They also had non-human primates, this time two, implanted with electrodes that measured the activity of individual neurons in their motor cortex. But the monkeys did not press buttons; rather, they had to move a cursor on a screen from a central

position to a marked position in one of the corners of the screen. Each monkey performed three series of experiments. First, they moved the cursor by moving their hand. Then, they moved the cursor using a brain-machine interface (BMI) that used a encoder, based on a mathematical model, similar to the one described in the previous example. In the last series, they used a BMI that encoded information using a trained neural network. Each monkey performed each of the three experiments hundreds of times. The researchers found that using this ANN based encoder significanty improved monkey's performance *vis-à-vis* the older model. This suggests that the BMI with an ANN was more successful in translating between neuronal activation and information about the outside world. We can thus assume that the ANN either approximated the mathematical function linking neuronal activation and external stimuli more accurately or it 'discovered' new properties of the input data that play an important role in the translation. But the ANN used was both h-opaque and w-opaque so despite its improved performance, it did not provide scientists with a better understanding of how the motor cortex works.

Turning to point (c), it should now be clear what Chirimuuta [5] is getting at when she says that the use of ANNs creates a divergence between the goals of predictive accuracy and understanding of neurological processes. Chirimuuta calls this divergence a trade-off between understanding and accuracy. The trade-off arises because a problem can either be tackled with models that are not the most accurate, but can be interpreted and can thus provide us with new understanding of the problem. Or it can be tackled with ANNs, which, although they achieve greater accuracy, are opaque and therefore do not bring new understanding to scientists.

In addition to presenting a dilemma from the point of view of epistemic goals of science, the trade-off also has some practical implications. For example, increasing reliance on ANNs to analyze data may be linked to issues related to trust in scientific findings. In his influential analysis of epistemic trust, T. Wilholt [16] argued that the reliance between the members of a scientific community is based on the "assumption that the results [the scientists are] relying upon were arrived at by means of professional methods suitably employed". Given the opacity of ANNs using deep learning, this assumption might be difficult to test. Furthermore, some researchers speculated that "hyping" scientific results (especially in the more directly applicative fields, such as biotechnology) can ultimately result in a loss of public trust in science. Although this connection between hype and public trust have not yet been empirically established [12], it is not hard to see how focusing on predictive accuracy, rather than understanding, could further increase the unwanted hype surrounding scientific research.

6 CONCLUSION

In this paper, I reviewed some of the emerging literature on the epistemological aspects of the problem of opaqueness in deep neural networks. First, I used Erasmus et al. [6] to point out that we need to distinguish between explainability and understandability of AI systems. I also presented their argument that the increasing complexity of these systems has a particular impact on our ability to understand them, not on their inherent explanaibility. Then, with the help of Boge [3], I distinguished between two dimensions of opacity of these systems. Finally, following Chirimuuta [5], I presented this problem using a concrete example of two studies in computational neuroscience. In this way, I have shown in more detail what philosophers mean when they talk about the trade-off between performance and intelligibility (or understandability) of AI systems in science.

ACKNOWLEDGMENTS

I would like to thank Olga Markič for her useful suggestions and encouragement. I would also like to thank Nejc for his help with the more technical aspects of the literature.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[2] Claus Beisbart and Tim Räz. 2022. Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, 17, 6, (June 2022). DOI: 10.1111/phc3.12830.

[3] Florian J. Boge. 2021. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, (Sept. 2021). DOI: 10.1007/s11023-021-09569-4.

[4] Cameron Buckner. 2019. Deep learning: A philosophical introduction. *Philosophy Compass*, 14, 10, (Oct. 2019). DOI: 10.1111/phc3.12625.

[5] M. Chirimuuta. 2021. Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, 199, 1-2, (Dec. 2021), 767–790. DOI: 10.1007/s11229-020-02713-0.

[6] Adrian Erasmus, Tyler D. P. Brunet, and Eyal Fisher. 2021. What is Interpretability? *Philosophy & Technology*, 34, 4, (Dec. 2021), 833–862. DOI: 10.1007/s13347-020-00435-2.

[7] Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 1-2, (Mar. 1988), 3–71. DOI: 10.1016/0010-0277(88)90031-5.

[8] Apostolos P. Georgopoulos, Andrew B. Schwartz, and Ronald E. Kettner. 1986. Neuronal Population Coding of Movement Direction. *Science*, 233, 4771, 1416–1419.

[9] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, (Oct. 2018), 80–89. ISBN: 978-1-5386-5090-5. DOI: 10.1109/DSAA.2018.00018.

[10] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23, 1, (Dec. 2020), 18. DOI: 10.3390/e23010018.

[11] Olga Markič. 2011. *Kognitivna znanost: filozofska vprašanja*. Aristej, Maribor.

[12] Zubin Master and David B. Resnik. 2013. Hype and Public Trust in Science. *Science and Engineering Ethics*, 19, 2, (June 2013), 321–335. DOI: 10.1007/s11948-011-9327-6.

[13] Tim Räz and Claus Beisbart. 2022. The Importance of Understanding Deep Learning. *Erkenntnis*, (Aug. 2022). DOI: 10.1007/s10670-022-00605-y.

[14] Emily Sullivan. 2022. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73, 1, (Mar. 2022), 109–133. DOI: 10.1093/bjps/axz035.

[15] David Sussillo, Paul Nuyujukian, Joline M Fan, Jonathan C Kao, Sergey D Stavisky, Stephen Ryu, and Krishna Shenoy. 2012. A recurrent neural network for closed-loop intracortical brain–machine interface decoders. *Journal of Neural Engineering*, 9, 2, (Apr. 2012), 026027. DOI: 10.1088/1741-2560/9/2/026027.

[16] Torsten Wilholt. 2013. Epistemic Trust in Science. *The British Journal for the Philosophy of Science*, 64, 2, (June 2013), 233–253. DOI: 10.1093/bjps/axs007.

[17] James Woodward and Lauren Ross. 2021. Scientific Explanation. (2021). Retrieved Jan. 29, 2022 from https://plato.stanford.edu/entries/scientific-explanation/.

Politizirana znanost in zaupanje v znanost kot politična uniforma

Politicized science and trust in science as a political uniform

Jar Žiga Marušič†
Oddelek za psihologijo

Famnit, Univerza na Primorskem
Koper, Slovenija
jar.marusic@famnit.upr.si

POVZETEK

Zaupanje v znanost je dandanes, sploh po dveh letih pandemije Covida-19, posebej družbeno relevanten problem. Izraz pa je nekoliko dvoumen, saj lahko znanost razumemo na več načinov, med drugim kot raziskovalni proces in kot institucije, na katerih se ta proces odvija. Zato je izraz zaupanje v znanost lebdeči označevalec, oznaka brez jasnega referentnega objekta.. Težava lebdečih označevalcev se pokaže, ko postanejo tarča politizacije. V tem primeru zaradi nejasnosti semantičnega pomena sociopolitične konotacije izraza postanejo njegov primarni pomen. V politizirani znanosti bi zato "zaupati v znanost" v resnici pomenilo podpirati obstoječi politični režim, izražanje tega zaupanja (ali njegovega pomanjkanja) pa bi služilo kot politična uniforma, ki izraža pripadnost enemu ali drugemu političnemu polu. V prispevku analiziram znanstveni diskurz zadnjih dveh let z namenom ugotavljanja, kaj je bil v tem obdobju družbeni pomen zaupanja v znanost – podpora procesa znanstvenega raziskovanja ali izraz politične konformnosti?

KLJUČNE BESEDE

Politizacija znanosti, Covid-19, zaupanje v znanost, socialno presojanje, psihološka inokulacija

ABSTRACT

Trust in science is especially relevant in today's society, given that we are living in the wake of the 2-year Covid-19 pandemic. The term itself is somewhat vague, as science has multiple definitions, mainly the process of scientific research as well as the institutions that engage in said process. Thus, trust in science is a floating signifier, a label without a clear referent. Such labels can be problematic if targeted by politicization. The vagueness of the floating signifier's semantic meaning allows the socio-political connotations to acquire primacy. In times of politicized science, "trusting in science" would then actually mean to endorse the established political regime. As for actions that signal this trust (or lack thereof), they would act as a political uniform – an expression of political allegiance to one's chosen side. This article analyses the state of scientific discourse during the pandemic, with the goal of establishing the precise meaning of trust in science in practice – endorsement of the process of scientific research, or an expression of political conformity?

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

KEYWORDS

Politicization of science, Covid-19, trust in science, social reasoning, psychological inoculation

1 Kaj pomeni zaupati v znanost?

Vse večjo relevantnost pojma "zaupanja v znanost" v sodobni družbi lahko jemljemo kot posledico *poznanstvenjenja družbe in družbenih praks* – vse večjega soodvisnost znanstveno-tehnološkega razvoja in vodenja sodobnih družbenh praks [22]. Še posebej pomemben pa je postal v zadnjih dveh letih, odkar se je svet soočil s pandemijo Covida-19. Narodne, mednarodne ali celo globalne zdravstvene krize, kamor spadajo tudi pandemije, so pogosto zaznamovane z določeno mero vključevanja medicinske znanosti v vodenje družbe in usmerjanje družabnega življenja in Covid kriza je bila še posebej izrazit primer tega. Tako smo bili priča vsesplošni uporabi slogana "zaupajmo v znanost" (včasih "zaupajmo znanosti"), v angleščini "trust the science" z namenom upravičevanja in izpostavljanja znanstveno podprtega značaja uradno sprejetih ukrepov za spopadanje s Covid epidemijo.

Kaj natanko pomeni zaupati v znanost? Drugače povedano, *kateri* znanosti naj bi se zaupalo? Znanost lahko razumemo kot metodo (znanstvena metoda), proces (znanstveno-raziskovalni proces), socialni sistem (skupnost znanstvenikov) ali institucijo (skupek akademskih institucij, kjer se izvaja znanstveno raziskovanje). Vidimo torej, da pojem zaupanja v znanost nima enoznačnega pomena – lahko pomeni zaupati kateri koli kombinaciji zgoraj naštetih vidikov znanosti – zato ga je tudi težko enoznačno vrednotiti in proučevati. Zaupanje v znanost je Hackingova *človeška vrsta*, je posplošitev oziroma klasifikacija neke človeške lastnosti oziroma vedenjske tendence, ki v svoji prisotnosti ali odsotnosti definira posebno kategorijo človeka [10]. Človeške vrste so podvržene učinku zanke, zaradi refleksije in samo-refleksije identifikacija neke socialne entitete z določeno človeško vrsto vpliva na lastnosti te socialne entitete, kar posledično vpliva tudi na pomen človeške vrste – oznake, s katero jo poimenujemo. Pomen besede "znanost" je relativen in dinamičen tudi v odsotnosti zankanja, zato to velja tudi za "zaupanje v znanost" – ko se spreminja pomen znanosti, se spreminja tudi pojem »zaupanje v znanost«.

Zato lahko trdimo, da je zaupanje v znanost *lebdeči označevalec* (ang. *floating signifier*), oznaka brez točnega ali splošno-sprejetega pomena, torej brez točnega referentnega

objekta [13]. Ravno v tej značilnosti se skriva moč lebdečih označevalcev – nejasnost njegovega pomena dopušča individualno konstrukcijo pomena. Tako je točen pomen lebdečega označevalca relativen – za eno osebo ali skupino ljudi pomeni nekaj, za drugo nekaj drugega.

Zaupanje v znanost torej nima enoznačnega pomena, kljub temu pa lahko to idejo ovrednotimo na podlagi različnih možnih definicij. Najprej si zamislimo dva ekstrema, znanstveni dogmatizem in radikalni skepticizem do znanosti. Dogmatik bo najverjetneje trdil, da *smo dolžni* zaupati vsem aspektom znanosti – v uporabnost znanstvene metode, zanesljivost znanstveno-raziskovalnega procesa pri odgovarjanju na raziskovalna vprašanja, verodostojnost znanstvenikov in nevtralnost oziroma apolitičnost znanstvenih institucij. Radikalni skeptik, v kolikor njegova pozicija ne temelji na a-priornem zavračanju, pa se bo najbrž skliceval na uvide Foucaulta [8] in Lyotarda [14], ki sta izpostavljala neko mero relativnosti znanstvenega spoznanja. Posledično bo trdil, da znanstvene institucije niso apolitične, znanstveniki niso racionalni in zato niti verodostojni, znanstveno-raziskovalni proces in znanstvena metoda pa nista univerzalno orodje za dostopanje do resnice, temveč orodje za perpetuacijo specifične jezikovne igre.

Na srečo lahko uberemo vmesno pot, ki ustreza klasični koncepciji razsvetljenske znanosti in temelji na egalitarnem odnosu do znanja in zavračanju dolžnosti laika, da zaupa intelektualni avtoriteti. To stališče dobro povzame izjava Richarda Feynmana, da je znanost *verjetje v nevednost strokovnjakov* [7]. Potemtakem "zaupanje v znanost" pomeni priznavanje uporabnosti znanstvene metode in zanesljivosti raziskovalnega procesa, hkrati pa ohranitev zdravega dvoma v verodostojnost znanstvenikov in institucij. Če strokovnjak ali institucija trdi da *p*, ni potrebno da temu slepo verjamemo, temveč lahko zahtevamo argumentacijo in vpogled v raziskovalni proces.

2 Politična znanost

Zagovarjam stališče, da bi "zaupanje v znanost" moralo pomeniti zaupanje v znanost kot proces in metodo, ne pa v njen človeški element (znanstveniki in institucije), ki je dovzeten za razne pristranosti in konflikte interesa, zaradi katerih trpi verodostojnost znanstvenih zaključkov. Znanstvenega procesa v praksi seveda ni brez človeškega elementa, ki ta proces izvaja, vendar človeški element v tej izvedbi tudi ni nezmotljiv. Zato velja zaupati v process, v človeški element pa ne povsem. Posledično moramo ugotoviti, ali se uporaba tega slogana v zahodni družbi sklada s tovrstnim razumevanjem ali ne. V kolikor se ne, in za tem stoji pričakovanje slepega zaupanja znanstvenikom in institucijam, je to znak dogmatizma in institucionalizacje znanosti, ki sta močno povezani s politizacijo.

Carl Schmit je znan po svoji definiciji politike kot presojanju na podlagi dihotomije prijatelj/sovražnik, pri čemer je prijatelj nekdo s komer si delim interese, sovražnikovim interesom pa nasprotujem [18]. Politično vrednotenje dogajanj in dejanj torej ne temelji na splošnih načelih, temveč poteka na podlagi identitete udeleženih subjektov in uporabnosti njegovih posledic za osebo, ki presoja. Če je politika razločevanje med prijatelji in sovražniki, potem je znanost politična kadarkoli primarni kriterij za razločanje med znanstveno in neznanstveno trditvijo ni kvaliteta argumentacije in podprtost z dokazi, temveč status njenega sporočevalca. Z drugimi besedami, dihotomija prijatelj/sovražnik se v znanosti odraža, ko je "kdo je to rekel?" pomembnejše vprašanje od "kako je bila izjava argumentirana?". V podrobnosti argumentacije se morda ne moremo popolnoma spustiti, lahko pa vsaj presodimo ali je argumentacija formalno-logično ustrezna.

Lebdeči označevalci so zaradi svoje nejasnosti in dvoumnosti idealna tarča za politizacijo. Politizirana oznaka poleg svojega semantičnega pomena dobi še sociopolitični pomen – prisotnost referentnega objekta označuje prijatelja ali sovražnika (režima) oziroma pripadnika ingrupe ali outgrupe. Ravno zaradi nejasnosti semantičnega pomena (oznaka pomeni različne stvari različnim skupinam) sociopolitični pomen nadvlada semantičnega in postane primarni. Tako potem lebdeči označevalec postane univerzalna oznaka za sovražnika režima – točen semantični pomen besede sicer vsak razume po svoje, njena čustvena in moralna valenca pa sta enoznačni. Znanost je v naši družbi pozitivna, torej bi primeru politizacije "zaupanje v znanost" v svoji lebdeči obliki označevalo pripadnike ingrupe oziroma prijatelje režima, njegova odsotnost pa njegove sovražnike oziroma pripadnike outgrupe.

S tega vidika je bila Covid kriza zelo zanimiva. Moja analiza se bo sicer osredotočala predvsem na dogajanje v mednarodni in ameriški znanosti, vendar so bili enaki ali podobni vzorci prisotni tudi v Sloveniji. Ekipa znanstvenikov iz MIT-ja je leta 2021 objavila pre-print študije *Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online*, ki je poročala o navadah, značilnostih, stališčih in vrednotah spletnih skupnosti *Covid-skeptikov* oziroma *anti-maskerjev*, ljudi, ki so tako ali drugače nasprotovali uradnim Covid ukrepom [12]. Intuitivno bi se nam zdelo, da so to skupine, ki ne "zaupajo znanosti", avtorji uporabijo termin "anti-znanost" (anti-science), obstaja tudi variacija "zanikalec znanosti" (science-denier). Vendar se je izkazalo, da ti ljudje niso klasični oziroma stereotipni zanikalci znanosti, v resnici sploh ne nasprotujejo znanosti kot taki in da so nadpovprečno znanstveno pismeni. Nasprotovali so *uradni* (politično podprti) znanosti, razlikovanju med uradno in neuradno znanostjo ter avtoritarnemu odnosu stroke do laikov. Zagovarjali so torej egalitarno znanost, kjer ima vsakdo dostop do podatkov in možnost oblikovanja svojih zaključkov [12].

Avtorji študije se s tem niso strinjali in so trdili, da Covid-skeptiki "spodkopavajo uradne znanosti s spretno manipulacijo podatkov". Ta trditev se mi zdi bizarna – kako lahko želja po intersubjektivnem preverjanju s strani visoko znanstveno pismenih posameznikov, ki želijo nepristransko ovrednotiti podatke [1] "spodkopava uradno znanost"? Ni to kvečjemu koristno, saj je po Popperju [2] ravno falsifikacija gonilo znanstvenega napredka, ki je v času pandemije še toliko bolj ključen? Rekel bi, da načeloma je, vendar ne v institucionalizirani znanosti, kjer akademske institucije želijo obdržati monopol nad produkcijo znanja. Institucionalizacijo

[1] Lee et al (2021): "These users want to understand and analyze the information for themselves, free from biased, external intervention.", str. 12

[2] Popperjev model falsifikacije ima sicer svoje težave, vsekakor pa gonilo znanstvenega napredka ni izogibanje možnostim falsifikacije.

sicer lahko razumemo kot mehko obliko politizacije, vendar to še ni indikator politizacije v pravem pomenu besede.

Žal pa je med Covid krizo prišlo tudi do slednje. V institucionalizirani znanosti vlada kredencializem – merilo ideje je znanstveni in akademski prestiž znanstvenika, ki jo predlaga. Vendar med Covid krizo niti znanstveni prestiž avtorja ni bil zadosten pogoj za sprejemanje neke ideje. Tako se je npr. dr. Robert Malone moral soočiti z deplatformiranjem zaradi "širjenja dezinformacij" – Twitter mu je deaktiviral račun [15] po nastopu na Roganovem podcastu, kjer je izrazil nestrinjanje z uradnim konsenzom glede Covida in zajezitvenih ukrepov, ter svoje stališče znanstveno argumentiral [3]. Malone je sicer znanstvenik – mednarodnega renomeja [4] – vendar očitno ni izpolnjeval kriterijev za "zaupanje znanosti". Kaj je torej znanost, na katero se je med Covid krizo nanašal slogan "zaupajmo znanosti"? Ugotovili smo, da se ne nanaša na proces znanstvenega raziskovanja in niti na individualne znanstvenike z dovoljšno mero prestiža. Moja teza je torej, da se je beseda "znanost" nanašala na uradno, torej politično-podprto znanost oziroma *znanost režima*. Tukaj se lahko navežem na Foucaltov režim resnice, kjer je ideja resnice politično in ideološko umeščena – diskurz in metode produkcije resnice so omejeni, hkrati obstaja skupina ljudi, ki ima monopol nad razglašanjem družbene resnice [8].

Med Covid krizo so vlogo razsodnika resnice prevzeli *znanstveniki režima* – uradni Covid komentatorji (kot npr. dr. Fauci v Ameriki, dr. Krek in dr. Beović v Sloveniji), vlogo "čuvaja" resnice pa mediji in socialni mediji, ki so tako ali drugače utišali znanstvenike, ki so želeli izraziti kakršno koli nestrinjanje z uradnim konsenzom. Covid krizo je torej zaznamovala močna politizacija znanosti, saj je pravico do širjenja (znanstvenih) resnic nudila predvsem podpora (prijateljstvo) režima, ki se je odražala v podpori uradnega konsenza glede spopadanja s pandemijo. Posledično trdim, da slogan "zaupajmo znanosti" ni predstavljal klica k epistemski racionalnosti in sistematičnemu presojanju znanstvenih izjav, temveč ravno nasprotno – emocionalno in politično prežet sklic na avtoriteto. Cilj je bil sprejemanje stališč intelektualnih avtoritet režima, ne pa samostojni razmislek.

3 Socialno presojanje

Tematika letošnje konference je "kognitivni vidiki zaupanja v znanost". Moj cilj je pokazati, da je zaradi politizacije znanosti in zaupanja v znanost večina teh kognitivnih vidikov pod vplivom socialnih pritiskov.

V socialni psihologiji obstaja veliko raziskav in teorij na temo oblikovanja in spremembe stališč ter presojanja novih informacij. Giner-Sorolila in Chaiken sta poimenovala koncept motiviranega sklepanja, kjer sistematično sklepamo z namenom potrditi točno določeno stališče [9]. Caccioppo in Petty sta postavila dvoprocesni model spremembe stališč, kjer centralno procesiranje upošteva predvsem vsebino sporočila, periferno pa lastnosti sporočevalca in socialni kontekst [2]. Festinger pa je postavil teorijo kognitivne disonance – ljudje se držimo očitno neresničnih stališč, ker težimo k ujemanju stališč, vedenja in samopodobe [6]. Za naštete fenomene predlagam nadpomenko *socialnega presojanja in sklepanja*[5] – presojanja in sklepanja v skladu s svojo skupinsko identiteto, konsenzom ingrupe ali stališčem ingrupne intelektualne avtoritete, kar pogosto vodi do fenomena, ki ga Perkins (po navedbi Barona) poimenuje *myside bias* [1]. Socialnega presojanja se po mojem mnenju poslužujemo na vseh družbeno-relevantnih področjih, kjer nimamo motivacije, sposobnosti ali predznanja za sistematično oblikovanje lastnega stališča.

V to kategorijo zaradi svoje kompleksnosti spada večina znanstvenih tem, še posebej tistih, ki so družbeno oziroma politično relevantne, vključno s pandemijo Covida-19 in z njo povezanimi ukrepi. Pinker govori o fokusnih točkah, javno vidnih in relevantnih dogodkih in dogajanjih, ki jih vidi posameznik in se hkrati zaveda, da so vidni tudi drugim prebivalcem družbe [17]. Fokusne točke, oziroma spektakli, pogosto postanejo politizirane – to so močno družbeno relevantna dogajanja, do katerih se je potrebno opredeliti. Že sama potreba po opredelitvi je političnega značaja, ker ne dopušča nevtralnosti, zgolj izbiro enega izmed dveh polov. Ko je prisotna binarna polarizacija, pa je prisotna tudi dihotomija prijatelja (podpornika uradnih ukrepov) in sovražnika (nasprotnika uradnih ukrepov). Fokusne točke torej aktivirajo in okrepijo vrojeno tendenco človeka po socialnem presojanju, v tem primeru o vsebini same fokusne točke. Ko je zaupanje znanosti postalo fokusna točka, kar se je zgodilo med Covid krizo (če ne še prej), se je torej navzelo političnih konotacij in postalo označevalec za prijatelje in sovražnike režima – definirane kot zaupnike in zanikalce znanosti (včasih teoretike zarote). Zaupanje v znanost je torej družbenopolitični problem. Stran, na kateri se nekdo nahaja, je prej merilo politične opredeljenosti kot samega zaupanja v znanost v klasičnem pomenu izraza, ali odraz globljih filozofskih načel. Drugače povedano, izražanje (ne)zaupanja v znanost v kakršnem koli socialnem kontekstu je *politična uniforma,* zato je to prej signal privrženosti ustaljeni politiki kot pokazatelj odnosa do raziskovalne dejavnosti, ki ji pravimo znanost.

4 Politični in spoznavni razhod

V obdobju politične polarizacije zaradi socialnega presojanja in politizacije znanosti pogosto pride do spoznavnega razhoda – na eni strani imamo množico ljudi, ki takorekoč zaupa znanosti oziroma uradnim virom in zgodbam, na drugi pa množico ljudi, ki "zanika znanost" – torej zavrača uradne vire in zgodbe, ter oblikuje svoja stališča s pomočjo alternativnih virov.

Pojavita se vsaj dve različni "socialni resničnosti", dve različni interpretaciji vsebine fokusne točke. Imamo torej ljudi, ki v grobem sprejemajo uradno zgodbo in ljudi, ki jo v grobem zavračajo (seveda pa sta to sprejemanje in zavračanje kontinuum), v primeru Covida se to nanaša na stališča do mask, cepljenja in drugih uradnih ukrepov. To je v veliki meri posledica

[3] https://open.spotify.com/episode/3SCsueX2bZdbEzRtKOCEyT

[4] Malone na https://www.rwmalonemd.com: "I am an internationally recognized scientist/physician and the original inventor of mRNA vaccination as a technology- I have approximately 100 scientific publications with over 12,000 citations of my work (per Google Scholar with an "outstanding" impact factor rating committees)."

[5] Ta koncept sem podrobneje razdelal v članku *Social Reasoning and the Politicization of Science During the Covid Pandemic*, ki bo objavljen Decembra v reviji Mankind Quarterly [16].

razlik v zaznavanju zaupanja vrednih oziroma verodostojnih virov v obeh (ali vseh) skupinah ljudi. Vir, ki je verodostojen za eno skupino nikakor ni verodostojen za drugo, to presojanje o verodostojnosti pa je politične narave. Torej, spoznavni razhod je posledica političnega razhoda, ne obratno. Oziroma, kot bi rekel Foucault, znanje izvira iz moči. In šele nato spoznavni razhod perpetuira političnega – sprejemanje ene ali druge interpretacije (označeno kot zaupanje znanosti ali teoriziranje zarote) je politična uniforma, ki signalizira pripadnost enemu od političnih polov.

Kljub temu pa pomanjkljivo znanje, do katerega pride v primeru cenzure nasprotujočih stališč, nosi svoje posledice – pogosto negativne. V zadnjih mesecih prihaja vedno več raziskav in medijskih objav, ki izpostavljajo destruktivne posledice določenih Covid ukrepov – ekonomska škoda, ki so jo povzročili lockdowni [20], zaviranje razvoja otrok zaradi obveznega nošenja mask [23] in njihova splošna neučinkovitost [19], neučinkovitost cepiv pri zaščiti pred okužbo s Covidom [5] in možnost nevarnih stranskih učinkov pri določenih demografskih skupinah, npr. nosečnicah [4]. Ameriški CDC je sicer pred kratkim spremenil svoje smernice za spopadnje s Covidom – zdaj so enake za cepljene in necepljene posameznike, kar implicira enako stopnjo tveganosti obeh skupin [3]. Vendar se moramo vprašati, zakaj šele zdaj? Različni ljudje in institucije po svetu so tako ali drugače opozarjali na morebitne negativne posledice uradnih Covid ukrepov, vendar so bili tako ali drugače utišani. Tukaj torej vidimo, da imata politizacija znanosti in dogmatični odnos do tako-imenovanega "strokovnega konsenza" v naši poznanstvenjeni družbi obsežne negativne posledice.

V svetu, kjer se zdi, da lahko motiviran laik z dovoljšno mero znanstvene pismenosti v enem tednu iskanja člankov na Google Scholar doseže osnovno razumevanje (ali vsaj aproksimacijo le-tega) nekega znanstvenega področja, uradne znanstvene institucije niso več edini možni vir znanja. In v skladu s tem se moramo tudi ravnati in priznavati veljavnost izvenkonsenzualnih stališč, v kolikor so podprta z argumenti in dokazi.

Na žalost pa Googlov think tank Jigsaw in Svet za družboslovno raziskovanje (Social Science Research Council, SSRC), tako kot Lee in kolegi povlečeta ravno obraten zaključek. Laikom ne želita prepustiti, da si sami ustvarijo stališče in sami presojajo med informacijami in dezinformacijami, oziroma med znanjem in lažnimi novicami. Nasprotno, Jigsaw predstavlja koncept "pre-bunkinga" oziroma *psihološke inokulacije*, vnaprejšnjega zavračanja možnih heterodoksnih stališč v obliki kratkih sporočil, ki predstavijo protiargumente in poslušalcu olajšajo zavračanje tega stališča v prihodnosti [11]. SSRC pa skuša ugotoviti kako maksimizirati povpraševanje po Covid cepivih – tako da dijake in študente nauči prepoznavati "dezinformacije o cepivih". sporočevalce opremi z ustreznimi "sporazumevalnimi strategijami" in na družbenih omrežjih oblikuje "(demografsko in geografsko) prilagojena sporočila" [21].

Spet se moramo vprašati, kdo razlikuje med informacijo in dezinformacijo, med ortodoksnimi in heterodoksnimi stališči. Je to znanost, politika ali politizirana znanost? In nadalje, ne bi to morala biti pravica in dolžnost vsakega odraslega državljana v demokratični in egalitarni državi? Če si posameznik ne more, oziroma *ne sme* sam ustvariti mnenja, čemu potem služi demokracija?

Odgovor je, seveda, režim in "znanost" režima. Vidimo torej, da je spoznavni razhod med podporniki uradne narative in kontranarative posledica aktivno ustvarjenega političnega razhoda s strani režima in njegovih ideoloških aparatov, ki v interakciji z javnostjo ustvarjajo koncept zaupanja znanosti, zanikanja znanosti in teorij zarote. Akademiki in drugi raziskovalci imamo edinstveno možnost izpostavljanja napak režima, ampak lahko to dosežemo zgolj, če znanost zaščitimo pred politizacijo. Prvi korak k depolitizaciji znanosti pa je po mojem mnenju prepoznavanje koncepta zaupanja v znanost kot politične uniforme in posledično zavračanje vseh dihotomij, ki jih ustvari.

LITERATURA

[1] Baron, J. (2000), Thinking and deciding (3rd ed.), New York: Cambridge University Press

[2] Cacioppo, J.T. in Petty, R.E. (1984). The Elaboration Likelihood Model of Persuasion. NA - Advances in Consumer Research, 11, 673-675.

[3] CDC (2022). CDC streamlines COVID-19 guidance to help the public better protect themselves and understand their risk. Sneto iz https://www.cdc.gov/media/releases/2022/p0811-covid-guidance.html

[4] Clark County Today (2022). FDA knew huge percentage of women in Pfizer trial suffered miscarriages. Sneto iz: https://www.clarkcountytoday.com/news/fda-knew-huge-percentage-of-women-in-pfizer-trial-suffered-miscarriages/

[5] Comber, J. in Madhava, S. (2022). After Data Show Vaccinated at Higher Risk of Dying from COVID, Canadian Province Ends Monthly Reports. Global Research. Sneto iz: https://www.globalresearch.ca/after-data-show-vaccinated-higher-risk-dying-from-covid-canadian-province-ends-monthly-reports/5790795

[6] Festinger, L. (2022). *A Theory of Cognitive Dissonance (Anniversary ed.)*. Stanford University Press.

[7] Feynman, R. (n.d.). What is Science? Sneto iz: http://www.feynman.com/science/what-is-science/

[8] Fontana A. & Pasquino, P. (n.d.) *Truth and Power (interview with Foucault)*. Sneto iz https://www2.southeastern.edu/Academics/Faculty/jbell/foucaulttruthpower.pdf

[9] Giner-Sorolila, R., & Chaiken, S. (1997). Selective Use of Heuristic and Systematic Processing Under Defense Motivation. Personality and Social Psychology Bulletin, 23(1), 84–97. doi:10.1177/0146167297231009

[10] Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). Clarendon Press/Oxford University Press.

[11] Jigsaw (2021). Psychological Inoculation: New Techniques for Fighting Online Extremism. Medium. Sneto iz: https://medium.com/jigsaw/psychological-inoculation-new-techniques-for-fighting-online-extremism-b156e439af23

[12] Lee, C., Yang, T., Inchoco, G. D., Jones, G. M., & Satyanarayan, A. (2021). *Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online.* Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. *https://doi.org/10.1145/3411764.3445211*

[13] Lévi-Strauss, C. (1987). *Introduction to Marcel Mauss*. Routledge.

[14] Lyotard, J.-F. (1984). *The Postmodern Condition: A Report on Knowledge*. University of Minnesota Press.

[15] Malone, R. (2021). Permanently suspended on Twitter. Who is Robert Malone. Sneto iz: https://rwmalonemd.substack.com/p/permanently-suspended-on-twitter

[16] Marušič, J. Ž. (2022). Social Reasoning and the Politicization of Science During the Covid Pandemic. Mankind Quarterly. Sprejeto v objavo.

[17] Pinker, S. (2021). Rationality: What Is It, Why It Seems Scarce, Why It Matters. Viking Press.

[18] Schmitt, C. (2007) *The Concept of the Political.* University of Chicago Press.

[19] Spira B (April 19, 2022) Correlation Between Mask Compliance and COVID-19 Outcomes in Europe. Cureus 14(4): e24268. doi:10.7759/cureus.24268

[20] Sumption, J. (2022). Little by little the truth of lockdown is being admitted: it was a disaster. The Times. Sneto iz: https://www.thetimes.co.uk/article/little-by-little-the-truth-of-lockdown-is-being-admitted-it-was-a-disaster-5b5lrlgwk

[21] The Rockefeller Foundation. Mercury Project to Boost Covid-19 Vaccination Rates and Counter Public Health Mis- and Disinformation in 17 Countries Worldwide. Sneto iz: https://www.rockefellerfoundation.org/news/mercury-project-to-boost-

covid-19-vaccination-rates-and-counter-public-health-mis-and-disinformation-in-17-countries-worldwide/
[22] Ule, A. (2006). Znanost, družba, vrednote. Aristej.
[23] Watson, S. (2022). CNN Medical Analyst Who Fiercely Advocated Masking Now Admits It 'Harmed' Her Own Son's Development. Summit News. Sneto iz: https://summit.news/2022/08/24/cnn-medical-analyst-who-fiercely-advocated-masking-now-admits-it-harmed-her-own-sons-development/

Filozofski in psihološki vidiki človeške racionalnosti

Philosophical and psychological aspects of human rationality

Nastja Tomat
Oddelek za filozofijo
Filozofska fakulteta UL
Ljubljana, Slovenija
nastja.tomat@ff.uni-lj.si

POVZETEK

Človeška racionalnost je kompleksen pojem, ki se nanaša na široko paleto našega spoznavanja in delovanja. Obstajajo številne opredelitve racionalnosti; Ronald de Sousa razlikuje med kategorično in normativno racionalnostjo, govorimo lahko o instrumentalni ali široki racionalnosti ali o racionalnosti kot logičnem sklepanju. Vprašanja o racionalnosti so tesno prepletena s preučevanjem odločanja. Normativne teorije odločanja racionalno vedenje opredelijo kot tisto, ki vodi do izida z največjo pričakovano koristnostjo, deskriptivne teorije pa preučujejo, kako se odločanje v vsakdanjem življenju dejansko poteka. K odmiku od idealiziranega pogleda na racionalnost so pripomogli program hevristik in pristranosti, ki sta ga osnovala Daniel Kahneman in Amos Tversky, koncept omejene racionalnosti, ki ga je predstavil Herbert A. Simon, ter delo Gerda Gigerenzerja in sodelavcev, ki preučujejo ekološko racionalnost. Poleg racionalnosti dejanj lahko govorimo tudi o racionalnosti prepričanj, kar preučevanje racionalnosti poveže s temeljnimi vprašanji s področja epistemologije.

KLJUČNE BESEDE

omejena racionalnost, ekološka racionalnost, racionalnost prepričanj, hevristike in pristranosti

ABSTRACT

Human rationality is a complex topic that encompasses a wide range of cognitive processes and behavior. Many definitions of rationality exist, one of them being Ronald de Sousa's notion of categorical and normative rationality. Some authors distinguish between instrumental and broad conception of rationality, while others define rationality in terms of logical reasoning. The study of rationality is intertwined with research in the field of decision making. Normative theories define rationality as behavior that leads to the outcome with the greatest expected utility, while descriptive theories examine how people actually make decisions in everyday life. Kahneman and Tversky's heuristics and biases program, Herbert A. Simon's concept of bounded rationality and Gerd Gigerenzer's study of ecological rationality all contributed to the shift from the idealized view of human rationality to a more moderate one. In addition to research on rational action, study of rational beliefs is another field of inquiry that connects investigation of rationality with fundamental questions in epistemology.

KEYWORDS

bounded rationality, ecological rationality, rationality of belief, heuristics and biases

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

1 UVOD

Vprašanje, ali smo ljudje racionalna bitja, še zdaleč ni enostavno. Odgovor se že stoletja izmika znanstvenikom različnih disciplin od ekonomije in psihologije do filozofije in kognitivne znanosti. Človeška racionalnost je tema, ki se je lahko lotevamo iz številnih vidikov in z uporabo različnih metod, zato ni nenavadno, da danes na tem področju obstaja ogromno polje razprav in raziskav. V veliki razpravi o racionalnosti, kot so to poimenovali v kognitivni znanosti, obstajata dva nasprotujoča si pogleda. Na enem polu so avtorji, ki zagovarjajo, da so človeško sklepanje, presojanje in odločanje, ki so del racionalnega vedenja, polni pomanjkljivosti in pristranosti ter da jih je mogoče izboljšati; zagovorniki takšnega pogleda v veliki meri izhajajo iz programa hevristik in pristranosti, ki sta ga osnovala psihologa Daniel Kahneman in Amos Tversky. Raziskovalci na drugem polu pa takšnemu pogledu na racionalnost nasprotujejo in trdijo, da so kriteriji normativnih teorij racionalnosti neustrezni ter da izsledki empiričnih raziskav, ki pričajo o sistematičnih odklonih od omenjenih kriterijev, še ne zadostujejo za sklep, da smo ljudje iracionalni [1, 2, 3].

Namen prispevka je podati pregled izbranih pogledov na človeško racionalnost. Začela bom z definicijo filozofa Ronalda de Sousa, nadaljevala pa z dvema opredelitvama racionalnosti, med katerima se v literaturi pogosto razlikuje: instrumentalno in široko. Na primeru Wasonove naloge izbire kart – ene najbolj uporabljenih nalog pri empiričnem preučevanju sklepanja – bom opisala pogled, ki racionalnost povezuje z logičnim sklepanjem ter je še vedno vpliven zlasti na področju filozofije. Poleg logike je področje, ki je prav tako prepleteno s preučevanjem racionalnosti, odločanje. Opisala bom, kakšno sliko racionalnosti prikazujejo normativne teorije odločanja ter kako se je kot kritika takšnega pogleda izoblikoval program hevristik in pristranosti, ki je še danes eden najvplivnejših okvirjev za preučevanje odločanja in presojanja. Nato bom predstavila koncept omejene racionalnosti, ki ga je oblikoval Herbert A. Simon in je

pomembno vplival na razumevanje in pojmovanje racionalnosti, ter koncept ekološke racionalnosti, ki ga preučujejo Gerd Gigerenzer in sodelavci ter se naslanja na Simonovo delo. V zadnjem delu se bom odmaknila od empiričnih raziskav odločanja in presojanja ter opisala nekatera vprašanja, ki jih odpira raziskovanje racionalnosti prepričanj – teme na presečišču preučevanja racionalnosti in epistemologije.

2 DE SOUSOVA OPREDELITEV RACIONALNOSTI

Filozof Ronald de Sousa najprej razlikuje med kategorično in normativno racionalnostjo. Pri kategorični racionalnosti je nasprotje racionalnega aracionalno vedenje. Racionalno je takšno vedenje, ki ga vodijo določeni razlogi, aracionalno pa takšno, ki ga ne vodi mišljenje ali izbira. Pri kamnu, ki ga vržemo skozi okno, ali človeku, ki se spotakne in pade v grm kopriv, ne govorimo o (i)racionalnosti – pri prvem gre namreč za pojav, ki uboga zakone fizike, pri drugem pa za dejanje, ki ga ni vodila izbira. Pri normativni racionalnosti pa razlikujemo med racionalnim in iracionalnim vedenjem. Racionalno vedenje je tisto, ki je ustrezno utemeljeno z določenimi razlogi, normami ali vrednotami, iracionalno pa tisto, ki se od temu pogoju na tak ali drugačen način ne zadostuje. De Sousa pravi, da lahko o ljudeh kot o racionalnih živalih govorimo samo, če sprejmemo, da smo ljudje racionalni v kategoričnem smislu in kot taki tudi sposobni iracionalnega vedenja [4].

Če kategorične racionalnosti ne pripisujemo dogodkom, ki jih lahko zadostno razložimo z naravnimi zakoni, ali to pomeni, da z njimi ne moremo razložiti človeškega vedenja? Zmernejša interpretacija pravi, da je človeško vedenje podvrženo naravnim zakonom, vendar ti ne ponujajo zadostne razlage. Kot primer de Sousa navaja igro šaha, ki ga moramo razložiti s pravili igre – in ta niso naravni zakoni. Močnejša interpretacija pa pravi, da vedenje racionalnih bitij, vključno s človekom, na nek način presega zakone narave. De Sousa meni, da je tako stališče absurdno, saj bi predpostavljalo čudež ali pa vsaj to, da zakonov narave ne razumemo pravilno. Zagovarja, da moramo človeka obravnavati kot bitje, ki je kot vsa ostala podvržen zakonom narave; razliko med človekom in ostalimi bitji je potrebno iskati v zakonih narave in ne v lastnostih, ki bi le-te na nek način presegale. Če privzamemo, da se racionalnost nanaša na misli in dejanja, lahko razlikujemo med dvema ključnima spremembama tako na nivoju evolucije kot razvoja posameznika: prva je razvoj od golega zaznavanja objektov do zmožnosti tvorbe reprezentacij, druga pa razvoj od avtomatskih vedenjskih odzivov do zmožnosti oblikovanja namer ter želja ter vedenja na podlagi le-teh [4].

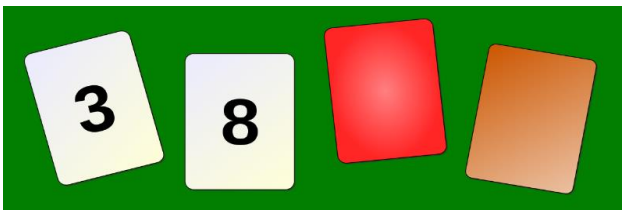
3 INSTRUMENTALNA IN ŠIROKA RACIONALNOST

V literaturi se pogosto pojavlja razlikovanje med ožjim, instrumentalnim in širokim pojmovanjem racionalnosti [1, 4]. Instrumentalno racionalnost opredelimo kot vedenje, ki nas približa doseganju zastavljenega cilja glede na mentalne in fizične vire, ki so nam na voljo. Povedano drugače, racionalno je tisto vedenje, ki optimizira doseganje ciljev, pri čemer se ne ukvarjamo s tem, kakšni ti cilji so in kako si jih posameznik postavlja. Prednost takšnega pristopa je v tem, da lahko postavimo norme, ki služijo kot kriterij racionalnosti, ter spremljamo, v kolikšni meri in pod kakšnimi pogoji ljudje od njih odstopamo. Po drugi strani pa se zdi preučevanje racionalnosti le iz instrumentalnega vidika preozko – če se osredotočamo samo na ciljno usmerjeno vedenje, izpustimo pa vprašanja o ciljih, normah in vrednotah, zanemarimo velik in pomemben del človeškega delovanja [1]. John Searle v svoji knjigi o racionalnosti navaja primer znanih raziskav o inteligentnosti opic, ki jih je psiholog Wolfgang Köhler izvajal na Tenerifih. V eksperimentih se je izkazalo, da so opice sposobne reševanja problemov z vpogledom; da bi dosegle na strop obešene banane, do katerih niso mogle priti s skakanjem, so uporabile škatle in palico [6]. Iz intrumentalnega vidika so se opice torej vedle racionalno in Searle meni, da tudi racionalnost človeka še vedno presojamo na podoben način. V klasičnih modelih racionalnosti je človeška racionalnost pravzaprav le kompleksnejša verzija šimpanzje. Searle v nadaljevanju opozarja na pomanjkljivosti takšnega pojmovanja racionalnosti in opozarja na pomembnost ločevanja vedenja na podlagi želja in na podlagi razlogov [7].

V odgovor na pomanjkljivosti instrumentalnega pristopa so se pojavila širša pojmovanja racionalnosti, ki upoštevajo tudi cilje, prepričanja, norme in vrednote, ki usmerjajo naše vedenje. Te teorije se med drugim ukvarjajo z vprašanji o racionalnosti samih ciljev [5] ter o vedenju, ki nima samo instrumentalne funkcije [1]. Filozof Robert Nozick na primer govori o konceptu simbolne koristnosti in pravi, da imajo naša dejanja neodvisno od instrumentalne tudi simbolno vrednost, ki bi jo morale vključevati vse formalne teorije racionalnosti in odločanja. Ker živimo v socialno in simbolno kompleksnem okolju, naša dejanja služijo tudi namenom, ki presegajo doseganje ozko zastavljenih ciljev, na primer temu, da sebi in drugim sporočamo, kakšne osebe smo [8]. Podobno ekonomist Shaun H. Heap kot protipol instrumentalni racionalnosti postavlja ekspresivno racionalnost. Ko izvajamo dejanja, ki so ekspresivno racionalna, opredeljujemo in raziskujemo lastna prepričanja in vrednote. Ne gre torej za enosmerno povezavo med vrednotami in delovanjem, temveč za povratno zanko, kjer z dejanji vrednote tudi konstruiramo, spremljamo in prilagajamo [9].

4 RACIONALNOST IN LOGIČNO MIŠLJENJE

Najbrž eden od najstarejših kriterijev racionalnosti je sledenje pravilom logičnega sklepanja in verjetnostnega računa [1]. Ena od najbolj preučevanih nalog, ki se uporablja v empiričnih raziskavah sklepanja, je Wasonova naloga izbire kart [10, 11], ki ima naslednjo obliko: »Na mizi so štiri karte. Vsaka ima na eni strani številko, na drugi pa barvo. Katere karte je potrebno obrniti, da testiraš pravilo: če je na eni strani sodo število, je na drugi strani rdeča barva?«



Slika 1: Primer Wasonove naloge izbire kart.

V zgoraj navedenem primeru je pravilni odgovor, da je potrebno obrniti karto s številko 8, s čimer preverimo modus ponens, in karto rjave barve, s čimer preverimo modus tollens. Večina udeležencev pri takšni nalogi poda odgovor, da je potrebno obrniti karto s številko 8 in karto z rdečo barvo, vendar gre pri slednjem za napako zatrjenega konsekvensa. V več kot petdesetih letih od izvirne objave je bila naloga uporabljena v ogromnem številu raziskav, kjer so avtorji manipulirali z različnimi spremenljivkami, ki bi lahko vplivale na izvedbo naloge, še danes pa ni enotne razlage za majhen delež pravilnih rešitev; ena od interpretacij je, da se večinoma osredotočamo na potrjevanje hipoteze, manj pa preverjanje pogojev, ki bi hipotezo ovrgli [12, 13, 14]. Še ena ugotovitev je, da so udeleženci pogosto nagnjeni k izbiri kart, ki so eksplicitno omenjene v navodilu [15]. Eno od opažanj je, da se delež pravilnih rešitev poveča, če namesto abstraktnih uporabimo konkretne primere, kar nakazuje na to, da se pri logičnem sklepanju oz. tesitranju hipotez ne zanašamo le na obliko argumentov, temveč tudi na vsebino [16]. Delež pravilnih rešitev je še večji, če uporabimo deontična pravila. Če morajo udeleženci na primer preverjati pravilo »Če piješ alkohol, moraš biti starejši od 18 let«, na mizi pa imajo karte s številkami 16 in 25 ter z napisi »pivo« in »kokakola«, večina pravilno izbere karti s številko 16 in napisom »pivo«. Testiranje hipotez nam gre očitno torej bolje, ko moramo preverjati morebitne kršitve socialnih pravil [17]. Ena od interpretacij, ki temelji na evolucijski psihologiji, je, da sklepanje ni le splošen, od vsebine neodvisen proces, temveč je v ozadju več specializiranih procesov, eden izmed katerih je namenjen reševanju problemov v kontekstu socialnih izmenjav in kršitev socialnih pogodb [17]; ta interpretacija je deležna številnih kritik [18]. Nekateri avtorji pa menijo, da je logično pravilna rešitev Wasonove naloge v konfliktu z načinom, kako v vsakdanjem življenju testiramo hipoteze, ter zagovarjajo, da je način, kako se udeleženci lotijo reševanja, v resničnem življenju adaptiven. Po njihovem neuspešno reševanje naloge torej ne služi kot dokaz iracionalnosti [19, 20]. To se sklada s pogledom, da logičnega mišljenja ne gre vedno in apriori enačiti z racionalnostjo, temveč je ustreznost takšnega mišljenja odvisna tudi od konteksta [1].

5 RACIONALNOST IN ODLOČANJE

Pojem racionalnosti je tesno prepleten s preučevanjem odločanja in presojanja. Znotraj vedenjskega preučevanja odločanja ločimo med normativnimi, deskriptivnimi in preskriptivnimi pristopi. Normativne teorije se osredotočajo na to, kako bi se ljudje morali odločati, da bi prišli do izida, ki ima zanje največjo koristnost, deskriptivne teorije preučujejo, kako človeško odločanje v resničnem življenju dejansko poteka, preskriptivne pa želijo zmanjšati vrzel med prvima dvema in osnovati predloge za izboljšanje odločanja [21].

Prevladujoč model normativnega odločanja pod pogojem tveganja je bila dolgo časa teorija pričakovane koristnosti, ki sta jo v knjigi Theory of Games and Economic Behaviour leta 1944 predstavila John von Neumann in Oskar Morgenstern. Teorija temelji na aksiomih, ki se nanašajo na odločevalčeve preference. Med drugim predpostavljajo, da ima posameznik popoln, urejen in tranzitiven nabor preferenc; to pomeni, da lahko za vsak par alternativ določi, v kakšnem odnosu sta, velja pa tudi, da v primeru, ko posameznik preferira alternativo A pred B in B pred C, preferira tudi A pred C. Če aksiomi držijo, lahko vsaki alternativi pripišemo določeno koristnost in racionalno odločanje je tisto, ki privede do izida z najvišjo koristnostjo [22].

Normativne teorije pred odločevalce torej postavljajo stroge zahteve in kmalu so se začela pojavljati vprašanja, če se ljudje v vsakdanjem življenju resnično odločamo na tak način. Kahneman in Tversky sta leta 1979 objavila članek, v katerem sta pokazala, da ljudje sistematično kršimo aksiome racionalnosti, na katerih slonijo normativne teorije. Svoje ugotovitve sta strnila v teorijo obetov, ki nadgrajuje teorijo pričakovane koristnosti in razlaga, kako se ljudje odločamo pod pogojem tveganja [23].

Kahneman in Tversky sta dolga leta preučevala presojanje in odločanje in osnovala raziskovalni okvir, ki ga poznamo pod imenom »program hevristik in pristranosti«. V številnih raziskavah sta pokazala, da ljudje v negotovih pogojih pogosto uporabljamo hevristike – miselne bližnjice, ki olajšujejo reševanje problemov, so hitre, varčne in zahtevajo manj napora – kar vodi do sistematičnih napak v presojanju in odločanju, ki sta jih poimenovala kognitivne pristranosti. Ljudje pogosto ne upoštevamo pravil logike in verjetnostnega računa, smo slabi intuitivni statistiki, zaključujemo brez ustreznih dokazov, slabo napovedujemo lastne prihodnje preference in smo podvrženi številnim dejavnikom, ki na tak ali drugačen način »neupravičeno« vplivajo na naše presoje. Doprinos programa hevristik in pristranosti je ravno v poudarjanju tega, da ljudje nismo racionalni v okviru normativnih teorij, temveč da odločanje in presojanje v vsakdanjem življenju potekata drugače [24, 25, 26].

Delo Kahnemana in Tverskyja je bilo skozi leta deležno različnih kritik. Če smo ljudje resnično tako podvrženi sistematičnim napakam v presojanju in odločanju, kako je sploh mogoče, da se dovolj učinkovito odzivamo na okolje, da preživimo? Različni avtorji so ponudili alternativne interpretacije izsledkov, ki naj bi izražali pristranosti v mišljenju. Ena vrsta interpretacij se ukvarja z razlago odgovorov na naloge znotraj laboratorijskih pogojev, druga pa z vprašanjem, kaj nam ti odgovori povedo o sklepanju, presojanju, odločanju in reševanju problemov v vsakdanjem življenju. Že znotraj laboratorijskega konteksta ni vedno enoznačno, ali je določen odgovor na nalogo pravilen ali napačen. Primer tega so različne interpretacije že omenjene Wasonove naloge izbire kart. Oaksford in Chater na primer menita, da naloga ne ocenjuje deduktivnega sklepanja, temveč verjetnostno. Če privzamemo, da kriterij za pravilne odgovore ni upoštevanje pravila falsifikacije, temveč izbira najbolj informativnih kart v skladu s teorijo optimalne selekcije podatkov (ang. *optimal data selection*), lahko nekatere odgovore udeležecev smatramo kot pravilne, tudi če ne sledijo pravilom formalne logike [19, 20]. Zagovorniki druge vrste interpretacij pa segajo izven laboratorija in menijo, da so »napačni« odgovori udeležencev iz evolucijskega, adaptivnega vidika pravzaprav smiselni. Odgovori, ki jih v umetno ustvarjenih problemih v laboratorijskem eksperimentiranju razlagamo kot napake, imajo v vsakdanjem življenju prilagoditveno vlogo in zato morda ni upravičeno, da jih jemljemo kot dokaz človeške iracionalnosti [27, 28, 29].

6 OMEJENA RACIONALNOST

Še en koncept, ki je pomembno vplival na odmik od idealiziranih, normativnih teorij odločanja, je bil koncept omejene racionalnosti, ki ga je v 50. letih prejšnjega stoletja predstavil Herbert A. Simon. Simon je menil, da je pojem globalne racionalnosti, ki naj bi jo posedoval človek v ekonomskih teorijah odločanja, potrebno nadomestiti s pojmom racionalnega vedenja, ki je kompatibilno z računskimi sposobnostmi in dostopnostjo do informacij, kot jo ima človek v lastnem okolju v resnici. Racionalnost po njegovem ne pomeni iskanje najboljše možne, temveč zgolj dovolj dobre rešitve, kar je poimenoval *satisficing*. Uporabil je prispodobo škarij, kjer eno rezilo ponazarja računske zmožnosti akterja, drugo pa strukturo okolja; zagovarjal je, da je pri preučevanju človeške racionalnosti pomembno upoštevanje in razumevanje obeh »rezil« [30, 31, 32, 33].

Simon je v svojih delih podrobno razdelal tako omejitve človekovega kognitivnega sistema kot značilnosti okolja. Menil je, da ni dokazov, ki bi pričali v prid temu, da človeško odločanje poteka na način, kot to predpostavljajo normativne teorije, in da ljudje v kompleksnih odločitvenih situacijah uporabljamo poenostavitve. Ena od njih je, da ne iščemo najboljše možne, optimalne rešitve, temveč si postavimo kriterij in izide nad njim obravnavamo kot zadovoljive, pod njim pa kot nezadovoljive. Seveda se ob tem poraja vprašanje, na kakšen način si postavljamo kriterij. Poleg tega pogosto nimamo popolnih informacij o tem, do kakšnih izidov bodo privedle različne alternative. Simon je zagovarjal, da v samem procesu odločanja postopoma pridobivamo informacije o tem in posodabljamo naše poznavanje odnosa med alternativami in izidi. Vrednotenje alternativ po njegovem mnenju poteka postopoma, zaporedno, in odločevalec lahko preprosto izbere prvo zadovoljivo. Kriterij za to, kaj je zadovoljiva rešitev, lahko po potrebi prilagajamo – če je previsok, ga znižamo in obratno, s čimer zagotovimo, da bomo v vsakem primeru prišli vsaj do ene rešitve [30].

Poleg zmožnosti organizma je za razumevanje racionalnosti potrebno upoštevati tudi strukturo okolja. Simon je menil, da se moramo osredotočiti na lastnosti okolja, ki so za odločevalca pomembne in ki predstavljajo njegov življenjski prostor. Ne gre torej preprosto za preučevanje fizičnih lastnosti sveta, ki nas obdaja; to, kaj smatramo kot okolje, je odvisno od zaznavnih sposobnosti, želja, potreb in ciljev organizma. Po Simonovem mnenju odločevalci nimajo le enega, temveč več različnih mehanizmov odločanja, ki so hierarhično urejeni, in vprašanje, ki si ga moramo zastaviti, je, katere procese odločanja bomo v posameznih situacijah še lahko označili za prilagoditvene [31].

Vprašanje, kaj pomeni racionalno obnašanje, je torej drugačno, če ga zastavimo z upoštevanjem omejitev odločevalca in njegovega okolja ali pa iz perspektive normativnih teorij racionalnosti. Ob upoštevanju vseh omejitev človeka, zlasti glede računskih in napovednih sposobnosti, je dejanska, človeška racionalnost lahko v najboljšem primeru le poenostavljen približek t. i. globalne racionalnosti, na kateri slonijo npr. modeli teorije iger [30].

Koncept omejene racionalnosti se je od izvirnih Simonovih del do danes razvijal in nadgrajeval ter še vedno močno vpliva na preučevanje odločanja in racionalnosti [34]. Na njem temelji tudi delo psihologa Gerda Gigerenzerja in sodelavcev, ki so osnovali raziskovalni program hitrih in varčnih hevristik ter so ostri kritiki programa hevristik in pristranosti. Zagovarjajo, da so hevristike lahko učinkovita orodja mišljenja in da poseganje po njih v nekaterih situacijah, sploh takšnih z visoko stopnjo negotovosti, lahko pojmujemo kot racionalno. Ukvarjajo se s tako imenovano ekološko racionalnostjo, kjer je poglavitno vprašanje, katera strategija v določeni situaciji vodi do boljših izidov kot druge. Boljše kot je ujemanje med strategijo, na primer določeno hevristiko, in strukturo naloge, bolj ekološko racionalni smo [35, 36].

7 RACIONALNOST PREPRIČANJ

Poleg racionalnosti dejanj lahko govorimo tudi o racionalnosti prepričanj. Prepričanje je eden od temeljnih pojmov v epistemologiji in je del klasične tripartitne definicije znanja, ki le-tega opredeli kot upravičeno resnično prepričanje. Eno od osrednjih vprašanj epistemologije je, kako priti do resničnih prepričanj. Vprašanje je neločljivo povezano s preučevanjem racionalnosti. Kakšen je odnos med racionalnostjo, upravičenostjo in resničnostjo prepričanj ter znanjem? So racionalna prepričanja tista, ki so upravičena, ali gre za ločena pojma? Kako ljudje oblikujemo svoja prepričanja in kako bi jih morali [37, 38]?

Tradicionalni pogled je, da je vprašanje, kako bi ljudje morali oblikovati prepričanja, v domeni epistemologije, vprašanje, kako dejansko jih, pa v domeni psihologije, in da naj bi disciplini delovali ločeno ena od druge. Do neke mere drži, da so normativna vprašanja epistemologije ločena od deskriptivnih vprašanj psihologije - če bi določena psihološka spoznanja na primer pričala o tem, da je proces oblikovanja prepričanj pretežno nezaveden in da ljudje večinoma stremimo k tem, da sprejmemo prepričanja, ki spadajo v že obstoječo mrežo prepričanj, to samo po sebi ne daje dodatne teže koherentistični teoriji upravičenja v epistemologiji. Vprašanji sta se začeli povezovati v 60. letih prejšnjega stoletja, ko je Willard V. O. Quine predstavil program naturalistične epistemologije, ki poudarja, da so pri preučevanju prepričanj in znanja potrebne tudi metode, izsledki in teorije empiričnih znanosti [38].

V literaturi se pogosto pojavlja izraz epistemska racionalnost. Pritchard jo opredeli kot obliko racionalnosti, katere cilj je pridobivanje resničnih prepričanj [37]. Po njegovem lahko človek, ki stremi k epistemski racionalnosti, privzame različne strategije. Ena od njih je maksimizacija števila resničnih prepričanj, druga pa minimizacija števila napačnih prepričanj, vendar pri obeh naletimo na težave: najboljši način za maksimizacijo števila resničnih prepričanj je, da verjamemo kar koli, s čimer neizogibno pridobivamo tudi napačna prepričanja, najboljši način za minimizacijo števila napačnih prepričanj pa, da ne verjamemo skoraj ničesar. Zdi se, da bi bilo najbolj smiselno privzeti vmesen, uravnotežen pristop med verjetjem vsemu in radikalnim skepticizmom [37]. Cilj epistemske racionalnosti pa ni postavljen v prihodnost, temveč v sedanjost – večino epistemologov zanima, kakšno je stanje naših resničnih prepričanj v tem trenutku, ne pa na primer čez eno leto. Za primer lahko vzamemo osebo, ki je brez ustreznih dokazov prepričana, da je dobra v matematiki. To prepričanje vodi v obiskovanje dodatnih ur matematike in zvišuje motivacijo ter količno učenja, kar na dolgi rok dejansko pripomore k večjemu številu resničnih prepričanj o matematiki. Kljub temu bi večina epistemologov

zavrnila idejo, da je posedovanje prvega prepričanja epistemsko racionalno [38].

Do zdaj omenjeni pogled prepričanja pojmuje kategorično, pri čemer imamo le tri možnosti: lahko smo prepričani, da p, prepričani, da ne-p, ali pa se prepričanja vzdržimo. Nekatera področja epistemologije, na primer bayesovska epistemologija, pa prepričanja obravnavajo kot stopenjska – prepričanje torej ni več propozicionalno stanje v smislu »vse ali nič«, temveč smo lahko v neko propozicijo prepričani bolj ali manj. V tem primeru se odpirajo številna nova vprašanja, na primer kakšen je odnos med dokazi za določeno propozicijo in našo stopnjo prepričanja vanjo ter kakšno stopnjo prepričanja potrebujemo, da lahko trdimo, da je posedovanje nekega prepričanja epistemsko racionalno [39, 40, 41]. S tem povezana so tudi vprašanja o tem, kako prepričanja posodabljamo ali spreminjamo, ko pridobivamo nove informacije. Obstajajo različni modeli, ki opisujejo te procese, na primer AGM model revizije prepričanj [42] in teorija rangiranja [43, 44].

Nadaljnja vprašanja, povezana z epistemsko racionalnostjo, se dotikajo epistemskih norm in odgovornosti. Pravila, ki nam narekujejo, kako oblikovati prepričanja, se imenujejo epistemske norme. Poraja se vprašanje, ali lahko agenta, ki prepričanja oblikuje v skladu z napačnimi epistemskimi normami, še vedno smatramo za epistemsko racionalnega. Šibkejši, deontični pogled na epistemsko racionalnost pravi, da ja – agentova prepričanja so epistemsko racionalna, če so v skladu z epistemskimi normami, ki jim agent sledi. V hipotetični situaciji, kjer bi bil agent sistematično zaveden glede epistemskih norm, ni odgovoren za morebitna napačna prepričanja; nasprotno pa v situaciji, kjer je bil seznanjen s pravimi epistemskimi normami, pa vseeno sledi napačnim, odgovornosti za napačna prepričanja ni razrešen. Močnejši, ne-deontični pogled pa kot kriterij za epistemsko racionalnosti postavlja, da agent sledi pravim epistemskim normam, torej tistim, ki dejansko vodijo do resnice. Težava ne-deontičnega pogleda je v tem, da agent nikoli ni odgovoren za napačna prepričanja – če sledi napačnim epistemskim normam, sicer ni epistemsko racionalen, vendar tudi ni odgovoren za svoje zmote [37].

Predmet razprave je tudi vprašanje o odnosu med epistemsko racionalnostjo in upravičenjem. Nekateri izraza »epistemsko racionalna prepričanja« in »epistemsko upravičena prepričanja« uporabljajo kot sinonima, drugi ju ločujejo. V drugem primeru ni jasno, kakšen je odnos med epistemsko racionalnim prepričanjem in znanjem. Ena od možnih pozicij je, da tudi če sprejmemo upravičenje vsaj kot nujen, če ne že zadosten pogoj za znanje, za epistemsko racionalnost to ne velja. Epistemsko racionalna prepričanja torej z znanjem niso povezana na enak način kot upravičena prepričanja. Če prekinemo povezavo med znanjem in epistemsko racionalnostjo, nam to omogoča, da slednjo preučujemo tudi izven okvirja epistemologije in jo povežemo z drugimi vidiki racionalnosti, na primer racionalnostjo odločitev in dejanj. Foley predlaga, da je odločitev (načrt, strategija) za osebo racionalna, če lahko oseba epistemsko racionalno verjame, da bo odločitev v zadovoljivi meri vodila v izpolnitev njenih ciljev [38].

Namen tega dela prispevka je bil nakazati le nekatera izmed številnih vprašanj, ki se odpirajo na presečišču preučevanja racionalnosti in epistemologije. Racionalnost prepričanj ali teoretsko racionalnost se pogosto prikazuje kot protipol praktični, instrumentalni racionalnosti in menim, da je za razumevanje celotne slike pomembno poznavanje obeh pogledov ali »vrst« racionalnosti. Osredotočila sem se predvsem na odnos med racionalnostjo in različnimi temeljnimi pojmi epistemologije, zlasti upravičenjem, ter na povezavo med racionalnostjo in epistemskimi normami. Seveda pa na področju racionalnosti prepričanj obstajajo še številna druga vprašanja in pogledi, opis katerih presega namen prispevka.

8 ZAKLJUČEK

Racionalnost je kompleksen pojem, ki zajema široko paleto človeškega spoznavanja in delovanja. Opredelitve racionalnosti, kriteriji zanjo in metode, s katerimi jo preučujemo, so tako številne in raznolike, da kategoričnega odgovora na vprašanje, ali smo ljudje racionalni, ni pričakovati. Hkrati so praktično vsa področja našega življenja prepredena vsaj z implicitnimi prepostavkami o lastni (i)racionalnosti in tako je preučevanje le-te pomembno ne le iz teoretskega, ampak tudi iz aplikativnega vidika. Preučevanje racionalnosti kot optimalnega doseganja ciljev lahko služi kot podlaga za oblikovanje spodbud in strategij, ki bi tako posameznikom v vsakdanjem življenju kot strokovnjakom z različnih področij, kot so zdravstvo, gospodarstvo in pravo, pomagale pri učinkovitem sprejemanju dobrih odločitev. Tu pa pridemo do naslednjega vprašanja, ki se odpre, ko presežemo instrumentalno pojmovanje racionalnosti – kaj so »dobre« odločitve ali »racionalni« cilji? In nenazadnje, zakaj bi si pravzaprav želeli biti racionalni – ker menimo, da je tako prav, ker racionalno delovanje izboljšuje naše možnosti za preživetje in uspeh, ker vodi v srečo in blagostanje? Tudi pri racionalnosti prepričanj se odpirajo podobna vprašanja; eno od njih je, ali je doseganje resnice vedno primarni epistemski cilj.

Pojmovanje racionalnosti je pomembno tudi pri razmislekih o različnih vidikih zaupanja v znanost. Na kakšen način je znanje, ki ga pridobivamo z znanstveno metodo, drugačno od znanja, ki ga pridobivajo laiki v vsakdanjem življenju? Koliko prostora za napake in kolikšno stopnjo negotovosti je smiselno dovoliti, ko preverjamo hipoteze? Kakšni dokazi so dovolj dobri, da bomo neko trditev sprejeli ali ovrgli? Odgovori na ta in podobna vprašanja so deloma odvisni od tega, kakšen pogled na racionalnost privzamemo.

Menim, da sta pri preučevanju racionalnosti pomembni tako filozofska analiza kot metode empiričnih znanosti, ki nam dajejo vpogled v procese in mehanizme v ozadju človeškega oblikovanja prepričanj, sklepanja, presojanja in odločanja. Integracija spoznanj različnih disciplin lahko pripomore k zmanjševanju vrzeli med normativnimi in deskriptivnimi teorijami ter pripomore k oblikovanju karseda celostne slike človeške racionalnosti.

VIRI

[1] Keith E. Stanovich. 2010. *Decision making and rationality in the modern world*. Oxford University Press, New York

[2] Phillip E. Tetlock and Barb Mellers, 2002. The great rationality debate. *Psychol Sci* 13, 1 (Jan, 2002), 94–99. DOI: 10.1111/1467-9280.00418

[3] Edward Stein. 1996. *Without good reason: the rationality debate in philosophy and cognitive science*. Clarendon Press, Oxford

[4] Ronald de Sousa. 2007. *Why Think?: Evolution and the Rational Mind*. Oxford University Press, New York

[5] Jon Elster. 1983. *Sour Grapes: studies in the subversion of rationality*. Cambridge University Press, New York

[6] Wolfgang Köhler. 1925. *The Mentality of Apes*. Percy Lund, Humphries & Co, London

[7] John R. Searle. 2001. *Rationality in Action*. MIT Press, London

[8] Robert Nozick. 1981. *The Nature of Rationality*. Princeton University Press, New Yersey

[9] Shaun H. Heap. 1989. *Rationality in economics*. Blackwell Publishing, London

[10] Peter C. Wason. 1966. Reasoning. In *New horizons in psychology*, Pelican, Harmondsworth

[11] Peter C. Wason, 1968. Reasoning about a rule. *Q J Exp Psychol*. 20, 3 (Apr, 1968), 273–281. DOI: 10.1080/14640746808400161

[12] Richard A. Griggs and James R. Cox, 1982. The elusive thematic-materials effect in Wason's selection task. *Br J Psychol* (July, 1982) 73, 407–420. DOI:10.1111/j.2044-8295.1982.tb01823.x

[13] Marco Ragni, Ilir Kola and Phillip N. Johnson-Laird. The Wason selection task: A meta-analysis. In *Proc 32nd Annu Conf Cogn Sci Soc* (July, 2017). Curran Assciates Inc., New York, 980–985

[14] Phillip N. Johnson-Laird and Peter C. Wason, 1970. A theoretical analysis of insight into a reasoning task. *Cogn Psychol* 1, 2 (Feb, 1970), 134–148. DOI:10.1016/0010-0285(70)90009-5

[15] Valerie A. Thompson, Jonathan St. B. T. Evans and Jamie I. D. Campbell, 2013. Matching bias on the selection task: It's fast and feels good. *Think Reason* (Oct, 2013), 19, 3–4, 431–452. DOI:10.1080/13546783.2013.820220

[16] Paul Pollard and Jonathan St. B. T. Evans, 1987. Content and context effects in reasoning. *AJP* 100, 1 (Spring, 1987), 41–60. DOI: 10.2307/1422641

[17] Leda Cosmides and John Tooby, 1992. Cognitive adaptations for social exchange. In *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford University Press, New York, 163–228

[18] Paul S. Davies, James H. Fetzer and Thomas R. Foster, 1995. Logical reasoning and domain specificity - A critique of the social exchange theory of reasoning. *Biol Philos* 10 (Jan, 1995), 1–37. DOI:10.1007/BF00851985

[19] Mike Oaksford and Nick Chater, 1994. A rational analysis of the selection task as optimal data selection. *Psychol Rev* 101, 4 (Oct, 1994), 608–631. DOI:10.1037/0033-295X.101.4.608

[20] Mike Oaksford and Nick Chater, 2003. Optimal data selection: Revision, review, and reevaluation. *Psychon Bull Rev* 10, 2 (May, 2003), 289–318. DOI: 10.3758/BF03196492

[21] Baruch Fischhoff, 2010. Judgment and decision making. *Wiley Interdiscip Rev Cogn Sci* 1, 5 (Sept/Oct, 2010), 724–735. DOI: 10.1002/wcs.65

[22] John von Neumann and Oskar Morgenstern. 1944. *Theory of games and economic behavior*. Princeton University Press, New Yersey

[23] Daniel Kahneman and Amos Tversky, 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47 (Mar, 1979), 263–291. DOI: 10.2307/1914185

[24] Daniel Kahneman, Amos Tversky and Paul Slovic. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge

[25] Daniel Kahneman and Amos Tversky. 2000. *Choices, values and frames*. Cambridge University Press, Cambridge

[26] Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York

[27] Aron K. Barbey and Steven A. Sloman, 2007. Base-rate respect: From ecological rationality to dual processes. *Behav Brain Sci* 30, 3 (Jun, 2007), 255–297. DOI: 10.1017/S0140525X07001653

[28] Gerd Gigerenzer, 1996. The psychology of good judgment: Frequency formats and simple algorithms. *Med Decis Mak* 16, 3 (Jul–Sept, 1996), 273–280. DOI: 10.1177/0272989X9601600312

[29] Gerd Gigerenzer, 1991. How to make cognitive illusions disappear: Beyond "heuristics and biases."*Eur Rev Soc Psychol.* 2, 1 (1991), 83–115. DOI: 10.1080/14792779143000033

[30] Herbert A. Simon, 1955. A behavioral model of rational choice. *Q J Econ* 69, 1 (Feb, 1955), 99–118. DOI: 10.2307/1884852

[31] Hebert A. Simon, 1956. Rational choice and the structure of the environment. *Psychol Rev* 63, 2 (Mar, 1956) 129–138. DOI: 10.1037/h0042769

[32] Herbert A. Simon, 1990. Invariants of human behavior. *Annu Rev Psychol* 41 (Feb, 1990), 1–19. DOI:10.1146/annurev.ps.41.020190.000245

[33] Herbert A. Simon, 1992. What is an explanation of behavior? *Psychol Sci* 3, 3 (May, 1992), 150–161. DOI: 10.1111/j.1467-9280.1992.tb00017.x

[34] Riccardo Viale (ur.). 2021. *Routledge handbook of bounded rationality*. Routledge, New York

[35] Gerd Gigerenzer and Peter M. Todd. 1999. *Simple heuristics that make us smart*. Oxford University Press, New York

[36] Gerd Gigerenzer, Ralph Hertwig and Thorsten Pachur (ur.). 2011. *Heuristics: the foundations of adaptive behavior*. Oxford University Press, New York

[37] Duncan Pritchard. 2018. *What Is This Thing Called Knowledge?* (4th edition.) Routledge: New York

[38] Sven Bernecker and Duncan Pritchard (Ed). 2011. *The Routledge Companion to epistemology*. Routledge, New York

[39] Richard Foley, 1992. The epistemology of belief and the epistemology of degrees of belief. *Am Philos Q* 29, 2 (Apr, 1992), 111–124

[40] Stephen Hartmann and Jan Sprenger, 2011. Bayesian epistemology. In *The routledge companion to epistemology*. Routledge, New York, 636–648

[41] Elizabeth G. Jackson, 2020. Relationship between belief and credence. *Philos Compass* 15, 6 (Mar, 2020), 1–13. DOI: 10.1111/phc3.12668

[42] Carlos E. Alchourrón, Peter Gärdenfors and David C. Makinson (1985). On the logic of theory change: Partial meet contraction and revision functions. *J Symb Log* 50, 2 (June, 1985), 510–530, DOI: 10.2307/2274239

[43] Wolfgang Spohn, 1988. Ordinal conditional functions: A dynamic theory of epistemic states. In *Causation in decision, belief change, and statistics II*. Dordrecht: Kluwer, 105–134

[44] Franz Huber, 2019. Ranking theory. In *The open handbook of formal epistemology*. PhilPapers Foundation, 397–437

Joint history of play provides means for coordination

Liubov Voronina†
MEi:CogSci
University of Ljubljana
Ljubljana, Slovenia
voronina.liuba@gmail.com

Christophe Heintz
Department of Cognitive Science
Central European University
Vienna, Austria
HeintzC@ceu.edu

ABSTRACT

In this study we investigate how joint history shapes strategic decisions for solving coordination problems. We show that coordinating partners use the history of their past interactions to select their strategies. More precisely, people accurately predict that a winning strategy used in the past is mutually salient and can be successfully used again in similar situations. Thus, joint history helps players form accurate mutual expectations about each others' choices and increase the rate of successful coordination.

We demonstrate that precedence is strongly relied upon and provides insights into the psychological bases of the social processes through which conventions emerge. By investigating the path dependence of the individual behaviour in the context of coordination, we experimentally confirm that conventions emerge because people systematically rely on their past interactions in order to coordinate successfully.

KEYWORDS

coordination games, path dependence, Schelling salience

1. INTRODUCTION

Coordination is the process of tacit convergence on a mutual strategy in the context of interdependent decisions. Coordinating partners can choose to do exactly same thing (drive on the right side of the road), exactly the opposite thing (wait while another person is calling back after the line is cut) or complement each other's actions to produce a common outcome (division of the household chores). In many everyday cases coordination is achieved by following an existing convention, by making an explicit verbal agreement or by performing actions in sequence, when the person initiating the interaction has the opportunity to make the first choice and express their preference. However, even in the absence of communication, successful coordination can be accomplished with the probability much higher than chance. Thomas Schelling first draw attention to this apparent paradox of coordination with his informal experiments [1], that were later successfully replicated in the controlled settings [2, 3].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

In pure coordination games (Schelling games) participants are asked to choose the same option from the set of equally attractive ones. Surprisingly, people tend to converge on one particular option at a rate significantly higher than chance. For example, choosing between «heads» or «tails» reveals consistent preference for «heads», much higher than mathematically implied equiprobability. Such recognisable prominence of one alternative over another, that results in a stable solution, is called a focal point or salience.

Pure coordination games therefore pose a question how to identify a unique solution to avoid coordination failure [4]. Although the exact reasoning behind the coordination process is open for debate [5, 6, 7, 8], Schelling's suggestion is to look for such selection rule among many, which can single out a successful coordination strategy. This rule should be mutually recognised by the interacting parties to be able to provide reliable means for coordination [1]. A focal point, emerged by applying such selection rule, is called Schelling salience.

Building on the logic of coordination games, David Lewis convincingly argued for the emergence of (linguistic) conventions [9]. According to his account, observed behavioural regularities that are commonly known among the population, create accurate mutual expectations that facilitate coordination by providing unambiguous solution to social coordination problems, resulting in stable equilibria.

We hypothesise that these behavioural regularities become salient by virtue of repeating precedence, which is used as Schelling salience, once the agents are confronted with the coordination problem.

The goal of the study is to show how the joint history of interactions in coordination problems shapes the choice of coordination strategies. At the cognitive level, this means that joint history is used by people as relevant information for choosing their strategy for coordination.

The following hypotheses were tested:

H1: Joint history facilitates accurate mutual expectations.

Players choose a coordination strategy in view of what they expect their partner to do. These expectations are informed by the knowledge they have of their joint history, which makes their prediction more accurate.

H2: Joint history determines coordination strategies.

When the situation does not provide any unambiguous clues for coordination, players choose a specific strategy that resulted in successful coordination in the past to resolve the ambiguity and avoid coordination failure.

2. METHOD

This research is based on the empirical methodology of experimental game theory. The economic game chosen for the experiment is pure coordination game [1, 2, 3].

In the experiment, participants were presented with various layouts of coloured tokens and asked to coordinate on the token of the same colour. Both sets of tokens were visible to both partners and the choice was simultaneous. The result of every interaction and individual players' choices were logged online in real time. The analysis was carried out for the particular type of rounds (at individual or dyadic level) with the condition as an independent variable.

2.1. Participants

One hundred and thirty-three participants took part in the "Mobile Coordination Games" experiment, which was conducted online in two parts. Game sessions for the baseline condition were organised during June and July, 2021 with a total of 51 participants (mean age = 26.2 years; 16 females and 35 males). Game sessions for the experimental conditions took place in January, 2022 with a total of 82 participants (mean age = 24.1 years; 24 females and 58 males). Participants were recruited online via the Sona Research Participation System of Central European University. There were no restrictions on participation for the adult participants, who needed basic English skills for understanding the instructions and a mobile device for accessing the Coordy research application. All the participants received compensation based on their performance level (average amount = 4,9 euros) in the form of an online voucher of their preference, either Amazon or PayPal.

2.2. Materials

To enable empirical investigation of the real-time coordination between the pairs of participants, a proprietary mobile research application named Coordy has been developed for both Android and iOS based mobile devices. Coordy was officially released and became available for download on Google PlayMarket and AppStore.

In the experiment, we used two different kinds of experimental scenarios:

- 30 single rounds of various difficulty in the baseline condition:

A. easy rounds with the symmetrical clues for coordination;
B. hard rounds with the clashing clues;
C. equiprobable rounds with no coordination clues;

- 46 games of five rounds (experimental conditions).

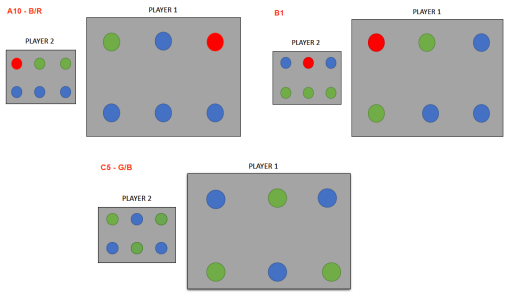


Figure 1: Examples of A, B and C rounds.

2.2. Experimental conditions

We used a between-subject design to examine the research hypotheses. Participants were presented with the experimental scenarios under the following three conditions:

(1) no joint history (baseline condition)

Baseline data reveals the rate of coordination in the absence of joint history of play. The baseline condition also helps to empirically differentiate various types of rounds that are used to construct scenarios in the experimental conditions. Rounds that reveal preference for one particular colour will become history rounds. Rounds, where the colour choices are equally distributed, will become test rounds. Coordination index [2, 3] is calculated to show the hypothetical coordination rate of the unpaired participants based on their individual responses.

Experimental setting: Participants play single rounds in pairs with their player IDs hidden. They connect with the new partner after each round and are aware of this. The participants pool is set to 2 players to allow random pairing.

Experimental stimuli: 30 individual rounds of various difficulty. Each round could be played for up to 2 times by any player (but not in sequence).

(2) random joint history (experimental condition)

Participants have the opportunity to build a joint history of play, consisting of randomly assigned rounds. This history of mutual interactions can provide them with the clues for successful coordination in the test round. Its coordination rate will be compared to the corresponding baseline rate and the coordination index.

Experimental setting: Participants play games, consisting of four random rounds and a test round, in dyads with their player IDs shown. They change their game partner after each game. The participants pool is set to 8 players to allow fixed pairing in order to avoid repetitions.

Experimental stimuli: 36 games of 5 rounds from the baseline condition (6 unique histories of four rounds combined with each of the 6 test rounds). Each game was played just once during the game session.

(3) specified joint history (experimental condition)

Joint history, provided by the designed scenario, increases the probability that a certain strategy is used during this history and, subsequently, in the test round. Individual player's strategy, operationalised as a choice of the specific colour, will be compared between different histories that end up with the same test round.

Experimental setting: Participants play games, consisting of four predefined rounds and a test round, in dyads with their player IDs shown. They change their game partner after each game. The participants pool is set to 8 players to allow fixed pairing in order to avoid repetitions.

Experimental stimuli: 10 games of 5 rounds (5 unique histories of four rounds combined with the corresponding test rounds). Each game was played twice during the game session.

All the scenarios (both single round and games of five rounds) appeared during the game sessions in the randomised order to avoid order effects. The order of rounds within a particular game was fixed. Both experimental conditions were tested together during the same game sessions.

3. RESULTS

Before reporting the results of the study, we would like to clarify the issue of the players' expertise and its potential influence on the outcome of coordination. In both conditions, all the participants would start playing without any prior experience (match number 0). We analysed the outcome of the coordination in the last round (success or failure) for match numbers below and above 7 (half of the experimental game sessions) and found no evidence for the improvement of the coordination success at the dyadic level. Coordination in the last round was successful in about half of the games irrespective of the participants' level of experience with the task.

3.1. Baseline results

A chi-square test of goodness-of-fit was performed to determine whether each of the three colours were equally chosen by the participants in the particular round.

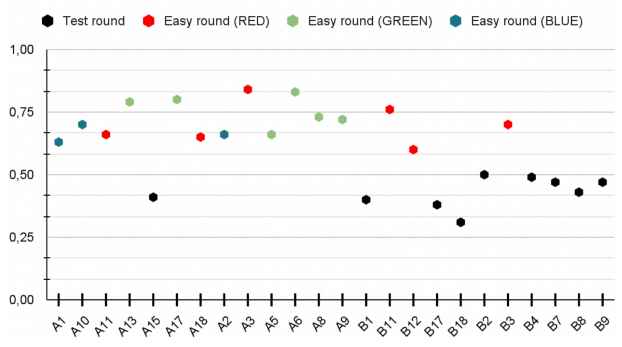


Figure 2: Baseline coordination rates for the A, B rounds.

A preference for a specific colour was found in the majority of A rounds (except for A15) and some B rounds (B3, B11, B12). The corresponding dots on the graph are coloured with the colour that was chosen the most (over ⅔ of the individual choices) in the particular round. Also the coordination rates (CRs) for those rounds were very high (mean CR = 0.72). These rounds were used to constitute history rounds.

A preference for the specific colour was not found in the four B rounds (B1, B8, B17, B18) and one A round (A15). While the choices for the three colours were not equally distributed in the rounds B2, B4, B7 and B9, the proportion of any particular colour did not exceed 60%. Their corresponding dots on the graph are therefore coloured in black. Also their CRs were significantly lower (mean CR = 0.43) than in the previous group of rounds with the focal points. These rounds were used as test rounds in the random history condition.

For the majority of C rounds (C1, C3, C4, C6) a colour preference was not established. Also the CRs for C rounds were not significantly higher than chance (mean CR = 0.57). Hence they are not depicted on the graph. These rounds were used as test round in the specified history condition.

3.2. Random history results

A chi-square test of independence was performed to examine the relation between the coordination rate in the test rounds (at the group level) and the history of previous interactions. The relation between these variables was significant, X^2 (1, N = 592) = 8.39, $p < .01$; $r = .12$.

Participants were more likely to successfully coordinate in the test rounds after the joint history of play than without it.

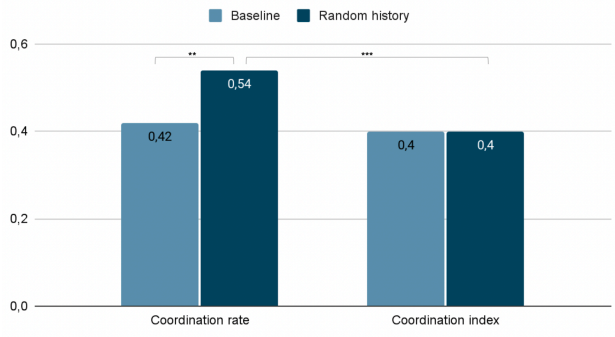


Figure 3: Change in the coordination rate and coordination index across conditions (group-level). Levels of significance: **: p < .01, ***: p < .001.

A two-proportion z-test was conducted to calculate the difference between the CR and coordination index (CI) in the last round of the games with random history. For the group of test B rounds CR was found to be significantly higher than CI after the joint history of play z (N = 1029) = 4.26, $p < .001$; $r = .13$.

Therefore the actual coordination rate exceeds the rate of the expected coordination, when the choices are made by the randomly paired participants.

3.3. Specified history results

A chi-square test of independence was performed to examine the relation between the individual player's choice in the same test round and the specific history preceding that round.

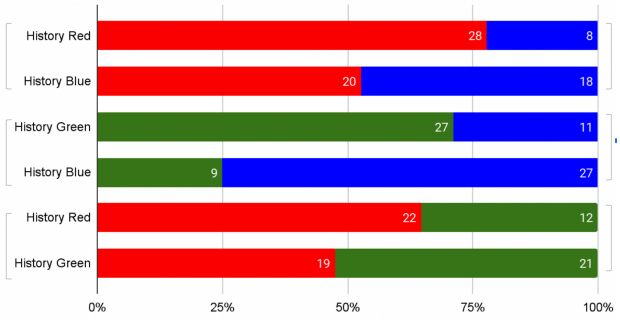


Figure 4: Individual player's choice of colour for coordination in the last round after the specified history.

For some pairs of histories, the relation between these variables was significant:

- for the test round C4 after the histories Red and Blue X^2 (1, N = 74) = 7.29, $p < .01$; $r = .3$;
- for the test round C2 after the histories Green and Blue X^2 (1, N = 74) = 15.69, $p < .0001$; $r = .45$.

In the same test round, participants were more likely to choose the modal colour of the history rounds (e.g. in the test round C4 participants were more likely to choose red colour after the history Red and blue colour after the history Blue).

For the histories Red and Green the relation between these variables was statistically insignificant.

4. DISCUSSION

In this study we aimed to investigate how previous interactions can influence the outcome of coordination for the pair of players. First, we let the participants play single rounds anonymously. Even though the participants played several rounds, they could not constitute joint history of interaction because they were — knowingly — paired with a new random participants for each round. This set-up helped us identify rounds with the «natural» focal points, i.e. colours that appealed to the participants as obvious to coordinate upon due to the specific layout of the scenario, irrespective of other factors.

We noticed that in the absence of communication and any explicit coordination rules, participants did manage to coordinate more than rational choice theory would predict. This is in line with previous results showing that people are able to rely on Schelling salience in order to coordinate successfully.

In our experiment participants converged on a tacit rule for coordination, which was «choose colour with the most tokens present on both players' layouts»**.** Those rounds, where this rule could not be unmistakably applied, demonstrated lower coordination rates and were chosen to be the test rounds for the subsequent experimental conditions. We wanted to explore the possibility that the history of interactions itself would provide Schelling salience and thus determine the choice of colour to increase the coordination rate.

We then created games with five rounds, which were played by the participants in dyads with their IDs shown and mutually known, thus letting them build the history of mutual interactions. In the games, which histories did not suggest a choice of any specific colour, we observed a significant improvement coordination in the last round. Interestingly, the coordination index for this round did not significantly changed between the conditions with and without joint history. It is only the actual coordination rate that changed. In other words, had the participants been paired randomly for the last round, no improvement would have occurred.

This suggests that the increase in coordination rate is due to players tracking what they have played with their own partner and using this information to make their future choices. This findings confirm our hypothesis that joint history of play facilitates coordination. When the game partners are aware of each other's previous choices, they tend to choose the focal point for coordination more accurately. However, the effect size of the observed differences remained small. One possible explanation is that randomness and variety of the history rounds created clashing focal points to converge on.

In the games from the third condition, where a history of rounds nudged the choice of a given colour, we observed that this same colour tended to be chosen in the last round. More precisely, participants had to select one of two colours in the last round of their joint history. They tended to select the same colour, on which they coordinated during their joint history. They did so significantly more than the participants, who were given a joint history that nudged towards another colour. We documented that effect for two test rounds with a moderate effect size. We did not observe a significant effect for the third test round.

Our post-hoc hypothesis is that the salience of the red colour overshadowed the salience shaped by the history of past interactions. This mixed result calls for the replication study with the different set of stimuli.

Overall, in our experiment we managed to observe how participants make use of the precedence by applying the following rule for coordination: «choose the colour that brought us successful coordination before». Though studying path dependence in the lab setting poses certain challenges [10], some researchers found the way to address them using the economic games [11]. In the future, it could be fruitful to empirically investigate the robustness of the coordination rules and the amount of common knowledge required for their emergence [12, 13, 14, 15].

ACKNOWLEDGMENTS

The authors acknowledge financial support by the European Research Council, under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 609819 (Somics project).

REFERENCES

1. Thomas Schelling. 1960. *The strategy of conflict* (1st. ed.). Harvard University Press, Cambridge, MA.
2. Judith Mehta, Chris Starmer, and Robert Sugden, 1994. The Nature of Salience: An Experimental Investigation of Pure Coordination Games. *The American Economic Review* 84, 3 (Jun, 1994), 658–73. JSTOR: http://www.jstor.org/stable/2118074.
3. Nicholas Bardsley, Judith Mehta, Chris Starmer and Robert Sugden, 2010. Explaining Focal Points: Cognitive Hierarchy Theory *versus* Team Reasoning, *The Economic Journal* 120, 543 (Mar, 2010), 40–79. DOI:https://https://doi.org/10.1111/j.1468-0297.2009.02304.x.
4. Giovanna Devetag and Andreas Ortmann, 2007. When and why? A critical survey on coordination failure in the laboratory. *Experimental economics* 10 (Aug, 2007), 331-344. DOI:https://doi.org/10.1007/s10683-007-9178-9.
5. Robert Sugden. The logic of team reasoning, 2003. *Philosophical explorations* 6, 3 (Sep, 2003), 165-181. DOI:https://doi.org/10.1080/10002003098538748.
6. Colin F. Camerer, Teck-Hua Ho and Juin-Kuan Chong, 2004. A Cognitive Hierarchy Model of Games, *The Quarterly Journal of Economics* 119, 3 (Aug, 2004), 861–898. DOI:https://doi.org/10.1162/0033553041502225.
7. Nicholas Bardsley and Aljaž Ule, 2017. Focal points revisited: Team reasoning, the principle of insufficient reason and cognitive hierarchy theory. *Journal of Economic Behavior & Organization* 133 (Jan, 2017), 74-86. DOI:https://doi.org/10.1016/j.jebo.2016.10.004.
8. Francesco Guala, 2018. Coordination, Team Reasoning, and Solution Thinking. *Revue d'économie politique* 128, 3, 355-372. DOI:https://doi.org/10.3917/redp.283.0355.
9. David Lewis. 1969. *Conventions: A Philosophical Study* (1st. ed.). Harvard University Press, Cambridge, MA.
10. Jean-Philippe Vergne and Rodolphe Durand, 2010. The missing link between the theory and empirics of path dependence: conceptual clarification, testability issue, and methodological implications, *Journal of Management Studies* 47, 4 (Aug, 2010), 736-759. DOI:https://doi.org/10.1111/j.1467-6486.2009.00913.x
11. Marc Knez and Colin Camerer, 2000. Increasing cooperation in prisoner's dilemmas by establishing a precedent of efficiency in coordination games, *Organizational Behavior and Human Decision Processes* 82, 2 (Jul, 200), 194-216. DOI:https://doi.org/10.1006/obhd.2000.2882
12. Margaret Gilbert, 1983. Agreements, conventions, and language. *Synthese* 54, 3 (Mar, 1983), 375-407. JSTOR: http://www.jstor.org/stable/20115846.
13. Ken Binmore, 2008. Do conventions need to be common knowledge? *Topoi* 27 (Jul, 2008), 17-27. DOI:https://doi.org/10.1007/s11245-008-9033-4.
14. Giacomo Sillari, 2008. Common Knowledge and Convention. *Topoi* 27 (Jul, 2008), 29-39. DOI:https://doi.org/10.1007/s11245-008-9030-7.
15. Kyle A. Thomas, Peter DeScioli, Omar Sultan Haque, and Steven Pinker, 2014. The Psychology of Coordination and Common Knowledge, *Journal of Personality and Social Psychology* 107, 4, 657–676. DOI:https://doi:10.1037/a0037037.

Predicting Trust in Science in the Context of COVID-19 Pandemic: The Role of Sociodemographics and Social Media Use

Žan Zelič
Department of Psychology
University of Ljubljana
Ljubljana, Slovenia
zan.zelic@gmail.com

Martin Berič
Department of Psychology
University of Ljubljana
Ljubljana, Slovenia
martinberic@yahoo.com

Darja Kobal Grum
Department of Psychology
University of Ljubljana
Ljubljana, Slovenia
darja.kobal@ff.uni-lj.si

ABSTRACT

Research in the context of COVID-19 pandemic has consistently shown that scientific distrust adversely affects health-related behavior. Therefore, the aim of our study was to identify the risk factors for the development of scientific distrust, with emphasis on the role of sociodemographic variables and social media use. A convenience sample of 490 Slovene speaking individuals was used to perform hierarchical linear regression analysis. In line with our hypotheses, the results showed that trust in science was negatively correlated with age, religiosity and use of social media as an information source about COVID-19, while it was positively correlated with male gender and total years of formal education. When only sociodemographic variables were entered into the prediction model, each of them explained a significant proportion of the variance in trust in science. However, after the inclusion of social media use, religiosity was no longer a significant predictor. In contrast to our expectations, the results also showed no significant interaction between education and social media use when predicting trust in science. Our findings are further discussed and additional implications are provided.

KEYWORDS

COVID-19, trust in science, social media use, education, religiosity

1 INTRODUCTION

When the new coronavirus (Sars-CoV-2) started to spread in 2020 it has quickly become evident that the world as we knew it was about to change. Ever increasing number of infections led to health system overloads, high mortality rates, mental health difficulties and great economic burden [1]. As adoption of social distancing measures and newly developed vaccines was crucial for reducing the spread of the new coronavirus and its adverse consequences, identification of factors influencing health-related behavior became of utmost importance. One of the variables that has been consistently found to predict preventive behavior as well as vaccine acceptance is trust in science [2].

According to Barber [3] public trust in science depends on the perceptions of scientists' compliance with technical and moral norms. Technical norms consist of expectations that the scientists will perform an assigned task with a certain level of competence and expertise, while moral norms are related to the anticipation that by doing so, they will also act in a way that puts the interest of the community before their personal advances. Similarly, Wintterlin et al. [4] argue that trust in science is rooted in expectations that scientists' claims are epistemically sound and that science has a prosocial stance. Overall, perhaps the most comprehensive definition of trust in science has been provided by Nadelson et al. [5], describing it as a multifaceted construct, which includes affective components, credibility and trustworthiness perceptions, knowledge and epistemic beliefs.

Since scientists were the main source of information on COVID-19 and its adverse consequences, and also the ones that helped governments develop preventive measures and vaccines, the findings that low trust in science negatively impacts health-related behavior [2] should not come as a complete surprise. However, not much has been researched about the predictors of trust in science in the context of the pandemic. As we believe this kind of knowledge is crucial to implement communication changes, which could accurately address those who are particularly prone to developing scientific mistrust, we conducted a study focusing on the sociodemographic predictors of trust in science as well as its connection to social media use.

1.1 Predictors of Trust in Science

Previous research on the relationship between trust in science and age has shown somewhat mixed results. For example, some researchers reported on non-significant correlations [6], while others found that scientists were more likely to be trusted by those who are younger [7]. The latter result could in part be explained by higher average levels of education among younger individuals, however age remained an important predictor even when education was accounted for [7].

Regarding gender, previous research has consistently shown that men generally have more positive attitudes towards science than women [6][8]. However, when possible reasons for these results were examined, other sociodemographic variables, such as education, religiosity and work status were found to explain this gender gap [8].

Throughout history, religion and science were often seen as epistemologically conflicted [9], which may have resulted in lower trust in science by those who are more religious. Indeed, previous research has shown that religiosity was associated with negative attitudes towards science as well as lower science literacy [5][10].

Another sociodemographic factor that has been consistently shown to predict trust in science is education [6]. One of the most prevalent explanations for the described relationship was that

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia
© 2022 Copyright held by the owner/author(s).

education indirectly influences positive attitudes towards science by increasing scientific knowledge [11]. However, further research showed that education remained an important predictor of trust even when controlling for scientific knowledge [12].

Although previous research has indicated that social media use positively predicts trust in science [13], we believe that the results might be different in the times of COVID-19 pandemic. Since social platforms enabled rapid misinformation dispersion [14], extensive social media use could lower trust in science by increasing conspiracy beliefs about scientists' involvement in the pandemic. Indeed, our previous research [15] showed that the extent of using social media as an information source predicted COVID-19 conspiracy beliefs, which were also highly inversely correlated with trust in science.

1.2 The Present Research

The aim of our study was to examine the importance of several sociodemographic variables and social media use in predicting trust in science. Based on the previous findings we hypothesized that trust in science will be higher among younger individuals (H1), men (H2), those who are less religious (H3), more educated (H4) and those who obtained less information about the coronavirus from the social media (H5). Additionally, we hypothesized that education would have a moderating role in the relationship between social media use and trust in science (H6). Since critical thinking has been found to develop through education [16], we assumed that even extensive social media use would not reflect in high levels of scientific distrust as long as individuals would be capable to critically evaluate the quality of obtained information. Furthermore, we also aimed to investigate the amount of the variance in trust in science that a combination of these variables could explain as well as their relative importance when entered into a multivariate prediction model.

2 METHOD

2.1 Sample

Data collection took place between March 29 and April 7, 2021, using an online survey. Convenience sample was used, consisting mostly of students at the University of Ljubljana and members of different COVID-19 related Facebook groups. Responses of 490 participants (397 women, 92 men and one non-binary), aged from 18 to 70 years (M = 35.7, SD = 13.2), were analyzed. The majority (56.5%) of the participants had a college degree, 41.8% reported on having a high school diploma and 1.6% completed only elementary school. Furthermore, 31.6% of them were students, 54.7% were employed, 9.0% were unemployed and 4.7% were retired.

2.2 Measures

Demographic data was obtained through a series of questions on age, gender, years of education and employment status.

Religiosity was measured by the participants' level of agreement with the statement "*I would define myself as a religious person."* on a 7-point Likert scale with anchors, 1 (*Strongly disagree*) and 7 (*Strongly agree*).

Use of social media as an information source about COVID-19 was measured by moving an interactive slider between values 0 and 100 to estimate the percentage of information about the new coronavirus they obtained through social media.

Trust in Science was measured by the Trust in Science and Scientists Inventory [5], which contains 21 items (e.g., *We can trust science to find answers that explain the natural world.*). Participants were asked to rate their agreement with the provided statements on a 5-point Likert scale with anchors, 1 (*Strongly disagree*) and 5 (*Strongly agree*). Confirmatory factor analysis (CFA) showed poor one-factor model fit, so we excluded item 11, which was semantically very similar to items 9 and 10. Additionally, we allowed for some residual covariances according to modification indexes. The fit of the modified 20-item scale was acceptable: $\chi 2(166) = 484.642$, $p < .001$, CFI = .939, TLI = .930, RMSEA = .070, 90% CI: [.063, .078], SRMR = .042. The shorter version of the scale also showed excellent internal consistency ($\alpha = .95$).

3 RESULTS

Firstly, the factor structure of the translated Trust in Science and Scientist Inventory was assessed by confirmatory factor analysis (CFA), using R package lavaan [17]. Since the data were non-normally distributed, we used the robust maximum likelihood method (MLM) of model estimation. After minor modifications were implemented to achieve an acceptable one-factor model fit, the total trust in science score was calculated as a mean value of all items. All further analyses were done in IBM SPSS version 25.0 [18].

Secondly, descriptive statistics and intercorrelations were calculated for all measured variables. The results showed that trust in science was negatively correlated with age (r = -.14, p = .002), religiosity (r = -.16, $p < .001$) and use of social media as an information source about COVID-19 (r = -.35, $p < .001$), while it was positively correlated with male gender (r_{pb} = .21, $p < .001$) and total years of formal education (r = .29, $p < .001$).

Thirdly, when we determined that all assumptions for multiple linear regression were met, hierarchical linear regression analysis was conducted. Trust in science was entered into the analysis as a criterion variable, while all other measured variables were consecutively added as predictors (see Table 1). In the first step age and gender together explained 6.5% of variance in trust in science with younger individuals and men exhibiting more trust. Both predictors were significant, although the relative importance of gender was greater. In the second step religiosity explained only 1.8% of additional variance in trust in science, however the change in R^2 was statistically significant. Those who were less religious showed significantly higher levels of trust even when age and gender were accounted for. Furthermore, both age and gender remained significant predictors of trust despite slight decrease of gender's β value. In the third step the years of formal education turned up to be the most important positive predictor of trust in science, additionally explaining 6.7% of its variance. Inclusion of education slightly lowered the β values of age and religiosity, however all included predictors remained statistically significant. In the fourth step the share of COVID-19 information obtained from social media was added into the equation, explaining an additional 5.8% of variance in trust in science. The results showed that those who relied more on social networks to obtain information were less likely to trust in science. Altogether, a combination of five

Table 1: The results of the hierarchical linear regression analysis

	Step 1			Step 2			Step 3			Step 4			Step 5		
Variable	***B***	***SE***	**β**	***B***	***SE***	**β**	***B***	***SE***	**β**	***B***	***SE***	**β**	***B***	***SE***	**β**
Age	-.01	.00	-.14**	-.01	.00	-.14**	-.01	.00	-.11**	-.01	.00	-.11**	-.01	.00	-.11**
Male gender	.46	.10	.21***	.42	.10	.19***	.41	.09	.19***	.28	.09	.13**	.28	.09	.13**
Religiosity				-.07	.02	-.14**	-.05	.02	-.11*	-.03	.02	-.07	-.03	.02	-.07
Education							.10	.02	.26***	.08	.02	.21***	.09	.03	.22**
SM information										-.01	.00	-.26***	-.01	.01	-.21
Edu x SM info													.00	.00	-.05
R^2			.064			.083			.149			.207			.207
$\triangle R^2$			.064***			.018**			.067***			.058***			.000

Note. *$p < .05$. **$p < .01$. ***$p < .001$.

predictors explained 20.7% of variance in trust in science. However, after the variable of social media use was included, β values of gender, education and religiosity decreased, thus designating religiosity as a non-significant predictor. Finally, analysis in the fifth step showed that there was no significant interaction between education and social media use when predicting trust in science.

4 DISCUSSION AND CONCLUSIONS

The aim of our research was to examine the predictors of trust in science in the context of COVID-19 pandemic since such knowledge could be used to implement communication changes that might motivate higher compliance with preventive measures and protect public health.

Regarding age, the results were in line with our assumption that younger individuals are more likely to trust in science (H1). Although our finding is supported by some of the previous research [7], it is still somewhat surprising, since general trust is known to increase with age [19]. Negative relationship between trust in science and age could be explained by lower average educational levels among the elderly, as both knowledge about science and certain cognitive skills, which are thought to be related to higher trust in science [6][20] are developed through education [16][21]. Indeed, in our study age and education were negatively correlated (r = -.12, p = .010), however age remained a significant predictor even when education was controlled for. Another possibility may be that the relationship between age and trust in science is underlaid by religiosity, as previous research showed that older individuals are more likely to be religious [22] and that religiosity also predicts lower trust in science [10]. However, our results showed that religiosity and age were not significantly correlated (r = -.03, p = .504), therefore undermining the described reasoning.

The results of our study were also in line with the assumption that male gender would be positively related to trust in science (H2). Even though some of the previous studies [8] indicated that this relationship could be entirely accounted for by other sociodemographic variables, we found that gender remained a significant predictor of trust in science even when age, religiosity and years of education were controlled. One possible explanation for this result may be that on average women have less specific science-related knowledge than men. Although in our research male gender was not significantly related to years of total education (r_{pb} = .04, p = .440), education of men and women might differ in terms of its type and field of interest. For example, Global Gender Gap Report 2022 showed that only 33% of STEM graduates in Slovenia are female [23]. In line with the above, Fox & Firebaugh [24] also found that years of education did not explain the gender gap in science confidence. Moreover, their research pointed out that gender differences can in large part be attributed to lower perceived utility of science by women.

Based on previous studies, which showed that religiosity predicts negative attitudes towards science [10], we also hypothesized that religiosity would be negatively associated with trust in science (H3). The results were in line with our assumption, however when in addition to all other sociodemographic variables, social media use was inserted into the model, religiosity was no longer a significant predictor of trust in science. Indeed, an unusual positive correlation could be observed between religiosity and social media use as an information source about COVID-19 (r = .21, $p < .001$). A possible explanation for this phenomenon may be that social media use is highly prevalent among religious individuals since social networks are often seen as channels that can be used to effectively minister to others [25]. Obtaining (mis)information from social media may thus be a side effect of extensive use of social networks for other purposes. An alternative explanation may also be that religious individuals are more likely to adopt conspiracy beliefs [15]. Since conspiracy ideation is likely to influence the perception of traditional media as deceiving [26], those who are more religious may thus be inclined to use informal sources of information, such as social media.

Regarding education and social media use, the results supported both of our hypotheses that trust in science would be positively related to years of education (H4) and negatively related to perceived share of information about COVID-19 that was obtained on social media (H5). Although more educated

individuals were also less religious (r = -.10, p = .023) and obtained smaller share of information on social media (r = -21, p < .001), education remained an important predictor of trust in science even when other variables were controlled. As previously suggested, this could be explained by the fact that critical thinking, which is thought to interrelate with trust in science [20], develops through education [16]. Furthermore, in contrast to previous research that reported on the positive relationship between social media use and trust in science [13], our results showed that in the times of the COVID-19 pandemic obtaining information from social media might in fact be detrimental for trust in science. Since social media's regulations on shared content are less strict compared to the traditional media, we believe the quick dispersion of COVID-19 conspiracy beliefs through social media could lower trust in science. Additionally, we hypothesized that social media use would not reflect in high levels of scientific distrust as long as the individuals would be sufficiently educated (H6). We assumed that well educated individuals would be able to critically evaluate the quality of obtained information due to their advanced critical thinking skills [16]. In contrast to our expectations, the results showed that there was no significant interaction between education and social media use when predicting trust in science. In our opinion, this finding could be based on the fact that: a) years of education are not a valid indicator of critical thinking skills, or b) that critical thinking abilities are somewhat irrelevant in the case when one's information space is so limited that they do not have any relevant data upon which information from social media could be judged.

To conclude, our findings suggest that in order to restore trust in science and reinforce health-related behavior in the context of the pandemic, it would be expedient to develop communication strategies that would specifically target older women, who are less educated, more religious and are extensive social media users. However, these findings are subjected to some limitations of our research design. Firstly, the data may not be entirely representative due to the convenience sampling method. Secondly, correlational design of our study does not allow for causal inferences. And thirdly, the used trust in science measure was one-dimensional, although some researchers argue that it is necessary to differentiate between trust in scientific methods and trust in scientific institutions [27]. Therefore, our suggestion for future research would be to examine how these two distinct forms of trust in science relate to health behavior and to identify which are the most important risk factors for either of them.

ACKNOWLEDGMENTS

Research was supported by the Slovenian research agency [grant number P5-0110].

REFERENCES

[1] Office for National Statistics, 2021. *Leaving no one behind – a review of who has been most affected by the coronavirus pandemic in the UK: December 2021*. Office for National Statistics, Newport. Available at: https://www.ons.gov.uk/economy/environmentalaccounts/articles/leavingnoonebehindareviewofwhohasbeenmostaffectedbythecoronaviruspandemicintheuk/december2021#economic-impact

[2] Yann Algan Eva Davoine, Martial Foucault and Stefanie Stantcheva, 2021. Trust in scientists in times of pandemic: Panel evidence from 12 countries. *Proceedings of the National Academy of Sciences of the United States of America* 118, 40 (Sep, 2021), Article e2108576118. DOI: 10.1073/pnas.2108576118

[3] Bernard Barber, 1987. Trust in science. *Minerva* 25 (Mar, 1987), 123–134 DOI: 10.1007/BF01096860

[4] Florian Wintterlin, Friederike Hendriks, Niels Mede, Rainer Bromme, Julia Metag and Mike Schäfer, 2022. Predicting public trust in science: The role of basic orientations toward science, perceived trustworthiness of scientists, and experiences with science. *Frontiers in Communication* 6 (Jan, 2022). DOI: 10.3389/fcomm.2021.822757

[5] Louis Nadelson, Cheryl Jorcyk, Dazhi Yang, Mary Jarratt Smith, Sam Matson, Ken Cornell and Virginia Husting, 2014. Trust in science and scientists. *School Science and Mathematics* 114 (Jan, 2014), 76–86. DOI: 10.1111/ssm.12051

[6] Fabienne Crettaz von Roten, 2004. Gender differences in attitudes toward science in Switzerland. *Public Understanding of Science,* 13 (Apr, 2004) 191–199. DOI: 10.1177/0963662504043870

[7] Paul Brewer and Barbara Ley, 2013. Whose science do you believe? Explaining trust in sources of scientific information about the environment. *Science Communication* 35, 1 (Feb, 2013), 115–137. DOI: 10.1177/1075547012441691

[8] Bernadette C. Hayes and Vicki N. Tariq, 2000. Gender differences in scientific knowledge and attitudes toward science: a comparative study of four Anglo-American nations. *Public Understanding of Science* 9 (Oct, 2000), 433–447. DOI: 10.1088/0963-6625/9/4/306

[9] John Evans and Michael Evans, 2008. Religion and science: Beyond the epistemological conflict narrative. *Annual Review of Sociology* 34 (Aug, 2008), 87–105. DOI: 10.1146/annurev.soc.34.040507.134702

[10] Jonathon McPhetres and Miron Zuckerman, 2018. Religiosity predicts negative attitudes towards science and lower levels of science literacy. *PLoS ONE* 13, 11 (Nov, 2018), Article e0207125. DOI: 10.1371/journal.pone.0207125

[11] Amitai Etzioni and Clyde Nunn, 1974. The public appreciation of science in contemporary America. *Daedalus* 103, 3 (Jan, 1974), 191–205. DOI: 10.1007/978-94-010-1887-6_15

[12] Hee-Je Bak, 2001. Education and public attitudes toward science: Implications for the "Deficit Model" of education and support for science and technology. *Social Science Quarterly* 82, 4 (Dec, 2001), 779–795. DOI: 10.1111/0038-4941.00059

[13] Brigitte Huber, Matthew Barnidge, Homero Gil de Zúñiga and James Liu, 2019. Fostering public trust in science: The role of social media. *Public understanding of science* 28, 7 (Sep, 2019), 759–777. DOI: 10.1177/0963662519869097

[14] Ramez Kouzy, Joseph Abi Jaoude, Atif Kraitem, Molly B. El Alam, Basil Karam, Elio Adib, Jabra Zarka et al., 2020. Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus* 12, 3 (Mar, 2020), Article e7255. DOI: 10.7759/cureus.7255

[15] Žan Zelič, Martin Berič and Darja Kobal Grum, 2022. Examining the role of Covid-19 conspiracy beliefs in predicting vaccination intentions, preventive behavior and willingness to share opinions about the coronavirus. *Studia Psychologica* 64, 1 (Mar 2022), 136–153. DOI: 10.31577/sp.2022.01.844

[16] Christopher R. Huber and Nathan R. Kuncel, 2016. Does college teach critical thinking? A meta-analysis. *Review of Educational Research* 86, 2 (Jun, 2016), 431–468. DOI: 10.3102/0034654315605917

[17] Yves Rosseel, 2012. lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 48, 2 (May, 2012), 1–36. DOI: 10.18637/jss.v048.i02

[18] IBM Corp. 2017. IBM SPSS Statistics for Windows, Version 25.0. IBM Corp., Armonk, NY.

[19] Michael Poulin and Claudia Haase, 2015. Growing to trust: Evidence that trust increases and sustains well-being across the life span. *Social Psychological and Personality Science* 6, 6 (Mar, 2015), 614–621. DOI: 10.1177/1948550615574301

[20] John Kleinig, 2016. Trust and critical thinking. *Educational Philosophy and Theory* 50, 2 (Jun, 2016), 1–11. DOI: 10.1080/00131857.2016.1144167

[21] Bryan Kennedy and Meg Hefferon, 2019. *What Americans know about science?* Pew Research Center, Washington, DC. Available at: https://www.pewresearch.org/science/2019/03/28/what-americans-know-about-science/

[22] Vern L. Bengtson, Merril Silverstein, Norella M. Putney and Susan C. Harris, 2015. Does religiousness increase with age? Age changes and generational differences over 35 years. *Journal for the Scientific Study of Religion* 54, 2 (Sep, 2015), 363–379. DOI: 10.1111/jssr.12183

[23] *Global Gender Gap Report 2022*. World Economic Forum, Geneva. Available at: https://www.weforum.org/reports/global-gender-gap-report-2022/

[24] Mary F. Fox and Glenn Firebaugh, 1992. Confidence in science: The gender gap. *Social Science Quarterly* 73, 1 (Mar, 1992), 101–113.

[25] Pamela J. Brubaker and Michel M. Haigh, 2017. The religious Facebook experience: Uses and gratifications of faith-based content. *Social Media + Society* 3, 2 (Apr, 2017), 1–11. DOI: 10.1177/2056305117703723

[26] Stephen Marmura, 2014. Likely and Unlikely Stories: Conspiracy Theories in an Age of Propaganda. *International Journal of Communication 8*, 1 (May, 2014), 2377–2395.

[27] Peter Achterberg, Willem de Koster and Jeroen van der Waal, 2017. A science confidence gap: Education, trust in scientific methods, and trust in scientific institutions in the United States, 2014. *Public understanding of science* 26, 6 (Dec, 2015), 704–720. DOI: 10.1177/0963662515617367

Indeks avtorjev / Author index

Kognitivna znanost

Cognitive Science

Uredniki • Editors:
Toma Strle, Borut Trpin, Olga Markič