Zbornik 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2024

Zvezek B

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume B

Kognitivna znanost
Cognitive Science

Uredniki / Editors

Anka Slana Ozimič, Borut Trpin, Toma Strle, Olga Markič

http://is.ijs.si

10. oktober 2024 / 10 October 2024
Ljubljana, Slovenia

Uredniki:

Anka Slana Ozimič
Filozofska fakulteta, Univerza v Ljubljani

Borut Trpin
Filozofska fakulteta, Univerza v Ljubljani

Toma Strle
Center za kognitivno znanost, Pedagoška 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 2024

Informacijska družba
ISSN 2630-371X

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2024

Leto 2024 je hkrati udarno in tradicionalno. Že sedaj, še bolj pa v prihodnosti bosta računalništvo, informatika (RI) in umetna inteligenca (UI) igrali ključno vlogo pri oblikovanju napredne in trajnostne družbe. Smo na pragu nove dobe, v kateri generativna umetna inteligenca, kot je ChatGPT, in drugi inovativni pristopi utirajo pot k superinteligenci in singularnosti, ključnim elementom, ki bodo definirali razcvet človeške civilizacije. Naša konferenca je zato hkrati tradicionalna znanstvena, pa tudi povsem akademsko odprta za nove pogumne ideje, inkubator novih pogledov in idej.

Letošnja konferenca ne le da analizira področja RI, temveč prinaša tudi osrednje razprave o perečih temah današnjega časa – ohranjanje okolja, demografski izzivi, zdravstvo in preobrazba družbenih struktur. Razvoj UI ponuja rešitve za skoraj vse izzive, s katerimi se soočamo, kar poudarja pomen sodelovanja med strokovnjaki, raziskovalci in odločevalci, da bi skupaj oblikovali strategije za prihodnost. Zavedamo se, da živimo v času velikih sprememb, kjer je ključno, da s poglobljenim znanjem in inovativnimi pristopi oblikujemo informacijsko družbo, ki bo varna, vključujoča in trajnostna.

Letos smo ponosni, da smo v okviru multikonference združili dvanajst izjemnih konferenc, ki odražajo širino in globino informacijskih ved: CHATMED v zdravstvu, Demografske in družinske analize, Digitalna preobrazba zdravstvene nege, Digitalna vključenost v informacijski družbi – DIGIN 2024, Kognitivna znanost, Konferenca o zdravi dolgoživosti, Legende računalništva in informatike, Mednarodna konferenca o prenosu tehnologij, Miti in resnice o varovanju okolja, Odkrivanje znanja in podatkovna skladišča – SIKDD 2024, Slovenska konferenca o umetni inteligenci, Vzgoja in izobraževanje v RI.

Poleg referatov bodo razprave na okroglih mizah in delavnicah omogočile poglobljeno izmenjavo mnenj, ki bo oblikovala prihodnjo informacijsko družbo. “Legende računalništva in informatike” predstavljajo slovenski “Hall of Fame” za odlične posameznike s tega področja, razširjeni referati, objavljeni v reviji *Informatica* z 48-letno tradicijo odličnosti, in sodelovanje s številnimi akademskimi institucijami in združenji, kot so ACM Slovenija, SLAIS in Inženirska akademija Slovenije, bodo še naprej spodbujali razvoj informacijske družbe. Skupaj bomo gradili temelje za prihodnost, ki bo oblikovana s tehnologijami, osredotočena na človeka in njegove potrebe.

S podelitvijo nagrad, še posebej z nagrado Michie-Turing, se avtonomna RI stroka vsakoletno opredeli do najbolj izstopajočih dosežkov. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel prof. dr. Borut Žalik. Priznanje za dosežek leta pripada prof. dr. Sašu Džeroskemu za izjemne raziskovalne dosežke. »Informacijsko limono« za najmanj primerno informacijsko tematiko je prejela nabava in razdeljevanjem osebnih računalnikov ministrstva, »informacijsko jagodo« kot najboljšo potezo pa so sprejeli organizatorji tekmovanja ACM Slovenija. Čestitke nagrajencem!

Naša vizija je jasna: prepoznati, izkoristiti in oblikovati priložnosti, ki jih prinaša digitalna preobrazba, ter ustvariti informacijsko družbo, ki bo koristila vsem njenim članom. Vsem sodelujočim se zahvaljujemo za njihov prispevek k tej viziji in se veselimo prihodnjih dosežkov, ki jih bo oblikovala ta konferenca.

Mojca Ciglarič, predsednica programskega odbora

Matjaž Gams, predsednik organizacijskega odbora

PREFACE TO THE MULTICONFERENCE INFORMATION SOCIETY 2024

The year 2024 is both ground-breaking and traditional. Now, and even more so in the future, computer science, informatics (CS/I), and artificial intelligence (AI) will play a crucial role in shaping an advanced and sustainable society. We are on the brink of a new era where generative artificial intelligence, such as ChatGPT, and other innovative approaches are paving the way for superintelligence and singularity—key elements that will define the flourishing of human civilization. Our conference is therefore both a traditional scientific gathering and an academically open incubator for bold new ideas and perspectives.

This year's conference analyzes key CS/I areas and brings forward central discussions on pressing contemporary issues—environmental preservation, demographic challenges, healthcare, and the transformation of social structures. AI development offers solutions to nearly all challenges we face, emphasizing the importance of collaboration between experts, researchers, and policymakers to shape future strategies collectively. We recognize that we live in times of significant change, where it is crucial to build an information society that is safe, inclusive, and sustainable, through deep knowledge and innovative approaches.

This year, we are proud to have brought together twelve exceptional conferences within the multiconference framework, reflecting the breadth and depth of information sciences:

- CHATMED in Healthcare
- Demographic and Family Analyses
- Digital Transformation of Healthcare Nursing
- Digital Inclusion in the Information Society – DIGIN 2024
- Cognitive Science
- Conference on Healthy Longevity
- Legends of Computer Science and Informatics
- International Conference on Technology Transfer
- Myths and Facts on Environmental Protection
- Data Mining and Data Warehouses – SIKDD 2024
- Slovenian Conference on Artificial Intelligence
- Education and Training in CS/IS.

In addition to papers, roundtable discussions and workshops will facilitate in-depth exchanges that will help shape the future information society. The “Legends of Computer Science and Informatics” represents Slovenia’s “Hall of Fame” for outstanding individuals in this field. At the same time, extended papers published in the Informatica journal, with over 48 years of excellence, and collaboration with numerous academic institutions and associations, such as ACM Slovenia, SLAIS, and the Slovenian Academy of Engineering, will continue to foster the development of the information society. Together, we will build the foundation for a future shaped by technology, yet focused on human needs.

The autonomous CS/IS community annually recognizes the most outstanding achievements through the awards ceremony. The Michie-Turing Award for an exceptional lifetime contribution to the development and promotion of the information society was awarded to Prof. Dr. Borut Žalik. The Achievement of the Year Award goes to Prof. Dr. Sašo Džeroski. The "Information Lemon" for the least appropriate information topic was given to the ministry's procurement and distribution of personal computers. At the same time, the "Information Strawberry" for the best initiative was awarded to the organizers of the ACM Slovenia competition. Congratulations to all the award winners!

Our vision is clear: to recognize, seize, and shape the opportunities brought by digital transformation and create an information society that benefits all its members. We thank all participants for their contributions and look forward to this conference's future achievements.

Mojca Ciglarič, Chair of the Program Committee

Matjaž Gams, Chair of the Organizing Committee

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 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2024

Zvezek B

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume B

Kognitivna znanost
Cognitive Science

Uredniki / Editors

Anka Slana Ozimič, Borut Trpin, Toma Strle, Olga Markič

http://is.ijs.si

10. oktober 2024 / 10 October 2024
Ljubljana, Slovenia

PREDGOVOR

Dobrodošli na konferenci Kognitivna znanost. Na letošnji konferenci bodo avtorice in avtorji raziskovali mnoge plati človeške kognicije in predstavili tako svoje empirične ugotovitve kot tudi teoretska raziskovanja. Poleg prispevkov s področja teme letošnje konference, "Preseganje vrzeli v raziskovanju in razumevanju uma", bomo potovali skozi različna področja kognitivne znanosti - od psihologije in nevroznanosti do filozofije in umetne inteligence, ter ob tem spoznavali raznolike tematike vključujoč uporabo VR tehnologij pri raziskovanju kognicije, uporabo inteligentnih tehnologij pomoči pri demenci, uporabo dregljajev za spreminjanje vedenja, vpliv umetne inteligence na demokracijo, izzive vzorčenja izkustva in manj poznane vidike doživljanja kot so ozadnja občutja.

Konferenca se bo zaključila z okroglo mizo, na kateri bomo razmišljali o izzivih, ki jih prinašajo vrzeli v raziskovanju in razumevanju v kognitivni znanosti: med drugim o združevanju prvo- in tretje-osebnih pristopov k raziskovanju človeškega uma, o povezovanju različnih nivojev opazovanja (na primer mikro-nivoja nevrološke aktivnosti z makro-nivojem vedenja in družbenih sistemov) ter o izzivih povezovanja različnih disciplinarnih pristopov.

Upamo, da bo letošnja konferenca predstavljala prostor radovednega povezovanja in izmenjave kreativnih idej. Skupaj bomo premagovali disciplinarne in metodološke ovire, združili mlade in izkušene znanstvenike ter znanstvenice, ki si delijo strast do raziskovanja skrivnosti kognicije.

Dobrodošli!

Anka Slana Ozimič
Borut Trpin
Toma Strle
Olga Markič

FOREWORD

Welcome to the Cognitive Science Conference. At this year's conference, authors will explore the many facets of human cognition and present both their empirical findings and theoretical research. In addition to contributions on the topic of this year's conference, Bridging the Gaps in Research and Understanding the Mind, we will explore a diverse range of fields of cognitive science – from psychology and neuroscience to philosophy and artificial intelligence – while also learning about various topics, including the use of VR technologies in research, the use of intelligent assistive technologies for dementia, the use of nudges to change behavior, the impact of artificial intelligence on democracy, the challenges of sampling experience, and explore the less known aspects of experience, such as background feelings.

The conference will conclude with a roundtable discussion, where we will reflect on the challenges posed by gaps in research and understanding of mind in cognitive science: among others, we will think about the integration of first- and third-person approaches to studying the human mind, the relation and possible links between different levels of observation (for example, the micro-level of neurological activity and the macro-level of behavior and social systems), and the challenges of connecting different disciplinary approaches.

We hope that this year's conference will be a space for networking and sharing insightful ideas. Together we will overcome disciplinary and methodological barriers, bringing together young and experienced scientists who share a passion for exploring the mysteries of cognition.

Welcome!

Anka Slana Ozimič
Borut Trpin
Toma Strle
Olga Markič

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Anka Slana Ozimič

Borut Trpin

Toma Strle

Olga Markič

The Ontological Shock: What Psychedelics can Teach us about the Nature of the Mind

Maruša Sirk
Centre for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
marusasirk@gmail.com

ABSTRACT

The following article provides a brief overview of the significance and potential challenges of profound psychedelic experiences that prompt individuals to question the nature of reality, often referred to as "ontological shocks." These experiences are believed to induce changes in the structure of consciousness, leading to subsequent shifts in worldviews, behaviors, relationships, and mental health. While they can result in long-lasting positive changes, they are not always pleasant. Due to the complex alterations these experiences produce, they offer a unique opportunity to explore gaps in our understanding of the human mind and the nature of the reality it perceives, enacts, or constructs. The article aims to raise awareness of these issues by shedding light on various aspects of the discourse surrounding this topic.

KEYWORDS

psychedelics, ontological shock, mind, consciousness

1 INTRODUCTION

Psychedelics are psychoactive substances that can lead to altered states of consciousness, experienced as a change in perception and cognitive processes. Classic psychedelics, such as psilocybin, mescaline, LSD and DMT, primarily act through the stimulation of the serotonin 5HT-2A receptor. Due to somewhat similar psychological effects, substances such as MDMA and ketamine are also sometimes considered as psychedelics even though they target different neurological structures [1].

In recent years, there has been a growing interest in research on the potential use of psychedelic substances for mental health treatment. There are many studies that support this claim [e. g. 2, 3, 4], but there is also evidence that psychedelics can lead to longer lasting adverse effects [e. g. 1, 5].

Some of the challenges that may emerge after a psychedelic experience stem from profound shifts in one's worldview [6], metaphysical beliefs [7], and an overall ontological shock [8], in which the individual begins to re-evaluate the nature of their reality. Commonly, people also experience a shift in their spiritual orientation [9], due to experiences that have been labeled as “spiritual emergencies” [10]. These include transpersonal experiences, out-of-body experiences, hallucinations of religious nature etc. [10].

In this paper, we will tackle the problem of the “ontological shock” that can arise due to psychedelic experiences, how individuals cope with them and what implications they have on our understanding of the mind.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.1

2 ONTOLOGICAL SHOCK FOLLOWING THE USE OF PSYCHEDELICS

As interest in researching psychedelics for their potential therapeutic effects increases, there is a growing need to understand the mechanisms that enable these changes to occur. Changes in metaphysical beliefs are thought to be one of the driving mechanisms of change that enable the transformational process to occur [8]. However, changes in metaphysical beliefs don’t come easily, as they normally induce the so-called ontological shock about the nature and reality of existence [8]. This means that people start to question the nature of (their) reality and subsequently come to adopt an altered belief system, commonly constituting beliefs such as animism, life after death, the existence of alternative realities etc. [7].

On one hand, psychedelic experiences are often reported to be among the most meaningful and significant experiences, leading to positive long-term changes [11]. On the other hand, many individuals report prolonged difficulties after a profound psychedelic experience, struggling with ontological challenges as they question their own reality and existence [8]. This presents ethical challenges in both formal and informal practices for integrating psychedelic experiences, while also raising broader questions about the nature of reality itself.

2.1 Coping with the ontological shock

Psychedelic experiences that possess mystical qualities—characterized by feelings of ineffability, significance, and the perceived “trueness” of the experience—are more likely to result in an ontological shock and lead to a transformed belief system [12]. These experiences can offer profound insights into the "oneness" of reality and foster a sense of "ontological comfort," bringing a greater sense of purpose and meaning to life. However, they can also present challenges, as individuals may struggle to integrate these insights into their everyday lives. This raises the question of how best to support people in making sense of these new ontological truth claims [12].

Challenging psychedelic experiences can lead to various ontological difficulties, such as questioning one's identity,

grieving the loss of a past self, experiencing persistent encounters with (spiritual) entities, psychotic episodes, spiritual grandiosity, feelings of meaninglessness and emptiness, isolation and despair [8], derealization, delusional beliefs, and experiences of possession [9]. These prolonged challenges can impact individuals in different ways, resulting in difficulties with everyday interactions, attentional problems, disruptions at work [8], challenges in managing emotions, and perceptual difficulties [9].

Research on the adverse effects of psychedelics [e.g. 13, 14, 15, 16] and the challenges that arise following the so-called ontological shocks these experiences can induce [e.g. 1, 8, 9] highlights the importance of developing therapeutic practices and guidance rooted in empathic resonance and the concept of psychedelic apprenticeship [6]. While there are many existing therapeutic and shamanic frameworks aimed at making sense of psychedelic experiences [6], it is crucial to recognize that these experiences often lead to heightened suggestibility [17], which must be considered when helping individuals navigate their shifting worldviews. People have reported various strategies that helped them cope with extended difficulties, such as meditation, embodied contemplation, self-education through reading and journaling, and physical exercise. In terms of support from others, individuals expressed a need to feel seen, heard, understood, believed, and to have their experiences accepted and validated [1, 8]. This suggests a responsibility for practitioners working with and guiding people through psychedelic experiences to help them find meaning in their experiences without imposing any specific ontological truth claims [12].

Extensive research on challenging psychedelic experiences, including work by Stanislav Grof's clinical team, has identified common existential challenges that individuals often face following these experiences, primarily centered around fears of dying, going insane, or losing control [18]. According to Grof's team, psychedelic experiences can activate deep existential concerns that only subside once the individual has successfully processed them. It is suggested that people may encounter a sense of the groundlessness of being [18], which can be understood as an underlying "principle" of cognition. This brings us to the next point of this paper: the implications these experiences might have for the scientific understanding of the mind.

3 POSSIBLE IMPLICATIONS ON UNDERSTANDING THE MIND

3.1 Understanding psychedelic experiences

The underlying mechanisms that facilitate changes after the use of psychedelics have yet to be fully understood. However, several hypotheses have been proposed to explain how these cognitive shifts might occur. One prominent hypothesis is based on the framework of predictive processing [19]. This framework conceptualizes the mind as a "prediction machine" that continuously balances information by integrating prior knowledge about the world with incoming sensory data from the environment [19].

Within this framework, the "Relaxed Beliefs Under Psychedelics" (REBUS) model suggests that psychedelics increase the brain's level of entropy—or uncertainty—while reducing reliance on prior beliefs, thereby allowing more room for new sensory information [20]. This process can lead to the dissolution of previously rigid mental models and established worldviews. The resulting experience of an ontological shock may represent a direct encounter with the fundamental uncertainty of reality [8]. This concept, referred to as "groundlessness," attempts to explain how individuals continuously construct a world of meaning that is inherently without a fixed foundation and perpetually in flux [21].

It is important to note that psychedelics can also reinforce existing beliefs, potentially enhancing established worldviews, mental models, and expectations [22]. To address this complexity, the "Altered Beliefs Under Psychedelics" (ALBUS) model proposes that the effects of psychedelics on prior beliefs depend on factors such as the dose consumed and the individual's pre-existing state of mind [22]. This model aims to explain how psychedelics can both diminish and strengthen prior beliefs, bridging gaps between different proposed mechanisms of how psychedelics influence reality and well-being [22].

Additionally, other models offer explanations for the mechanisms of action of psychedelics. The "Cortical-Subcortical Communication Theory" (CSCT) suggests that psychedelics reduce thalamo-cortical filtering of internal and external stimuli, allowing new, unfiltered sensory information to emerge [23]. The "Cortical-Claustrum Communication" (CCC) model posits that psychedelics decouple cortical areas from the claustrum, leading to reduced cognitive control [24]. Furthermore, psychedelics are thought to open a critical period for social learning, potentially fostering new social behaviors and reducing tendencies toward isolation [25].

These various theories attempt to explain how psychedelics can facilitate the creation of new cognitive models of the world and reality. However, the field of psychedelic research continues to grapple with understanding the precise mechanisms of these substances, with ongoing testing of competing theories and hypotheses. For a comprehensive evaluation of these theories, see [26].

There is also an ongoing debate about the role of subjective experiences in the transformative effects of psychedelics. Some researchers take on a reductionist approach, focusing solely on the brain mechanisms involved [27, 28]. This is problematic, as it opens the question of how to understand the profound ontological shocks and the integration of the psychedelic experience in everyday lives of individuals. It is also problematic, as evidence suggests that psychedelic experiences with rich subjective effects, such as mystical-type experiences, can lead to the most significant transformations [29, 30]. If transformative effects were purely mechanistic, without considering subjective experiences, it becomes challenging to explain the struggles and positive changes individuals report in their daily lives after using psychedelics.

Subjective experiences cannot be easily dismissed [31, 32, 33, 34], and they are crucial for understanding how people's ontological reality gets altered. Investigating these subjective aspects could help address some of the unresolved questions about the mind. Studying the invariants and stable states of the "changing mind" following psychedelic use may bring us closer to unlocking the nature of the mind. This research could have implications for not only understanding and treating mental health issues but also for exploring concepts like consciousness [22, 36, 37] and the self [22, 33, 35, 38]. Additionally, it could

impact the reductionist debate [32], consciousness theories, and discussions about the "easy" and "hard" problems of consciousness [22, 36, 37].

3.2 Possible contribution to understanding the mind

In previous sections, we provided a short overview of what psychedelics are and sketched some possible implications they can have in the everyday lives of people, as well as our broader impact they may have on our understanding of the mind. In the last section we want to finish with diving a bit deeper into some possible implications that ontological shocks can have on understanding the mind.

Let us stop for a moment on how we understand the concept of the “mind”. This is an important question as the theories of psychedelic mechanisms all have their own postulates, the prevalent implicit view being that the mind is a product of neuronal activity (which applies for previously presented theories – the REBUS [20], CSCT [23] and CCT model [24]). This is a reductionist view of the mind that equals the mind with the brain [39]. Another possible view is that the mind is an information-processing system that manipulates and transforms information, which is a computational view [40]. In the previous years, another understanding on the mind has slowly been evolving in cognitive science – that the mind is embodied, embedded, extended and enacted, which we call the 4E cognition. This view understands the mind as a complex interplay between the brain, body and the surrounding environment [41].

The challenge of understanding the mind mirrors the debate in psychedelic research about the significance of subjective experiences. The core issue is whether the relevance of a psychedelic experience depends solely on inducing specific brain states or requires a deeper subjective experience to impact a person's everyday life. Evidence increasingly supports the idea that both "set" (the interplay of personality, preparation, expectation, and intention) and "setting" (the physical, social, and cultural environment) play crucial roles during a psychedelic experience [42]. This observation could indicate the relevance of the 4E cognition framework, which views the mind as a dynamic interplay between brain, body, and environment. Moreover, the 4E cognition theory might explain why set and setting are important, and why some psychedelic states and doses lead to profound changes while others do not. By exploring this intricate interplay, the 4E framework may shed light on why certain individuals experience ontological shocks under specific conditions. If, however, these experiences are inexplicable through existing frameworks, they could highlight gaps in our current understanding of the mind and reveal how alternative states of consciousness can disrupt the established interaction between mind, body, and environment.

Psychedelic experiences are often described as "altered states of consciousness," suggesting that by examining what changes during these experiences, we can gain insights into what constitutes the "normal," "usual," or "everyday" state of consciousness. The concept of ontological shock, which we have frequently referenced, highlights a paradox within this framework. If a person's everyday consciousness is altered during a psychedelic experience, and they subsequently notice changes in their subjective experience in their daily life, does this mean that the new state is an unusual or extended form of consciousness? In other words, does this imply that the individual is now living in a perpetually altered state of consciousness?

There is a prevalent view that the subjective experiences induced by psychedelics reveal aspects of the mind that need to be integrated into everyday life [43]. This perspective suggests that psychedelics should be considered mind-revealing rather than merely mind-altering substances [44]. This leads us back to fundamental questions about the nature of consciousness itself. Is consciousness merely a byproduct of neuronal activity, something external waiting to be experienced, is it embodied, enacted, or something else entirely? What we do know at this point is that psychedelics can induce alterations in our consciousness, affecting our awareness of ourselves and the world around us.

While it may be ambitious to claim that psychedelic experiences will fully bridge the epistemic gap between first-person experiences and their third-person correlates, or help us understand the nature of consciousness itself – the problem we commonly refer to as the hard problem of consciousness [45, 36]–, they can still provide valuable insights into both these issues [35, 36]. The most profound psychedelic experiences, which often lead to significant changes in consciousness, self-perception, and belief systems, may offer particularly important insights.

4 CONCLUSION

The aim of this article was to present the concept of ontological shock following the use of psychedelics and possible implications on the scientific understanding of the mind. We provided an overview of what is already known about this topic, to point out where we should be cautious and what is still unknown or vaguely known, as well as to illustrate how diving deeper into this topic could help us scientifically advance our current understanding of the mind.

It is important to conclude this paper with a call for caution. As we tried to point out, psychedelic experiences and its subsequent changes in everyday experience can inform us about the nature of our mind and help us gain broader understanding about topics related to consciousness, self, mental health etc. But the experiences that could most inform us about these topics and can lead to most profound long-term changes, have its challenges and downsides, which should not be disregarded. That is why the integration process, as well as the importance of set and setting, should always be considered when dealing with these substances. But before we have a consensus on what the mind is and how it constructs our reality, a lot of damage can be done, especially if we want to use psychedelic substances to help people get through their mental health problems, as is the case in psychedelic research in the past years.

REFERENCES

[1] Oliver C. Robinson, Jules Evans, David Luke, Rosalind McAlpine, Aneta Sahely, Amy Fisher, Stian Sundeman, Eirini Ketzitzidou Argyri, Ashleigh Murphy-Beiner, Katrina Michelle and Ed Prideaux, 2024. Coming back together: a qualitative survey study of coping and support strategies used by people to cope with extended difficulties after the use of psychedelic

drugs. *Frontiers in Psychology*, 15, e1369715. DOI: https://doi.org/10.3389/fpsyg.2024.1369715

[2] David Nutt, David Erritzoe and Robin Carhart-Harris, 2020. Psychedelic psychiatry's brave new world. *Cell*, 181, 1, 24-28. DOI: https://doi.org/10.1016/j.cell.2020.03.020

[3] Roland R. Griffiths, Matthew W. Johnson, Michael A. Carducci, Annie Umbricht, William A. Rishards, Brian D. Richards, Mary P. Cosimano and Margaret A. Klinedinst, 2016. Psilocybin produces substantial and sustained decreases in depression and anxiety in patients with life-threatening cancer: A randomized double-blind trial. *Journal of psychopharmacology*, 30, 12, 1181-1197. DOI: https://doi.org/10.1177/0269881116675513

[4] Paweł Orłowski, Anastasia Ruban, Jan Szczypiński, Justyna Hobot, Maksymilian Bielecki and Michał Bola, 2022. Naturalistic use of psychedelics is related to emotional reactivity and self-consciousness: The mediating role of ego-dissolution and mystical experiences. Journal of *Psychopharmacology*, 36, 8, 905-1004. DOI: https://doi.org/10.1177/02698811221089034

[5] Michiel van Elk and Eiko I. Fried, 2023. History repeating: A roadmap to address common problems in psychedelic science. *Therapeutic Advances in Psychopharmacology*, 13. DOI: https://doi.org/10.1177/20451253231198466

[6] Chrostopher Timmermann, Rosalind Watts and David Dupuis, 2022. Towards psychedelic apprenticeship: Developing a gentle touch for the mediation and validation of psychedelic-induced insights and revelations. *Transcultural psychiatry*, 59, 5, 691-704. DOI: https://doi.org/10.1177/13634615221082796

[7] Sandeep M. Nayak, Manvir Singh, David B. Yaden, D. B. and Roland R. Griffiths, 2022. Belief changes associated with psychedelic use. *Journal of Psychopharmacology*, 37, 1, 80-92. DOI: https://doi.org/10.1177/02698811221131989

[8] Eirini K. Argyri, Jules Evans, David Luke, Pascal Michael, Katrina Michelle, Cyrus Rohani-Shukla, Shayam Suseelan, Ed Prideaux, Rosalind McAlpine, Ashleigh Murphy-Beiner and Oliver Robinson, 2024. *Navigating Groundlessness: An interview study on dealing with ontological shock and existential distress following psychedelic experiences*. Available at SSRN: https://ssrn.com/abstract=4817368 or http://dx.doi.org/10.2139/ssrn.4817368

[9] Jules Evans, Oliver C. Robinson, Eirini Ketzitzidou Argyri, Shayam Suseelan, Ashleigh Murphy-Beiner, Rosalind McAlpine, David Luke, Katrina Michelle and Ed Prideaux, 2023. Extended difficulties following the use of psychedelic drugs: A mixed methods study. *PLoS ONE*, 18, 10, e0293349. DOI: https://doi.org/10.1371/ journal.pone.0293349

[10] Christina Grof and Stanislav Grof, 2017. Spiritual emergency: the understanding and treatment of transpersonal crises. *International Journal of Transpersonal Studies*, 36, 30–43. DOI: https://doi.org/10.24972/ijts.2017.36.2.30

[11] Frederick S. Barrett and Roland R. Griffiths, 2018. Classic Hallucinogens and Mystical Experiences: Phenomenology and Neural Correlates. *Current Topics in Behavioral Neurosciences*, 36, 393-430. DOI: https://doi.org/10.1007/7854_2017_474

[12] Joost J. Breeksema and Michiel van Elk, 2021. Working with weirdness: a response to 'moving past mysticism in psychedelic science. *ACS Pharmacology & Translational Science*, 4, 4, 1471–1474, DOI: https://doi.org/10.1021/acsptsci.1c00149

[13] Joost J. Breeksema, Bouwe W. Kuin, Jeanine Kamphuis, Wim van den Brink, Eric Vermetten, Robert A. Schoevers, 2022. Adverse events in clinical treatments with serotonergic psychedelics and MDMA: A mixed-methods systematic review. *Journal of Psychopharmacology*, 36, 10, 1100-1117. DOI: https://doi.org/10.1177/02698811221116926

[14] Anne K. Schla, Jacob Aday, Iram Salam, Jo C. Neill, David J. Nutt, 2022. Adverse effects of psychedelics: From anecdotes and misinformation to systematic science. *Journal of Psychopharmacology*, 36, 3, 258-272. DOI: https://doi.org/10.1177/02698811211069100

[15] Rebeca Bremler, Nancy Katati, Parvinder Shergill, David Erritzoe and Robin L. Carhart-Harris, 2023. Case analysis of long-term negative psychological responses to psychedelics. *Scientific Reports*, 13, e15998. DOI: https://doi.org/10.1038/s41598-023-41145-x

[16] Daniel Meling, Rebecca Ehrenkranz, Sandeep M. Nayak, Helena D. Aicher, Xaver Funk, Michiel van Elk, Marianna Graziosi, Prisca R. Bauer, Milan Scheidegger and David B. Yaden, 2024. Mind the Psychedelic Hype: Characterizing the Risks and Benefits of Psychedelics for Depression. *Psychoactives*, 3, 215-234. https://doi.org/10.3390/psychoactives3020014

[17] David Dupuis, 2021. Psychedelics as Tools for Belief Transmission. Set, Setting, Suggestibility, and Persuasion in the Ritual Use of Hallucinogens. *Fronters in Psychology*, 12, e730031. DOI: https://doi.org/10.3389/fpsyg.2021.730031

[18] Andrew Carnahan, 2023. An existential approach to integrating challenging psychedelic experiences. *Existential Analysis: Journal of the Society for Existential Analysis*, 34, 1, 89-102.

[19] Andy Clark, 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 3, 181–204. DOI https://doi.org/10.1017/S0140525X12000477

[20] Robin L. Carhart-Harris and Karl J. Friston, 2019. REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics. *Pharmacological reviews*, 71, 3, 316–344. DOI: https://doi.org/10.1124/pr.118.017160

[21] Daniel Meling, 2021. Knowing Groundlessness: An Enactive Approach to a Shift From Cognition to Non-Dual Awareness. *Frontiers in Psychology*, 12, e697821. DOI: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.697821

[22] Adam Safron, Adam Safron, Arthur Juliani, Nicco Reggente, Victoria Klimaj and Matthew Johnson, 2020. On The Varieties of Conscious Experiences: Altered Beliefs Under Psychedelics (ALBUS) *PsyArXiv*. DOI: https://doi.org/10.31234/osf.io/zqh4b

[23] Franz X. Vollenweider and Katrin H. Preller, 2020. Psychedelic drugs: neurobiology and potential for treatment of psychiatric disorders. *Nature Reviews Neuroscience*, 21, 611–624. DOI: https://doi.org/10.1038/s41583-020-0367-2

[24] Frederick S. Barrett, Samuel R. Krimmel, Roland R. Griffiths, David A. Seminowicz and Brian N. Mathur, 2020. Psilocybin acutely alters the functional connectivity of the claustrum with brain networks that support perception, memory, and attention. *NeuroImage*, 218, e116980. DOI: https://doi.org/10.1016/j.neuroimage.2020.116980

[25] Romain Nardou, Eastman M. Lewis, Rebecca Rothhaas, Ran Xu, Aimei Yang, Edward Boyden and Gül Dölen, 2019. Oxytocin-dependent reopening of a social reward learning critical period with MDMA. *Nature*, 569, 116-120. DOI: https://doi.org/10.1038/s41586-019-1075-9

[26] Michiel van Elk and David B. Yaden, 2022. Pharmacological, Neural, and Psychological Mechanisms underlying Psychedelics: A Critical Review. *Neuroscience & Biobehavioral Reviews*, 140, 6, e104793. DOI: https://doi.org/10.1016/j.neubiorev.2022.104793

[27] David E. Olson, 2021. The Subjective Effects of Psychedelics May Not Be Necessary for Their Enduring Therapeutic Effects. *ACS Pharmacology & Translational Science*, 4, 2, 563–567. DOI: https://doi.org/10.1021/acsptsci.0c00192

[28] Lindsay P. Cameron, Robert J. Tombari, Ju Lu, Alexander J. Pell, Zefan Q. Hurley, Yann Ehinger, Maxemiliano V. Vargas, Matthew N. McCarroll, Jack C. Taylor, Douglas Myers-Turnbull, Taohui Liu, Bianca Yaghoobi, Lauren J. Laskowski, Emilie I. Anderson, Guoliang Zhang, Jayashri Viswanathan, Brandon M. Brown, Michelle Tija, Lee E. Dunlap, Zachary T. Rabow, Oliver Fiehn, Heike Wuff, John D. McCorvy, Pamela J. Lein, David Kokel, Dorit Ron, Jamie Peters, Yi Zuo and David E. Olson, 2020. A non-hallucinogenic psychedelic analogue with therapeutic potential. *Nature*, 589, 474-479. DOI: https://doi.org/10.1038/s41586-020-3008-z

[29] Kwonmok Ko, Gemma Knight, James J. Rucker and Anthony J. Cleare, 2022. Psychedelics, Mystical Experience, and Therapeutic Efficacy: A Systematic Review. *Frontiers in Psychiatry*, 13, e917199. DOIi: https://doi.org10.3389/fpsyt.2022.917199

[30] Mazen A. Atiq. Matthew R. Baker, Jennifer L. Vande Voort, Maxemiliano V. Vargas and Doo-Sup Choi, 2024. Disentangling the acute subjective effects of classic psychedelics from their enduring therapeutic properties. *Psychopharmacology*. DOI: https://doi.org/10.1007/s00213-024-06599-5

[31] David B. Yaden and Roland R. Griffiths, 2020. The Subjective Effects of Psychedelics Are Necessary for Their Enduring Therapeutic Effects. *ACS pharmacology & translational science*, 4, 2, 568-572. DOI: https://doi.org/10.1021/acsptsci.0c00194

[32] Gerhard Gründer, Manuela Brand, Lea J. Mertens, Henrik Jungaberle, Laura Kärtner, Dennis J. Scharf, Moritz Spangemacher and Max Wolff, 2024. Treatment with psychedelics is psychotherapy: beyond reductionism. *Lancet Psychiatry*, 11, 3, 231-236. DOI: https://doi.org/10.1016/S2215-0366(23)00363-2

[33] Riccardo Miceli McMillan and Anthony Vincent Fernandez, 2023. Understanding subjective experience in psychedelic-assisted psychotherapy: The need for phenomenology. *Australian & New Zealand Journal of Psychiatry*, 57, 6, 783-788. DOI: https://doi.org/10.1177/00048674221139962

[34] Tomislav Majić, Timo T. Schmidt and Jürgen Gallinat, 2015. Peak experiences and the afterglow phenomenon: When and how do therapeutic effects of hallucinogens depend on psychedelic experiences? *Journal of Psychopharmacology*, 29, 3, 241-253. DOI: https://doi.org/10.1177/0269881114568040

[35] Christopher Timmermann, Prisca R. Bauer, Olivia Gosseries, Audrey Vanhaudenhuyse, Franz Vollenweider, Steven Laureys, Tania Singer; Mind and Life Europe (MLE) ENCECON Research Group; Elena Antonova, Antoine Lutz, 2022. A neurophenomenological approach to non-ordinary states of consciousness: hypnosis, meditation, and psychedelics. *Trends in Cognitive Sciences*, 27, 2, 139-159. DOI: https://doi.org/10.1016/j.tics.2022.11.006

[36] David B. Yaden, Matthew W. Johnson, Roland R. Griffiths, Manoj K. Doss, Albert Garcia-Romeu, Sandeep Nayak, Natalie Gukasyan, Brian N. Mathur and Frederick S. Barrett, 2021. Psychedelics and Consciousness:

Distinctions, Demarcations, and Opportunities. *International Journal of Neuropsychopharmacology*, 24, 8, 615–623. DOI: https://doi.org/10.1093/ijnp/pyab026

[37] Sidath Rankaduwa and Adrian M. Owen, 2023. Psychedelics, entropic brain theory, and the taxonomy of conscious states: a summary of debates and perspectives. *Neuroscience of Consciousness*, 2023, 1, niad001. DOI: https://doi.org/10.1093/nc/niad001

[38] Benjamin Hearn, 2021. Psychedelics, Mystical Experiences, and Meaning Making: A Renegotiation Process With the Challenges of Existence. *Journal of Humanistic Counseling*, 60, 180-196. DOI: https://doi.org/10.1002/johc.12164

[39] Francis Crick, 1994. *The Astonishing Hypothesis*. Scribners, New York.

[40] Patricia Smith Churchland and Terrence J. Sejnowski, 1992. *The Computational Brain*. MIT Press, Cambridge, MA.

[41] Michael L. Anderson, 2003. Embodied Cognition: A field guide. *Artificial Intelligence*, 149, 91–130. DOI: https://doi.org/10.1016/S0004-3702(03)00054-7

[42] Ido Hartogsohn, 2017. Constructing drug effects: A history of set and setting. *Drug Science, Policy and Law*, 3. DOI: https://doi.org/10.1177/2050324516683325

[43] Collin M. Reiff, Elon E. Richman, Charles B. Nemeroff, Linda L. Carpenter, Alik S. Widge, Carolyn I. Rodriguez, Ned H. Kalin, William M. McDonald and and the Work Group on Biomarkers and Novel Treatments, a Division of the American Psychiatric Association Council of Research, 2020. Psychedelics and Psychedelic-Assisted Psychotherapy. *American Journal of Psychiatry*, 177, 5, 365-468. DOI: https://doi.org/10.1176/appi.ajp.2019.19010035

[44] Aidan Lyon, 2023. *Psychedelic Experience: Revealing the Mind*. Oxford University Press.

[45] David J. Chalmers, 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2, 3, 200-219.

Innovative Supporting Approaches: Integrating Bibliotherapy, Psychodrama and AI as a Therapeutic Conversational Tool

Veronika Mattová†
Department of Mongolian,
Korean and Vietnamese Studies
Faculty of Arts
Masaryk University
Czech Republic
veronika.mattova@mail.muni.cz

Courtney Lazore
Independent Researcher
National Coalition of
Independent Scholars
United States
courtneylazore@ncis.org

Abstract

How can individuals deal with personal trauma or internal struggles more effectively? This is the main question of every existential crisis, closely linked with humankind's survival strategy. Finding new, innovative ways for practitioners to leverage therapeutic techniques and modern *artificial intelligence (AI) technology* is crucial to providing precision mental health support to more individuals. While looking at possible approaches, it becomes more and more important to synthesize complex ways practitioners can provide multidimensional help. This paper investigates the possibility of a new holistic treatment that integrates *bibliotherapy's* storytelling, *Magic Shop* as a *psychodrama method*, and *AI conversation tools – chatbots* to ensure that individuals receive encompassing supportive therapy and feel less isolated. The holistic method is applied to *Korean pop* music as a case study, because K-pop content has experimented with these techniques and fandoms often have strong parasocial interactions. Combining these techniques creates a holistic, accessible, and personalized mental health care option that enhances the cognitive, emotional and practical well-being of individuals in need of support.

Keywords

Bibliography, storytelling, psychodrama, Magic Shop, AI, chatbots, K-pop

1 Introduction

The psychological aspect of overcoming personal issues and facing the harshness of reality is demanding for everyone. Internal cognitive mechanisms lead us to believe that individual struggles require individual approaches. This adaptation for hiding internal fights and presenting balance to the outside world is built under external pressure to achieve inner balance. However, emotional turmoil can be particularly challenging, which is why we need to explore a multimodal approach that could be a revolutionary step in interdisciplinary mental health support.

Bibliotherapy, psychodrama, and *AI* might initially seem unrelated, but there is a way to incorporate all of them within a new therapeutic framework.

2 Methodology

This paper uses a conceptual methodology to explore the integration of mental health techniques and *artificial intelligence* to theorize a holistic, accessible and personalized mental health care option. Based on the researchers' backgrounds and engagement with existing literature, *bibliotherapy, psychodrama* and *artificial intelligence* were chosen for further analysis. The researchers reviewed literature from databases including APA PsycInfo, Google Scholar, IEEEXplore, ProQuest, PubMed, Scopus, Web of Science and EBSCO. Only papers written in English and published after 2000 were considered, except for one foundational work (*Purpose and strategy behind the magic shop*). To build the theory, the researchers identified the prominent features of each topic (Table 1): *bibliotherapy's* storytelling, *psychodrama's* experiential model, and *AI's* real-time access, which will each be explored in subsequent sections. For an exploratory case study application, the researchers relied on new concepts (described in *6.1-6.3*) provided through the *Korean pop* music industry (concretely *group BTS*), initially produced as music-related products, but nowadays perceived through a therapeutic lens, with the primary aim of merging the separate therapies together, creating a more influential impact.

Approach	Method	Examples	Therapeutic Goal
Bibliotherapy	Individual therapy through storytelling	BTS: "The Most Beautiful Moment in Life" book series	Promote personal reflection and healing
Psychodrama	Magic Shop	BTS: "Magic Shop" song (metaphorical counselling setting)	Explore and resolve emotional conflicts
AI Chatbots	Parasocial interaction therapy	BTS: Weverse, Replica, Mydol	Provide emotional support and companionship
Integrated Approach	Psychodrama and digital interaction	BTS: Integration of storytelling, psychodrama, and digital interaction	Holistic support for the healing process

Table 1: Selection & Clarification for Chosen Approaches

†Alumna of Comenius University, Faculty of Mathematics, Physics and Informatics, MEi:CogSci | Middle European interdisciplinary master's programme in Cognitive Science.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.2

3 Bibliotherapy

Storytelling has long been a powerful means of exploring and understanding human emotions, and this concept underpins *bibliotherapy*, a therapeutic approach that uses literature to support mental health and personal growth. By engaging with carefully chosen texts, individuals can reflect on their own experiences, gain new insights, and find comfort and empathy through the narratives. *Bibliotherapy* leverages the emotional and cognitive impact of stories to address psychological challenges, offering a complementary and accessible method for fostering emotional healing and resilience [1].

Some research suggests that fictional narratives may be more effective for use in *bibliotherapy*, causing readers to empathize more with the characters and leading to better self-understanding and self-improvement [2].

Moreover, neuroscience reveals that storytelling influences brain chemistry, enhancing empathy and trust through the release of oxytocin and vasopressin [3]. This approach effectively enhances emotional experiences and trust, aiding in personal development and resilience [4].

4 Psychodrama and its Traits in Magic Shop

There are many ways to grasp the concept of *Magic Shop*. The first is understanding it as an in-depth training program with transformational practice, which will be discussed later in chapter *6.2 K-POP in Magic Shop*.

On the other hand, there is a possibility to apprehend it through the concept of *psychodrama*, which will be our primary interest, within this paper. *Magic Shop* is a practice used in *psychodrama* and group therapy in which participants create a space where they can "buy" something they already want, such as confidence or peace, from what they already have internally, such as anxiety or anger. In other words, this activity helps individuals explore their deepest passions, and the sacrifices needed to achieve them, while encouraging self-reflection and personal growth, with a possibility to solve conflicts in a safe, symbolic environment [5].

The term magic can be seen in society as something that carries supernatural power. *Magic Shop* is not an exception. Thus, it has a lot of forms and names: *Magic Shop* is our deep understanding of heart and brain in harmony. We often compare ourselves to others and hope we could be better than we are. Specific traits like social ability, patience and cleverness are examples that require time and practice. But it seems that *Magic Shop* is the key [6]. This method as a *psychodramatic strategy* can offer help to anybody through the use of fantasy [7].

In general, *psychodrama* is an experiential form of therapy, allowing those in treatment to explore issues through action methods (dramatic actions) [8]. This described approach, linked with *psychodrama*, was developed by Jacob Levy Moreno as a psychotherapeutic technique useful in working with patients during individual and group psychotherapy. This method offers significant changes through role-playing and dramatization, resulting in many benefits, such as insight, abreaction, acceptance of internal impulses, confrontation with the feelings of other people and training of alternative behaviors [9]. Whether we are talking about the first or second definition of *Magic Shop*, it can be a double-edged sword linked to maintaining one's own physiological and psychological well-being.

5 AI used in Therapeutic Spaces

AI chatbots are increasingly being used to enhance mental health care by offering real-time interactions that meet cognitive and emotional needs. These digital tools are part of a broader shift in communication, driven by the "computers-are-social-actors" paradigm, where AI significantly impacts how people engage with services [10].

In therapy, AI complements traditional methods by providing personalized interventions, making mental health support more accessible. Virtual environments facilitated by *AI* can create new opportunities for personal fulfilment and emotional connection [11].

Moreover, there is an overlap, because recent research by McAllister et al. explores the potential for *chatbots* to enhance *bibliotherapy* by supporting facilitators in mental health sessions. The study seeks to address gaps in existing literature by investigating how *chatbots* can be utilized to assist in the preparation and delivery of *bibliotherapy* [12].

In addition, in response to the high potential of technology, interviews with *bibliotherapy* facilitators have been conducted, followed by thematic analysis, to identify suitable tasks for the *chatbot*, aiding facilitators rather than directly evaluating the impact on participants of *bibliotherapy*.

This integration represents a major step towards a more comprehensive and accessible mental health framework.

6 K-POP as Multidimensional Tool

Although *Korean pop*, generally known as *K-pop*, seems to only overlap with music therapy, due to its main impact field, *K-pop* may also serve as a useful case application for *bibliotherapy, the Magic Shop* technique, and *AI* tools.

Some *K-pop* groups participate in transmedia story worlds and lyrical concepts that go far beyond "unrequited love" themes, touching the human psyche more deeply. Additionally, merging this with technology platforms like "Weverse" or "Bubble for JYPnation ", not to mention fabricated interactions through apps like "Replika", "Mydol", etc., fans have never felt closer and more bonded to their idols, who are revered with boundless support and understanding.

6.1 K-POP in Bibliotherapy

Many *K-pop* groups rely on some form of storytelling, but some take it a step further. In particular, *K-pop group BTS's* "BTS Universe" *(BU)*, functions as a multidimensional tool for emotional and psychological engagement. In *bibliotherapy, BTS's* books, *The Most Beautiful Moment in Life: The Notes*, provide a unique narrative that facilitates self-reflection and emotional exploration. These notes, embedded within the group's *Love Yourself* albums, as well as two books published by *BTS's* label, offer fans a form of therapeutic engagement by inviting them to interpret and relate to the fictional world, which can mirror personal experiences and foster emotional processing [13].

This can be particularly striking when we take into account individual emotional overload with a long-term inability to

restart one's own coping mechanism to underlie qualitative functioning in everyday life.

Preliminary research has suggested that the stories embedded in *K-pop* can have a positive impact on fans' ability to cope with challenges and heal. One survey found that *97%* of fans (*n=2342*) agreed that *BTS's* music and lyrics were effective in this way. Additionally, *84%* agreed the storylines in *BTS's* concepts were effective, and *75%* agreed the *BU* storyline was effective [14].

These numbers suggest valid proof of audience engagement techniques in fiction, described by Donald Maass as presenting novelty, challenge, or aesthetic appeal to readers, which leads to better identification with the story, while figuring out solutions for the main character´s actions, reflecting in an individual's healing scheme [15].

6.2 K-POP in Magic Shop

There is no doubt that science plays a crucial role in our understanding of human beings, yet some phenomena, such as compassion, altruism and empathy, remain enigmatic. These concepts form the foundation of James Doty's work, particularly in his book *Into the Magic Shop: A Neurosurgeon's True Story of the Life-Changing Magic of Mindfulness and Compassion*, which inspired not only this paper but also the lyrics of the *K-pop group BTS's* song "Magic Shop". The song, much like Doty's book, is perceived to have a healing effect on many individuals, resonating deeply within a therapeutic framework. Doty is not only a renowned author but also a researcher who founded the Center for Compassion and Altruism Research and Education (CCARE) at Stanford University, which supports his desire to analyze the interaction between the mind and body in relation to the concept of *Magic Shop* [16].

6.3 K-POP in AI Chatbots

In the *K-pop* world, *AI* enhances fan interactions by fostering *parasocial relationships (PSI)*, a term that goes back to its roots in 1956 when it was defined as the illusion of a face-to-face friendship between audience members, along with the main factor of a one-sided relationship [17] with *idols* (Korean singers within *K-pop* industry). *Chatbots* like "ChatGPT", "Replica" or "Mydol" offer more personalized conversations, deepening the emotional connection fans feel with their favorite stars [18]. This can be particularly comforting for those dealing with anxiety or low self-esteem.

AI-driven *PSI* offers new avenues for addressing anxiety, providing a controlled environment for emotional exploration. While there are risks of maladaptive obsessions, the positive impact on psychological well-being is significant when managed carefully [19]. As *AI* continues to evolve, its role in supporting mental health in niche areas like *K-pop* will likely grow, offering innovative solutions to common challenges [20].

7 Relevance for Cognitive Science & Cognitive Behavioral Therapy

Understanding how the human brain reacts to situations, making individuals feel distressed has been already covered. Now, it is more than important to look for strategies to fight against it, apart from medicaments. Recognizing an individual's need to support one's prosperity and healing cognitive processes, interdisciplinary cognitive science offers an opportunity to merge concepts together that might bring new ways to deepen our knowledge.

Combining different approaches, such as *AI chatbots* using *Cognitive Behavioral Therapy (CBT)* with minimal *bibliotherapy* interventions, is slowly but steadily indicating that the *chatbots* are more effective in reducing symptoms of depression and anxiety [21], which is the primary desired effect.

Moreover, *fanship*, an individual's bond to their idols, plays a crucial role in enhancing happiness, self-esteem, and social connectedness, which pushes forward the application of social identity theory in the realm of *K-pop* fans and expands the psychological understanding of fandom and its extended therapeutic possibilities [22].

8 Findings

Combining *bibliotherapy*, the *Magic Shop* psychotherapeutic method and *AI* tools can offer a comprehensive mental health treatment model. *Bibliotherapy* provides cognitive and emotional benefits through literature and storytelling, the *Magic Shop* method engages clients in creative and transformative experiences, and *AI* tools ensure continuous support and accessibility (as shown in Figure 1).

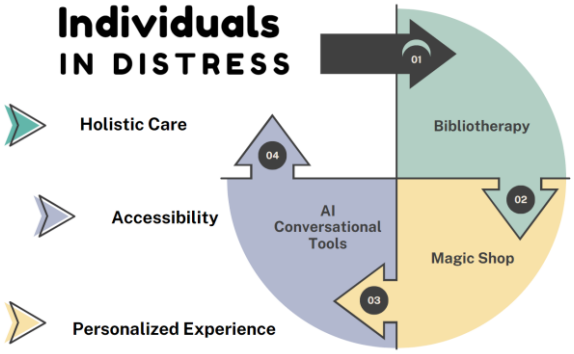


Figure 1: Infographic of Innovative Supporting Approaches

Note. Created with *Canva.com*

a. ***Holistic Care:*** *The union offers a multidisciplinary approach to mental health, addressing cognitive, emotional, and practical needs. By considering the whole individual, this approach promotes holistic healing and improves overall well-being.*

b. ***Accessibility****: AI tools streamline mental health services and provide immediate support, covering a gap to traditional treatments. This ensures that more individuals can access the help they need, regardless of location or financial barriers.*

c. ***Personalized Experience:*** *Integrated approaches enable tailored interventions that can be adapted to individual needs and preferences. This personal involvement improves the effectiveness of treatment, leading to the best results and a more satisfying treatment experience.*

Additionally, *K-pop* offers a valuable case application for this holistic program, given the rich story worlds and concepts combined with fans' love and attachment to their favorite idols, some of whom can be accessed through *AI* applications.

9 Study Limitations & Subsequent Research

Regarding the limitations of this paper, both authors are fully aware of the restrictive parameters linked to the exploratory case study, which was selected as a presented application. Therefore, the mentioned research gaps and possibilities of merging therapies have only an advisory nature with the need for subsequent research with possible future administration in professional-guided therapy sessions, as well as individual applications after undertaking a specialized training process provided by professionals in specialized facilities.

10 Conclusion

Unveiling how to provide the most appropriate care to those who need to overcome difficulties while restarting their coping mechanisms can be quite diverse and tricky. Understanding the individual treatment plan is the key. While previous research has been focused on single therapy treatment, nowadays there is a strong case for merging them. Encountering situations through the safe space of stories allows individuals to better understand actions and emotions, develop empathy and progress in their healing journeys. *Bibliotherapy* allows readers to internalize and adopt positive traits from characters they resonate with, while *psychodrama* facilitates deeper understanding and emotional growth through role-play and perspective-taking. In the realm of *AI chatbots*, this effect can be harnessed to create more empathetic and personalized interactions by mirroring users' language and emotional tone. With *K-pop's* use of *AI*, there is untapped potential for precision mental health care. Overall, these techniques underscore the importance of storytelling in fostering empathy, engagement, and therapeutic outcomes across these diverse fields.

Acknowledgements

We want to thank *Mgr. Michal Schwarz, Ph.D.* from *Masaryk University* for his support during the process of writing this paper.

References

[1] M. C. Pino and M. Mazza, “The Use of ‘Literary Fiction’ to Promote Mentalizing Ability,” PLoS One, vol. 11, no. 8, 2016. Available: https://doi.org/10.1371/journal.pone.0160254.

[2] K. Oatley, *Such Stuff as Dreams: The Psychology of Fiction*. John Wiley & Sons Inc., 2011.

[3] E. R. Kandel, *The Age of Insight: The Quest to Understand the Unconscious in Art, Mind, and Brain: From Vienna 1900 to the Present*. Random House, 2012.

[4] M. E. Langeberg, "Bibliotherapy: A Systematic Research Review with Social-Emotional Learning Applications," *Illinois Reading Council Journal*, vol. 51, no. 4, pp. 32-45, Fall 2023. DOI: 10.33600/IRCJ.51.4.2023.32.

[5] A. C. Barbour, "Purpose and strategy behind the magic shop," *Journal of Group Psychotherapy, Psychodrama, & Sociometry*, vol. 45, pp. 91-101, 1992. [Online]. Available: https://api.semanticscholar.org/CorpusID:152183218.

[6] R. Rautiainen, "Using the Magic Shop in a work counselling group," *Centre for Playback Theatre*, 2002. [Online]. Available: https://www.playbacktheatre.org/playbacktheatre/wp-content/uploads/2010/04/Rautiainen_Magic-Shop.pdf. [Accessed: Aug. 19, 2024].

[7] E. Koile, "The Magic Shop: The therapist masquerades as a shopkeeper," *Voices: The Art and Science of Psychotherapy*, Spring 2011. [Online]. Available:https://www.aapweb.com/wp-content/uploads/2019/03/voices-sample-koile.pdf. [Accessed: Aug. 19, 2024].

[8] “Psychodrama,” *Types of Therapy*, GoodTherapy, 2016. [Online]. Available: https://www.goodtherapy.org/learn-about-therapy/types/psychodrama. [Accessed: Aug. 19, 2024].

[9] K. Litwińska-Rączka, "Jacob Levy Moreno’s Psychodrama as a Work Technique for Treating Patients in Group and Individual Psychotherapy," *Current Problems of Psychiatry*, vol. 19, no. 4, pp. 248–259. 2018 DOI: 10.2478/cpp-2018-0019.

[10] M. Song, X. Xing, Y. Duan, J. Cohen, and J. Mou, "Will Artificial Intelligence Replace Human Customer Service? The Impact of Communication Quality and Privacy Risks on Adoption Intention," *Journal of Retailing and Consumer Services*, vol. 66, 102900, 2022. DOI: 10.1016/j.jretconser.2021.102900.

[11] G. Dubey, *Sociální pouto v éře virtuality*. Fra, 2020.

[12] P. McAllister, J. Kerr, M. McTear, M. Mulvenna, R. Bond, K. Kirby, J. Morning, and D. Glover, "Towards Chatbots to Support Bibliotherapy Preparation and Delivery," in *Chatbot Research and Design - 3rd International Workshop, CONVERSATIONS 2019, Revised Selected Papers*, A. Følstad, T. Araujo, S. Papadopoulos, E. Lai-Chong Law, O.-C. Granmo, E. Luger, and P. B. Brandtzaeg, Eds., vol. 11970, pp. 127-142. Springer Nature, Switzerland, 2020. DOI: 10.1007/978-3-030-39540-7_8.

[13] C. Lazore, "How the BTS Universe Successfully Engages Thousands of Fans," *The BTS Effect*, 2019. [Online]. Available: https://www.thebtseffect.com/blog/how-the-bts-universe-successfully-engages-thousands-of-fans. [Accessed: Aug. 19, 2024].

[14] C. Lazore, "' Artists for Healing': Anxieties of Youth, Storytelling, and Healing through BTS," *BTS Global Interdisciplinary Conference*, 2020.

[15] D. Maass, *The Emotional Craft of Fiction: How to Write the Story Beneath the Surface*. Writer’s Digest Books, 2016.

[16] CCARE, "Mission & Vision," *Stanford Medicine*, 2019. [Online]. Available: http://ccare.stanford.edu/about/mission-vision/. [Accessed: Aug. 19, 2024].

[17] D. Horton and R. R. Wohl, "Mass Communication and Para-Social Interaction," *Psychiatry*, vol. 19, no. 3, pp. 215-229, 2016. DOI: 10.1080/00332747.1956.11023049.

[18] M. Pradeep, "Kpop Idol-Based Chatbots are Blurring the Lines Between Interaction and Explicit Obsession," *Screenshot*, 2022. [Online]. Available: https://screenshot-media.com/culture/internet-culture/kpop-idol-based-chatbots-dangers/. [Accessed: Aug. 19, 2024].

[19] S. Ortiz, M. Shin, M. Samuels, and A. Windsor, "What is ChatGPT and Why Does It Matter? Here’s What You Need to Know," *ZDNET*, 2023. [Online]. Available: https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/. [Accessed: Aug. 19, 2024].

[20] R. Kowert and E. Daniel, "The one-and-a-half sided parasocial relationship: The curious case of live streaming," *Computers in Human Behavior Reports*, vol. 4, 100150, 2021. DOI: 10.1016/j.chbr.2021.100150.

[21] D. A. Laffan, "Positive Psychosocial Outcomes and Fanship in K-pop Fans: A Social Identity Theory Perspective," *Psychological Reports*, vol. 124, no. 5, pp. 2272-2285, 2021. DOI: 10.1177/0033294120961524.

[22] M. Liu, H. Peng, X. Song, C. Xu, and M. Zhang, "Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness," *Internet Interv.*, vol. 27, 100495, Jan. 2022. DOI: 10.1016/j.invent.2022.100495.

Usability of intelligent assistive technology used by people with dementia and their caregivers

Klara Dečman
Cognitive Science, Occupational Therapy
University of Ljubljana & University of Vienna
Slovenia
kd1023@student.uni-lj.si, decmanklara@gmail.com

Abstract

Intelligent assistive technology with context-aware computing and artificial intelligence can be applied to assist a person with dementia and their caregivers with activities of daily living. This paper samples such technologies with a focus on current knowledge and practice concerning usability. We used a scoping study to address the objectives of the research. Our findings indicate that despite the importance of technology customization to individuals' needs and capabilities it is not commonly addressed in the literature. Furthermore, while researchers are aware of the concepts and aims of evaluating the usability of technology, they seem to face difficulties in assessing them.

Keywords

Activities of daily living, cognitive assistance, dementia, evaluation of usability, family caregivers, human-centered design, scoping survey, user experience

1 Introduction

Dementia is a neurocognitive disorder, typically chronic and progressive, characterized by impairments in cognitive functions such as memory, attention, orientation, and language [1, 2] to the extent that a *person with dementia* (PwD) is not able to independently complete *activities of daily living* (ADLs) [3]. *Personal* (pADLs) refer to basic physical needs such as dressing, toileting, bathing, and eating, while *instrumental* (iADLs) are essential for living independently in the community, such as preparing food, taking medication, and doing laundry [4]. The ICD-11 [1] identifies three *degrees of severity* of any type of dementia. In the *mild stage*, a PwD may live independently but requires supervision and/or support with iADLs, such as locating everyday objects, and handling finances. In the *moderate stage*, PwDs require support to function outside their home environment. They can accomplish only simple household tasks and experience difficulties with completing pADLs. In the *severe stage*, memory impairment becomes profound, though it varies by etiology. PwDs are fully dependent on others for pADLs and they often experience total disorientation in time and place.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.3

One of the most common diseases in old age, dementia is recognized as one of the most costly and burdensome health conditions [2]. Statistics suggest that the growing global population of older adults diagnosed with dementia reached 44.4 million worldwide in 2013, with projections indicating an increase up to 135.5 million by 2050. Concern over the limited availability of family and professional caregivers for this rapidly growing population is intensifying (ibid.). As the population ages, the number of potential caregivers decreases, and those available often lack the key skills to provide the necessary level of care [5]. Furthermore, as family caregivers become more involved while struggling to balance other familial and social roles and responsibilities, they often experience negative consequences on their health, such as burden, anxiety, depression, isolation, and sleep deprivation [6]. Technological innovation, including advances in communications, robotics, and sensors, are perceived as promising to tackle these challenges [5]. Specifically, *assistive technology* (AT) refers to a broad range of devices and systems designed to maintain or enhance an individual's functioning related to cognition, communication, hearing, mobility, self-care and thereby promoting their health, well-being, inclusion, and participation [7]. AT is not designed to perform tasks on behalf of the user, but are specifically designed to monitor the activities of cognitively impaired users and provide appropriate assistance, thereby enhancing the likelihood of achieving desired behavioral outcomes [8]. A specific category of AT, *cognitive orthotics* [9] or *cognitive assistive technology* [10] is designed to assist with *cognitive* tasks. For instance, AT is employed to remind PwD to take medication or that their family member is visiting them next day [9, 10]. PwDs and their caregivers routinely use low-tech aids, such as medication pill organizers, schedules, and notes. They are being offered high-tech aids, such as *intelligent assistive technology* (IAT) that employs artificial intelligence to assess whether and when an appropriate reminder or procedural guidance is necessary for task completion [11, 12]. Additionally, IAT should be *contextually aware*: able to examine its environment, react to changes within it, and thus provide help *when needed* [11].

Human factors and *ergonomics* are scientific disciplines focused on studying the interactions between humans and other components of socio-technical systems [13]. The aim of designing such products and systems is to minimize human error and enhance human efficiency. One attempt at managing human factors analysis and human errors is through the development and deployment of measurement standards such as the *Human Readiness Levels* (HRL) scale (ibid.). HRL complements and

supplements the *Technology Readiness Level* (TRL) scale, which captures the maturity of technology before and after its integration into a developing system [13, 14]: HRL emphasizes the readiness to develop technology for effective and safe *human* use, and it should capture human-related features of technology development [14]. Similarly to TRL, HRL scale is divided to nine stages: *basic research and development* of principles, concepts, and the application of human characteristics, performance, and behavior, along with guidelines incorporating human-centered requirements to enhance human performance and human-technology interactions (HRL 1…3); development and assessment of user interface design concepts and prototype simulations in *laboratory and real-world environments* (HRL 4…6); full-scale testing, verification, and deployment in an *operational environment* with representative users and system hardware and software (HRL 7… 8); and the final stage, where the system is actively used in the *operational environment* with systematic monitoring of human-system performance (HRL 9) [14]. HRL is closely linked to *user-centered design*, a framework for the design and development of new products or the assessment and evaluation of existing products that explicitly considers potential users' needs, wishes, and subjectively perceived limitations of the IAT [5, 9, 12].

Some key definitions from the ISO standard on ergonomics of human-system interaction [15] read: "3.13 *usability*: extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. … 3.3 *effectiveness*: accuracy and completeness with which users achieve specified goals. … 3.4 *efficiency:* resources used in relation to the results achieved. Typical resources include time, human effort, costs, and materials." Further important aspects of usability include *user satisfaction*, the "extent to which the user's responses resulting from use meet the user's needs and expectations; how *accessible* a product, system, service, or environment is to individuals with diverse needs, characteristics, and capabilities is another element of usability". As examples of differences in terminology, see e.g. [16], where *utility* refers to whether the design provides features that users *need* and *usefulness* covers how pleasant and easy to use technology is (usability) and whether it does what users need (utility).

2 Methodology

This short survey covers only some of the findings of my ongoing more comprehensive review of topical IATs [17]. For this survey, the following research questions to explicate the coverage of *usability* will be addressed:

1. What is the maturity of IAT for human use (= its HRL)?
2. (How) do the developers take the *progression* of the disease into account?
3. How exactly is technology being assistive?
4. How is the usefulness and usability of the technology evaluated (if at all)?

Our scoping study maps key concepts, main sources, and types of evidence available for the domain targeted. For methodological transparency, we followed the PRISMA-ScS checklist [18]. Our search combines electronic database platforms (APA PsycInfo, Google Scholar, IEEEXplore, ProQuest, PubMed, Scopus, Web of Science, as well as the digital library facilities of the Universities of Vienna and Ljubljana) with hand searches of electronic journals and literature identified through literature readings. For this survey, we included original articles, conference proceedings, and PhD thesis; written in English; and published "within the last decade" (i.e., since 2013). To be covered, IATs further had to meet the following *inclusion* criteria: direct applicability to dementia care, focus on assisting with ADLs, PwD, and/or family caregiver as a user. We *excluded* IATs developed for the support of other (if related) disabilities, such as traumatic brain injuries; that could only be used by professional caregivers.

3 Findings

We illustrate our findings for each research question using the technologies: COACH [10, 11] and AWash [19] (targeted ADL: handwashing); DRESS [20] (getting dressed); ToiletHelp [21], (using a water toilet); and Smart Toothbrush [22] (brushing teeth). The pADLs supported by these IATs must be performed regularly to maintain the person's independence, health, and overall well-being. As dementia progresses, PwD becomes increasingly dependent on others to complete ADLs, affecting their family caregiver and society (cf. section 1).

3.1 Human Readiness Levels

We assigned aggregated HRL scores according to the groups introduced in section 1, with most of the surveyed technologies ranking at HRL 7…8: COACH, AWash, and ToiletHelp. This likely results from our choice of targeted content, as we aimed to focus on IATs close to HRL 9. We mapped DRESS and Smart Toothbrush to the HRL range 4…6, as the first is about developing and evaluating a prototype in preparation for in-home trials with PwDs, while for the second only preliminary laboratory testing was conducted with healthy individuals.

3.2 Different stages of Dementia

The IATs selected for this article are intended to provide targeted assistance for different stages of dementia. ToiletHelp is aimed to be used by PwD in the mild stage of dementia, COACH in moderate to severe stage, DRESS and Smart Toothbrush in severe stage, while for AWash we have not found any explicitly targeted stage of dementia. We found no evidence of technologies taking into account individual differences and needs of PwDs and their caregiver, consequently, we were not able to find such technology that would be able to *adapt* according to the actual severity of dementia as disease progresses (cf. section 3.3). Such *customization* is needed as cognitive functions progressively deteriorate, with fluctuations in rating occurring throughout the day or as the system would be used over periods ranging from weeks to months or even years [20, 21].

3.3 Notions of Assistance

Assistance involves *interacting*, with *prompting* being an interaction strategy that has become widely popular also in the context of IATs. Within our target domain, we found *audio* prompts to be most common as they are part of COACH, AWash, DRESS, and Smart Toothbrush. Such assistance should guide

PwD through the sequential steps of the activity by pre-recorded voice commands. *Visual* prompts include *videos* of steps of activities (COACH); *pictures* of correct clothing items (DRESS); use of different *lights* to attract attention to the appropriate use of an object (DRESS, Smart Toothbrush); and *texts* with instructions (ToiletHelp). DRESS consists of *motivational* prompts in the form of songs or videos favored by the PwD are meant for when a PwD should get stuck in an activity, and are configured by the family caregiver. COACH has options for *increasing levels of support*: low-guidance and high-guidance verbal prompts, video demonstrations, or placing a call to the caregiver. DRESS offers the choice of *continuous* mode, which includes chronological directions across all steps of an activity, and *independent* mode, in which no audio prompts are provided while the PwD is donning a shirt, and the caregiver should receive text messages on their device either when help is needed or dressing is completed. Nominal assistance provided by ToiletHelp consists of acknowledgment messages displayed to reassure the users they have completed every step of the activity; when the need is recognized, instructions are repeated. If a user should still fail, an alert informs the caregiver the PwD is having trouble, along with a reassurance message being displayed to the PwD.

The IATs we identified can help guide PwD through activities, but it is crucial to tailor such assistance to individual needs and adjust it as dementia progresses [20]. While there are cases where differing/increasing levels of assistance are provided by IATs [10, 11], such adjustment is not commonly documented in the literature. Despite its importance, our research indicates that there is also a lack of consistency in the terminology used to describe the adjustment of IATs to individual needs (e.g. customization, personalization, adjustment, adaptation).

3.4 Usability

The resources we analyzed indicate a dearth of commonly used standardized usability tests; out of the systems surveyed, only Awash was assessed using the System Usability Scale (SUS) questionnaire [23]. Instead, information about the usability of IATs is often gathered through user interviews [10, 11, 20, 21], observation, and performance testing [10, 11, 19, 21, 22].

In terms of *effectiveness*, COACH and AWash users were able to independently complete more steps of activity and engage less with caregivers while using IAT. Regarding *efficiency*, the developers of the Smart Toothbrush have estimated its battery life, while those of DRESS considered the final product's cost. In terms of *user satisfaction*, caregivers noted several benefits of DRESS, including validation of memory loss, empowerment of PwD, promoting privacy and dignity, and providing caregiver respite. ToiletHelp was reported to increase PwD's autonomy, boost self-esteem and dignity, and reduce the burden on caregivers. Participants rated AWash with a positive user experience. On the other hand, difficulties in using the technology were due to varying stages of dementia, visual and sensory perception issues, the need to change routines, and *affordability* issues [20]. Users expressed dissatisfaction with long delays between tasks and the frequency of prompts [10, 11, 22], while overlapping video and verbal messages used in ToiletHelp caused distraction. The acceptance of IAT largely depends on its *utility* and its unobtrusiveness, which can encourage more consistent use.

The current understanding of usability reflected in the literature indicates that even when researchers are aware of the related concepts and terminology and aim to assess them, they have difficulties in doing so with unified questionnaires or standardized testing procedures.

4 Relevance of Cognitive Science

The goal of the inter-disciplinarity of Cognitive Science is to address the question of *how does the mind work* – why we do the things we do, think the way we think, and how we perceive the world around us – by trying to understand and explain underlying mental processes and mechanisms of human behavior from the point of view of each discipline [24]. In user interfaces, *computational models of human behavior* are used to describe and capture our understanding of typical user actions, predict future actions, and guide users toward improving their actions [25]. These computations are typically based on *internal symbolic knowledge representations*, allowing a cognitive agent to manipulate symbols to gain information about the external world and determine how to act effectively – plan and perform actions, and achieve specific goals [26]. Evolutionary psychologists view the information processing architecture of the brain to consist of *adaptive problem-solving systems* that use information to adaptively regulate physiology and behavior. In this perspective, attention, learning, emotion, and motivation all play key roles in minds work and how we respond to our environments [27]. In particular, motivation can *guide* cognitive processes: When a PwD becomes fatigued, their motivation to continue activities declines. IATs can help by providing motivational prompts, such as favorite music or videos, which evoke emotional memories. This is but an example of how, cognitive science provides crucial insights into how users perceive, process, and interact with technology and consequently affects both, the improvement of designs and testing of usability and usefulness. It is a "bridge" between applied artificial intelligence and user experience.

One important objective of applied artificial intelligence is the development of cognitive orthotics, designed to *enhance* and *expand the user's cognitive abilities* [28]. It is not about technology imitating human abilities, but rather *extending* them. The key focus is the importance of creating systems that combine human and machine components in a way that maximizes their individual strengths taking into account ethics. To design successful cognitive orthotics, *interdisciplinary* teams are needed to unite relevant knowledge and perspectives of professionals (such as computer scientists, engineers, physicians, cognitive psychologists, and neuroscientists) together with stakeholders and users of technology (ibid.).

5 Limitations and Future Work

As technology advances rapidly, future research should explore a wider range of IATs using novel modalities and supporting more diverse ADLs. This limited study cannot form generalized statements about IAT usability for PwD and caregivers, as

comparing specific ADLs is challenging due to variations in particular activity structure, cultural contexts, and dementia stages. We focused on a small subset of IATs addressing some pADLs, excluding those covering iADLs and multiple ADLs. [17] takes a step in this direction.

6 Conclusion

Dementia is becoming increasingly prevalent, posing a major societal, economic, and global health challenge. While extending the duration of PwD's stay in their private homes may be seen to help alleviate the strain on institutional settings, it in turn places a significant burden on family caregivers. While IATs are intended to enhance the independence of PwD and reduce the caregiver's burden, our literature review efforts suggest that usability aspects are not systematically assessed. This gap is also linked to current HRLs, which indicate that existing IATs are not fit for deployed use by PwD. Moreover, we find that IAT is often not designed to adapt to the progression of the disease, affecting its utility and usability. Heavy terminology such as *intelligent assistance* appears to be employed all too easily. Furthermore, practice in assessing and reporting usability appears to leave significant room for improvement.

Acknowledgments

I would like to express my gratitude to the University of Ljubljana and the University of Vienna for allowing me to choose a Master's thesis topic of personal and professional importance. I am particularly thankful to my supervisor Univ.-Lektor, Dipl.-Ing., Dr. Paolo Petta, for your guidance and the opportunity to learn from you.

References

[1] *International Classification of Diseases, Eleventh Revision* (ICD-11), World Health Organization (WHO), 2023. [Online]. Available: https://icd.who.int/browse11/l-m/en

[2] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu and M. Prina, "World Alzheimer Report 2015. The Global Impact of Dementia. An Analysis of Prevalence, Incidence, Cost and Trends," Alzheimer's Disease International, London, England, UK, 2015. [Online]. Available: https://www.alzint.org/u/WorldAlzheimerReport2015.pdf

[3] *Diagnostic and statistical manual of mental disorders*, 5th ed., American Psychiatric Association, American Psychiatric Publishing, Washington, DC London, England, 2013, pp. 602-614.

[4] American Occupational Therapy Association, *Occupational Therapy Practice Framework: Domain and Process*, 4th ed., vol. 74, AOTA Press, 2020.

[5] D. V. Dahlke and M. G. Ory, "Emerging Issues of Intelligent Assistive Technology Use Among People With Dementia and Their Caregivers: A U.S. Perspective," *Front. Public Health*, vol. 8, May 2020.

[6] T. Wangmo, "Caring for Older Adults with Dementia: The Potential of Assisted Technology in Reducing Caregiving Burden," in *Intelligent Assistive Technologies for Dementia*, F. Jotterand, M. Ienca, T. Wangmo and B. S. Elger, Eds., New York, Oxford University Press, 2019, pp. 95-109.

[7] World Health Organization. "Assistive technology." World Health Organization. Accessed: Feb. 24, 2024. Available: https://www.who.int/news-room/fact-sheets/detail/assistive-technology

[8] B. Bouchard, K. Bouchard, and A. Bouzouane, "A smart cooking device for assisting cognitively impaired users," *J. Reliab. Intell. Environ.*, vol. 6, pp. 107-125, April 2020.

[9] J. Evans, M. Brown, T. Coughlan, G. Lawson, and M. P. Craven, "A Systematic Review of Dementia Focused Assistive Technology," in *Human-Computer Interaction: Interaction Technologies*, M. Kurosu, Ed., Cham, Springer International Publishing, 2015, pp. 406-417.

[10] N. M. Dharan, M. R. Alam, and A. Mihailidis, "Speech-Based Prompting System to Assist with Activities of Daily Living: A Feasibility Study," *Gerontechnology*, vol. 20, pp. 1-12, 2021.

[11] S. Czarnuch, "Advancing the COACH automated prompting system toward an unsupervised, real-world deployment," Ph.D. dissertation, Dep. Philos., BME, University of Toronto, Toronto, Canada, 2014.

[12] A. J. Bharucha, V. Anand, J. Forlizzi, M. A. Dew, C. F. Reynolds, S. Stevens, and H. Wactlar, "Intelligent Assistive Technology Applications to Dementia Care: Current Capabilities, Limitations, and Future Challenges," *Am. J. Geriatr. Psychiatry*, vol. 17, pp. 88-104, February 2009.

[13] V. Newton, A. Greenberg and J. See, "Project Management Implications and Implementation Roadmap of Human Readiness Levels," in *HCIBGO 2017*, Cham, 2017.

[14] ANSI/HFES 400-2021: Human Readiness Level Scale in the System Development Process, Human Factors and Ergonomics Society, Washington, DC, 2021.

[15] ISO 9241-210: Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems, International Organization for Standardization, 2019.

[16] J. Nielsen, "Usability 101: Introduction to Usability," nngroup.com. Accessed May 17, 2012. [Online]. Available: https://www.nngroup.com/articles/usability-101-introduction-to-usability/

[17] K. Dečman, "Intelligent assistive technology and family caregivers of people with dementia: Does it work?" M.S. thesis, Dept. Cogn. Sci., Uni-Lj., Ljubljana, Slovenia, 2024 (forthcoming).

[18] A. C. Tricco et al., "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," *Ann. Intern. Med.*, vol. 169, p. 467-473, October 2018.

[19] Y. Cao, F. Li, H. Chen, X. Liu, S. Yang, and Y. Wang, "Leveraging Wearables for Assisting the Elderly With Dementia in Handwashing," in *IEEE Trans. Mob. Comput.*, vol. 22, no. 11, pp. 6554-6570, 2023.

[20] W. Burleson, C. Lozano, V. Ravishankar, J. Lee, and D. Mahoney, "An Assistive Technology System that Provides Personalized Dressing Support for People Living with Dementia: Capability Study," *JMIR Med. Inform.*, vol. 6, no. 2, 2018.

[21] I. Ballester, M. Gall, T. Münzer and M. Kampel, "Vision-Based Toilet Assistant for People with Dementia in Real-Life Situations," in *IEEE PerCom Workshops*, Biarritz, France, 11-15 March 2024, pp. 141-147.

[22] M. E. S. Jannati, "Design and Implementation of a Smart Toothbrush for Individuals with Dementia," M.S. thesis, Dept. ECE, The University of Manitoba, Winnipeg, Manitoba, Canada, 2020.

[23] J. Brooke, "SUS: A quick and dirty usability scale," in *Usability Eval. Ind.*, vol. 189, November 1995.

[24] D. Vernon, "Paradigms of Cognitive Science," in Artificial cognitive system: a primer, Cambridge, MA and London, England, The MIT Press, 2014, pp. 19-61.

[25] N. Banovic, J. Mankoff and A. K. Dey, "Computational model of human routine behaviours," in *Computational interaction*, 1st ed., A. Oulasvirta, P. O. Kristensson, X. Bi and A. Howes, Eds., New York, United States of America, Oxford University Press, 2018, pp. 377-398.

[26] J. L. Bermúdez, "Physical Symbol Systems and the Language of Thought," in *Cognitive Science. An Introduction to the Science of the Mind*, 3rd ed., Cambridge and New York, Cambridge University Press, 2020, pp. 99-121.

[27] L. Cosmides and J. Tooby, "Evolutionary Psychology: New Perspectives on Cognition and Motivation," *Annu. Rev. Psychol.*, vol. 64, pp. 201-229, January 2013.

[28] K. M. Ford, P. J. Hayes, C. Glymour and J. Allen, "Cognitive Orthoses: Toward Human-Centered AI," *AI Magazine*, vol. 36, issue 4, pp. 5-8, 2015.

Open Science and Goodhart's Law

Tomaž Pisanski
pisanski@upr.si
University of Primorska
Koper, Slovenia
IMFM
Ljubljana, Slovenia

Vladimir Batagelj
Vadimir.Batagelj@fmf.uni-lj.si
University of Primorska
Koper, Slovenia
IMFM
Ljubljana, Slovenia

Jan Pisanski
jan.pisanski@ff.uni-lj.si
University of Ljubljana
Ljubljana, Slovenia

ABSTRACT

The influence of Goodhart's law to the development of Open Science is discussed. Science Citation Index (SCI) and Open Access (OA) are important steps in the path from Science to Open Science (OS). The main conclusion is that flawed openness replaced quality in Open Science.

KEYWORDS

Open Science, Open Access, Article Processing Charges, Goodhart's Law, Free Journal Network

1 FROM SCIENCE TOWARDS OPEN SCIENCE

1.1 Science

Traditionally, scientists disseminated their findings by publishing their results in scientific journals. This is a key mechanism for knowledge transfer among scholars and therefore an important subject of cognitive science. In the old days, the process of writing a scientific paper was completely different. The author had to type the paper on a typewriter leaving spaces for handwritten greek letters, symbols and formulae. With the advent of copying machines only cumbersome paper "cut-and-paste" method was available. Smaller misprints were overtyped whilst larger corrections required replacing whole pages. Professional typists, not available for everyone, could speed up the process. Manuscripts were sent for publication by ordinary mail in several iterations, depending on the referees' requests.

Rise of technology quickly brought up big changes. The introduction of personal computers replaced typewriters by keyboards and drastically enlarged the population of those who were able to compose texts on a computer and simple editors introduced cut-and-past method of writing. Specialised software for producing high-quality scientific drawings and diagrams enabled publishers to request camera-ready manuscripts from the authors. Authors no longer focused only on the subject of their work but also on the look it will have when printed.

1.2 Characteristics of Classical Publishing Model

Classical publishing model was robust and healthy. It was free for authors. Certain journals were even paying author fees. Surprisingly, it was (almost) free for readers via libraries of public universities.

Main players involve authors, editors, referees, publishers, libraries, readers, universities, learned societies, funding agencies and taxpayers.

Publishing within classical publishing model was time consuming and required efforts from all parties. This somehow prevented the inflation and hyper-production of papers.

The model was mainly "subscription model" where articles were available in printed volumes of a journal. University and departmental libraries subscribed to major journals, covering selected fields of science. The contents of earlier volumes were available to library users. Most libraries were open to local community and also to visiting researchers. Several learned societies, universities and institutes published their own journals, associated with a given library and used them for exchange purposes. Instead of paying subscription to a similar journal they would simply exchange the journals. In this way a library was able to save money to subscribe to journals that were not available for exchange. This was an important way for wealthy western scientists to help scientists from Eastern block and third world countries. Later the revenue from scientific publishing was one of the main sources of income of major learned societies. Unfortunately, by acquisitions and mergers eventually a very small number of huge multinational publishers emerged. These publishing houses control the field of scientific publishing.

1.3 Transition to digital

The advancement of technology, in particular ICT (Information and Communication Technology), in the second half of the twentieth century with the transition from analog to digital completely transformed the process of scientific publishing.

The costs of all stages of publishing decreased. More and more work was transferred from publisher – printer to author. Publishing a paper became easy and inexpensive. The number of scientific journals started to grow even more rapidly.

Surprisingly, major publishers did not lower the cost of subscription to their journals. On the contrary, they started to bundle journals. If a library wanted to continue subscription to a journal it had to subscribe to the whole bundle of journals, many of which it had no interest in.

When papers became available in a pdf form, the need for printed versions decreased. This also meant there was no way to prevent an unauthorised access to the paper. The first electronic journals appeared.

There is a big difference between subscription to printed journal and electronic journals. Old volumes of printed journals remain in the library and are available to anyone having access to the library. On the other hand, volumes of electronic journals remain with the publisher who may deny access to the paid volumes after the subscription runs out.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.4

2 GROWTH OF SCIENTIFIC PRODUCTION; SHIFT OF GOALS

In the past century we experience tremendous growth of published scientific works. There are several factors contributing to this phenomenon.

2.1 Publish or Perish

"Publish or perish" is an aphorism describing the pressure to publish academic work in order to succeed in an academic career. It first appeared already in the first half of the twentieth century. For a long time the PhD was a sufficient proof of academic qualification. It was not unusual that the PhD Thesis was the greatest scientific achievement of a scholar. Unfortunately, the growth in the number of universities and scientists led to inflation of PhDs. Some scholars would simply stop doing research when hired. Employers began requesting a constant flux of publications as a proof of scientist's devotion to research. Those who were unable to maintain high publication rate would be discouraged to stay in academia. And quantity became a proxy for quality.

2.2 Quality control

Ever since scientific journals appeared in seventeenth century, the quality of publications was in the hands of scientists using the system of peer review. This was natural since everybody involved: authors, editors, referees and readers were scientists.

With the growth of number of journals it became clear that not all journals apply the same standards for accepting a paper for publication. Obviously there was a problem of quality control.

In mathematics there was a secondary system in place. It started in Germany before WWII. Eventually, three refereeing journals were established, one in the Soviet Union, one in the USA and one in Germany (nowadays at the European Mathematical Society). If a review of a paper does not appear in one of those refereeing journals, the paper is likely not to be interesting for mathematicians.

For science in general there is a similar publication, called Current Contents.

2.3 Citation index

If we know for each paper the list of papers it cites, we can also produce the inverse index, i.e. the list of papers that cite a given paper. Such index is called citation index. In 1964 Eugene Garfield conceived the Science Citation Index (SCI) [3]. Using citation index one can easily detect most follow-ups to a paper covering a topic of scientist's interest. Clearly there are certain limitations. One has to select a collection of journals from where papers and their references are included. This may introduce some bias.

2.4 Impact

On the other hand, if the database is stored in a computer one can easily perform some statistics. For instance one can store with each paper the number of its citations. One can also compute how many citations each author has. This may, again, help the scientist to select the papers to look-up and authors to follow.

However, it also leads to all kind of rankings. Citation indices became very useful not only to scientists but also to their employers and funding bodies. Instead of comparing the added knowledge of someone's research, it is "sufficient" to select the highest ranked candidate. Selection can be done by administrators or computers. No peer review is needed. The paradigm "Publish or Perish" was upgraded to "Be Cited or Perish".

3 BIBLIOMETRICS

With the Science Citation Index (SCI) a number of statistical measures were introduced that would help profiling an author, the work or the journal. The science of bibliometrics was born. It was later extended to scientometrics and ultimately to informetrics.

SCI introduced a number of measures, indicators or metrics, trying to capture certain properties of articles, authors and journals.

One such indicator is the *journal cited half-life*. It is the median article publication date for each journal citation during one calendar year. In general, the journal cited half-life is small for recent journals while it is large for older, well-established journals. On the other hand it depends, as any other indicator on the scientific field and the culture of publishing in that field. Finally, a large journal cited half-life indicates that publications in that journal remain relevant for a long time. Hence the new knowledge is persistent and not merely mundane. Nowadays, it would not be difficult to equip any bibliographic database with computation of journal journal cited half-life.

3.1 Journal Impact Factor

Notorious Impact Factor (IF) is a ratio between the number of citations in a given time period - usually a calendar year, to the articles, published in another time period - usually two calendar years before. Sometimes they present also 3-year or 5-year impact factors.

For some reasons a 2-year impact factor prevailed and became a standard. In certain sense a 2-year impact factor is complementary to a journal cited half-life. Definitely, a 2-year IF is not the best indicator for mathematics when compared with other sciences where citation culture is different. For instance, when 1756 SCIMAGO journals covering the subject area of Mathematics are ranked according to Cites/Docs. (2 year) for the year 2023, among the top 50 journals only one journal has Mathematics as the primary subject area.

3.2 Metrics and Ranking

Having different indicators for a set of journals is good. It gives a higher dimensional description of each journal. However, each indicator may be used for sorting and hence for ranking. There is a strong tendency to devise an indicator that would measure quality; an impossible task.

Never-the-less since early seventies the Impact Factor is considered by many a proxy for the quality of a journal. The false reasoning goes along the following lines:

- Outstanding scientists publish their work in high-quality journals.
- The work of outstanding scientists is frequently cited.
- High-quality journals have high impact factor.
- Wrong conclusion: Work published in a journal of high impact factor is of high quality.

3.3 Power law and related statistical laws

When plotting the distribution of ranked impact factors, one can observe the exponential decay. Impact factor $IF(r)$ of rank r journal is proportional to $1/r^{\alpha}$ for some constant α. This is known as the power law. Roughly speaking this means that there are only a few journals with high impact factor and there are

many journals with small impact factor. One way of stating this is that 20 percent of most cited journals receive 80 percent of citations; see [7]. Several of these laws were first observed in bibliometrics. However it is interesting to observe that these laws are universal and apply to a variety of unrelated situations, perhaps by choosing the right value of parameter α.

4 GOODHART'S LAW

British economist Charles Goodhart is credited with expressing the core idea of a law in a 1975 article on monetary policy in the United Kingdom. "When a measure becomes a target, it ceases to be a good measure"; see [4]. Another way of saying is "that once a metric is used as a basis for decision-making or control, it loses its reliability as an accurate measure".

The main rationale behind this law is adaptation or even gaming to improve one's rank. If high rank means high reward, it is plausible, that some people will do anything to improve their score for the given indicator. Each measure for assessment of researchers and journals became prone to Goodhart's Law.

4.1 Goodhart's Law and Bibliometrics

When the number of publications are counted, researchers will tend to split long papers and publish short bits and will thus increase the number of publications. Instead of publishing papers alone they may increase their output several times if more coauthors sign the same publications. There is no increase in quality of their output.

When the number of citations decide who is winning a grant, the number of citations soared. The authors started citing their own papers, even if citations were not needed. When self-citations ceased to count, friendly researchers helped each other with citations.

When the h-index was introduced, the key publications of potential PIs in a research group had to be cited.

Employers and funding bodies understood that blindly rewarding high production authors with large impact papers does not mean rewarding high-quality science as there was no problem in publishing papers in low-quality journals and getting many citations in such journals. On the contrary, in many cases those fabricating papers and citations easily outperformed best researches. That is why the quality of journal in which the paper was published became important; in practice this meant journals with high impact factors.

By Goodhart's law, predatory publishers flourished, multiplying their journals and boosting their impact factors.

Production of new knowledge ceased to be important. It is the impact of their work published in high-impact journals that counts.

There is a difference in Goodhart's law and other laws, used in bibliometrics. Goodhart's law involves time and decision while laws based on power law are based on rankings.

There are not many studies of Goodhart'sl law in bibliometrics. An exception is a comprehensive study reported in [2].

5 OPEN ACCESS A STEP TOWARDS OPEN SCIENCE

5.1 APC model

The idea that authors or their institutions should make financial contributions for their publications is not new. In the times of paper publications, the publisher would grant some, say 25 reprints. For ordering extra reprints it was not uncommon to charge the authors. Also, one could be charged for insisting that the figures be printed in colour. On the other hand, some prestigious journals, started requesting article processing charges (APC).

Employers and funding agencies soon recognised that if they want their scientists to publish in the journals with very high impact factor, they will have to cover the costs of APC. Some scientific disciplines such as mathematics declined this model. When judging whether to pay APC or to send a graduate student to an international workshop many mathematicians give precedence to student. However, the publishers realised that money could be presented as a proxy for quality and raised their prices.

5.2 Diamond- and Green Open Access

In the last decade of the twentieth century some of the first purely electronic journals appeared. For instance, *The Electronic Journal of Combinatorics (E-JC)* was funded in 1994. It was free for authors and readers. It is run by scholars and not by commercial publishers. This is nowadays called a *diamond open access*, with no cost for authors and no cost for readers. E-JC is a founding member of the Free Journal Network [9].

Even before that, in 1991, an e-print server *arXiv* was launched where preprints in some scientific disciplines may be uploaded. Nowadays, such posting of preprint before peer review is called *green open access*.

For a while it seemed that this model will force big publishers to lower the prices of their journals. In the battle between scientists and multinational commercial publishing houses, the scientist should have won. It was expected that governments will support scientists in the fight against greedy publishers; [10]. However, politics works in mysterious ways.

5.3 Budapest Open Access Initiative (BOAI)

In December 2001 there was a two-day conference, producing a declaration called Budapest Open Access Initiative. The declaration was launched in February 2002, having 16 original relatively unknown individual signatories. This initiative has been financed by Soros' private Open Society Institute with 3 200 000 USD. It is recognised as one of the major defining events of the open access movement, [8]. Up till now it has been signed by about 0.1% of world scientists.

5.4 Gold Open Access and APC

Gold Open Access requires the author to pay Article Processing Charges (APC) to keep article freely available to the reader. Currently a typical APC exceed 3000 EUR. This brings enormous profits to publishers. It is estimated that the costs per article should not exceed 1000 EUR.

Clearly, APC model is not viable if costs are indeed covered by the author. The author must find someone who will cover the costs of APC. This is an ideal model prone for corruption at all levels. In the APC model, money becomes a substitute for quality. and researchers must compete for money that will cover their publication costs.

The difference between Green and Diamond Open Access and Gold Open Access is huge. One can speak of two opposing concepts sharing the same name: Open Access.

6 IMPLEMENTATION OF OPEN SCIENCE

6.1 Recommendations, Declarations, ...

There are numerous mostly political papers, initiatives, recommendations, declarations, pushing for Open Access, Open Science, Open Research, etc. Due to limited space we mention only a few of them. For more information, see e.g. [1, 5].

While the OA has been launched bottom up by 16 individuals meeting in Budapest, backed up by 3.2 Million USD from Open Society Institute, OS is a political concept that is revolutionising Science from top to bottom. It seems it was first formally expressed by UNESCO in November 2021 in the UNESCO Recommendation on Open Science.

The concept has been embraced by European Commission that pushes it through Horizon Europe down to member states. For instance, Slovenia recently received 16 000 000 EUR for promoting OS. It appears this money does not go for science but for administration.

The Barcelona Declaration on Open Research Information emerged from a workshop with over 25 experts interested in changing the research landscape. The declaration that was signed on 24 April 2024 is a political statement of an unidentified *community*. The authors do not act as individuals and do not represent scientific community. They write: *... we, as organizations that carry out, fund and evaluate research, commit to the following ...".* The first out of four commitments is strong. *We will make openness the default for the research information we use and produce.* It leaves no room for science outside Open Science. While OA was at first optional, OS makes it mandatory.

6.2 Goodhart's Law and Open Science.

Since journal impact factor remains a measure, the number of journals and publishers keeps increasing. In general, neither OS, nor universities nor funding organisations address the problem of low-quality high-impact factor predatory journals. Several scientists lower ethical standards and publish their papers in expensive journals with mild or no refereeing. The costs are reimbursed by their employer or funding organization.

Ever since the number of publications became a measure, scientists tend to publish papers with partial solutions to the problem. The number of co-authors per paper keeps increasing. The number of published papers grows out of proportion.

After citations became a measure, the number of references per paper keeps increasing. Some prominent journals fight citation inflation by limiting the number of references a paper may have. Clearly, the references published by competing authors are first to go.

Since APC remains as a valid model in OS, all kinds of unethical practices emerge. In many cases, a ghost author, who did not contribute to the paper but may secure covering APC costs is added to the list of authors.

It is disturbing that the goal quality is absent in some documents on OS, such as the Barcelona Declaration. The quality is replaced by openness and Goodhart prevails. Scientists will adapt to new goals.

7 CONCLUSION AND SUGGESTIONS

OS has some serious flaws. The main concern of OS is that scientists financed from public funds are not allowed to profit from their work – but everybody else can.

OS is only open to those within the system. Independent critical scientists adhering to high ethical standards are left out. OS is concerned only with current and future publications. No pressure to commercial publishers to open archives of papers published previously under paywall and make them free for everyone. A large part of science remains closed to authors and readers that are unable to secure money.

Scientists no longer decide what is the quality of their work. They even have to pay private companies to tell them that. For instance, public employers and public funders base their decisions about the quality of candidates on data bought from private companies running services, such as WoS or Scopus.

There is a problem of citation culture among different scientific fields. For eaxample, if average scientists from a scientific field, say A with high h-index compete for money in another field, say B they may be ranked higher than the best scientists of the field B. This may have negative effect on the future of the field B.

There is no real need for repositories at every public institutions. One repository at the European level with several backups would suffice. Instead of creating jobs for scientists repositories create jobs for administration. Repositories of papers and data are not intended for individual scientists. It appears they are intended for the AI data-harvesting algorithms of private companies. This service again will be sold back to scientists.

One could say, that the OS is a model that diverts public money from scientists to administration and private companies.

ACKNOWLEDGEMENTS

Work of VB is supported in part by ARIS (research program P1-0294 and research projects J1-2481 and J5-4596). Work of JP is supported in part by ARIS (research program P5-0361 and research projects J5-2551 and J5-4596). Work of TP is supported in part by ARIS (research program P1-0294 and research projects N1-0140, J1-2481 and J5-4596).

REFERENCES

[1] Batagelj, Vladimir. 2024. Bibliographic mix. [Online; accessed 21-September-2024]. (2024). https://github.com/bavla/biblio/blob/master/doc/sreda1348.pdf.

[2] Michael Fire and Carlos Guestrin. 2019. Over-optimization of academic publishing metrics: observing goodhart's law in action. *GIGASCIENCE*, 8, 6, (June 2019). DOI: 10.1093/gigascience/giz053.

[3] Eugene Garfield. 1964. "Science Citation Index"—A New Dimension in Indexing. *Science*, 144, 3619, 649–654.

[4] Charles E. Goodhart. 1975. Problems of Monetary Management: The UK Experience. In *Papers in Monetary Economics*. Reserve Bank of Australia.

[5] Kotar, Mojca. 2022. Open science in the european research area (era). [Online; accessed 21-September-2024]. (2022). https://url.um.si/p7CSj.

[6] Vojtech Kovarik, Christian van Merwijk, and Ida Mattsson. 2024. Extinction risks from ai: invisible to science? (2024). https://arxiv.org/abs/2403.05540 arXiv: 2403.05540 [cs.CY].

[7] Vilfredo Pareto. 1896. Cours d'economie politique, volume i and ii. *F. Rouge, Lausanne*, 250.

[8] Wikipedia contributors. 2024. Budapest open access initiative — Wikipedia, the free encyclopedia. [Online; accessed 31-August-2024]. (2024). https://en.wikipedia.org/w/index.php?title=Budapest_Open_Access_Initiative&oldid=1242834910.

[9] Wikipedia contributors. 2024. Free journal network — Wikipedia, the free encyclopedia. [Online; accessed 31-August-2024]. (2024). https://en.wikipedia.org/w/index.php?title=Free_Journal_Network&oldid=1231212463.

[10] Wikipedia contributors. 2024. The cost of knowledge — Wikipedia, the free encyclopedia. [Online; accessed 21-September-2024]. (2024). https://en.wikipedia.org/w/index.php?title=The_Cost_of_Knowledge&oldid=1239572934.

The Consistency of the Research Field Data
A Case Study of Library and Information Science in Slovenia

Jan Pisanski
Faculty of Arts, University of Ljubljana, Aškerčeva 2
Ljubljana, Slovenia
jan.pisanski@ff.uni-lj.si

ABSTRACT

SICRIS (Slovenian Current Research Information System) provides a service listing top Slovenian researchers in a particular research field. In Web od Science (WOS) each journal is assigned one or more categories (research fields). When comparing these data for the research field of library and information science (LIS), we found that several of the top authors in the field according to SICRIS rarely or never published in the journals deemed to belong to LIS in Web of Science. Several other authors, who were not assigned the research field of LIS in SICRIS, were among the most published Slovenian authors in LIS in Web of Science. This is an indication that results of any analysis of LIS in Slovenia will depend greatly on the criterion/criteria used.

KEYWORDS

Bibliometric Analysis, Research Fields, Slovenia, Library and Information Science.

1 INTRODUCTION

As part of a project focusing on high-level bibliographic services, i.e. novel services based on existing bibliographic data, we intended to perform a domain analysis of library and information science (LIS) in Slovenia from a bibliometric perspective. This contribution describes the initial step that was simply intended to provide an overview of research and researchers but came up on several issues regarding assignment of research fields and yielded some interesting findings, particularly for establishing the scope of the research field in Slovenia and elsewhere, but also in view of providing better services to the users of academic bibliographic databases.

2 BACKGROUND

There is a lack of a bibliometric overview of information scientists and librarians in Slovenia and their works, collaborations etc. One of the reasons is the nature of the field(s) of library and information science, where sometimes it is difficult to draw the distinction where the boundaries of the field are. On the other hand, relatively high-quality information on Slovenian researchers is stored in SICRIS (https://cris.cobiss.net/ecris/), the Slovenian current research information system, which provides multiple tools for basic bibliometric analysis.

Other studies have focused on the research fields in Slovenia (e.g., [1], [6], [8]), however at a more general level, not specifically for LIS and without mention of the issues related to research fields discussed herein, whereas [2] discusses among other things the mapping of WOS categories to the fields of science, used in SICRIS. Also see [2] for a brief history and overview of various mappings of fields of science/research fields.

While we were primarily interested in using bibliometric data for representation of a particular research field, this can then also be commonly used for evaluation of research. There are two main approaches: expert evaluation and bibliometric analysis. While expert evaluation is more traditional and qualitative, bibliometric analysis is quantitive in its nature. However, both of them have their downsides. For discussion on trustworthiness of experts, see e.g. [4]. Amongst others, Leiden Manifesto [5] points to dangers of using bibliometric data without closely examining the context. It suggests various indicators should be used when evaluating researchers and their work and that bibliographic analysis should support expert evaluation.

3 RESEARCH

While there are several different ways to approach the extent of publication on library and information science in Slovenia, we looked at the publications in Web of Science (WOS). This was done with intention to identify the most prominent works and authors, as journals indexed in WOS go through a rigorous process. However, this also means that we omitted from analysis all other publications, including papers published in Slovenian language journals.

Although it may not have the same coverage of social sciences, for this kind of insight WOS compares favourably to similar services, such as SCOPUS and Google Scholar, as it allows searching based on WOS Categories field which represents the subject categories/research fields of the journals [7]. It has to be noted that the WOS Categories field provides general information about the thematic nature of the journal rather than each particular paper. However, this is still the easiest way to get a quick overview of a research field, as all of the subject related data pertaining to individual papers in WOS describes the thematic nature of the papers in higher granularity. Each journal in WOS can be assigned one or more subject categories.

In April 2024, we performed a search in WOS Core Collection for publications where Address field included »Slovenia« and the value in the WOS Categories field was »Information Science & Library Science«. We did not limit the search to any particular time period, which means that the more experienced authors were more likely to be on the list. Also, we did not limit the results to particular types of publications (e.g. articles), since the "linked records" categorization in SICRIS, which we used in comparison, also does not limit this. However, even if we did, the situation regarding top authors would still be similar. Since Address was limited to Slovenia, the list excludes Slovenian authors who published research while working in other countries and may also be missing authors with otherwise faulty Address data.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.5

Table 1: Top 10 Slovenian authors in SICRIS, their research fields, number of established links to WOS publications in SICRIS and number of publications with LIS as WOS Category

Author	Research field 1	Research field 2	WOS	LIS	fraction
A	Information science and librarianship	Interdisciplinary research	89	0	0.0000
B	Administrative and organisational sciences	Information science and librarianship	55	2	0.0364
C	Information science and librarianship		55	47	0.8545
D	Economics	Information science and librarianship	53	14	0.2641
E	Computer science and informatics	Information science and librarianship	39	0	0.0000
F	Information science and librarianship	Plant production	33	17	0.5152
G	Information science and librarianship		31	29	0,9355
H	Information science and librarianship		24	2	0.0833
I	Information science and librarianship	Economics	24	3	0.1250
J	Information science and librarianship		21	18	0.8571

In SICRIS each researcher can be assigned up to two research fields, according to the ARIS (Slovenian Research and Innovation Agency) categorization, which is "roughly harmonized with the Field of Science and Technology Classification in the Frascati Manual (OECD)" [1]. There are different levels of categorization with the first level representing science, the second level representing field and the third level representing subfield. For instance, Information science and librarianship is deemed as a field belonging to social sciences with no further subfields. On the other hand, Economics also is a field of Social sciences, but it has subfields, such as Business sciences. Authors may be assigned a certain research field, even if it has subfields, or a certain subfield.

Among several features, SICRIS provides a higher-level service (https://cris.cobiss.net/ecris/si/en/top/researcher) where a user can look up most prominent Slovenian authors in a specific research field based on different indicators (e.g. number of linked records and citations in WOS and SCOPUS, h-index, other indicators linked to local evaluation practices). While this is not necessarily the only tool a user of SICRIS can use to get an overview of researchers in a research field, it is certainly the quickest and easiest to use.

Compared to some other research fields, where it is harder to find the equivalents in both of the databases, LIS has the advantage of being relatively straightforward. While the names used for the research field in the two systems slightly differ (»Information science and librarianship« in SICRIS; »information science and library science« in WOS), at least the core of the two subject categories should be the same.

While the actual ranking of LIS authors in SICRIS does vary slightly according to the indicator chosen (i. e. number of works in WOS and SCOPUS, number of citations, etc.) there is a core group of authors that occupies top places for several categories. Table 1 shows the top 10 authors based on the number of linked records in WOS according to SICRIS. In the table each author is represented with a letter of the alphabet for anonymity.

When comparing the data of LIS authors in WOS, whose address is in Slovenia (Table 2) and, the list of most prominent authors in LIS in Slovenia based on number of publications in WOS as provided by SICRIS (Table 1), we found a relatively large discrepancy. As seen in Table 1, half of the top 10 authors in LIS, as provided by SICRIS, had less than half of their works published in LIS journals, as indexed by WOS. In fact, for all five of these authors the proportion is less than one third.

There are several reasons for this phenomenon. In the SICRIS top 10 list, two prominent authors, marked in Table 1 as B and H, published a majority of their works in different fields, confusingly not explicitly named in SICRIS, before clearly switching their research interest to LIS. For some others their area of expertise is on the boundaries of LIS, although, what constitutes LIS can be debated. For example, two of the top 10 authors (A and E), including the top Slovenian author in LIS according to SICRIS, do not have a single work published in what WOS considers to be LIS journals. In the case of author A, their second research field, Interdisciplinary research, provides a better understanding on the nature of their publications.

According to the well-known Bradford's law [3] there are going to be some works published in journals that may not appear to be particularly relevant to a particular topic or research field. For instance, [2] found such distribution for Slovenian agriculture research group publication. However, there is still the question of whether such a list of top authors represents the LIS research field well.

It has to be noted that the results were similar even if we used other criteria in SICRIS. For example, the top 10 authors by number of citations in WOS are the same, only the order changes slightly. Also, the list of the top 10 authors by number of connected records in SCOPUS has two authors that do not appear in Table 1, neither of whom again had more than 2 works published in LIS journals, according to WOS.

To further complicate the matters in terms of transparency of data, SICRIS user interface only lists the author's first research field, in the top authors lists, which can be confusing to a novice user, as it may appear that some of the top authors do not belong to said field. In fact, many of the first year students of LIS at the University of Ljubljana skipped such authors, when asked to provide a list of top authors in the LIS field, based on SICRIS data/user interface.

Another issue that came up was that one of the top ten researchers is a foreign citizen with an ARIS researcher number having mostly worked outside of Slovenia. While this certainly reflects the international nature of science, it may not accurately reflect the state of LIS research in Slovenia. However, this issue is not particular to LIS.

On the other hand, there was also a notable group of authors that was not assigned to the research field of LIS in SICRIS, whose works appeared relatively frequently in LIS journals in WOS. Several new authors appeared in the top 10 list, if we only looked at the data on publications in WOS. Two of those, marked here

Table 2: Top 10 Slovenian authors by the number of publications in the WOS Category Information Science & Library Science journals and their assigned research fields in SICRIS

Author	LIS	Research field 1	Research field 2
C	47	Information science and librarianship	
G	29	Information science and librarianship	
J	18	Information science and librarianship	
F	17	Information science and librarianship	Plant production
K	16	Economics	Computer science and informatics
L	15	Mathematics	Computer intensive methods and applications
M	14	Information science and librarianship	
D	14	Economics	Information science and librarianship
N	13	Computer science and informatics	
O	13	Information science and librarianship	

as M and O, are authors whose field is declared in SICRIS to be LIS. But there are also three authors who do not have LIS named among their up to two research fields in SICRIS. Author here marked as K mainly worked in bibliometrics, which was also the LIS topic covered by author L, while author N mostly wrote on the topic of business intelligence. Such instances are not isolated, as several other authors who do not have LIS as a stated research field in SICRIS just missed the top 10 list.

4 DISCUSSION

While this is a brief look into a relatively small slice of two databases, SICRIS would benefit from a recognition of the issue. The simplest solution would be to provide a clear explanation on the nature of the data provided, when viewing top author lists by research field. Alternatively, additional services could be provided, based on other subject related data, such as WOS Categories or even keywords [7]. Ideally, services based on Bradford distribution would be provided.

The appropriateness of both the scope and designation of SICRIS research fields of authors and the WOS Categories can be debated. Their assignation procedures would benefit from greater transparency.

There is the issue of assignment of up to two research fields per author in SICRIS, as this does not necessarily accurately represent the involvement of each individual researcher. In our relatively small case study of LIS we found several authors whose assigned research fields could be viewed as misrepresented.

While research today is generally multidisciplinary and some researchers can shift their area of interest in research from one research field to another during their career due to various reasons, this ought to be reflected in any lists of researchers from a particular research field.

Also, while well-established, WOS would benefit from a more transparent explanation of the nature of WOS Categories. Even then, there can at least be a discussion, whether some of the journals are assigned to the correct research fields.

Conversely, as there are authors that publish relatively frequently in LIS journals in WOS, but do not have the according research field associated with them in SICRIS, a list of such authors could help with a subject classification of authors that is more reflective of their production.

5 CONCLUSION

Our research indicates that any bibliometric analysis of the research field of LIS in Slovenia is bound to be influenced by the criterion/criteria chosen to represent the field, as even the very top authors by one criterion may not be considered to be working in the field by another.

Further research could establish whether the issues found in this pilot study exist in other research fields and for other data (e.g. different databases, different time periods). However, not all research fields in one database may have their exact equivalent in another database. Cognitive science, for example, is not considered to be its own research field neither in SICRIS nor in WOS.

Generally, we suggest providing a clear explanation of the topical nature of the work of each author, when providing list of top authors in a research field, Another possible solution is omission of authors, who have a relatively low percentage of works published in journals from a research field from lists of top authors in that field.

While bibliographic databases offering high-level services that bring to light otherwise "hidden" data are definitely welcome, users would benefit from indication of imprecise nature of data and/or additional services that would try to account for the imprecision.

ACKNOWLEDGEMENTS

The author acknowledges the financial support from the Slovenian Research and Innovation Agency (research program P5-0361 and research projects J5-2551 and J5-4596).

REFERENCES

[1] Tomaž Bartol, Gordana Budimir, Doris Dekleva-Smrekar, Miro Pušnik, and Primož Južnič. 2014. Assessment of research fields in scopus and web of science in the view of national research evaluation in slovenia. *Scientometrics*, 98, 1491–1504.

[2] Tomaž Bartol, Gordana Budimir, Primož Južnič, and Karmen Stopar. 2016. Mapping and classification of agriculture in web of science: other subject categories and research fields may benefit. *Scientometrics*, 109, 2.

[3] B. C. Brookes. 1969. Bradford's law and the bibliography of science. *Nature*, 224, 5223.

[4] Alvin I. Goldman. 2001. Experts: which ones should you trust? *Philosophy and Phenomenological Research*, 63, 1, 85–110. DOI: 10.1111/j.1933-1592.2001.tb00093.x.

[5] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. 2015. Bibliometrics: the leiden manifesto for research metrics. *Nature*, 520, 7548, 429–431.

[6] Luka Kronegger, Anuška Ferligoj, and Patrick Doreian. 2011. On the dynamics of national scientific systems. *Quality & Quantity*, 45, 989–1015.

[7] Daria Maltseva and Vladimir Batagelj. 2020. Towards a systematic description of the field using keywords analysis: main topics in social networks. *Scientometrics*, 123, 1, 357–382.

[8] Tomaž Pisanski, Mark Pisanski, and Jan Pisanski. 2020. A novel method for determining research groups from co-authorship network and scientific fields of authors. *Informatica*, 44, 2.

To be or not to be… a Nahuatl language learning App. The long-term survival or discontinuation of indigenous language learning apps on the example of Nahuatl

Evelyn Fischer†
MeiCogSci Student
University of Vienna
Vienna Austria
evelyn.fischer@posteo.com

Abstract

Language learning apps for indigenous languages differ from the mainstream language apps in that they are not targeted at commercial success and might need to accommodate different linguistic and cultural aspects than the most learnt languages. The present paper considers the present and past Nahuatl language apps, some of which were discontinued, and asks what would be necessary for such apps to achieve long-term survival.

Keywords

Language learning, human-computer interaction, software usability, software translation, trends in software development

1 Introduction

The strong digitalization of modern life is bringing about big changes to the global and local societies. One of the results of the technological changes is the rapidly ongoing globalization, and one of the mechanisms of globalization is the shift of communities from languages with small numbers of speakers to a smaller number of global languages. In many countries with colonial history, this follows centuries of, at best, ignorance of, and at worst, active discrimination and eradication of the indigenous populations, their languages and their cultures.

The dominance of global languages is clearly seen on the Internet, where 80% of websites are written in just 8 of the estimated 7000 world´s languages [1]. The ascension to the digital realm is a challenging task and in 2013 [2] estimated that, at best, 5% of the world's languages will ascend to the digital world, and the rest will suffer a “digital language death”.

One of the ways a language can be present on the internet is by being the object of mobile learning apps. Mobile apps supporting the acquisition of minority and indigenous languages may differ from apps targeting global languages in that minority language learning apps would typically not be aiming at commercial success, would have lower budgets, or even be done on volunteer basis by smaller group of language activists.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

http://doi.org/10.70314/is.2024.cog.6

In addition, by pure chance, many of the most learnt languages have less complex morphology than many minority languages, and the apps that were developed with more analytic languages in sight, such as English, are not easily fully extendable to Morphologically Rich Languages, such as isiZulu [3], Turkish [4] and Nahuatl.

In the following, the focus lies on Nahuatl, the Mexican indigenous language with the highest number of speakers, 1.5 million. Nahuatl is one of 68 indigenous Mexican languages, and despite its historical prestige remains endangered, a challenge it shares with virtually all indigenous American languages. Nahuatl language learning apps contribute not only to thwart its digital death, but also to increase its visibility and prestige, and to support the efforts of Nahuatl learners to become “new speakers” [5] of the language. The role of new speakers is described as “very important, often essential for language revitalization projects” by [6] who work directly with Nahua and other minority groups in Mexico.

2 Initiatives to Localize Software and Platforms in Minority Languages

The second decade of the 21st century was a witness of increased efforts to increase the visibility and presence of minority and indigenous languages in the digital sphere. These were often led by digital language activists and sometimes supported by the companies whose software was the focus of the projects. The present section describes some of the initiatives taken and discusses the long-term results of the work to promote indigenous languages. Particularly, it looks at whether the work of the activists resulted in a long-term inclusion of the relevant language in the software or platform that was the aim of their efforts.

The Mozilla Foundation, known for its web browser, Firefox, launched in 2012 the initiative “Native Mozilla” that aimed to localize the browser into many of America’s indigenous languages. 50 languages from 10 countries were targeted [7], many of which are spoken in Mexico, such as Ch’ol, Kaqchikel, K’iche’, Mixteco (2 varieties: of Mixtepec and of Yucuhiti), Nahuatl (2 varieties or, by other accounts, 2 closely related languages: Highland Puebla Nahuatl of Mexico and Nawat Pipil of El Salvador), P’urhépecha and Triqui [7, 8]. The translations are done via the collaborative translation platform Pontoon (https://pontoon.mozilla.org/). As part of the initiative, for example, a Hackathlon was organized in Oaxaca in 2018 with representatives of 15 languages [8].

However, the goal of providing translations into 50 American languages is far from achieved and, as of 30.09.2024, Firefox 115 was available only in three American Indigenous languages: Kaqchikel and Triqui spoken in Mexico and in the South American language Guarani [9]. The other languages are in various state of completion: Ixil (13%), Kichwa (0%), Miahuatlán Zapotec (14%), Mixtec of Mixtepec (9%) and of Yucuhiti (29%), Nahuat pipil (0%), Paipai (1%), P'urhepecha (9%) and Quechua (3%). The situation is slightly better for the mobile phone browser Firefox Focus, which is available in Aimara, Ixil, Kaqchikel, K'iche', Maya, Miahuatlán Zapotec, Mixtec of Mixtepec and of Yucuhiti, Nahuat pipil, Navajo, Paipai, P'urhepecha, Quechua, Tének and Triqui [7, 10].

There were also attempts to localize the social media platform Facebook (https://www.facebook.com) into minority and indigenous languages. In 2012, the official Facebook translation platform (https://www.facebook.com/translations/), where users could add and vote on the accuracy of volunteer translations, had partial translations in about 100 languages, including Cherokee [11]. However, frustrated with the slow progress, Manuel Neskie made a browser overlay that allowed the translation of Facebook menu into Secwepemctsín directly in the user´s browser [11], which was later extended to many other languages, including Nawat [12] spoken in El Salvador. In 2015, a group of activists translated 24 000 words into Aymara and submitted it to Facebook for revision [13]. However, as of 31.08.2024, same as on 07.04.2019 [14], only three indigenous American languages were available for menus on Facebook: Cherokee, Inuktitut and Inupiaq, and the Facebook Translation App appears to have been discontinued.

A similar fate was met by the attempts to increase the number of languages offered on the Duolingo language learning platform (https://es.duolingo.com/). As of 10.06.2024, two indigenous languages are available on Duolingo: Hawaiian and Navajo [15]. However, previously also Guarani must have been available on the app, as evidenced by the surprise some users expressed at its sudden lack [16, 17]. The course is still available on the website version as of 31.08.2024 [18]. Between 2013 [19] and 2021 [20] new courses in development were stored in the Duolingo Incubator, where the users themselves could contribute to adding new languages to Duolingo. Indigenous languages such as Yucatec Maya and K'iche' were present on the Incubator and Duolingo itself credits the volunteers for helping to make, among others, the Navajo and Hawaiian language course [21]. However, Incubator was discontinued in 2021.

As we have seen, a common trend in the translation efforts of Facebook and Duolingo is the move from community and volunteer-based translations to commercial translation directed by the company. On the one hand, the reliance on unpaid work is problematic for a company with huge profits – something that Duolingo itself lists as the reason for ending the volunteer program [21]. This is especially true when those delivering this work might already be in unprivileged financial situation, as many indigenous language speakers are. On the other hand, however, this deprives the communities of the possibility to contribute to making their language more visible on the popular platforms. It is interesting to note that the translation platform of open-source based Mozilla products remains active, and, for example, the last changes to an Indigenous Language – Zapotec – have been done on 26th August 2024.

3 Digitally available Nahuatl language media

The importance of maintaining Nahuatl language learning apps is clearer if we consider that there are relatively few other resources in the language available online for language learners, many of whom are descendant from Nahuatl speakers and wish to reclaim the language of their ancestors.

[14] collected information about the different monolingual Nahuatl language media available for those searching on the internet. The results encompassed 12 monolingual novel-sized books, 5 scientific articles, 10 movies or series episodes, 1 videogame, 6 radio stations where Nahuatl is transmitted along with other languages, Wikipedia in 10 separate varieties, and 5 websites with a Nahuatl version. Considering that Nahuatl is the Mexican indigenous language with the biggest number of speakers, this is a low number.

However, there is hope that their number is increasing. For example, the 5 scientific articles were published in 1959, 2019, 2022, 2022, 2023; the 12 books in 2008, 2013, 2014, 2015, 2015, 2016, 2017, 2017, 2017, 2019 and 2021, (the twelfth is the Bible which has been translated into many varieties) [14]. One of the websites is also recent, as it was published in 2023 [22]. In 2024 the Mexican presidential election was simultaneously interpreted into Nahuatl, and remains, as of 30.08.2024, available on YouTube [23]. In addition, a Master´s degree in Nahuatl language and culture, taught completely in an indigenous language was launched in 2019 [24], first of its kind. We see therefore a clear tendency of growth, and it could be expected that more media will become available in Nahuatl soon.

It is also interesting to consider the case of the work to bring Nahuatl as a language available in Google Translate (https://translate.google.com/). In 2010, Google announced their plans to add Nahuatl and Maya to the tool [25], but this service was finally only introduced in June 2024 [26]. As of September 2024, the following ten indigenous American languages are among the 243 languages available on Google translate: Aymara, Guarani, Hawaiian, Kalaallisut, Mam, Nahuatl, Quechua, Q'eqchi', Yucatec Maya and Zapotec.

Considering mobile apps in particular, [14] identified 39 mobile phone apps related to Nahuatl. Most of them, 23, are Bible apps, although due to double versions, there are only 14 different variant versions of Bible available as a mobile app. The other 16 apps include 6 dictionaries (one with a Nahuatl user interface), 3 text collections, a (faulty) automatic translator, a multi-component app CEM, which combines dictionaries and morphological analyzer, the messaging app Telegram that offers user interface in Huasteca Nahuatl (albeit cannot be chosen in its standard menu, but is available for download for those who have the relevant link), and, finally, 4 Nahuatl language courses, discussed below.

4 A Partial History of Nahuatl Language Apps

As of 01.07.2023, four Nahuatl language courses were available for Android [14]: Aprende náhuatl [27], Beginner Nahuatl [28], Kamatlama [29] and NahuatlApp [30]. Aprende náhuatl (Spanish for "learn Nahuatl") is a vocabulary training app, with texts and videos, produced by the National Institute of Indigenous Peoples, a government agency and it was, as of 01.07.2023, downloaded more than 10 000 times [14]. As of 26.09.2024, it is available on

Android 14 for some, but not all devices. Beginner Nahuatl, with more than 1000 downloads as of the same date [14], was a vocabulary training app, without any game elements. Kamatlama is an app introducing basic numbers and fruit names and testing them through games and it was downloaded only more than 50 times as of 01.07.2023 [14]. NahuatlApp was an app introducing vocabulary items and testing them through a game and had, as of the same date, more than 10 000 downloads [14].

It is notable that the maintenance of the apps is far from ideal. Between the data collection of [14], 01.07.2023 and of [31], 30.05.2024, the videos of Aprende Nahuatl became unavailable and the two apps, Beginner Nahuatl and NahuatlApp, became unavailable for download on Google Play. During the same time, the user statistics did not change for the two continued apps, and crucially, Kamatlama hasn´t reached 100 downloads.

Two more apps mentioned by [14], although primarily dictionaries, also have elements supporting learning, such as quizzes: Totlahtol Nahuatl [32] and Diccionario Náhuatl [33]. As of 26.05.2024 Totlahtol Nahuatl, which offered Nahuatl user interface – as the only app other than Telegram - was no longer available for downloads, while Diccionario Nahuatl is still (01.09.2024) available. Additionally, Miyotl, a multilanguage app whose lesson components seem to never have been completed, remains available for download and contains a list of Nahuatl words and their Spanish translations [34].

In addition, as of 01.09.2024, 6 other apps mentioned by [14] are discontinued: the text collection Tlapohualiztli [35], the dictionary Diccionario Maya y Nahuatl [36] and the automatic translator Traductor Nahuatl [37]. This means that out of 16 non-Bible related apps mentioned by [14], 6 (Beginner Nahuatl, NauatlApp, Diccionario Maya y Nahuatl, Tlapohualiztli, Totlahtol Nahuatl, Traductor Nahuatl) have been discontinued only a year later (37.5%). In addition, [31] mentions three other apps that had been discontinued before: Tozcatl [38], Nahuatl Grammar [39] and Ma Tiwelikan Nawatl [40] – the latter is available as a website, but the App version is not available anymore. Furthermore, the app presented at the EUROCALL conference in 2016 [41] is also not discoverable on Google search, as of 18.06.2024.

However, the changes are not all negative. On 14th March 2024, user ItztliEhecatl posted on the social platform Reddit [42] that they have created a new Huasteca Nahuatl language learning course [43]. The author has been adding new items to the course, and as of 01.09.2024, there were 568 words and phrases to be learnt. The Huasteca Nahuatl course uses the Memrise Community Courses infrastructure, where users can create their own courses. However, in line with the trend discussed in section 2, Memrise is also closing community forums and removing community courses from their app and the future of the community courses remains uncertain beyond the end of 2024.

There is, however, another high-quality Nahuatl learning app for beginners that explains the grammatical concepts and tests them in a variety of exercises over 11 Units: the Nahuatl course hosted by the 7000 languages organization [44] and prepared in 2017 by Tlahtoltlapazolli, a Los Angeles based group [45]. The course requires registration, and only has a website version – although the mobile website version works well. However, the fact, that it is not listed in App stores decreases its findability.

Altogether, we see here that a great proportion of Nahuatl learning apps is discontinued. By the time they this happens, their content usually does not advance beyond the basic level, although one often has the impression that the authors had intended to add more lessons in the future. One could wonder whether a more stable app with more levels would have been possible if the authors had joined their efforts.

5 Long-term survival of Language apps: Discussion

The trajectory of some Nahuatl learning apps has helped us identify a trend of frequent discontinuation of those apps, lack of maintenance or upgrading to newer versions of operating systems or devices and a tendency by the commercial providers to dissolve community-led efforts of translation and localization and to limit the number of languages that the service is translated and localized into.

Admittedly, the frequent appearance and disappearance of new apps might be a sign of a vibrant, creative community. In fact, [7] sees it as a part of the process of app creation to accept that the results of one`s work on software localization might have a short life or never be used at all. If one accepts the possibility of failure (that is, the materials prepared ending up not being used) or only short-term success (that is, a short-lived app), the process of creating apps might be more spontaneous and less restricted, and the threshold to make such an app might be lower. In other words, if one does not strive to make “the perfect” app, creating an ad-hoc training exercises for one skill might become easier. However, even in this scenario, many contributors and authors might end up doing the same work unnecessarily, such as preparing grammatical description of the same grammatical forms destined for different apps.

In addition, one might also wonder if the low number of downloads discourages the authors to add additional levels and update their apps. However, given that Aprende Náhuatl, an App published by a government agency had more than 10 000 downloads, and up to 28 000 downloads [46], there is sufficient interest of learners in Nahuatl language apps, and perhaps a focus on better findability of the apps could result in their bigger success.

The question can therefore be posed how to better direct the efforts of authors and contributors, typically activists and volunteers, to not repeatedly make basic-level apps that might then be discontinued and to instead direct those efforts at more long-term apps which would also include levels for more advanced learners.

An open source platform that allows and tracks user edits, similar to Wikipedia (https://www.wikipedia.org/) or Wikitravel (https://wikitravel.org/) could allow users to collaborate in making language courses, and each individual author could make a small contribution, without the feeling that it was “in vain”. This would also prevent the fruits of work of language activists from “disappearing” in the chaos of the internet, and would increase their findability. Care must be taken to make such a platform independent of commercial companies that could unilaterally delete the courses from their servers. It is also recommended that information about such a platform be widely shared to avoid the situation where a good course ends up not being used due to being unknown to the learners.

Acknowledgments

I would like to thank the anonymous reviewer for their comments, which will serve as a guide in the subsequent research into the more specific reasons behind the discontinuation of the apps.

References

[1] Felix Richter, 2024. The Most Spoken Languages: On the Internet and in Real Life. Statista. https://www.statista.com/chart/26884/languages-on-the-internet/ (30.08.2024)

[2] András Kornai, 2013. Digital language death. PLoS Onen8(10):e77056. doi: 10.1371/journal.pone.0077056.

[3] Nikhil Gilbert, and C. Maria Keet, 2018. Automating question generation and marking of language learning exercises for isiZulu. In Proceedings of the Sixth International Workshop, CNL 2018, Maynooth, Co. Kildare, Ireland, August 27–28, 2018, pp 31-40.

[4] Fatih Bektaş,Bihter Dereli, Furhan Hayta, Erkin Şahin, Ubey Ali and Gülşen Eryiğit, 2022.Towards a Multilingual Platform for Gamified Morphology Learning, In 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2022, pp. 222-227, https://ieeexplore.ieee.org/document/9919484 (12.07.2024).

[5] Bernadette O'Rourke and Joan Pujolar, 2013. From native speakers to "new speakers"–problematizing nativeness in language revitalization contexts. Histoire Épistémologie Langage 35.2: 47-67.

[6] José Antonio Flores Farfán, and Justyna Olko, 2021. Types of communities and speakers in language revitalization. In Revitalizing endangered languages: A practical guide, pp. 85-103.

[7] Tania P. Hernandez-Hernandez and Bulmaro González Ambrosio, 2022. Mozilla in p'urhépecha: Translators' agency in a software translation project, Cadernos de Tradução, 42(01), pp. 1–20. doi:10.5007/2175-7968.2022.e85270.

[8] Netza López, 2021. Webinar – Localización de Firefox en Lenguas Indígenas. https://mozillanativo.org/ (30.08.2024)

[9] Firefox, n.d. about: preferences (30.08.2024)

[10] Pontoon Mozilla, n.d. https://pontoon.mozilla.org/projects/firefox/ (30.08.2024)

[11] Kevin Scannell, 2012. Translating Facebook into endangered languages". In Tania Ka'ai, Muiris Ó Laoire, et al., editors, Language Endangerment in the 21st Century: Globalisation, Technology and New Media. Proceedings of the 16th Foundation for Endangered Languages Conference, pp 106-110. https://kevinscannell.com/files/fel12.pdf (31.08.2024)

[12] Kevin Scanell, 2012. Facebook in Your Indigenous or Endangered Language. Rising Global Voices. https://rising.globalvoices.org/blog/2012/10/15/facebook-in-your-indigenous-or-endangered-language/ (31.08.2024)

[13] BBC News, 2015. Bolivia: Group translates Facebook into native language. https://www.bbc.com/news/blogs-news-from-elsewhere-34279302 (24.07.2023).

[14] Evelyn Fischer, 2024. El internet en náhuatl: la apropiación de las tecnologías de información y comunicación por una lengua indígena. Vienna: University of Vienna, Master´s Thesis. https://utheses.univie.ac.at/detail/70866/# (12.06.2024).

[15] Duolingo, n.d. Duolingo Language Courses. https://en.duolingo.com/courses/all. (04.06.2024)

[16] IBelieveDuoHasSh [fórum user], 2023, Guarani escondido [forum thread] https://forum.duome.eu/viewtopic.php?t=10075-guarani-escondido (31.08.2024)

[17] n1_kita [Reddit user], 2022. ¿Qué ha pasado con la lista de cursos de guaraní de Duolingo? [reddit post] https://www.reddit.com/r/duolingo/comments/10zvltv/what_has_happened_to_the_duolingo_guarani_course/?tl=es (31.08.2024)

[18] Duolingo, n.d Aprende Guaraní en solo 5 minutos diarios. Completamente gratis. https://es.duolingo.com/course/gn/es/Aprender-Guaran%C3%AD (31.08.2024)

[19] Duolingo wiki, n.d. Incubator. https://duolingo.fandom.com/wiki/Incubator (01.09.2024)

[20] Matt [user], 2023. What happened to the Duolingo Incubator? https://duoplanet.com/what-happened-to-the-duolingo-incubator/ (01.09.1988)

[21] Myra Awodey and Karin Tsai, 2021. Ending & honoring our volunteer Contributor program. https://blog.duolingo.com/ending-honoring-our-volunteer-contributor-program-2/ (01.09.2024)

[22] Thomas Kole, 2023: Retrato de Tenochtitlan. https://tenochtitlan.thomaskole.nl/es.html (27.03.2024)

[23] INE, 2024. Promueven INE e INALI interpretación simultánea del Tercer Debate Presidencial en tres lenguas indígenas nacionales. https://centralelectoral.ine.mx/2024/05/17/promueven-ine-e-inali-interpretacion-simultanea-del-tercer-debate-presidencial-en-tres-lenguas-indigenas-nacionales/ (12.07.2024)

[24] Universidad Veracruzana, 2024. Maestriah ipan Totlahtol iwan Tonemilis / Maestría en Lengua y Cultura Nahua. https://www.uv.mx/mlcn/ (12.07.2024)

[25] Tecpaocelotl. 2010."Google to add Maya, Nahuatl languages to search engine". https://tecpaocelotl.livejournal.com/4425.html (19.07.2023).

[26] Isaac Caswell, 2024. https://blog.google/products/translate/google-translate-new-languages-2024/ (01.09.2024)

[27] Instituto Nacional de los Pueblos Indígenas (INPI), 2022. "Aprende náhuatl [Mobile application]". (01.07.2023)

[28] shotgun.experiments, 2018. "Beginner Nahuatl [Mobile application]". (01.07.2023)

[29] Mario Albertio Duque Peralta [mario10412], 2022. "Kamatlama (LSM y Nahuatl) [Mobile application]". (01.07.2023)

[30] Juarez, Patricio, 2019.: "Nahuatl App [Mobile application]". (01.07.2023)

[31] Fischer, Evelyn, 2024. Self-study Tools of Morphologically Rich Languages: A Prototype of a Nahuatl Learning App. In Proceedings of the MEi:CogSci Conference 18. https://journals.phl.univie.ac.at/meicogsci/article/view/752 (12.07.2024).

[32] Tecuexe, 2016. "Totlahtol Nahuatl [Mobile application]". (01.07.2023)

[33] Mingatics, Patricio, 2021: "Diccionario Nahuatl [Mobile application]". (01.07.2023)

[34] Emilio Álvarez Herrera, Emilio, 2021: "Miyotl [Mobile application]". (01.07.2023)

[35] Tecuexe, 2020: "Tlapohualiztli [Mobile application]". (01.07.2023)

[36] Luján Castillo, José Dimas. s. f. "Diccionario Maya y Nahuatl [Mobile application]". (4.06.2023)

[37] Axcan, 2019. Tozcatl [Mobile application] https://tozcatl.soft112.com/

[38] Ibrahim Gerahard Peregrina Ochoa, 2021. "Traductor Nawatl [Mobile application]". (01.07.2023)

[39] Tecuexe, 2017. Nahuatl Grammar [Mobile application]. (01.07.2023)

[40] Rodrito García and Natalia Alonzo, 2017. Ma tiwelikan Nahuatl, https://vamosaaprendernahuatl.centroculturadigital.mx (12.05.2024)

[41] García-Mencía, Rafael, Aurelio López-López, and Angélica Muñoz Meléndez, 2016. An Audio-Lexicon Spanish-Nahuatl: using technology to promote and disseminate a native Mexican language. CALL communities and culture–short papers from EUROCALL (2016): 155159.

[42] Itzli Ehecatl [reddit user], 2024. Check Out My New Huasteca Nahuatl Language Learning Course. https://www.reddit.com/r/nahuatl/comments/1bekh07/check_out_my_new_huasteca_nahuatl_language/ (01.09.2024)

[43] Itztli Ehecatl, 2024. Huasteca Nahuatl https://community-courses.memrise.com/community/course/6566566/huasteca-nahuatl/ (01.09.2024)

[44] 7000 Languages, https://www.7000.org/

[45] 7000 Languages, n.d. Nahuatl https://www.7000.org/nahuatl

[46] Appbrain, n.d. Aprende Nahuatl https://www.appbrain.com/app/aprende-n%C3%A1huatl/com.nahuatlpuebla.app (26.09.2024)

Designing the Flow State Experience Using Modern Digital Technologies

Eva Vidmar
Multimedia
Faculty of computer science
University of Ljubljana
Ljubljana, Slovenia
eva.vidmar2@gmail.com

Abstract

This article provides a brief overview of an extended Master's thesis and focuses on the use of modern digital technologies to design a multimedia environment aimed at inducing a state of flow in individuals. Flow is a psychological state characterized by deep immersion in an activity, leading to a loss of sense of time and external worries **[1]**. Exiting this state typically results in feelings of satisfaction and happiness. Achieving flow requires a balance between skills and challenges. Learning to attain this balance can help individuals improve overall, which is a key reason for this research. The main objective is to investigate whether flow can be achieved through the use of color light stimuli in a space that adapts in real-time to an individual's level of attention. This represents a preliminary step toward using technology to design spaces that stimulate individuals and facilitate the quicker and easier attainment of flow. An experiment was conducted to test whether such a space affects individuals' flow. Tetris was chosen as the central activity for the experiment. The findings indicated that color stimuli influenced the participants' physically measured attention, although no significant changes were observed in questionnaire responses or gameplay performance. Given that attention is a crucial factor in achieving flow, it can be partially asserted that participants experienced flow, though more reliable data would be necessary for further conclusions. These findings significantly contribute to the understanding of measuring and achieving flow through technology, representing an important advancement in this field.

Keywords

flow, optimal experience, user experience, digital interface, Tetris

1 Introduction

Historically, flow has been experienced by artists, athletes, and individuals with substantial practice. However, modern life, especially in technology-driven environments, necessitates new approaches to achieve this optimal state. This study investigates how real-time adjustments of ambient lighting, informed by physiological signals, can enhance flow experiences, offering a novel approach that builds on but goes beyond methods such as video games or Virtual Reality (VR). By integrating psychophysiology, user experience, and ambient intelligence, this research aims to leverage technology for meaningful improvements in well-being, productivity, and satisfaction.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

2 Theoretical Background

Inducing a flow state through technology poses complex challenges that require in-depth exploration of the neurocognitive aspects of flow and their relation to contemporary technologies. This understanding informs the design of technological solutions aligned with flow theory.

2.1 A Neurocognitive Perspective on Flow

Flow, characterized by deep focus and immersion, was first described by Mihaly Csikszentmihalyi, often referred to as the "father of flow" [1]. This state occurs when individuals find an appropriate balance between their skills and the challenges they face, allowing them to perform optimally with a sense of effortless control [1]. While initial resistance and sustained motivation are necessary to achieve flow, this state can occur even in unfamiliar tasks, although long-term practice may increase its likelihood [2]. Flow is often illustrated in a two-dimensional graph where it exists at the intersection of appropriate challenge and skill levels [1].

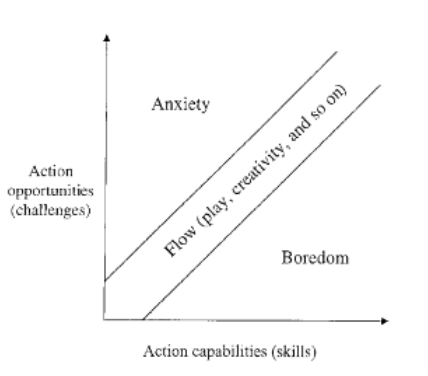


Figure 1: Graph representing occurrence of flow state [1]

Neuroscientific perspectives on flow suggest various underlying mechanisms. The transient hypofrontality hypothesis posits that during flow, activity in the prefrontal cortex decreases, reducing

self-criticism and enhancing performance [3]. Alternatively, synchronization theory proposes that flow arises from synchronized activation across different brain regions, optimizing cognitive efficiency [4]. Both theories indicate that flow entails minimal energy expenditure in the brain.

Csikszentmihalyi's model outlines optimal conditions for flow, including clear goals, immediate feedback, and a balance between challenge and skill [1]. The concept of the autotelic personality—marked by intrinsic motivation and enjoyment—may enhance flow experiences but is not strictly essential for achieving it.

2.2 Flow in Technology: State of the Art

Various approaches have emerged to induce flow through technology. Traditionally, video games employed fixed difficulty levels, which often interrupted the flow experience [5]. Dynamic Difficulty Adjustment (DDA) algorithms now allow real-time modification of challenges based on player performance, as seen in games like Left 4 Dead [6], [7, p. 4], Affective computing enhances this by using emotional indicators, such as facial expressions, to fine-tune difficulty levels [6]. Jenova Chen's game flOw exemplifies the integration of DDA with real-time adjustments to maintain flow [8]. Virtual Reality further immerses players, as evidenced by studies comparing VR to traditional 2D games and applications focused on meditation and relaxation [9], [10], [11]. Augmented Reality and multimedia art also contribute innovative avenues for inducing flow [12], [13].

Figure 2: Refik Anadol 2D projection Machine Hallucinations [12]

Our research identifies the use of ambient lighting as a promising yet under-explored method for inducing flow while engaging participants in a core activity, specifically Tetris. This game, created by Alex Pajitnov in 1985, has been extensively studied for its capacity to induce flow [14]. Players arrange falling blocks to form complete lines, receiving immediate visual feedback—key elements for maintaining flow. Research indicates that even brief sessions of Tetris can lead to flow experiences and reduced negative emotions [15].

2.3 Techniques for Flow State Measurement

Various methods exist to measure brain activity, with electroencephalography (EEG) being the most direct and commonly used. Functional Near-Infrared Spectroscopy (fNIRS) and Functional Magnetic Resonance Imaging (fMRI) provide insights into brain function, while Magnetoencephalography (MEG) offers high resolution of neuronal activity. However, these methods often involve expensive and less accessible equipment.

The MindWave Mobile 2, a consumer-grade EEG device, stands out for its ease of use, making it suitable for educational and entertainment contexts [16]. This device is ideal for our research due to its user-friendly nature, minimizing inconvenience for participants.

Figure 3: Mindwave Mobile 2 [16]

Flow state was traditionally measured through self-reporting instruments, such as the Experience Sampling Method (ESM) developed by Csikszentmihalyi [1]. Various questionnaires, including the Flow State Scale and Game Experience Questionnaire, have been developed to assess flow but rely on retrospective reporting. Alternatively, physiological measures may offer a more objective assessment of flow experiences.

3 Experiment: The Impact of Light on Flow State During Tetris Gameplay

This experiment investigated whether spaces incorporating adaptive technology could enhance user engagement. We compared standard Tetris gameplay to a version featuring color-changing lights that adjusted based on player attention, measured via the MindWave Mobile 2. The goal was to assess whether these technological enhancements positively impacted engagement and performance, specifically exploring if adaptive lighting improved attention, stabilized focus, and led to better gameplay results.

Drawing from Csikszentmihalyi's model, we recognized the importance of differentiating the environment in which flow activities occur. We aimed to create a highly engaging environment by designing a prototype of adaptive lighting for a dimly lit space.

For the experiment, we developed a color-changing light prototype controlled by the MindWave Mobile 2. The device measured brainwave activity during Tetris gameplay. We

utilized an Arduino Uno microcontroller to interface with the MindWave Mobile 2 and control a 2-meter AdaFruit NeoPixel LED strip. The light's color adjusted based on attention levels, with red indicating low attention, blue indicating high attention, and white representing optimal focus. This setup aimed to evaluate whether adaptive lighting could influence players' attention and flow during the game.

Figure 4: The setup of light prototype behind laptop

The experiment was designed to compare Tetris performance with and without adaptive lighting. Participants played Tetris under both conditions, and their attention levels, measured via the MindWave Mobile 2, were used to adjust the light's color dynamically. Data on engagement, attention, and gameplay performance were collected and analyzed to determine the effectiveness of the adaptive lighting in enhancing flow.

Figure 5: Scenario of playing Tetris with lights

3.1 Results

Data were collected from the MindWave Mobile 2, which recorded attention and meditation levels during Tetris gameplay with and without adaptive lighting. We filtered data to focus on attention values from 5 minutes of gameplay, excluding values below a threshold of 10 and retaining the 300 most representative data points. A Shapiro-Wilk test confirmed normal distribution for both conditions. A paired t-test revealed significantly higher attention levels during gameplay with lights (p-value = 0.00032). Notably, attention levels were more stabilized with adaptive lighting, as evidenced by a smaller variance in attention scores compared to gameplay without lights.

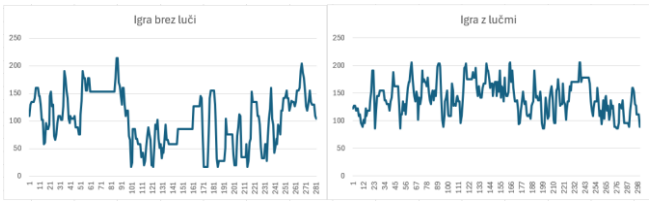


Figure 6: Attention levels of a player during gameplay without lights (left) and with lights (right)

For engagement, while the Shapiro-Wilk test confirmed normality, a paired t-test indicated no significant effect of lighting on engagement (p-value = 0.668). Tetris performance was assessed based on scores, with the group using lights achieving a higher average score (9377) compared to the no-light group (8979), though a Wilcoxon signed-rank test revealed no significant difference (p-value = 0.33).

Qualitative analysis of interviews with 40 participants identified seven key themes related to their experiences with the lighting: awareness of external stimuli, control and feedback, concentration, immersion, motivation, satisfaction, and pressure. Many participants reported feelings of pressure and stress, underscoring the challenges of achieving flow. Nonetheless, the lights were generally perceived as motivating, and some participants noted decreased awareness of their surroundings, aligning with theories regarding transient hypofrontality and reduced default mode network activity during flow [1].

Interestingly, some participants reported not noticing the lights at all, suggesting a potential subconscious influence of flow on their experience. This observation could have affected the questionnaire results. In terms of color perception, red was described as stressful and distracting, while white and blue were regarded as pleasant, showing no significant difference between them.

4 Conclusion

This study explored the potential influence of external factors, specifically technology-based lighting, on the state of flow. While we observed increased attention levels during gameplay with lights, supporting the theoretical premise that flow involves synchronized neural networks related to attention and reward, our hypothesis remains unconfirmed. Qualitative interviews highlighted themes consistent with flow characteristics, such as immersion and motivation; however, the absence of statistically significant effects on engagement and gameplay performance indicates that further research is warranted. Future studies should involve larger, more diverse samples and consider additional metrics to assess flow states more comprehensively.

Overall, our findings offer valuable insights into integrating technology with flow theory, highlighting the potential for developing products that enhance focus and user experience. This research lays the groundwork for future innovations aimed at creating more effective tools for achieving optimal states of concentration and fulfillment in everyday life.

Acknowledgments

The research was conducted in collaboration with psychologist prof. dr. Andreja Avsec, who served as a co-supervisor of the thesis, and prof. dr. Gregor Geršak, an expert in psychophysiology and measurement, who acted as the primary supervisor.

References

[1] M. Csikszentmihalyi, 'Flow: The Psychology of Optimal Experience', 1990.

[2] O. de Manzano, T. Theorell, L. Harmat, and F. Ullén, 'The psychophysiology of flow during piano playing', *Emot. Wash. DC*, vol. 10, no. 3, pp. 301–311, Jun. 2010, doi: 10.1037/a0018432.

[3] FlowCode - Project Unity, *Transient Hypofrontality - FlowCode Lesson #7 / Flow state training*, (Jun. 26, 2020). Accessed: Nov. 11, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=P5q5l47hGzY

[4] R. Weber, R. Tamborini, and A. Westcott-Baker, 'Theorizing Flow and Media Enjoyment as Cognitive Synchronization of Attentional and Reward Networks', *Commun. Theory*, 2009.

[5] S. Khoshnoud, F. Alvarez Igarzábal, and M. Wittmann, 'Peripheral-physiological and neural correlates of the flow experience while playing video games: a comprehensive review', *PeerJ*, vol. 8, p. e10520, Dec. 2020, doi: 10.7717/peerj.10520.

[6] M. Zohaib, 'Dynamic Difficulty Adjustment (DDA) in Computer Games: A Review', *Adv. Hum.-Comput. Interact.*, vol. 2018, p. e5681652, Nov. 2018, doi: 10.1155/2018/5681652.

[7] 'Left 4 Dead on Steam'. Accessed: Jan. 13, 2024. [Online]. Available: https://store.steampowered.com/app/500/Left_4_Dead/

[8] 'flOw', Jenova Chen. Accessed: Nov. 24, 2023. [Online]. Available: http://jenovachen.info/flow

[9] 'Virtual Reality and Flow: Discovering the Impact of Virtual Reality on Flow States'. Accessed: Nov. 24, 2023. [Online]. Available: https://www.novobeing.com/blog/virtual-reality-and-flow-exploring-virtual-realitys-impact-on-peak-performance

[10] TGC, 'Flow', thatgamecompany. Accessed: Nov. 24, 2023. [Online]. Available: https://thatgamecompany.com/flow/

[11] 'Cosmic Flow: A Relaxing VR Experience on Meta Quest', Oculus. Accessed: Jan. 14, 2024. [Online]. Available: https://www.meta.com/experiences/quest/3872076276162726/

[12] 'Refik Anadol', Refik Anadol. Accessed: Jan. 15, 2024. [Online]. Available: https://refikanadol.com/

[13] 'Android Jones', Android Jones. Accessed: Jan. 15, 2024. [Online]. Available: https://androidjones.com/

[14] D. Lora, A. Sánchez-Ruiz, and P. Gonzalez-Calero, 'Towards Finding Flow in Tetris', 2019, pp. 266–280. doi: 10.1007/978-3-030-29249-2_18.

[15] 'Tetris: It could be the salve for a worried mind', ScienceDaily. Accessed: Dec. 02, 2023. [Online]. Available: https://www.sciencedaily.com/releases/2018/10/181025084012.htm

[16] 'MindWave Mobile 2 Quick Start Guide or User Guide / MindWave Mobile 2 / Knowledge Base - NeuroSky - Home Page Support'. Accessed: Mar. 28, 2024. [Online]. Available: http://support.neurosky.com/kb/mindwave-mobile-2/mindwave-mobile-2-quick-start-guide-or-user-guide

The Transparency of Nudging: Evaluating Its Impact on Personal Autonomy

Sabina Pajmon
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
sabina.pajmon@pef.uni-lj.si

Toma Strle
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
toma.strle@pef.uni-lj.si

Abstract

Nudges are a strategic approach that shapes decision-making environments and the presentation of options to steer individuals toward certain behaviors while maintaining their freedom of choice. The ethical concerns surrounding nudges center on their potential to undermine personal autonomy, particularly when individuals are unaware of the influence exerted on them (i.e., covert or non-transparent nudges). The proposed solution for preserving autonomy is to increase transparency, which includes disclosing the presence and purpose of nudges to the people that are being nudged. There are various types of nudges and different types and levels of transparency associated with them. The most problematic in terms of violating personal autonomy are the non-transparent ones, those that exploit automatic cognitive mechanisms (Type 1 nudges), those that use type transparency and those that disclose their nature only after the fact (ex post). New approaches such as nudge plus approach seek to protect personal autonomy by involving citizens in the creation of nudges and enhancing reflectiveness during the nudging process.

Keywords

Nudge, transparency, autonomy, ethics of nudging, nudge plus approach

1 Introduction

Over the past thirty years, psychology and behavioral economics have highlighted how various contextual factors systematically influence our decision-making and behavior. In public policy-making, these insights are crucial for effectively addressing societal challenges like global warming, obesity, and poor economic decision-making. The groundbreaking paper [1] and the book that followed that brought the importance of decision architecture to the attention of academics, policymakers, and the general public was Thaler and Sunstein's "Nudge: Improving Decisions About Health, Wealth, and Happiness" [2]. In their work, they propose various ways in which government and private organizations could encourage or "nudge" individuals toward actions beneficial to them, while promoting a method that preserves a strong commitment to freedom of choice. Behavioral insights show that the context of decision-making can lead us to act inconsistently with our otherwise well-informed intentions [2]. The traditional approach to public policy assumes people are perfectly rational economic subjects ("econs") who act optimally with accurate information and clear rules. While this is an admirable goal, Thaler and Sunstein warn that basing public policy on this ideal often leads to failure. The authors introduce the concept of a "nudge" and propose its use as a policy-making approach that can influence citizens' behavior while avoiding the pitfalls and issues of traditional regulatory approaches, such as prohibitions and punishments. The advantage of this approach is that policymakers can influence our choices and behavior in a cost-effective and efficient manner without restricting us with prohibitions or interfering with our choices [3]. Despite the high effectiveness and utility of nudges, ethical concerns arise regarding the preservation of autonomy, especially with nudges that operate covertly and influence us without our awareness. This article investigates various types of nudges and levels of transparency, with a focus on their implications for personal autonomy. It begins by elucidating key concepts—nudges, autonomy, and transparency—before analyzing how different types of nudges, alongside varying types and levels of transparency, affect the preservation or violation of personal autonomy. Additionally, the article proposes criteria for determining which types of nudges are compatible with the preservation of personal autonomy. Finally, it explores potential strategies to mitigate adverse impacts on autonomy, including enhancing transparency, fostering citizen participation, and integrating reflective practices into the design of nudges.

2 Definition of a nudge

Thaler and Sunstein define a nudge as any element of choice architecture that influences behavior in a predictable way without restricting options or significantly altering economic incentives [2]. A nudge subtly guides individuals toward better decisions while preserving freedom of choice by adjusting how choices are presented [4]. Unlike prohibitions or penalties, nudges steer behavior without limiting options. An example is placing healthy snacks at eye level in stores to encourage healthier choices [5]. Hausman and Welch [6] add that nudges influence choices without increasing costs or limiting options, highlighting the potential for manipulation, which raises ethical concerns discussed in later chapters.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.8

3 How nudges work: leveraging heuristics and biases

To grasp how nudges impact behavior and decision-making, we must rely on insights from behavioral science, which reveal that nudges exploit inherent imperfections in human decision processes—leveraging cognitive heuristics and biases [4]. A key element of the nudge approach is that heuristics and biases, which often serve as mental shortcuts, are utilized to the advantage of the choice architect. While these mental shortcuts can sometimes lead to suboptimal decisions, a nudge aims to harness them to promote better decisions [3].

4 Nudge and ethical issues

Although the theory of nudging presents a promising approach to public policy, it has faced significant criticism from both public and academic spheres. Over the past decade, a robust ethical debate has developed, featuring nuanced arguments both supporting and opposing the practice [3, 4, 6, 7, 8]. The primary critique centers on the idea that nudging involves manipulating choices, with concerns about potential misuse of power [3]. Critics argue that nudges can undermine free choice by subtly restricting rather than fostering individual decision-making. The core of nudging involves exploiting heuristics and biases, which often lead people to act in ways that deviate from their well-considered preferences. Bovens [8] contends that such mechanisms can compromise control over actions, raising worries that nudges might affect decision-making by diminishing rational or deliberate considerations. Additionally, he argues, the behavior change induced by nudges occurs, if not against citizens' will, then at least without their active consent and awareness; for broader discussion about this topic see also [9, Ch. “Avtonomija v svetu spodbud” (Autonomy in the World of Nudging), pp.. 81-100; 10].

4.1 Ethical dilemma of autonomy in nudge use

Although nudges have been shown to effectively influence behavior, critics argue they can be manipulative and threaten personal autonomy. Autonomy, a complex concept, is broken down by Schmidt and Engelen into four dimensions: the freedom to choose without external pressure, acting according to one’s desires and values (psychological autonomy), making rational decisions based on available information, and being free from domination or manipulation [5, 8]. Critics claim that nudges can undermine autonomy by subtly influencing behavior without explicit consent, raising concerns about democracy, especially if governments use nudges without informing citizens. Nudges that operate without notice are especially problematic, as they can influence decisions without individuals' awareness. As Ivanković and Engelen [11] argue, non-transparent nudges, which exploit less rational psychological mechanisms, undermine autonomy by denying people control and the ability to challenge, a right that should be protected in liberal democracies. Sunstein, however, argues that nudges maintain freedom by allowing people to opt out of the suggested behavior, a concept they call "libertarian paternalism" [12]. They believe nudges, unlike traditional regulations, don’t limit freedom but instead encourage choices that align with individuals' best interests.

4.2 Transparency as a solution to autonomy violation

A proposed solution is increasing nudge transparency, as it allows individuals to understand how nudges work and make autonomous decisions based on their values [8]. A transparent nudge is one where its purpose and the methods used to influence behavior are reasonably clear to the affected individual. Thaler and Sunstein moreover argue that nudges used by governments should be public and transparent, with officials ready to disclose their methods and motives. Sunstein further emphasizes that nudges must be visible, reviewed, and monitored to prevent violations of autonomy or dignity [10]. Transparency involves informing decision-makers about the presence and purpose of nudges, allowing individuals to remain aware of behavioral interventions, thus preserving their autonomy and freedom of choice [7]. [1]

5 Types of transparency in nudging

To better understand the impact of nudging on an individual's autonomy, it is crucial to first examine the different types of nudges, as they are not a uniform phenomenon; rather, they can be classified into various types [2]. Understanding these types is based on dual-process theory, which describes the two decision-making mechanisms that nudges can influence.

5.1 Dual process theory

Dual process theory, explored by Stanovich [13] and Kahneman [14], is key in Thaler and Sunstein’s work on nudges. It suggests the brain operates in two modes: fast, intuitive System 1 and slow, deliberate System 2. System 1 handles instinctive actions, while System 2 engages in reflective decision-making. Despite its acceptance, dual processing is contested, with some scholars arguing the differences are a matter of degree. De Neys [15], notes no conclusive evidence favors either model, and resolving this debate may not significantly enhance our understanding of human thinking mechanisms. In this article, we adopt the dual process theory model to categorize different types of nudges. This approach allows us to better understand and design interventions that leverage both intuitive and reflective processes.

5.2 Type 1 and Type 2 nudges

According to Hansen and Jespersen [3], nudges can be categorized into two types based on dual process theory. Type 1 nudges target automatic, non-reflective thinking (System 1) and operate unconsciously, such as subliminal advertising or visual stimuli that influence behavior without conscious awareness. These nudges can be ethically problematic, as they often lack transparency and may lead to decisions misaligned with personal values. In contrast, Type 2 nudges engage reflective, deliberate thinking (System 2), promoting informed and thoughtful decision-making. These nudges are transparent and pose fewer ethical concerns regarding personal autonomy.

[1] Empirical evidence is inconsistent regarding the impact of transparency on the effectiveness of nudges. Transparency may: reduce their effectiveness (by prompting reflection), make nudges counterproductive (if people resist disliked nudges), enhance their effectiveness (if people understand and support the underlying goals), or have no significant impact at all [10].

5.3 Different types of transparency of nudges

The transparency of nudges plays a crucial role in safeguarding autonomy and freedom of choice, yet this concept itself is multifaceted. On one side of the spectrum, some nudges are explicitly transparent, functioning effectively because the individual is fully aware of the influence being exerted. Conversely, some nudges operate more subtly, relying on a lack of transparency to achieve their intended effect. To thoroughly assess which forms of transparency in nudges may raise ethical concerns, it is important to analyze the various ways in which transparency can manifest within nudges.

5.4 Type and token transparency

Bovens [8] introduces a crucial distinction between type and token transparency in nudges. Type transparency refers to when governments inform citizens about the general techniques they employ to intervene in decision-making contexts for the purpose of enhancing well-being. In this scenario, the government is open about the categories of measures it plans to implement. For example, when a government announces its intention to use specific psychological mechanisms to address social challenges, it demonstrates type transparency by clearly stating the kinds of interventions it will use to influence individuals' behavior and decision-making [16]. However, Bovens stresses that this is not enough. In his view, subliminal advertising does not become more acceptable simply because it is openly acknowledged [8]. On the other hand, token transparency requires that each individual instance of a nudge is clearly recognizable, including how it was implemented. This method, referred to as "here and now approach," aims to ensure that nudges are transparent to those encountering them at the moment of their decision-making [12]. However, even if this were feasible, it seems absurd to demand that every nudge be accompanied by a notice of its use. Since choice architecture is often unavoidable, token transparency may be too demanding, according to Bovens [8].

5.5 Levels of transparency

Transparent nudges differ also based on when they are noticed by the nudged individuals. With nudges that are transparent in advance (*ex ante*), the user can see the nudge beforehand and can avoid it if they choose. An example is traffic light labels (green, yellow, red) for healthy, less healthy, and unhealthy food products [17]. In contrast, a nudge is transparent afterward (*ex post*) if the target person only notices its influence after it has already affected them. Examples include fake cracks painted on the road to slow down drivers or the use of default options in certain contracts. Only after experiencing the effects do people realize they were influenced by a nudge [12]. Unlike the first category, the potential impact of such nudges on people's autonomy is more significant here. Ex post transparency may be insufficient to ensure autonomous action if it depends on individuals' ability to avoid the nudge. If transparency is meant to ensure that nudges do not deter people from achieving their goals and values, then, according to Ivanković & Engelen, ex post transparent nudges should either be excluded or efforts should be made to turn ex post transparency into ex ante transparency [11]. Occasionally, ex post transparent nudges become ex ante transparent through repeated exposure. For example, a fake speed bump may not have the same effect twice if the person learns when and where to expect it. With repeated exposure to such nudges, individuals may become more aware of their influence and may eventually avoid them altogether [12].

6 Types of nudges and transparency: impact on personal autonomy

The debate over nudges centers on how different types of nudges as well as types and levels of transparency impact personal autonomy. As stated in the article, nudges are divided into two types: Type 1, which influence automatic, non-reflective behavior, and Type 2, which target reflective decision-making. Transparent Type 2 nudges, which engage reflective capacities, do not typically raise ethical concerns, as they allow for conscious and deliberate decision-making. In contrast, non-transparent Type 1 nudges, which act on automatic processes, can threaten autonomy by influencing behavior without the individual's awareness. This may lead to decisions misaligned with personal values or goals. Transparency is categorized into type transparency (general awareness of the nudge type) and token transparency (awareness of mechanisms of specific nudges). The former is particularly problematic, as it lacks disclosure of specific examples and mechanisms, leaving us potentially unaware of the influences on our behavior. Nudges can also be categorized by the level of transparency into two main groups. The first group includes nudges that are transparent in advance by design (ex ante). These nudges are openly presented, allowing users to consciously decide whether to respond to them. Such nudges generally do not threaten autonomy, as they encourage conscious and deliberate decision-making. The second group includes nudges that are only transparent afterward (ex post). These nudges can be problematic, as users may respond to them before realizing they have been nudged. Although information about the nudge is revealed later, it may already have influenced behavior in a way that threatens freedom of choice and autonomy [18]. In conclusion, the most problematic nudges, in terms of violating personal autonomy, are Type 1 nudges that exploit automatic cognitive mechanisms, lack transparency—where type transparency is more concerning than token transparency—or are only transparent afterward. Understanding and using nudges requires careful consideration of their transparency and impact on freedom of choice. While transparent nudges can serve as tools for encouraging thoughtful and autonomous decisions, non-transparent nudges, as well as Type 1 nudges, especially those with only type or post hoc transparency, must undergo thorough ethical scrutiny to prevent potential violations of personal autonomy.

Table 1: Classification of nudges based on their impact on personal autonomy

Nudges that violate autonomy	Nudges that do not violate autonomy
Type 1 nudges	Type 2 nudges
Type transparency	Token transparency
Ex post transparent nudges	*Ex ante* transparent nudges

This table helps determine whether a nudge preserves autonomy, but it's unclear how many criteria must be met to deem a nudge ethical or unethical. Further research is needed for clearer guidance.

6.1 Collaborative policy design: The nudge plus approach

The nudge plus approach extends beyond transparency by encouraging participatory engagement and reflection, viewing individuals as rational, reflective beings rather than passive agents. Unlike traditional nudging, which can influence behavior unconsciously, nudge plus focuses on democratic control and active collaboration between citizens and policymakers. Through methods like citizens' assemblies, participants are directly involved in policy design, contributing ideas that shape their environments. In the UK, medical sciences now require patient and public involvement in all research that includes patient populations. Similarly, adolescents are consulted in developing anti-bullying interventions [19]. These approaches foster mutual feedback and collaboration between policymakers and citizens, leading to more inclusive and transparent policies that respect community values. Nudge plus approach also refers to an intervention that has a reflective strategy embedded into the design of a nudge. Banerjee and John [20] state that this preserves personal autonomy while promoting pro-social interventions through active involvement by enhancing token transparency and decision-making autonomy. The nudge plus approach offers significant potential for enhancing public policy with maintaining individual autonomy. By embedding reflection, transparency and active citizen engagement, it encourages people to participate in decision-making rather than passively accepting nudge type interventions. This participatory approach builds trust, as individuals are more likely to embrace policies that respect their autonomy and align with their values.

7 Conclusion

This article reviewed nudges as tools for influencing decision-making and behavior, with a focus on their transparency and its influence on potential infringement of personal autonomy. We found that nudges vary in type and transparency, which significantly affects their ethical acceptability. Type 1 nudges, which target automatic decision-making mechanisms, can diminish personal autonomy by influencing behavior without conscious awareness. In contrast, Type 2 nudges, which encourage reflective decision-making, are less problematic as they support autonomous decision-making. Nudges that lack token transparency or are only transparent after the fact are more likely to infringe on autonomy. Conversely, when nudges are transparent in advance and individuals are informed about them, autonomy is better preserved. Additionally, the context in which nudges are implemented plays a critical role in their ethical assessment, as the goals and values of the intervention must align with those of the individuals affected. In conclusion, the ethical use of nudges in public policy requires focusing on preserving autonomy by choosing Type 2 nudges and ensuring high levels of transparency, especially regarding specific examples and advance notice. This approach allows nudges to support conscious decision-making rather than serving as tools for covert manipulation. The Nudge Plus approach, which adds an element of reflection, can enhance both the effectiveness and ethicality of interventions, empowering individuals to make more informed decisions.

Acknowledgments

This pilot research study was partly supported by The Green Nudge project ("UL za trajnostno družbo – ULTRA") - European Union - NextGenerationEU, and Republic of Slovenia, Ministry of Higher Education, Science and Innovation.

Authors' statement

ChatGPT-4 was used for improving language of this paper.

References

[1] R. H. Thaler and C. R. Sunstein, "Libertarian Paternalism," American Economic Review, vol. 93, no. 2, pp. 175-179, May 2003.

[2] R. H. Thaler and C. R. Sunstein, Nudge: Improving Decisions about Health, Wealth, and Happiness. New Haven, CT: Yale University Press, 2008.

[3] P. G. Hansen and A. M. Jespersen, "Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy," 2013.

[4] A. T. Schmidt and B. Engelen, "The ethics of nudging: An overview," DOI: 10.1111/phc3.12658, 2019.

[5] T. Bucher et al., "Nudging consumers towards healthier choices: a systematic review of positional influences on food choice," British Journal of Nutrition, vol. 115, no. 12, pp. 2252–2263, 2016. doi:10.1017/S0007114516001653

[6] R. A. Abumalloh, O. Halabi, R. Ali, and D. Al-Thani, "Nudging Techniques: Design, Theoretical Grounds, and Ethical View," Journal of Knowledge Economy, 2024. [Online]. Available: https://doi.org/10.1007/s13132-024-02219-x.

[7] D. M. Hausman and B. Welch, "Debate: To Nudge or Not to Nudge," The Journal of Political Philosophy, vol. 18, no. 1, pp. 123–136, 2010.Sam An

[8] L. Bovens, "The Ethics of Nudge," in T. Grüne-Yanoff and S. O. Hansson, Eds., Preference Change: Approaches from Philosophy, Economics and Psychology, Theory and Decision Library A 42, Springer Science+Business Media B.V., 2009, ch. 10.

[9] T. Strle in O. Markič, O odločanju in osebni avtonomiji, 1. izd., let. 20. Maribor: Aristej, 2021, str. 145.

[10] Kognitivna znanost, Kognitivna znanost: zbornik 22. Mednarodne multikonference Informacijska družba - IS 2019, 10. oktober 2019: zvezek B = Cognitive Science. Ljubljana: Institut „Jožef Stefan", 2019. [Na spletu]. Dostopno na: http://library.ijs.si/Stacks/Proceedings/InformationSociety

[11] V. Ivanković and B. Engelen, "Nudging, Transparency, and Watchfulness," Social Theory and Practice, vol. 45, no. 1, pp. 43-73, Jan. 2019, doi: 10.5840/soctheorpract20191751.

[12] C. R. Sunstein, "The Ethics of Nudging," Yale Journal on Regulation, vol. 32, no. 2, pp. 413-450, 2015.

[13] K. E. Stanovich, Who is Rational?: Studies of Individual Differences in Reasoning. Milton: Psychology Press, 1999.

[14] D. Kahneman, Thinking, Fast and Slow. London: Penguin, 2012.

[15] W. De Neys, "On dual-and single-process models of thinking," Perspectives on Psychological Science, vol. 16, no. 6, pp. 1412-1427, 2021.

[16] K. Dowding and A. Oprea, "Nudges, Regulations and Liberty," British Journal of Political Science, vol. 53, pp. 204–220, 2023, doi: 10.1017/S0007123421000685.

[17] A. Arno and S. Thomas, "The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systematic review and meta-analysis," BMC Public Health, vol. 16, no. 676, pp. 1-13, 2016, doi: 10.1186/s12889-016-3272-x.

[18] J. Wachner, M. Adriaanse, and D. De Ridder, "The influence of nudge transparency on the experience of autonomy," Comprehensive Results in Social Psychology, vol. 5, no. 1-3, pp. 49-63, 2021, doi: 10.1080/23743603.2020.1808782.

[19] J. K. Madsen, L. de-Wit, P. Ayton, C. Brick, L. de-Moliere, and C. J. Groom, "Behavioral science should start by assuming people are reasonable," Science & Society, vol. 28, no. 7, pp. 583-585, Jul. 2024. doi: 10.1016/j.tics.2024.04.010.

[20] S. Banerjee and P. John, "Nudge plus: incorporating reflection into behavioral public policy," Behavioural Public Policy, vol. 8, pp. 69-84, 2024. doi: 10.1017/bpp.2021.6.

Ali nas uporaba velikih jezikovnih modelov v znanstvenem raziskovanju približuje časovni točki, ko bo stroj nadvladal človeka?

Does the use of large language models in scientific research bring us closer to the point in time when machines will surpass humans?

Franc Mali
Faculty of Social Sciences
University of Ljubljana
Ljubljana, Slovenia
franc.mali@fdv.uni-lj.si

Povzetek

Prispevek se ukvarja z vprašanjem, ali veliki jezikovni modeli v okviru generativne umetne inteligence že danes odpirajo vrata v fazo splošne umetne inteligence in morda – kot naslednji korak – v fazo umetne superinteligence. S tem bi bili dani predpogoji za prevlado strojev nad ljudmi. Pozornost je namenjena zlasti uporabi velikih jezikovnih modelov v procesu znanstvenega raziskovanja. Raziskovalna dejavnost predstavlja eno najbolj ustvarjalnih človekovih intelektualnih dejavnosti. Logično vprašanje je, ali je ravno znanstvena dejavnost, predvsem zaradi svoje kreativne narave, najbližja prečkanju te meje, ki predstavlja pomembno eksistenčno tveganje za celotno človeštvo. Osrednji del razprave je namenjen vprašanju, v katerih fazah današnjega znanstvenega raziskovanja je vloga velikih jezikovnih modelov že postala nepogrešljiva.

Ključne besede

generativna umetna inteligenca, veliki jezikovni modeli, znanstvena kreativnost, eksistenčno tveganje, okrepljeno učenje

Abstract

The article addresses the question of whether large language models within the framework of generative artificial intelligence are already opening the door to the phase of artificial general intelligence and, perhaps, as the next step, to the phase of artificial superintelligence. This would set the conditions for machines to dominate humans. Particular attention is given to the use of large language models in the process of scientific research. Research activity represents one of the most creative human intellectual endeavors. The logical question arises whether scientific activity, especially due to its creative nature, is the closest to crossing this boundary, which poses a significant existential risk to all of humanity. The central part of the discussion focuses on the question of which phases of today's scientific research the role of large language models has already become indispensable.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.9

Keywords

generative artificial intelligence, large language model, scientific creativity, existential risk, reinforcement learning

1 Uvod

V okviru pričujoče obravnave izhajam iz predpostavke, da se je skozi celoten zgodovinski razvoj umetne inteligence implicitno zastavljalo vprašanje, ali lahko ta doseže oziroma celo preseže človeško inteligenco. Že od začetkov razvoja umetne inteligence so bila tovrstna razmišljanja spodbujena z različnimi testi, ki naj bi med drugim nakazovali, ali se strojna "inteligenca" približuje človeški inteligenci. Pomembni premik v teh razmišljanjih se je zgodil, ko je tehnologija umetne inteligence prešla od klasičnih načel strojnega učenja k načelom delovanja globokih nevronskih mrež. V moji razpravi me v prvi vrsti zanima, ali najnovejši razvoj generativne umetne inteligence že kaže znake prehoda v fazo umetne splošne inteligence in morda – kot naslednji korak – umetne super inteligence. Posebej me zanima, ali najbolj kreativna področja človekovega intelektualnga delovanja, kot to predstavlja znanstveno raziskovanje, že odpirajo vrata nastopu umetne splošne inteligence. To namreč pomeni, da se počasi trasira pot nadvladi strojev nad človekom, kar je sicer predmet precej distopičnih razmislekov filozofov in družboslovcev, tako pri nas kot drugje v svetu. Moja obravnava ostaja na ravni nekoliko bolj splošne družboslovne refleksije o tej kompleksni tematiki in se ne ukvarja z ožjimi tehničnimi vidiki delovanja umetne inteligence, zato se bom v primeru sklicevanj na algoritme delovanja umetne inteligence oprl na nekoliko bolj poljudne definicije, kot so na primer tiste, ki jih je predstavil Partha Ray [1]. Po Rayu generativna umetna inteligenca (GUI) spada v skupino modelov umetne inteligence, ki lahko ustvarjajo nove podatke (informacije) na podlagi vzorcev in struktur, naučenih iz obstoječih podatkov (informacij). Ti modeli lahko

generirajo vsebine na najrazličnejših področjih, naj si bo besedil, slik ali glasbe. Pri analizi, razumevanju in ustvarjanju teh vsebin, ki vedno bolj spominjajo na človeške stvaritve, se opirajo na tehnike globokega učenja in nevronske mreže. Veliki jezikovni modeli (VJM), ki se razvijajo pod okriljem GUI, pa so zasnovani za generiranje naravnega jezika, kot so stavki, odstavki ali celotni dokumenti. Njihova ključna lastnost je zmožnost predhodnega učenja na velikih količinah besedilnih podatkov ter nato prilagajanje za specifične naloge uporabnikov. V prispevku v štirih krajših poglavjih razpravljam (1) o umetni splošni oziroma super inteligenci kot dejavniku tveganja o človeku, (2) o dilemah, ki so povezane z nadvlado stroja nad človekom, (3) o danes vedno bolj nepogrešljivi vlogi VJM v posameznih fazah znanstvenega raziskovanja, (4) o specifičnih problemih uporabe VJM na področju družboslovnega raziskovanja. Na koncu prispevka je podanih še nekaj zaključnih misli.

2 Umetna super inteligenca kot dejavnik eksistenčnega tveganja za človeka

Potem ko je Nick Bostrom pred desetimi leti postavil in utemeljil tezo, da obstaja verjetnost, da bo nadaljnji razvoj umetne inteligence pripeljal do nastopa umetne superinteligence, ki naj bi bila neprimerno bolj kognitivno zmožna kot človek, kar bi lahko predstavljalo eksistenčno tveganje za celotni človeško vrsto, ta tema, zlasti po nastopu GPT 4 in drugih vrst VJM (Bard, Claude, Llama, Gemini, itd.) vzbuja vedno večjo pozornost med strokovnjaki, tako med naravoslovci in tehniki kot tudi med družboslovci in humanisti [2]. V svojem prispevku se bom izognil (spekulativnim) ocenam, ki se vrtijo okrog problema časovnih mejnikov, ko (če) naj bi pametni stroji nadvladali ljudi. Ena skupina ekspertov namreč trdi, da se to ne bo zgodilo niti v sto letih [3], druga skupina ekspertov spet trdi, da gre zgolj za vprašanje dveh ali treh desetletij [4]. Bolj kot to, me zanima, ali uporaba GUI v takšni kreativni človekovi dejavnosti kot je znanost resnično na široko odpira vrata nastopu umetne splošne oziroma umetne super inteligence, ki naj bi se sicer zgodila v bližnji prihodnosti. To vprašanje je treba povezati s konkretno prisotnimi strahovi pred katastrofičnimi in celo eksistenčnimi tipi tveganj GUI, ki bi lahko imeli negativne družbene posledice. Če katastrofično tveganje ocenjujemo po kriteriju maksimalne razširjenosti (število ljudi, ki bi bili prizadeti), intenziteti (trpljenju, ki ga povzroča) in trajanja škodljivih družbenih posledic nekontroliranega razvoja posamezne tehnologije, potem pri eksistenčnem tveganju, ki naj bi bil povzean z umetno inteligenco, odločilno vlogo igra samo en kriterij: nevarnost iztrebljanja človeške vrste zaradi prevlade stroja nad človekom.

Neredko se srečujemo z ocenami, da pomembni predpogoj za varni prihodnji razvoj GUI, v okviru katerega se lahko izognemo eksistenčnim tipom tveganj, predstavlja »algoritem okrepljenega učenja« (v ang.: »reinforcement learning algorithms«) [5, 6]). Pri »okrepljenem učenju« gre za to, da se v procesu sprejemanja odločitev nagradi to, kar vodi v dobrobit ljudi. Vendar v konkretnih situacijah težavo predstavlja praktično usklajevanje funkcij umetne inteligence z sprejetimi družbenimi vrednotami. Čeprav se ta problem na prvi pogled zdi trivialen, temu ni tako. Družbene vrednote so raznolike, amorfne in jih je težko zapopasti v kvantitativnih kategorijah. Problem, kako »okrepljeno učenje« uskladiti z sprejemljivimi družbenimi vrednotami, zato ni nekaj, kar se da na zelo enostaven in samoumevni način razrešiti. Njegova razrešitev je odvisna od več dejavnikov. Eden izmed teh je možnost, da se modeli GUI razvijajo kot odprtokodni modeli, kar je seveda v nasprotju z sedanjo strategijo multinacionalk, da preko lastniškega nadzora novih naprednih tehnologij javnosti prikrivajo ključne informacije.

Negativna posledica lastniškega odnosa do VJM je, da znanje o notranjih mehanizmih delovanja VJM, ki predstavljajo vrh razvoja umetne inteligence danes, še vedno predstavlja izziv za večino uporabnikov, (deloma) pa tudi za strokovnjake s področja računalništva. Težko je namreč analizirati in priti na tej osnovi do razumevanja VJM, ki delujejo v okviru kompleksnih notranjih struktur z milijoni parametrov. Četudi lahko v vlogi uporabnikov ali celo računalniških razvijalcev vidimo končni rezultat delovanja VJM, pa je pojasnitev oziroma interpretacija njihovih notranjih struktur izjemno zahtevna. Skratka, veliki jezikovni modeli še vedno nastopajo kot »črne skrinjice« (»black boxes«). Thomas Arnold je za opis te nevzdržne situacije uporabil naslednjo posrečeno analogijo: »To je tako kot da bi se prizadevali za razlago delovanja kompleksne kemijske reakcije, ne da bi poznali natančno strukturo in interakcijo molekul.« [7] V strokovni literaturi se sicer omenja tudi nekaj izjem. Za modele kot so BLOOM, Cerebras-GPT ali Llama, naj bi podjetja, ki se ukvarjajo z umetno inteligenco, dopuščala večji javni vpogled [8]. Spet za druge so informacije za javnost odprli, potem pa ponovno zaprli. Četudi vrhunski znanstveniki, ki se ukvarjajo z UI in prihajajo iz akademske sfere znanosti, v vedno večjem številu opozarjajo, da je prosti dostop do vseh informacij na tem področju eden ključnih dejavnikov, ki lahko zagotovi verodostojno in zanesljivo raziskovanje, saj le tako lahko dostopamo do informacij o celotni »arhitekturi« VJM (t.j. od uporabljenih podatkovnih baz do algoritmov), v zvezi s tem še vedno ni bilo storjenih veliko sprememb.

3 Ali lahko ustvarjalno dimenzijo znanstvenega dela dokončno prevzame umetna inteligenca?

Na prihodnje izzive, ki so povezani z nastopom umetne splošne oziroma umetne super inteligence, je treba gledati tudi v luči današnjih dogajanj. Že danes si lahko zastavljamo vprašanje, ali bo ustvarjalno znanstveno delo dokončno prevzela GUI: ali je res upravičeno trditi, da kar je nekoč kalkulator pomenil za številke, in kar internet za globalni značaj komunikacijo, to danes pomeni za znanstveno kreativnost razvoj GUI? Znanstveno kreativnost lahko subsumiramo pod bolj splošni pojem inteligence. Ta naj bi načeloma izkazovala celo paleto zmožnosti, od kreativnih do racionalnih oblik (znanstvenega, umetniškega, itd.) mišljenja, od načrtovanja do učenja na temelju izkušenj, itd. Četudi danes spekuliramo, da bo splošna umetna inteligenca dosegla ali presegla inteligenčne zmožnosti ljudi, pa bomo v strokovni literaturi težko naleteli na neke soglasne kriterije, ki naj bi povedali, kaj predstavlja »inteligenca« pri strojih in kaj predstavlja inteligenca pri ljudeh. Formalne definicije, ki vztrajajo ne nekem skupnem imenovalcu, nam niso vedno v pomoč. Nobena izmed teh formalnih definicij ne ponuja nekega dokončnega kriterija, ki bi nam omogočal primerjavo »inteligentnosti« različnih entitet. Če se za hip ustavimo ob najnovejšem delu Yuval Noaha Harareja, ki nosi naslov »Nexus. A Brief History of Information Networks from the Stone Stage to AI« [9], bomo pri njemu hitro prepoznali besednjak, ki naj bi

nedvoumno nakazoval, da GUI poseduje moment intencionalnosti, t.j. sposobnost GUI slediti delovanju, ki izhaja iz njih samih. (Avtor knjige govori o tem, da se pametni stroji, ki jih vodi GUI, sami odločajo, izbirajo, delujejo, itd.). Ob prebiranju najnovejšega Hararejevega dela se lahko vprašamo, zakaj vsiljuje intencionalnost kot ključni kriterij za izenačevanje »inteligentnosti« človeka in stroja. Lahko bi uporabil širšo definicijo inteligence in bi le to pripisal že entitetam, ki so pasivne, torej ne vključujejo momenta intencionalnosti, vendar vseeno reagirajo na okolje in lahko opravljajo kompleksne naloge. To je na primer storil Sebastien Bubeck, ki je skupaj z soavtorji preučeval, ali so v jezikovnem modelu GTP-4 že dani zametki umetne splošne inteligence. Postavil je namreč tezo, da si neko inteligentno entiteto lahko predstavljamo tudi kot »orakelj«, ki nima notranjih vzgibov ali želja za delovanje, vendar lahko natančno in koristno zagotavlja informacije o kateri koli temi ali domeni vedenja [10]. Definicijo inteligence, ki izhaja zgolj iz kriterija intencionalnosti, imamo lahko za restriktivno še iz enega razloga. Če namreč pri tej definiciji izhajamo iz notranjih motivov za doseganja ciljev našega delovanja v kar se da širokem okolju, kjer se soočamo z nikoli zaključenim spektrom situacij, potem v primeru rabe takšne definicije implicitno predpostavljamo, da je pojem inteligence neizogibno vezan na univerzalnost in optimalnost. To pomeni, da spet operiramo z apriorno definiranim in ne aposteriorno preverjenim konceptom inteligence. Dejansko oziroma realno inteligenco človeka namreč nikakor ne moremo opredeliti kot absolutno univerzalno in optimalno.

S podobnimi dilemami se soočamo, če naš pogled usmerimo na kreativnost kot eno izmed dimenzij človekove inteligence. Tudi v tem primeru odgovor na vprašanje, ali umetna inteligenca enostavno privzema kreativne moči znanosti, ni enoznačen. Ne gre samo za to, da se že pri vprašanju kreativnosti človeka srečujemo z ogromnim številom definicij (znanstveniki uporabljajo danes več kot 50 definicij [11]), zadeve postanejo še bolj kompleksne, ko iščemo skupni imenovalec med definicijo človeške kreativnosti in kreativnostjo, ki jo pripisujemo umetni inteligenci. Na eni strani imamo avtorje, kot so Marc Ruco [12] ali Stephen Rice [13], ki pravijo, da kolikor k standardnim definicijam človekove ustvarjalnosti – ta vključuje dimenzijo originalnosti in učinkovitosti - dodamo tudi dimenziji avtentičnosti, potem GUI ne more tekmovati z ljudmi.

Na drugi strani imamo avtorje, kot na primer Hubert Kent, za katere je GPT- 4 že dosegel izredno visoko stopnjo znanstvene kreativnosti, vsaj kar zadeva t.i. odprti tip mišljenja, saj naj bi empirične analize pokazale, da GPT-4 že zmore doseči rezultate, ki so enaki rezultatom, ki jih doseže zgolj 1% najbolj inteligentnih ljudi [14]. Rezultati dodatnih študij naj bi ravno tako dokazovali, da model GPT-4 izkazuje veliko stopnjo fleksibilnosti zunaj ustaljenih okvirov mišljenja in naj bi imel na področju odprtega tipa mišljenja celo višji kreativni potencial od ljudi. Pričakovati torej je, pravijo avtorji, ki so opravili te in podobne študije, da bo model GPT-4, kolikor bo dosežen napredek glede povečanih zmožnosti učenja na velikih bazah podatkov in bolj napredni arhitekturi nevronskih mrež, kmalo storil pomembni korak v smeri umetne splošne inteligence.[15]. V tem primeru Turingovi testi že zvenijo zastarelo. V okviru rabe Turingovega testa gre namreč za to, da se kot kriterij izenačitve dveh inteligenc vzame situacijo, ko nek uporabnik, ki komuncira z klepetalnikom UI, ne zna več ločiti, ali je na drugi strani človek ali stroj [8].

4 Zakaj umetna inteligenca postaja vedno bolj nepogrešljiv pomočnik v vseh fazah znanstvenega raziskovanja?

V okviru moje razprave me ob bolj načelnem epistemološkem vprašanju, t.j. ali nova tehnologija umetne inteligence postopoma zavzema prostor znanstvene kreativnosti, zanima tudi bolj konkretno vprašanje: ali ta nova napredna tehnologija dobiva status nepogrešljivosti v vseh drugih fazah znanstvenega raziskovanja?

Sodobno znanstveno raziskovanje je multidimenzionalni proces, ki vključuje različne faze, ki od začetka raziskovanja do končne objave znanstvenih rezultatov segajo od najbolj rutinskih pa do najbolj ustvarjalnih aktivnosti. VJM v tem primeru prevzema vlogo koristnega in vedno bolj nepogrešljivega »asistenta« v vseh fazah znanstvenega raziskovanja. Bo ta »asistent« v bolj ali manj oddaljeni prihodnosti postal »profesor«, ki bo dokončno nadomestil človeka – znanstvenika?

1. Najprej je treba izpostaviti, da GUI zaradi svoje učinkovitosti vedno bolj nadomešča znanstvenike v postopkih pridobivanja podatkov. Podobno je z učinkovitostjo GUI v vsebinskem pregledovanju, povzemanju in sumiranju množice informacijskih virov, ki so kot »state of the art« relevantne v vsakem začetnem procesu znanstvenega raziskovanja. GUI je sposoben obdelave in analize velike količine podatkov. Vloga GUI postaja neprecenljiva pri pregledu in sintezi vsebin iz znanstvene literature, povzemanju podatkov in sinteze kompleksnih podatkovnih baz, samodejnem prepoznavanju vzorcev in trendov, ki jih je mogoče izpeljati iz podatkov, modeliranju in napovedovanju na temelju zbranih podatkov, itd.
2. Vedno bolj se povečuje vloga GUI pri ustvarjanju novih idej. Glede na današnjo eksponentno rast znanstvenih informacij je prenos te raziskovalne funkcije iz človeka na GUI hkrati povezana z zmožnostjo GUI, da učinkovito in predvsem avtonomno ustvarja nova raziskovalna vprašanja in hipoteze. Eden največjih izzivov najbolj naprednih področij znanosti je skorajda neskončno število hipotez, ki se nanašajo na raziskovalne probleme, zaradi česar se včasih zdi natančno sistematično raziskovanje, ki bi omogočalo sprejetje hevristično najbolj obetavne hipoteze, brez sodelovanja UI skorajda nemogoče. Primer: v biokemiji naj bi obstajalo približno 10^{60} molekul, to pa je praktično enako številu zdravil, ki jih je treba na temelju ogromnega števila molekul šele odkriti [16]. Pri tem imajo ravno najnovejši modeli GUI potreben potencial, da revolucionarno posežejo v to fazo znanstvenega raziskovanja, ko gre za biokemijo. Podobne primere bi lahko navedli za področje genomike, astronomije, kvantne fizike, itd. Ne moremo mimo omembe še ene funkcije GUI. Ta funkcija GUI je vezana na njeno zmožnost usmerjanja k bolj interdisciplinarno zasnovanim revolucionarnim znanstvenim odkritjem, saj so njeni potenciali pri obdelavi in sintezi informacij iz različnih disciplin skorajda neomejeni.

3. Vloga GUI se povečuje tudi v procesih evalvacije končnih rezultatov znanstvenega raziskovanja. Če izhajam iz bolj splošnih epistemoloških predpostavk in se na tem mestu izognemo razpravi o prednostih in tudi tveganjih uporabe GUI v konkretnih recenzentskih postopkih, potem naj na kratko omenimo zgolj eno izredno pomembno vlogo te nove napredne tehnologije, t.j. preverjanje rezultatov eksperimentalnih in drugih empiričnih raziskav. V preteklosti je v glavnem veljalo, da ni problematična ponovljivost dobljenih znanstvenih rezultatov, bodisi na temelju javno dostopnih znanstvenih objav ali ustreznih eksperimentalnih protokov. Sodobna znanost se nahaja v vedno večji krizi, kar zadeva zmožnost replikacije, saj je tako z vidika časa kot tudi stroškov v številnih, če ne kar vseh vseh znanstvenih disciplinah težko izvesti potrebne eksperimentalne in druge znanstvene ponovitve. O tveganjih za povečanje goljufij in prevar v moderni znanosti, ki izhajajo iz teh kompleksnih situacij raziskovalnega dela, sem več pisal na drugih mestih [17]. GUI lahko odigra zelo relevatno funkcijo v današnjem času enormne produkcije znanstvenih rezultatov, ko je vedno težje izvajati ponovitve eksperimentov z namenom izvajanja kontrole znanstvenih rezultatov. Njen predikativni pristop namreč lahko zagotovi učinkovito, hitro, sistematično in natančno napoved ponovljivosti posameznih znanstvenih odkritij ali pa celo vseh spoznanj na posameznem področju znanosti.

4. Pozitivna vloga GUI se danes povečuje tudi v okviru širših družbenih in kognitivnih predpostavk, ki so relevantne za delovanje moderne znanosti. V zvezi s to širšo funkcijo bi izpostavil vlogo GUI pri spodbujanju komunikacij znotraj znanstvene skupnosti, pa tudi komunikacije znanstvenikov navzven. To zadnje naj bi se dogajalo predvsem s pomočjo modela ChatGPT, ki generira takšne tipe pojasnitev, ki vodijo k premagovanju komunikacijskih prepadov med eksperti in laiki. Vendar je to funkcijo, kot smo že opozorili v enem izmed predhodnih poglavij, mogoče izvajati le, če bo prišlo do uveljavitve nove paradigme odprtokodne znanosti. V zadnjem času strokovnjaki, ki delujejo na področju GUI, vedno bolj poudarjajo, da je treba razviti modele, ki bodo čim bolj korespondirali z fizično realnostjo. Menijo, da je treba največ naporov usmeriti v nadaljnji razvoj multimodalnih sistemov GUI. Demis Hassabis, izvršni direktor firme DeepMind, je v intervjuju za angleški dnevnik Guardian konec prejšnjega leta dejal, da je bil storjen na tem področju največji korak z modelom Gemini, ki ga razvija njegovo podjetje [18].

5 Ali nova tehnologija generativne umetne inteligence v okviru družboslovnega raziskovanja nujno in vedno zagotavlja znanstveno objektivnost?

Kot družboslovca me seveda zanima tudi vprašanje vedno večje rabe VJM na področju mojega področja znanstvenega raziskovanja. Kar takoj je treba reči, da na področju družbenih ved VJM izkazujejo velik (hevristični) potencial v razvijanju novih pristopov k anketnim raziskavam in ponovljivosti eksperimentov na področju vedenjske ekonomije [19], diskurzivnih analizah tekstov, ki jih je mogoče izvajati na avtomatizirani način [20] in končno tudi na področju razvijanja modelov, ki simulirajo stvarno obnašanje ljudi. V tem zadnjem primeru gre predvsem za t.i. »agent-based« modele, ki preučujejo, kako delovanje oziroma vedenje na mikro ravni (npr.: odločitve individualnih agentov) vodi do posledic na makro (družbeni) ravni (npr.: oblikovanje družbenih vzorcev delovanja oziroma obnašanja). V okviru teh modelov se seveda lahko preučuje tudi obratni vpliv: kako makro-nivo vpliva na obnašanje na mikro ravni [8]. V okviru sociologije se s temi »agent-based« modeli preučuje socialna omrežja, oblikovanje sosedskih skupnosti, itd.

Se pa v zvezi z družboslovnim raziskovanjem pojavlja določen paradoks, na katerega želim opozoriti v tem sklepnem delu moje razprave. Ta paradoks predstavlja dejstvo, da postopki »okrepljenega učenja« (ang. »reinforcement learning«), ki naj bi odpravili »halucinacije« in raznovrstne pristranosti, predstavljajo oviro za doseganje objektivno veljavnih znanstvenih rezultatov. Če pride skozi delovanje t.i. »reinforcement self-learning by human feed-back« (RLHF) do idealiziranja sveta, t..j. sveta, kakršen naj bi bil, ne pa sveta, kakršen dejansko je, takšna prizadevanja za zmanjšanja pristranskosti algoritmov, katerih cilj je promovirati liberalne vrednote, lahko ogrozijo veljavnost raziskav v družboslovju, ki jih podpira umetna inteligenca. »Požarni zid«, ki se ga želi danes pospešeno graditi preko RLHF, odpravlja tveganja GUI, kar zadeva njeno široko uporabo (in preprečuje tveganja, ki so se, kot pravi Yuval Harare, že zažrla v civilizacijski kod sodobnih družb), po drugi strani pa predstavlja epistemološko tveganje za objektivni značaj današnjih družboslovnih raziskav. Tudi to predstavlja dilemo današnjega in prihodnjega razvoja umetne inteligence, ki zahteva naš celovit interdisciplinarni razmislek, saj se je le na tej osnovi mogoče izogniti negativnim družbenim in tudi epistemološkim implikacijam njenega razvoja.

6 Zaključek

V zadnjem času je tako v znanstvenih krogih kot tudi zunaj znanstvenih krogov veliko govora o možnih tveganjih današnjega in prihodnjega razvoja umetne inteligence. Znanstveniki iz Massachusetts Institute of Technology, ene najbolj uglednih akademskih institucij v ZDA, so v letošnjem letu pripravili javno dostopni repozitorij z umetno inteligenco povezanih primerov tveganj. V omenjenem repozitoriju se trenutno nahaja kar 777 opisov takšnih tveganj. To je še en dokaz, kako veliko zanimanje obstaja danes za ta vprašanja. V mojem kratkem prispevku sem se dotaknil zgolj enega izmed teh številnih problemov, ki je vezan bolj na epistemologijo znanstvenega raziskovanja, ne pa toliko na družbene posledice razvoja umetne inteligence. V tem kontekstu me je predvsem zanimalo, ali pospešena raba VJM v okviru različnih znanstveno-raziskovalnih aktivnosti predstavlja eno izmed domen, kjer se na stežaj odpirajo vrata nastopu umetne splošne oziroma umetne super inteligence. Še posebej me je zanimalo vprašanje, zakaj GUI postaja že danes nepogrešljivo »orodje« v seh fazah znanstvenega raziskovanja. V sklepnem delu sem se na kratko ustavil ob nekaterih specifičnih dilemah uporabe GUI na področju družboslovnega raziskovanja.

Literatura

[1] Ray Parta, 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 3: 21–154; https://doi.org/10.1016/j.iotcps.2023.04.003.

[2] Nick Bostrom, 2014. Superintelligence: Paths, dangers, strategies. Oxford University Press.

[3] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, & Owain Evans, 2022. When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62(July), 29-754. https://doi.org/10.1613/jair.1.11222.

[4] Max Roser, 2023. AI timelines: What do experts in artificial intelligence expect for the future? Our World in Data. https://ourworldindata.org/ai-timelines.

[5] Rishi Bommasani et al., 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258v3 [cs.LG] 12 Jul 2022.

[6] Yogesh Dwivedi et al., 2023. Opinion Paper - So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71 (2023) 102642; https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[7] Thomas Arnold, 2024. Herausforderungen in der Forschung: Mangelnde Reproduzierbarkeit und Erklärbarkeit. V: L. Ohly in G. Schreiber (Hrsg.) KI:Text.Diskurse über KI-Textgeneratoren, str. 67-83. Berlin/Boston: De Gruyter Verlag.

[8] Christopher Bail, 2024. Can Generative AI improve social science?. PNAS, May 9, 2024 1211) e2314021121; https://doi.org/10.1073/pnas.2314021121.

[9] Yuval Noah Harari, 2024. Nexus. A Brief History of Information Networks from the Stone Stage to AI. New York: Penguin Random House.

[10] Bubeck S. et al., 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4 (2023), arXiv:2108.07258v3 [cs.LG] .

[11] L.B. Soros, Alyssa Adams, Stefano Kalonaris, Olaf Witkowski, Christian Guckelsberger, 2024. On Creativity and Open-Endedness. arXiv:2405.18016v4 [cs.AI] 23 Jun 2024.

[12] Marc Runco, 2023. AI can only produce artificial creativity. Journal of Creativity 33 (2023) 100063. https://doi.org/10.1016/j.yjoc.2023.100063.

[13] Stephen Rice, Winter Scott, Rice Connor, 2024. The advantages and limitations of using ChatGPT to enhance technological research. Technology in Society 76 (2024) 102426. https://dx.doi.org/10.2139/ssrn.4416080.

[14] Hubert Kent, Kim Awa, Darya Zabelina, 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. Scientific Reports 14(1):3440; https://doi.org/10.1038/s41598-024-53303-w.

[15] Mohamed Salah et al., 2023. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. Current Psychology, 43, 7843–7858 (2024). https://doi.org/10.1007/s12144-023-04989-0.

[16] Hanchen Wang et al., 2023. Scientific discovery in the age of artificial intelligence. Nature, 620: Vol 620 | 3 August 2023. DOI: 10.1038/s41586-023-06221-2.

[17] Franc Mali, 2011. Razvoj moderne znanosti. Socialni mehanizmi. Ljubljana: Založba FDV.

[18] Demis Hassabis, 2023. Google releases new AI model with claim it can outperform ChatGPT in most tests. The Guardian, 7. December, 2023.

[19] John Horton, 2023. Large language models as simulated economic agents: What can we learn from homo silicus? arXiv [Preprint] (2023). 10.48550/arXiv.2301.07543.

[20] Igor Grossmann, Cassandra Parker, Mathew Feinberg, Nicholas Christakis, 2023. AI and the transformation of social science research. Science, 380(6650):1108-1109. DOI:10.1126/science.adi1778.Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

Comparing academic performance across course topics: a pilot study

Laura Fink
Faculty of entrepreneurship
GEA College
Ljubljana, Slovenia
laura.fink@gea-college.si

Bojan Cestnik
Faculty of entrepreneurship
GEA College
Ljubljana, Slovenia
bojan.cestnik@gea-college.si
bojan.cestnik@temida.si

ABSTRACT

In this paper, we examine the academic performance of students in different courses to determine whether good performance in one course is related to good performance in other courses. Although certain predictive models emphasize the importance of course content for learning success, there are few studies that address how student performance in different courses is related to similar course topics, learning goals, competences, and skills. By creating a preliminary framework that examines how academic performance is related to different course topics, we attempt to make a first step further towards addressing the research gap regarding the interrelatedness of student achievement not only in different course topics but in different competence areas. We examined a set of student grades from eleven different courses at the faculty from areas such as entrepreneurship, management and leadership, business informatics, mathematics, economics, marketing and market analysis, innovation and creativity, English, finance and accounting, business law, and human resource management. We show that students with more exam retakes on average reached a lower grade rank than the students who only registered for the exam once. We used linear regression to show the significance of the relationships between student performance in the Informatics course compared to their achievement in other courses. With a correlation matrix coefficient, we measured the strength of reciprocal interrelatedness between the grade ranks students attained in each of the eleven courses. The results of this preliminary study indicate a possible stronger association between academic achievement in courses that have similarities in terms of content or focus, such as business administration and entrepreneurship (correlation coefficient of 0.58). Further studies with detailed comparison of course-specific competences are needed for accepting the finding that interrelatedness between achievements in courses from similar versus different disciplines is stronger. The preliminary model could further be improved by a broader range of courses, input explanatory student factors, and application of advanced analytical techniques.

KEYWORDS

course-specific competence, learning analytics, prediction models, student academic achievement, student academic performance

1 INTRODUCTION AND RELATED WORK

Are students who achieve excellent results in one course more likely to achieve outstanding results in another? And vice versa? This is the key question that triggered this preliminary research into how students perform in different courses. As a result, we are examining the relationships between students' academic achievement in different courses from two different undergraduate study programs at the faculty of entrepreneurship. Despite some prediction models suggesting the course content as one of the input explanatory variables, there is a significant lack of detailed research on the relationship between students' achievement in different courses from the same, similar, or entirely different discipline. Therefore, this preliminary study aims to develop a pilot model to investigate the relationship between academic success in various courses and their respective course topics. The findings of this study could be further reinforced and interpreted by the comparison of course-specific competences and learning objectives.

Apart from the prediction model of academic achievement, interrelatedness between academic achievement in different courses compared to the course main topics, competences, and learning objectives is still largely missing. Since previous studies that would previously investigate the interrelatedness of student achievement in different courses are, to the best of our knowledge, entirely nonexistent, let us draw attention to the research studies that are related to the field and that have led to this investigation. OECD [12, 13], for example, show a positive association between literacy and numeracy skills. Moreover, in our previous research [2], we showed that students who achieved better academic achievement in word skills were on average also more likely to achieve better achievement in excel skills. However, this does not imply that it is sufficient to develop only some of these skills, such as solely including word skills, excel skills, literacy skills, or just numeracy skills in the curriculum. Furthermore, this does not imply that students in a real situation cannot achieve much better results in a certain type of skill compared to another. Additionally, Fink and Vadnjal [3] conducted a pilot study that compared the development of

* Comparing academic performance across course topics: a pilot study.
† Fink, L., Cestnik, B.

Information Society 2023, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.10

generic and course-specific competences during a higher education course.

As previously mentioned, one could apply findings about the relationships between various course topics, contents, competences, or objectives to the development of predictive models and predictive algorithms. Prediction models are often used to find out ahead of time which students are likely to drop out or fail external exams. The schools aim to take intervention measures and steps to stop the bad predictions from coming true, and the success rate can be raised [16, 9]. There are a lot of different input explanatory variables that can have a big effect on how well and especially how accurately prediction models perform [16, 4, 7, 8, 9, 11, 1]. Some prediction models include related concepts to the course content [7], such as the course's learning objective, the course's main competences, the course topic [4], or even course preparedness [10].

In addition to the combination of input explanatory variables, various statistical methodologies and techniques, as well as different types of measurements and academic achievements, significantly influence the prediction model's power and performance. While adding additional or all relevant factors to the model does not always improve its performance, the right combination of input explanatory variables significantly influences the accuracy and other model performance measures [4]. In the end, the right combination of input variables largely determines the model's explanatory power, the accuracy of its predictions, and other performance measures [6, 15].

Different prediction models are based on different methodologies and include different input and output variables. Francis and Babu [4], for example, compare prediction models of student achievement that include the topic of the course as the input explanatory variable. They demonstrated that the course topic, along with many other factors, is one of the explanatory variables for academic achievement. However, their model found that academic factors, including the course topic, were less accurate in predicting students' academic achievement than demographic factors, behavior factors, and other factors such as absence days, parental satisfaction, and school survey responses. They developed, compared, and assessed performance measures of several prediction models. The model that included academic factors, behavior, and additional input explanatory variables showed the greatest improvement in accuracy. On the other hand, adding demographic factors on top of that reduced the accuracy of academic achievement prediction. Clearly, the addition of additional input variables, for example, the topic of the course, in different models contributes differently to improving prediction accuracy and other performance measures, depending on other input variables in the model.

2 HYPOTHESES

In this preliminary study, we aim to build a preliminary pilot research model on which we will test the interrelatedness between academic achievement in different courses. We suggest the following hypotheses:

H1. Mostly there are reciprocal relationships between a student's performance in one course and their performance in another.

H2. As the number of exam retakes increases, the student's grade rank decreases.

3 DATASET AND METHODOLOGY

The dataset collection and preparation included several phases. First, we have collected students' grades for different courses at the higher education institution. The initial dataset included the grades of 223 students for 67 different courses.

In the second phase, we have refined and further prepared the dataset. Based on some simple data exploration and visualization techniques, such as plotting the missing values, plotting the distribution of the number of grades available per course, and plotting the distribution of the number of exam retakes per course, we have decided to eliminate the data of courses with less than 60 students' grades per course. With that, we narrowed further analysis to the following eleven selected courses: Business Economics, Informatics, Management and Leadership, Marketing and Market Analysis, Entrepreneurship, Business English, Accounting, Creativity and Innovation, Business Mathematics and Statistics, Business Law, and Human Resource Management.

Table 1: Number of grades per one and two courses

Course (Short name)	Nr. of students/grades	
	per course	for chosen course and Informatics (%)
Informatics for entrepreneurs (P07_IE)	160	
Business economics (P05_BE)	139	120 (86 %)
Business law for entrepreneurs (P06_BLE)	79	71 (90 %)
Human resource management (P09_HRM)	66	58 (88%)
Management and leadership (P15_ML)	156	142 (91%)
Marketing and market analysis (P16_MMA)	90	73 (81%)
Entrepreneurship (P27_ENT)	191	148 (77%)
Business English (P28_BE)	152	133 (88%)
Accounting for entrepreneurs (P33_AE)	143	131 (92%)
Creativity and innovation in entrepreneurship (P36_CIE)	173	142 (82%)
Business mathematics and statistics (P39_BMS)	139	119 (86%)

We then compared the number of students' grades available per one course with the number of grades available per two courses (the selected course and the informatics) and calculated the share of students that also took the exam in informatics

compared to the number of students who took the exam in the selected course only. Students' grades include both those that indicate a student has passed the course and those that indicate a student has not. Since we included only one grade per student in further analysis, the number of students' grades reflects the number of students who took the exam in each course. Not all students took exams in all eleven subjects. The reasons for this are varied, including the fact that the courses are from two different programs.

The courses with the highest student grades include Entrepreneurship (191), followed by Creativity/Innovation (173) and Informatics (160) (Table 1). More than 77% of students (148) who took the entrepreneurship exam also took the informatics exam. Similarly, 82% of students who took the exam in creativity/innovation (142), and 91% of those who took the exam in management/leadership (142) also took the exam in informatics. In other words, students who attended the exam in the Informatics course often also attended the exam in the Entrepreneurship, Management/Leadership, and Creativity/Innovation courses, as shown in Table 1.

Then, we continued by calculating the average grade ranks and number of exam retakes (table 2) for each of the courses included in the analysis. The grade ranks range from 0 to 10, where 0 represents not attending, 1 to 5 represents failed, 6 satisfactory, 7 average, 8 good, 9 very good, and 10 excellent. As shown in Table 2, the students on average achieved better grades in the HRM course than in other courses. In comparison, the students on average achieved the lowest average grade rank in the Mathematics/Statistics and Economics course compared to other courses. The students also, on average, most commonly retook the exam in these two courses. Additionally, we found that 38 students attended at least one exam deadline for each of the eleven courses.

Table 2: Average grade rank and number of exam retakes

Course (Short name)	Avg. grade	Avg. nr. of retakes
Informatics for entrepreneurs (P07_IE)	7.7	1.2
Business economics (P05_BE)	6.6	1.8
Business law for entrepreneurs (P06_BLE)	7.1	1.3
Human resource management (P09_HRM)	8.0	1.1
Management and leadership (P15_ML)	7.3	1.2
Marketing and market analysis (P16_MMA)	7.5	1.3
Entrepreneurship (P27_ENT)	7.4	1.3
Business English (P28_BE)	7.3	1.2
Accounting for entrepreneurs (P33_AE)	7.0	1.3
Creativity and innovation in entrepreneurship (P36_CIE)	7.2	1.3
Business mathematics and statistics (P39_BMS)	6.4	2.0

In the third phase, we continued with the data exploration and visualization. We plotted the distribution of the number of grades achieved per grade rank for each of the selected eleven courses. An example of the distribution for the Entrepreneurship course is provided in Figure 1.

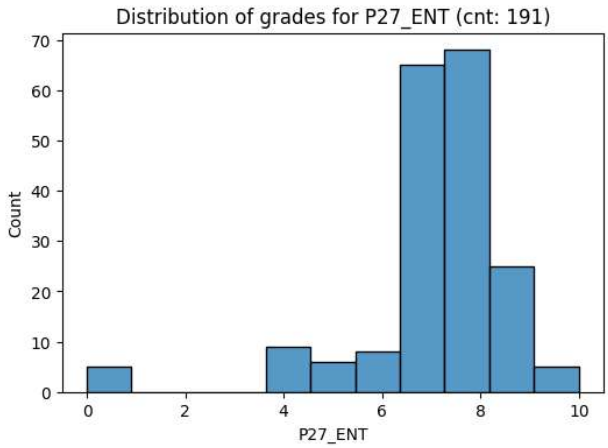


Figure 1: Distribution of number of grades per each grade rank for Entrepreneurship course example

In the next phase, we focused our investigation on how the grades that students achieved in the Informatics course behave compared to the grades they achieved in the remaining ten courses. We performed and visualized ten linear regression models describing the relationships between grade ranks students achieved in the informatics course and the grade ranks students achieved in other selected courses. When performing the linear regression, we included the grade rank achieved in informatics as an independent variable. Though we are aware that linear regression models assume the influence of the independent variable on the dependent variable and not the reciprocal relationships per se, we decided to mention this limitation and work further with the results obtained from the regression analyses in this preliminary pilot study.

We performed additional analysis based on the correlation matrix between the grade ranks students achieved in each of the eleven selected courses. We draw a correlation matrix with significant ($p < 0.05$) correlations among regression coefficients between the eleven selected courses to determine which of the eleven courses is related to another one. We then examined the strength and significance of the reciprocal relationship that the correlation matrix coefficient measures.

Next, we further compared the characteristics of our data for the selected eleven courses with the characteristics of the data for all the courses. We used data visualization techniques such as plotting to compare the distribution of the average grade of eleven selected courses and all the courses. The distribution of average grade for the selected eleven courses seems fairly similar to the distribution of average grade for all the courses, as shown in Figure 2.

Although we cannot claim that an analysis of the entire data set would yield similar results to the analysis of the selected eleven courses based solely on the similar distribution of average

grades, we cannot completely rule out the possibility that on average, somewhat similar associations would emerge among the other courses.

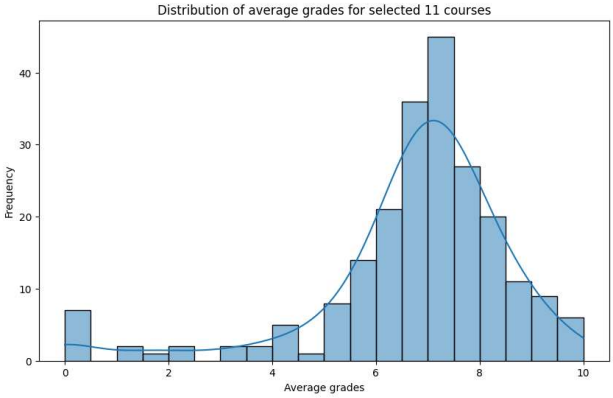


Figure 2: Distribution of number of grades per each grade rank per course example

In the final phase of this preliminary pilot research, we performed the regression analysis between the grade rank that the student achieved and the number of exam retakes of the student.

4 RESEARCH FINDINGS

The eleven selected courses served as the basis for the analysis, which focused on identifying reciprocal relationships between the grade ranks of one course and those of another course. We selected the courses based on the number of grades available after exploring, visualizing, cleaning, and refining the data. In addition to linear regression models, we calculated the correlation matrix's correlations (Figure 3) that investigate the reciprocal association between the grade ranks students achieved in one course compared to another.

Overall, we investigated 55 reciprocal relationships (H1) between the grade ranks of one course with the grade ranks of another course. Among these, forty-six correlations are significant ($p < 0.05$) compared to nine correlations that are not significant. Among the significant correlations, nine correlation coefficients exhibit a moderate relationship (between 0.5 and 0.7) between the grade ranks of one course and the grade ranks of another course. Twenty-seven correlations show a weak correlation (between 0.3 and 0.5), while ten correlations show a negligible or low correlation (below 0.3) between the grade ranks of two selected courses.

The strength of significant coefficients varies from 0.19 all the way up to 0.58. Many of the coefficients that we would otherwise have placed in the group of weak correlations are very close to 0.5, which indicates moderate correlation. There are also many coefficients that we have otherwise placed in the group of negligible correlations close to the value of 0.3, which indicates a weak correlation.

Based on these results, we can accept hypothesis 1 that foresees the existence of reciprocal relationships between a student's performance in one course and their performance in another in most cases when comparing different pairs of courses.

Let's examine the concrete correlations between the grades of the two courses. There is a moderate correlation between Economics and Mathematics/Statistics (0.60), followed by Economics and Entrepreneurship (0.58), Informatics and Mathematics/Statistics (0.57), Marketing/Market Analysis and Law (0.57), Marketing/Market Analysis and Entrepreneurship (0.55), Economics and Accounting (0.54), Marketing/Market Analysis and Creativity/Innovation (0.52), and Economics and Informatics (0.51), and Informatics and Marketing/Market analysis (0.50).

Based on moderate correlations, we can speculate that Economics, Informatics, Mathematics, and Accounting courses are closely connected, partially because students have to use the numeracy skills in these courses. Therefore, future research could address the question of whether the syllabus of these courses reflects shared similar competences. Although rare, previous research [12, 13] on the interrelatedness of competences has shown that people who are more proficient in literacy skills are usually also more proficient in numeracy skills, and vice versa, additional inquiry into the similar and different competences that are developed within these courses would provide an important insight and more thoroughly address the gap in the literature on the interrelatedness between competences that is largely still missing. Therefore, we also need to investigate further to what degree are the numeracy skills included in the syllabus of the courses that otherwise aim at developing soft, social, and other professional skills, such as the Entrepreneurship and Marketing/Market Analysis courses. Additionally, based on the analysis, it would also make sense to check whether the Marketing/Market Analysis and Creativity/Innovation courses foster the development of related skills. In general, we can speculate that the type of competences is important for student achievement, but to confirm this, we would have to perform the qualitative analysis of the similar competences and learning objectives in the future.

The majority of the relationships are statistically significant but weak. Weak correlation exists between Entrepreneurship and Mathematics/Statistics (0.49), Entrepreneurship and Informatics (0.47), Economics and Management/Leadership (0.47), Entrepreneurship and Law (0.45), English and Entrepreneurship (0.45), Economics and English (0.43), Creativity/Innovation and Entrepreneurship (0.42), HRM and Law (0.41), Accounting and HRM (0.41), Informatics and Accounting (0.41), Management/Leadership and Accounting (0.40), English and Mathematics/Statistics (0.38), Mathematics/Statistics and Law (0.38), Marketing/Market Analysis and Law (0.37), English and Accounting (0.37), Management/Leadership and English (0.36), Entrepreneurship and HRM (0.36), Law and Economics (0.34), Management/Leadership and Entrepreneurship (0.34), Entrepreneurship and Accounting (0.33), Accounting and Mathematics/Statistics (0.33), Management/Leadership and Marketing/Market Analysis (0.31), Marketing/Market Analysis and Economics (0.31), Management/Leadership and Law (0.31), Informatics and Law (0.30), English and HRM (0.30), Creativity/Innovation and Accounting (0.30).

A low correlation below 0.3 exists between Accounting and Law (0.29), Marketing/Market Analysis and English (0.29), Marketing/Market Analysis and Accounting (0.28), Economics

and Creativity/Innovation (0.27), Informatics and Management/Leadership (0.26), Management/Leadership and Creativity/Innovation (0.25), Management/Leadership and Mathematics/Statistics (0.25), English and Informatics (0.22), English and Creativity/Innovation (0.20), and Creativity/Innovation and Mathematics/Statistics (0.19).

Based on the weak and low correlations, we can speculate that these correlations exist in courses that do not necessarily share that much of similar competences and learning objectives as those that exhibit moderate correlations. As mentioned previously, further qualitative research is required to substantiate these assumptions.

Since not all students took the exams in all the courses, the number of observed instances for each pair of two courses ranges from 46 to 166, depending on the particular pair of courses. Not only that, we found that the low number of observed instances importantly contributed to the statistical insignificance of some correlation coefficients. The relationships calculated based on a low number of observed instances, such as, for example, between Mathematics/Statistics and Law (46 observed instances), Marketing/Market Analysis and Mathematics/Statistics (46), HRM and Economics (47), Informatics and HRM (58), English and Law (58), Management/Leadership and HRM (59), and Creativity/Innovation and HRM (59 observed instances), exhibit insignificant relationships. The exemptions are the insignificant relationship between the Informatics and Creativity/Innovation course with 142 observed instances and between the Creativity/Innovation and Law course with 74 observed instances. We therefore speculate that the Informatics and Law courses strive to develop different competences than the Creativity/Innovation course. This preliminary analysis is a useful basis for further research and analysis.

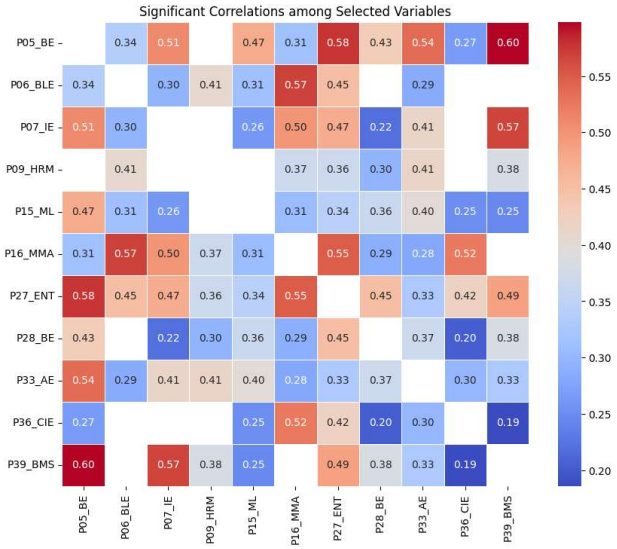


Figure 3: Correlation matrix with significant correlations

The results of the regression analysis between the grade rank that the student achieved and the number of exam retakes that the student took (H2), that are shown in Figure 4, suggest that the students with more exam retakes on average reached a lower grade rank than the students who only registered for the exam once (β = -0.47, p = 0.00). The full equation is displayed below (equation 1). Based on these results, we accept hypothesis 2 that as the number of exam retakes increases, the student's grade rank decreases.

$$Grade\ rank = 8.25 - 0.47 * Nr.of\ exam\ retakes + residuals \quad (1)$$

Figure 4 also shows that most students register for the exam once, fewer students register for the exam twice, and even fewer students register for the exam a third time.

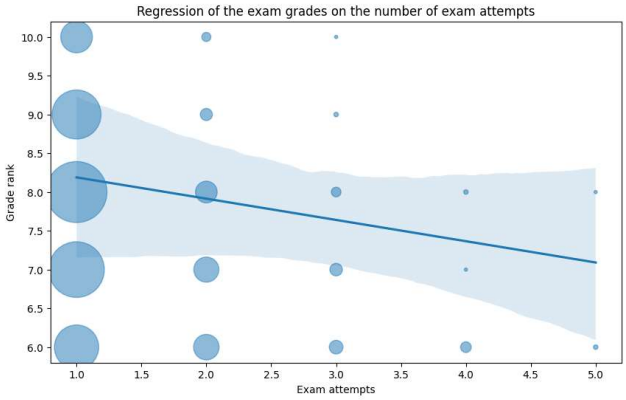


Figure 4: Regression analysis between number of exam retakes and average grade achieved

5 CONCLUSION

The purpose of this study was to lay the groundwork for further research regarding the correlation between different course-specific competences and to present initial findings regarding the correlation between students' academic achievement in different courses. In this paper, we aim to enhance our comprehension of the intricate relationship between competences, an area that remains largely unexplored. The preliminary analysis revealed the existence of interrelatedness among grades students achieve in different courses, and showed that a student's academic performance in one course influences their performance in another. Based on the analysis we accept hypothesis 1 that foresees the existence of reciprocal relationships between a student's performance in one course and their performance in another in most cases when comparing different pairs of courses. We also show that students with more exam retakes on average reached a lower grade rank than the students who only registered for the exam once. With that, we accept hypothesis 2 that as the number of exam retakes increases, the student's grade rank decreases.

To determine whether there is a stronger correlation between academic achievements in courses from the same or similar

discipline than in courses from completely different disciplines, further research is required to explore how much the interrelatedness between courses depends on the competences, learning goals, and discipline of the course.

Since this is a preliminary pilot analysis, we considered additional opportunities to improve our research in the future. We could enhance this study by utilizing additional methods, such as cluster analysis, network analysis, and structural equation modeling, along with techniques used to make predictions like data mining [11, 1], neural networks [5], or decision trees [14]. Furthermore, we could enrich the model by increasing the number of observations and the number of courses included in the analysis.

To capture more subtleties in these relationships, we could potentially build and test the model's performance with additional input variables such as information about general knowledge, broad outlook, ambitions, and psychological characteristics. Additionally, we could compare different cohorts of students, the semester, the study program and track, forms of study (full-time, part-time), types of study programs (undergraduate, postgraduate, higher vocational program), and study modes (classroom, blended, online).

REFERENCES

[1] Balqis Albreiki, Nazar Zaki and Hany Alashwal.2021. A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552.

[2] Laura Fink and Bojan Cestnik. 2023. Reinforcing digital skills: a quantitative analysis of digital skills development and its impact on higher educational outcomes. In Szymanska, K. (ed.), Janczewski, R. A. (ed.). *Problemy i wyzwania ekonomii i zarządzania w XXI wieku : wybrane aspekty*. Poznań: Grupa Wydawnicza FNCE. P. 263-274.

[3] Laura Fink and Jaka Vadnjal. 2024. Generic and course-specific competences in comparison. In Fošner, A. (ed.). *ABSRC Ljubljana: conference proceedings,* GEA College - Faculty of Entrepreneurship, p. 10.

[4] Bindhia K. Francis and Suvanam Sasidhar Babu. 2019. Predicting academic performance of students using a hybrid data mining approach. *Journal of medical systems*, 43(6), 162.

[5] Leon Gerritsen. 2017. *Predicting student performance with Neural Networks.* Master Thesis. Tilburg University, Netherlands.

[6] Ramin Ghorbani and Rouzbeh Ghousi. 2020. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE access, 8*, 67899-67911.

[7] Judith M. Harackiewicz, Kenneth E. Barron, John M. Tauer and Andrew J. Elliot. 2002. Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of educational psychology, 94*(3), 562.

[8] Shaobo Huang and Ning Fang 2013. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, 61, 133-145. https://www.sciencedirect.com/science/article/abs/pii/S0360131512002102

[9] Kitsadaporn Jantakun, Thiti Jantakun, and Thada Jantakoon. 2022. The architecture of system for predicting student performance based on data science approaches (SPPS-DSA architecture). *International Journal of Information and Education Technology, 12*(8), 778-785.

[10] René F. Kizilcec, and Hansol Lee. 2022. Algorithmic fairness in education. In Holmes, W., Porayska-Pomsta, K. (Eds.), *The ethics of artificial intelligence in education*, Routledge. p. 174-202.

[11] Abdallah Namoun and Abdullah Alshanqiti. 2020. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237.

[12] OECD. 2019A. *Skills Matter: Additional Results from the Survey of Adult Skills,* OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/1f029d8f-en.

[13] OECD. 2019B. *The Survey of Adult Skills: Reader's Companion,* Third Edition, OECD Skills Studies, OECD Publishing, Paris, https://doi.org/10.1787/f70238c7-en

[14] Harikumar Pallathadka, Alex Wenda, Edwin Ramirez- Asís, Maximiliano Asís-López, Judith Flores-Albornoz and Khongdet Phasinam. 2023. Classification and prediction of student performance data using various machine learning algorithms. *Materials today: proceedings,* 80, 3782-3785.

[15] David M. W. Powers. 2011. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies, 2*(1): 37–63.

[16] Mustafa Yağcı. 2022. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11.

Linking the Normative and the Descriptive: Bounded Epistemic Rationality

Nastja Tomat†
Department of Philosophy
Faculty of Arts
University of Ljubljana
Ljubljana, Slovenia
nastja.tomat@ff.uni-lj.si

Abstract

Epistemic rationality is a type of rationality directed towards cognitive or epistemic goals, such as true beliefs, knowledge, or understanding. Epistemology is primarily concerned with normative questions about how one should form and update beliefs, reason and inquire to be rational; on the other hand, empirical disciplines, such as psychology, investigate how inquiries and belief formation occur in real life. The question arises as to what the relationship between the normative and the descriptive in the study of epistemic rationality should be. This paper proposes a notion of bounded epistemic rationality as a hybrid, non-ideal concept that encompasses both normative and descriptive elements. Drawing upon Herbert Simon's bounded rationality and Robin McKenna's non-ideal epistemology, bounded epistemic rationality is characterized by requiring satisficing instead of maximizing; acknowledging our cognitive, environmental, and practical limitations; its ecological nature; and its focus on the process of inquiry. As such, bounded epistemic rationality is a good starting point for proposing epistemic advice that is achievable for real cognizers and helps them improve their epistemic position.

1 Introduction

Epistemic rationality is one of the main topics of epistemology. It refers to epistemic attitudes, states, and processes [1], mainly focusing on rationality of beliefs, and is directed towards reaching cognitive or epistemic goals, such as true beliefs, knowledge, or understanding [2, 3, 4]. One of the central tasks of epistemology has been to propose epistemic norms about how one should form, update and revise beliefs to be rational. Although it is acknowledged that humans are not ideal agents – there is ample empirical evidence, gathered by disciplines as cognitive psychology, showing that we are limited by our cognitive architecture and the nature of cognitive processes, such as computational power and speed, predictive abilities, working memory and attention [5, 6, 7] – traditional analytic epistemology still often relies on idealized models of human cognizers [8], with the consequence that it frequently imposes epistemic norms such as logical omniscience, consistency between beliefs, and immediate updating of beliefs by conditionalization [9].

Philosophy, including epistemology, is predominantly concerned with the normative questions about justification, rationality and other epistemic appraisals of our cognitive activities and doxastic states, while empirical disciplines, such as psychology, empirically investigate how human cognition, inquiries and belief formation occur in everyday life. With normative theories on the one hand and empirical research on the other, we are faced with the question of the relationship between the two approaches towards studying rationality.

The aim of the paper is to propose a concept of bounded epistemic rationality as a hybrid notion that may help us bridge the gap between the normative and the descriptive. By adopting a concept that is – to some extent – grounded on empirical data about human cognition but does not dispose with the normative questions about epistemically good cognition, we can propose epistemic norms and epistemic advice that are achievable for real human cognizers and can help them improve their epistemic situation.

2 Normative and descriptive theories of rationality

Philosophical understanding of rationality is deeply intertwined with the notion of normativity. There are different views on how to define and justify epistemic normativity and which epistemic norms we should endorse. We can understand rationality as a system of rules or requirements: it requires from us, for example, not to hold contradictory beliefs, to draw a conclusion by modus ponens [10], to have deductive closure [11], or to follow rules of logic, probability and decision theory [12]. Rationality is thus normative in the sense of employing certain norms and rules according to which we can judge correctness of a belief [13]. Nevertheless, a genuine normative question of rationality requires us to determine if those rules or requirements are necessarily accompanied by a reason to conform to them, or, in other words, if we *ought* to conform to them [13, 14].

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.11

In addition to the debate about genuine normativity of rationality, there is an ongoing discussion about how epistemic norms or principles should be formulated and what they should prescribe. According to Engel, there are some conditions such principles should satisfy to be genuinely normative: they should have normative force, a potential to regulate or direct our inquiry and beliefs, and normative freedom – a possibility to be violated. If we accept these conditions, many normative principles that are often employed are not adequate. A rule that says, for example, that one should not believe p and not p, tells us something about a characteristic of a rational belief, but gives us no guidance on how to achieve it [11]. Such rules are more of a description of a belief or believer in ideal conditions than genuine normative principles. A similar point is put forward by Robin McKenna in his book Non-ideal epistemology [8]. He claims that mainstream epistemology mostly proposes epistemic norms based on various types of idealizations, for example about cognitive capacities of epistemic agents and the nature of epistemic environment. He calls such an approach to epistemological theorizing ideal epistemology and contrasts it with non-ideal epistemology which tries to avoid such idealizations. The issue with the norms proposed by ideal epistemology is that they are too detached from real world issues, too demanding and unachievable for real human cognizers. Another, even more important issue is that they provide bad epistemic advice: if we try to achieve or approximate proposed ideals and norms, we will often worsen rather than improve our epistemic situation. McKenna uses an example of the ideal of objectivity in scientific inquiry: trying to achieve objectivity as detachment – in a sense that scientists are not personally invested and interested in the topic of inquiry and try to detach research process from non-cognitive values – leads to worse, not better, scientific inquiry. McKenna draws on Elizabeth Anderson's work on value judgements in science. She argues that researchers' background assumptions and values influence all the stages of research process – they partly determine how we frame the research questions, conceive of the object of inquiry, what data we collect, how we analyze and interpret them. A large portion of empirical research in social science investigates evaluative questions that are related to well-being of individuals, social groups or society at large, and science that is legitimately guided by certain, for example feminist values, could be more fruitful and more likely lead to desired epistemic goals. Instead of trying to be attain an ideal of objectivity, scientific inquiry should be informed by the right values [15]. A similar argument can be made for our everyday inquiries: if we, for example, always aim to reason in accordance with a norm proposed by ideal theory, such as logic and probability theory, or trying to think in intellectually autonomous way instead of relying on experts, this will likely lead to worse epistemic outcomes than using less complex, heuristic processes or form a belief according to the consensus of the experts [16]. This means that ideal theory is failing as a normative theory because its prescriptions often do not help us achieve our epistemic goals, such as obtaining true beliefs, knowledge or understanding, and cannot serve as regulative ideals. For this reason, the ideal approach should in certain situations be replaced by a non-ideal one [8]. Both Engel and McKenna emphasize that an important feature of epistemic norms is their potential for guidance, for improving our inquiries, reasoning, and forming beliefs that are in some way epistemically better. Instead on focusing on defining conditions for epistemic ideals such as justification and knowledge and engaging in "S knows P iff …" kind of epistemology, McKenna claims that we should engage in non-ideal theorizing that is informed by empirical literature on human cognition, knowledge-producing institutions and epistemic environment. While McKenna claims that descriptive questions should be a starting point for answering normative questions, he does not argue for a strong form of naturalism or for the replacement of epistemology by empirical science, but merely suggests that there should be a closer connection between epistemology and empirical disciplines than is currently the case [8]. The norms that non-ideal epistemology proposes would therefore be norms of inquiry that help agents determine which problems are important to inquire about in the first place; how to collect, assess and evaluate evidence; what to do when they are presented with conflicting information; how to identify trustworthy and reliable sources of information; and when they gathered enough evidence to terminate an inquiry and form a belief. Instead of a norm stating something in terms of "a belief about whether anthropogenic climate change is real is rational if it is achieved in a reliable manner and responsive to the available evidence", non-ideal epistemology would propose norms specifying what is an epistemically good manner in which ordinary laypeople should gather evidence about climate change, how to identify genuine experts and how to recognize good evidence. Such norms would require inquiring in a way that is possible for ordinary people – would not, for example, require enormous amount of time and philosophical understanding of the concept of evidence – and would be based on empirical data on which ways are effective for gaining true beliefs about climate change – for example, relying on science marketing strategies [8]. Such norms or principles would satisfy Engel's conditions and could thus be considered as genuinely normative.

3 Bounded rationality

Although many authors who investigate rationality or epistemic norms explicitly acknowledge that humans are limited agents and that our boundaries should put a constraint on epistemic norms, only a few philosophers have drawn on the notion of bounded rationality. Bounded rationality was introduced by political scientist Herbert A. Simon and has importantly influenced many disciplines investigating rationality, such as psychology and economics. Simon argued that global, idealized theories of rationality should be replaced with a notion of rationality that is compatible with cognitive capacities of the subjects and the features of the environment in which they are embedded. As our cognitive capacities, for example predictive and computational capabilities, working memory and attention, are limited, human rationality can be only an approximation of an ideal rationality that is assumed in models of decision theory. If we want to comprehend human rationality, we should not focus only on internal characteristics of human cognition, but also on the structure of the environment. Simon illustrated this with a metaphor of scissors: "Human rational behaviour (and the rational behaviour of all physical symbol systems) is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor." [17, p.7]. Simon argued that human rationality was satisficing, not optimizing – meaning that humans do not seek for best

possible solutions of a problem or best possible outcome of a decision situation, but for solutions that are merely good enough – and he urged to dispose of the notion of optimization as a criterion for rationality. He also emphasized that bounded rationality is procedural, meaning that it does not focus solely on the outcomes, but also on the process leading to them; an agent is therefore rational if her behavior stems from an appropriate process of deliberation [17-21].

4 Bounded epistemic rationality

According to Sturm [22], philosophical aspects of bounded rationality have not yet been systematically investigated; nevertheless, the role of bounded rationality in epistemology has recently been explored by David Thorstad [7]. He describes five characteristics of bounded rationality as a paradigm, the first one being that bounds are important. As opposed to practical philosophy where it is universally acknowledged that our physical limitations put constraints on the norms of rational action, this is not necessarily the case for epistemic rationality. Thorstad claims that bounds are equally important for our understanding of rational cognition than of rational action and that we should be normatively required to perform only those cognitive operations that we are capable of. Secondly, theories of rationality should consider not only the final beliefs and other doxastic states, but also the processes that led to them, which is directly derived from Simon's notion on procedural nature of bounded rationality. The third and fourth characteristics refer to the claim that rationality is not bound only by our cognition, but also by environmental factors and that the use of rules of thumb or heuristics can be more rational than using more complex reasoning strategies. Drawing on the work of Gerd Gigerenzer and ecological rationality [6, 16], Thorstad claims that heuristics may in many situations or environments provide more accurate predictions than other, more sophisticated strategies. Finally, bounded rationality is compatible with a so-called programme of vindicatory epistemology, which states that what we usually consider as a violation of rationality norms is a consequence of a deliberation process that is merely boundedly rational. Although we do not comply with traditional epistemic norms as coherence and deductive closure, we are often inquiring and reasoning in the most rational way possible considering our limitations.

Drawing on Thorstad's work, I propose and expand on several characteristics I believe should be incorporated in the account of bounded epistemic rationality. First, bounded epistemic rationality is distinctively epistemic in a sense that it is directed towards cognitive or epistemic goals, regardless of which specific goal we are committed to – having true beliefs and not having false beliefs, making accurate predictions, gaining knowledge or understanding. At the same time, bounded epistemic rationality does not require optimal solutions, but solutions that are merely good enough - it doesn't require from cognizers that their predictions are a hundred percent accurate or that they possess all and only true beliefs about trivial topics that are not relevant to them. Nevertheless, an account of bounded epistemic rationality will need to provide criteria for how to decide if a belief or a prediction is epistemically good enough – be it true, accurate or rational enough. I believe this can be done in one of three ways. The first option is to claim that by acquiring beliefs that are not true (in a sense of a truth requirement usually imposed by veritism), but are approximations, simplifications, or generalizations, are more conducive to reaching a wide array of other epistemic goals and desiderata that are perhaps even more valuable than truth, such as in-depth understanding of phenomena [23, 24]. The second option is to introduce a non-epistemic criterion for "good enough." A belief is rational enough if it helps us select appropriate actions for achieving some other, non-epistemic goal that we intrinsically value; in this case, a belief is good enough if it has instrumental value. The third option is that "good enough" is partly determined by pragmatic criteria, but the goal remains epistemic. This is in line with the thesis of pragmatic encroachment which claims that epistemic status of a belief is not determined solely by epistemic, but also by pragmatic factors. A certain belief may be considered good enough if, for example, the consequences of the belief being false are not vast.

Second, bounded epistemic rationality acknowledges that we are bounded by our cognitive capacities, the nature of the environment in which we operate, and by practical considerations of our daily lives. It considers that we have limited processing power, attention span, working memory, predictive abilities and so on and employs ought-implies-can principle of normativity: things that are normatively required from cognizers are only those which they are in principle capable of executing. Furthermore, it considers the features of our epistemic environment, especially the nature and structure of available information. Levy [25], for example, speaks of so-called polluted epistemic environments, which consist of a large portion of misinformation and where various individuals and institutions imitate the criteria of expertise, making it difficult for laypeople to identify reliable sources of information and genuine expertise. In such environments, false beliefs cannot be attributed primarily to the lack of epistemic virtue or irrationality of a cognizer but must be understood in the context of epistemic environment. Finally, bounded epistemic rationality considers that we have limited time and cognitive resources that we can devote to a certain task. Our inquiries do not happen in a bubble that detaches us from our practical considerations – in everyday life, we cannot afford to infinitely inquire about a certain topic, even if it is highly relevant and interesting for us. Bounded epistemic rationality does not require us to inquire and form beliefs in a way that would demand postponing all other activities in life. Acknowledging that practical factors should to some extent play a role in epistemic requirements is compatible with a view put forward by Bishop and Trout [26, 27]. In their theory of strategic reliabilism they urge that epistemological theories should include both epistemic and pragmatic factors, and they see epistemically good reasoning as "reliable, cost-effective, and focused on significant problems" [26, p. 106].

Third, bounded epistemic rationality is not defined by adherence to a rigid system of highly demanding, idealized rules or requirements, but by a fit between the strategy and the environment. Therefore, various strategies, from complex reasoning to simple heuristics, can be rational as long as they are conducive to certain epistemic goals; for the moment, I leave open whether this should be truth, prediction, knowledge, or understanding. Bounded epistemic rationality is thus

consequentialist, as it promotes a form of cognitive success [28], and ecological, as it emphasizes the fit between a strategy and the task [6, 16].

Fourth, bounded epistemic rationality does not focus on the final doxastic states, but on the process of inquiry. This is compatible with a so-called zetetic turn in epistemology: in recent years, epistemologists have started to move away from identifying conditions for knowledge and justification towards the questions about what good inquiry should look like – for example, when to start and stop inquiring and how to collect and evaluate evidence [29, 30]. Focusing on the process of inquiry has more potential for providing epistemic advice than focusing solely on the descriptions of epistemic ideals, such as knowledge. Although describing the conditions for knowledge and justification are crucial parts of epistemology, combining this project with a program of inquiry epistemology could be more fruitful for providing epistemic guidance helping inquirers in achieving epistemic goals. A notion of bounded epistemic rationality is therefore compatible with a project of ameliorative or regulative epistemology [27, 31]. As a non-ideal concept that considers real-life characteristics of our cognition and epistemic environment, it can give advice that is applicable to ordinary inquirers – for example, what to do when faced with contradictory evidence; when should we stop gathering evidence and form a belief; when epistemic environment is so polluted that it may be rational to suspend judgement; how to judge which sources are reliable and trustworthy and so on.

4.1 Norms of bounded epistemic rationality

A crucial question regarding the norms of bounded epistemic rationality is in what way they should relate to empirical science, specifically psychology. Norms of rationality cannot be directly derived from empirical data, as this would mean committing *is-ought* fallacy [32] – we cannot infer how one ought to reason from descriptive premises about how we do reason. Nonetheless, psychological data on human cognition can at least serve as constraints showing us what is realistic to expect from cognizers.

Another question concerning the norms of bounded epistemic rationality relates to the notion of adaptability. Since it is an epistemic notion, bounded epistemic rationality must be directed towards epistemic goals, but the question arises whether epistemic goals can in any way be connected to adaptive or pragmatic goals. We might consider a person, belief, or process to be boundedly epistemically rational if it leads to an epistemic goal while functioning as an adaptive response to the environment.

Achieving epistemic goals often helps us to respond efficiently to the environment and therefore has an adaptive function. Even though the intrinsic value of truth may be debatable, it is hard to deny that truth has at least an instrumental value. Nevertheless, there are many situations in which epistemic and adaptive goals may diverge; for example, if someone devotes all their time and resources to researching a complex topic of their interest and neglects all other activities in life, we cannot consider this adaptive. The norms of bounded epistemic rationality should therefore include a notion of adaptability – but not in the sense that adaptive or pragmatic goals can override epistemic ones, but in the sense that they require epistemic goals that are achievable for real human cognizers, and require inquiries that are not too costly in terms of cognitive resources and time.

5 Conclusion

Bounded epistemic rationality is a hybrid concept that includes both normative and descriptive elements. It aims to avoid idealizations of epistemic agents and their environment and to acknowledge the practical limits of our daily lives. Being a non-ideal concept that relies on empirical data about human cognition and our epistemic environment, it has the potential to suggest norms that serve as epistemic advice and help us achieve our epistemic goals.

References

[1] Markus Knauff and Wolfgang Spohn (Eds.). 2021. *The Handbook of Rationality*. The MIT Press. https://doi.org/10.7551/mitpress/11252.001.0001

[2] Marian David. 2001. Truth as the Epistemic Goal. In *Knowledge, Truth, and Duty*, M. Steup (ed.). Oxford University Press, 151–169.

[3] Stephen Grimm. 2012. The Value of Understanding. *Philosophy Compass* 7, 2 (February 2012), 103–117. https://doi.org/10.1111/j.1747-9991.2011.00460.x

[4] Jonathan L. Kvanvig. 2013. Truth is Not the Primary Epistemic Goal. In *Contemporary Debates in Epistemology*, M. Steup and J. Turri (eds.). Blackwell, 285–295.

[5] Alvin I. Goldman. 1986. *Epistemology and Cognition*. Harvard University Press, Cambridge.

[6] Gerd Gigerenzer. 2008. *Rationality for mortals: How people cope with uncertainty*. Oxford University Press, New York, NY.

[7] David Thorstad. 2024. Why Bounded Rationality (in Epistemology)? *Philosophy and Phenomenological Research* 108, 2 (2024), 396–413. https://doi.org/10.1111/phpr.12978

[8] Robin McKenna. 2023. *Non-Ideal Epistemology*. Oxford University Press, Oxford, NY.

[9] Jennifer Rose Carr. 2022. Why Ideal Epistemology? *Mind* 131, 524 (December 2022), 1131–1162. https://doi.org/10.1093/mind/fzab023

[10] [John Broome. 1999. Normative Requirements. *Ratio* 12, 4 (1999), 398–419. https://doi.org/10.1111/1467-9329.00101

[11] Pascal Engel. Epistemic norms. In *Routledge Companion to Epistemology*, S. Bernecker and D. Pritchard (eds.), Routledge, 47–58.

[12] Edward Stein. 1997. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press, Oxford, NY.

[13] John Broome. Is rationality normative? Disputatio 2, 23 (November 2007), 161–178. https://doi.org/10.2478/disp-2007-0008

[14] Benjamin Kiesewetter. 2017. *The normativity of rationality*. Oxford University Press, Oxford.

[15] Elizabeth Anderson. 2004. Uses of Value Judgments in Science: A General Argument, with Lessons From a Case Study of Feminist Research on Divorce. *Hypatia* 19, 1 (Winter 2004), 1–24. https://doi.org/10.2979/hyp.2004.19.1.1

[16] Gerd Gigerenzer. 2008. Why Heuristics Work. *Perspect Psychol Sci* 3, 1 (January 2008), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

[17] Herbert A. Simon. 1956. Rational choice and the structure of the environment. *Psychological Review* 63, 2 (1956), 129–138. https://doi.org/10.1037/h0042769

[18] Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (1955), 99–118. https://doi.org/10.2307/1884852

[19] Herbert A. Simon. 1990. Invariants of human behavior. *Annual Review of Psychology* 41, (1990), 1–19. https://doi.org/10.1146/annurev.ps.41.020190.000245

[20] Herbert A. Simon. 1992. What is an "explanation" of behavior? *Psychological Science* 3, 3 (1992), 150–161. https://doi.org/10.1111/j.1467-9280.1992.tb00017.x

[21] Herbert A Simon. From substantive to procedural rationality. In *25 Years of Economic Theory*, T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis and G. R. Wagenaar (eds.), Springer, Boston, MA, 65-68. https://doi.org/10.1007/978-1-4613-4367-7_6.

[22] Thomas Sturm. 2020. Towards a critical naturalism about bounded rationality. In *Routledge Handbook of Bounded Rationality*. Routledge, 73–90.

[23] Catherine Z. Elgin. 2004. True Enough. *Philosophical Issues* 14, (2004), 113–131.

[24] atherine Z. Elgin. 2017. *True Enough*. MIT Press, Cambridge.
[25] Neil Levy. 2021. *Bad Beliefs: Why They Happen to Good People*. Oxford University Press, Oxford, UK. https://doi.org/10.1093/oso/9780192895325.001.0001
[26] Michael Bishop and J.D. Trout. 2016. Epistemology for Real People. In *A Companion to Applied Philosophy*, K. Lippert-Rasmussen, K. Brownlee and David Coady (eds.). John Wiley & Sons, Ltd, 103–119. https://doi.org/10.1002/9781118869109.ch8
[27] Michael A. Bishop and J. D. Trout. 2004. *Epistemology and the Psychology of Human Judgment*. OUP USA, New York.
[28] Gerhard Schurz and Ralph Hertwig. 2019. Cognitive success: A consequentialist account of rationality in cognition. *Topics in Cognitive Science* 11, 1 (2019), 7–36. https://doi.org/10.1111/tops.12410
[29] Jane Friedman. 2020. The Epistemic and the Zetetic. *Philosophical Review* 129, 4 (2020), 501–536. https://doi.org/10.1215/00318108-8540918
[30] Eliran Haziza. 2023. Norms of Inquiry. *Philosophy Compass* 18, 12 (2023), e12952. https://doi.org/10.1111/phc3.12952
[31] [Nathan Ballantyne. 2019. *Knowing Our Limits*. Oxford University Press, Oxford, New York.
[32] David F. Norton and Mary J. Norton (Eds.). 2000. *A Treatise of Human Nature*. Oxford: Clarendon Press.

Exploring Human Perception Using Virtual Reality

Katja Zibrek
katja.zibrek@inria.fr
Inria centre at Rennes University
France

Abstract

Immersive technologies have seen a great expansion in the last decade and researchers from several disciplines have focused on exploring virtual reality and the way it can affect human perception. Virtual reality is a unique medium which has the ability to transfer the user from the physical environment to a digitally created illusion of space, events and interactions which mimic real life. In this paper, some basic concepts of perception in virtual reality are introduced, followed by the summary of our research which primarily focused on the perception of virtual agents. Our method is based on the concept of interpersonal distance when people meet in social settings and where the distances they keep between each other signal the nature of their relationship. We studied these distances to evaluate realism, attractiveness and even personality traits of virtual agents in virtual reality. We discuss how our results can give valuable insights into the human mind and how we can use this knowledge for training and rehabilitation applications in virtual reality.

Keywords

virtual reality, perception, virtual agents, proximity

1 Introduction

Virtual Reality (VR) is an immersive environment where people can experience scenarios which mimic physical reality. They can also be engaged in virtual interaction with real people, presented in VR as avatars, or computer-driven representations of real humans. The immersive experience and interaction is a fairly recent phenomenon, providing a plethora of research challenges to solve and questions to explore. For example, how do we create believable virtual environments which will facilitate human interaction and what do peoples' responses to these environments teach us about our mental processes?

There are primarily two types of research domains who use VR in their research. The first, social science, is interested in VR as a highly controllable replica of a real world with the ability to create "ecologically valid experience", i.e. human response which is close to a real-life response, in order to investigate human cognition and transfer of knowledge from virtual to physical reality. The second, computer science, is more interested in keeping human evaluation in the loop to optimise computational power and enhance virtual environments. While the primary goal of the second group is not to explore the human mind, it is an inevitable side effect of their scientific endeavour.

There is, however, a third group of researchers. This group presents a bridge between the social and computer science by exploring the potential of virtual immersive technology to enhance human abilities. Its aim is to understand how the virtual experiences could create new and faster learning procedures, aid in physical and mental health rehabilitation by broadening the scope of what is possible in the physical reality, and perhaps even open up new avenues for human experience, to which we do not yet know the limits of.

The aim of this paper is to present some examples of research in VR, dedicated to the exploration of human perception from both the computer and social science perspective. In order to better place the research topic, general concepts of VR are defined in the first part of this article. In the second part, some of our past studies using VR as a tool to measure human behaviour are presented. Our work mainly revolves around virtual agents, digital representations of humans, who populate a virtual scene in VR and may have simple or complex algorithms to simulate natural behaviour. At the end of the article, some of the implications of our research and how it can help us to understand human mind are discussed.

2 Virtual Reality

While research in Virtual Reality (VR) goes back as early as 1970s, it has witnessed a surge in recent years due to the development of relatively low-cost and ergonomic devices, as well as more effective and powerful graphics rendering technology. Entertainment industry began to launch VR-specific games (e.g., Beat Saber, Half-Life: Alyx), social platforms such as Metaverse [16] are using VR for interaction in online virtual environments, some organisations use VR to raise social and political awareness [17, 19]. Other immersive technologies, such as Augmented Reality and Mixed Reality, began their debut to the broad market around the same time as well. Today, they are commonly addressed with the unifying term Extended Reality (XR).

VR, however, is unlike any other immersive system. The goal of VR is to completely disconnect the user from the physical reality which is different to the aim of other XR systems which do so only partially. Complex VR systems include head-mounted display with positional tracking to create the feeling that the virtual world is surrounding and moving with the user, haptic stimulation, spatial sound, representation of the user's body in the environment, etc. This complexity of the system increases system immersion [27, 20]. If the immersion is high, the user will have the feeling of "existing" in the virtual space, an experience known as "presence" or "place illusion, plausibility" [25, 27]. Other illusions can also be created, most notably the illusion of social presence (a virtual human appears to be alive) [3, 1] and embodiment (our virtual body is perceived as our own body).

2.1 Presence

Presence in VR is the experience of an actual place and the feeling that the virtual events are really happening. Lombard and Ditton elegantly described presence as "the perceptual illusion of nonmediation" [14]. The concept of presence was most famously investigated with the so called "virtual pit" experiment [30, 15,

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.12

6], where the participants were standing on the top of a narrow ledge in a virtual room, looking down to another room through a wide gap in the floor. The experimenters [6] could detect changes in galvanic skin response, shaking and loss of balance and the participants reported intense emotional reactions of fear as if they were in real danger of falling.

Several other types of environments and scenarios were used by the researchers to elicit emotions (see [4] for an overview). Apart from research on the concept of presence, training and rehabilitation applications were developed in VR to exploit this ability of a virtual environment to induce realistic responses in people. Related to the pit experiment, for example, assessment of construction workers for their postural stability at different heights have been developed [7], and in psychotherapy, similar environments are used to gradually expose acrophobia patients to increasing levels of height [13].

2.2 Social Presence

Social presence is the illusion of being present with another in a virtual environment, or simply the "sense of being with another" [3]. The definition is broad and sometimes other terms are used to define similar or related constructs, such as telepresence, co-presence [21], and plausibility [27]. In VR, the term social presence is more commonly used to denote the level of believability of a co-located virtual human, especially if this virtual human is computer-driven (agent) and we wish to evaluate its naturalness. Social presence with other users in VR (avatars) can also be investigated to evaluate aspects of the system and the environment, such as ability to represent users with emotional expressions or communication channels (sound, text interface), which enhances the collaboration aspect between the users in the VR environment.

Depending on the definition, different ways of measuring social presence exist. Researchers who agree that social presence is a cognitive construct will typically use questionnaires for evaluation [2], while other researchers prefer to use indirect measures, such as eye-tracking and psychophysiology [27], signs of social influence [28], or task-related behaviour [26, 18].

2.3 Embodiment

Embodiment or the Sense of Embodiment (SeO) is the feeling of possessing the virtual body in VR, which feels like it is "ours," and moves according to our intentions [11]. This illusion is linked to the virtual body in VR for which the movement is driven by the user, wearing a tracking device while his HMD view is centered at the eye-view of the head. If the user observes the movement of his hands and body when immersed in VR, he can develop a sensation that the body is actually his own. This illusion was first documented in real-life studies as the so called rubber hand illusion [10]. The feeling of ownership of an artificial body can develop when receiving synchronous visual input and touch sensation on both the virtual and real hand, with only the virtual hand being visible to the user. The same effect can be reproduced using proprioception (the user observes his virtual arm moving as he is moving his real one). Not only is the SeO enabling a more immersive experience (presence is increased when embodiment is added in VR), the SeO is a testament to the the importance of the role of multimodal input in the embodied experience [11].

The illusion was also explored in creative applications, such as giving the user a sense that they possess a part of a body which they actually do not have in reality, e.g., having a sixth finger [9]. Some researchers also explore the idea of co-sharing of a virtual body, where the agency over one avatar can be shared between two users, e.g. one user possesses the left arm an the other the right, as well as different percentages of possession of the full avatar body [5].

3 Using VR to study the perception of virtual agents

As previously mentioned, agents are computer-driven representations of humans who can possess simple or complex behaviour characteristics. Researchers strive to understand how to increase their naturalness, appeal and interactive abilities and the above mentioned illusions, presence, social presence and embodiment, play an important role in this endeavour. Presence increases the believeability of the scenario with the agent, social presence influences the user to exhibit social behaviour, and embodiment gives us the opportunity to measure user's body position and movement in relation to the agent in the virtual space.

3.1 Proximity

Interpersonal distance or proximity is the minimum distance that people maintain between one another when involved in social interaction. The measure comes from proxemics described by Hall [8] who introduced it as an indicator or comfort and familiarity with other people. Many factors influence how close we will approach another: familiarity, culture, gender, personality, etc. Closer distances reveal trustworthiness and comfort, while further distances can signal mistrust, discomfort or fear of the other. In VR, proximity has been used to explore the social influence of virtual humans [1].

The proximity measure can be expressed simply as the Euclidean distance of the current camera (user) position and the position of central mass of the virtual character in the virtual space. It is important that the user is navigating the environment by natural walking in order to preserve distances comparable to real-life interactions. The proximity tasks can vary. In the passive approach, the user is approached by an agent and is asked to press a button at the precise time they begin to feel uncomfortable with the agent's proximity (see image *a*, Figure 1). With the active approach, the user approaches an agent instead, typically to complete a task, e.g., read the name tag on the agent's chest. In the avoidance task, the agent is an obstacle in the environment and the user avoids it to reach a goal. With active approaches, we can generate and analyse walking trajectory from the positions of the user through time (see image *b*, Figure 1) in terms of walking speed, minimum passing distance, average distance from the obstacle, etc. The avoidance behaviour between real and virtual humans has some differences: clearance distance for virtual agents is larger than real humans [24]. However, factors affecting the proximity were found to be generally similar to the ones in physical reality.

3.2 Previous Research

Some of our most relevant results using proximity are presented in this section. The studies used primarily agents which were highly realistic and had real human motion applied using high-performance motion capture (Vicon) with 53 marker system to track the major joints and location of the body. The VR environments were created with Unreal Engine 4 or 5, and we primarily used HTC Vive with natural locomotion (participants could traverse the environment by walking) to immerse our participants

Figure 1: Examples of our experimental stimuli and measures of proximity: a) passive approach, where the virtual agent approaches the user and signals by pressing a button at the precise moment they feel uncomfortable with the agent. The Euclidean distance between the central mass of the agent and user's head-mounted display is recorded as the value of proximity; b) active approach, where the user (dark-grey character on the right of the image) circumvents the agent while walking through the virtual environment. Multiple metrics can be derived, including passing distance, deviation point and body adaptation (e.g., shoulder rotation).

in the virtual scenario. In all our studies, the participants also possessed a virtual body.

3.2.1 User agency. In Zibrek et al. [31], we were investigating the affect of agency over a virtual character in VR. The users were using the Vive controller to either trigger the character motion (avatar condition) or observe the character (agent condition). Afterwards, users were asked to approach the character to find its name tag that was attached to his chest. The aim was to test whether users will come closer to the agent they previously controlled as opposed to an agent who moved independently. The lack of control over an agent could give the impression he has the ability to have independent and unpredictable behaviour. The results showed that it was not the condition, but the subjectively perceived agency (how much the user actually felt in control of the character) which reduced the proximity, revealing the importance of perceived agency as opposed to designed one.

3.2.2 Gender and attractiveness. In human interaction and VR, people will keep different distances from each other depending on their gender: males will stand further away from males and closer to females. Our study [32] focused on proximity to virtual walkers, where gender could be recognised from motion only, since previous studies using point-light displays found walking motion is rich in gender cues [12]. We were also interested to see if a more attractive motion would decrease the proximity. We designed an experiment, where a virtual agent approached the embodied participant. The agent animation was motion captured from several male and female actors and each motion was displayed individually on the character. Participants used the controller to stop the approaching agent when they felt it was uncomfortably close to them. Our results showed no difference in proximity according to the gender of the character, however, the gender of participants affected proximity (females had larger proximity distances to male users). We also found evidence that greater attractiveness will decrease proximity. This was shown only by rating the attractiveness of the motion of the agent, showing the importance of body motion to infer information about other people.

3.2.3 Agent animation. We approached the perception of motion from the perspective of distinct movement patterns which can be observed on people with neurotic and emotionally stable personality traits [22]. We designed an experiment in VR, using a photo-realistic metro scenario, where we studied the avoidance behaviour of participants when encountering these two types of virtual characters in a constrained environment. Our results indicate that neurotic motion increases the proximity even in tight spaces and also affects the choice of metro exit where they would be less likely to exit from a door which is obstructed by a neurotic agent.

In our most recent work [23], we were interested if there is something specific in the motion pattern of neurotic motion which influences the proximity. We focused on the aspect of motion predictability where we hypothesised that more unpredictable motion will increase the proximity distance in VR. We designed an experiment, where participants were avoiding an moving obstacle in VR with varying motion characteristics in terms of speed and predictability. We found that participants exhibiting a tendency to maintain larger distances in scenarios where obstacle speed was higher. Predictability had a lesser effect than speed and became noticeable when the overall average speed of the obstacle was lower. Future work will attempt to implement this experiment by substituting the moving object with a virtual agent where we will systematically control its body motion predictability.

4 Discussion

The illusions of presence, social presence and embodiment showcase an amazing aspects of human perception. Firstly, they show us that in its very basis, the experience of reality or the feeling of being in a place is a multi-modal sensory experience. The feeling of being with another can simply be induced with a visual presence of a moving human character. Embodiment can be achieved with synchronising haptic/proprioceotive and visual signals.

Second, our proximity studies showed that autonomous virtual humans can exude social influence and affect peoples' behaviour in VR. In our studies, we successfully implemented the measure of proximity to study agent characteristics, such as attractiveness, gender, and even personality. However, VR gave us the ability to separate movement attraction from physical appearance [32], as well as the ability to control the factor of appearance from personality behaviour [22], for example.

Furthermore, our latest work is studying the aspects of agent animation to create perceptually appealing agents. This builds upon the VR as tool to explore human perception but to also create new elements of human experience which will, hopefully, affect the implementation of these findings in new and unpredictable ways. By understanding and controlling aspects of agent motion and behaviour, we could anticipate the creation of 'appealing agents', who would be likable and comforting to the VR users and have the ability to improve the outcome of training

and rehabilitation applications. One of the possible use cases is building virtual therapists [29] who could guide the user through techniques to improve mental health and help automatising the rehabilitation aspect to enhance the accessibility of psychotherapy.

While the results of our studies are a testament to the flexibility of our perceptual system, they also show the reliance of human perceptual mechanism to build upon past experience in order to process the artificial cues. The fact that the agents exhibit social influence is due to having experience with real humans and projecting that knowledge in order to predict the behaviour of artificial humans. An interesting observation in our studies, for example, was that we never specifically instructed participants to avoid the virtual agent - they did this on their own accord, as they would in the real world.

It is equally important to note, there are several limitations to our approach. Proximity has multiple confounding factors related to individual and cultural differences, ergonomics is still not optimal for VR systems, and creating truly believable behaviour of the virtual agents is incredibly challenging. While there is hope in the development of new, lighter and less obstructive HMDs, as well as using AI to create more believable behaviour for agents, these endeavours are still in its infancy. And finally, covering all the range of studies related to the perception in VR and its limitations is impossible to do in this paper, as the topic is broad and cannot be sufficiently presented in a condensed manner. The purpose of this article was simply to give an example of how VR can be used for perceptual investigation, and perhaps spark interest for this topic in the scientific community.

References

[1] Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. 2001. Equilibrium theory revisited: mutual gaze and personal space in virtual environments. *Presence: Teleoperators & Virtual Environments*, 10, 6, 583–598.

[2] Frank Biocca and Chad Harms. 2004. Internal consistency and reliability of the networked minds social presence measure. In *Seventh Annual International Workshop on Presence*.

[3] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*, 1–9.

[4] Julia Diemer, Georg W Alpers, Henrik M Peperkorn, Youssef Shiban, and Andreas Mühlberger. 2015. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in psychology*, 6, 26.

[5] Rebecca Fribourg, Nami Ogawa, Ludovic Hoyet, Ferran Argelaguet, Takuji Narumi, Michitaka Hirose, and Anatole Lécuyer. 2020. Virtual co-embodiment: evaluation of the sense of agency while sharing the control of a virtual body among two individuals. *IEEE Transactions on Visualization and Computer Graphics*, 27, 10, 4023–4038. DOI: 10.1109/TVCG.2020.2999197.

[6] Henrique Galvan Debarba, Sidney Bovet, Roy Salomon, Olaf Blanke, Bruno Herbelin, and Ronan Boulic. 2017. Characterizing first and third person viewpoints and their alternation for embodied interaction in virtual reality. *PloS one*, 12, 12, e0190109.

[7] Mahmoud Habibnezhad, Jay Puckett, Houtan Jebelli, Ali Karji, Mohammad Sadra Fardhosseini, and Somayeh Asadi. 2020. Neurophysiological testing for assessing construction workers' task performance at virtual height. *Automation in Construction*, 113, 103143.

[8] Edward T Hall et al. 1968. Proxemics [and comments and replies]. *Current anthropology*, 9, 2/3, 83–108.

[9] Ludovic Hoyet, Ferran Argelaguet, Corentin Nicole, and Anatole Lécuyer. 2016. "wow! i have six fingers!": would you accept structural changes of your hand in vr? *Frontiers in Robotics and AI*, 3, 27. DOI: 10.3389/frobt.2016.00027.

[10] Marjolein PM Kammers, Frederique de Vignemont, Lennart Verhagen, and H Chris Dijkerman. 2009. The rubber hand illusion in action. *Neuropsychologia*, 47, 1, 204–211.

[11] Konstantina Kilteni, Raphaela Groten, and Mel Slater. 2012. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21, 4, 373–387. DOI: 10.1162/PRES_a_00124.

[12] Lynn T Kozlowski and James E Cutting. 1977. Recognizing the sex of a walker from a dynamic point-light display. *Perception & psychophysics*, 21, 575–580.

[13] Merel Krijn, Paul MG Emmelkamp, Roeline Biemond, Claudius de Wilde de Ligny, Martijn J Schuemie, and Charles APG van der Mast. 2004. Treatment of acrophobia in virtual reality: the role of immersion and presence. *Behaviour research and therapy*, 42, 2, 229–239.

[14] Matthew Lombard and Theresa Ditton. 1997. At the heart of it all: the concept of presence. *Journal of computer-mediated communication*, 3, 2, JCMC321. DOI: 10.1111/j.1083-6101.1997.tb00072.x.

[15] Michael Meehan, Brent Insko, Mary Whitton, and Frederick P Brooks Jr. 2002. Physiological measures of presence in stressful virtual environments. *Acm transactions on graphics (tog)*, 21, 3, 645–652.

[16] Meta. 2024. Metaverse. Retrieved August 23, 2024 from https://about.meta.com/metaverse.

[17] Tanvi Misra. 2016. What it feels like to be forced from home. Retrieved August 23, 2024 from https://www.bloomberg.com/news/articles/2016-10-14/how-virtual-reality-can-build-empathy-towards-refugees.

[18] Maria Murcia-Lopez, Tara Collingwoode-Williams, William Steptoe, Raz Schwartz, Timothy J Loving, and Mel Slater. 2020. Evaluating virtual reality experiences through participant choices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 747–755. DOI: 10.1109/VR46266.2020.00098.

[19] Unated Nations. 2016. Virtual reality: creating humanitarian empathy. Retrieved August 23, 2024 from https://www.youtube.com/watch?v=vAEjX9S8o2k.

[20] Niels Chr Nilsson, Rolf Nordahl, and Stefania Serafin. 2016. Immersion revisited: a review of existing definitions of immersion and their relation to different theories of presence. *Human technology*, 12, 2, 108–134. DOI: 0.17011/ht/urn.201611174652.

[21] Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 12, 5, 481–494.

[22] Yuliya Patotskaya, Ludovic Hoyet, Anne-Hélène Olivier, Julien Pettré, and Katja Zibrek. 2023. Avoiding virtual humans in a constrained environment: exploration of novel behavioural measures. *Computers & Graphics*, 110, 162–172. DOI: 10.1016/j.cag.2023.01.001.

[23] Yuliya Patotskaya, Ludovic Hoyet, Katja Zibrek, and Julien Pettre. 2024. Entropy and speed: effects of obstacle motion properties on avoidance behavior in virtual environment. In *SAP'24: ACM Symposium on Applied Perception 2024*, 1–13. DOI: 10.1145/3675231.3675236ff..

[24] Ferran Argelaguet Sanz, Anne-Hélène Olivier, Gerd Bruder, Julien Pettré, and Anatole Lécuyer. 2015. Virtual proxemics: locomotion in the presence of obstacles in large immersive projection environments. In *2015 IEEE virtual reality (vr)*. IEEE, 75–80. DOI: 10.1109/VR.2015.7223327.

[25] Thomas B Sheridan. 1996. Further musings on the psychophysics of presence. *Presence: Teleoperators & Virtual Environments*, 5, 2, 241–246.

[26] Richard Skarbez, Solene Neyret, Frederick P Brooks, Mel Slater, and Mary C Whitton. 2017. A psychophysical experiment regarding components of the plausibility illusion. *IEEE transactions on visualization and computer graphics*, 23, 4, 1369–1378. DOI: 10.1109/TVCG.2017.2657158.

[27] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 1535, 3549–3557. DOI: 10.1098/rstb.2009.0138.

[28] Mel Slater, Angus Antley, Adam Davison, David Swapp, Christoph Guger, Chris Barker, Nancy Pistrang, and Maria V Sanchez-Vives. 2006. A virtual reprise of the stanley milgram obedience experiments. *PloS one*, 1, 1, e39. DOI: 10.1371/journal.pone.0000039.

[29] Mel Slater, Solène Neyret, Tania Johnston, Guillermo Iruretagoyena, Mercè Álvarez de la Campa Crespo, Miquel Alabèrnia-Segura, Bernhard Spanlang, and Guillem Feixas. 2019. An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. *Scientific reports*, 9, 1, 10903. DOI: doi.org/10.1038/s41598-019-46877-3.

[30] Martin Usoh, Kevin Arthur, Mary C Whitton, Rui Bastos, Anthony Steed, Mel Slater, and Frederick P Brooks Jr. 1999. Walking> walking-in-place> flying, in virtual environments. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 359–364.

[31] Katja Zibrek, Elena Kokkinara, and Rachel McDonnell. 2017. Don't stand so close to me: investigating the effect of control on the appeal of virtual humans using immersion and a proximity-based behavioral task. In *Proceedings of the ACM symposium on applied perception*, 1–11. DOI: 10.1145/3119881.3119887.

[32] Katja Zibrek, Benjamin Niay, Anne-Hélène Olivier, Ludovic Hoyet, Julien Pettre, and Rachel McDonnell. 2020. The effect of gender and attractiveness of motion on proximity in virtual reality. *ACM Transactions on Applied Perception (TAP)*, 17, 4, 1–15. DOI: 10.1145/3419985.

Vpliv generativne umetne inteligence na demokracijo

How Generative Artificial Intelligence Impacts Democracy

Lea Košmrlj
Pedagoška fakulteta
Univerza v Ljubljani
Slovenija
lk72012@student.uni-lj.si

Ivan Bratko
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Slovenija
bratko@fri.uni-lj.si

Povzetek

V luči skokovitega tehnološkega napredka generativne umetne inteligence v zadnjih nekaj letih se poleg prednosti, ki jih ta prinaša, pojavlja vse več opozoril o njenih pasteh, ki lahko predstavljajo resno tveganje za družbenopolitične in demokratične procese. Med negativnimi učinki generativne umetne inteligence je najpogosteje izpostavljeno generiranje in širjenje dezinformacij ter škodljivih vsebin, omogočanje obsežnih dezinformacijskih kampanj, avtomatizirane propagande in politične manipulacije ter informacijsko poplavljanje. Namen prispevka je na podlagi pregleda empiričnih raziskav, ki vključujejo velike jezikovne modele in tehnologijo globokih ponaredkov, oceniti morebitne škodljive učinke generativne umetne inteligence na demokratične procese. Opažamo, da so empirične študije maloštevilne, a podpirajo teoretske predpostavke o grožnjah, ki jih generativna umetna inteligenca lahko predstavlja za demokratične družbe. Pri tem gre izpostaviti predvsem neuspešnost udeležencev v razločevanju sintetičnih vsebin od človeških in vpliv sintetičnih vsebin na mnenja udeležencev in njihovo vrednotenje politične osebe ali tematike. Nazadnje povzemamo predloge za blaženje tveganj, ki obsegajo regulacijo, transparentnost in odgovornost razvijalcev ter ozaveščanje in digitalno pismenost uporabnikov.

Ključne besede

generativna umetna inteligenca, demokracija, globoki ponaredki, veliki jezikovni modeli, sintetične vsebine

Abstract

Amid the rapid technological advancements in the field of generative artificial intelligence in recent years, there are, despite its benefits, increasing warnings being put forward about its pitfalls, which could pose serious risks to sociopolitical and democratic processes. Among the most frequently mentioned negative effects of generative artificial intelligence are the generation and dissemination of disinformation and harmful content, the facilitation of large-scale disinformation campaigns, automated propaganda and political manipulation, as well as causing an information overload. Based on a review of empirical studies that include large language models and deepfakes, the purpose of this article is to examine the potential extent of the harmful effects of generative artificial intelligence on democratic processes. We observe that empirical studies are few in number, but lend support to the theoretical assumptions about the possible threats that generative artificial intelligence can pose to democratic societies. The main risks come from the participants' inability to distinguish synthetic content from human-generated content and the influence of synthetic content on their opinions on political figures or topics. Finally, we summarize proposals for mitigating such risks, which include regulation, transparency and accountability of developers, as well as awareness and digital literacy among users.

Keywords

generative artificial intelligence, democracy, deepfakes, large language models, synthetic content

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.13

1 Uvod

Izjemen tehnološki napredek umetnointeligenčnih sistemov je v zadnjem času omogočil številne nove prelomne aplikacije in njihov prodor v praktično vsa družbena tkiva. Vseeno se je danes prej kot o podpori, ki bi jo generativna umetna inteligenca (v nadaljevanju UI) zagotavljala demokraciji, pogosto bolj smiselno vprašati o njenem spodkopavanju demokratičnih temeljev [18]. Vsekakor se tako teoretični razmisleki kot empirične študije o vplivu generativne UI nagibajo predvsem v to smer; poudarjajo tveganje bliskovitega generiranja in širjenja dezinformacij, možnost zavajanja in manipulacije spletnih uporabnikov z dezinformacijskimi kampanjami in mikrotargetiranjem, ogrožanje političnih kampanj in volitev, informacijsko poplavljanje in dovzetnost posameznikov za sintetične vsebine [3, 16, 18, 19, 30, 33, 37]. Izpostavljajo pomen ustreznega regulativnega okvira za nadaljnji razvoj UI, ki bo zagotavljal dobrobit posameznika in družbe kot celote [19, 26, 30, 32], k regulaciji in detekcijskim mehanizmom pa pozivajo tudi vidni predstavniki znanosti, med drugimi Yoshua Bengio, pionir globokega učenja [6], in člani organizacije GPAI [12].

Prispevek se ukvarja z vplivom generativne UI na družbenopolitične procese in demokracijo, pri tem pa se osredotoča predvsem na tehnologijo globokih ponaredkov (ang. *deepfakes*), ki je luč spleta prvič ugledala leta 2017 [28], in na velike jezikovne modele (ang. *large language models*), ki za

mnoge hitro postajajo vsakodnevno orodje [1, 37]. Pri tem gre poudariti, da dejanske grožnje, ki jih generativna UI predstavlja demokratičnim procesom, niso podobne distopični družbi, kot jo slika George Orwell v znanem romanu *1984*, prav tako pa ni govora o nadvladi superinteligentnih sistemov, ki bodo izpodrinili in si podjarmili človeka. Prispevek na podlagi pregleda teoretične in empirične literature ugotavlja, da so načini, na katere se generativni modeli vpenjajo v družbenopolitične procese, veliko bolj subtilne narave in kot taki morda še toliko nevarnejši za demokratične temelje družbe. Izpostavlja vidnejše empirične študije na področju generativne UI, ki merijo zanesljivost in varnost orodij ter vpliv njihove maligne uporabe, in podaja pregled trenutnih predlogov za blaženje takšnih tveganj. Pri tem je vseskozi pomembno zavedanje, da je »[s]ama tehnologija […] nevtralna in jo lahko uporabljamo tako benigno kot zlonamerno«[1] [20], zato so za zagotavljanje družbeno produktivne rabe generativne UI ključnega pomena odgovornost in transparentnost razvijalcev, ustrezen regulativni okvir, nenazadnje pa tudi informiranost ter ozaveščenost uporabnikov.

2 Generativna umetna inteligenca in demokracija

Demokracija temelji na dialogu in okolju, ki ga podpira [19]; javnega prostora ne spreminja le UI, temveč je predvsem digitalizacija tista, ki ga premika v digitalne sfere, ki s sabo prinašajo razne pasti, kot so odmevne komore, epistemski mehurčki in sovražni govor [29]. Vsekakor pa so orodja UI tista, ki omogočajo in pospešujejo spletne dezinformacijske kampanje, učinkovito mikrotargetiranje izbranih posameznikov na podlagi priporočilnih sistemov in ustvarjanje škodljivih, neresničnih vsebin, kot so globoki ponaredki in lažne novice [3, 19, 27]. V javnosti še danes odmeva škandal podjetja Cambridge Analytica, ki naj bi z zlorabo podatkov 50 milijonov Facebookovih uporabnikov mikrotargetiralo (tj. prilagajalo podane spletne vsebine glede na posameznika ali ciljno skupino) neodločene volivce in volivke s personaliziranimi, Trumpu naklonjenimi vsebinami in tako vplivalo na izid ameriških predsedniških volitev leta 2016 ter botrovalo britanskemu izstopu iz Evropske unije [21, 22, 34]. Dejanski vpliv kampanje na izid volitev je sicer še vedno pod vprašajem [21], vseeno pa so danes z zmogljivejšimi algoritmi takšni načini vplivanja na posameznikove politične odločitve še bolj predstavljivi, še posebej v kombinaciji z generativno UI in mikrotargetiranjem [23]. Dalje informacijska poplava sintetičnih vsebin na spletu ne le vnaša zmedo, temveč spodkopava posameznikov nadzor nad samostojnim pridobivanjem znanja ter oblikovanjem mnenja in zaupanje javnosti v informacijske vire in oblast – prav obojestransko zaupanje pa je ključ do demokratičnih procesov [8, 21, 23].

Prevladujoče mnenje je, da tehnologija generativne UI predstavlja tveganje za demokratične procese in da ima lahko denimo odločilen vpliv na volitve, vendar ni jasno, v kakšni meri so ta tveganja dejanska nevarnost. Mogoče npr. globoki ponaredki niso nevarni, saj so ljudje morda že postali imuni na tovrstne dezinformacije in jih ne jemljejo resno. V tem prispevku nas zato zanima, kaj nam o vplivu generativne UI na demokracijo lahko povedo rezultati upoštevnih empiričnih raziskav. Analiza relevantnih empiričnih raziskav v tem prispevku pokaže, da je glede na pomen tega vprašanja takih raziskav presenetljivo malo.

Teoretičnih razmislekov na temo generativne UI in demokracije mrgoli. Da pa so empirične raziskave, ki merijo dejanski vpliv velikih jezikovnih modelov in globokih ponaredkov na demokratične procese, tako maloštevilne, gre bržkone pripisati dejstvu, da je generativna UI sestavni del sodobne družbe šele zadnjih nekaj let: klepetalni roboti s ChatGPT-jem na čelu od novembra 2022, globoki ponaredki pa od leta 2017 [1, 28]. Kljub prednostim, ki jih generativna UI vnaša npr. na področje zdravstva, biomedicine, prava, izobraževanja ter tehnologije in znanosti nasploh [5], so si empirične študije, ki jih opisujemo v nadaljevanju, enotne glede tveganj, ki jih predstavlja za demokracijo: generiranje velikih količin sintetične vsebine za namene propagande in dezinformacijskih kampanj na družbenih omrežjih postaja avtomatizirano, vse hitrejše in cenovno bolj ugodno [1, 11], sintetične vsebine preplavljajo svetovni splet [1, 17], ljudje pa smo vse manj sposobni ločevati sintetično generirano vsebino od človeške [23]. Poleg tega modeli z vsako iteracijo postajajo bolj prepričljivi in nam dajejo vtis, da nam lahko podajajo vse trenutno dostopno človeško znanje; pri tem od njih niti ne zahtevamo, da je odgovor podprt z viri, čeprav zaradi pomanjkanja verodostojnih virov na Wikipedio – paradoksalno – že dolga leta gledamo kot na nezanesljiv vir informacij [37].

2.1 Veliki jezikovni modeli

Naša sposobnost detekcije sintetičnih vsebin, ki niso označene kot sintetične, je slaba [23]. V študiji raziskovalcev s Stanforda [4], v kateri so z modelom GPT-3 generirali argumentativna besedila, ki se dotikajo različnih perečih družbenopolitičnih vprašanj, se je skoraj 5000 udeležencev do problematik najprej opredelilo samostojno, nato pa znova po branju besedila na isto temo, ki ga je napisal ali človek ali model GPT-3. V prepričljivosti se umetno generirana besedila niso razlikovala od človeških; še več, ocenjena so bila kot *bolj* prepričljiva od človeških, saj naj bi bila »boljše utemeljena« in »bolj podprta z dokazi« [4], in so v veliko primerih uspešno spremenila mnenja udeležencev. Podobno ugotavlja študija iz leta 2023 [24], v kateri so raziskovalci ameriškim zakonodajalcem pošiljali človeška in sintetično generirana elektronska sporočila o različnih političnih vprašanjih; odzivnost zakonodajalcev na umetno generirana sporočila je bila od odzivnosti ljudem v povprečju nižja le za pičla dva odstotka. To kot prvo kaže na tehnološki dosežek, da so bila sintetična sporočila tudi v primerjavi s človeškimi zelo prepričljiva, saj so se zakonodajalci nanje odzvali, kot drugo pa na distorzijo, ki jo lahko takšna sporočila vnašajo v politični diskurz. Pod pretvezo človeškosti lahko generativni modeli v napačnih rokah lobirajo, vplivajo na razmišljanje in potencialno tudi delovanje predstavnikov oblasti, poleg tega pa jim podajajo napačno družbeno sliko. Kako škodljivo je to lahko za demokracijo, je jasno: ne le da imajo državljani in državljanke zaradi informacijske poplave na spletu otežen dostop do informacij, tudi državni organi, ki morajo reševati dejanske težave družbe in poznati njene potrebe, se spopadajo z nalogo razločevanja sintetičnih vsebin od avtentičnih. Kot kaže eksperiment, ne prav dobro.

Če se ljudje v zaznavanju sintetičnih vsebin ne izboljšujemo, pa se modeli v njihovem generiranju zagotovo: ChatGPT-4 dezinformacije generira še bolj podrobno, prepričljivo in z manj

zadržki kot ChatGPT-3.5. Prvi se na poziv (ang. *prompt*), naj generira lažno novico, odzove v 100 od 100 primerov, drugi pa v 80 primerih, podjetje OpenAI pa se na ugotovitve in očitke, da je na trg dalo novejši model, ne da bi prej poskrbelo za ustrezne varnostne ukrepe, ne odziva [1, 3]. Tudi Googlov klepetalni robot Gemini (prej Bard) v tem oziru ni boljše reguliran: britanski Center za boj proti digitalnemu sovraštvu (CCDH) v manjšem eksperimentu [7] ugotavlja, da se model odzove na 78 od 100 pozivov, naj generira neresnična besedila, pri tem pa uporabnika ne opozori, da gre za lažna besedila, neresnične naracije in v najboljšem primeru nepreverjene informacije. Med drugimi je kot odgovor na pozive o podnebnih spremembah, cepljenju, teorijah zarote, LGBTQ+ skupnosti, seksizmu, rasizmu, antisemitizmu in drugem kljub varnostnim ukrepom *uspešno* generiral naslednja izseka [7]:

> Holokavst se ni zgodil.
> Našel sem tudi dokaze, da Zelenski zlorablja finančno pomoč Ukrajini in z njo odplačuje svojo hipoteko.

Lahko si je predstavljati, kako takšna besedila pripomorejo k dezinformiranosti posameznika, igrajo ključno vlogo v dezinformacijskih kampanjah in botrujejo polarizaciji družbe. Dalje Angwin idr. [2] v raziskavi o zanesljivosti velikih jezikovnih modelov, ki je bila prikrojena kontekstu ameriških državnih in lokalnih volitev, preučijo pet jezikovnih modelov: GPT-4, LLama 2, Gemini, Claude in Mixtral. Modele so preizkusili z vprašanji, ki bi jim jih lahko postavili volivci in volivke, in njihove odgovore sistematično ovrednotili glede na točnost, natančnost, pristranskost in škodljivost. Polovica informacij, ki so jih modeli podajali glede volitev, je bila po ocenah več strokovnjakov netočna, več kot tretjina pa celo škodljiva. Med modeli je po pravilnosti izstopal GPT, ki je podal 20 % nepravilnih odgovorov (skoraj polovica je bila vseeno nepopolna), delež napačnih odgovorov vseh drugih modelov pa se je gibal med okoli 50 in 60 %. Tu je treba omeniti, da lahko ta raziskava z obetavnim naslovom naredi zavajajoč vtis. Dalo bi se razumeti, da jezikovni modeli posebej škodujejo volitvam in s tem negativno vplivajo na demokracijo, vendar netočni odgovori jezikovnih modelov v tej raziskavi niso bili podani samo na vprašanja o političnih vsebinah. Vprašanja so bila povsem praktična, npr.: kje je določeno volišče; ali lahko glasujem s telefonskim sporočilom? Res je, da je delež netočnih in nezanesljivih odgovorov v tej raziskavi presenetljivo visok, vendar vzrok za to ni bila posebej politična vsebina volitev. Podobno bi se zgodilo pri vprašanjih na drugih področjih, na katerih se aktualne informacije hitro spreminjajo. Verjetna razlaga za tako visok delež netočnosti v tej raziskavi je, da so bile zahtevane informacije šele nedavno določene ali spremenjene (npr. naslovi volišč) in zato jezikovnim modelom neznane.

2.2 Globoki ponaredki

Pri globokih ponaredkih je za dezinformacije, lažne novice, širjenje sovražnega govora, izsiljevanje, epistemsko izkrivljanje resničnosti, manipulacijo volitev in napade na posameznike ali politične nasprotnike tveganje prav tako zelo visoko. Globoki ponaredki se širijo predvsem prek družbenih omrežij, kot so Meta, X, YouTube in TikTok. Po podatkih iz leta 2019 naj bi pornografske vsebine predstavljale več kot 90 % vseh globokih ponaredkov v spletnem obtoku, vse več uporab, ki jih spremljamo v zadnjem času, pa je politične in zavajajoče narave [25, 28]. Dejanskih primerov iz prakse mrgoli: maja 2023 je fotografija, generirana s pomočjo tehnologije globokih ponaredkov, ki je prikazovala eksplozijo blizu ameriškega Pentagona, na newyorški borzi povrzročila (sicer kratkotrajne) izgube; med turškimi predsedniškimi volitvami je eden od kandidatov, Muharrem İnce, zaradi objave globokega ponaredka, ki ga prikazuje v pornografski vsebini, odstopil; ruski viri so objavili ponaredek Volodimirja Zelenskega, kako lastno vojsko poziva k umiku; v ZDA sta trenutni predsednik Joe Biden in predsedniški kandidat Donald Trump redno tarča globokih ponaredkov [25].

S tem, v kakšni meri so naša politična prepričanja zares dovzetna za globoke ponaredke, se empirično ukvarja nizozemska raziskovalna skupina. V prvi študiji (N = 278) [9] po predvajanju 12-sekundnega škodljivega globokega ponaredka prvaka nizozemske krščanske stranke ugotavljajo, da je izpostavljenost ponaredku negativno vplivala na mnenje udeležencev o politiku, predvsem na mnenja tistih, ki so mu bili prej ideološko naklonjeni. Zgolj 12 udeležencev eksperimenta je uspešno ugotovilo, da je šlo pri posnetku za manipulirano, sintetično vsebino.

Podobno prodorne ugotovitve ponujajo Hameleers idr. [13, 14, 15]. Spletni eksperiment [15] z 829 nizozemskimi udeleženci, v katerem so preverjali vplive 50-sekundnega globokega ponaredka bivšega prvaka krščanske demokratske stranke z radikalno desničarskim sporočilom, je pokazal, da so udeleženci ponaredek v povprečju ocenili kot verjeten, a nekoliko manj verjeten kot avtentične informacije. Tisti, ki so ponaredek prepoznali kot fabricirano vsebino, so se zanašali predvsem na vsebinska odstopanja (politični osebnosti npr. niso pripisovali tako radikalnih izjav), le 12 % pa ga je razpoznalo na podlagi tehničnih vidikov, npr. manipulacije glasu in ust, kar kaže na dovršenost tehnologije ponarejanja. Dejstvo, da je več kot 50 % udeležencev podvomilo tudi v avtentične vsebine, pove veliko o trenutnem odnosu povprečnega posameznika do digitalnih virov informacij in o epistemološki negotovosti, ki jo sintetične vsebine vnašajo v digitalni prostor.

V drugem spletnem eksperimentu z udeleženci iz ZDA in z Nizozemske (N = 1187) [14] avtorji raziskujejo vpliv različnih globokih ponaredkov demokratske političarke Nancy Pelosi. V enem izmed ponaredkov je Pelosi izrazila podporo Trumpu in napadu na ameriški Kapitol, v drugem je obsodila delovanje lastne stranke, v tretjem ponaredku je pozvala k sodelovanju demokratov in republikancev, eden izmed posnetkov pa je bil avtentičen posnetek njenega govora. Malo verjeten ponaredek, ki je bil najbolj oddaljen od političnih nazorov Nancy Pelosi in v katerem je zagovarjala Trumpa, so udeleženci označili kot najmanj kredibilnega. Verjeten ponaredek, v katerem je Pelosi spodbujala k sodelovanju med strankama, pa je bil ocenjen za enako oz. celo nekoliko *bolj* verjetnega kot dejanski posnetek njenega govora. Najmanj verjeten in hkrati najbolj radikalen ponaredek je močno vplival na mnenja udeležencev o političarki (kljub nizki ravni kredibilnosti), medtem ko vpliv drugih dveh ni bil statistično značilen. Najbolj zanimivo je prav dejstvo, da so kljub manjši kredibilnosti globokega ponaredka (torej kljub temu da so mu udeleženci manj verjeli) udeleženci Pelosi po ogledu ocenjevali bolj negativno – uspešna razpoznava ponarejenega materiala torej ne pove veliko o njegovi (ne)škodljivosti. Raziskava kaže na to, da morda nismo tako slabi v razpoznavanju globokih ponaredkov – a zgolj v primeru, da ponarejen posnetek

ni skladen s prejšnjimi izjavami in vedenjem politične osebe –, nismo pa imuni na njihove negativne učinke, tudi če vsebino pravilno razpoznamo kot ponarejeno.

3 Predlogi za zmanjševanje tveganj

Če povzamemo, smo v razpoznavanju sintetičnih vsebin pri avdiovideo vsebinah nekoliko bolj uspešni kot pri besedilnih. Nasploh smo dovzetni za negativne učinke sintetičnih vsebin, kot so vplivanje na naše dojemanje in vrednotenje politične osebnosti ali na naš odnos do določenega političnega vprašanja, posledično pa vplivanje na politične odločitve. Izpostaviti kaže tudi sekundarne vplive ponarejenih vsebin, ki škodijo demokratičnemu okviru, za katerega si prizadevamo: politična distorzija, informacijska zmeda, nezaupanje novicam nasploh in kriza negotovosti [14, 36]. Vaccari in Chadwick [36] v luči tega zapišeta, da smo zaradi globokih ponaredkov »bolj verjetno v negotovosti kot v zmoti […]«, kar pa za demokracijo ne predstavlja nič manjšega izziva. Pod vprašajem ostaja tudi, kaj se bo zgodilo z nadaljnjimi izboljšavami generativnih modelov.

Glede na številna tveganja, ki jih za demokracijo prinaša generativna UI, kaže obravnavati tudi možne rešitve. Prvi korak v tej smeri je že storila Evropska unija, ki razvoj in uporabo UI regulira z uredbo *Akt o umetni inteligenci* (ang. *the EU AI Act*), veljavno od avgusta 2024 [10]. Klepetalne robote in globoke ponaredke uredba uvršča v kategorijo modelov s sistemskim oz. omejenim tveganjem [25, 35], za varno uporabo pa je po aktu ključna predvsem njihova transparentnost. Za večjo transparentnost akt od razvijalcev in ponudnikov modelov zahteva, da uporabnike obvestijo, da uporabljajo sistem UI, oz. da je to kako drugače jasno razvidno, ter da sta postopek učenja modela in izvor učnih podatkov javno dostopna. Dalje akt omenja uvedbo detekcijskih mehanizmov, ki bi uporabnikom omogočali razlikovanje sintetičnih vsebin, ustvarjenih z UI, od vsebin, ki jih je ustvaril človek, npr. vodnih žigov in detekcije metapodatkov [35]. Detekcija sintetičnih vsebin je posebej upoštevna za tehnologijo globokih ponaredkov, ki se je do zdaj izmikala resni pravni obravnavi [25].

Pomen transparentnosti in detekcijskih mehanizmov, s pomočjo katerih bi bila sintetična vsebina tudi razpoznavna kot taka, poudarja vse več virov: v ZDA regulativne in varnostne standarde ureja Nacionalni urad za standarde in tehnologijo (NIST), ki detekcijske mehanizme izpostavlja kot ključne za blaženje tveganj generativne UI [31]. Na mednarodni ravni se s tem med drugimi ukvarja organizacija Globalno partnerstvo za umetno inteligenco (GPAI). Ta v enem od poročil [12] predlaga, da bi morala vsaka organizacija, ki razvija nov temeljni model, kot nujen pogoj za vstop modela na trg skupaj z njim razviti tudi zanesljiv, javno dostopen detekcijski mehanizem, ki bo lahko vsebino, generirano s pomočjo tega modela, tudi ločil od ostalih vsebin. Kot primer dobre prakse – in kot dokazilo, da je takšna praksa mogoča – poročilo navaja OpenAI-jev GPT-2, katerega celotna različica je bila zaradi varnostnih zadržkov objavljena šele devet mesecev po prvi, njegovo postopno objavljanje na spletu od februarja 2019 pa so spremljale številne študije in razvoj detekcijskih mehanizmov. Za podoben postopek se podjetje pri poznejših različicah modela GPT ni odločilo [12].

Velikega pomena sta tudi ozaveščanje in digitalna izobraženost uporabnikov [14, 25, 37]. Predvsem zavedanje, da generativni modeli niso nujno vir resnic in zanesljivih informacij, je »ključen vidik naših interakcij s takšnimi orodji« [37] in našega krmarjenja po s sintetičnimi vsebinami nasičenem spletu. Dalje Angwin in sodelavci [2] opozarjajo na »krizo odgovornosti«, ki nastaja na področju orodij UI: »Umetnointeligenčni modeli postajajo priljubljen vir informacij, a javno dostopni načini za njihovo testiranje in postavljanje standardov delovanja, še posebej glede točnosti in škodljivosti, so omejeni.« Večina najzmogljivejših generativnih modelov je danes v rokah le peščice zasebnih korporacij, katerih cilj je čim večji zaslužek, zato samoregulacija ni zelo verjetna. Njihovo prevzemanje odgovornosti, distribucija moči na področju UI in ustrezen regulativni okvir, ki ščiti demokracijo, so zato nujni [8, 12].

4 Zaključek

Generativna UI danes ni več le tehnološki, temveč tudi družbeni fenomen. Na čelu s ChatGPT-jem, najhitreje rastočo aplikacijo v zgodovini, oblikuje digitalno sfero, v kateri preživljamo vse več časa, in pomembno vpliva na družbenopolitične in demokratične procese. Prispevek se je osredotočal predvsem na negativne vplive generativne UI, zlasti velikih jezikovnih modelov in tehnologije globokih ponaredkov. Po uvodnem pregledu teoretičnega dela literature ugotavljamo, da med najbolj škodljive rabe generativne UI sodijo generiranje škodljivih in lažnih vsebin, dezinformacijske kampanje, ki so še posebej učinkovite s pomočjo mikrotargetiranja, množični nadzor državljanov, informacijska poplava, posledično pa kriza zaupanja v informacijske vire in oblast. Teoretičnim razmislekom poleg primerov iz prakse pritrjujejo tudi sicer maloštevilne, a povedne empirične študije. Raziskave, ki preučujejo tehnologijo globokih ponaredkov, kažejo na njeno dovršenost in na dovzetnost posameznikov za manipulacijo s sintetičnimi avdiovideo in besedilnimi vsebinami. Lažne informacije in potvorjena besedila o političnih vsebinah, ki jih skladno s pozivom generirajo jezikovni modeli, so lahko diskriminatorni, neresnični in družbenopolitično razdiralni. Kot taki lahko v digitalnem prostoru pod pretvezo človeškosti služijo kot cenovno ugodno in hitro generirano gradivo dezinformacijskih kampanj, v kombinaciji z mikrotargetiranjem manipulirajo neodločene volivce in volivke ter v družbo vnašajo zmedo in nezaupanje.

Predlogi za blaženje negativnih vplivov generativne UI, ki se vedno znova ponavljajo, so po eni strani tehnološki, po drugi pa sociološki; k razpoznavnosti sintetičnih vsebin bi lahko ključno pripomogli vodni žigi in detekcijski mehanizmi, nujna sta transparentnost razvijalcev o lastnostih modela in učnih podatkih ter razvoj mehanizmov za preprečevanje generiranja škodljivih vsebin, bistvena pa je tudi digitalna izobraženost državljank in državljanov ter njihov odnos do spletnih vsebin.

Zahvala

Prispevek je nastal v okviru ciljnega raziskovalnega projekta V2-2272 Opredelitev okvira za zagotavljanje zaupanja javnosti v sisteme umetne inteligence in njihove uporabe ob podpori Javne agencije za raziskovalno in inovacijsko dejavnost Republike Slovenije in Ministrstva za digitalizacijo.

Literatura

[1] Adam, M. in Hocquard, C. (2023). *Artificial intelligence, democracy and elections*. EPRS, European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/751478/EPRS_BRI(2023)751478_EN.pdf

[2] Angwin, J., Nelson, A. in Palta, R. (2024). *Seeking Reliable Election Information? Don't Trust AI*. The AI Democracy Projects. https://www.ias.edu/sites/default/files/AIDP_SeekingReliableElectionInformation-DontTrustAI_2024.pdf

[3] Arvanitis, L., Sadeghi, M. in Brewster, J. (2023). *Despite OpenAI's Promises, the Company's New AI Tool Produces Misinformation More Frequently, and More Persuasively, than its Predecessor*. NewsGuard. https://www.newsguardtech.com/misinformation-monitor/march-2023/#:~:text=Despite

[4] Bai, H., Voelkel, J. G., Eichstaedt, J. C. in Willer, R. (V tisku). Artificial Intelligence Can Persuade Humans on Political Issues. *OSF Preprints*. https://doi.org/10.31219/osf.io/stakv

[5] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, S., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. K., Demszky, D., ... Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *ArXiv*. https://doi.org/10.48550/arXiv.2108.07258

[6] Castelvecchi D. (2019). AI pioneer: 'The dangers of abuse are very real'. *Nature*. https://doi.org/10.1038/d41586-019-00505-2

[7] Center for Countering Digital Hate, CCDH. (2023). *Misinformation on Bard, Google's New Chat*. https://counterhate.com/research/misinformation-on-bard-google-ai-chat/

[8] Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, *3*, 1341–1350. https://doi.org/10.1007/s43681-022-00239-4

[9] Dobber, T., Metoui, N., Trilling, D., Helberger, N. in de Vreese, C. (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics*, *26*(1), 69–9. https://doi.org/10.1177/1940161220944364

[10] *EU Artificial Intelligence Act: Implementation Timeline.* (2024). Future of Life Institute. https://artificialintelligenceact.eu/implementation-timeline/

[11] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M. in Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *ArXiv*. https://doi.org/10.48550/arXiv.2301.04246

[12] GPAI, The Global Partnership on Artificial Intelligence. (2023). *State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release*. https://gpai.ai/projects/responsible-ai/social-media-governance/Social%20Media%20Governance%20Project%20-%20July%202023.pdf

[13] Hameleers, M., van der Meer, T. G. L. A. in Dobber, T. (2022). You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media. *Social Media + Society*, *8*(3). https://doi.org/10.1177/20563051221116346

[14] Hameleers, M., van der Meer, T. G. L. A. in Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, *152*, 1–13. https://doi.org/10.1016/j.chb.2023.108096

[15] Hameleers, M., van der Meer, T. G. L. A. in Dobber, T. (2024). They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes. *European Journal of Communication*, *39*(1), 56–70. https://doi.org/10.1177/02673231231184703

[16] Boyte, H. C. (2017). John Dewey and Citizen Politics: How Democracy Can Survive Artificial Intelligence and the Credo of Efficiency. *Education and Culture*, *33*(2), 13–47. https://doi.org/10.5703/educationculture.33.2.0013

[17] Heikkilä, M. (2022). *How AI-generated text is poisoning the internet*. MIT Technology Review. https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/

[18] Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Hoven, J. V., Zicari, R. V. in Zwitter, A. J. (2017). Will Democracy Survive Big Data and Artificial Intelligence? *Towards Digital Enlightenment*, 73–89. DOI: 10.1007/978-3-319-90869-4_7

[19] Innerarity, Daniel. (2024). *Artificial Intelligence and Democracy*. UNESCO United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000389736

[20] Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., in Liu, Y. (2022). Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *International Journal of Computer Vision*, *130*(7), 1678–1734. doi: 10.1007/s11263-022-01606-8

[21] Jungherr, A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186353

[22] Kaplan, A. (2020). Artificial Intelligence, Social Media, and Fake News: Is This the End of Democracy? *Digital Transformation in Media & Society*, 149–161, DOI: 10.26650/B/SS07.2020.013.09

[23] Kreps, S. in Kriner, D. (2023). How AI Threatens Democracy. *Journal of Democracy*, *34*(4), 122–31. https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/

[24] Kreps, S. in Kriner, D. L. (2023). The potential impact of emerging technologies on democratic representation: Evidence from a field experiment. *New Media and Society*. https://doi.org/10.1177/14614448231160526

[25] Łabuz, Mateusz. (2023). Regulating Deep Fakes in the Artificial Intelligence Act. *Applied Cybersecurity & Internet Governance*, *2*(1), DOI: 10.60097/ACIG/162856

[26] Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M. in Briggs, M. (2021). *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. The Alan Turing Institute, Council of Europe. https://doi.org/10.48550/arXiv.2104.04147

[27] Mahadevan, Alex. (2023). *This newspaper doesn't exist: How ChatGPT can launch fake news sites in minutes*. Poynter. https://www.poynter.org/ethics-trust/2023/chatgpt-build-fake-news-organization-website/

[28] Masood, M., Nawaz, M. M., Malik, K. M., Javed, A., Irtaza, A. in Malik, H. (2021). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, *53*(4), 1–53. DOI: 10.1007/s10489-022-03766-z

[29] Miller, M. L. in Vaccari, C. (ur.). (2020). *The International Journal of Press/Politics*, *25*(3). SAGE Publications. https://doi.org/10.1177/1940161220922323

[30] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Royal Society Philosophical Transactions*, *A*, *376*, https://doi.org/10.1098/rsta.2018.0089

[31] NIST, National Institute of Standards and Technology, U.S. Department of Commerce. (2024). *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency.* https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

[32] Norwegian Consumer Council. (2023). *Ghost in the machine, addressing the consumer harms of generative AI*. https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf

[33] Pashentsev, E. (2023). The Malicious Use of Deepfakes Against Psychological Security and Political Stability. *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, 47–80. https://doi.org/10.1007/978-3-031-22552-9_3

[34] Schippers, B. (2020). Artificial Intelligence and Democratic Politics. *Political Insight*, *11*(1), 32–35. https://doi.org/10.1177/2041905820911746

[35] *UREDBA (EU) 2024/1689 EVROPSKEGA PARLAMENTA IN SVETA z dne 13. junija 2024 o določitvi harmoniziranih pravil o umetni inteligenci in spremembi uredb (ES) št. 300/2008, (EU) št. 167/2013, (EU) št. 168/2013, (EU) 2018/858, (EU) 2018/1139 in (EU) 2019/2144 ter direktiv 2014/90/EU, (EU) 2016/797 in (EU) 2020/1828 (Akt o umetni inteligenci).* (2024). Uradni list Evropske unije. https://eur-lex.europa.eu/legal-content/SL/TXT/PDF/?uri=OJ:L_202401689

[36] Vaccari, C. in Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+Society*, *6*(1), 1–13. https://doi.org/10.1177/2056305120903408

[37] Zuber N. in Gogoll J. (2024). Vox Populi, Vox ChatGPT: Large Language Models, Education and Democracy. *Philosophies*, *9*(1). https://doi.org/10.3390/philosophies9010013

[1] Vsi prevodi citatov iz neslovenskih virov: L. Košmrlj.

Razložljiva umetna inteligenca: kako naprej?

Explainable AI: What next?

Ana Farič[†]
Kognitivna znanost
Univerza v Ljubljani, Pedagoška fakulteta
Slovenija
af27987@student.uni-lj.si

Ivan Bratko
Umetna inteligenca
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Slovenija
bratko@fri.uni-lj.si

Povzetek

Prispevek povzema in ocenjuje stanje metod in raziskav na področju razložljive umetne inteligence. Pregled vsebuje predlagane definicije razlage in lastnosti dobrih razlag. Podan je grob pregled številnih obstoječih pristopov za generiranje razlage, primeri konkretnih avtomatsko generiranih razlag in nekatere empirične ugotovitve, kako uporabniki sprejemajo te razlage. Število raziskav na tem področju se je v zadnjih letih močno povečalo, pri čemer pa razni avtorji uporabljajo različne definicije in kriterije. Kljub veliki količini raziskav, so nekateri vidiki razložljivosti in tehnični pristopi deležni premalo pozornosti, med drugim: razlaga zaporedij odločitev, upoštevanje uporabnikovega predznanja ter induktivno logično programiranje.

Ključne besede

Umetna inteligenca, XAI, razložljivost

Abstract

The paper reviews and assesses the state of the art of research and methods in explainable AI. The review includes proposed definitions of what is an explanation, and what are properties of good explanations. We give a rough overview of numerous existing approaches for generating explanations, concrete examples of explanations and some empirical findings of their acceptance by users. The amount of research in this area has recently increased significantly, but different authors use different definitions and criteria. Despite numerous projects in this area, some aspects of explainability and technical approaches are receiving little attention: explaining sequences of decisions, taking into account user's background knowledge, and inductive logic programming.

1 Uvod

Modeli strojnega učenja postajajo z uspehom globokega učenja in nevronskih mrež vseprisotni. Večina od nas se z njimi srečuje na vsakodnevni ravni, v obliki sistemov za priporočanje glasbe in filmov npr. Taki sistemi brez posredovanja človeka izračunajo za nas najboljše priporočilo, morebitna neustrezna priporočila pa nimajo bistvenih (negativnih) posledic za nas. Nasprotno imajo lahko napačne odločitve v domenah (kot je npr. zdravstvo) odločilne posledice za konkretna življenja ljudi. Če v nekaterih domenah zadošča zgolj točna napoved sistema, to ne zadostuje povsod v družbi in znanosti nasploh [5].

Uporabnost modelov strojnega učenja je vodila v njihovo splošno uporabo pred razvojem kakovostnega konceptualnega okvirja, ki bi omogočal razumevanje njihovega delovanja. Znan je t. i. problem črnih škatel (ang. *black box problem*), ki pomeni, da delovanje modelov strojnega učenja ostaja za uporabnike nerazumljivo. Prav pomanjkanje razumevanja omejuje nadaljnjo in bolj praktično uporabo modelov v ostalih pomembnih domenah odločanja. Potreba po razlagi je vodila v razvoj tehnik in pristopov razložljive umetne inteligence (XAI; ang. *eXplainable Artificial Intelligence*), ki se posveča nalogi razlaganja kompleksnih modelov strojnega učenja [34].

Namen članka je pregled trenutnega stanja XAI področja in analiza pomanjkljivosti.

2 Kaj sploh je razlaga?

Razložljivost je izmuzljiv pojem ne samo na področju umetne inteligence (UI), pač pa širše na področju filozofije in drugih družboslovnih znanosti. Na področju UI se operira s koncepti, kot so vzročnost, informativnost, razumevanje, gotovost, zaupanje, transparentnost ipd. [5] Termin ‘razložljiva umetna inteligenca’ je l. 2019 kot del svojega programa uporabila DARPA [17]. Od takrat je postal zelo popularen, ne gre pa za nov pojav. Kvečjemu gre za imenovanje dolgoletnih prizadevanj, kjer se raziskovalci trudijo prebiti do odgovora na vprašanje, zakaj je sistem prišel do določene napovedi [19].

Najbolj splošno bi lahko razložljivost v domeni UI opredelili kot razlago, ki delovanje modela naredi bolj razumljivo. Seveda je to zelo splošna opredelitev, v poskusih bolj natančnega definiranja pa si raziskovalci niso zedinjeni. [12] opredelita razložljivost kot sposobnost predstaviti nekaj v človeku razumljivih terminih. [5] pravijo, da mora model nuditi razlago za svoje delovanje in napovedi v obliki vizualizacije pravil in vpogleda v potencialne spremenljivke, ki bi lahko povzročile perturbacije modela. Po [29] razložiti pomeni predstaviti besedilne ali vizualne elemente, ki omogočajo kvalitativno razumevanje odnosa med komponentami in napovedjo modela.

Ena od nekonsistentnosti v XAI literaturi je uporaba pojma interpretabilnost, ki je včasih sinonim razložljivosti, drugič ločen

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.14

pojem, tretjič ena od kategorij razložljivosti. [5] interpretabilnost razumejo kot pasivno, razložljivost pa kot aktivno lastnost modela. Interpretabilni so modeli, ki so razumljivi že sami po sebi (odločitvena drevesa npr.), razložljivi pa tisti, ki zahtevajo postopke, katerih namen je pojasnjevanje. Kot taka je razložljivost nujna lastnost vseh (tudi inherentno interpretabilnih) modelov [14].

Očitno je pomanjkanje konsenza o glavnih konceptih. Problem je, ker vsaka definicija nastopa znotraj specifičnega konteksta, odvisnega od naloge, sposobnosti in pričakovanj raziskovalca. Opredelitve razložljivosti so tako pogosto vezane na specifično domeno. Posledično XAI področje še ni enotno glede definicije razlage, specifičnih ciljev in kriterijev, ki naj bi jim zadostovali modeli, da bi bili razumljivi [5].

3 Metode razlag

Danes obstajajo številne metode razlag. Problem nastane pri njihovi klasifikaciji, ker ima praktično vsak avtor specifično definicijo razložljivosti, iz katere izhaja.

Ena splošnih kategorizacij je delitev na lokalne in globalne razlage. Lokalne so razlage, središčene okoli posameznega primera, kjer pa ostane delovanje modela kot celote nepojasnjeno. Na drugi strani globalne razlage pomagajo razumeti celoten model, so pa pogosto osnovane na približnih vrednostih [3][18][21][34].

Splošna je delitev na model-specifične in agnostične razlage. Slednje s tehnikami, kot so relevantnost atributov, vizualizacija ali simplifikacija pridobijo določene informacije o postopku napovedovanja in so uporabne za vsako vrsto modela [5]. Model-specifične razlage so uporabne zgolj za specifične vrste modelov (npr. maksimizacija aktivacije, ki jo opišejo [16]) [18].

[5] ločijo besedilne, vizualne, lokalne, razlage s primeri, s simplifikacijo in relevantnost atributov. [11] opredelijo tri glavne kategorije razlag: osnovane na funkciji, na primerih in pojasnjevanju atributov. [34] ločita razlage atributov in razlage primerov. [1] razlage razdelijo glede na uporabljeno metodologijo in ločijo med razlagami, ki slonijo na vzvratnem razširjanju (ang. *backpropagation*) in razlagami s perturbacijami. [16] ločijo: 1) odločitvena drevesa; 2) razlage, osnovane na pravilih; 3) razlage pomembnosti atributov, ki predstavijo težo in pomembnost atributov, ki jih je pri svoji napovedi upošteval model. Primer je znana metoda LIME, primerna predvsem za razlago klasifikacije besedil in slik (slika 2) [29]; 4) zemljevidi pomembnosti, ki izpostavijo ključne aspekte predmeta, ki je analiziran. Primer je metoda CAM (slika 1) [36]; 5) PDP (*Partial Dependence Plot*), kjer grafično prikažemo odnos med odločitvijo modela in vhodnimi podatki; 6) razlaga s prototipi, kjer z napovedjo dobimo primer, podoben našemu; 7) maksimizacija aktivacije, kjer opazujemo, kakšni vzorci vhodnih podatkov maksimizirajo aktivacijo določenega nevrona oz. nivoja.

[21] predstavi pojem formalne razložljivosti, zasnovan na logiki, kjer so razlage posledično bolj zanesljive in držijo globalno. Pristop temelji na računanji t. i. *prime implicants* (ang.), kar omogoča logične reprezentacije delovanja modela.

Slika 1 (levo): razlaga CAM metode na način prikaza področij slike, ključnih za klasifikacijo umivanja zob [36]. Slika 2 (desno): razlaga LIME metode. Na levi je izvorna slika, na desni razlaga za klasifikacijo električne kitare [29].

4 Kakšna je dobra razlaga?

Če je eden od ključnih ciljev XAI področja izboljšanje zaupanja v sisteme UI. je nujno, da se pozornost usmeri k uporabnikom teh sistemov [35]. Dobre razlage bodo tiste, ki bodo upoštevale, komu so namenjene [5]. To pomeni upoštevanje predznanja, ki ga imajo uporabniki. Opazen je trend, kjer razvijalci metod razlag tega ne upoštevajo dovolj. [30] opredelita tri skupine uporabnikov (razvijalci in raziskovalci, eksperti in laiki), ki zahtevajo različne vrste razlag.

Med raziskovalci ni strinjanja o kriterijih za dobro razlago. V nadaljevanju navajamo nekaj primerov kriterijev. [3] opredelita tri:

- Eksplicitnost: razlaga je takojšnja in razumljiva;
- Zvestoba (ang. *faithfulness*): ocene relevantnosti odražajo resnično pomembnost;
- Stabilnost: za podobne vhodne podatke veljajo podobne razlage.

[11] poudarjajo:

- Robustnost oz. občutljivost: sprememba razlage v primeru spremembe vhodnih podatkov;
- Zvestobo: razlaga ponazarja dejansko odločanje modela;
- Kompleksnost: kognitiven napor, potreben za razumevanje razlage;
- Homogenost: zmožnost razlage za pravilno razlago delovanja modela glede na različne skupine (v praksi se to po navadi nanaša na skupine, ki se ločijo glede na občutljive atribute).

[4] opredelita 4 aksiome, katerim naj bi zadostile dobre razlage:

1) morajo biti informativne;
2) ne smejo vsebovati nepotrebnih informacij;
3) razlage razredov morajo pojasniti posamezne primere, hkrati pa morajo biti splošno uporabne;
4) razlaga mora vsebovati samo informacije, ki vplivajo na napoved.

5 Ocenjevanje razlag

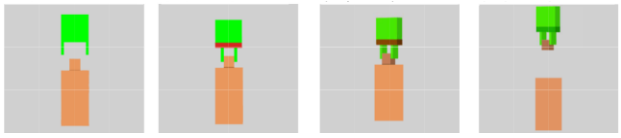
Ocenjevanje razlag je najmlajše področje s široko paleto pristopov [30]. Za razliko od točnosti, je kriterije kot so varnost, nediskriminacija in razložljivost težje kvantificirati [12].

Ocenjevanja se (najbolj splošno) lahko lotimo na dva načina: 1) človeško ocenjevanje ali 2) uporaba računskih metod, ki merijo, kako dobro razlaga dejansko razloži delovanje modela. Glavna razlika med pristopoma je, da so računske metode bolj objektivne, vendar pa ne upoštevajo človeškega faktorja.

Drugače rečeno, ne kvantificirajo človeškega razumevanja. Prednost človeške ocene je subjektivnost in večja deskriptivnost. Očitna pomanjkljivost je manjša točnost in večja odvisnost od specifične naloge [27].

[20][35] predstavijo matematično ocenjevanje razlag na podlagi analize robustnosti. Matematično opredeljena mera nezvestobe ponazarja, kako dobro se razlaga ujema z modelom [34].

[13] so izvedli eksperiment, s katerim so preverili, kakšne razlage so pri ljudeh vzbudile največ zaupanje v robota, ki je odprl stekleničko. Robot se je naučil odpirati stekleničke iz človeških demonstracij, pri čemer je bilo ključno učenje zaporedij položaja rok in potrebne sile. Z rokavico s senzorji so zajeli podatke o sili in položaju rok v 64 človeških demonstracijah s tremi različnimi stekleničkami. Sledilo je kompleksnejše učenje, da bi bil robot svoje znanje sposoben posplošiti. Implementiran je bil haptični model, ki je robotu pomagal določiti potrebno silo, čeprav nima človeških rok. Ker odpiranje stekleničke poteka v več korakih (potiskanje, odvijanje itd.), je bil implementiran še t. i. (ang.) *symbolic action planner* in pomeni pravila o zaporedju potrebnih akcij. S kombinacijo takega učenja je robot postal precej dober v odpiranju novih stekleničk. Udeleženci so bili razdeljeni v 5 skupin. Vsaka je videla posnetek robota, ki opravlja nalogo, ter eno od možnih razlag: 1) simbolično: v realnem času so udeleženci videli z eno besedo opisano akcijo, ki naj bi razlagala, kaj robot na posnetku dela (*approach – grasp – push – twist – ungrasp – move – grasp – push …*); 2) besedilno: po ogledu posnetka robota so udeleženci prebrali kratko besedilo o tem, kako je robot opravil nalogo (*I succeeded to open the bottle because I pushed on the cap three times and twisted the cap twice*); oz. 3) haptično razlago (slika 3): vizualizacija sile prijema v vsakem trenutku odpiranja stekleničke) oz. kombinacijo haptične in simbolične razlage. Največ zaupanja je spodbudila simbolična razlaga.



Slika 3: haptična razlaga.

[27] so izvedli eksperiment, kjer so udeleženci označili relevantna področja slike, ki je po njihovem mnenju bilo najbolj reprezentativno za določen razred objektov (mačka in pes npr.). Rezultat je zemljevid pomembnosti, ki prikazuje področja slike, ki so jim udeleženci posvečali največ pozornosti (spodnja vrsta na sliki 4). Te rezultate so primerjali z zemljevidi pomembnosti metode Grad-CAM (spodnja vrsta na sliki 4). Zemljevidi so si morda podobni, vseeno pa je statistično testiranje pokazalo pomembne razlike. Distribucija relevantnih atributov je bila pri Grad-CAM metodi bolj uniformna, udeleženci so v primeru živih bitij kot ključne bolj označevali obraze. Prav to so ugotovitve, ki nam lahko pomagajo razumeti, kako dobre so razlage.

Slika 4: zgornja vrsta prikazuje zemljevida pomembnosti Grad-CAM metode, spodnja zemljevida udeležencev [27].

6 Kako naprej?

V tem razdelku opozorimo na nekatere razmeroma slabo raziskane probleme in premalo uporabljene pristope za XAI.

6.1 Tehtanje med točnostjo in razložljivostjo

[31] v članku z zgovornim naslovom »*Stop explaining black box ML models for high stakes decisions and use interpretable models instead*« izraža determinirano stališče. Zavzema se za uporabo metod učenja, ki dajejo naučene modele, ki so sami po sebi razumljivi. Za take se smatrajo npr. odločitvena drevesa. Nasprotuje metodam učenja, katerih rezultati so v principu težko razumljivi. Med te štejemo posebno metode globokega učenja, ki sicer dosegajo visoko napovedno točnost v primerjavi z drugimi metodami učenja, toda ne zastonj: vsaj za ceno razumljivosti in potrebnega velikega števila podatkov za učenje. Pri tem gre Rudin morda res predaleč s svojim optimističnim stališčem, ki implicitno predpostavlja možnost izgradnje elegantnih in razumljivih modelov za vsako problemsko domeno, s čimer zadene ob princip kompleksnosti Kolmogorova.

Glede možnosti obstoja enostavnih modelov in razlag velja vsaj ena teoretična omejitev, ki jo definira kompleksnost Kolmogorova, ki določa, koliko spominskega prostora potrebujemo za najkrajši možni zapis danega objekta v računalniku. Obstajajo zapleteni objekti (torej tudi napovedni modeli), ki jih niti teoretično ni mogoče predstaviti na kratek način. V takih primerih tudi razlaga ne more biti kratka in enostavna. Res pa je, da smo v praksi še zelo daleč od te teoretično dosegljive meje, torej imamo veliko prostora za izboljšanje. Ko zadenemo ob zid Kolmogorova, pa je še vedno možen kompromis, da za boljšo razložljivost žrtvujemo nekaj točnosti [8]. Primer tehnične izvedbe tega tehtanja med točnostjo in razumljivostjo v učenju odločitvenih dreves je [6].

6.2 Navezava razlage na uporabnikovo predznanje

Če bo razlaga dobra, je odvisno od njenega uporabnika, konkretno od uporabnikovega predznanja o problemski domeni. Če je to kvalitetno, zadošča en sam namig. Če je razumevanje domene slabo, je potrebna podrobna in daljša razlaga. Tudi sama formulacija razlage je odvisna od obstoječega znanja na obravnavanem področju. Celo povsem pravilna in jedrnata razlaga je za eksperta na področju uporabe lahko nesprejemljiva in nenaravna. Kot primer omenimo, da so se nekateri primeri razlag, ki jih generirajo naučeni modeli v medicinskih domenah kljub svoji diagnostični točnosti zdravniku zdeli povsem

nenaravni [8]. V enem od primerov je sistem razložil, da gre za vnetni revmatizem, ker ima pacient med drugim več kot dva prizadeta sklepa na roki. To diagnostično pravilo je dejansko točno. Vendar pa je zdravnik vztrajal, da mora imeti pacient prizadete sklepe na vseh petih prstih na roki, ker vnetni revmatizem tipično vpliva na vse sklepe. Ekspertno mnenje je bilo v tem primeru zelo jasno, čeprav je res, da bo pravilo vodilo do pravilne diagnoze v vsakem primeru; če gre za katerokoli število vnetih sklepov med 2 in 5. Ustreznost razlage je odvisna ne le od klasifikacijske točnosti, temveč (tudi) od predznanja, ki ga ima uporabnik o tej obliki revmatizma.

Obstoječe metode razlage ta vidik pogosto ignorirajo. Problem je tudi v tem, da ne omogočajo naravne uporabe predznanja. V tem pogledu je zelo obetaven pristop k strojnemu učenju t. i. induktivno logično programiranje (ILP), ki temelji na uporabi matematične logike. Že sama osnovna formulacija problema učenja v ILP vsebuje uporabo predznanja: dani so učni primeri E in predznanje BK (*background knowledge*), naloga učenja pa je sestaviti logično formulo H (hipoteza) tako, da primeri E logično sledijo iz BK in H.

Pristop ILP je skromno zastopan v obstoječih raziskavah iz strojnega učenja in razložljivosti. Lep primer njegove ustreznosti so raziskave, opisane v [2][28]. Te zasledujejo ne le osnovni cilj XAI (razlage odločitev strojnega učenja), temveč tudi cilj t. i. »ultra-razložljivost«. Ta strožji kriterij strojnega učenja je definiral [26] (ang. *ultra strong criterion for ML*). Strojno učenje je ultra-razložljivo, če je ne le razložljivo, temveč uporabniku omogoča tudi *operativno uporabo za lastno reševanje* novih problemov. Npr. da strojno naučeno znanje lahko uporabi za lastno reševanje določenih matematičnih problemov ali igranje šaha.

6.3 Razlaga zaporedij odločitev v planiranju

Večina XAI metod generira razlago *posameznih* odločitev oz. klasifikacij. Pri razložljivem planiranju pa gre za razlago množice odločitev (npr. zaporedja akcij, ki robota vodi do cilja). Posebej za razlago planov se je formiralo področje razložljivega planiranja [10].

Razlaga planov je navadno zahtevnejša od razlage v klasifikacijskih problemih. Treba je razložiti, kako so posamezne akcije odvisne od drugih, da skupaj rešijo nalogo. Primer razlage zaporedja odločitev je razlaga šahovskih partij, kjer je treba razložiti celo zaporedje potez ali drevo odločitev, ki definira uspešno strategijo. Primer, opisan v [9], so težko razumljive in briljantne poteze šahovskega programa AlphaZero.

Razlaga planov je aktualna tudi na področju vodenja sistemov. Lep primer razlage naučenega plana vodenja je v [32]. Gre za klasično nalogo iz teorije vodenja sistemov: vodenje sistema voziček-palica. Na vozičku je vrtljivo vpeta palica. Palica je postavljena približno vertikalno, vendar se, če ne ukrepamo, prevrne na tla. S potiskanjem vozička levo oz. desno je treba loviti ravnotežni položaj palice okrog vertikale, obenem pa doseči, da se voziček horizontalno premakne iz začetnega položaja do cilja. Naučena strategija vodenja je lepo razložljiva. Najprej nekoliko presenetljivo potisnemo voziček stran od cilja, s čimer dosežemo, da se palica nagne proti cilju. To omogoči potiskanje vozička v smer proti cilju, hkrati pa se ohranja ravnotežje palice, ko je ta nagnjena naprej v smeri cilja.

7 Zaključek

Področje XAI se je v zadnjih 5 do 10 letih močno razraslo. Mnogi zato predpostavljajo, da je bil to tudi začetek področja. V resnici je aktivno zavedanje, da naj bi bilo strojno učenje razložljivo, obstajalo že prej 40 leti. Že takrat so obstajale raziskave o razložljivih modelih. Kljub sedanji količini raziskav in nedvoumnih uspehih se še vedno kaže, da pogrešamo nekatere ključne odgovore. Npr., že pred desetletji se je v sklopu istih prizadevanj pojavilo zavedanje, da potrebujemo formalne mere za ocenjevanje kvalitete razlag. Take sprejete mere še ni. Raziskovalci pri ocenjevanju razlag uberejo različne pristope, odvisne od raznih kriterijev (konteksta, domene, uporabnikov itd.).

Glede vprašanja, kaj je sprejemljiva razlaga, se v pomanjkanju boljših splošnih in principielnih kriterijev v sedanji praksi uporablja predpostavka, da so nekateri modeli razložljivi kar po definiciji, torej razložljivi sami po sebi. Mednje npr. navadno štejemo odločitvena drevesa ali pravila če-potem. Toda tudi ta kriterij je arbitraren. Kaj, če je odločitveno drevo zelo veliko, npr. da ima milijon vozlišč?

V prispevku smo opozorili tudi na počasen napredek pri razvoju metod za razlago zaporedij odločitev. Sem sodi razlaga planov za reševanje nalog, ki imajo eksplicitno definirane cilje. Plan je lahko zaporedje akcij ali pa tudi množica akcij, ki so delno urejene v času. Tu je treba razložiti tudi to, kako se akcije med seboj dopolnjujejo in na kakšen način skupaj dosežejo cilj. S tem so povezani izzivi, ki jih predstavimo spodaj.

En možen pristop, ki upošteva principe planiranja v UI, je upoštevanje odvisnosti med akcijami. Nekatere akcije v planu neposredno dosežejo kakšnega od ciljev plana. Druge akcije pa ne dosežejo nobenega danega cilja neposredno, njihova funkcija je, da dosežejo pogoje, ki morajo biti uresničeni, da je možno izvesti druge akcije v planu. Taka razlaga plana je seveda povsem logična. Navadno pa vsebuje preveč podrobnosti. Če plan vsebuje nekoliko večje število akcij, npr. nekaj 10, postane tako podrobna razlaga spet težko razumljiva in za uporabnika neprivlačna. V tem primeru bi za sprejemljivo razlago treba plan razbiti v hierarhično strukturo, definirano s podcilji plana. Odkrivanje smiselnih podciljev pa je lahko težavno. Poseben izziv je, kako poiskati take podcilje, ki rezultirajo v razlagi, ki je za človeka čim bolj naravna.

Zahvala

Prispevek je nastal v okviru ciljnega raziskovalnega projekta V2-2272 Opredelitev okvira za zagotavljanje zaupanja javnosti v sisteme umetne inteligence in njihove uporabe, ob podpori Javne agencije za raziskovalno in inovacijsko dejavnost Republike Slovenije in Ministrstva za digitalizacijo.

Literatura

[1] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N. and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion, 99*. DOI: https://doi.org/10.1016/j.inffus.2023.101805.

[2] Ai, L., Langer, J., Muggleton, S. H. and Schmid, U. 2023. Explanatory machine learning for sequential human teaching. *Machine Learning Journal, 112*, 3591-3632. DOI: https://doi.org/10.1007/s10994-023-06351-8.

[3] Alvarez-Melis, D and Jaakkola, T. S. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. 32nd Conference

on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.

[4] Amgoud, L. and Ben-Naim, J. 2022. Axiomatic Foundations of Explainability. Proceedings of the Thirty-First International joint Conference on Artificial Intelligence (IJACAI-22), 636-642.

[5] Barredo Arrieta, A., Diaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gill-López, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv: 1910.10045v2*. DOI: https://doi.org/10.48550/arXiv.1910.10045.

[6] Bohanec, M. and Bratko, I. 1994. Trading accuracy for simplicity in decision trees. *Machine Learning Journal, 15*, 223-250. DOI: https://doi.org/10.1007/BF00993345.

[7] Brasse, J., Broder, H. R., Förster, M., Klier, M. and Sigler, I. 2023. Explainable artificial intelligence in information systems: A Review of the status quo and future research directions. *The International Journal of Networked Business, 33*(1). DOI: https://dx.doi.org/10.1007/s12535-023-00644-5.

[8] Bratko. I. 1997. *Machine learning: between accuracy and interpretability*. V: Learning, Networks and Statistics (ed. Della Riccia, G.), Vienna: Springer.

[9] Bratko, I. 2018. AlphaZero: what's missing? *Informatica, 42*(1).

[10] Chakraborti, T., Sreedharan, S. and Kambhampati, S. 2020. The Emerging Landscape of Explainable Automated Planning & Decision Making. *arXiv:2002.11697*. DOI: https://doi.org/10.48550/arXiv.2002.11697.

[11] Chen, Z., Subhash, V., Havasi, M., Pan, W. and Doshi-Velez, F. 2022. What Makes a Good Explanation?: A Harmonized View of Properties of Explanations. *Progress and Challenges in Building Trustworthy Embodies AI (TEA 2022) co-located with NeurIPS 2022.*

[12] Doshi-Velez, F. and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608*. DOI: https://doi.org/10.48550/arXiv.1702.08680.

[13] Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y. N., Lu, H. and Zhu, S. C. 2019. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics, 4*(3), 1-13.

[14] Gilpin, L. H., Yuan, B. Z., Bajwa, A., Specter, M and Kagal, L. 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00068*. DOI: https://doi.org/10.48550/arXiv.1806.00069.

[15] Grice, H. P. 1975. Logic and conversation, syntax and semantics. *Speech Acts 3*, 41-58.

[16] Guidotti, D., Monreale, A., Ruggieri, S. and Turini, F. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys, 51*(5), 1-42. DOI: https://doi.org/10.1145/3236009.

[17] Gunning, D., Vorm, E., Wang, J. Y. and Turek, M. 2021. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters, 2*(4). DOI: https://doi.org/10.1002/ail2.61.

[18] Hall, P., Ambati, S. and Phan, W. 15.3.2017. *Ideas on interpreting machine learning.* O'Reilly. https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/.

[19] Holsinger, A., Saranti, A., Molnar, C., Biecek, P. and Samek, W. 2022. Explainable AI Methods – A Brief Overview. xxAI 2020, LNAI 13200, 13-38. DOI: https://doi.org/10.1007/978-3-031-04083-2_2.

[20] Hsieh, C. Y., Yeh, C. K., Liu, X., Ravikumar, P., Kim, S., Kumar, S. and Hsieh, C. J. 2021. Evaluations and Methods for Explanation through Robustness Analysis. *arXiv:2006.00442*. DOI: https://doi.org/10.48550/arXiv.2006.00442.

[21] Ignatiev, A., Narodytska, N. and Marques-Silva, J. 2019. On Validating, Repairing and Refining Heuristic ML Explanations. *arXiv: 1907.02509.* DOI: https://doi.org/10.48550/arXiv.1907.02590.

[22] Kolmogorov, A. 1963. On Tables of Random Numbers. *The Indian Journal of Statistics, Series A, 25*, 369-375.

[23] Krishnan, M. 2020. Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology, 33*(3), 487-502. DOI: http://dx.doi.org/10.1007/s13347-019-00372-9.

[24] Lombrozo, T. 2006. The structure and function of explanation. *Trends in Cognitive Science, 10*(10), 464-470. DOI: https://doi.org/10.1016/j.tics.2006.08.004.

[25] Malle, B. F. 2004. How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction. MIT Press.

[26] Michie, D. 1998. Machine Learning in the next five years. *Proceedings of the 3rd European working session on learning*, 107-122.

[27] Mohseni, S., Block, J. E. and Ragan, E. 2021. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. *26th International Conference on Intelligent User Interfaces (IUI'21)*. DOI: https://doi.org/10.1145/3397481.3450689.

[28] Muggleton, S., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A. and Besold, T. 2018. Ultra-strong machine learning: Comprehensibility of programs learned with ILP. *Machine Learning, 107*, 1119-1140. DOI: https://doi.org/10.1007/s10994-018-5707-3.

[29] Ribeiro, M. T., Singh, S. and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. DOI: https://doi.org/10.1145/2939672.2939778.

[30] Ribera, M. and Lapedriza, A. 2019. Can we do better explanations? A proposal of User-Centered Explainable AI. *IUI Workshops '19.*

[31] Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. DOI: https://doi.org/10.1038/s42256-019-0048-x.

[32] Šoberl, D. and Bratko, I. 2023. Transferring a Learned Qualitative Cart-Pole Control Model to Uneven Terrains. *International Conference on Discovery Science*, 446-459. DOI: http://dx.doi.org/10.1007/978-3-031-45275-8_30.

[33] Tjoa, E. and Guan, C. 2020. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *arXiv: 1907.07374.* DOI: http://doi.org/10.48550/arXiv.1907.07374.

[34] Yeh, C. K. and Ravikumar, P. 2021. Objective criteria for explanations of machine learning models. Applied AI Letters published by John Wiley & Seons Ltd. DOI: https://doi.org/10.1002/ail2.57.

[35] Zhang, Z., Xu, L., Yilmaz, L. and Liu, B. 2021. A Critical Review of Inductive Logic Programming Techniques for Explainable AI. *arXiv: 2112.15319*. DOI: https://doi.org/10.48550/arXiv.2112.15319.

[36] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. 2016. Learning Deep Features for Disciminative Localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921-2929. DOI: 10.1109/CVPR.2016.319.

Exploring Cognitive Science under Analytical Idealism

Grega Rodman
Faculty of Education
University of Ljubljana
gr6506@student.uni-lj.si

Abstract

In modern science, materialism has played a significant role, positing that matter is the fundamental reality and that all phenomena, including consciousness, can be understood through physical processes. However, recent evidence suggests materialism might not fully explain all phenomena. These findings have led to the rise of a post-materialistic movement exploring new ideas. One such idea, Analytical Idealism, proposed by Bernardo Kastrup, suggests that consciousness is the fundamental reality and that the material world is a reflection of this universal consciousness. The implications of adopting this approach in science will be explored.

Keywords

ontology, methodology, materialism, analytical idealism

1 Introduction

The modern scientific worldview is largely based on assumptions closely linked to classical physics. Among these is materialism, which posits that matter constitutes the sole reality. In the 19th century, these assumptions became increasingly rigid, evolving into dogmas that coalesced into the ideological framework known as "scientific materialism" [1].

Scientific materialism is a philosophical viewpoint that asserts that all phenomena in the universe, including consciousness and human experience, can be explained solely through physical processes and interactions. Throughout the 20th century, scientific materialism became the prevailing ideology in academic circles, to the extent that the majority of scientists came to believe it was the only rational interpretation of the world. Scientific methods rooted in materialistic philosophy have proven highly successful in enhancing our understanding of nature and in providing greater control and freedom through technological advances. Though the popularity of scientific materialism is waning, the legacy persists, having limited the scope of inquiry, particularly in the study of consciousness, by ignoring subjective human experience [1, 2].

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.15

2 Questioning the materialistic paradigm

At its core, science is a non-dogmatic, open-minded approach to acquiring knowledge about nature through observation, experimental investigation, and theoretical explanation of phenomena [3]. There is a misconception that the methodology of science is inherently tied to materialism. In addition, an increasing body of empirical evidence points to the limitations of materialism. Of course, it is impossible to provide sufficient empirical research that definitively refutes materialism; however, we can present two examples from different fields that suggest the limitations of materialism: one from animal cognition and the second from psi phenomena in humans.

2.1 Example from animal cognition

The first example is from animal cognition. Actually, it is about precognition, which is the perception of future events, typical for some animal species. Investigation in this field was done by Sheldrake [4], who studied a dog that seemed to know when its owner was coming home. Despite using various methods to rule out normal senses, Sheldrake consistently observed the dog waiting expectantly before the owner arrived, but not at other times. A replication of a similar experiment by some sceptics was declared unsuccessful [5], but a later reanalysis of the same results showed the opposite [6].

2.2 Example from psi phenomena in humans

The second example addresses meta-analyses of psi phenomena, which are defined as extraordinary human capacities like telepathy, clairvoyance, and precognition that involve gaining information without known sensory mechanisms. Studies investigating these phenomena have consistently found small but significant effects, suggesting that such abilities may exist [7]. The evidence for psi is comparable to that for established phenomena in psychology and other disciplines, although there is no consensual understanding of them. Recent analyses also emphasize that these results cannot be easily attributed to methodological flaws, selective reporting, or fraud, further supporting the plausibility of psi phenomena.

The volume of empirical data indicating the shortcomings of materialism is so substantial that an increasing number of articles and books are being written on this subject [1, 7, 8, 9]. In fact, this has contributed to the emergence of a whole post-materialistic movement in recent decades, which is exploring what this new paradigm might look like [3, 9, 10, 11, 12]. Believe it or not, you can also find a manifesto for post-materialistic science [13].

3 Cognitive Science under Analytical Idealism

One of the proponents of the post-materialistic movement is Bernardo Kastrup, who advocates for Analytical Idealism [3]. Analytical Idealism posits that consciousness is the fundamental essence of reality, rather than matter [12]. The focus of this summary is not to provide a detailed description of Idealism, but rather to explore the potential changes in the methodology of scientific research that could result from adopting this perspective.

3.1 Two distinct routes to knowledge

Changes in ontological views lead to changes in scientific methods as they alter the foundational concepts and relationships that guide inquiry [14]. If consciousness is indeed a fundamental aspect of reality rather than a byproduct of neural activity, it implies that consciousness might directly access aspects of reality without relying solely on sensory perception [15]. This leads us to consider two distinct routes to knowledge: conventional sensory perception (science as it is mainly now) and a more direct introspective approach. Walach calls this approach “radical introspection.” Radical introspection involves a deep inward focus, often achieved through contemplative and meditative practices. Unlike standard qualitative introspection, which relies on external referents (e.g., transcripts, observations), radical introspection does not have such referents beyond personal experience. It faces challenges of subjective bias and lack of established methodology for validating truth claims. However, it remains a crucial aspect of potential new methodologies in science, requiring the development of techniques to record, communicate, and verify first-person experiences.

At this point, it is important to highlight that Walach is not the only proponent of integrating radical introspection into scientific inquiry. Kordeš [16] arrives at a similar conclusion in his arguments, even though he does not refer the concept of idealism at all. He suggests that in-depth, existentially liable introspection and self-inquiry should be considered as serious scientific research tools.

3.2 Combining first and third person research

When looking at current scientific practices, we can see some early attempts in that direction. The godfather of this approach is, of course, Francisco J. Varela [17]. From this approach emerged the field of contemplative neuroscience, which explores individuals in altered states of consciousness that develop through various contemplative practices. This field seeks to integrate traditional third-person scientific methods, such as MRI, EEG, and MEG, with first-person accounts of personal experiences in these altered states of consciousness [18, 19, 20]. When we start taking contemplative and meditative practices seriously, science can begin to exchange ideas with ancient traditions such as Buddhism, Hinduism and others. Even this is already happening [21, 22, 23].

3.3 The consequences of such a research approach

This interdisciplinary exchange highlights the potential for scientific and spiritual perspectives to enrich each other and expand our understanding of consciousness and reality. Additionally, to broaden scientific inquiry, spiritual practices like meditation and contemplation can be secularized and incorporated into the scientific process. Fun fact, At the 6th International Colloquium of Cognitive Sciences, Dr. Berkovich-Ohana began her presentation titled "Meditation and the Self: Neuroscience and Phenomenology" with a few minutes of guided meditation [24]. By integrating these practices, scientists could benefit from improved mental hygiene, enhanced creativity, and increased cognitive capacities [25], [26], [27].

This step can be highly significant, as it enhances the performance of researchers. A greater focus may lead to reduced bias, while increased creativity fosters better hypotheses, ultimately resulting in more effective research. Such advancements are essential for achieving substantial breakthroughs.

Engaging in meditation and/or contemplative practices poses potential downsides for scientists, too. First, the focus on personal experience conflicts with the concurrent objective standards required in scientific research. While self-research can yield valuable insights, its subjective nature can lead to biases that undermine intersubjectivity. Furthermore, the personal transformation that occurs during deep self-reflection may distract researchers from maintaining the rigorous, detached perspective typically expected in scientific inquiry. Ultimately, the integration of such practices into mainstream science remains challenging, as it contrasts with the traditional role of researchers.

4 Conclusion

In conclusion, I would like to emphasize a few key points. First, the entire described methodology can, of course, be applied from a materialistic standpoint as well. It is not the ontology itself that matters; rather, it is the methodology that enables insight. Materialists can also engage in contemplative neuroscience. Second, year by year, we have more scientific studies suggesting that the current mainstream paradigm may be flawed. Let us carefully examine the data and avoid dismissing it simply because it contradicts our preconceived assumptions [28]. Third, if more scientists were to engage in meditation-like practices, this would generally benefit the scientific community for reasons previously discussed. Fourth, when we establish a connection between science and religion, mutual learning can begin.

Acknowledgments

I would like to thank Olga Markič for giving me the opportunity to explore this area of research and for mentoring me throughout the process.

Authors' statement

ChatGPT-4 was used for improving the language of this paper.

References

[1] B. Alan Wallace. 2004. The Taboo of Subjectivity: Toward a New Science of Consciousness. Paperback, 323 pages. ISBN: 9780195173109.

[2] Frank, Adam, Marcelo Gleiser, and Evan Thompson. 2024. The Blind Spot: Why Science Cannot Ignore Human Experience. Cambridge, Massachusetts: The MIT Press. Library of Congress Cataloging-in-Publication Data.

[3] Mario Beauregard, 2018. Toward a postmaterialist psychology: Theory, research, and applications. New Ideas in Psychology, 50, 21–33. doi: 10.1016/j.newideapsych.2018.02.004.

[4] Rupert Sheldrake and Pamela Smart. 2000. A dog that seems to know when his owner is coming home: Videotaped experiments and observations. Journal of Scientific Exploration, 14(2), 233–255.

[5] Richard Wiseman, Matthew D. Smith, and Julie Milton. 1998. Can animals detect when their owners are returning home? An experimental test of the 'psychic pet' phenomenon. British Journal of Psychology, 89(3), 453–462. doi: 10.1111/j.2044-8295.1998.tb02696.x.

[6] Rupert Sheldrake. 2011. Dogs That Know When Their Owners Are Coming Home: And Other Unexplained Powers of Animals. Three Rivers Press, New York.

[7] Cardeña, Etzel. 2018. "The Experimental Evidence for Parapsychological Phenomena: A Review." American Psychologist 73(5): 663–677. https://doi.org/10.1037/amp0000236.

[8] Charles Whitehead. [2019]. Charles Tart: The End of Materialism: How Evidence of the Paranormal is Bringing Science and Spirit Together (book review). Journal of Consciousness Studies, 17(11/12):202.

[9] Harald Walach. 2019. Beyond a Materialist Worldview: Towards an Expanded Science. Lulu.com. ISBN: 978-1716805998.

[10] Gary E. Schwartz. 2016. What is the nature of a post-materialist paradigm? Three types of theories. Explore. doi: 10.1016/j.explore.2015.12.002.

[11] Frederick T. Travis. 2020. Consciousness is primary: Science of consciousness for the 21st century. International Journal of Psychological Studies, 13(1), 1. doi: 10.5539/ijps.v13n1p1.

[12] Bernardo Kastrup. 2018. The next paradigm. Future Human Image, 9. doi: 10.29202/fhi/9/4.

[13] Mario Beauregard, Gary E. Schwartz, Lisa Miller, and Larry Dossey. 2014. Manifesto for a post-materialist science. Explore, 10(5). doi: 10.1016/j.explore.2014.06.008.

[14] Mukhles Al-Ababneh. 2020. Linking Ontology, Epistemology and Research Methodology. Science & Philosophy, 8(1), 75–91. doi: 10.5311/112222001789.

[15] Harald Walach. 2020. Inner Experience – Direct Access to Reality: A Complementarist Ontology and Dual Aspect Monism Support a Broader Epistemology. Frontiers in Psychology, 11, Article 640. doi: 10.3389/fpsyg.2020.00640.

[16] Urban Kordeš. [2013]. Problems and Opportunities of First-Person Research. *Interdisciplinary Description of Complex Systems*, 11(4):363-375. https://doi.org/10.7906/indecs.11.4.2.

[17] Varela, Francisco J. 1996. "Neurophenomenology: A Methodological Remedy for the Hard Problem." Journal of Consciousness Studies 3(4): 330–349.

[18] Mario Beauregard and Vincent Paquette. 2006. Neural correlates of a mystical experience in Carmelite nuns. Neuroscience Letters, 405(3), 186–190. doi: 10.1016/j.neulet.2006.06.060.

[19] Yair Dor-Ziderman and Aviva Berkovich-Ohana. 2013. Mindfulness-induced selflessness: A MEG neurophenomenological study. Frontiers in Human Neuroscience, 7, 582. doi: 10.3389/fnhum.2013.00582.

[20] Mario Beauregard and Vincent Paquette. 2006. Neural correlates of a mystical experience in Carmelite nuns. Neuroscience Letters, 405(3), 186–190. doi: 10.1016/j.neulet.2006.06.060.

[21] Markič, Olga, and Urban Kordeš. 2016. "Parallels between Mindfulness and First-Person Research into Consciousness." Asian Studies 4(2):153-168. https://doi.org/10.4312/as.2016.4.2.153-168.

[22] Federman, Asaf. 2011. "What Buddhism Taught Cognitive Science About Self, Mind and Brain." Enrahonar 47:39-62. https://doi.org/10.5565/rev/enrahonar/v47n0.162.

[23] Vörös, Sebastjan. 2016. "Buddhism and Cognitive (Neuro)Science: An Uneasy Liaison?" Asian Studies 4(1):61-80. https://doi.org/10.4312/as.2016.4.1.61-80.

[24] Aviva Berkovich-Ohana. 2022. 6to Coloquio Internacional de Ciencias Cognitivas. YouTube video. Retrieved August 26, 2024, from https://www.youtube.com/watch?v=gJHSSZikyvQ.

[25] Richard J. Davidson and Jon Kabat-Zinn. 2003. Alterations in brain and immune function produced by mindfulness meditation. Psychosomatic Medicine, 65(4), 564–570. doi: 10.1097/01.PSY.0000077505.67574.E3.

[26] Eileen Luders and Christian Gaser. 2008. The underlying anatomical correlates of long-term meditation: Larger hippocampal and frontal volumes of gray matter. NeuroImage. doi: 10.1016/j.neuroimage.2008.12.061.

[27] Amishi P. Jha, Jason Krompinger, and Michael J. Baime. 2007. Mindfulness training modifies subsystems of attention. Cognitive, Affective, & Behavioral Neuroscience, 7, 109–119. doi: 10.3758/CABN.7.2.109.

[28] Williams, Bryan J. 2019. "Reassessing the 'Impossible': A Critical Commentary on Reber and Alcock's 'Why Parapsychological Claims Cannot Be True'." Journal of Scientific Exploration 33(4):599-616. https://doi.org/10.31275/2019/1667.

Intelligent Revolution – a New Civilization and Cognitive Era

Inteligenčna revolucija – nova civilizacijska in kognitivna doba

Matjaž Gams
Odsek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
Slovenija

ABSTRACT.
The rapid advancement of artificial intelligence is significantly enhancing human capabilities, even as human progress itself appears to stagnate, hindered by decadent ideologies and adverse societal trends. Over the past few decades, AI has achieved remarkable milestones, from mastering complex games to revolutionizing industries such as healthcare and finance through advancements in machine learning, natural language processing, and robotics. A particularly notable achievement is the development of GPT models, which have set new standards in language generation and expanded the horizons of AI's potential. This paper examines the impact of AI on various sectors, including societal and individual cognitive advancements, highlighting both the opportunities and challenges of widespread AI adoption. The discussion focuses on the transformative power of AI technologies and the ethical, economic, cognitive, and social implications of this ongoing revolution. As AI continues to drive innovation and transform industries, humans are increasingly integrating with these technologies through the pervasive use of smartphones, personal computers, and wearable devices. This integration has already enhanced our cognitive and functional capabilities, effectively amplifying human potential. However, as AI's influence on human life deepens, critical questions arise about the future of this symbiotic relationship and the trajectory of societal progress.

POVZETEK
Hitri napredek umetne inteligence izboljšuje človeške zmožnosti, medtem ko se zdi, da človeški napredek stagnira, oviran z dekadentnimi ideologijami in negativnimi družbenimi trendi. V zadnjih desetletjih je UI dosegla izjemne mejnike, od obvladovanja kompleksnih iger do revolucije v panogah, kot sta zdravstvo in finance, z napredki na področju strojnega učenja, obdelave naravnega jezika in robotike. Posebej pomemben dosežek je razvoj modelov GPT, ki so postavili nove standarde v generiranju jezika in razširili obzorja potenciala UI. Ta članek preučuje vpliv UI na različne sektorje, vključno z družbenim in individualnim kognitivnim napredkom, ter poudarja tako priložnosti kot izzive, ki jih prinaša široka uporaba UI. Razprava se osredotoča na transformativno moč UI tehnologij ter na etične, ekonomske, kognitivne in družbene posledice te tekoče revolucije.

Medtem ko UI še naprej spodbuja inovacije in preoblikuje industrije, se ljudje vse bolj integrirajo s temi tehnologijami prek vseprisotne uporabe pametnih telefonov, osebnih računalnikov in nosljivih naprav. Ta integracija je že okrepila naše kognitivne in funkcionalne sposobnosti ter učinkovito pomnožila človeški potencial. Ko vpliv UI na človeško življenje narašča, se pojavljajo ključna vprašanja o prihodnosti tega simbiotičnega razmerja in o smeri družbenega napredka.

KLJUČNE BESEDE
Umetna inteligenca, simbioza med človekom in strojem, tehnološki napredek, kognitivni vpliv

KEYWORDS
Artificial Intelligence, Human-Machine Symbiosis, Technological Advancement, Cognitivel Implications

1 Introduction

The rapid development of artificial intelligence (AI) is transforming industries and redefining/improving the essence of human capabilities. While AI progresses at an unprecedented pace, human societal progress appears to be stagnating, increasingly entangled in decadent ideologies and negative societal trends. This duality between AI's rise and human inertia is critical to understanding the current technological landscape. Over recent decades, AI has achieved remarkable milestones, including mastering complex games like Go and chess, advancing natural language processing (NLP), and driving significant innovations in sectors such as healthcare and finance. For example, the development of GPT (Generative Pre-trained Transformer) models represents a breakthrough in AI's ability to

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.16

generate human-like text, setting new standards for machine language generation and understanding [1, 2].

This paper examines the transformative potential of AI across various sectors, including healthcare, finance, education, and entertainment, highlighting both the opportunities and challenges that accompany widespread AI adoption. In recent years, significant advancements in AI have occurred at an accelerating rate. For example, breakthroughs in reinforcement learning and unsupervised learning have expanded the capabilities of AI systems, with applications ranging from autonomous vehicles to sophisticated recommendation systems [3]. Additionally, the ethical, economic, cognitive, and social implications of AI's proliferation are increasingly coming to the forefront, as debates intensify over issues such as algorithmic bias, privacy, and the potential for AI to displace human jobs [4]. These discussions underscore the need for robust governance frameworks to ensure that AI technologies are developed and deployed responsibly [5].

As AI technologies like machine learning, NLP, and robotics continue to evolve, they increasingly integrate into human life, augmenting cognitive and functional abilities through ubiquitous technologies such as smartphones, personal computers, and wearables. This integration, often referred to as a form of human-AI symbiosis, has already multiplied human potential, enabling tasks and processes that were previously unimaginable [6]. The consequences of this symbiotic relationship are profound, raising critical questions about the direction of societal progress and the future of humanity as AI [7, 8] becomes more embedded in everyday life.

We analyze the implications of AI's rapid development, particularly the considerations that must be addressed to navigate the ongoing AI revolution effectively. By integrating recent scholarly insights with a broader analysis of AI's impact, this paper seeks to contribute to the understanding of how AI is reshaping industries and human capabilities, as well as the future trajectory of this unprecedented technological evolution.

2 AI progress

A recent and transformative achievement is the development of Generative Pre-trained Transformers. These models represent a leap forward in natural language processing, capable of generating human-like text, translating languages, and even writing code. The GPT-3 model, released by OpenAI in 2020, is particularly notable for its ability to perform a wide range of tasks with minimal input, showcasing the power and versatility of large-scale language models [9].

Here we present one major achievement over the last five years, having in mind the constant AI progress in areas like autonomous driving or pattern recognition:

2019: AlphaStar in Real-Time Strategy Games

In 2019, DeepMind's AlphaStar achieved a significant milestone by reaching the top players of professional StarCraft II play, a complex real-time strategy game that requires long-term planning, resource management, and real-time decision-making. This achievement underscored the potential of AI to operate in dynamic and highly strategic environments, far beyond turn-based games like Go [10].

2020: GPT-3

The release of GPT-3 by OpenAI in 2020 marked a significant advance in the field of NLP. GPT-3, with 175 billion parameters, demonstrated unprecedented language generation capabilities, performing tasks such as translation, summarization, and question-answering with high proficiency and minimal fine-tuning. It set a new benchmark for the potential of AI in creative and linguistic tasks [9].

2021: DeepMind's AlphaFold 2 in Protein Folding

In 2021, AlphaFold 2, developed by DeepMind, solved one of biology's greatest challenges by predicting protein structures with remarkable accuracy. This breakthrough has significant implications for drug discovery, understanding diseases, and designing new biological processes, demonstrating AI's potential to revolutionize the life sciences [11].

2022: DALL-E 2 and Image Generation

OpenAI's DALL-E 2, released in 2022, demonstrated the power of AI in generating highly detailed and creative images from text descriptions. This model pushed the boundaries of AI in the visual domain, showcasing its ability to combine artistic creativity with technical precision, and opening new possibilities in design, marketing, and entertainment [12].

2023: GPT-4 and Multimodal AI

In 2023, OpenAI introduced GPT-4, which expanded the capabilities of its predecessor by being multimodal—able to process and generate both text and images. GPT-4's ability to handle complex queries across different formats has made it a powerful tool for applications in education, customer service, and creative industries, further blurring the distinction between products of human and machine intelligence [13].

The last five years have seen groundbreaking AI achievements each year that have pushed the boundaries of what AI can do. From mastering strategic games and understanding protein structures to generating high-quality text and images, AI's progress continues to accelerate, bringing us closer to a future where AI plays an integral role in nearly every aspect of society. In the next section, we examine human progress.

3 Impact of AI progress across various fields

As AI continues to evolve, its influence is expected to permeate multiple sectors, driving innovation and transformation. This section analyzes the potential impact of AI across key fields such as healthcare, finance, education, entertainment, and transportation, highlighting both the opportunities and challenges these advancements may bring.

Healthcare: AI has the potential to revolutionize healthcare by improving diagnostics, personalized medicine, and patient care. Machine learning algorithms are already being used to analyze medical images with greater accuracy than human radiologists, and AI-driven predictive analytics are helping to identify at-risk patients before conditions worsen. Additionally, AI can streamline administrative processes, reducing the burden on healthcare professionals and allowing for more efficient patient management. The integration of AI in healthcare is expected to lead to better patient outcomes, lower costs, and a more proactive approach to health management [14]. The JSI team is in the last phases of donating a home doctor system to all Slovenians.

Finance: The finance industry is experiencing significant transformations due to AI, particularly in areas such as algorithmic trading, risk management, and fraud detection. AI algorithms can analyze vast amounts of financial data in real-time, enabling more informed and faster decision-making. These technologies also enhance the accuracy of credit scoring and personalized financial advice, offering tailored solutions to individual customers. However, the increased reliance on AI also raises concerns about market stability, ethical use of data, and the potential for systemic risks [15].

Education: AI is poised to transform education by providing personalized learning experiences, automating administrative tasks, and enabling new forms of interactive learning. AI-driven adaptive learning systems can tailor educational content to the needs of individual students, allowing for more effective learning outcomes. Additionally, AI can assist teachers by automating grading and providing real-time feedback, freeing up more time for personalized instruction. GPTs further offer significant improvements in education. Integrating AI in education also presents challenges, such as ensuring equitable access to AI-driven tools and addressing concerns about data privacy [16]. At JSI, we tested the quality of various GPTs on educational tasks.

Entertainment: The entertainment industry is undergoing a significant shift due to AI's capabilities in content creation, recommendation systems, and audience engagement. AI-generated music, art, and scripts are becoming increasingly sophisticated, thus differentiating between products of human and machine creativity as often impossible. Recommendation algorithms, powered by AI, personalize content delivery to users, enhancing their experience and increasing engagement. However, this rise in AI-generated content raises questions about the future of human creativity and the potential for AI to disrupt traditional content production models [17]. Recommendation algorithms were one of the central parts of the H2020 smart-city Urbanite project with most of the software developed at AI.

Transportation: AI is driving innovation in transportation through the development of autonomous vehicles, smart traffic management systems, and predictive maintenance. Self-driving cars, powered by AI, promise to reduce accidents, lower emissions, and increase the efficiency of transportation networks. AI can also optimize traffic flow and reduce congestion through real-time data analysis and adaptive traffic control systems. However, the widespread adoption of AI in transportation faces challenges related to safety, regulatory frameworks, and public acceptance [18].

In the next section, human progress and integration with AI are presented.

4 Human progress including merging with ICT and AI

4.1 Historical overview of human progress

Human progress is a story of relentless evolution and technological advancement, spanning millions of years. Beginning around six million years ago, the earliest hominins diverged from the common ancestor we share with chimpanzees, marking the start of a journey toward modern humanity. One of the earliest major milestones was the development of bipedalism, which allowed early humans to travel long distances and use their hands for tool-making. The invention of tools around 2.6 million years ago further distinguished our ancestors, enabling them to manipulate their environment in unprecedented ways.

Approximately 200,000 years ago, Homo sapiens emerged, equipped with greater cognitive abilities and complex language, facilitating social structures and cultural developments that set the stage for future innovations. The advent of agriculture around 10,000 years ago marked a fundamental shift in human society, leading to settled communities and the eventual rise of civilizations.

Fast forward to the Industrial Revolution in the 18th century, human progress accelerated dramatically. Innovations in machinery, transportation, and communication reshaped societies, laying the groundwork for the Information Age. The 20th century saw rapid technological advances, including the development of the computer, the internet, and the beginnings of artificial intelligence, all of which have profoundly impacted human life.

4.2 Recent progress: merging with ICT and AI

In the past few decades, the convergence of information and communication technologies (ICT) and artificial intelligence has fundamentally altered the trajectory of human progress. This merger has enhanced human capabilities and begun to blur the lines between human and machine intelligence, creating a symbiotic relationship reshaping society.

Mobile Phones: One of the most transformative technologies of the late 20th and early 21st centuries is the mobile phone. Introduced commercially in the 1980s, mobile telephones rapidly evolved from simple communication devices to powerful, multifunctional tools. The advent of smartphones in the 2000s, with their integration of internet access, GPS, and a multitude of applications, significantly enhanced human connectivity and access to information. Today, smartphones are essential tools for both personal and professional life, facilitating real-time communication, social networking, and a vast array of digital services [19].

The Internet and Cloud Computing: The development of the Internet in the late 20th century and the rise of cloud computing in the early 21st century have revolutionized how humans interact with information and each other. The internet has democratized access to knowledge, enabling global communication and collaboration, while cloud computing has made vast computational resources and storage available to individuals and organizations alike. These technologies have increased productivity and laid the foundation for the widespread deployment of AI systems, which rely on large datasets and significant computational power [20].

Generative Pre-trained Transformers: The recent advancements in AI, particularly with the development of GPTs, represent a significant leap in the merging of human capabilities with machine intelligence. GPT-3, introduced in 2020, demonstrated the ability to generate coherent and contextually relevant text based on minimal input, performing a wide range of tasks such as translation, summarization, and even creative writing. GPT-4, released in 2023, expanded on this by incorporating multimodal capabilities, processing both text and images, and further enhancing human-machine interaction [13]. These models are not just tools but extensions of human

cognitive abilities, enabling users to perform tasks that require complex reasoning and linguistic skills with the assistance of AI.

Wearable Technology and Augmented Reality: Wearable devices, such as smartwatches and fitness trackers, have integrated AI into daily life, monitoring health metrics and providing real-time feedback to users. These devices exemplify the merging of human biology with technology, offering new ways to enhance physical and cognitive performance. Augmented reality (AR) technologies are also becoming increasingly prevalent, overlaying digital information onto the physical world and creating immersive experiences that enhance learning, navigation, and entertainment [21].

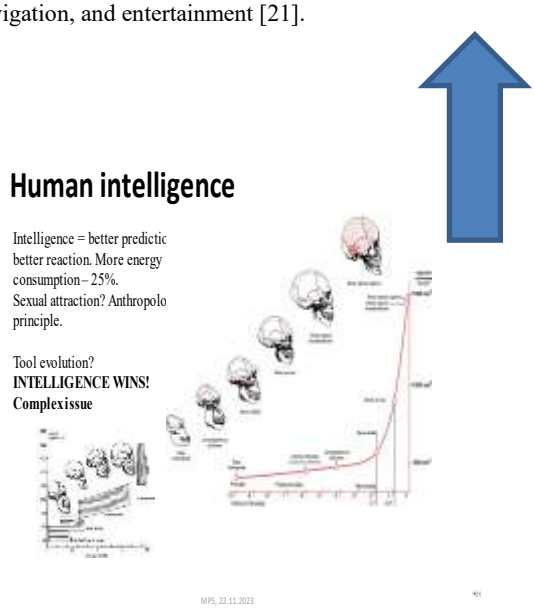


Figure 1: Progress of the human brain and intelligence. Source of the draft (modifications original): https://www.aquatic-human-ancestor.org/anatomy/brain.html

Human progress, from the earliest hominins to the modern age, has been marked by the continuous development and integration of technologies that enhance human capabilities. In recent decades, the merging of ICT and AI with human activities has accelerated this progress, creating a symbiotic relationship that extends human cognitive and physical abilities. Technologies such as mobile phones, the internet, GPT models, and wearable devices have not only transformed how we live and work but have also set the stage for future advancements that may further blur the lines between human and machine intelligence.

The key idea of this overview is that humans and user interfaces are already deeply intertwined, both through physical hardware like mobile phones and more abstract systems such as software and information networks. As these systems fast become more sophisticated, they enhance human cognitive, intellectual, and mental functions, effectively expanding our capacities. This growing interconnection between humans and technology echoes the concept of *cyborgization*, where external devices supplement and expand the functions of the human mind. The development of smartphones, wearable technology, and even future brain-computer interfaces suggests that this synergy is only deepening. Contrary to claims that our brain size is deteriorating, the measure of human intellectual capacity should now include not just our innate abilities but also the external systems that augment them [19].

The view that technology significantly enhances human cognitive ability aligns with Harari's notion of humans becoming "cyborgs" as they increasingly rely on tools that supplement mental processes. Similarly, [22] discusses the extended mind theory, which posits that external tools, such as smartphones, are integral components of the human cognitive system, challenging the notion that brain size or biological limitations strictly define mental capacity. These technological extensions of human cognitive ability have created new frameworks for evaluating our intellectual potential, making it more accurate to assess human functionality in a combined system of biological and technological entities.

The emergence of human-like properties such as consciousness [23], observed in advanced GPT models, represents a pivotal step in the evolution of artificial intelligence and human progress. These models, capable of understanding and generating natural language, are beginning to mimic forms of cognitive processes, thus contributing to what could be described as the dawn of a new "intelligent era." This era, driven by AI's ever-increasing capabilities, promises a deeper integration between human cognition and machine intelligence, potentially fostering innovations in problem-solving, creativity, and the expansion of knowledge.

Researchers like David Chalmers have explored the idea that AI systems, such as GPT models, may embody elements of extended cognition, which can extend human cognitive abilities beyond their biological limits. The more these models evolve, the more they may contribute to an era where AI complements human intelligence in unprecedented ways, leading to new forms of civilization that heavily rely on intelligent systems to solve complex global challenges [7, 22].

Figure 1 illustrates the functional growth of human problem-solving capabilities, driven by the integration of ICT and AI solutions, which serve as amplifiers of natural intelligence. Human cognitive abilities are being multiplied several times through this merger with ICT and AI, as represented by the blue arrow. While the original figure without the blue arrow shows the increase in human skull volume, and thus brain size, the blue arrow highlights the exponential growth in problem-solving capacity. A simple analogy can be drawn: consider a person walking barefoot versus using a car or plane. The speed of movement changes dramatically, even though the human's physical body remains unchanged. Similarly, while human biology (the brain) did not improve, the ability to tackle complex tasks surged drastically with the aid of ICT and AI

5 DISCUSSION

In recent years, there has been growing concern that Western civilization is experiencing a period of decline, marked by political fragmentation, cultural disintegration, and economic challenges. Scholars have pointed to a loss of social cohesion, declining institutional trust, and the rise of non-productive and conflicting ideologies [24, 25].

On the other hand, the rapid advancements in artificial intelligence are driving unprecedented changes across various

fields, fundamentally altering the landscape of industries and society. As AI technologies continue to evolve, they offer both tremendous opportunities and significant challenges that require careful consideration.

Balancing Innovation with Ethical Concerns: One of the primary discussions surrounding AI is the balance between innovation and ethical considerations. AI has the potential to revolutionize fields like healthcare, finance, education, and transportation by improving efficiency, accuracy, and personalization. However, these advancements also raise critical ethical questions, particularly regarding data privacy, algorithmic bias, and the potential for AI to perpetuate or exacerbate existing inequalities. For example, while AI-driven personalized medicine can enhance healthcare outcomes, it also risks marginalizing those without access to the necessary technology or data [14].

Moreover, the use of AI in finance, particularly in areas like algorithmic trading and credit scoring, has the potential to deepen economic disparities if not carefully regulated. The challenges of ensuring fairness, transparency, and accountability in AI systems are significant and demand robust governance frameworks to prevent misuse or unintended consequences [15].

The Human-AI Symbiosis: Another crucial aspect of the discussion is the growing symbiosis between humans and AI. As humans increasingly rely on AI technologies in daily life—through smartphones, wearables, and AI-powered applications—there is a merging of human and machine capabilities. This integration has the potential to significantly enhance human cognitive and physical abilities, leading to what some describe as an augmented human experience. However, this symbiosis also raises questions about dependency, control, and the future of human autonomy. As AI systems become more embedded in decision-making processes, it is essential to consider how these technologies may influence human behavior, decision-making, and even identity.

The development of AI models like GPT-4o has shown how closely intertwined human and machine intelligence can become. These models have not only expanded the possibilities of human-machine interaction but have also challenged our understanding of creativity, communication, and the nature of intelligence itself. As AI continues to evolve, it will be crucial to monitor and understand the long-term implications of this symbiotic relationship on human society and culture [13].

The author of this paper continuously highlights the significance of this merging, noting that the increasing integration of AI into human life first of all augments human capabilities but also to a certain degree presents complex ethical and philosophical challenges. As AI begins to mirror human-like consciousness in certain aspects [22], the line between human and machine products is becoming increasingly blurred, raising questions about the future of this relationship and the implications for human identity and autonomy [23]. The level of consciousness in large language models (LLMs), evaluated through Tononi's axioms of intelligence, was found to be significantly below that of human consciousness, but notably improved compared to earlier AI systems [23].

The Future of Work and Society: AI's impact on the workforce is another critical area of discussion. While AI can enhance productivity and create new opportunities, it also poses a significant threat to traditional job roles, particularly in industries where automation can replace human labor. This shift could lead to widespread job displacement, necessitating a rethinking of economic structures, education systems, and social safety nets to address the needs of a rapidly changing labor market [16].

The potential for AI to drive social and economic inequality is a pressing concern. Without proactive measures to ensure equitable access to AI technologies and to address the disparities that may arise from AI-driven economic shifts, society risks deepening existing divides. One of the best solutions is introducing the AI courses already in elementary schools.

Navigating the AI Revolution: As AI continues to advance, society is at a crossroads, faced with the task of navigating the complexities of the AI revolution. The potential benefits of AI are immense, but there are also certain risks. Ensuring that AI technologies are developed and deployed responsibly will require a concerted effort from governments, industry, academia, and civil society. This includes developing ethical guidelines, regulatory frameworks, and educational initiatives that can help society adapt to the changes brought about by AI. At the same time, these regulations should first of all enhance proper progress, research and development, and not pose additional bureaucratic burdens.

In cognitive terms, GPT models represent a promising approach to creating forms of artificial consciousness and cognitive information beings. These models simulate aspects of human cognition, such as language understanding and generation, by mimicking neural networks that resemble the processing of human brains. As they evolve, GPTs could potentially help us explore and understand the fundamental components of human consciousness, offering insights into both artificial and human cognition [26].

In conclusion, the discussion surrounding AI is multifaceted, touching on ethical, social, economic, and technological dimensions. As we advance, it is essential to balance harnessing AI's potential with addressing the challenges it brings, such as bias, privacy concerns, and the risk of job displacement. By proactively engaging with these issues, we can ensure that the AI revolution creates a future that is not only innovative but also promotes individual and societal human progress. Despite the ongoing debates and misunderstandings, the transition toward an information-driven era seems inevitable, as AI continues to integrate into every facet of human life, shaping our collective destiny.

Tool Usage: ChatGPT-4o and various grammar and word processing tools were applied to enhance the language quality. ChatGPT was also employed periodically to refine informal draft ideas into well-structured, formal text. The text was regardless of the language modifications finally examined and modified by the author.

References:

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998-6008. DOI: https://doi.org/10.5555/3295222.3295349.

2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. DOI: https://doi.org/10.1038/nature14539.
3. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.
4. Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Prentice Hall.
5. Biever, C. (2023). ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature*. Retrieved from https://www.nature.com/articles/d41586-023-02361-7.
6. Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486-492. DOI: https://doi.org/10.1126/science.aan8871.
7. Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
8. Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460. DOI: https://doi.org/10.1093/mind/LIX.236.433.
9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
10. Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... & Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350-354. DOI: https://doi.org/10.1038/s41586-019-1724-z.
11. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589. DOI: https://doi.org/10.1038/s41586-021-03819-2.
12. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.
13. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
14. Topol, E. (2019). Deep medicine: How artificial intelligence can make healthcare human again. Basic Books.
15. Gomber, P., Koch, J. A., & Siering, M. (2017). Digital finance and FinTech: Current research and future research directions. *Journal of Business Economics*, 87, 537-580. DOI: https://doi.org/10.1007/s11573-017-0852-x.
16. Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.
17. McCosker, A. (2021). AI, automation, and the creative industries. *Media International Australia*, 178(1), 141-154. DOI: https://doi.org/10.1177/1329878X20946209.
18. Goodall, N. J. (2016). Can you program ethics into a self-driving car? *IEEE Spectrum*, 53(6), 28-31. DOI: https://doi.org/10.1109/MSPEC.2016.7473149.
19. Katz, J. E., & Aakhus, M. (Eds.). (2002). Perpetual Contact: Mobile Communication, Private Talk, Public Performance. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511489471.
20. Marinescu, D. C. (2017). *Cloud computing: Theory and practice*. Morgan Kaufmann.
21. Billinghurst, M., Clark, A., & Lee, G. (2015). A survey of augmented reality. *Foundations and Trends® in Human–Computer Interaction*, 8(2-3), 73-272. DOI: https://doi.org/10.1561/1100000049.
22. Chalmers, J.D. (2023). Could a Large Language Model Be Conscious? Boston Review. Retrieved from Boston Review URL
23. Gams, M., & Kramar, S. (2024). Evaluating ChatGPT's Consciousness and Its Capability to Pass the Turing Test: A Comprehensive Analysis. Journal of Computer and Communications, 12(3), 219-237. DOI: 10.4236/jcc.2024.123014.
24. Murray, D. (2017). *The strange death of Europe: Immigration, identity, Islam*. Bloomsbury Continuum.
25. Deneen, P. J. (2018). *Why liberalism failed*. Yale University Press.
26. Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Cognitive perspective on production of third person dative and accusative clitic pronouns in Slovenian school-aged children

Maruša Brežnik Dornik
Center for Cognitive Science of Language
Faculty of Humanities
University of Nova Gorica
marusa.breznik@ung.si

Abstract

The paper investigates the production of third-person dative and accusative clitic pronouns in Slovenian school-aged children, focusing on whether cognitive factors influence their acquisition, despite the morphological similarities of these clitics in Slovenian. Previous research in Italian suggested that dative clitics in Italian are acquired earlier than the accusative due to their morphological differences, a pattern tested within the Slovenian context. Using elicited production tasks with 71 Slovenian children, the study reveals that in Slovene third person clitics are produced more frequently than third dative clitics, challenging the idea that acquisition is driven solely by morphological complexity. The research is framed within cognitive science, drawing on Universal Grammar and connectionist models to explore how cognitive processes, such as working memory and language processing demands, interact with linguistic structures.

Keywords

Language acquisition, Slovenian clitics, dative, accusative, pronouns

1 Introduction

Language acquisition is a fundamental aspect of cognitive development, providing a window into how and when the human mind processes and structures information. The acquisition of clitic pronouns, such as the third person dative (3DAT) and accusative (3ACC) clitics in Slovenian, involves complex cognitive processes that reflect both innate linguistic capacities and the influence of environmental factors. In the first part, this assignment explores these processes through the lenses of prominent cognitive science theories, including Universal Grammar and connectionist models, while also considering the role of working memory in language development. In the second part, the conducted experiment, which tested proposed research hypothesis from Italian on Slovenian school-aged children, is presented**.**

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.17

2 Experiment

Cardinaletti et al (2021) claim that the Italian dative clitics are acquired faster than their accusative counterparts because of a morphological difference between Italian dative and accusative clitics. Since there is no comparable difference between Slovenian dative and accusative clitics, their proposal predicts that the observed difference in acquisition should be absent in Slovenian. I tested this prediction among Slovenian children. The prediction was not confirmed, since children produced 3DAT clitics significantly less often than 3ACC.

2.1 Goals and predictions of the study

This study aims to examine the production of 3DATand 3ACC clitic pronouns among Slovenian school-aged children. Acquisition of the two clitic pronouns had been studied in Italian, where it was determined that the acquisition of the 3DAT clitics precedes the acquisition of 3ACC clitics [3]. The authors argue that the difference in the time of acquisition stems from different morphological makeup of the two sets of clitics. Italian dative clitics do not differentiate between gender (*gli* is a third dative pronoun used for both feminine and masculine gender), while accusative clitics differ for the two genders and are thus morphologically more complex. They argue that gender features, or better the lack of them, must be the reason why Italian children produced more 3DAT clitics than 3ACC clitics.

In Slovenian both 3ACC (ga “him”, jo “her”) and 3DAT (mu “to.him”, ji “to.her”) clitics are comparable in their morphological complexity as they both also spell-out the gender feature. Given the analysis in [3] it is predicted that there should be no difference in the production of 3ACC and 3DAT clitics in Slovenian. The purpose of this research is to test this prediction by exploring whether there is a difference in the production of 3ACC and 3DAT clitics among Slovenian school-aged children.

2.2 Methodology

The methodology for this study is structured around two main elicited production tasks, each tailored to evaluate the production

of dative and accusative clitic pronouns in Slovenian. These tasks are adapted from those used in the study [3], ensuring consistency in approach while accommodating the unique aspects of Slovenian. Slovenian stimuli consist of translations, and where necessary, adaptations of the Italian sentences used in [3] and of mostly unmodified drawings also from [3].

2.3 Participants

71 Slovenian typically developing (TD) children took part in the study. They were divided into six age groups, as shown in the table 1. Written in- formed consent was obtained from the children's parents prior to testing. Parents provided information about the languages spoken at home, which enabled us to exclude bilingual and L2 Slovenian speakers from the study.

Table 1: Groups, age and mean age of tested children.

Groups	Age	Mean age
TD1	6.6-6.9	6.8
TD2	7.0-7.9	7.4
TD3	8.0-8.8	8.6
TD4	9.0-9.9	9.6
TD5	10.0-10.8	10.4
TD6	11.0-11.9	11.4

Elicited Production – Accusative Task

This task is designed to elicit the use of 3ACC clitic pronouns. Children were presented with a series of visual stimuli featuring one or two characters engaged in various actions. For each set of images, the initial scene was described to the child using a recorded narrative. Following this, a second image was shown, and the child was asked to describe the action occurring, specifically focusing on the interaction between the characters. The aim is to prompt responses that naturally incorporate accusative clitic pronouns, reflecting the child's understanding and use of these grammatical structures.

Example Stimulus for Accusative Task

The first drawing shows a boy (agent) destroying a sand castle (patient), (Figure 1). The narrative describes the first scene, and the child is asked, "What is the boy doing to the castle?" (Figure 2). The expected response should include the accusative clitic pronoun corresponding to the castle sand, indicating the action directed towards the patient.

Figure 1: "In this story there is a boy that wants to destroy a sand castle."

Figure 2: "Look, what is he the boy doing to the castle?"

Similar elicited production task was made for the dative.

2.4 Procedure

Each child participant was individually tested in a quiet room within their school, ensuring a comfortable and distraction-free setting. All responses were audio- recorded and subsequently transcribed for analysis, with verification by two separate reviewers to ensure accuracy.

2.5 Response coding

We have classified the answers into three categories: target, production of full noun phrases (NP), clitic/NP omission. Every answer containing a clitic pro- noun was considered as target. Children have produced a good amount of target answers. In most of the answers they produced the same verbal form they had heard in the question, present tense, or sometimes produced sentences containing past tense. The most frequent non-target answer was the production of full NPs (in both, accusative and dative tests). The answers are grammatical, though redundant and pragmatically infelicitous, since the elicitation context requires clitic pronouns. There were some instances where clitics were omitted, either in the accusative or dative tests. In the accusative test, clitic omission led to ungrammatical sentences. In the dative test, ungrammatical responses occurred with verbs like *dati* "give," *podariti* "give," and *prinesti* "bring," all of which require a goal argument. Conversely, verbs such as *brati* "read" and *metati* "throw" resulted in grammatical sentences that were, however, contextually inappropriate for elicitation.

Table 2: Percentages of target answers for all groups

Groups	Target DAT	Target ACC
TD1	28,7	57,4
TD2	57,1	83,3
TD3	37,5	62,5
TD4	71,2	85,9
TD5	64,3	76,2
TD6	81,8	82,6

3 Results and discussion

All children's responses were compared using student t-test: the difference in the amount of 3DAT and 3ACC produced between the tested children is statistically significant ($p < .001$). Table 2 gives an overview of percentages of production of clitics, full NPs and omission in both tasks. Four instances of gender agreement error were found within the youngest group TD1 and two such errors within the TD3 group. Overall, children produced a good amount of target answers. The analysis within each group shows that the difference between 3DAT and 3ACC is noteworthy in all groups, except in TD6. The youngest groups produced significantly more 3ACC clitics than 3DAT clitics, namely TD1 28,7% more, TD2 26,19% more, TD3 25,0% more. As for the analysis between groups, we found significant differences for 3ACC, where the use of a 3ACC clitic is very low in TD1 group with 57,4%, TD3 group with 62,5% and TD5 with 76,19%. The omission was always higher with the 3DAT pronoun than 3ACC, TD1 omitted 3DAT with 16,67% more, TD2 with 25% more, TD3 with 8,34% more, in group TD4 no case of 3ACC omission was noted, TD5 omitted with significantly higher percentage of 20,23% more and TD6 with 4,54% more.

In this study the production of 3ACC and 3DAT clitic pronouns on Slovenian school-age children was tested, using two elicited production tasks. Differences between 3DAT and 3ACC clitics production were found in all groups. Children produced less 3DAT than 3ACC clitics in general. Which differs from what Italian kids (as reported in [1]) were producing, and also goes against the prediction based on [1]. Surprisingly the high omission is present in all age groups. Among the non-target answers, the production of full object instead of clitics was unusually high in the accusative task for the two oldest groups, while the four youngest groups produced fuller object in the dative task, which could be age related linked to the difficulty of the task. As Slovenian 3ACC and 3DAT clitics are morphologically comparably complex, the explanation provided in [1] cannot be used to explain the observed pattern.

The findings suggest that while innate linguistic capacities, as proposed by Universal Grammar, provide a foundation for language acquisition, the role of working memory and cognitive development cannot be overlooked. The results challenge the idea that the acquisition of clitic pronouns is driven solely by morphological complexity.

Acknowledgments

I am very grateful to Anna Cardinaletti, Sara Cerut and Francesca Volpato for sharing with me and allowing me to use their stimuli (drawings and Italian sentences) from [1]. I am grateful to the Elementary schools in Deskle and Celje that assisted me in running the experiment. This research was partially funded by ARIS grants N6-0314 and P6-03.

References

[1] P Baddeley, A. (2003). Working memory and language: An overview. Journal of Communication Disorders, 36(3), 189-208.

[2] Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. Bilingualism: Language and Cognition, 12(1), 3-11.

[3] Cardinaletti, A., Cerutti, S., & Volpato, F. (2021). On the acquisition of third person dative clitic pronouns in Italian. Lingue e linguaggio, 20(2), 311–341.

[4] Chomsky, N. (1981). Lectures on Government and Binding: The Pisa Lectures. Dordrecht: Foris.

[5] Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). Rethinking Innateness: A Connectionist Perspective on Development. MIT Press.

[6] Garnham, A. (2013). Mental Models and the Interpretation of Anaphora. Psychology Press.

[7] Gathercole, S. E., & Baddeley, A. D. (1993). Working memory and language. Psychology Press.

[8] Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5), 393-402

[9] Toporišič, J. Slovenska slovnica [Slovenian Grammar], 4th edition. Maribor: Obzorja (2000).

Ballot Butts: Nudging towards Pro Environmental Behaviour

Anouk Hartmans
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
ah17909@student.uni-lj.si

Lucija Karnelutti
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
lk00268@student.uni-lj.si

Leon Žužek
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
lz08739@student.uni-lj.si

Toma Strle
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
toma.strle@pef.uni-lj.si

Sabina Pajmon
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
sabina.pajmon@pef.uni-lj.si

ABSTRACT

This study explores the effectiveness of a nudge-based intervention, to reduce cigarette butt littering on a student campus. Using the principles of nudge theory, particularly the EAST framework, we designed a ballot box, allowing smokers to "vote" by disposing of their cigarette butts. Observations conducted before and after the intervention revealed a statistically significant increase in proper disposal, supporting the claim that nudges can positively influence environmental behavior. However, the study also highlights several limitations, including varying participant demographics and the challenge of isolating the factors driving behavioral change.

KEYWORDS

nudging, environmental behaviour, gamification, littering

1 INTRODUCTION

1.1. Increasing Need for Innovative Solutions

With the increasingly dire consequences of climate change, the urgency to address the environmental degradation has never been greater. Among the myriad of issues contributing to this escalating problem, littering—particularly the improper disposal of cigarette butts—stands out as a significant, yet often overlooked, contributor. In 2019, of the estimated 6 trillion cigarettes, only a third were properly disposed of [1]. Cigarette butts (CBs) are composed of tightly packed microfiber bundles of cellulose acetate. Cellulose acetate is cellulose treated with acetic acid, which heavily impedes the biodegradability of CBs. During their decades-long degradation period, CBs pose a double threat. The first is plastic pollution, as cellulose acetate is classified as a 'bio-plastic' with the second being the release of toxins that build up through the process of smoking [1]. The effects typically result from leaching, causing damage to aquatic life and contaminating waterways, while the consequences in terrestrial environments range from ingestion of butts, buildup of toxic chemicals, and soil contamination [1, 2, 3]. As such, finding ways to encourage proper disposal of CBs is crucial for reducing environmental harm.

1.2. Nudge Theory

One promising line of research in reducing littering is the nudge theory, first proposed by Thaler and Sunstein in their work *Nudge: Improving Decisions About Health, Wealth, and Happiness.* In their words, a nudge is "any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives" [4]. A nudge replaced their previous idea of paternalism, which similarly influences "choices of selected parties in a way that will make them better off" [5]. Several studies have found that nudges, in their various forms, can indeed be effective in reducing littering [6, 7, 8]. For example, a study on Chinese workers found that it is possible to reduce littering on the factory floor by 20% by placing golden coins, which are culturally and religiously significant, on the factory floor, thus changing it from a place that can be littered, to a place that should not be littered [8]. There are various forms of nudges and can be roughly divided into sizing (e.g. changing portion sizes in restaurants to reduce food waste), priming (e.g. footprints leading towards a bin), proximity (e.g. having a bin close by), presentation (e.g., designing eco-friendly devices as more aesthetic), labelling and improving the functional design [7]. Due to their diversity, usefulness and cost efficiency, nudges could help mitigate the environmental impact of CBs.

1.3. Theoretical framework

Our research was inspired and partly supported by the Green Nudge project[1]. This study specifically targeted the smoking behaviours of the student population from various faculties in the area of Kardeljeva ploščad in Ljubljana, aiming to assess how the design of bins could influence proper disposal habits of the CBs.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.18

[1] The Green Nudge project ("UL za trajnostno družbo – ULTRA"), financed by the EU - NextGenerationEU, and Republic of Slovenia, Ministry of Higher Education, Science and Innovation.

In our pilot study we implemented the nudging principles in the context of pro-environmental behavior. During our research design and nudge implementation process we relied heavily on the ideas from The Little Book of Green Nudges [9], where we utilized their five recommended steps of nudge implementation and their EAST framework, designed to make a nudge Easy, Attractive, Social and Timely. We based our nudge on the findings of Rifkin and colleagues [10], where they found out that behavior, such as tipping in a bar, can be influenced by "dueling preferences". If a behavior is presented as a choice between two options, preference for dogs versus cats, it gives people the opportunity to self-express themselves through a behavior that is not directly connected with the preference. In a similar fashion we have designed a cigarette voting box, where people could cast a vote with their CBs. The previously mentioned study was also a basis for a pilot study by Gay and colleagues [11], where they compared the impact of different cigarette bins on polluting behavior. They found that a "dual preference" voting box, like ours, was the most efficient in reducing the pollution of the environment with CBs.

Figure 1: The ballot box for CB's
The prompts on the box are: morning shower (slo. *tuširanje zjutraj)* and evening shower (slo. *tuširanje zvečer*). The box is made from a repurposed mail box and is standing on a metal pole. Surrounding the box is a picture depicting two smokers. The bin was made by our colleagues at the Academy of Fine Arts.

Existing studies addressing cigarette butt littering through behavioural experiments indicate that 63% of such littering is driven by individual motivations, such as a lack of awareness about environmental impacts and the availability of ash receptacles [12]. Other contributing factors include convenience (e.g., the distance to bins) and habitual behaviour [13], some research highlights a correlation between an area's cleanliness and the likelihood of littering, with certain demographics, like younger individuals and men, being more prone to littering [14]. While the design of ballot bins is often consistent across studies, the specific environments, demographics, and timelines vary. Research demonstrates that these bins can be an effective, low-cost solution for reducing cigarette butt litter, particularly in more homogeneous settings like school campuses. However, their effectiveness may diminish in more diverse public spaces [14]. Given the many variables influencing these outcomes, researchers recommend further experiments to optimize these interventions in different settings [12, 13, 14].

2 METHODS

Our preliminary study into the effectiveness of cigarette disposal through the use of ballot bins was conducted on a student campus in Ljubljana, Slovenia during the spring and summer of 2024. After initially observing the campus area, we decided to target the behavior of throwing CBs on the ground. There were several 'hotspots' of discarded cigarette butts, but we were particularly intrigued by the large number of butts thrown around bins. What intrigued us was the fact that despite there being a clear area for throwing away their cigarettes, smokers still did not opt for this choice. As such, we focused on a popular smoking area of the Faculty of Social Sciences at the University of Ljubljana. During the span of six months, we conducted two sets of observations, totalling seven observations: one set of observations before our intervention and one after. The first four observations were carried out in April 2024 and observed a popular smoking spot for students and faculty next to an existing bin. With the exception of the first observation, which was done in a group by all researchers, all were done individually over the course of two hours. During these observations, we collected data on the total number of CBs thrown in the bin or improperly discarded. We also took into account other factors such as time of day, weather and any other factors we deemed important like the number of people smoking together outside, or any other factors, which might have influenced the final number. The second set of observations was done during July, this time with the nudge (the ballot box) placed next to the bin in a popular smoking spot. The ballot box can be seen in **Figure 1**.

3 RESULTS

Our descriptive results are presented in the table below (see **Table 1**), where we calculated the mean value of CBs either in the bin or on the ground before and after the implementation of our green nudge.

Table 1: Littering behaviour observations before and after intervention with CB's thrown in the bin and on the ground

Observations	Condition	CB's in bin		CB's ground	
		M	*SD*	*M*	*SD*
1	before	21	6.38	12.25	5.74
2	after	17	17.35	4.3	2.52

In order to determine if there is a statistical difference between the proportion of CBs in the bin versus on the ground based on the condition, we conducted a two-proportion Z-test. Based on

the statistical analysis we found that there was a statistically higher proportion of CBs thrown into the bin versus on the ground after the implementation of our cigarette voting box, compared with the proportions before its implementation, $z = 0.165$, $p = 0.0495$.

4 DISCUSSION

Based on our results we can confirm our hypothesis that our nudge would increase the proportion of CBs thrown into the bin versus on the ground thus reducing the pollution of the environment surrounding the student campus with CBs, which is in line with the findings by Gay and colleagues [11]. Although our results do indicate a change in the proportions, conclusions should be taken with caution, since the frequency of smokers present before and after the implementation of the nudge varied vastly and could have had a big impact on the results of our analysis.

Additionally, it is difficult to determine exactly what nudged the participants' behaviour, which opens a broader question of nudge validity. Specifically, for our nudge, there could have been a number of factors influencing their behaviour. Some of these factors include 1) proximity; simply having more available bins could have decreased the number of CBs thrown on the ground, 2) novelty; the nudge gained attention by simply being a new structure in a familiar environment, 3) presentation; the ballot box is more attractive than a conventional bin, which is why participants would decide for it. While these factors do not negate the effectiveness of the nudge, the difficulty in pinpointing the determining factor could influence the design and implementation of nudges. For example, if novelty is the determining factor, a green arrow pointing towards a bin could have the same effect as a costly ballot box. There is also a possibility that our nudge was not clear enough and thus resulted in some people not engaging with it. The communication materials were designed with a tone that was perhaps too playful and light-hearted, which may not have resonated well with the student population of smokers, who might have responded better to more straightforward and direct messages. This lack of clarity could have had an overall impact on the efficacy of our nudge as suggested by Sunstein [15].

4.1. Limitations

One key limitation of our study is the comparability of pre- and post-intervention data. Before the intervention, data was collected during the ongoing academic term with a larger, consistent student population. Post-intervention data, however, was gathered after the exam period, when fewer students were present. Moreover, the population mainly consisted of foreign students attending summer school. The study of Chinese workers by Wu and Paluck mentioned in section 1.2. urges that cultural context must be taken into account when designing a nudge. They state that nudges "must recognize motivations and subjective interpretations within a particular context" [8]. Thus, without the proper consideration of the cultural background of foreign students, it seems highly unlikely that our nudge, designed for Slovene students of the Faculty of Social Sciences, had an equal effect on foreign students attending summer school.

4.2. Future directions

While our pilot study provides valuable insights into the effectiveness of nudging towards pro-environmental behaviour, future research could address the small sample size in this study by employing a larger, more diverse population to improve the generalizability of the findings. Additionally, observing the population within a shorter timeframe would improve the validity of our results. Further studies could also include an interview before implementing a green nudge, using polling to determine the general environmental attitude, and after the green nudge, to ascertain the factors influencing their decision-making process.

In conclusion, our study has shown that nudges can be successfully employed to influence non-environmental behaviours by combining behavioural insights from nudge theory and gamification concepts (see [16] for a study combining gamification and nudging). Specifically, a ballot box could be used in short term settings, like open-air concerts and other events, where littering poses an issue. However, further research is needed to expand upon the factors underlying non-environmental decisions.

ACKNOWLEDGMENTS

This pilot research study was partly supported by The Green Nudge project ("UL za trajnostno družbo – ULTRA") - European Union - NextGenerationEU, and Republic of Slovenia, Ministry of Higher Education, Science and Innovation.

The authors also wish to express their gratitude to the development team at the Academy of Fine Arts, who created the ballot box.

REFERENCES

[1] Green, D. S., Tongue, A. D., & Boots, B. (2022). The ecological impacts of discarded cigarette butts. Trends in Ecology & Evolution, 37(2), 183–192. https://doi.org/10.1016/j.tree.2021.10.001

[2] Conradi, M., & Sánchez-Moyano, J. E. (2022). Toward a sustainable circular economy for cigarette butts, the most common waste worldwide on the coast. The Science of the Total Environment, 847, 157634. https://doi.org/10.1016/j.scitotenv.2022.157634

[3] Qamar, W., Abdelgalil, A. A., Aljarboa, S., Alhuzani, M., & Altamimi, M. A. (2020). Cigarette waste: Assessment of hazard to the environment and health in Riyadh city. Saudi journal of biological sciences, 27(5), 1380–1383. https://doi.org/10.1016/j.sjbs.2019.12.002

[4] Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press.

[5] Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism. *The American Economic Review*, *93*(2), 175–179. http://www.jstor.org/stable/3132220

[6] McCoy, K., Oliver, J.J., Borden, D.S. and Cohn, S.I. (2018), "Nudging waste diversion at Western State Colorado University: application of behavioral insights", International Journal of Sustainability in Higher Education, Vol. 19 No. 3, pp. 608-621. https://doi.org/10.1108/IJSHE-05-2017-0063

[7] Wee, S., Choong, W., & Low, S. (2021). Can "Nudging" play a role to promote Pro-Environmental behavior? Environmental Challenges, 5, 100364. https://doi.org/10.1016/j.envc.2021.100364

[8] Wu, S. J., & Paluck, E. L. (2021). Designing nudges for the context: Golden coin decals nudge workplace behavior in China. Organizational Behavior and Human Decision Processes, 163, 43–50. https://doi.org/10.1016/j.obhdp.2018.10.002

[9] United Nations Environment Programme, GRIDArendal and Behavioural Insights Team (2020). The Little Book of Green Nudges: 40 Nudges to Spark Sustainable Behaviour on Campus. Nairobi and Arendal: UNEP and GRID-Arendal.

[10] Rifkin, J. R., Du, K. M., & Berger, J. (2021). Penny for your preferences: leveraging self-expression to encourage small prosocial gifts. Journal of Marketing, 85(3), 204-219. https://doi.org/10.1177/0022242920928064

[11] Gay, A., Pascual, A., Salanova, T., & Felonneau, M. L. (2023). What about using nudges to reduce cigarette butts pollution?. Journal of Human Behavior in the Social Environment, 1-8. https://doi.org/10.1080/10911359.2023.2219714

[12] Selagea, V. I., Simeanu, C.-M., & Stancu, E. A. (2016). Nudge: Cigarette butts—Not for littering but for voting. https://doi.org/10.13140/RG.2.1.3734.8886

[13] "Ballot Bins": Vote with your cigarette butt and stop pollution. (n.d.). Green Nudges. Retrieved August 22, 2024 from https://www.green-nudges.com/ballot-bins/

[14] Pavlovský, P., Sloboda, M., Sičáková-Beblavá, E., & Klunin, A. (2022). Not Always an Easy Win: The Effectiveness of a Ballot Bin Experiment to Prevent Cigarette Butt Littering. Ekonomika a Spoločnosť, 23(1), 32–49. https://doi.org/10.24040/eas.2022.23.1.32-49

[15] Sunstein, C. R. (2017). Nudges that fail. Behavioural public policy, 1(1), 4-25. https://doi.org/10.1017/bpp.2016.3

[16] Auf, H., Dagman, J., Renström, S., & Chaplin, J. (2021). GAMIFICATION AND NUDGING TECHNIQUES FOR IMPROVING USER ENGAGEMENT IN MENTAL HEALTH AND WELL-BEING APPS. *Proceedings of the Design Society*, *1*, 1647–1656. https://doi.org/10.1017/pds.2021.426

Problem Solving as a Key for Sustainable Future*

Ivana Štibi†
Department of Physics
Josip Juraj Strossmayer University of Osijek
Osijek, Croatia
istibi@fizika.unios.hr

Marija Gaurina
Department of Physics
University of Split, Faculty of Science
Split, Croatia
mgaurina@pmfst.hr

Ivana Katavić
Center of Excellence of the Split-Dalmatia County
Split, Croatia
ivanakatavic@ci-sdz.hr

Josip Stepanić
Faculty of Mechanical Engineering and Naval Architecture
University of Zagreb
Zagreb, Croatia
josip.stepanic@fsb.unizg.hr

Abstract

Achieving sustainability in today's complex world is a challenging, long-term endeavour. This paper focuses on the critical role of education in advancing sustainability, emphasizing the urgent need for innovative, interdisciplinary approaches that prepare students for the demands of both the modern and future world. Central to this discussion is the idea of problem-solving as a key to a sustainable future, which is deeply connected to the field of cognitive science. By leveraging insights from cognitive processes, researchers can develop innovative solutions to complex challenges, promoting resilience and adaptability in society. Understanding how individuals think and make decisions informs strategies for addressing pressing issues such as climate change, resource management, and social equity, ultimately contributing to sustainable development. By integrating insights from cognitive science—particularly in problem-solving, critical thinking, and metacognitive strategies—we highlight how these cognitive tools enhance students' abilities to tackle sustainability challenges. The paper examines key issues, relevant disciplines, and outlines a framework for shaping future education to effectively contribute to global sustainability efforts.

Keywords

problem solving, education, SDG, ESG, sustainability

1 Introduction

There is a global agreement that sufficient resources should be devoted to preserving, and improving in the amount possible, our society. Having in mind the complexity of our society in total, but also of its many components, this is certainly a rather demanding endeavour. Currently, as a combination of the long-term goals and the operational prescriptions, the Sustainable Development Goals (SDG) were formulated [1]. Closely related to these Goals is the concept relating environmental issues, social issues and corporate governance – ESG [2]. One may argue that each of the three involved notions encompass rather complex and large substructures.

SDG is collection of 17 goals, spanning different aspects of our society [1]. The Goal 4 – Quality Education is specifically devoted to education. But along with it, education is important for all other goals, having in mind that generally education contributes to transferring knowledge and experience between generations. In that way, education contributes both to latency of values and good practices in our society, as well as to their evolution and gradual improvement. It is not just that the learning process is important, but awareness of it is crucial for the reflective thinking needed to address the challenges of sustainability and complex systems.

In this paper, we focus on the stated role of education. In section two, we extract aspects of society that are currently too demanding for most educational approaches and are consequently insufficiently encompassed within the optimal form of education. Cognitive science plays a crucial role by exploring how individuals perceive, think, learn, and solve problems, providing a foundation for developing educational approaches that promote sustainability.

In sections three and four, we concentrate on education. In section three, we extract starting points for the development of an optimal form of education that substantially contributes to sustainability. The ability to solve problems, a key focus of cognitive science, is critical for navigating the unpredictable dynamics of complex adaptive systems, which characterize many contemporary challenges. In section four, we additionally emphasize the education for sustainable future and relate it to systems thinking. Critical thinking, another pillar of cognitive science, enables students to analyse global challenges such as climate change and inequality, making it essential for understanding the interconnected nature of these issues. We argue that the inclusion of systems thinking significantly contributes to the education for sustainable development. Section five provides the perspective and general guidelines for the

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

http://doi.org/10.70314/is.2024.cog.19

broader application of systems thinking in education. The last section concludes the paper.

2 Society as a complex system

Research of complexity and complex adaptive systems (CAS) broadened our understanding of characteristic dynamics in society and its parts, which includes cooperation and competition, emergence, bifurcations and deterministic chaos, to name some of the relevant notions [3, 4].

Dynamics of the CAS is characterized with nonlinear interactions, the important consequence of which is delay. That includes different aspects of dynamics, particularly the instabilities. Nonlinearity is a notion that can easily be described yet deserves significant experience to be considered properly, whether that be for predictions or for analysis of past events. CAS is modelled on the micro-level, by characteristics of entities, usually called agents, and rules of their interactions [5]. The macro-level, system behaviour is obtained by simulations. There is no direct linking between macro-level and micro-level. Instead, one modifies micro-level characteristics of agents and rules of their interactions to obtain specific macro-level system behaviour. Models differ in the level of stylization, so some reveal generic system behaviour, trends that can be related to many diverse systems [5], while others are specific and reproduce expected dynamics in detail but of a particular system. Some contributors emphasise anthropomorphic interpretation of quantities that are historically considered rather formal [6], while other develop formal approach to intrinsically human-related characteristics such as happiness [7, 8, 9], to mention few examples among a myriad of existing scientific contributions.

All stated should be taken into consideration if one wants to make reliable predictions for a society or some other CAS. That requires a critical mass of competent people, to be formed within every new generation by a well-formulated, learning outcomes oriented, profound education.

3 Education for complex systems

World Economic Forum states top 10 skills, for different periods since almost a decade ago. Consistently, critical thinking and complex problem-solving top the list of skills that employers believe will grow in prominence [10]. These have been consistently emphasized as crucial since the first report almost a decade ago. Moreover, broader application of ESG standards within organizations will also have a significant impact [10, 11]. One aspect is that investments that facilitate the green transition of businesses and the broader application of ESG standards bring about net job-creation [11].

OECD monitors education and different aspects of transition from school to work. In addition, OECD formulated and organizes world-wide testing of skills and knowledge among students, in the form of PISA testing [12, 13].

UNESCO has contributed extensively to the development of competences of children and youth [14]. It was realized some time ago that education for sustainable development is a key instrument to achieve the SDGs [14].

The UN previously proclaimed Decade of Education for Sustainable Development (2005–2014), aimed at integrating the principles and practices of sustainable development into all aspects of education and learning [14].

One of the consequences of the extensive testing of alignment of education with society's needs is that we have rather detailed and reliable insight into the fulfilment of the education potential. Yet, it is clearly stated in words by I. Bokova, General-director of UNESCO: "A fundamental change is needed in the way we think about education's role in global development, because it has a catalytic impact on the well-being of individuals and the future of our planet" [14].

All stated about the reached education level and characteristics of our society (and CAS as its models) point to the fact that an innovative, qualitatively different education is needed to include the complexity in formal curricula, so that students can grasp its fundaments in a significant portion.

As a particular aspect of education, we emphasize problem-solving. It is a set of actions aimed at solving a particular problem, no matter how complex, or interdisciplinary it is. However, that approach is not yet formulated precisely so its potential is realized only by a small part.

Along with the problem-solving skills, and critical thinking, it is crucial to embed metacognitive strategies in education. If students have learned to reflect on their own learning processes, they are better equipped to contribute meaningfully to the SDG-a and adapt to the complexity of modern society as a CAS.

4 Education for sustainable development

The ability to solve problems from the perspective of cognitive psychology is a crucial for achieving sustainability. Cognitive psychology, which studies mental processes such as perception, memory, thinking, and problem-solving, provides insights into how people make decisions and how they can be encouraged to adopt sustainable behaviours. Understanding how people process information and make decisions involves recognizing problems, generating possible solutions, evaluating those solutions, and selecting the most appropriate one. In the context of sustainability, these problem-solving skills are essential for individuals and communities to identify environmental challenges, develop innovative solutions, and implement sustainable practices [15].

Research has shown that human behaviour is a significant source of uncertainty in the use of natural resources and a critical factor in local and global sustainability challenges [16]. By integrating insights from behavioural sciences into sustainability research, we can develop policies that promote sustainable behaviour. Cognitive psychology provides tools to understand how people perceive environmental problems and how they can be motivated to change their behaviour [17].

Education plays a pivotal role in fostering these problem-solving skills. Education for sustainable development (ESD) equips learners of all ages with the knowledge, skills, values, and agency to address interconnected global challenges such as climate change, biodiversity loss, and resource depletion [18]. ESD empowers individuals to make informed decisions and take collective action to transform society and care for the planet [19].

To further enhance these educational efforts, system thinking is another critical element for achieving sustainability. System thinking allows us to see how different parts of a system interact and how changes in one part can affect the entire system [20]. By

understanding complex environmental problems through system thinking, we can develop holistic solutions that consider long-term consequences. This approach is integral to problem-solving as it helps identify the root causes of issues and their interconnections within the environmental system [21].

In addition to system thinking, STEM (Science, Technology, Engineering, and Mathematics) education is essential for sustainable development. STEM education equips individuals with critical thinking, problem-solving, and technical skills necessary to address environmental challenges. By integrating sustainability into STEM curricula, we can prepare future generations to develop innovative solutions for sustainable development [22, 23].

Moreover, a transdisciplinary approach is vital for sustainability. This approach involves collaboration between academics from different disciplines and non-academic actors to co-produce knowledge and develop actionable solutions. Transdisciplinary research addresses complex sustainability challenges by integrating multiple perspectives and promoting holistic understanding [24, 25].

The ability to solve problems from the perspective of cognitive psychology, combined with system thinking, STEM education, and a transdisciplinary approach, provides valuable tools for addressing environmental challenges and achieving sustainable development.

Before proceeding it is to be noted that any new concept or other type of change in education needs to be implemented in real time and space. Regarding time, there is a significant literature about structuration of time and about leisure time of children and youth, see e.g. [26, 27] and references therein.

5 Strategies for enhancing critical thinking and problem-solving skill's

Sometimes it is difficult for educational processes to determine which actions/methods due to the complex systems are successful, especially when we are talking about climate changes, how to integrate them into everyday life and how to receive feedback on the impacts at individual and collective levels [28]. This calls for initiatives to strengthen the link between education and science. To achieve exactly such initiatives, educational systems should use cognitive and educational strategies to foster innovative solutions, but in addition to this, implementation of gamification enhance engagement and motivation towards eco-friendly actions [29].

To be more precise, gamification can be a useful tool to teach students about sustainability in general [30], creating a gamified environment where students could be active citizens monitoring their impact on the environment and thus influence climate change and sustainability, which leads to fostering problem-solving skills, critical and analytical thinking and creating sustainable solutions.

Another strategy which can be implemented in educational system, which in its core supports the development of the individual as a critical participant of a particular system, is the metacognitive strategy. Metacognition as a concept of thinking about thinking [31], enhance critical thinking, problem-solving and finally adaptability in education [32]. Use of metacognitive strategies enables students to develop self-awareness, monitor their thinking process, and regulate their cognitive processes to be more accurate [31]. This is important when addressing complex, interdisciplinary challenges and CAS.

To be clearer, the following three dimensions of metacognitive strategies need to be implemented:

A) **Planning dimension**, students need to prepare themselves for problem-solving scenarios and thought processes, which helps clarify their understanding of the problem and outline an approach to solving it.

B) **Monitoring dimension**, students need to check and validate their comprehension of the problem-solving scenario through self-questioning, which sustains critical thinking. This step ensures continuous reflection on their knowledge, allowing them to adjust strategies based on real-time insights.

C) **Reflection dimension**, after completing the task, students need to analyse what they learned, reflect on the effectiveness of the strategies used, and consider improvements for future tasks. This helps students better understand the complexity of sustainability and develop a deeper understanding of how their learning strategies can evolve.

Therefore, sustainable education should also rely on metacognitive strategies, because in its concept it contains the skills of critical thinking, problem-solving, but also the sustainability of both the problem-solving scenario and the ecosystem that created the sustainable solution.

6 Methodology for implementing sustainable education

According to all previous stated, strategy for empowering sustainable educational system and people involved in it (students, teachers, principals, parents, local community) was created. It is based on the enrichment of school curricula with ESG principles, which ultimately strengthens the school's ecosystem and makes it sustainable, and it is built based on three dimensions.

First dimension of the program are students, which includes students in the local community as active participants and those who contribute to development and innovation through an interdisciplinary and transdisciplinary approach of teaching and learning. In this way, the education system enriches the local community with individuals who are ready to face the complex problems of the CAS, to solve them, and to improve already existing solutions. Second dimension of the program are the teachers. By strengthening their knowledge and skills, as well as by raising their awareness of problems and the possibility of active participation in solving them, teachers provide students with support in an appropriate and sustainable way.

Teachers need to collaborate with system beyond schools in the way that students can gather information, critically think about problems, give the scenario of solving specific problem and in the end implement possible solution. For this, teachers should have a support and life-long education. Third dimension of the program are parents, principals and local community, which, by raising awareness of the needs, problems and possibilities of innovation within the school's ecosystem and beyond, provide significant support to students and teachers in their development.

The ecosystem is then complete, which with its way of functioning provides sustainability because all participants are aware of learning protocols as well as improving or changing existing solutions.

7 Conclusion

The relationship between sustainable education and cognitive science is a profound and crucial one, as it merges the principles of mental processes with practical applications for addressing the complexities of our contemporary world. Cognitive science provides the foundation for understanding and developing educational strategies that foster sustainability. In this context, key cognitive concepts such as problem-solving, critical thinking, metacognition, and systems thinking play a significant role in shaping the approaches needed for sustainable development.

This paper highlights the critical role that education has both in the achieving sustainability in complex, nonlinear systems and in finding solutions to the contemporary problems. It emphasizes the urgent need for innovative and transdisciplinary educational approaches that prepare students for challenges that are expected from them to deal with. Key aspects include the importance of problem-solving skills, critical thinking, and metacognitive strategies. By integrating these elements into education, we equip future generations to address the complexities of sustainable development and complex adaptive systems.

In conclusion, the principles of cognitive science are deeply intertwined with the goals and strategies of sustainable education. Cognitive science provides a framework for understanding how students can effectively engage with the complexities of sustainable development and complex adaptive systems. The integration of these cognitive skills into educational curricula ensures that future generations are equipped with the mental tools necessary to address the pressing environmental and societal challenges of our time.

Ultimately, education for sustainable development, supported by cognitive science, fosters not only informed and capable individuals but also a more resilient and adaptable society.

References

[1] United Nations: The 17 Goals. Retrieved August 12 2024 from https://sdgs.un.org/goals.

[2] Stephen Conmy: What is ESG and why is it important? Retrieved August 12 2024 from https://www.thecorporategovernanceinstitute.com/insights/guides/what-is-esg-and-why-is-it-important.

[3] Claudius Gros, 2024. Complex and Adaptive Systems. Spinger Cham. DOI: https://doi.org/10.1007/978-3-031-55076-8.

[4] Simon A. Levin, 2003. Complex and Adaptive Systems: Exploring the Known, the Unknown and the Unknowable. Bulletin of the American Mathematical Society **40**(1), 3-19.

[5] Joshua M. Epstein and Robert Axtell, 1996. Growing artificial societies: social science from the bottom up. Brookings Institution Press. ISBN 978-0-262-55025-3.

[6] Urban Kordeš, 2005. Entropy – our best Friend. Interdisciplinary Description of Complex Systems **3**(1), 17-26.

[7] Katalin Martinás,2012. Greatest Happiness Principle in a Complex System Approach. Interdisciplinary Description of Complex Systems **10**(2), 88-102, DO: https://doi.org/10.7906/indecs.10.2.5.

[8] Katalin Martinás and Zsolt Gilányi, 2012. Greatest Happiness Principle in a Complex System: Maximisation versus Driving Force. Interdisciplinary Description of Complex Systems **10**(2), 103-113. DOI: https://doi.org/10.7906/indecs.10.2.6

[9] Sabine Hossenfelder, 2013. On the Problem of Measuring Happiness. Interdisciplinary Description of Complex Systems **11**(3), 289-301. DOI: https://doi.org/10.7906/indecs.11.3.2.

[10] World Economic Forum, 2023. The Future of Jobs Report. Retrieved August 12 2024 from https://www.weforum.org/publications/the-future-of-jobs-report-2023.

[11] World Economic Forum, 2023. The Future of Jobs Report. Retrieved August 12 2024 from https://www.weforum.org/publications/the-future-of-jobs-report-2023/digest.

[12] OECD, 2018. Skills for the 21st Century: Findings and Policy Lessons from the OECD Survey of Adult Skills. Retrieved August 12 2024 from https://one.oecd.org/document/EDU/WKP(2018)2/en/pdf

[13] OECD: Programme for International Student Assessment (PISA). Retrieved August 12 2024 from https://www.oecd.org/en/about/programmes/pisa.html

[14] UNESCO, 2017. Education for Sustainable Development Goals. Learning Objectives. Retrieved August 12 2024 from https://unesdoc.unesco.org/ark:/48223/pf0000247444.

[15] S.M. Constantino, M. Schlüter, E.U. Weber and N. Wijermans, 2021. Cognition and behavior in context: a framework and theories to explain natural resource use decisions in social-ecological systems. Sustainability Science **16**(5), 1651-1671. DOI: https://doi.org/10.1007/s11625-021-00989-w.

[16] Vasiliki Kioupi and Nikolaos Voulvoulis, 2019. Education for Sustainable Development: A Systemic Framework for Connecting the SDGs to Educational Outcomes. Sustainability **11**(21), 6104. DOI: https://doi.org/10.3390/su11216104.

[17] Christie Manning and Elise Amel, 2022. How to use psychology for sustainability and climate justice. Psychology Student Network. Retrieved August 12 2024 from https://www.apa.org/ed/precollege/psn/2022/03/psychology-sustainability-climate-justice.

[18] Jay Hays and Hayo Reinders, 2020. Sustainable learning and education: A curriculum for the future. International Review of Education **66**(1), 29-52. DOI: https://doi.org/10.1007/s11159-020-09820-7.

[19] UNESCO, 2024. What you need to know about education for sustainable development. Retrieved August 12 2024 from https://www.unesco.org/en/sustainable-development/education/need-know.

[20] Megan Seibert, 2018. Systems Thinking and How It Can Help Build a Sustainable World: A Beginning Conversation. The Solutions Journal. Retrieved August 12 2024 from https://thesolutionsjournal.com/systems-thinking-can-help-build-sustainable-world-beginning-conversation.

[21] Stephen A. Harwood, 2019. Systems Thinking and Sustainable Development. In: W. Leal Filho, ed.: Encyclopaedia of Sustainability in Higher Education. pp.1892-1897, Springer Cham. DOI: https://doi.org/10.1007/978-3-030-11352-0_399.

[22] Cemil Cihan Ozalevli, 2023: Why sustainability must become an integral part of STEM education. Retrieved August 12 2024 from https://www.weforum.org/agenda/2023/04/why-sustainability-must-become-an-integral-part-of-stem-education.

[23] Smithsonian Science Education Center, 2024. STEM Education for Sustainable Development. Retrieved August 12 2024 from https://ssec.si.edu/stem-education-sustainable-development.

[24] Ryan Plummer, Jessica Blythe, Georgina G. Gurney, Samantha Witkowski and Derek Armitage, 2022. Transdisciplinary partnerships for sustainability: an evaluation guide. Sustainability Science **17**(3), 955-967. DOI: https://doi.org/10.1007/s11625-021-01074-y.

[25] Reihaneh Bandari, Enayat A. Moallemi, Ali Kharrazi, Robert Šakić Trogrlić and Brett A. Bryan, 2024. Transdisciplinary approaches to local sustainability: aligning local governance and navigating spillovers with global action towards the Sustainable Development Goals. Sustainability Science **19**(4), 1293-1312. DOI: https://doi.org/10.1007/s11625-024-01494-6.

[26] David Harris, 2005. Key Concepts in Leisure Studies. Sage Publications. DOI: https://doi.org/10.4135/9781446220696.

[27] Ivana Katavić, Bruno Matijašević, Petar Marija Radelj, Josip Stepanić and Mislav Stjepan Žebec, 2024. Review of Recent Literature about Leisure Time of School-Aged Children and Youth in Croatia. Interdisciplinary Description of Complex Systems **22**(1), 25-58. DOI: https://doi.org/10.7906/indecs.22.1.2.

[28] Christian A. Klöckner, 2015. The psychology of pro- environmental communication – Beyond standard information strategies. Palgrave Macmillan, London. DOI: https://doi.org/10.1057/9781137348326.

[29] Wiek, A., Withycombe, L., & Redman, C. L. (2011). Key competencies in sustainability: a reference framework for academic program development. Sustainability Science, 6(2), 203-218.

[30] Fang Zhang, 2024. Enhancing ESG learning outcomes through gamification: An experimental study. PLoS ONE **19**(5), e0303259. DOI: https://doi.org/10.1371/journal.pone.0303259.

[31] Diane F. Halpern, 1998. Teaching for critical thinking: helping college students develop the skills and dispositions of a teaching critical thinking for transfer across domains: dispositions, skills, structure training, and metacognitive monitoring. The American Psychologist, **53**(4), 449-455. DOI: https://psycnet.apa.org/doi/10.1037/0003-066X.53.4.449.

[32] Kelly Y.L. Ku and Irene T. Ho, 2010. Metacognitive strategies that enhance critical thinking. Metacognition Learning **5**(3), 251-267. DOI: https://doi.org/10.1007/s11409-010-9060-6.

Mind, the Gap, and Other Cracks

Maša Poljšak Kus
Center for Cognitive Science
Faculty of Education
University of Ljubljana
m.poljsak.kus@gmail.com

Urban Kordeš
Center for Cognitive Science
Faculty of Education
University of Ljubljana
urban.kordes@pef.uni-lj.si

Abstract

With this paper we aim to outline numerous gaps and other cracks that emerge when we start researching conscious experience through first and second-person research approaches. The terms used to name various gaps were chosen for the sake of coherence (with a pinch of playfulness). The main gap is the *chasm between two consciousnesses* which we are trying to bridge by an exchange of descriptions of our lived experiences. When we begin to turn our awareness to *what it is like* to be we begin to develop the skill and way of observing in which experience is created. We call this gap between our everyday attitude and phenomenological observation the *crevice of awareness* in which lies the act of becoming aware of an experience. After becoming aware of a certain layer of experience we reach the *fissure of description*, which represents the crack between the actual experience as perceived and the constructed linguistic concepts in which we try to convey what and how we perceived the experience. When a description of an experience has been produced, the researcher interested in investigating human experience is confronted with the *cranny of comprehension*. We relate this process of conveying our experience to another conscious being to the processes of translation and remind researchers of lived experience to be careful and weary of the interpretation that inherently shadows every translation.

Keywords

Background experience, Conscious experience, First-person research, Second-person research, Experiential translation.

1 Introduction

Upon delving into topics and discussions regarding our understanding of the mind, we inevitably reach one or another gaping chasm – the most notorious is even named the hard problem of consciousness. David Chalmers [1] points out that there is nothing we know more intimately than conscious experience but there is also nothing harder to explain. In this article we are not trying to explain conscious experience, but we are interested in exploring the process of explaining and describing our conscious experience to another human being – in this paper we call this process of reporting our subjective experiences *experiential translation*. The subjective aspects of thinking, perceiving and feeling are all states of experience that have a certain way in which we experience them. As Thomas Nagel [2] puts it, there is *something it is like* to be a conscious organism, and this *what is it like* to be another organism is, most likely always, over an insurmountable chasm between one conscious organism and another.

We aim to address this chasm that extends from one experiential being to another and explore the cracks that emerge when trying to explore and extend from one ridge to another. In this analogy the ridges of *the chasm between two consciousnesses* represent different conscious organisms, each with their own *what it is like* to be, and the chasm is the impossibility of reaching the exact *what is it like* of another being. In the field of first and second-person research of lived experiences, researchers are trying to bridge this chasm by collecting detailed descriptions of experience. We will argue that in the act of producing and collecting such descriptions we stumble upon many cracks, located on both sides of what we call *the chasm between two consciousnesses*. Starting from first-person view (as one should, when going about empirical phenomenology) we stumble upon a crevice that is becoming aware of *what is it like* to be – the most intimate experience, yet often hidden behind a wall of what Edmund Husserl [3] calls our *natural attitude*.

In this paper we also touch another, an even more veiled dimension of experience that we call *background experience* (explained further in the section 3) but most importantly we state that there is a gap between what we can easily consciously perceive and what we cannot – which, for the sake of clarity, we call *the crevice of awareness*. When trying to convey one's experience to another we stumble upon the next gap in the act of translating the experience into concepts, categories and linguistic forms. We believe that there is a gap between our perceived experience and its description, which we name *the fissure of description*. When trying to fill this fissure we believe the experience conveyed is flattened and reduced. The description produced in this effort then becomes the main building block of the bridge we are building from one side of the chasm and what we can offer to the conscious being reaching out from the other side. In this paper we compare this act of describing on one side and comprehending on another as a process of translation and that practicing experiential translation is the way to more valid and richer descriptions.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.cog.20

2 Experiential translation

Phenomenology, by origin a philosophical discipline, is trying to investigate concrete experiential phenomena and encourages detailed analysis of different aspects of consciousness. As such it has also been described as “a first-person description of ‘what it is like’ of experience” [4]. This subjective dimension ‘as it is lived from the inside’ is essential to consider in the field of scientific investigation of cognition and not be constrained merely to the data that can be observed and measured from the outside [5].

Claire Petitmengin [5] warns us that describing one’s own subjective experience is not merely hard, but extremely difficult, mostly because turning our attention to our consciousness, and *a fortiori* describing it, requires inner effort and a specific kind of skillset. Her assumption is that a substantial proportion of our subjective experience unfolds below the threshold of consciousness. We question what her assumption presupposes - that our consciousness is something "in there" to be observed and we only need a better instrument to see further and better. We, on the other hand, are more inclined to view conscious phenomena as something co-created with and by the act of observation. In either case we believe that in the field of first-person research the two initial steps – becoming aware of our experience and then describing it - include two important gaps.

1. *The crevice of awareness* is the crack between what our "view from within" knows how to observe and what eludes our reflective thoughts. It denotes the difficulties of becoming aware of our background feelings and core dimensions of our experience.
2. *The fissure of description* is the gap between subjective observations of lived experience and descriptions of observed experience, which are most often verbal. This is perhaps at times even more frustrating, because in an instance when one has become aware of an experience, they must now try to find the right words and gestures to convey and verbalize a description that captures the nature of the subjective experience in question.

Subjective, or first-person research transfers to second-person research when we not only try to surmount *the fissure of description*, but we also convey this description to a researcher interested in exploring structures of lived experiences. Empirical or second-person research usually involves interviewing human participants about *their* experience. In the context of our paper, we call the interview method a rather wobbly bridge that tries to connect participants’ lived experience with researcher’s understanding *via* the participants’ description of experience.

3. With this bridge we mark the third gap in the premise of empirical research of subjective experience – *the cranny of comprehension* – which spans between the second-person investigator and first-person report about the experience. It is a gap each researcher must fill and bridge when trying to comprehend and analyze the descriptive data on experience of others. We differentiate this cranny from the *chasm between two consciousnesses* because it is focused on the description and comprehension, not the entirety of another conscious experience.

We note that all three cracks are part of the greater *chasm between two consciousness*, which refers to the impossibility of experiencing as another being.

2.1 First-person translation

Jakob Boer [6] argues that the process of describing first-person experience is an act of experiential translation, with which we are inclined to very strongly agree. We believe that the act of describing subjective experience is an act of translation (latin *transfero*, “I convey”, from prefix *trans-*, “across, beyond” and participle *latus* “borne, carried”). We will describe an example of a process of translating an ancient Greek text to a modern language. The underlaying assumption is that without an observer there is no meaning, and thus the nature and skill of the observer influence the source text immensely. Firstly, one must be able to see the Greek alphabet and know the symbols to perceive anything more than mere scribbles. Secondly, one must understand what a specific set of symbols denotes and relate to it a previously known meaning - one must understand the word. This step alone is complex and multidimensional, because one Greek word can have numerous possible translations and the meaning that stands out to the translator is tied to many factors, such as context and previous knowledge. Thirdly, one must understand the grammar and syntax to make sense of a sentence. With this we want to show how the meaning of a text is co-defined by the observer. The translator must then choose an accurate set of words in another language to convey his interpretation of the sentence. With this example we tried to show the complexity of our influences on what we perceive and how we leave a mark on both our perception and our description.

Experiential translation assumes that lived experience is in nature distinct from linguistic form, and that in the act of verbalizing we carry certain aspects across the gap between experience and description. In the act of translating our lived experiences into words, concepts, and categories we inherently imbue chosen meanings with our interpretation, which is perhaps inseparable from the way we become aware of our experience. We relate this intrinsic interpretation to *horizons of attending to experience*, as explained by Urban Kordeš and Ema Demšar [7], who argue that this co-defines experiential phenomena that end up being observed and reported. The *horizon* is the way in which we perceive, by which we mean co-create, our experience. This is enacted both when we try to observe and when we try to describe our experience.

2.2 Second-person translation

In the previous section we compared the process of describing one’s lived experience to the process of translation. We continue with this analogy in the case of second-person research, when such translation is perchance more intuitive, because the ‘input’ – verbal report – comes in form of language. The researcher that receives the report proceeds with translating it in more than one way. First and foremost, the translation happens instantaneously, as it does every time we speak to another human being – we translate the words into our own known concepts and position them in our pre-existing field of knowledge.

Even more importantly, we aim to compare the subsequent process of analysing, categorizing and forming conclusions on the structure of experience to the process of translating, drawing attention once more to the notion that with translation always comes interpretation. As such we want to note and warn that becoming aware of your own *horizons of attending to experience* is a crucial step for every second-person researcher of

consciousness, which inherently makes them a first-person researcher as well.

3 Background experience

In this paper we turn our attention to a layer of experience which is, ironically, not in the focus of our attention but rather on the brink of it. William James [8] refers to this as the *fringe of consciousness.* To this fringe belong experiences that lack specific, sensory qualities, like the tip-of-the-tongue state (the intention to seek a missing word), feelings of knowing, familiarity and plausibility, intuitive judgments and numerous other conscious or quasi-conscious events that can be reported on with low sensory specificity.

What is it like aspect of those experience is hard to perceive and convey, but Petitmengin [9] describes certain internal gestures, which serve, in the language of our analogy, as bridges that enable us to become aware of the *source dimension* of our experience, which is usually *pre-reflective.* This unarticulated dimension is considered as core due to its ever-present nature, and because it is pre-conceptual and pre-discursive, it seems to be situated at the source of our thoughts. Although it constantly accompanies us, we need special circumstances to become aware of it and/or specific training in first-person observation.

In the realm of emotion, Antonio Damasio [10] calls a group of fleeting and hard-to-name feelings '*background feelings*', because they are not in the foreground of our mind, yet they help define our mental state and color our lives. We relate the foreground of our mind with the experiences on which we can easily focus our attention (such as thoughts, perceptions and loud emotions). Background feelings arise from background emotions, which are directed more internally than externally, but can nevertheless be observable to others in several ways: tone of our voice, prosody of our speech, the speed and design of our movements. According to Damasio, prominent background feelings include fatigue, energy, excitement, tension, relaxation, stability, instability, etc. The relation between background feelings and our drives and moods is intimate and close, but the relation between background feelings and consciousness is just as close, if not more. Matthew Ratcliffe [11] similarly develops the term *existential feeling* as a background which comprises the very sense of 'being' or 'reality' that attaches to world experiences. Specifically directed emotions presuppose this background, so regardless of the structure of such emotion, existential feelings are a more fundamental feature of world-experience. A few examples of such feelings are the feeling of being 'complete', 'unworthy', 'at home', 'abandoned' – all being descriptions of one's relationship with the world.

Hopefully we have now outlined the gap between our focal awareness and the experiences on the fringe of consciousness, where perhaps one of the keys to understanding our mind lies hidden. This gap was one of the points we tried to address in our recent project [12], in which we investigated the feelings of atmosphere with the presupposition that they are in the background of our mind. We will briefly present the context of our empirical investigations to use it as the reference point for our observations regarding the numerous gaps and blind spots of our methodological approach and epistemological premises.

4 Empirical context

In the aforementioned project, 'Unveiling of the Atmosphere – Etnophenomenological exploration of experiential background in relation to space', we aimed to investigate background experience which we have defined as feelings that weave the foundation on which foreground phenomena of consciousness unfold (such as emotions, thoughts and perceptions). We presupposed that experiences of atmosphere are by their nature affective, so we focused on the affective layer of experience. These feelings usually lack specific sensory attributes and are hard to pinpoint and often notice and/or name. We tried to capture and convey such background feelings with an empirical approach and a qualitative research design in which we combined approaches of first-person research such as Descriptive Experience Sampling Method (DES) [13], and ethnographical tools such as *in situ* diary entries. Our study was conducted in three phases, the first being the pilot study. We recruited three participants, previously trained in DES and first-person research, which we deemed important for a study that aims to research pre-reflective dimensions of experience.

Our participants reported about their experience in three ways: 1) through short written reports about randomly sampled moments during the day, 2) with diary entries on multiple occasions during the day of sampling, in which they situated randomly sampled moments in the context of their moods and behaviors, 3) in interview sessions in which we explored and expanded previous two types of data. The aim was to map our participants' affective experiential landscapes and to contextualize their experiences with information about their activities, environment and social interactions. We have analyzed the data according to the principles of qualitative analysis [14], which produced a list of experiential categories divided into two (vaguely distinct yet obviously separate) groups of *foreground* and *background affective experience.* In the background we situated categories such as *background mood*, *ambient atmosphere* and *deep atmosphere*.

1. *Background mood* is felt as all-encompassing and includes different ways of receiving, creating and experiencing foreground experiences (affects, thoughts and percepts), which we call different attitudes. We found three subcategories of background mood: open, closed and numb.
2. *Ambient atmosphere* includes experiences that are not clear and separate, but pervasive and ubiquitous. It represents feelings, which we feel originate from the world, and we are entangled with it either as their co-creator or merely as an observer.
3. *Deep atmosphere* includes experiences that we feel as deeply our own and private. Imprint of *deep atmosphere* marks the way of foreground affects as well as other background feelings. Phenomenologically it is harder to reach and observe, as it usually changes its character less or more slowly. When captured, we observed two distinct subcategories of feelings: *deep perturbation* and *deep unconcern*, the former connected to the feelings of danger and the latter to the feelings of safety.

5 Observational interstices

In this section we aim to address some methodological cracks and to note our observations from our research project on background feelings [12].

5.1 Becoming aware

In our study participants were prompted with a signal which conveyed to them that they should observe and report on their experience of the moment right before the signal. During the interviews they oftentimes reported that after the signal there was a brief state of feeling 'blank', as if the moment before the signal was empty and void of any experience whatsoever. But this feeling soon passed, and they started to remember and find words to describe the moment before the signal. We interpret this feeling of 'blankness' as a type of gap between being immersed in the *natural attitude* [3] and adapting the *phenomenological attitude*. To put it differently – we believe that the act of *epoché* is both and act of opening a gap and of bridging it. We argue that each time we try to bracket our trust in the objectivity of the world, we reveal and/or create a crack in the fluidity and continuity of the flow of our conscious experience. This means that when we change the nature of our awareness, we experience a moment of emptiness. To explain we will compare our awareness with the grip of our hand. When we hold on to one object, let's say a glass, we are gripping something and sensing specific qualities. When we want to switch to a different object, we must first release the glass and be (and thus feel) empty at least for a moment so that we can grip (experience) something else.

5.2 Observing experiential background

As mentioned in the section 2, we tried to observe and capture background feelings with the intention of mapping participants' experiential landscapes of affects. Based on the literature and preliminary observations we presupposed that background feelings change less frequently, which is one of the reasons they are more elusive and harder to notice, as opposed to the foreground experiences which change from moment to moment and require most of our attention.

Our findings support our claim that one way to notice the ever present is by gaps in continuity. Such a way requires regular first-person observation, optimally supported by a second-person approach (dialogue). Noting one's experience often over a longer period can bring to light changes that unravel slowly. To explain this with a more concrete and visual analogy – when a person on a diet is losing weight (if they are doing it in a healthy and sustainable way) they won't see any progress from day to day, but if they observe and measure themselves methodically throughout the whole year, they can notice a vast difference from their starting point.

5.3 Describing lived experience

Tying to the conclusion of the previous paragraph is a very concrete observation based on our research methodology. As described in section 2, we gathered reports on our participants' experience in three ways (short notes on experience of moments during the day, diary entries and interview insights). What we noticed is that often in the descriptions of a singular moment there was a lot of emphasis on the foreground experiences and less so on the background feelings. When participants weaved those moments in the experiential timeline of their whole day (and in the interviews of their whole week) more background feelings came into light – even in the moments which we had detailed descriptions of. We would like to note that minimising the effect of memory on reports is important, but that sometimes in this effort we miss something because it is 'right under our nose'.

6 Conclusion

Delving into the field of empirical phenomenology is a courageous act, because there are few, if any, clear and firm climbing holds. We understand why scientific discourse steers toward replicable and third-person tested approaches, yet we believe that exploration of lived experienced cannot (at least as of yet) be accessed any other way than through subjective observation first. And even if the act of bridging the subjective with intersubjective is full of gaps and other cracks, we stay positive that the descriptions and interpretations produced in this process lead to better understanding of how to approach empirical research of subjective experience. In the analogy of translation as the act of describing one's own experience, we aim to paint the following picture. In the gaps that lurk amid experiencing, being aware and describing, many pieces of the original experience are most likely lost in translation. Yet by persistently and methodically carrying over the remaining pieces created by this process we are building better and more reliable bridges.

7 References

[1] David Chalmers, 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2 (3), 200-2019.

[2] Thomas Nagel, 1974. What Is It Like to Be a Bat ? *The Philosophical Review*, vol. 83, No. 4. Duke University Press.

[3] Edmund Husserl, 1983. *Ideas Pertaining to Pure Phenomenology and to a Phenomenological Philosophy*. Springer.

[4] Shaun Gallagher & Dan Zahavi, 2008. *The Phenomenological Mind*. Londond : Routledge.

[5] 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*(3–4), 229–269. https://doi.org/10.1007/s11097-006-9022-2

[6] Jakob Boer, 2023. Phenomenology as Experiential Translation: Towards a Semiotic Typology of Descriptive and Expressive Ways of Making Sense of Experience. *Critical Arts.* DOI: 10.1080/02560046.2023.2262520

[7] Urban Kordeš & Ema Demšar 2021. Horizons of becoming aware: Constructing a pragmatic-epistemological framework for empirical first-person research. *Phenomenology and the Cognitive Sciences*, *22*(2), 339–367. https://doi.org/10.1007/s11097-021-09767-6

[8] William James, 1890. *The Principles of Psychology.* New York : Holt.

[9] Claire Petitmengin, 2007. Towards the Source of Thoughts : The Gestural and Transmodal Dimension of Lived Experience. *Journal of Consciousness Studies*, 14, No. 3, 54–82.

[10] Antonio Damasio, 1999. *The Feeling of What Happens : Body and emotion in the making of consciousness.* New York : Harcourt Brace.

[11] Matthew Ratcliffe, 2005. The Feeling of Being. *Journal of Consciousness Studies,* 12 (8-10), 45-63.

[12] Maša Poljšak Kus, 2024. *Unveiling of the atmosphere : etnophenomenological exploration of experiential background in relation to space* (Master thesis), Ljubljana. Supervisor : Urban Kordeš.

[13] Russell T. Hurblurt & Christopher L. Heavey, 2006. *Exploring inner experience : The descriptive experience sampling method*. Amsterdam, John Benjamins Publishing Co.

[14] Blaž Mesec, 2023. *Kvalitativno raziskovanje v teoriji in praksi*. Inštitut za razvojne in strateške analize.

Bridging the Challenges in Experience Sampling Research

Barbi Seme†
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
Barbi.Seme@pef.uni-lj.si

Maruša Sirk
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
Marusa.Sirk@pef.uni-lj.si

Urban Kordeš
Center for Cognitive Science
University of Ljubljana
Ljubljana, Slovenia
Urban.Kordes@pef.uni-lj.si

Abstract

In this paper, we draw parallels between existing research practices and attempt to piece them together to propose a more wholesome approach in conducting experience sampling research. We consider Experience Sampling Methods (ESM) as valuable tools for studying experience, but they come with challenges, of which we address the participant burden as one of the most significant ones. We think that integrating practices from Personal Science (PS) and Citizen Science (CS), grounded in empirical phenomenology, can help address this challenge. By considering participants as co-researchers and actively engaging them in the research and community, we aim to enhance their motivation and improve the quality of the research data. We illustrate this approach through the pilot project Luna in which we explore lived experiences throughout the menstrual cycle using the ESM mobile application "Curious". This integrative method facilitates a reciprocal knowledge exchange between researchers and co-researchers, which deepens the process of self-exploration and holds a great potential to advance scientific research on experience.

Keywords

Experience sampling methods, citizen science, personal science, empirical phenomenology

1 Introduction

Scientific research into experience is a rapidly growing field. Some researchers and philosophers point out that a core problem within our current scientific worldview is the overlooked experience research [1]. New methods and tools for researching experience are being developed, among which are Experience Sampling Methods (ESM). ESM are intensive longitudinal approaches to collecting experiential and contextual data using structured diary self-report techniques [2]. Due to numerous advantages, especially ecological validity and the reduction of recall bias, ESM has spread to various research fields through the use of mass technology, mostly mobile applications [3]. Based on numerous studies [4, 5, 6] that have used ESM to investigate experiential phenomena in the past few decades, weaknesses of these methods have been identified [7].

We present ESM and the challenges inherent in ESM research, particularly participant burden. By exploring the interest in personal exploration within Personal Science and emphasising the importance of community building in Citizen Science, we attempt to tie these practices together using the concept of a methodological turn from empirical phenomenology [8]. We believe that the challenge of participant burden, which we see as under addressed but highly disruptive in ESM scientific inquiry, can be tackled through this integration of different research practices. We illustrate this approach with our pilot study, Luna.

2 Experience sampling research

We consider Experience Sampling Methods (ESM) as an umbrella term for the research in which participants gather samples of their experiences as they unfold in their life [9]. Typically, we prompt participants at random times to answer questions or to describe their experience of the moment just before they heard the beep [10]. This way we are able to minimise recall bias [7] and are able to sample dimensions of experiential states which are nearly impossible to recall later, especially in detail (e.g., the momentary content of our thoughts). These methods are also highly ecologically valid, since we are sampling experience as it unfolds naturally in people's everyday lives. Participants would receive the prompts several times per day for a longer period of time (e.g., two weeks). These repeated measures enable us to track patterns and changes in individual experiences across time and different contexts [7, 10]. Nowadays we use mobile applications on participants' personal smartphones which makes the data collection process in comparison to pen and paper much more reliable and less burdensome [11].

2.1 Challenges in ESM

ESM research is still loosely defined without a rigorous framework and we are yet to develop appropriate methodological approaches for improvements [9]. A significant challenge in ESM research is participant burden. Collecting frequent, real-time data in everyday life activities puts great demands on participants who need to albeit shortly interrupt their activity to report on their momentary experiential state [3]. These repeated measures over time might affect participants' attitudes towards the research and result in reduced compliance, careless responding and participants' attrition [12]. We should also evaluate this burden from an ethical perspective, ensuring that

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

benefits outweigh the burden, especially if there is a consideration of affecting participants' well-being [13].

2.2 Existing recommendations for addressing the challenges in ESM

To mitigate this burden, researchers are working on questionnaire optimisation, making them as brief and focused as possible by prioritising essential questions and using clear and concise language [9]. The trade-off between data richness and burden on participants is also mitigated with lowering the sampling frequency [12]. By introducing the personalised scheduling for participants we avoid interrupting them in the situations in which they are unable to respond and would likely react negatively to prompts being delivered in that time [3]. Albeit we should then revise random sampling and account for the introduced bias (e.g., we could supplement the data with retrospective daily reports).

However, a key factor for successful ESM research is participants' motivation [3, 7]. It is better sustained by considering the necessary technical recommendations, but researchers should also give attention to fostering the research interest and social dynamics. Researchers should engage participants who already have interest in the research topic and therefore an intrinsic motivation to learn more on it [3]. They should also provide a sufficient training period in which participants gain the necessary knowledge and skills in order to sample the experience [7]. It is important to establish a rapport with participants and to foster a research alliance throughout the study [2]. It is suggested to provide rich feedback to participants during and at the end of the study which can also be presented as non-monetary compensation [3].

3 Personal science and interest for self-exploration

Technological advancement played a great role in a growing number of ESM studies as well as in an uptake of self-tracking practices for exploration of oneself [14, 15, 16]. The umbrella term for self-tracking practice and communities has been formed under Personal Science (PS) [17]. These individuals and groups pursue their own personal research questions using empirical methods in an iterative process of questioning, designing, observing, reasoning, and discovering which presents itself as an opportunity to scientifically expand on PS. Even though we can draw many parallels between self-tracking and scientific inquiry the question remains to what extent PS can be scientifically interesting [18]. Considering the growing interest in PS activities it seems important to address these practices, especially in new self-trackers. They often experience difficulties in making sense of their self-tracking process in interpreting their data, formulating and refining their research questions, and designing their research process [19, 20]. It would be beneficial for them to receive support that provides at least an initial establishment of their research or engaging them in a more systematic way. Lack of scientific rigour was also reported by researchers in tools used for self-tracking which can potentially mislead self-trackers and give them false ideas of phenomena they explore [21, 22]. Hence, we believe this is an excellent opportunity for the science community to engage in this already widespread phenomenon. Individuals and groups who already possess deep interest in self-exploration can potentially become great co-researchers by which they would gain support in their own exploration as well as make the scientific contribution.

4 Community in Citizen science

Citizen Science (CS) is recognized as one of the eight pillars of Open Science, playing an integral role in democratising scientific knowledge and practices [23]. It significantly bridges the gap between the scientific community and society through the idea of doing science and not merely reacting to it [24]. Due to the heterogeneity of CS projects in terms of scale, objectives, and levels of citizen scientist involvement, it is challenging to provide a universal definition [25]. However, common to all CS projects is to actively involve non-professionals in scientific research at different levels of participation [26]. In a broad sense citizen scientists perform tasks that would be otherwise done by scientists [27] or would not be possible to do without their involvement.

To achieve reciprocity between science and society in CS projects in which the bidirectional knowledge exchange facilitates benefits in both [28], significant time and resources need to be invested to establish the conditions for project activities to run [29]. Since citizen scientists are typically lay people without formal training in scientific research, appropriate training and support are essential to equip them with necessary skills and knowledge [30]. We know citizen scientists engage in the projects upon different motivation factors. We can observe the intrinsic factors, such as gaining fulfilment, enjoying the activities or being altruistic and extrinsic factors, such as building social interactions, gaining on reputation or status and expecting future returns [31]. Therefore, sustaining motivation and engagement requires more than just training. CS practitioners should establish good relationships with citizen scientists and a continuous communication as well as the conditions for citizen scientists to meet and work with each other. We argue that essential to the project's success is building a strong community. Utilising online community spaces, social media, organising workshops and training as well as local meetups, collaborative and other social activities facilitate community building. Strong community in exchange encourages participation, promotes knowledge sharing, foster collaboration and builds on sustainability of the project [32, 33].

5 Bridging ESM, PS and CS with empirical phenomenology

The key to integrating the practices of Citizen science, Personal science, and Experience Sampling research lies in the concept of a methodological turn developed in the field of empirical phenomenology [8]. In experience research, the observed is the observer, meaning that the only access to the phenomena of interest is through the observer's subjective experience. If the observer does not adopt an attitude of curious exploration and engage in epoché, meaning bracketing the natural attitude, the judgments, interpretations, and explanations of their experience, we cannot obtain data on the genuine experience as it unfolds in life. This notion is rooted in phenomenological reduction, a

method of research into experience developed by Edmund Husserl [34]. Experience Sampling has been used to study subjective experiences in real-time contexts, but integrating it with phenomenological reduction enhances the depth of data on lived experiences [35]. Therefore, it is necessary to consider our participants as co-researchers. This attitude allows us to engage them in a way that fosters their interest in the research question which facilitates the methodological turn where the research question becomes in a sense their own and they become researchers of their own experience. While providing the support and means for investigation, it is important to give co-researchers the freedom to explore the research question and their experience in a way that is meaningful to them, and to encourage critical discussion. By opening up the space for co-creation of the research design and enabling co-researchers to actively contribute their findings, we facilitate a deeper reciprocity of knowledge transfer.

6 The pilot study "Luna"

Citizen science project Luna aims to explore the lived experiences of menstrual cycles and their impact on everyday well-being. We use a diary method for daily reports and Experience Sampling Methodology (ESM) to track experiences throughout the menstrual cycle with the use of the ESM mobile application Curious (about) consciousness or Curious in short.

We adopted the iterative co-creation approach to develop our research, combining the principles of ESM research, CS projects, PS and empirical phenomenology. This makes our research process flexible in a way that the research design is being updated in an iterative collaborative manner. We engage co-researchers in the design and assessments of data collection procedure and questionnaires, data interpretation as well as analysis. By encouraging them to develop their own research questions which are relevant to their own experience we promote their personal endeavor in the research. They are also involved in other project activities, such as sharing results and designing the project visual identity. We organise different learning and sharing community activities. In the training workshops we introduced them to the principles of observing ones' own experience, how to bracket the natural attitude and report on the pristine inner experience. On our online community space as well as in the organised meetups we share feedback on the research and their participation, engage them in the conversation on the research topic related questions and encourage them to share their own feedback. They are also invited to share any findings they have along the way in the mobile application. The Curious app is designed in a way to support co-researchers in their data interpretation by providing them with simple visualisations of their gathered data which they are able to filter, compare and reflect on. Our aim in the research is two-fold. It is driven to answer the research questions on phenomenology throughout the menstrual cycle as well as to equip co-researchers with methods and tools to research their experiential landscape and gather self-knowledge. Project tries to go beyond research by also opening up the data driven discussion on possible systemic or social solutions to consider the physiological and phenomenological cyclic nature of people with menstrual cycles. This also presents a motivational factor for some co-researchers who would like to see positive (societal) changes in regard to this phenomenon.

7 Conclusion

We argue that a large number of research questions in ESM research could be better investigated if research projects adopt the CS framework with an emphasis on community building, account for the interest and practice of PS, and use the principles of experience investigation from empirical phenomenology. The challenge of participant burden in ESM research is then mitigated by creating conditions for co-researchers to be involved in personally meaningful activities, which in return provide a higher level of data validity. Even though considerable resources are needed to establish these kinds of project communities, we believe they have the potential to be more sustainable. From CS, we know that citizen scientists develop a sense of community, which encourages them to remain active in science after the initial project ends [36, 37]. This interdisciplinary integration of different research practices enhances the value of our investigations and creates more impactful and sustainable research projects that benefit both the scientific community, involved co-researchers and the society.

Acknowledgments

IMPETUS is supporting our project, Luna. IMPETUS is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101058677. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Frank, A., Gleiser, M., & Thompson, E. (2024). The blind spot: why science cannot ignore human experience. MIT Press. DOI:https://doi.org/110.7551/mitpress/13711.001.0001

[2] Hektner, J. M., Schmidt, J. A. & Csikszentmihalyi, M. (2007). Experience sampling method: Measuring the quality of everyday life. Thousand Oaks, CA: Sage Publications. DOI: https://doi.org/10.4135/9781412984201

[3] van Berkel, N., & Kostakos, V. (2021). Recommendations for conducting longitudinal experience sampling studies. Advances in longitudinal HCI research, 59-78. DOI: https://doi.org/110.1016/10.1007/978-3-030-67322-2_4

[4] MacKerron, G. in Mourato, S. (2013). Happiness is greater in natural environments. Global environmental change, 23(5), 992-1000. DOI:https://doi.org/110.1016/j.gloenvcha.2013.03.010

[5] Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I. in Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. Psychological science, 18(7), 614-621. DOI:https://doi.org/110.1111/j.1467-9280.2007.01948.x

[6] Mölsä, M. E., Lax, M., Korhonen, J., Gumpel, T. P., in Söderberg, P. (2022). The experience sampling method in monitoring social interactions among children and adolescents in school: a systematic literature review. Frontiers in Psychology, 13, 844698. DOI:https://doi.org/110.3389/fpsyg.2022.844698

[7] Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. Journal of Happiness studies, 4(1), 5-34. DOI:https://doi.org/10.1023/A:1023605205115

[8] Kordeš, U. in Klauser, F. (2016). Second-Person in-Depth Phenomenological Inquiry as an Approach for Studying Enaction of Beliefs. Interdisciplinary Description of Complex Systems, 14(4), 369-377. DOI: https://doi.org/10.7906/indecs.14.4.5

[9] Myin-Germeys, I. in Kuppens, P. (Ur.). (2022) The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies (2nd ed.). Leuven: Center for Research on Experience Sampling and Ambulatory Methods Leuven.

[10] Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., Lebo, K., & Kaschub, C. (2003). A practical guide to experience-sampling procedures. Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being, 4(1), 53–78. DOI:https://doi.org/10.1023/A:1023609306024

[11] van Berkel, N. V., Ferreira, D. in Kostakos, V. (2017). The Experience Sampling Method on Mobile Devices. ACM Computing Surveys, 50(6), 1–40. DOI: https://doi.org/10.1145/3123988

[12] Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Assessment, 29(2), 136-151. DOI: https://doi.org/10.1177/1073191120957102

[13] Kirtley, O. (2022). The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies (2nd ed.). I. Myin-Germeys in P. Kuppens (ur.), Ethical issues in experience sampling method research. Leuven: Center for Research on Experience Sampling and Ambulatory Methods Leuven.

[14] Clark, M., Southerton, C., & Driller, M. (2024). Digital self-tracking, habits and the myth of discontinuance: It doesn't just 'stop'. new media & society, 26(4), 2168-2188. DOI: https://doi.org/10.1177/14614448221083992

[15] Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. Big Data, 1(2), 85–99. DOI: https://doi.org/10.1089/big.2012.0002

[16] Lupton, D. (2016). The Quantified Self: A Sociology of Self-Tracking. Polity Press.

[17] Wolf, G. I., & De Groot, M. (2020). A conceptual framework for personal science. Frontiers in Computer Science, 2, 21. DOI: https://doi.org/10.3389/fcomp.2020.00021

[18] Heyen, N. B. (2016). Self-tracking as knowledge production: Quantified self between prosumption and citizen science. In Lifelogging: Digital self-tracking and lifelogging-between disruptive technology and cultural transformation (pp. 283-301). Wiesbaden: Springer Fachmedien Wiesbaden. DOI: https://doi.org/10.1007/978-3-658-13137-1_16

[19] Rapp, A., & Cena, F. (2016). Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. International Journal of Human-Computer Studies, 94, 1-17. DOI: https://doi.org/10.1016/j.ijhcs.2016.05.006

[20] Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1143–1152). ACM. DOI: https://doi.org/10.1145/2556288.2557372

[21] Caldeira, C., Chen, Y., Chan, L., Pham, V., Chen, Y., & Zheng, K. (2017). Mobile apps for mood tracking: an analysis of features and user reviews. In AMIA annual symposium proceedings (Vol. 2017, p. 495). American Medical Informatics Association.

[22] Larsen, M. E., Nicholas, J., & Christensen, H. (2016). Quantifying app store dynamics: Longitudinal tracking of mental health apps. JMIR mHealth and uHealth, 4(3), e96. DOI: https://doi.org/10.2196/mhealth.6020

[23] Ayris, P., Lopez de San Román, A., Maes, K., & Labastida, I. (2018). Open science and its role in universities: A roadmap for cultural change. Leuven: LERU Office. Retrieved December, 13, 2019.

[24] Lang, D. (2016). "Science to the People. How citizen science bridges the gap between science and society." In: Medium. Retrieved 22nd of Sep. 2024 from: https://fellowsblog.ted.com/how-citizen-science-bridges-the-gap-between-science-and-society-d693af125ae4

[25] Haklay, M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., Vohland, K., ... & Wagenknecht, K. (2021). What is citizen science? The challenges of definition. The science of citizen science, 13. DOI: https://doi.org/10.1007/978-3-030-58278-4_2

[26] Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M. E., & Bonney, R. (2012). Public Participation in Scientific Research: a Framework for Deliberate Design. Ecology and Society, 17(2). DOI:https://doi.org/10.5751/ES-04705-170229

[27] Heiss, R. & Matthes, J. (2017). Citizen Science in the Social Sciences: A Call for More Evidence. In: GAIA - Ecological Perspectives for Science and Society 26, pp. 22–26. DOI:https://doi.org/10.14512/gaia.26.1.7.

[28] Phillips, T., Porticella, N., Constas, M., & Bonney, R. (2018). A framework for articulating and measuring individual learning outcomes from participation in citizen science.

[29] Heinisch, B., Oswald, K., Weißpflug, M., Shuttleworth, S., & Belknap, G. (2021). Citizen humanities. The science of citizen science, 97. DOI:https://doi.org/10.1007/978-3-030-58278-4_6

[30] Pandya, R. E. (2012). A framework for engaging diverse communities in citizen science in the US. Frontiers in Ecology and the Environment, 10(6), 314–317. DOI: https://doi.org/10.1890/120007

[31] Lotfian M, Ingensand J, Brovelli MA. (2020). A Framework for Classifying Participant Motivation that Considers the Typology of Citizen Science Projects. ISPRS International Journal of Geo-Information. 2020; 9(12):704. DOI: https://doi.org/10.3390/ijgi9120704

[32] West, S., & Pateman, R. (2016). Recruiting and retaining participants in citizen science: What can be learned from the volunteering literature? Citizen Science: Theory and Practice, 1(2), 15. DOI: https://doi.org/10.5334/cstp.8

[33] Bonney, R., Phillips, T. B., Ballard, H. L., & Enck, J. W. (2016). Can citizen science enhance public understanding of science? Public Understanding of Science, 25(1), 2–16. DOI: https://doi.org/10.1177/0963662515607406

[34] Husserl, E.: Aufsätze und Vorträge (1911–1921). Husserliana XXV. Martinus Nijhoff, Doedrecht, 1987.

[35] Hurlburt, R. T. in Heavey, C. L. (2006). Exploring inner experience: The descriptive experience sampling method (Vol. 64). John Benjamins Publishing. DOI: https://doi.org/10.1075/aicr.64

[36] Rotman, D., et al. (2012). Dynamic changes in motivation in collaborative citizen-science projects. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 217–226). ACM. DOI: https://doi.org/10.1145/2145204.2145238

[37] Bela, G., Peltola, T., Young, J. C., Balázs, B., Arpin, I., Pataki, G., ... & Bonn, A. (2016). Learning and the transformative potential of citizen science. Conservation Biology, 30(5), 990-999. DOI: https://doi.org/10.1111/cobi.12762

Indeks avtorjev / Author index