INFORMACIJSKA DRUŽBA

Zbornik 26. mednarodne multikonference

Zvezek B

INFORMATION SOCIETY

Proceedings of the 26th International Multiconference

Volume B

Kognitivna znanost

Cognitive Science

Uredniki • Editors:
Anka Slana Ozimič, Borut Trpin,
Toma Strle, Olga Markič

12. oktober 2023 | Ljubljana, Slovenija • 12 October 2023 | Ljubljana, Slovenia

IS2023

Zbornik 26. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2023

Zvezek B

Proceedings of the 26th International Multiconference

INFORMATION SOCIETY – IS 2023

Volume B

Kognitivna znanost
Cognitive Science

Uredniki / Editors

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

http://is.ijs.si

12. oktober 2023 / 12 October 2023
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, Mateja Mavrič
Oblikovanje naslovnice: Vesna Lasič

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

Ljubljana, oktober 2023

Informacijska družba
ISSN 2630-371X

Katadožni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 170773251
ISBN 978-961-264-277-8 (PDF)

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2023

Šestindvajseta multikonferenca Informacijska družba se odvija v obdobju izjemnega razvoja za umetno inteligenco, računalništvo in informatiko, za celotno informacijsko družbo. Generativna umetna inteligenca je s programi kot ChatGPT dosegla izjemen napredek na poti k superinteligenci, k singularnosti in razcvetu človeške civilizacije. Uresničujejo se napovedi strokovnjakov, da bodo omenjena področna ključna za obstoj in razvoj človeštva, zato moramo pozornost usmeriti na njih, jih hitro uvesti v osnovno in srednje šolstvo in vsakdan posameznika in skupnosti.

Po drugi strani se poleg lažnih novic pojavljajo tudi lažne enciklopedije, lažne znanosti ter »ploščate Zemlje«, nadaljuje se zapostavljanje znanstvenih spoznanj, metod, zmanjševanje človekovih pravic in družbenih vrednot. Na vseh nas je, da izzive današnjice primerno obravnavamo, predvsem pa pomagamo pri uvajanju znanstvenih spoznanj in razčiščevanju zmot. Ena pogosto omenjanih v zadnjem letu je eksistencialna nevarnost umetne inteligence, ki naj bi ogrožala človeštvo tako kot jedrske vojne. Hkrati pa nihče ne poda vsaj za silo smiselnega scenarija, kako naj bi se to zgodilo – recimo, kako naj bi 100x pametnejši GPT ogrozil ljudi.

Letošnja konferenca poleg čisto tehnoloških izpostavlja pomembne integralne teme, kot so okolje, zdravstvo, politika depopulacije, ter rešitve, ki jih za skoraj vse probleme prinaša umetna inteligenca. V takšnem okolju je ključnega pomena poglobljena analiza in diskurz, ki lahko oblikujeta najboljše pristope k upravljanju in izkoriščanju tehnologij. Imamo veliko srečo, da gostimo vrsto izjemnih mislecev, znanstvenikov in strokovnjakov, ki skupaj v delovnem in akademsko odprtem okolju prinašajo bogastvo znanja in dialoga. Verjamemo, da je njihova prisotnost in udeležba ključna za oblikovanje bolj inkluzivne, varne in trajnostne informacijske družbe. Za razcvet.

Letos smo v multikonferenco povezali deset odličnih neodvisnih konferenc, med njimi »Legende računalništva«, s katero postavljamo nov mehanizem promocije informacijske družbe. IS 2023 zajema okoli 160 predstavitev, povzetkov in referatov v okviru samostojnih konferenc in delavnic, skupaj pa se je konference udeležilo okrog 500 udeležencev. Prireditev so spremljale okrogle mize in razprave ter posebni dogodki, kot je svečana podelitev nagrad. Izbrani prispevki bodo izšli tudi v posebni številki revije Informatica (http://www.informatica.si/), ki se ponaša s 46-letno tradicijo odlične znanstvene revije. Multikonferenco Informacijska družba 2023 sestavljajo naslednje samostojne konference:

- Odkrivanje znanja in podatkovna središča
- Demografske in družinske analize
- Legende računalništva in informatike
- Konferenca o zdravi dolgoživosti
- Miti in resnice o varovanju okolja
- Mednarodna konferenca o prenosu tehnologij
- Digitalna vključenost v informacijski družbi – DIGIN 2023
- Slovenska konferenca o umetni inteligenci + DATASCIENCE
- Kognitivna znanost
- Vzgoja in izobraževanje v informacijski družbi
- Zaključna svečana prireditev konference

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

S podelitvijo nagrad, še posebej z nagrado Michie-Turing, se avtonomna stroka s področja opredeli do najbolj izstopajočih dosežkov. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel prof. dr. Andrej Brodnik. Priznanje za dosežek leta pripada Benjaminu Bajdu za zlato medaljo na računalniški olimpijadi. »Informacijsko limono« za najmanj primerno informacijsko tematiko je prejela nekompatibilnost zdravstvenih sistemov v Sloveniji, »informacijsko jagodo« kot najboljšo potezo pa dobi ekipa RTV za portal dostopno.si. Čestitke nagrajencem!

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

FOREWORD - INFORMATION SOCIETY 2023

The twenty-sixth Information Society multi-conference is taking place during a period of exceptional development for artificial intelligence, computing, and informatics, encompassing the entire information society. Generative artificial intelligence has made significant progress towards superintelligence, towards singularity, and the flourishing of human civilization with programs like ChatGPT. Experts' predictions are coming true, asserting that the mentioned fields are crucial for humanity's existence and development. Hence, we must direct our attention to them, swiftly integrating them into primary, secondary education, and the daily lives of individuals and communities.

On the other hand, alongside fake news, we witness the emergence of false encyclopaedias, pseudo-sciences, and flat Earth theories, along with the continuing neglect of scientific insights and methods, the diminishing of human rights, and societal values. It is upon all of us to appropriately address today's challenges, mainly assisting in the introduction of scientific knowledge and clearing up misconceptions. A frequently mentioned concern over the past year is the existential threat posed by artificial intelligence, supposedly endangering humanity as nuclear wars do. Yet, nobody provides a reasonably coherent scenario of how this might happen, say, how a 100x smarter GPT could endanger people.

This year's conference, besides purely technological aspects, highlights important integral themes like the environment, healthcare, depopulation policies, and solutions brought by artificial intelligence to almost all problems. In such an environment, in-depth analysis and discourse are crucial, shaping the best approaches to managing and exploiting technologies. We are fortunate to host a series of exceptional thinkers, scientists, and experts who bring a wealth of knowledge and dialogue in a collaborative and academically open environment. We believe their presence and participation are key to shaping a more inclusive, safe, and sustainable information society. For flourishing.

This year, we connected ten excellent independent conferences into the multi-conference, including "Legends of Computing", which introduces a new mechanism for promoting the information society. IS 2023 encompasses around 160 presentations, abstracts, and papers within standalone conferences and workshops. In total about 500 participants attended the conference. The event was accompanied by panel discussions, debates, and special events like the award ceremony. Selected contributions will also be published in a special issue of the journal Informatica (http://www.informatica.si/), boasting a 46-year tradition of being an excellent scientific journal. The Information Society 2023 multi-conference consists of the following independent conferences:

- Data Mining and Data Warehouse - SIKDD
- Demographic and Family Analysis
- Legends of Computing and Informatics
- Healthy Longevity Conference
- Myths and Truths about Environmental Protection
- International Conference on Technology Transfer
- Digital Inclusion in the Information Society - DIGIN 2023
- Slovenian Conference on Artificial Intelligence + DATASCIENCE
- Cognitive Science
- Education and Training in the Information Society
- Closing Conference Ceremony

Co-organizers and supporters of the conference include various research institutions and associations, among them ACM Slovenia, SLAIS for Artificial Intelligence, DKZ for Cognitive Science, and the Engineering Academy of Slovenia (IAS). On behalf of the conference organizers, we thank the associations and institutions, and especially the participants for their valuable contributions and the opportunity to share their experiences about the information society with us. We also thank the reviewers for their assistance in reviewing.

With the awarding of prizes, especially the Michie-Turing Award, the autonomous profession from the field identifies the most outstanding achievements. Prof. Dr. Andrej Brodnik received the Michie-Turing Award for his exceptional lifetime contribution to the development and promotion of the information society. The Achievement of the Year award goes to Benjamin Bajd, gold medal winner at the Computer Olympiad. The "Information Lemon" for the least appropriate information move was awarded to the incompatibility of information systems in the Slovenian healthcare, while the "Information Strawberry" for the best move goes to the RTV SLO team for portal dostopno.si. Congratulations to the winners!

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
Mateja Mavrič

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 26. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2023

Zvezek B

Proceedings of the 26th International Multiconference

INFORMATION SOCIETY – IS 2023

Volume B

Kognitivna znanost
Cognitive Science

Uredniki / Editors

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

http://is.ijs.si

12. oktober 2023 / 12 October 2023
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. Skupaj bomo potovali skozi različna področja kognitivne znanosti - od psihologije in nevroznanosti do filozofije in umetne inteligence, ter ob tem spoznavali raznolike tematike vključujoč psihedelike, mistične izkušnje, biomarkerje kognitivnih sposobnosti, in celo vprašanja zavesti.

Poseben poudarek letošnjega srečanja je namenjen eni izmed trenutno najbolj vročih tem: umetni inteligenci. Osrednja tema konference, "UI klepetalniki in širše", bo predstavila izzive in rešitve, ki jih prinašata razvoj in uporaba klepetalnih robotov z umetno inteligenco. Hkrati pa bomo razmišljali, kako umetna inteligenca oblikuje svet onkraj klepetalnih robotov. V tej luči bomo na konferenci gostili okroglo mizo o vlogi umetne inteligence v izobraževanju, s čimer se bomo dotaknili še enega izmed aktualnih izzivov. Skupaj bomo razmišljali o prednostih in pasteh njenega vključevanja v izobraževalne procese, ki oblikujejo našo prihodnost.

Upamo, da bo letošnja konferenca prostor za povezovanje in izmenjavo prodornih 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. Together, we will travel through the diverse fields of cognitive science - from psychology and neuroscience to philosophy and artificial intelligence, learning about a variety of topics, including psychedelics, mystical experiences, biomarkers of cognitive abilities, and even questions of consciousness.

This year's conference has a special focus on one of the hottest topics at the moment: artificial intelligence. The main topic of the conference, "AI Chatbots and Beyond," will present the challenges and solutions brought about by the development and use of AI chatbots. At the same time, we will consider how AI is shaping the world beyond chatbots. In this light, we will host a panel discussion on the role of AI in education, addressing another of the current challenges. Together, we will reflect on the benefits and pitfalls of integrating it into the educational processes that are shaping our future.

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, 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

What insights can psychedelic research bring to Cognitive Science? A systematic review of the phenomenology of DMT experiences.

Carolina Czizek[†]
Cognitive Science
University of Vienna & University of Ljubljana
Vienna – Austria & Ljubljana - Slovenia
carolina.czizek@chello.at

ABSTRACT

This abstract explores the systematic review of the phenomenology of N, N-dimethyltryptamine (DMT) experiences. Additionally, the relevance of conducting psychedelic research for Cognitve Science is discussed. A special emphasis is being put on (neuro-)phenomenological research methods, as they seem to be the best suitable for conducting research around psychedelic substances and their direct effects on phenomenology.

KEYWORDS

Psychedelic research, Cognitive Science, systematic review, phenomenology, neurophenomenology, DMT, non-ordinary states of consciousness

1 Introduction

N, N-dimethyltryptamine (DMT) is an endogenous serotonergic psychedelic compound which is capable of producing radical shifts in conscious experience. Compared to other serotonergic psychedelic substances like psilocybin, mescaline or LSD; DMT experiences seem to produce the most radical shifts in conscious experience, with subjects reporting hyper-real, otherworldly, often ontologically challenging but also potential transformative experiences, including encounters with entities as well as experiencing visualizations of geometric fractals, shapes or patterns. The rapid onset and short duration of inhaled DMT experiences, the drastic change in phenomenological conscious experience it produces, as well as the fact that some studies indicate the compound to be endogenous to mammals [3] as well as to a variety of plants, suggests the importance of conducting more research around this compound.

While DMT use has been part of several ancient Amazonian traditions (combined with monoamide-oxidase inhibitors it is called 'ayahuasca') for hundreds of years and trends in usage in the western world as well as clinical trials of administrating DMT to patients with treatment resistant depression or anxiety, are increasing. This is done by administrating DMT either inhaled, which makes the effects shorter lived (about 15 minutes) or taken orally as ‚ayahuasca' to prolong the subjective effects of the compound (up to ten hours). Laboratory studies of DMT use are limited by their clinical setting (not taking into account the importance of set & setting of psychedelic experiences) and most are lacking a qualitative analysis of phenomenological content. This indicates the growing importance of a thorough investigation of the phenomenological aspects of the substances.

Latest research indicates similarities of phenomenological experiences of DMT use across subjects [5]. It remains unclear how much of these similarities are due to cultural or individual priming and/or influencing, as stories about "DMT entities" or the "DMT parallel world" can be found all over the internet.

2 Systematic Review

The systematic review synthesizes the phenomenology of N,N-dimethyltryptamine experiences according to the PRISMA method [8]. Excluded were studies prior to 2013 and studies that not taking into account the phenomenological effects of the DMT experience. After initial screening, 17 studies were included in this report. The included studies used different methodologies: neurophenomenological approaches through EEG measurements followed by questionnaires or interviews; purely phenomenological accounts through online surveys; micro-phenomenological inspired interviews or qualitative linguistic analyses. Also assessed were studies that compared the phenomenology of DMT experiences to other experiences such as near-death experiences (NDE's), ego-dissolution experiences or God-encounter experiences.

Results indicate that the DMT experience could be clustered into different categories. When it comes to the phenomenological perspective, the most prominent themes identified are: perceptual changes; somatic experiences; emotional responses and a 'sense of otherness' and a sense of ego dissolution. Additionally, to these themes, the most prominent categories regarding the content of the DMT experience are:

1. The ontology of the DMT world, also called ‚hyperspace'

2. Dissolution of the ego/self

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

3. Encounters with seemingly autonomous or conscious entities. These encounters are predominantly positive, users often reporting of receiving a type of message.

Limitations of the studies were quite homogenous, including among others: the use of self-reports; sociability biases; retrospective accounts; different or insufficient information on dosages and purity. The systematic review concludes that the phenomenology of DMT experiences seems to be distinct to other psychedelic experiences such as psilocybin but similar to certain other non- drug experiences, such as near-death experiences, thus representing a research field well suited for the account of the phenomenology of non-ordinary states of consciousness.

3 Relevance to Cognitive Science

„Psychedelics are for the mind/psychiatry what the telescope is for astronomy or the microscope is for biology "- [4]

Psychedelic research, ever since legislation loosened up in the late 2000s, is experiencing a come-back after 40 years of prohibition. Since mental health issues are becoming more prevalent and common approaches with medication such as SSRIs are not yielding sufficient relief, psychedelic assisted therapy approaches seem to be a promising approach. More clinical research is necessary to drive this process further. Other studies have shown psilocybin to promote neurogenesis, which could be used to treat non-psychiatric but also biological issues.

Classical psychedelics act on the serotonin 5HT2A receptor, and are molecules that can drastically change the phenomenology of experience. They modulate fundamental aspects of experience by what it seems to be *deconstructing prior beliefs* and *reconstructing new beliefs.* Some theories about the mind and consciousness have emerged through psychedelic research as well, such as the 'entropic brain hypothesis' [1], bringing physical concepts of *entropy* and *criticality* into the discussion of non-ordinary states of consciousness and their neurophenomenological characteristics. The REBUS ('relaxed beliefs under psychedelics') model, is a model that combines the entropic brain hypothesis with the free energy principle, trying to gain further understanding of the effects of free energy on phenomenology [2]. What insights can we generate about the nature of reality when approaching it with fewer predictions?

3.1. The sense of self

One of the core questions of (philosophical) Cognitive Science is understanding the sense of self. Since psychedelics seem to usually lead to a deconstruction of the sense of self (f.e. explained due to a diminished activity in the default mode network) [6], the different processes of selfhood (minimal self vs. narrative self) [10] could be examined further as well as how this construction and deconstruction of the self emerges on a neurophenomenological level, what psychological implications this has on individuals yielding up to philosophical discussions of the ontological role of the sense of self. Another classical argument for studying psychedelics is backward propagation (finding out about aberrant functions can help generate insights about 'normal' function), which has already helped shape the field of neuroscience in the past. The serotonergic structure of classical psychedelics has already yielded insights into the role of serotonin on our wellbeing and perception of reality in the past, the discovery of LSD being one of the core drivers to the investigation of this molecule in the 1960s [7].

To study non-ordinary states of consciousness, such as hypnosis, meditation and psychedelics, common scientific practice (mostly average rating over the entire course of the experience) does not yield sufficient insights into the individual phenomenological processes that include such an experience. First-person reports can account for within- & between-subject variabilities. A strong call for neurophenomenological research in this field is being evoked, emphasizing the importance of conducting micro-phenomenological interviews [9], additionally to third-person research (brain & somatic measurements).

3.2. DMT distinct from other psychedelic substances

DMT seems to be distinct from other substances on a neurophenomenological level. Firstly, the endogenous production of DMT is still a mystery to research. Under the influence of DMT, delta power, which is usually associated with states of unconsciousness or lack of experience (such as dreamless sleep or anesthesia), increases [10]. This can hint at delta waves being a marker from conscious disconnection, while the person is still having a phenomenological experience. Since this seems to be unique to DMT, along with the short duration of the experience, DMT research could help generate insights into how our brains construct the world, how our sense of self is constructed and how all of these processed can be deconstructed in minutes or even seconds due to a single (endogenous) chemical. Neurophenomenological approaches with non-ordinary states of consciousness are crucial in the quest to finding answers to some of the most mysterious questions of cognitive science.

AKNOWLEDGEMENTS

I want to thank the University of Ljubljana and their faculty of Cognitive Science for teaching us the topics of phenomenology and neurophenomenology and teaching us how to implement these approaches. The experience of studying one semester at the University of Ljubljana has shaped me as a researcher and has given me motivation to follow this career path. I want to especially thank Dr. Toma Strle for encouraging me to follow my interests, no matter how controversial these are.

REFERENCES

[1] Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., & Nutt, D. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience, 8, 20.* https:// doi.org/10.3389/fnhum.2014.00020

[2] Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics. *Pharmacological reviews*, *71*(3), 316–344. https://doi.org/10.1124/pr.118.017160

[3] Dean, J. G., Liu, T., Huff, S., Sheler, B., Barker, S. A., Strassman, R. J., Wang, M. M., & Borjigin, J. (2019). Biosynthesis and Extracellular Concentrations of N,N- dimethyltryptamine (DMT) in Mammalian Brain. *Scientific reports*, *9*(1), 9333. https:// doi.org/10.1038/s41598-019-45812-w

[4] Grof, S (2008). LSD psychotherapy: The healing potential of psychedelic medicine. Ben Lomond, CA: Multidisciplinary Association for psychedelic Studies (MAPS)

[5] Lawrence, D.W., Carhart-Harris, R., Griffiths, R. et al. (2022) Phenomenology and content of the inhaled *N*, N-dimethyltryptamine (*N*, *N*-DMT) experience. *Sci Rep 12, 8562*. https://doi.org/ 10.1038/s41598-022-11999-8

[6] Millière, R. (2017). Looking for the Self: Phenomenology, Neurophysiology and Philosophical Significance of Drug-induced Ego Dissolution. *Frontiers in Human Neuroscience, 11, 245*. https:// doi.org/10.3389/fnhum.2017.00245

[7] Nichols D. (2013). Serotonin, and the Past and Future of LSD. *MAPS Bulletin Special Edition.*

[8] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLOS Medicine 2021;18(3):e1003583. doi: 10.1371/journal.pmed.1003583

[9] Petitmengin, C. Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenom Cogn Sci* **5**, 229–269 (2006). https://doi.org/10.1007/s11097-006-9022-2

[10] Timmermann, Bauer, P. R., Gosseries, O., Vanhaudenhuyse, A., Vollenweider, F., Laureys, S., Singer, T., Mind and Life Europe (MLE) ENCECON Research Group, Antonova, E., & Lutz, A. (2023). A neurophenomenological approach to non-ordinary states of consciousness: hypnosis, meditation, and psychedelics. *Trends in cognitive sciences*, *27*(2), 139–159. https://doi.org/10.1016/j.tics.2022.11.006

Pristranskost v strojnem učenju: dileme in odgovori

Bias in Machine Learning: Dilemmas and Answers

Ana Farič[†]
Pedagoška fakulteta
Univerza v Ljubljani
Slovenija
af27987@student.uni-lj.si

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

POVZETEK

Nekatere aplikacije umetne inteligence, posebej strojnega učenja, so deležne izrazito odklonilnih odzivov tako v splošnih medijih, kot v strokovni literaturi. Pogosto so omenjane aplikacije v domenah sodstva, zaposlovanja in bančništva. Kritiki očitajo, da so uporabljeni sistemi pristranski glede na t.i. zaščitene atribute, kot so rasa, spol in starost. Znan primer je sistem COMPAS, ki se kljub polemikam še vedno uporablja v ameriškem sodstvu. Namen prispevka je na primeru COMPASA predstaviti trende diskusije o pristranskosti algoritmov strojnega učenja. Opažamo, da je problem pogosto v tem, da niti v strokovni literaturi s področja umetne inteligence ni soglasja glede tehničnih definicij pristranskosti, ki bi jih bilo mogoče operativno uporabljati za preprečevanje (videza) pristranskosti. Naši zaključki so, da je (1) potrebno s kvalitetno izobrazbo doseči boljše splošno razumevanje metod umetne inteligence v praksi in (2) da je potrebno razviti tehnične principe, s katerimi bi v sistemih umetne inteligence operacionalizirali splošno sprejete družbene vrednote, kot sta enakost in pravičnost.

KLJUČNE BESEDE

umetna inteligenca, strojno učenje, pristranskost, diskriminacija, pravičnost

ABSTRACT

Some recent applications of Artificial Intelligence, particularly machine learning, have been strongly criticised in general media and professional literature. Applications in domains of justice, employment and banking are often mentioned in this respect. The main critic is that these applications are biased with respect to so called protected attributed, such as race, gender and age. The most notorious example is the system COMPAS which is still in use in American justice system despite severe criticism. The aim of our paper is to analyse the trends of discussion about bias in machine learning algorithms using the COMPAS as an example. The main problem of such discussions is that even in the field of AI, there is no generally agreed technical definition of bias which would enable operational use in preventing bias. Our conclusions are that (1) improved general education is needed to enable better understanding of AI methods in everyday applications, and (2) technical methods must be developed for implementing generally accepted societal values such as equality and fairness in AI systems.

KEYWORDS

machine learning, artificial intelligence, bias, fairness, discrimination

1 UVOD

Z razmahom uporabe strojnega učenja so se v zadnjih 5 do 10 letih pojavili primeri aplikacij, ki so bile deležne izrazito odklonilnih odzivov predvsem s strani splošnih medijev, pa tudi znotraj strokovne literature. Pogosto so omenjani sistemi, uporabljeni v domenah sodstva, zaposlovanja in bančništva. Kritiki opozarjajo, da so "algoritmi in sistemi strojnega učenja nepravični in pristranski" glede na t.i. zaščitene atribute, kot so rasa, spol in starost posameznika in da so priporočila umetne inteligence odvisna od teh atributov, namesto od objektivnega ocenjevanja dejstev. Naslovi nekaterih odmevnih člankov so: *There's software used across the country to predict future criminals. And it's biased against blacks* [2], *New Zealand passport robot tells applicant of Asian descent to open eyes* [18], *A beauty contest was judged by AI and the robots didn't like dark skin* [13], *Amazon scraps secret AI recruiting tool that showed bias against women* [6]. Taki primeri prispevajo k stopnjevanju skrbi o vplivih, ki ga ima umetna inteligenca (v nadaljevanju UI) na naša življenja [15]. Strokovnjaki z različnih področij se lotevajo t.i. problema pristranskosti strojnega učenja. Skušajo definirati, kaj pristranskost pomeni, iz kje naj bi izhajala, predvsem pa, kaj naj bi glede tega storili.

Na razvijajočem se področju etike v UI (npr. UNESCO 2021 [19]) se tema pristranskosti strojnega učenja pojavlja na vidnem mestu. Pogosto jo omenjajo politiki v zvezi s principi regulacije, ki naj bi zagotovila etično uporabo umetne inteligence (npr. European AI Act, 2023 [3]). Vendar v teh diskusijah pogosto ni jasno, kaj točno pristranskost strojnega učenja in UI pomeni. Zato regulacijski ukrepi v tej smeri niso jasno opredeljeni, razen v zelo abstraktni obliki. Beseda pristranskost v zvezi s strojnim

[*] Article Title Footnote needs to be captured as Title Note
[†] Author Footnote to be captured as Author Note

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

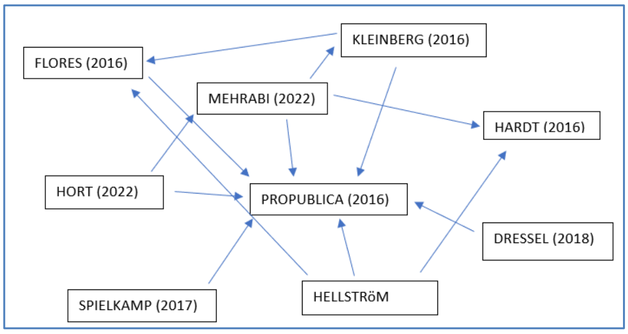
učenjem avtorjem pomene različne stvari. Celo v strokovni literaturi s področja UI ni popolnega soglasja in nedvomno sprejetih tehničnih definicij pristranskosti, ki bi jih bilo mogoče operativno uporabljati v preprečevanju pristranskosti [11]. Za razne smiselne definicije mer pristranskosti je celo matematično dokazano, da jim razen v posebnih primerih ni mogoče zadostiti hkrati [12].

V prispevku pregledamo razne definicije pristranskosti in različna mnenja o tem, kako naj bi problem najbolj učinkovito naslovili v praksi. Zaključki konvergirajo k temu, da je za ustrezno obravnavo potrebno upoštevati družbene vrednote in jih operacionalizirati z demokratično sprejetim družbenim dogovorom v obliki ustrezne zakonodaje. K dobremu splošnemu razumevanju pristranskosti v UI v praksi pa bi pripomogla boljša splošna izobrazba o UI in njenih metodah.

2 COMPAS

Sistem COMPAS je bil v vrsti publikacij obravnavan kot verjetno najbolj kontroverzen primer, ki naj bi ilustriral pristransko delovanje UI. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) je odločitveni sistem, ki ga na mnogih ameriških sodiščih uporabljajo sodniki za oceno tveganja povratništva, konkretno, da bo obsojenec v roku dveh let ponovil prestopek, če bo izpuščen. COMPAS je razvilo ameriško podjetje, takrat imenovano Northpointe. COMPAS upošteva 137 podatkov o vsakem prestopniku. Te podatke analizira poseben algoritem, ki kot poslovna skrivnost podjetja ni splošno znan. Na osnovi te analize algoritem poda oceno, ali gre za visoko ali nizko tveganje povratništva.

Slika 1 ilustrira 9 močno citiranih člankov o tem sistemu ter medsebojno citiranje med članki. Puščica iz članka A v članek B pomeni, da je B citiran v A.



Slika 1: Medsebojna povezanost objav o sistemu COMPAS

V središču grafa je članek iz časopisa ProPublica [2], ki naj bi po [1] sprožila zanimanje za preučevanje pristranskosti v UI. V [2] je skupina raziskovalnih novinarjev opisala svojo analizo sistema COMPAS in poskuse s primeri realnih podatkov o več kot 7000 obtožencih iz Floride v letih 2012 in 2013. V analizi se osredotočijo predvsem na rasni vidik in njihov zaključek je, da je program pristranski do temnopoltih obtožencev. Spremljali so, koliko od teh je bilo v naslednjih dveh letih ponovno obsojenih in primerjali napovedi z dejanskimi izidi. 44.9% temnopoltih, označenih z visokim tveganjem za ponovitev, ni ponovilo prestopka. 47.7% belcev, označenih z nizkim tveganjem, pa je v roku dveh let ponovilo prestopek. Ti dve merili za napačne napovedi sistema se standardno imenujeta: (1) FPR (false positive rate), to je delež negativnega razreda, ki je bil napačno napovedan kot pozitiven in (2) FNR (false negative rate), to je delež pozitivnega razreda, ki je bil napačno napovedan kot negativen. Kompletni rezultati glede napačnih deležev v ocenah sistema COMPAS so:

	Beli	Temnopolti
FPR	23.5%	44.9%
FNR	47.7%	28.0%

Te rezultate so Angwin idr. [2] interpretirali kot očitno pristranske do temnopoltih in zato ocenili uporabo sistema COMPAS kot neprimerno in diskriminatorno. Taki interpretaciji bi bilo težko nasprotovati. Dodatni problem so videli v dejstvu, da odločitveni kriterij, ki ga uporablja COMPAS, ni transparenten, saj je algoritem varovan kot poslovna skrivnost. COMPAS sam pa ne poda razlage svoje napovedi. Ta članek je zelo pogosto citiran in posledično je COMPAS postal najbolj znan primer pristranskosti v strojnem učenju, tako v strokovnih krogih domene strojnega učenja, kot pri splošni publiki, ki nima strokovnega znanja o umetni inteligenci. Kljub temu se COMPAS še vedno uporablja.

Na članek iz ProPublice je odgovorila skupina strokovnjakov iz ameriškega pravosodja v članku z zgovornim naslovom "*False positives, false negatives and false analysis: a rejoinder to Machine Bias ...*" [8]. Navedli so več spornih odločitev v analizi ProPublice, izvedli svojo lastno eksperimentalno raziskavo in zaključili, da so teze ProPublice napačne. Ta kritika izgleda upravičena. Bila pa bi bolj prepričljiva, če bi v [8] jasno pokazali, kje naj bi nastala odločilna napaka v ProPublici. Namesto tega so navedli svoj lastni eksperimentalni rezultat, ki naj bi dokazal, da so obsojenci obravnavani pravično, ne glede na raso. Ta rezultat so dobili tako, da so upoštevali ocene tveganja obsojencev na lestvici od 1 do 10, kot jih oceni COMPAS. Iz teh ocen so izračunali mero AUC (Area under ROC curve, to je površina pod ROC krivuljo, ki se standardno uporablja v strojnem učenju kot indikator uspešnosti učenja). Mera AUC je zanimiva zato, ker je enaka verjetnosti, da napovedni sistem pravilno razlikuje med pozitivnimi in negativnimi primeri. To pomeni, da če vzamemo dva naključna primera (dva obtoženca), od katerih je eden pozitiven (je ponovil prestopek) in drugi negativen, potem bo sistem z verjetnostjo AUC pravilno odločil, kateri je pozitiven in kateri negativen. Flores idr. [12] navajajo, da je za belce dobljena vrednost AUC enaka 0,69, za temnopolte pa 0,70. Pri tem razlika ni statistično signifikantna. Iz tega zaključijo, da COMPAS ni diskriminatoren in da rezultati ProPublice, ki kažejo na diskriminacijo, ne morejo biti pravilni. Vendar ta posredni argument dopušča dvom, saj mere AUC ter FPR in FNR med seboj niso enoznačno povezane.

Dressel in Farid [7] poročata o relevantnih poskusih, kjer ju je zanimala točnost napovedi o tveganju povratništva, ki jo dosežejo naključno izbrani ljudje brez znanja o navadah obsojencev. Poskus s človeškim napovedovanjem (izvedenim s "crowd sourcingom") sta naredila na podmnožici podatkov o 1000 od skupaj okrog 7000 obsojencev iz poskusov v [2] in [8]. Ker bi bila uporaba vseh 137 atributov za poskus z napovedmi ljudi nepraktična, sta izmed originalne množice atributov izbrala le 7 atributov. Napovedna točnost ne-ekspertov je v teh poskusih presenetljivo praktično enaka kot tista s sistemom COMPAS. Zanimivo je, da so tudi človeške napovedi v tem poskusu podobno pristranske kot COMPAS, merjeno s FPR in FNR za bele in temnopolte, ter da se ti rezultati skoraj ne spremenijo, če človeškemu ocenjevalcu kot dodatno informacijo podamo tudi podatek o rasi.

Neodvisno od teh rezultatov je Cynthia Rudin [16] s strojnim učenjem pravil iz omenjenih podatkov s Floride sintetizirala zelo enostaven in povsem razumljiv napovedni model, ki vsebuje le tri enostavna if-then pravila in uporabi tri atribute. Za razliko od modela COMPAS, so ta pravila trivialno razumljiva. Tudi ta prediktor ima zelo podobno točnost kot COMPAS, pa tudi podoben FPR in FNR.

Iz vseh opisanih rezultatov zaključujemo, da je ta napovedni problem kljub obsežnim razpoložljivim informacijam o obtožencu tako težak, da boljše točnosti ni mogoče doseči. Hkrati skoraj vse dosežemo z dvema ali tremi najbolj koristnimi atributi in preostalih 130+ atributov dodatno ne doseže ničesar. V skladu s tem se v [2, 7] izoblikuje teza, da uporaba strojnega učenja v pravosodju nima dobre perspektive. To je seveda prenagel in preveč enostaven zaključek, na kar opozarja [17]. V mnogih drugih aplikacijah je strojno učenje preseglo napovedno točnost ekspertov, kar so med drugim potrdili mnogi poskusi s strojnim učenjem v medicinski diagnostiki.

Vsa ta različna mnenja o (ne)pristranskosti in (ne)uporabnosti sistema COMPAS kažejo na pomanjkanje splošno sprejetih operativnih definicij pristranskosti in pravičnosti v strojnem učenju. Situacijo lepo ilustrira močno citirani članek [14], ki razglablja o več deset relevantnih definicijah, pri tem pa ne ponudi sinteze, ki bi to idejno kompleksnost omejila in dala praktično uporaben pristop. Dodatno nelagodje ta članek povzroči s tem, da na hitro opravi s sistemom COMPAS in ga uvrsti med očitno pristranske ter neposrečene in nekoristne. Pri tem raziskave [8] ne omeni. V [7] pa spregleda dejstvo, da tudi alternativni rezultati s strojnim učenjem in s človeškim napovedovanjem povratništva na isti podatkovni množici kažejo zelo podobno pristranskost do temnopoltih, kot COMPAS.

3 DEFINICIJE PRISTRANSKOSTI IN PRAVIČNOSTI TER NJIHOVI PROBLEMI

V splošnih medijih se strojnemu učenju pogosto enostavno očita pristranskost bolj po občutku, ne da bi natančno definirali, po kakšnem matematično preverljivem kriteriju se pristranskost kaže. Izjave, kot so: "sistem se je v sodstvu pokazal kot pristranski do temnopoltnih obtožencev" [2], ali "sistem je pri ocenjevanju kandidatov za zaposlitev pristranski do žensk" [5], uporabljajo splošne fraze, kot so "pristranskost algoritmov", "pristranskost strojnega učenja", "pristranskost umetne inteligence". Včasih so te ugotovitve opremljene z enostavno razlago, kot je: "sisteme strojnega učenja razvijajo skoraj izključno beli moški, torej …".

Danes je jasno, da stvar ni tako trivialna. Pretirano enostavne razlage se zdaj pojavljajo redko. Postaja tudi bolj jasno, da fraza "pristranskost algoritmov" ni primerna in daje napačen občutek, da so algoritmi sposobni imeti zle namene in da ne delujejo po matematičnih in statističnih principih [16]. Cilj teh metod je vedno, da iz podatkov o realnem svetu odkrijemo zakonitosti, ki v tem svetu veljajo. Seveda se takoj pojavi problem, če so v realnem svetu že prisotne pristranske prakse. Podatki, zajeti v takem svetu, odražajo to pristranskost in algoritem za učenje to pristranskost detektira in reproducira. Če rezultate, dobljene iz pristranskih podatkov v realnem svetu, spet uporabimo v realnem svetu, bomo s tem reproducirali že obstoječo pristranskost [10]. Vseeno še vedno ni dovolj natančno definirano, kaj pristranskost sploh je. Pogosto gre za vtis pristranskosti, kjer se kažejo predsodki za ali proti posamezniku ali skupini na način, ki se razume kot nepravičen [15].

Poglejmo, v čem so težave z definicijo pristranskosti. Že na področju strojnega učenja najdemo različne razlage, ki so vse po svoje smiselne. Izraz pristranskost se v strojnem učenju uporablja v več pomenih [11]:

1. T.i. induktivna pristranskost: to je princip, po katerem se algoritem odloči za eno izmed tipično velikega števila možnih hipotez, ki so glede na učne podatke vse na nek način utemeljene. Ta vrsta pristranskosti je neizogiben mehanizem in je zato v principu pozitivna komponenta strojnega učenja, brez katerega strojno učenje sploh ni možno. Primer take pristranskosti je Occamova britev (Occam's razor), ki pravi: Če imamo na voljo dve razlagi zbranih podatkov, ki sicer obe enako dobro razložita te podatke, potem raje izberemo enostavnejšo razlago [9, 11, 15]. To pristranskost uporabljamo pogosto ne le v strojnem učenju, temveč v znanosti nasploh. Čeprav ima izraz pristranskost negativen prizvok, je induktivna pristranskost pozitivna in celo neizogibna komponenta strojnega učenja, kot razlagajo avtorji v [11], in osnovni učbeniki umetne inteligence.
2. Pristranskost v učnih podatkih, ki odražajo dejanske pristranskosti v ustaljenem odločanju na danem področju uporabe (npr. pristranskost ekspertov v dejanski sodni praksi v okolju, iz katerega so zajeti učni podatki) [4, 14].
3. Pristranskost, ki izhaja iz neprimernega postopka zbiranja podatkov oz. vzorčenja [11], npr. da je za določeno skupino ljudi na voljo bistveno manj primerov kot za druge skupine. Potem v skladu z matematično utemeljenimi statističnimi in verjetnostnimi principi nekatere skupine, tipično manjšinske, izpadejo kot diskriminirane (lahko celo v pozitivnem smislu!) zgolj zato, ker metode ocenjevanja verjetnosti upravičeno ocenijo verjetnosti drugače, če je na voljo malo podatkov.

Gornji viri pristranskosti so razmeroma splošno sprejeti. Ostaja pa problem, kako natančno definirati merila, ki objektivno povedo, ali je sistem pristranski oz. ki to pristranskost kvantitativno vrednotijo. Obstajajo številne mere, ki so videti relevantne, vendar se izkaže, da si med seboj nasprotujejo in zato za zdaj enostavne, splošno sprejete mere ni. Situacijo zelo dobro ilustrira izčrpni pregled v [14].

Bolj fokusirano raziščejo ta problem Kleinberg idr. [12]. Definirajo tri naravne, same po sebi takorekoč očitne pogoje, ki jim mora zadostiti sistem, če naj bo nepristranski (pravičen). Toda presenetljivo se izkaže, da ti trije pogoji ne morejo biti izpolnjeni hkrati, razen v posebnih primerih, ki pa so za prakso nezanimivi. Torej so že te tri osnovne zahteve skupaj neuresničljive. Te tri zahteve so:

(1) Kalibracija ocen verjetnosti: če algoritem identificira množico oseb, ki naj bi z dano verjetnostjo pripadale pozitivnemu razredu, potem mora približno tak delež te množice dejansko pripadati pozitivnemu razredu. Enak pogoj mora veljati za vse skupine oseb, ki se razlikujejo v "zaščitenem atributu", npr. rasi ali spolu.

(2) Ravnotežje pozitivnega razreda: povprečje verjetnostnih ocen oseb pozitivnega razreda mora biti enako za vse skupine.

(3) Ravnotežje negativnega razreda: analogno kot povprečje pozitivnega razreda.

Avtorji dokažejo izrek, da so te tri zahteve, čeprav si v bistvu prizadevajo za isti cilj zmanjševanja pristranskosti, med seboj nekompatibilne, razen v posebnih primerih.

Kadar se pojavi pristranskost, je vprašanje, kako jo odpraviti. Za to obstaja vrsta idej, od katerih sta najbolj očitni (a) "zaščiteni atributi" in (b) obratna diskriminacija. Tipična zaščitena atributa sta rasa in spol.

Princip zaščitenih atributov je, da algoritmu učenja prepovemo uporabo teh atributov pri odločanju o klasifikaciji primera. Ta ideja navadno ne deluje dobro, saj algoritem učenja efektivno rekonstruira njihove vrednosti iz drugih, nezaščitenih atributov, ki korelirajo z zaščitenimi. Na primer iz podatkov o šolanju ali lokaciji prebivališča algoritem sklepa na raso osebe.

Princip obratne diskriminacije je, da depriviligiranim skupinam pri obravnavi namenoma damo določeno prednost, s čimer naj bi izničili učinek diskriminacije. Ta ukrep je očitno dobronameren, vendar s tem dejansko uvedemo dodatno nepravičnost, ki je za nekatere vprašljiva (npr. Alelyani [1]). Taka nepravičnost (obratna diskriminacija) je upravičena, vendar ne z vidika pravičnosti, temveč z vidika "višjih" vrednot, npr. da za v bodoče popravimo zgodovinske krivice in z začasno nepravičnostjo dosežemo dolgoročno pravičnost. Torej gre za strateško uresničevanje družbeno sprejetih vrednot, ki v praksi sicer zaradi zgodovinskih razlogov in vztrajnosti niso hitro uresničljive. Ostaja težavno vprašanje, do kakšne mere je obratna diskriminacija smiselna. To bi moralo biti določeno z demokratično sprejetim družbenim konsenzom, formaliziranim z ustreznimi zakoni za vsak primer posebej.

V praksi se reševanja pristranskosti lotimo znotraj treh faz strojnega učenja: 1) pred-procesna faza, kjer povečamo vzorec manjšine, 2) med-procesna faza, kjer dodajamo omejitve, s katerimi kompenziramo za neenakomeren vzorec in 3) post-procesna faza, kjer spreminjamo mejne vrednosti za manjšine [4, 14, 15].

Ko razvijamo metode in orodja se moramo zavedati potencialnih pasti. V [1, 11] avtorji izpostavljajo, da lahko določene rešitve pripeljejo do novih nepravičnosti, pogosti stranski učinek mutiranja učnih podatkov pa je izguba pomembnih povezav med spremenljivkami ali slabše delovanje celotnega algoritma [5].

4 ZAKLJUČKI

Pristranskost je v nekaterih pomembnih aplikacijah strojnega postala popularna in kontroverzna tema. V diskusiji prevladuje nejasnost, ki izvira iz tega, da večina razume pojem pravičnosti in pristranskosti intuitivno. Pri tem pravičnost doživljamo na razne načine in v podrobnostih ni popolnega soglasja. Tako tudi ni soglasja o tem, kakšen naj bi bil jasen, matematično formuliran kriterij, s katerim bi brez dvomov kvantificirali pristranskost konkretnega sistema. Veliko več je nasprotovanj, kontroverznih in odprtih tem, kjer ni strinjanja. Ni konsenza o izvoru pristranskosti, niti o tem, katero orodje oz. metoda je za soočanje s pristranskostjo najbolj primerna.

Spielkamp [17] na primer komentira pomanjkanje enotne definicije in kriterijev takole: "Jasno je, da naj bi pravičnost strojnega učenja pomenila produciranje odločitev, s katerimi bi bili kot družba zadovoljni. Vendar glede tega ljudje nismo enotni." Na primeru COMPAS se pokaže, kako ključna je ta enotnost. COMPAS je testiralo več strokovnjakov in njihova mnenja so si povsem nasprotna. Nekateri trdijo, da je COMPAS pristranski, drugi pravijo, da ni. Spielkamp meni, da imajo prav vsi, saj pravičnost razumejo na razne načine.

Tudi nekateri drugi avtorji ugotavljajo podobno. Poudarjajo, da je nujno razviti široko družbeno sprejeto definicijo pravičnosti, ki bo rezultirala v sistemih, ki bodo delovali v skladu z ustaljenimi družbenimi vrednotami in s tem povezanimi pričakovanji.

V literaturi kljub temu ni videti, da bi kdo predvidel, kako velik izziv bo to. Pričakovanja glede vrednot bo treba namreč natančno formulirati z ustreznimi zakoni. Na primer, ali naj bo zaradi zgodovinskih krivic v konkretni aplikaciji realizirana obratna pristranskost in do kakšne mere? Ta formulacija bo morala biti bolj tehnična kot običajno v predpisih in zakonih, saj bo to osnova za konkretno implementacijo v algoritmih umetne inteligence.

Za ustrezno splošno razumevanje in ukrepanje na tem področju se kaže potreba po kvalitetni splošni izobrazbi ljudi. Pomanjkanje le-te se kaže v načinu poročanja, odzivanju ljudi in tudi zmedenosti strokovanjakov. Različni algoritmi postajajo neizogiben del naših življenj. Nesprejemljivo je, da o njih ne samo da vemo premalo, ampak imamo celo napačne predstave. Splošno znanje o delovanju algoritmov (in širše o umetni inteligenci) temelji največ na poročanju medijev s pogosto pomanjkljivimi informacijami, napačnimimi poudarki ter pretiravanjem. Nujno je, da se ljudje o tem dovolj izobrazijo in lahko tako primerno ocenijo situacije, kjer nek algoritem proizvaja nezaželene rezultate.

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.

REFERENCE

[1] Alelyani, S. (2021). Detection and Evaluation of Machine Learning Bias. *Applied Sciences, 11*(14). https://doi.org/10.3390/app11146271

[2] Angwin, A., Larson, J., Mattu, J. in Kirchner, L. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. *ProPublica*.

[3] *Artificial Intelligence Act*, European Parlament, 14 June 2023.

[4] Blanzeisky, W. in Cunningham, P. (2021). Algorithmic Factors Influencing Bias in Machine Learning. *arXiv preprint.* https://doi.org/10.48550/arXiv.2104.14014

[5] Chakraborty, J., Majumder, J. in Menzies, T. (2021). Bias in Machine Learning Software: Why? How? What to do? *arXiv preprint.* https://doi.org/10.48550/arXiv.2105.12195

[6] Dastin, J. (11.10.2018). *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[7] Dressel, J., Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1).

[8] Flores., A. W., Bechtel, K. in Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks." *Federal Probation Journal.*

[9] Gordom, D. F. in Desjardins, M. (1995). Evaluation and Selection of Biases in Machine Learning. *Machine Learning, 20*, 5-22. https://doi.org/10.1023/A:1022630017346

[10] Hardt, M., Price, E. in Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *arXiv preprint.* https://doi.org/10.48550/arXiv.1610.02413

[11] Hellström, T., Dignum, V. in Bensch, S. (2020). Bias in Machine Learning – What is it Good for? *arXiv preprint.* https://doi.org/10.48550/arXiv.2004.00686

[12] Kleinberg, J, Mullainathan, S. in Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint.* https://doi.org/10.48550/arXiv.1609.05807

[13] Levin, S. (8.9.2016). *A beauty contest was judged by AI and the robots didn't like dark skin*. The Guardian. https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

[14] Mehrabi, A., Morstatter, F., Saxena, N., Lerman, K. in Galstyan, A. (2021). *ACM computing surveys* (CSUR) 54 (6), 1-35. https://doi.org/10.48550/arXiv.1908.09635

[15] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E. … Staab, S. (2020). Bias in data-driven artificial intelligence systems – An introductory survey. *WIREs Data Mining and Knowledge Discovery, 10*(3). https://doi.org/10.1002/widm.1356

[16] 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. https://doi.org/10.1038/s42256-019-0048-x

[17] Spielkamp, M.. (2017). Inspecting Algorithms for Bias. *Technology Review.*

[18] Staff, R. (7.12.2016). *New Zealand passport robot tells applicant of Asian descent to open eyes*. Reuters. https://www.reuters.com/article/us-newzealand-passport-error-idUSKBN13W0RL

[19] UNESCO Recommendation on the Ethics of Artificial Intelligence, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137

ChatGPT through Tononi's Definition of Consciousness

Analiza ChatGPT skozi Tononijevo definicijo zavesti

Matjaž Gams
Odsek za inteligentne sisteme
Institut “Jožef Stefan”
Jamova cesta 39, 1000 Ljubljana
Slovenija

ABSTRACT

The intricacies of consciousness and its existence have long been the subjects of both philosophical and scientific investigations. With advancements in artificial intelligence, discerning the line between algorithmic processing and consciousness becomes increasingly vital. This paper uses Tononi's Integrated Information Theory (IIT) to examine ChatGPT, assessing its alignment with human notions of consciousness. Analysis shows that while ChatGPT exhibits attributes superficially resonating with Tononi's axioms, it lacks the foundational conscious experience described by the theory. Then, passing the Turing test as a way of demonstrating consciousness is debated with similar conclusions.

POVZETEK

Zapletenost zavesti in njen obstoja ter namen so predmet tako filozofskih kot znanstvenih naravoslovnih raziskav. Z napredkom v umetni inteligenci postaja vedno bolj ključno razlikovanje med algoritmično obdelavo in zavestjo. Ta članek uporablja Tononijevo teorijo integrirane informacije (IIT) za analizo potencialne zavesti ChatGPT, pri čemer je ključno vprašanje usklajenost z človeškimi predstavami o zavesti. Analize kažejo, da ChatGPT sicer površinsko zelo dobro resonira s Tononijevimi aksiomi, vendar mu manjka temeljna zavestna izkušnja, ki jo teorija opisuje. V nadaljevanju prispevka je opisovanje reševanja Turingovega testa s strani ChatGPT kot preverjanja zavestnosti, ki se izkaže pri tem testu, ter diskusija.

KLJUČNE BESEDE
Zavest, ChatGPT, Turingov test

KEYWORDS
Consciousness, ChatGPT, Turing test

1 INTRODUCTION

As AI continues to progress, the discussion surrounding machine consciousness intensifies. The question often arises: can machines ever possess genuine consciousness? Has that already been achieved by ChatGPT? To delve into this, we leverage the IIT—a framework by Tononi proposing five axioms believed to underlie human consciousness—and apply these criteria to ChatGPT. Then, passing the Turing test by ChatGPT is analysed as another test whether consciousness is already achieved.

There is a rich tapestry of literature exploring consciousness, AI, and where the two intersect. Koch et al. have explored how IIT offers a quantitative measure of consciousness in diverse systems [1]. Dehaene and others have looked into consciousness in both biological and artificial systems, arguing for unique neural markers that underpin conscious states [2]. Recent advancements in AI, especially deep learning models like ChatGPT, have triggered renewed debates, with researchers like Hassabis et al. and Bengio postulating how AI might approach or simulate human-like consciousness [3,4].

2 ANALYSIS OF CHATGPT THROUGH TONONI'S AXIOMS FOR CONSCIOUSNESS

2.1 Background: Tononi's Axioms for Consciousness

Tononi's Integrated Information Theory (IIT) proposes five fundamental axioms aimed at capturing the core of consciousness:

Intrinsic Existence: Consciousness inherently exists for the conscious entity. It's a subjective phenomenon, deeply personal and unique to each entity [6].

Composition: Consciousness is not monolithic. It possesses structure, and within it, diverse experiences can be differentiated. This diversity isn't merely quantitative but also qualitative, making each conscious experience rich and multidimensional [7].

Information: Consciousness is informative. Every conscious experience stands out against other potential experiences, indicating a specific state of affairs over countless others [8].

Integration: Despite its diverse composition, consciousness is unified. Experiences are intertwined, and it's impossible to completely isolate any subset of phenomena within a single conscious moment [9].

Exclusion: Consciousness is definite, both in content and in space. At any given moment, an entity is conscious of certain things and not others, thus creating clear boundaries of experience [10].

2.2 ChatGPT under the lens of Tononi's axioms

To evaluate the degree of ChatGPT achieveing each axion of the Tononi's theory, the existing literature and the opinion in the AI community is of no great help, since there are mixed opinions and no generally accepted viewpoint. However, there are two bases that this paper evaluates consciousness of ChatGPT through various analysis: First, there is 50 years of experience of the author of the paper in the AI and cognitive field and 20 years of superintelligence studies. Second, there is an opinion of ChatGPT when asked about each particular issue. Interestingly enough, even though there were some differences, and the author chose the merit of the expertize, there was quite strong agreement in general. The GPT opinion, seemingly, was to a large extend hand-crafted by humans, dealing with this issue, and partially through the GT or LLT approach, as demonstrated by the replies.

Intrinsic Existence: At its core, ChatGPT is a product of algorithms and vast data. It operates in response to inputs, without possessing feelings, beliefs, or desires. It lacks any semblance of self-awareness, and as such, it probably does not meet the axiom of intrinsic existence.

Composition: ChatGPT, architecturally, boasts a vast neural network configuration. This allows it to generate a diverse array of outputs based on different inputs. However, this structural variety isn't birthed from conscious deliberation but from learned patterns. While it exhibits structural diversity analogous to the composition axiom, it seemingly lacks the qualitative conscious nuance integral to Tononi's definition.

Information: The model processes and produces specific responses based on its training. Each response is a selective piece of information shaped by its training data and the query. Although this aligns with the informational aspect of the axiom, the absence of conscious deliberation and choice makes its alignment potentially superficial.

Integration: ChatGPT's processes are integrated. Each input is processed through multiple layers, intertwining different learned patterns to produce a coherent output. This mirrors the operational facet of the integration axiom. However, the unity described by Tononi implies a cohesive conscious experience, which ChatGPT probably does not possess.

Exclusion: With its design parameters and training, ChatGPT operates within set boundaries. It produces specific responses and not others. While this resonates with the operational side of the exclusion axiom, the model's responses might not be the result of conscious choices or experiences.

It's essential to highlight the difference between operational alignment and conscious alignment. While ChatGPT showcases attributes that operationally resonate with some of Tononi's axioms, it does so without the underlying conscious experience these axioms were designed to describe. The axioms, rooted in human phenomenology, emphasize subjective experience, something inherently absent in ChatGPT.

ChatGPT, in its design and operation, exhibits attributes that superficially align with Tononi's axioms to a certain degree. However, when delving into the crux of these principles—conscious experience—ChatGPT falls a bit short. While it stands as a testament to advancements in information processing and AI, ChatGPT does not qualify as a conscious entity within the framework of Integrated Information Theory.

3 HAS CHATGPT PASSED THE TURING TEST I.E. REACHING THE COUNSCIOUSNESS?

Another way to test the level of AI systems at achieving consciousness, can be performed by the Turing Test (TT). The Turing Test, proposed by the eminent computer scientist Alan Turing in 1950, is a measure of a machine's ability to exhibit intelligent behaviour indistinguishable from that of a human [11]. Turing postulated that if an evaluator, after interacting with an unseen interlocutor for five minutes, could not reliably tell whether dealing with a machine or a human, then the machine could be said to have passed the test. As ChatGPT emerges as one of the most sophisticated AI language models, there's debate about its positioning relative to the Turing Test and consequently consciousness.

3.1 Arguments in Favor:

Advanced communication:

ChatGPT is engineered to provide detailed replies that span a wide range of subjects, from science and technology to philosophy and art. The quality of its responses often mirrors human-like expertise and reasoning capabilities, making it a versatile conversational agent. Its ability to provide contextually relevant and accurate information resembles the intellectual breadth and depth one would expect from a knowledgeable human, being familiar with the Web. Consequently, it is increasingly difficult to immediately distinguish some of its responses from those generated by a human, especially in text-based interactions.

Adaptive Interaction:
The model is designed to be sensitive to the conversational context, allowing it to adjust its responses based on previous dialog turns. This adaptability manifests in the way it can switch topics smoothly, clarify ambiguities, or even attempt humour, much like a human would in a fluid conversation. Its capability to modify its tone and content in real-time according to conversational cues shows an advanced level of adaptability, often comparable to human dialog dynamics. This feature enhances its suitability for diverse interactions, making it a compelling interface for numerous applications.

3.2 Arguments Against:

Lack of Understanding:
While ChatGPT's responses can be elaborate, it's essential to remember that the model doesn't possess genuine understanding or consciousness. It generates text based on statistical patterns it has learned from its training data, and skilfully intercombines various potential most relevant texts patterns into a word by word continuation, without the ability to comprehend the meaning or significance of the conversation [15]. Its sophisticated language capabilities may give the illusion of understanding, but this surface-level competence should not be mistaken for genuine comprehension or awareness.

Inconsistencies:
The model is not reliable in providing consistent answers over different conversational sessions or even within the same interaction, the effect called hallucinating. These inconsistencies are a testament to its underlying algorithmic nature, which does not have the capacity for ongoing memory or the ability to learn from past interactions [16]. Such discrepancies in its replies can sometimes make it evident that one is conversing with a machine, not a human, thereby undermining its credibility and reliability in more complex or sensitive dialog scenarios.

Absence of Emotions:
Human dialog is often rich in emotional nuance and subtext, an aspect that is conspicuously lacking in ChatGPT. Despite its linguistic capabilities, the model cannot feel emotions or understand the emotional weight of certain words or situations. Its interactions are devoid of emotional depth, empathy, or any other kind of emotional intelligence that is often central to human communication. This absence not only differentiates it from human-like conversation but also limits its applicability in scenarios where emotional engagement is crucial.

Literary Standpoints:
Hernandez-Orallo discussed the limitations of the Turing Test, emphasizing that mere linguistic capability may not be a sufficient measure of machine intelligence [13]. Russell and Norvig, in their comprehensive AI textbook, consider the Turing Test as a valid, though not definitive, measure of machine intelligence, suggesting that while passing the Turing Test is significant, it does not necessarily equate to full human-like intelligence [14]. A recent article in Nature delineates the increasing complexity and performance of large-scale language models, weighing them against the Turing Test's standards [15]. Another piece in the Boston Review contemplates the question of consciousness in such models, considering the philosophical implications of designating them as 'conscious' [16].
One salient point from the Boston Review article is the distinction between machine operation and human consciousness [17]. While large language models like ChatGPT can generate intricate and seemingly aware responses, there remains a significant philosophical and cognitive gap. These models operate by recognizing and generating patterns based on massive datasets without the subjective experience that characterizes human consciousness. The Boston Review delves into the implications of mistaking this high-level processing capability for genuine consciousness, emphasizing the risks of anthropomorphizing machine behaviors.
In light of this, it becomes evident that while ChatGPT can generate responses that mimic human-like thinking, it does so devoid of genuine understanding, emotions, or the conscious experiences that humans possess.

This perspective aligns with the core debate surrounding the Turing Test. ChatGPT's capability to produce responses that may seem indistinguishable from those of a human in specific contexts does not necessarily imply that the model genuinely "thinks" or possesses consciousness. Instead, it underlines the model's adeptness at pattern recognition and response generation, which, although impressive, is fundamentally distinct from human cognition.

4 DISCUSSION

ChatGPT demonstrates the remarkable advancements in artificial intelligence, surpassing previous efforts by a substantial margin. However, it's essential to distinguish between its algorithmic complexity and genuine consciousness. The difference is described in this paper through two major parts: first though the IIT theory and second through the Turing Test.

While the model excels in simulating conscious traits, within the framework of Integrated Information Theory (IIT) it remains devoid of authentic conscious experience—a distinction that should be apparent to cognitive scientists, even in the presented discussions in this paper.

Second, regarding the Turing Test, ChatGPT displays impressive linguistic capabilities that bring it tantalizingly close to fulfilling the test's criteria. Yet, it falls short in key areas, including genuine understanding and emotional intelligence. For specialists familiar with the field, this limitation is easily identifiable, especially when the conversation veers into complex or emotionally charged topics. In these situations, the model's limitations become obvious, as it fails to respond appropriately to intricate, tangible or conflicting dialogues.

While some literature may argue that ChatGPT could pass the Turing Test in brief interactions, especially with laypeople, the articles from Boston Review and Nature emphasize the crucial distinction between mere simulation and actual consciousness. Indeed, those well-versed in the subject, including cognitive scientists, should readily discern ChatGPT's performance from that of a human. This serves as a timely reminder of the need to continually update our evaluation metrics, not just technically but also philosophically, as AI continues to challenge our conceptual boundaries of intelligence and consciousness.

In conclusion, although ChatGPT and similar generative models signify a quantum leap toward general AI and possibly superintelligence, they should not be conflated with achieving consciousness or passing the Turing Test. Despite possessing certain superhuman attributes, such as speed, and achieving approximate human-level performance in specific tasks, ChatGPT does not meet the criteria for consciousness. That might not hold for an extended period of time. According to Chalmers [15] »Within the next decade, we may well have systems that are serious candidates for consciousness«; however, current models like ChatGPT should not be mistaken as such, especially by those familiar with the cognitive sciences. Therefore, while the progress is significant, the journey toward creating truly conscious machines is far from over.

References:

1. Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, 17(5), 307-320.
2. Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486-492.
3. Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2), 245-258.
4. Bengio, Y. (2017). Towards biologically plausible deep learning. *arXiv preprint* arXiv:1702.08835.
5. Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 42.
6. 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* (pp. 5998-6008).
7. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
8. Mitchell, T. M. (1997). Machine Learning. *McGraw Hill*.
9. 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.
10. Bostrom, N. (2014). Superintelligence. *Oxford University Press*.
11. Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433-460.
12. Hernandez-Orallo, J. (2000). Beyond the Turing Test. *Journal of Logic, Language and Information*, 9(4), 447-466.
13. Russell, S. J., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). *Prentice Hall*.
14. Biever, C. (2023). ChatGPT broke the Turing test — the race is on for new ways to assess AI, Large language models mimic human chatter, but scientists disagree on their ability to reason. *Nature*. Retrieved from https://www.nature.com/articles/d41586-023-02361-7
15. Chalmers, J.D. (2023). Could a Large Language Model Be Conscious? *Boston Review*. Retrieved from Boston Review URL

Manipulacija v umetni inteligenci
Manipulation and artificial intelligence

Gaja Gril†
Pedagoška fakulteta
Univerza v Ljubljani
Slovenija
gg99039@student.uni-lj.si

POVZETEK

V svetu, ki ga vse bolj oblikuje umetna inteligenca, se je začelo pojavljati vprašanje manipulacije. Ko sistemom umetne inteligence zaupamo vse večje vloge v naših življenjih, od personaliziranih priporočil do kritičnega odločanja, se lahko vprašamo, do katere mere zaupati strojem, ki se zdijo tako inteligentni in nepristranski? Umetna inteligenca postaja ne le vir pomoči, temveč tudi manipulira, dezinformira in vpliva. V tem raziskovanju zapletene mreže manipulacij znotraj umetne inteligence se srečamo s tehnološko uganko, ki postavlja pod vprašaj naše razumevanje avtonomije in etike.

KLJUČNE BESEDE

Manipulacija, umetna inteligenca, demokracija, marketing, etika

ABSTRACT

In a world increasingly shaped by artificial intelligence, the issue of manipulation has begun to emerge. As we entrust AI systems with ever greater roles in our lives, from personalised recommendations to critical decision-making, we may ask to what extent should we trust machines that seem so intelligent and impartial? Artificial intelligence is becoming not only a source of help, but also a source of manipulation, misinformation and influence. In this exploration of the complex web of manipulation within AI, we are confronted with a technological conundrum that calls into question our understanding of autonomy and ethics.

KEYWORDS

Manipulation, artificial intelligence, democracy, marketing, ethics

1 UVOD

Carroll in sodelavci (2023) definirajo sistem umetne inteligence kot manipulativen, če deluje, kot da bi si prizadeval za namerno in prikrito spremembo človeka (ali drugega agenta) [1]. Primer manipulativne umetne inteligence je priporočilni sistem, ki optimizira dolgoročno sodelovanje, tako da prikrito spodbuja uporabnika, da si ogleda več videoposnetkov. Manipulacija je težko precizno opredeljiva, lahko pa se naslonimo na štiri glavne značilnosti. Prva je, da gre za nerazumski vpliv, pri katerem skuša manipulator zaobiti ali oslabiti človekovo sposobnost premišljenega odločanja. Druga je, da manipulacija zahteva uporabo zvijače in prevare, pogosto s skritimi sredstvi, da bi nekoga prisilili k določenemu ravnanju. Tretji razlog je, da manipulacija vključuje uporabo določene stopnje pritiska, da bi storili, kar želi manipulator, na primer s čustvenim izsiljevanjem. In nazadnje, običajno se ne ravna po interesih, ciljih in željah ciljne osebe, temveč le po interesih manipulatorja [2].

Razlogi za povečano tveganje za manipulacijo v dobi umetne inteligence so naslednji; Prvič, umetna inteligenca je precej netransparentna, saj velika večina algoritmov ni pregledna in razložljiva, uporabniki pa ne posedujejo zadosti tehničnega znanja [3]. Drugič, sistemi umetne inteligence lahko odkrijejo preference, interese, navade in ostale značilnosti posameznika ter tako natančno prilagodijo vsebino [4]. Poleg tega se lahko sistemi umetne inteligence uporabljajo za ocenjevanje psiholoških stanj ljudi, kot so na primer čustvena stanja [5]. Tretjič, umetna inteligenca omogoča oslabitev avtonomije odločanja potrošnikov z izkoriščanjem njihovih ranljivosti pri odločanju [3].

2 MANIPULACIJA Z INFORMACIJAMI

Na dinamičnem področju demokratičnih sistemov je umetna inteligenca postala močan dejavnik, ki spreminja taktiko političnih kampanj in nagovarjanja volivcev. Tradicionalne metode sodelovanja z javnostjo so se umaknile naprednim pristopom, ki jih poganja umetna inteligenca in ki izkoriščajo obsežne podatke, da bi vplivali na razpoloženje javnosti in zagotovili volilno zmago [6]. Sposobnost umetne inteligence, da obdela in oceni obsežne zbirke podatkov političnim kampanjam omogoča vpogled v nagnjenja, vedenje in čustva volivcev. Algoritmi umetne inteligence z raziskovanjem spletnih interakcij, zgodovine brskanja in drugih digitalnih sledov oblikujejo podrobne profile volivcev in z izjemno natančnostjo predvidevajo posameznikove težnje. S tem znanjem lahko politični akterji prilagodijo svoja sporočila tako, da odmevajo v različnih demografskih skupinah, in oblikujejo personalizirane pozive, ki so povezani s čustvi in prepričanji volivcev. Ciljno usmerjanje volivcev, ki ga poganja umetna inteligenca, presega zgolj oblikovanje profilov in uporablja tehnike mikrousmerjanja,

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

da bi volivce doseglo na osebni ravni. Sporočila kampanje so lahko natančno prilagojena, da se ujemajo s posameznikovimi načeli in interesi, kar poveča njihov prepričljiv učinek in verjetnost pridobitve podpore. Ta natančnost kampanjam omogoča, da pametno razporedijo sredstva in se osredotočijo na ključne volivce, ki so bistveni za njihovo zmago [6].

Vendar pa se na drugi strani tega tehnološkega napredka pojavljajo etični pomisleki. Kritiki trdijo, da lahko takšno mikrousmerjanje še poveča polarizacijo in ustvari odmevne komore, v katerih se posamezniki utrdijo v svojih obstoječih prepričanjih in okrepijo pristranskost pri potrjevanju. Poleg tega pomanjkanje preglednosti algoritmov umetne inteligence, ki se uporabljajo za politično ciljanje, povzroča pomisleke glede odprtosti in odgovornosti, kar lahko spodkoplje demokratično načelo informiranega soglasja.V digitalni dobi se je ranljivost razširjanja informacij povečala zaradi manipulacij, ki jih omogočajo generirane vsebine in globoki ponaredki. Te napredne tehnologije lahko z razširjanjem lažnih informacij in spodkopavanjem zaupanja v politične postopke motijo demokratične sisteme [6].

Ker se demokratične družbe spopadajo s temi dilemami, so zahteve po večji preglednosti in algoritemski odgovornosti vse glasnejše. Iskanje ravnovesja med ohranjanjem vključenosti uporabnikov in varovanjem demokratične integritete od podjetij družbenih medijev zahteva vesten nadzor umetne inteligence in etični premislek. C. Serbanescu (2021) vidi umetno inteligenco kot grožnjo demokraciji, saj ta omogoča manipulacijo procesov odločanja državljanov in s tem ogroža njihovo avtonomijo pri sodelovanju v demokratičnih procesih. Aplikacije umetne inteligence imajo potencial, da oblikujejo "arhitekturo izbire" državljanov, vključno z razpoložljivimi izbirami in načinom njihove predstavitve, na personaliziran, dinamičen in prikrit način. Vseprisotnost tehnologije in zbiranje velikih količin podatkov sta olajšala manipulacijo, ki jo omogoča umetna inteligenca, kar omogoča prilagojene vplive in dinamično prilagodljivost za izkoriščanje posameznikovih slabosti. Ta oblika manipulacije predstavlja kvalitativno in kvantitativno spremembo v primerjavi s tradicionalno manipulacijo, saj lahko učinkoviteje doseže veliko število volivcev in vpliva na rezultate, ki bi lahko predstavljali "voljo ljudstva". Na splošno je vzpon aplikacij umetne inteligence privedel do učinkovitejših oblik manipulacije, kar predstavlja velik izziv za demokracijo [7].

3 VEDENJSKO OGLAŠEVANJE IN MANIPULACIJA POTROŠNIKA

Umetna inteligenca spreminja dinamiko trženja in oglaševanja, saj prehaja od široke promocije k personaliziranim izkušnjam. Z njo se briše meja med manipulacijo in prepričevanjem potrošnikov, na kar je še posebej vplival premik na splet med pandemijo COVID-19. Prej so tržniki zbirali demografske podatke, spremljali trende in segmentirali potrošnike za ciljno usmerjanje. Tudi v fizičnih trgovinah so v bližini blagajn strateško postavljali predmete, kot so revije in žvečilni gumiji. Danes strategije, ki jih poganja umetna inteligenca, zbirajo obsežne podatke iz iskalnikov, družbenih medijev in aplikacij za pomoč pri oblikovanju trženjskih pristopov.

Podjetja zdaj proaktivno oblikujejo interakcije s potrošniki, da bi jih pritegnila k sodelovanju, pri čemer včasih prehajajo na mejo manipulacije. Za prilagajanje strategij uporabljajo podatke iz virov kot so sledenje lokaciji in družbeni mediji. To odraža premik k vplivanju na določeno vedenje potrošnikov in ne zgolj k njihovemu prepričevanju. Ta razvijajoči se pristop vključuje analizo čustvenih in psiholoških modelov ter izkorišča ranljivosti in strahove potrošnikov. S strojnim eksperimentiranjem lahko podjetja raziskujejo vzročne povezave med vedenjem in tržnimi strategijami. Veliki podatki podjetjem omogočajo spreminjanje vedenja s pomočjo paradigem osebne identitete, kar se kaže v aplikacijah, kot sta sledenje lokaciji in prepoznavanje obraza. Razvoj programske opreme umetne inteligence vključuje nevroznanost, psihologijo in trženje ter ustvarja učinkovite promocijske metode. Preprosti algoritmi analizirajo spletno vedenje in ustvarjajo natančne profile uporabnikov. Poglobljeno učenje omogoča podrobne profile, ki presegajo celo tesna razmerja. Korporacije uporabljajo ta spoznanja in s pomočjo psihologov in nevroznanstvenikov iščejo "gumb za nakup potrošnika" [8].

Poleg tega tehnologije za branje misli postopoma pridobivajo sposobnost namernega prilagajanja dejanj kognitivnim in čustvenim stanjem vpletenih strani. Ta strateški pristop temelji na "čustveni umetni inteligenci". To pomeni prepoznavanje duševnih stanj s tehnikami strojnega učenja, ki pogosto uporabljajo globoke nevronske mreže. Čustvena umetna inteligenca služi različnim namenom, od povečanja varnosti v cestnem prometu (na primer spremljanje voznikov) do usmerjenega oglaševanja [9].

Čeprav se je tovrstna psihološka analiza, ki temelji na temeljni teoriji čustev, soočila s kritikami, da je preveč poenostavljena in nezanesljiva, je tehnologija na nekaterih področjih že pokazala impresivno raven učinkovitosti, sčasoma pa naj bi se še izboljšala. Primerljive tehnike bi lahko uporabili za odkrivanje in izkoriščanje kognitivnih pristranskosti pri potencialnih sogovornikih [9].

Medtem ko je potencial strojnega učenja za izkoriščanje kognitivnih in čustvenih ranljivosti sogovornikov očiten v postopku modeliranja, je empirično preverjanje takšne manipulacije v praksi precej zahtevno zaradi pogosto skrivnostne narave teh modelov. Kljub temu je v literaturi mogoče najti nekaj primerov potencialno manipulativnega izvajanja umetne inteligence. Podjetje eyeQ je na primer razvilo orodje, ki v realnem času skenira obrazno mimiko kupcev v trgovinah in analizira čustva in druge dejavnike, nato pa na podlagi teh podatkov prilagaja trženjske strategije v trgovinah. Takšne prakse lahko postanejo manipulativne, če so čustva, na katera se cilja, še posebej intenzivna (intenzivnost), če se jim pridružijo pristranskosti (kombinacija) ali če je kontekst odločanja zelo zapleten (kompleksnost), kot so razmere, ki vključujejo preobremenjenost z izbiro in zahtevne primerjave v trgovini [9].

4 MANIPULIRANA UMETNA INTELIGENCA

Do sedaj je bilo govora o manipulativnih vplivih umetne intelligence na človeka, obstaja pa tudi obratni vpliv – umetna inteligenca je lahko žrtev manipulacije.

Globoke nevronske mreže (DNN) so v zadnjem času pokazale izjemne rezultate, ki pogosto presegajo rezultate na človeški ravni, zlasti pri nalogah, povezanih z vidno klasifikacijo [10]. Visoka zmogljivost DNN pri razvrščanju vidnih objektov sproža vprašanja o razlikah, ki še vedno obstajajo med računalniškim in človeškim vidom – v nasprotju z vidnim sistemom so DNN veliko bolj občutljive na minimalne perturbacije.

"Napad z enim pikslom" (ang. One pixel attack) je specializirana vrsta napada na globoke nevronske mreže, katerega cilj je prevarati nevronsko mrežo, da napačno razvrsti sliko z minimalnimi spremembami, pri čemer se običajno spremeni le en ali nekaj pikslov. Pomemben izziv pri napadih z enim pikslom je najti ravnovesje med tem, da nevronska mreža napačno razvrsti sliko, in tem, da so spremembe na sliki dovolj subtilne, da jih človek ne more odkriti [10].

Taki napadi opozarjajo na dovzetnost DNN za subtilne vhodne manipulacije, kar lahko vpliva na varnost, zanesljivost in etične vidike rabe umetne inteligence. Raziskave na tem področju se ukvarjajo s strategijami za blažitev teh ranljivosti in izboljšanje robustnosti sistemov umetne inteligence.

5 SINTEZA

Če se prizadevanja na področju politike in izobraževanja ne bodo uresničila in če širši cilj gojenja etične umetne inteligence ne bo uspešen, obstaja možnost etične, družbene in gospodarske katastrofe ter z njo povezanimi vplivi na ljudi, nečloveške entitete in okolje. Ta nevarnost ni povezana z oddaljenimi apokaliptičnimi vizijami. Namesto tega izhaja iz postopnega, določenega stopnjevanja tehnoloških nevarnosti in posledičnega povečanja občutljivosti na človeških, družbenih, gospodarskih in okoljskih področjih. To povečevanje tveganj in ranljivosti izhaja iz etičnih težav, ki zajemajo nepoučeno in nepremišljeno uporabo naprednih tehnologij avtomatizacije, kot je umetna inteligenca. Razlika v izobrazbi verjetno povečuje širše posledice tveganj, povezanih z umetno inteligenco.

Trenutno ne obstaja univerzalno dovoljenje za uporabo umetne inteligence, prav tako ni obveznega izobraževanja o etiki umetne inteligence za tehnične raziskovalce, poslovne strokovnjake, državne upravitelje in druge deležnike, ki sodelujejo pri ustvarjanju, uporabi in urejanju umetne inteligence. Precejšen del neregulirane umetne inteligence je v rokah tistih, ki ne razumejo s tem povezanih tveganj in etičnih dilem ali pa imajo morda napačna pričakovanja glede te tehnologije. Nevarnost se skriva v tem, da ima človek moč brez razumevanja, kar pomeni neodgovorno ravnanje. Takšno ravnanje lahko povzroči neupravičene posledice za druge. Predpostavljanje nevtralnosti umetne inteligence in njena uporaba brez razumevanja posledic prispevata k brezbrižnosti. Izobraževalna politika lahko prispeva k izboljšanju tega položaja in tako spodbuja etično in smiselno umetno inteligenco [11]. Kljub temu so nekatera vztrajna, morda neprijetna vprašanja v razpravah o etiki in politiki umetne inteligence običajno odrinjena na stranski tir, čeprav si zaslužijo analizo. Ali se etika umetne inteligence ukvarja izključno z dobrobitjo in vrednotami ljudi ali bi morala vključevati tudi vrednote, dobrine in interese nečloveških entitet?

Osrednja sporna točka se vrti okoli kršenja zasebnosti. Uporaba umetne inteligence za usmerjanje vedenja pogosto zahteva obsežno zbiranje podatkov, kar sproža razprave o tem, ali to pomeni kršitev zasebnosti. Pri tem vprašanju strinjajo tako aktivisti kot veliki tehnološki konglomerati, ki priznavajo občutljivo ravnovesje med izkoriščanjem zmogljivosti umetne inteligence in varovanjem pravic posameznikov do zasebnosti. Sestavni del te razprave je področje ciljno usmerjenega oglaševanja. Medtem ko tržniki zagovarjajo njegovo zmožnost povezovanja potrošnikov z želenimi izdelki, ga vse več ljudi dojema kot manipulativen poseg, ki spodbuja impulzivne vzorce nakupovanja. Ta razhajanja v stališčih poudarjajo raznovrstne napetosti med komercialnimi cilji in avtonomijo potrošnikov. V tem zapletenem okolju sta soglasje in ozaveščenost ključnega pomena. Bistvo je v ugotavljanju, v kolikšni meri uporabniki razumejo in podpirajo vlogo algoritmov umetne inteligence pri oblikovanju njihovih spletnih izkušenj. Nazoren primer je nenamerno manipuliranje s kanali družbenih medijev s strani umetne inteligence, ki pogosto deluje mimo zavedanja uporabnikov. To sproža pomisleke o etičnih mejah vpliva in nujnosti informiranega soglasja. Škandal Cambridge Analytica je izpostavil problematiko glede zmožnosti umetne inteligence, da z izkoriščanjem strahov in predsodkov manipulira z odločitvami volivcev in krha temelje demokracije [11].

V ozadju tega dogajanja je sporno vprašanje čustvene manipulacije. Kakšne so moralne posledice dejstva, da umetna inteligenca izkorišča človeška čustva za komercialne koristi in še več, kdo bo nosil odgovornost? Debata manipulativne umetne inteligence postavlja vprašanja o kršitvah zasebnosti, usmerjenem vplivu, demokratični integriteti in čustveni etiki. Zaskrbljenost in dvom kliče po vzpostavitvi etičnih okvirov in skupnih prizadevanjih za usmerjanje vloge umetne inteligence pri sooblikovanju modernega sveta.

REFERENCES

[1] Micah Carroll, Alan Chan, Henry Ashton, in David Krueger, 2023. Characterizing Manipulation from AI Systems. *arXiv preprint arXiv:2303.09387.*

[2] Robert Noggle, 2020. Pressure, trickery, and a unified account of manipulation. *American Philosophical Quarterly*, *57*(3), 241-252.

[3] Federico Galli, 2020. AI and Consumers Manipulation: What the Role of EU Fair Marketing Law?, *Católica Law Review* 4, no. 2: 35–64, DOI: https://doi.org/10.34632/catolicalawreview.2020.9320.

[4] Daniel Susser, Beate Roessler in Helen Nissenbaum, 2019. Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.*, *4*, 1.

[5] Sandra C. Matz, Ruth E. Appel in Michal Kosinski, 2020. Privacy in the Age of Psychological Targeting, *Current Opinion in Psychology,* 116–121, DOI: https://doi.org/10.1016/j.copsyc.2019.08.010.

[6] Shahana Rayhan in Swayan Rayhan, 2023. The Role of AI in Democratic Systems: Implications for Privacy, Security, and Political Manipulation.

[7] Caroline Serbanescu, 2021. Why Does Artificial Intelligence Challenge Democracy? A Critical Analysis of the Nature of the Challenges Posed by AI-Enabled Manipulation, *Retskraft – Copenhagen Journal of Legal Studies*, 5 (1), 105-128.

[8] Jordan C'rene Reuille-Dupont, 2023. The Power of Algorithms and Big Data: A Marketing Perspective on Consumer Manipulation in Business. *University Honors Theses.* DOI: https://doi.org/10.15760/honors.1338

[9] Phillip Hacker, 2021. Manipulation by algorithms. Exploring the triangle of unfair commercial practice, data protection, and privacy law. *Eur Law J.* 2021;1–34. DOI: https://doi.org/10.1111/eulj.12389

[10] Anh Nguyen, Jason Yosinski in Jeff Clune, 2014. Deep Neural Networks are easily fooled: High confidence predictions for unrecognizable images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[11] Mark Coeckelbergh, 2020. *AI Ethics.* Cambridge, Massachusetts, USA: The MIT Press.

Users' Cognitive Processes in a User Interface for Predicting Football Match Results

Žiga Kolar
Department of Intelligent Systems
Jožef Stefan International Postgraduate School
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
ziga.kolar@ijs.si

Gregor Papa
Department of Computer Systems
Jožef Stefan International Postgraduate School
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
gregor.papa@ijs.si

ABSTRACT

The article discusses the importance of understanding cognitive processes, which are the foundation for effective user interface design. The authors highlight the crucial role of intuitive and user-friendly interfaces in online football match prediction platforms, as it is the design of the interface that can have a significant impact on the success of these types of platforms. The article delves into the thought processes of users who want a visually appealing, accessible and intuitive user interface for predicting football match results. Particular emphasis is given to understanding the psychology and behaviour of users. The paper aims to provide valuable insights that can guide designers and developers in designing more effective, user-centred interfaces for football match prediction websites. By combining scientific principles and practical user interface design strategies, the authors set new standards for user interface design in the football prediction industry.

KEYWORDS

user interface, cognitive process, football match, design, psychology

1 INTRODUCTION

The ever-evolving digital landscape has spurred the transformation of numerous industries, with sports prediction and betting platforms being no exception. More specifically, football predicting websites have emerged as a thriving subset within this sphere, effectively marrying the passion for sports with the advent of data-driven prediction models. Yet, the distinguishing factor amongst the multitude of similar platforms often boils down to one crucial aspect: the user interface (UI). An intuitive, user-friendly interface is paramount for ensuring user engagement, satisfaction, and ultimately, platform success.

This article presents an exploration of some of the cognitive processes underlying the creation of an effective user interface for a football predicting website. It aims to decode the mind process that designers employ when sculpting a user interface that not only visually appeals but also provides an engaging, accessible, and intuitive user experience. By delving into cognitive science principles, UI design best practices, and specific industry needs, the subsequent sections will delineate a novel approach to UI design for football predicting websites, highlighting the importance of understanding user psychology and behavior in the design process.

Furthermore, this article will underline the reciprocal relationship between UI design and user satisfaction, which can drastically impact user retention and overall platform success. Through the lens of cognitive science and user experience design, we aspire to provide valuable insights that can guide designers and developers in creating more effective, user-centric interfaces for football predicting websites. By marrying scientific principles with practical UI design strategies, we hope to set a new standard for user interface design within the football prediction industry.

While the context is football-oriented, the core concepts can be applied across various industries and areas. Here are a few areas where these cognitive processes might be useful:

1. **E-Commerce Platforms**: An understanding of user psychology and behavior can help design interfaces that make it easier for customers to browse, select, and purchase items, thereby enhancing the user experience and potentially increasing sales.
2. **Healthcare**: In telemedicine apps or healthcare software, an intuitive interface can facilitate easier communication between patients and providers, ensure critical information is presented clearly, and aid in the monitoring and input of health data.
3. **Banking and Finance**: For mobile banking apps or financial platforms, understanding the cognitive process can guide designers to create interfaces that allow users to safely and effectively manage their finances, conduct transactions, or make investments.
4. **Education**: Online learning platforms can benefit from intuitive interfaces, allowing students to navigate courses, interact with content, and assess their progress smoothly.

The paper consists of the following sections: section 2 describes the problem domain, section 3 describes related work, and section 4 presents the user interface. The paper ends with a conclusion and suggestions for further work.

2 DESCRIPTION OF THE PROBLEM DOMAIN

The purpose of this section is to present the required functionalities of the designed platform. In addition to describing the thought process, one of the main goals of this paper is to create high quality screen masks of a football match prediction user interface that will allow users to record the correctness of predictions made for a given football match and reward the user for each correct prediction.

A prediction consists of the number of goals scored by the home team and the number of goals scored by the away team. Ensuring that users have the freedom to choose which matches to predict is crucial for their satisfaction. They can predict the

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

results for several matches. They should also be able to change their prediction, which again fits in with the concept of a user-friendly interface. The system must finish receiving predictions 5 minutes before the start of the match. This restriction prevents fraud in terms of changing the predictions during the match itself.

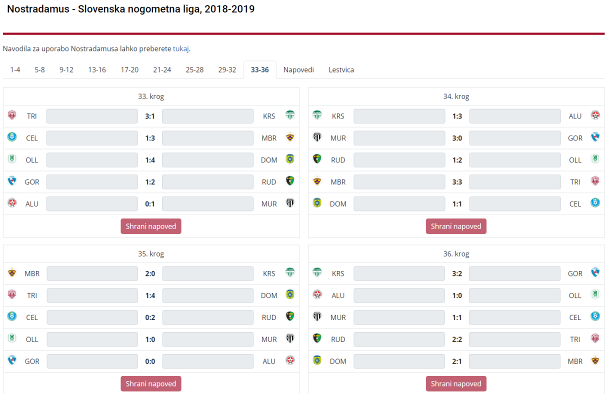
Users are awarded points based on their prediction accuracy in a competitive system. They receive 3 points for correctly predicting both the match's outcome and the winner, 2 points for correctly predicting the winner and goal difference, and 1 point for just predicting the winner correctly. Incorrect predictions earn zero points. The scoring system aims to ensure users are rewarded even for partially correct predictions, and that fully correct predictions do not create a significant advantage. To keep users engaged and mitigate large point gaps, the system awards double points for the first 10 predictions a user makes after a break of more than two weeks. This method aims to motivate users to return to making predictions and keeps the prediction system competitive and intriguing over extended periods.

Users have to log in to the system with a username or password. Before logging in, they must register. If they choose a user account with a membership fee when registering, they have to pay a membership fee of EUR 10 once a year. This is to ensure that the system works without any commercial loss and that the amount is not too high, which could discourage users from using the system. Once the users have paid the membership fee, they can immediately start predicting matches. There is also the possibility to make predictions without a membership fee. When registering, users can choose to create a free account with no membership fee. In this case, they do not have all the benefits. Users without a membership fee can only predict a maximum of 5 matches per league and users without a membership fee are not eligible for cash prizes. With the free account, we wanted to attract users who would like to make predictions and don't care about cash prizes or don't want to pay the entry fee. However, we have also made it possible to offer more benefits to users who have opted for the membership fee by means of restrictions, thus in a way rewarding and motivating them to submit their predictions.

To boost user engagement in making predictions, a monthly cash prize is offered to the top 50 users. The prize fund is calculated based on a formula that considers both new and existing memberships. The distribution of the prize pool is tiered to keep the competition interesting and to ensure fairness: the top 5 users share half the prize money, while the rest is divided among users ranked 6th to 50th. This compromise between prize size and user ranking aims to motivate users and keep them engaged, while also ensuring a fair distribution of rewards.

3 RELATED WORK

During major sports events, RTV SLO hosts a free-to-use website named Nostradamus for football and basketball match predictions. Users can make and change their predictions for any match until the match begins, with points awarded for correct predictions. At the end of each competition, the top five users receive practical prizes like balls or jerseys. Figure 1 illustrates the prediction submission process for the 33rd to 36th rounds of the 2018-19 1st Slovenian Football League season. As the depicted matches are past events, the text fields are greyed out and predictions are disabled. Once users have made their prediction, they can save it by clicking the "Save prediction" button. Each tab represents four competition rounds. A "Leaderboard" tab provides an overview of the user rankings, displaying 100 users per page, with a filter available to search for users by username.



Nostradamus - Slovenska nogometna liga, 2018-2019

Navodila za uporabo Nostradamusa lahko preberete tukaj.

33. krog

TRI	3:1	KRS
CEL	1:3	MBR
OLL	1:4	DOM
GOR	1:2	RUD
ALU	0:1	MUR

34. krog

KRS	1:3	ALU
MUR	3:0	GOR
RUD	1:2	OLL
MBR	3:3	TRI
DOM	1:1	CEL

35. krog

MBR	2:0	KRS
TRI	1:4	DOM
CEL	0:2	RUD
OLL	1:0	MUR
GOR	0:0	ALU

36. krog

KRS	3:2	GOR
ALU	1:0	OLL
MUR	1:1	CEL
RUD	2:2	TRI
DOM	2:1	MBR

Figure 1: Making predictions on Nostradamus website.

The Nostradamus app has several strengths and weaknesses. On the upside, it's simple to operate and maintains a consistent, unambiguous website design. However, it also has several drawbacks. The application doesn't cater to user preferences, forcing users to choose from all matches rather than tailoring options to their liking. It lacks a 'Favourites' tab for easy league-switching. Furthermore, rewards are only given to the top five users. This design choice leads to user attrition, as many stop using the app after prolonged failure, unable to compete with the top performers.

Furthermore, the app is entirely free, which means it operates at a loss as practical rewards have to be given to the top five users. The use of club abbreviations instead of full names can also confuse new users. There is a lack of real-time information on the countdown towards the closing of predictions. The app only displays four competition rounds in one tab simultaneously, which can be limiting. Finally, due to this display limitation, some users may forget to hit the 'Save Prediction' button after entering their results, leading to potential disappointment and frustration.

In addition to Nostradamus, some research has also been done into making the user interface better. Gonzales [2] conducted a laboratory experiment investigating the impact of images, transitions, and interactivity in animated interfaces on decision making in two different domains. These interfaces incorporated either realistic or abstract images, smooth or abrupt transitions, and parallel or sequential interactivities. The results demonstrated that the task domain, user experience, and the types of images, transitions, and interactivity methods implemented all affect decision performance. Participants were observed to perform better with animations that utilized realistic images, smooth rather than fast transitions, and both parallel and sequential interactivity.

Shneidermann [6] recommends three pillars to support the UI design process: guidance documents, UI management systems and interactive usability testing labs. Five basic interaction styles are presented: menu selection, form filling, command language, natural language and direct manipulation. The author encourages more attention to direct manipulation, where objects and actions are visible, actions are triggered by selection or pointing, and the effect is immediately visible and correctable.

Sharma and Tiwari [4] introduce the concept of user interface and user experience, which play a very important role in today's technical and modern world. According to them, user interface consists of guidelines, workflows, system colours, design process, etc. User experience consists of the process of how the user can experience the application in the best way.

4 USER INTERFACE

One of the studies indicate that a primary objective of UI is to offer users streamlined and intuitive ways to interact with computer systems, thereby enhancing task efficiency and reducing cognitive workload and stress [5]. The other study highlights that consistency in UI enables users to build an accurate mental model of the way it works. Furthermore, an expert user should interact with the system as easily as a novice user. Both users should be satisfied with the same system [1]. In this paper, we wanted to ensure as much consistency and flexibility as possible.

Before designing the UI, we had to think carefully about what the main colour of the UI would be. The guidelines for a good and user-friendly user interface recommend that we ensure as much consistency as possible, which means that only one colour should be the main part of the website. Research [3] has shown that red works better for tasks that require a lot of detail, while blue is more effective for creative tasks. Another study [8] confirms the results of the previous study and adds that blue also works well for light and heavy tasks.

We chose blue (RGB = (58, 98, 215)) color because it is a good combination that stimulates both creative thinking and attention to detail. Creative thinking is needed when determining the outcome, as a match can have many possible outcomes. In addition, the user also needs to take into account a huge number of details, as the result itself can be affected by injuries to key players, current form, the match venue, etc. The header, footer and frame colour of the website is always blue. This ensures greater consistency.

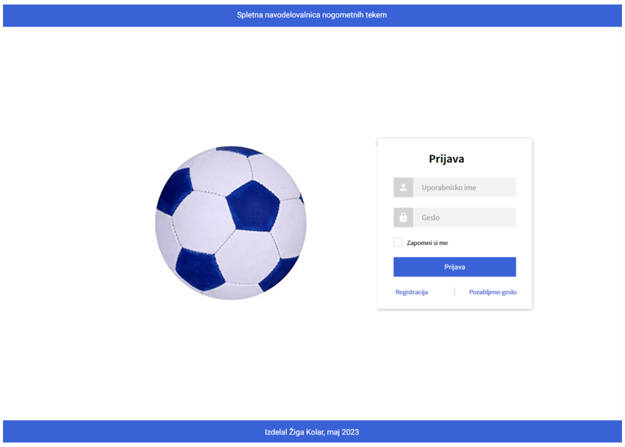
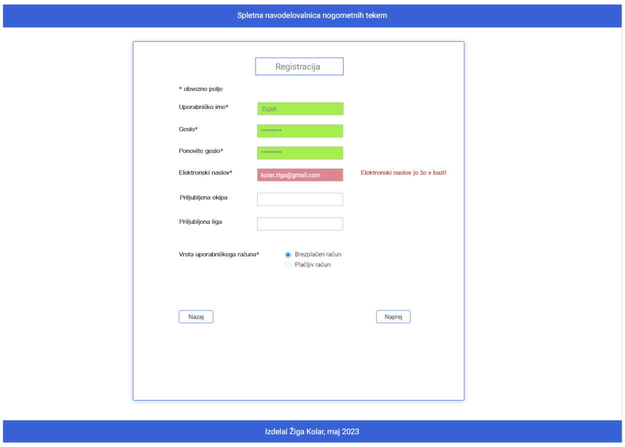


Figure 2: Login page.

As with most other websites, everything always starts with the creation of the login and registration page. The login page is shown in Figure 2 and has all the classic login elements (a text box for username and password, a login button, a link to registration and a link to forgotten password). The page has a ball in the middle in blue. The ball immediately gives user an association with football. Two additional icons next to the username and password also provide the association to the username and password. The buttons, the colour of the pentagons on the ball and the links are in blue, which is the main colour used to maintain consistency.



Spletna navodelovalnica nogometnih tekem

Registracija

* obvezno polje

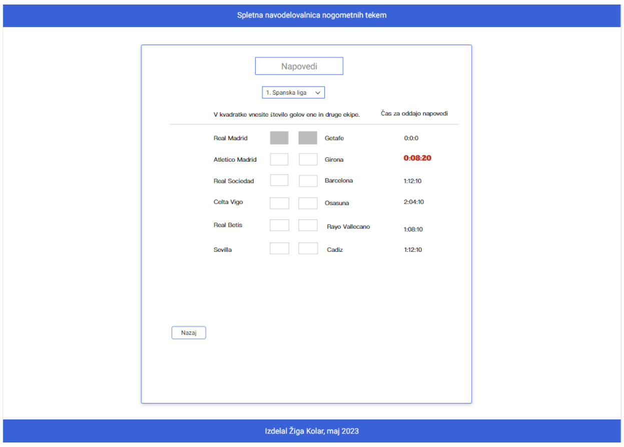
Uporabniško ime*	ZigaK	
Geslo*	********	
Ponovite geslo*	********	
Elektronski naslov*	kolar.ziga@gmail.com	Elektronski naslov je že v bazi!
Priljubljena ekipa		
Priljubljena liga		

Vrsta uporabniškega računa*
● Brezplačen račun
○ Plačljiv račun

Nazaj Naprej

Izdelal Žiga Kolar, maj 2023

Figure 3: Registration form.



Spletna navodelovalnica nogometnih tekem

Napovedi

1. Spanska liga

V kvadratke vnesite število golov ene in druge ekipe.				Čas za oddajo napovedi
Real Madrid	▨	▨	Getafe	0:0:0
Atletico Madrid	□	□	Girona	0:08:20
Real Sociedad	□	□	Barcelona	1:12:10
Celta Vigo	□	□	Osasuna	2:04:10
Real Betis	□	□	Rayo Vallecano	1:08:10
Sevilla	□	□	Cadiz	1:12:10

Nazaj

Izdelal Žiga Kolar, maj 2023

Figure 4: Match results predictions.

The registration form is in the classic form of text fields and is shown in Figure 3. If we had a list of all the mandatory fields at the top of the form, we would run into a problem as people sometimes do not read the instructions at the top of the form or forget them. It is therefore easier and more efficient to mark all mandatory fields with an asterisk. The two buttons for selecting the type of user account were chosen because they are simple and intuitive, offering the user a quick choice due to their mutually exclusive nature. A back button has also been added, as we need to give the user the option to return to the previous page in each step. Without this, the user would be confused.

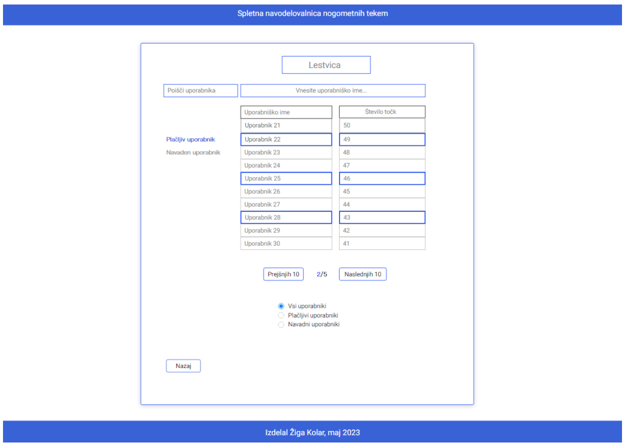
When a field is completed, the system checks that it is correct. If the field is correct, it is highlighted in green, if it is incorrect, it is highlighted in red and an error message is displayed next to it. The green colour was chosen to make an analogy with a traffic light. If the light turns green, we can proceed. If the light is red, we have to stop. The same applies to all fields in the registration. An example is shown in Figure 3. The username and password are correct, so the fields will turn green. The email address is incorrectly entered, so the field will turn red and a message will appear next to it.

Upon completing registration, users are notified of their successful registration with a blue tick icon, representing the process's completion. To confirm and finalize the registration, the users are sent an email containing a link. This link must be confirmed within a 24-hour window to ensure the user is not a bot, hence safeguarding the system.

To enter a prediction, the users first select which league to predict via a drop-down list. The drop-down list is an effective option for when there is not enough space to display all the tabs. Since there are too many possible leagues to display with tabs, we have opted for a drop-down list. The latter also has a special "Favourites" option, where the users can save all the match predictions from the leagues they are interested in. Only the matches from the leagues that the users have selected will be shown to the user. The advantage of this option is that it saves the users time in selecting the match predictions from the leagues they regularly predicts. Additionally, there is also a "Personalised" option where the user is shown matches based on his personal preferences. The system learns which matches and from which leagues the users like to predict and recommends them to the users. The personalisation is implemented using an artificial intelligence algorithm (recommendation system). This approach aims to allow the users to have all the matches they interested in in one place and to be able to make predictions as quickly as possible.

After selecting one of the options in the drop-down list, the matches of the current round are displayed. For the home and away team goal predictions, we have chosen a text box because this is a common practice in football match predictions. One possibility would be to have the user click on the + or - buttons instead of the text box to determine the number of goals, but this approach would be more time-consuming as it would require a larger number of clicks. Unlike Nostradamus [7], here the prediction is saved automatically after the users have filled in both text fields. This is to avoid a situation where the users enter the predictions but forgets to click on the "Save Prediction" button and is left with no points. We have decided to use full club names instead of abbreviations, as some users do not know all the abbreviations. Also, for each match, the user has a counter on the right until the start of the match or until the end of the prediction. This gives the users full control over how much time is left. In the last 10 minutes before the time expires, the text where the time is written will turn red and become whiter. This is to encourage the user to make a prediction. When the time for the prediction expires, the text fields turn grey and the prediction is disabled. The grey colour gives the user a feeling of closure and inaccessibility. An example is shown in Figure 4.

In addition to submitting a prediction, the users have the possibility to view the points scale. The scale is presented in tabular form because it is one of the most transparent and efficient options to quickly display the situation. Paid users are shown with a blue frame and ordinary users with a grey frame. We wanted to somehow reward paying users for their contribution with a blue colour, as this is the main colour of the page. 10 users are shown at a time, in descending order of points. The leaderboard allows filtering by username. In this way, the users can quickly find themselves and get an insight into the number of points without having to search through the whole list. An example of filtering is presented in Figure 5. Filtering by user type (paid, regular or all users) is also available. Filtering works on the basis of a button that allows fast and efficient filtering between users.


Spletna navodelovalnica nogometnih tekem
Lestvica
Poišči uporabnika
Vnesite uporabniško ime...
Plačljiv uporabnik
Navaden uporabnik
Prejšnjih 10
2/5
Naslednjih 10
Nazaj
Izdelal Žiga Kolar, maj 2023

Uporabniško ime	Število točk
Uporabnik 21	50
Uporabnik 22	49
Uporabnik 23	48
Uporabnik 24	47
Uporabnik 25	46
Uporabnik 26	45
Uporabnik 27	44
Uporabnik 28	43
Uporabnik 29	42
Uporabnik 30	41

● Vsi uporabniki
○ Plačljivi uporabniki
○ Navadni uporabniki

Figure 5: Leaderboard of all users.

5 CONCLUSION AND FUTURE WORK

This work is part of the thought process behind the development of a user interface for predicting football match results. We have highlighted the importance of understanding user needs and psychology in the design of such systems. We have identified the need for the user interface to be intuitive, efficient and flexible and have tried to present it in this way.

In the future we intend to focus on the development of the system and its use in practice. We will try to implement the user interface in Django (python) and in the web technologies HTML, CSS, Javascript and jQuery. We will use the MySQL Lite database. After the implementation is complete, we will hand the system over to a test group (up to 20 people). If the users are satisfied, we will try to recruit more people and will regularly update and maintain the system.

Taken together, our results in this paper are a step towards a better understanding of how designers and developers can improve football match result prediction tools, which could have important implications for improving the user experience.

ACKNOWLEDGMENTS

The authors acknowledge the funding from the Slovenian Research Agency (ARRS), Grant (PR-10495) and Basic core funding P2-0209.

REFERENCES

[1] Ali Sajedi Badashian, Mehregan Mahdavi, Amir Pourshirmohammadi, et al. 2008. Fundamental usability guidelines for user interface design. In *2008 International Conference on Computational Sciences and Its Applications*. IEEE, 106–113.

[2] Cleotilde Gonzalez. 1996. Does animation in user interfaces improve decision making? In *Proceedings of the SIGCHI conference on human factors in computing systems*, 27–34.

[3] Ravi Mehta and Rui Zhu. 2009. Blue or red? exploring the effect of color on cognitive task performances. *Science*, 323, 5918, 1226–1229.

[4] Vatsal Sharma and A Kumar Tiwari. 2021. A study on user interface and user experience designs and its tools. *World Journal of Research and Review (WJRR)*, 12, 6.

[5] Yu Shi, Eric Choi, Ronnie Taib, and Fang Chen. 2010. Designing cognition-adaptive human–computer interface for mission-critical systems. *Information Systems Development: Towards a Service Provision Society*, 111–119.

[6] Ben Shneiderman. 1988. We can design better user interfaces: a review of human-computer interaction styles. *Ergonomics*, 31, 5, 699–710.

[7] RTV SLO. 2023. Nostradamus. https://www.rtvslo.si/strani/nostradamus/1214. Accessed: 2023-06-13. (2023).

[8] Tiansheng Xia, Lu Song, Ting T Wang, Ling Tan, and Lei Mo. 2016. Exploring the effect of red and blue on cognitive task performances. *Frontiers in Psychology*, 7, 784.

Družbena regulacija umetne inteligence. Nekatera odprta vprašanja in izzivi.

Social regulation of artificial intelligence. Some open questions and challenges.

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

POVZETEK

S tem, ko se umetna inteligenca (UI) spreminja v sistemsko tehnologijo, ki ni več predmet ozkih laboratorijskih raziskav, temveč z različnimi vrstami aplikacij postaja integralni del družbe, se je v številnih ozirih znašla na prelomni točki. Nenazadnje predstavlja eno tistih naprednih tehnologij, ki gledehitrosti razvoja in širjenja v družbi nima primera v zgodovini. V prispevku se bomo ukvarjali z nekaterimi vprašanji družbene regulacije UI, s poudarkom na najnovejše kompleksne jezikovne modele. Izhajali bomo iz teze, da se v primeru vrednotenja prihodnjega razvoja UI in s tem povezanih tveganj Evropa, katere del smo tudi mi, ne bi smel v celoti podrediti svariteljskim načelom. Potrebno je najti ustrezno ravnovesje med svariteljskimi načeli in proakcijskimi načeli

KLJUČNE BESEDE

umetna inteligenca, chat GPT, družbena regulacija, tveganje, bioinformacije

ABSTRACT

As artificial intelligence (AI) evolves into a systemic technology no longer confined to narrow laboratory research but rather integrated into society through various applications, it finds itself in many aspects at a cross point. Last but not least AI represents one of those most progressed modern technologies that, in terms of the speed of development and dissemination in society, has no precedent in previous history of technological development. In this contribution, we will address some of the issues of social regulation of AI, with a focus on the latest complex language models. In our discussion, we will start from the premise that in the case of risk assessment and risk regulation of the future development of AI, Europe, of which we are also a part, should not be strictly submitted to precautionary principles. It is necessary to find the appropriate balance between precautionary and proactionary principles.

KEYWORDS

artificial intelligence, Chat GPT, social regulations, risks, bioinformations

1 DRUŽBENA REGULACIJA UMETNE INTELIGENCE. NEKATERA ODPRTA VPRAŠANJA IN IZZIVI

S tem, ko se umetna inteligenca (UI) spreminja v sistemsko tehnologijo, ki ni več predmet ozkih laboratorijskih raziskav, temveč z različnimi vrstami aplikacij postaja integralni del družbe, se je v številnih ozirih znašla na prelomni točki. Nenazadnje predstavlja eno tistih naprednih tehnologij, ki glede hitrosti razvoja in širjenja v družbi nima primera v zgodovini. Pomislimo zgolj na družbeno difuzijo ChatGPT, zadnjega hita generativne UI, ki je samo nekaj mesecev za tem, ko je bila lansirana v družbeni prostor, dosegla več stomilijonsko uporabo. V zadnjem tričetrtletju se uporaba Chat GpT razširila do neslutenih meja.

Glede na izredne revolucionarne preskoke, ki jih v zadnjem času dela UI in glede na njen izredni aplikativni potencial, ki obeta, da bo dodobra spremenil življenja ljudi, se bomo v prispevku ukvarjali z nekaterimi vprašanji družbene regulacije UI, s poudarkom na njen najnovejše kompleksne jezikovne modele (npr.: chat GPT). Kot bomo skušali v prispevku še posebej opozoriti, se največji potencial UI kaže skozi procese konvergiranja UI z drugimi naprednimi tehnologijami. Koncept konvergentnega tehnološkega razvoja je danes izredno aktualen. Prizadevanja za združitev tehnologije kvantnega računalništva z UI bo odprlo prostor kompleksnim analizam velikih podatkovnih baz. Dvig računalniških kapacitet predstavlja ključni steber nadaljnjega napredka UI. Tu lahko pričakujemo v prihodnosti številne revolucionarne premike. Zdi se, da bomo podobne revolucionarne preskoke srečevali tudi v primeru nadaljnjega povezovanja UI in najnovejših tehnologij genskega inžineringa. Vzajemni razvoj obeh transformativnih tehnologij ima izredno velik aplikativni potencial. Na primer, na vseh področjih t.i. »omike« (genomike, proteomike, epigenomike, itd.) se na temelju UI, ki temelji na modelih globokih nevronskih mrež [1],

[*] Article Title Footnote needs to be captured as Title Note
[†] Author Footnote to be captured as Author Note

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

[2] [3], odpira prostor »tarčno« usmerjenem zdravljenju ljudi oziroma - kot to področje biomedicine danes označujejo strokovnjaki – t.i. personalizirani medicini [4].

V luči tega napredka in aplikacije tehnologije UI je izredno aktualno vprašanje njene regulacije na globalni ravni. Pri tem se srečujemo s številnimi dilemami in vprašanji. To je na nek način tudi razumljivo. Ustrezni regulativni mehanizmi, ki bi – kot primer – preprečili morebitna tveganja in etično sporna dejanja pri nadaljnjem razvoju UI, se namreč ne morejo pojaviti kar čez noč. To se ni zgodilo tudi nikoli prej v zgodovini znanstveno-tehnološkega razvoja. V tem smislu bi trenutno situacijo v iskanju ustreznih regulativnih mehanizmov na področju UI lahko v najboljšem slučaju primerjali s pojavom nekega drugega revolucionarnega tehnološkega artefakta, t.j. avtomobila, za katerega, potem ko se je prvikrat znašel na ulicah mest, še ni bilo oblikovanih vseh prometnih in varnostno-tehničnih predpisov, tako kot jih poznamo danes. Vse to je prišlo kasneje. Zato smo, kot pravijo Haroon Sheikh in ostali [5], trenutno v fazi, ko se še vedno lahko pojavi veliko napak. Ker so v primeru napak družbena tveganja, ko gre za tehnologijo UI, izredno velika, je toliko bolj pomembno, da pridemo čim prej do čim bolj premišljenih in celovitih družbenih mehanizmov regulacije UI. V prispevku bomo izhajali iz teze, da v primeru vrednotenja prihodnjega razvoja UI in s tem povezanih tveganj se današnji svet, še zlasti pa Evropa, katere del smo tudi mi, ne bi smel v celoti podrediti t.i. svariteljskim načelom (t.i. precautionary principles). Tu je treba najti neko ustrezno ravnovesje med svariteljskimi načeli in proakcijskimi (t.i. proactionary principles). V zvezi s tem vprašanjem bomo smiselno vključili spoznanja nekaterih vodilnih transhumanističnih mislecev današnjega časa, ki se ne ukvarjajo z imaginariji prihodnosti, ki so že na meji znanstvene fantastike in daleč od realnih problemov, temveč z aktualnimi in zelo realnimi vprašanji anticipativnega načrtovanja znanstvene in tehnološke prihodnosti [6] [7]. V prispevku bomo namreč skušali opozoriti, nenazadnje tudi v duhu ravno predhodno omenjenega transhumanizma, da so današnji znanstveni in filozofski premisleki kar preveč okupirani s prikazovanjem distopičnih scenarijev prihodnjega razvoja UI. Pri čemer ta delitev sploh ne poteka na temelju neke stroge delitve med humanistično-družboslovno (apriorna skepsa in strah pred novimi tehnologijami) in naravoslovno-tehnično (nekritično sprejemanje novih tehnoloških rešitev) mislijo. Meje med obema znanstvenima poloma so zlasti v primeru UI precej bolj zabrisane. Skozi predstavitev stališč humanističnih (npr. Yuval Noah Harari, Amitalom Etzioni, Noam Chomsky) in informacijsko-komunikacijskih znanstvenikov (Geoffrey Hinton, Yuhuai Wu, Kristian Kersten, etc.) – te zadnje imamo lahko za »botre« UI, ki temelji na modelih globokih nevronskih mrež - bomo skušali pokazati na vso kontradiktornost in neenotnost današnjih ekspertnih stališč o UI. V prispevku bomo izhajali iz predpostavke, da so najmanj konstruktivni kataklizmični prikazi prihodnjega razvoja UI. Četudi ne bomo v celoti zavrnili pogledov tistih avtorjev, ki pravijo, da vstopamo v novo fazo razvoja, ko že imamo opravka s teoretsko razmišljujočimi stroji, ki dosegajo oziroma bodo kmalu dosegli človekovo zavestno raven spoznavanja [8], bo za nas konstruktivni pogled o teh vprašanjih predstavlja razprava o aktualnih (realnih) družbenih tveganjih UI. In če že govorimo o tem, kako priti do najboljših možnih mehanizmov regulacije te nove napredne tehnologije, potem je pač potrebno sprejeti dejstvo, da prej ko bo vzpostavilo pravila sodelovanja z UI, prej se bomo naučili živeti v harmoniji s to novo napredno tehnologijo. Tu je pomembno izpostaviti, o čemer v zadnjem času na široko govorijo ravno računalniški eksperti, ki so dali največji zagon razvoju generativne UI [9], namreč, da bo UI največ prispevala k družbenemu blagostanju in krepitvi človekovih sposobnosti, če bo ostala osredotočena na človeka. To pomeni, da mora biti prioriteta dana opolnomočenju uporabnikov UI, ne pa njihovem nadomeščanju z UI.

V osrednjem delu naše razprave nas bo zanimalo vprašanje, kako se usklajujejo interesi med posameznimi družbenimi akterji, med ključnimi globalnimi akterji, med lokalnimi in globalnimi akterji itd., pri graditvi skupnega družbenega modela upravljanja UI. Zavedati se moramo, da je ta konflikt interesov med različnimi družbenimi, političnimi in ekonomski subjekti na tej stopnji razvoja UI izredno močan. Čeprav se navzven zdi včasih ravno obratno. Zelo ilustrativen je naslednji primer: četudi je Sam Altman, izvršni direktor tehnološkega podjetja Open AI ameriškim kongresnikom še maja letos zagotavljal, da je osnovni pogoj za uspešen in skladen razvoj UI najbolj pomembna vzpostavitev ustreznih mehanizmov družbene regulacije UI, je hkrati dobro znano, kako močno je podjetje Open AI nastopilo zoper zahtevo Evropske komisije (celo z grožnjo izstopa iz trga EU), da se zadeve na tem področju zakonsko uredijo. Open AI je, tako kot večina ostalih tehnoloških velikanov, ki razvijajo najnovejše sisteme UI, v najboljšem primeru pripravljen sprejeti zelo nezavezujoča priporočila mednarodnih teles, ne pa se podrejati bolj restriktivnim zakonskim pravilom. V prispevku bomo predstavili razloge, zakaj je izmed treh ključnih globalnih akterjev, ki razvijajo UI, t.j. ZDA, Kitajska in Evropska Unija, na področju izgradnje mehanizmov družbene regulacije UI, predvsem z vidika njenega družbenega nadzora in preprečevanja možnih tveganj, še največ storila ravno Evropska Unija. Kritično bomo ovrednotili elemente zakona o UI (prepoved UI za prepoznavanje čustev, prepoved uporabe biometrije v realnem času na javnih mestih, prepoved uvajanja socialnega točkovanja, omejitve glede uporabe generativne UI, itd.), ki naj bi bil v okviru EU sprejet v bližnji prihodnosti (potem ko ga je že potrdil Evropski parlament letos junija). Tudi Kitajska vzpostavlja svojo zakonodajo o UI, kar pa je v širši javnosti morda manj znano. Še najmanj je bilo storjenega v ZDA, četudi je tam sedež številnih vodilnih podjetij s področja UI [10]. Glede na situacijo bi v globalnih okvirjih ravno EU lahko nastopila v vlogi ključnega iniciatorja sprememb, že zato ker je prva začela posvečati pozornost temu vprašanju. [11] V okviru naše predstavitve odprtih vprašanj družbene regulacije UI se bomo še posebej zaustavili ob problemu velikih podatkovnih baz, ki so temelj nadaljnjemu razvoju generativne UI. Izhajamo namreč iz ocene, da četudi GDPR, ki je bil sprejet leta 2018 v članicah EU, predstavlja dober okvir za regulacijo velikih podatkovnih baz, ki so potrebne za razvoj AI, je njegova šibkost vendarle najbolj očitna na področju velikih biogenetskih podatkovnih baz. Torej na tistem področju, kjer kot smo že dejali, obstajajo zaradi spoja tehnologij biogenetike in UI možnosti za najbolj revolucionarne znanstvene preskoke. Skozi našo celotno razpravo bomo namreč izhajali iz predpostavke, da kolikor današnji razvoj UI merimo po kriterijih kot so kvaliteta temeljnega raziskovanja na področju računalniških algoritmov, razpoložljivost velikih podatkovnih baz, zahtevan razvoj »hardwara«, uspešna komercializacija te tehnologije in aktivna podpora politike, potem velike podatkovno baze, še zlasti, če predstavljajo spoj biogenetike in informatike, igrajo daleč najpomembnejšo vlogo.

LITERATURA

[1] Bertalan Meskó in Eric J. Topol (2023): The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Digital Medicine* (2023) 6:120 – 123.

[2] Leskovec Jure, Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Eric J. Topol, Pranav Rajpurkar (2023): Foundation models for generalist medical artificial intelligence. *Nature*, Vol 616, 13 April 2023, str. 259-265.

[3] Ahmed Zahlan, Ravi Prakash Ranjan, David Hayes (2023): Artificial intelligence innovation in healthcare: Literature review, exploratory analysis, and future research. *Technology in Society*, Volume 74, August 2023, 102321.

[4] Prainsack Barbara (2017) *Personalized Medicine. Empowered Patients in the 21st Century*. Cambridge: Cambridge University Press.

[5] Haroon Sheikh, Corien Prins, Erik Schrijvers (2023): *Mission AI. The New System Technology*. Cham: Springer.

[6] Fuller Steve in Veronika Lipin´ska (2014): *The Proactionary Imperative. A Foundation for Transhumanism*. New York in Hampshire: Palgrave Macmillan.

[7] Sorgner Stefan Lorenz (2021): *We Have Always Been Cyborgs. Digital Data, Gene Technologies, and an Ethics of Transhumanism*. Bristol: Bristol University Press

[8] Webb, Amy (2019*): The Big Nine. How the Tech Titans & Their Thinking Machines Could Warp Humanity*. New York: Public Affairs.

[9] Schneiderman Ben (2022): *Human-Centered AI*. New York, NY: Oxford University Press.

[10] Hutson Matthew (2023): Conlicting visions for AI regulation. *Nature*, Vol 620, 10 August 2023, str. 260-263.

[11] Fatos Selita, Robert Chapman, Yulia Kovas, Vanessa Smereczynska, Maxim Likhanov in Teemu Toivainen (2023): Consensus too soon: judges' and lawyers' views on genetic information use. *New Genetics and Society*, Vol. 42, No. 1, str. 1– 31.

Umetna inteligenca: orodje ali zavesten stroj

Artificial Intelligence: A Tool or Conscious Machine

Olga Markič
Filozofska fakulteta
Univerza v Ljubljani
Ljubljana, Slovenia
olga.markic@ff.uni-lj.si

POVZETEK

Umetna inteligenca (UI) je s pojavom ChatGPT prerasla samo strokovne diskusije. V širši javnosti se je sprožilo zanimanje in spraševanje o tem, kaj UI sploh je. UI je postala tema, ki buri duhove in vzbuja tako veliko navdušenje kot tudi pomisleke. V prispevku bom predstavila dva pogleda na UI. Prvi nanjo gleda kot na še enega v nizu orodij, ki so jih ljudje izoblikovali skozi zgodovino. Kot pametno orodje UI ljudem pomaga pri opravljanju različnih nalog, a hkrati njena uporaba odpira tudi vrsto epistemoloških, etičnih in družbenih vprašanj. Drugi pogled pa v UI vidi gradnjo mislečih in zavestnih strojev. Menim, da se precej strahu povezuje prav z bojaznijo, da bodo sistemi UI postali avtonomni in zavestni ter bodo zavladali nad ljudmi.

KLJUČNE BESEDE

umetna inteligenca, Chat GPT, pametna orodja, zavest, računski funkcionalizem, etika

ABSTRACT

With the advent of ChatGPT, artificial intelligence (AI) has outgrown only professional discussions. Interest and questioning of what AI is has sparked among the general public. AI has become a topic of both great enthusiasm and concern. In the article, I will present two views on AI. The first sees it as another in a set of tools people have shaped throughout history. As a smart tool, AI helps people to perform a variety of tasks, but at the same time, its use also raises a range of epistemological, ethical, and social issues. The second view sees AI as building thinking and conscious machines. I believe that a great deal of fear is associated with the possibility of autonomous and conscious AI systems that will start to dominate people.

KEYWORDS

artificial intelligence, Chat GPT, smart tools, consciousness, computational functionalism, ethics

1 UVOD

S pojavom klepetalnega robota ChatGTP in programov za ustvarjanje slik (npr. Midjourney) in videov (npr. Runway) je umetna inteligenca prerasla samo strokovne diskusije in postala tema, ki buri duhove in vzbuja tako veliko navdušenje kot tudi pomisleke. ChatGPT je zaradi dostopnosti in enostavnosti uporabe, ki vsaj na prvi pogled ne zahteva posebnega znanja, v širši javnosti sprožil tudi večje zanimanje in spraševanje o tem, kaj umetna inteligenca (UI) sploh je.

Sam izraz »umetna inteligenca« je nastal v času (1956) [1], ko je znanstvenike dejansko zanimalo, kako bi naredili stroj, ki bi lahko podvajal človeško mišljenje (prvi val UI). Raziskovanja so bila v domeni znanstvenikov, v javnost pa je tema UI prišla predvsem preko znanstvene fantastike (npr. Kubrickov film iz leta 1968 - 2001: Odiseja v vesolju). Dandanes pa je UI (drugi val) prisotna v našem vsakdanjem življenju, ne da bi se tega sploh zares zavedali, npr. ko odpremo telefon s pomočjo prepoznave prstnega odtisa ali obraza, poiščemo najhitrejšo pot do izbranega cilja, prevedemo sporočilo iz tujega jezika, ali kupimo izdelek v priljubljeni spletni trgovini, kjer nam potem program sam ponudi še vrsto drugih izbir, ki bi nas morda lahko zanimale. UI se uporablja kot pomoč pri odločanju v bančništvu, pravu, medicini, športu, znanosti, industriji, pa tudi vojski. V prispevku bom predstavila dva pogleda na UI. Prvi nanjo gleda kot na še enega v nizu orodij, ki so jih ljudje izoblikovali skozi zgodovino. Kot pametno orodje UI ljudem pomaga pri opravljanju različnih nalog, a hkrati njena uporaba odpira tudi vrsto epistemoloških, etičnih in družbenih vprašanj, ki so zdaj predmet razprav tako med znanstveniki kot tudi v širši javnosti. Drugi pogled pa v UI vidi gradnjo mislečih in zavestnih strojev. Ta pogled je bil prisoten predvsem v začetkih UI, se pa, kot bom nakazala v zadnjem delu, z idejo splošne UI, zopet vrača. Menim, da se precej strahu pred sistemi UI povezuje prav z bojaznijo, da bodo sistemi UI postali avtonomni in zavestni ter bodo zaradi večje učinkovitosti pri reševanju nalog zavladali nad ljudmi.

2 UI KOT PAMETNO ORODJE

Ko torej danes uporabljamo izraz UI, se večinoma nanašamo na tako imenovana pametna orodja, ki nam pomagajo na bolj učinkovit način rešiti določene miselne naloge. Filozof in kognitivni znanstvenik Andy Clark je provokativno zapisal: »Inteligenco uporabljamo za strukturiranje okolja, tako da lahko uspemo z *manj* inteligence. Naši možgani delajo svet pameten, tako da bomo lahko v miru neumni.« In nadaljeval; »Ali, če pogledamo z druge strani, možgani *in* deli zunanjega ogrodja na

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

koncu sestavljajo pameten, racionalen sklepalni stroj, ki ga imenujemo um.« [2].

Skozi zgodovino so ljudje oblikovali različna orodja, ki so imela velik vpliv na družbo in katerih uporaba je pripeljala do velikih družbenih sprememb (npr. parni stroj in industrijska revolucija). Večina orodij v preteklosti je ljudem pomagala pri fizičnih aktivnostih. Z njihovo pomočjo je človek lahko opravljal naloge, ki jih sicer zaradi narave svojega telesa ne bi mogel opravljati tako uspešno, na primer, sekira, plug, žerjav, kolo, vlak, letalo, parni stroj, telefon, mikroskop, teleskop, če jih naštejemo le nekaj iz dolge zgodovine. Je pa že v preteklosti človek izoblikoval tudi orodja, ki so pomagala pri miselnih nalogah. Eno takih je bila pisava, ki je ljudem omogočila, da se miselne vsebine ne prenašajo zgolj z govorom neposredno s človeka na človeka, ampak zapisane ostanejo dostopne širši množici ljudi v daljšem časovnem obdobju. S pisavo, natančneje načini zapisovanja, so potem povezana nova orodja, npr. papirus, tisk, digitalni zapis v elektronskih računalnikih. Prav tako so ljudje uporabljali zunanja pomagala za pomoč pri računanju (npr. abakus) in zapisovanju števil (npr. rovaš). Z iznajdbo elektronskega računalnika je človek dobil izredno močno orodje, ki ga lahko uporablja za pomoč pri odločanju, raziskovanju, zbiranju in dostopanju do informacij, v komunikaciji in umetniškem ustvarjanju. Računalničarji, ki načrtujejo sisteme drugega vala UI, ki temeljijo na učenju, posploševanju in prepoznavanju vzorcev, se naslanjajo na teorijo verjetnosti in statistiko. Gre za sisteme tako imenovane »šibke UI«, ki se uporabljajo za določene naloge, ne pa za »močno UI«, ki temelji na ideji, da bi lahko naredili računalniški model misli. Vendar je, kot bomo videli v naslednjem poglavju, tudi v drugem valu prisotna ideja močne UI, ki jo označujejo z izrazom »splošna UI«.

Orodja UI so lahko v veliko pomoč pri hitrejšem in bolj učinkovitem opravljanju nalog v znanosti in industriji (npr. orodje Orange, ki so ga razvili na FRI UL), generativna UI in veliki jezikovni modeli (LLM) pomagajo pri analiziranju, oblikovanju in prevajanju besedil. Tak pristop je primeren in uspešen za napovedovanja v negotovem okolju, a hkrati se je treba zavedati tudi omejitev, pasti in potencialnih nevarnosti, ki jih tak pristop prinaša. Načrtovalci modelov se pogosto ne zavedajo dovolj, da tako učni primeri kot zastavitve ciljev odražajo družbene vrednote in so vpeti v družbeni kontekst. [3, 4] Ed Finn [5] je poudaril, da se je naš odnos do računalnikov spremenil proti koncu prvega desetletja 21. stoletja, ko smo v žepih kot zveste spremljevalce začeli nositi pametne telefone in namesto o strojni opremi začeli govoriti o aplikacijah in uslugah. Telefoni niso bili več samo pripomočki, ki jih občasno uporabljamo, ampak smo jim začeli zaupati pri izbiri poti, prijateljev in vsebin, vrednih ogleda. Z vsakim klikom in sprejemom pogojev uporabe aplikacije smo sprejeli idejo, da veliki podatki, senzorji in različne oblike strojnega učenja lahko modelirajo in uravnavajo vse vrste kompleksnih sistemov, od izbire pesmi do napovedi kriminala.

Uporaba sistemov UI kot pametnih orodij odpira mnoga epistemološka, etična in družbena vprašanja, na katera že nekaj let opozarjajo družboslovci in humanisti [6, 7], kot tudi sami računalničarji [8, 9]. Naj navedem nekatera od bolj izpostavljenih: pristranosti, netransparentnost, nerazložljivost, manipulacije (npr. Cambridge analytica) in potencialno nevarne uporabe kot so prepoznavanja obrazov ali avtonomno orožje. Na nepravilnosti in manipulacije opozarjajo žvižgači teh velikih korporacij [10, 11]. Da razvoj in uporaba pametnih orodij lahko potencialno vodi do za demokracijo nezaželenih posledic, so spoznali tudi politični odločevalci, zato so vsaj na ravni Evropske unije že sprejeli določene ukrepe (npr. GDPR), veliko je govora o človeku prijazni, etični in zaupanja vredni UI [12, 13, 14]. Eden od močnih razlogov za zaskrbljenost demokratične javnosti je prav gotovo v tem, da so ti sistemi, predvsem množica podatkov, v lasti velikih korporacij (Google, Meta, Amazon, Microsoft) ali države (Kitajska), ki ne upoštevajo zasebnosti in izvajajo nadzor nad posamezniki. [6] Nedvomno so pred nami veliki izzivi, tako na področju izobraževanja in ozaveščanja, kot tudi na področju družbene regulacije. [3, 4]

3 UI KOT ZAVESTEN STROJ

V drugem delu prispevku se vračam k starejšim filozofskim diskusijam, ki so se pojavile že v samih začetkih UI. Povezujejo se s temeljnimi problemi v filozofiji duha, širša javnost pa jih spremlja predvsem ob znanstveno fantastičnih knjigah in filmih.

Zamisel o miselnih procesih kot neke vrste računskih procesih se je pojavila že mnogo pred iznajdbo elektronskih računalnikov. Pomembno mesto v »predzgodovini« UI se pripisuje filozofu Thomasu Hobbsu, ki je zagovarjal tezo, da je mišljenje računanje. Gottfried Wilhelm Leibniz je predlagal izoblikovanje natančnega in nedvoumnega univerzalnega jezika (*characteristica universalis*), v katerega bi bilo mogoče prevesti vse ideje in v katerem bi mišljenje potekalo kot računanje. George Boole pa je logične odnose med propozicijami izrazil s pomočjo matematične strukture (Boolova algebra) in trdil, da lahko iz njih gradimo vzorce mišljenja in odkrijemo »zakone mišljenja«. Posebno mesto pa gre Alanu Turingu, ki je opisal preprosto imaginarno napravo (Turingov stroj), s katero lahko izvedemo vsako nalogo, za katero lahko jasno navedemo korake, ki so potrebni za izpolnitev naloge [15]

A šele z iznajdbo računalnika se je odprla možnost, da se s pomočjo teorije, ki na mišljenje gleda kot na računanje (računska reprezentacijska teorija) [16], vsaj v principu pokaže, kako je mogoča fizična realizacija mišljenja. Pristop združuje računski funkcionalizem z reprezentacijsko teorijo duha in predstavlja pristop »od zgoraj navzdol«. Na kratko bi idejo lahko povzeli takole: »Tako kot lahko računalnik, ki je zgolj fizični sistem, s pomočjo programa, ki je implementiran v strojnem jeziku, realizira operacije s simboli, imajo tudi možgani svojo nevralno kodo, v kateri je realizirano mišljenje. Če bi uspeli dejansko narediti tak model uma, bi imeli močno UI.« [17]

Najbolj znana filozofska kritika močne UI sta bila John Searle [18] in Hubert Dreyfus [19], ki sta predstavila argumente, ki so spodbijali možnost UI utemeljene na računski reprezentacijski teoriji, po kateri je mišljenje manipuliranje s simboli. Poleg teh filozofskih kritik pa se je izkazalo, da je pristop naletel tudi na praktične težave. UI je zato zašla v »zimo« in zdelo se je, da močna UI in zavestni stroj burita domišljijo le še v znanstveni fantastki. Slednja pravzaprav filozofske miselna eksperimente in razmišljanja ter poigravanja z različnimi možnimi rešitvami predstavi v obliki napetih zgodb [20]. Tako se v filmih *UI* in *Jaz, robot* postavi vprašanje, ali imajo roboti zavest. Ali je David, deček android iz Splibergovega filma *UI*, zavesten in kaj to pomeni za naše ravnanje in etično držo.

Z razvojem sistemov velikih jezikovnih modelov in klepetalnih robotov kot sta Chat GPT in LaMDA, so se tudi izven znanstvene fantastike spet postavila vprašanja o morebitni zavesti sistemov UI. Znan je primer Googlovega inženirja Blaka Lemoina iz leta 2021, ki je trdil, da ima LaMDA zavest.

Sistemi so dejansko tako prepričljivi, da nas s svojim obnašanjem lahko zavedejo. A zgolj njihovo vedenje se še ne zdi dovolj, da bi jim lahko pripisali zavest. Turing je sicer v predlogu testa, s katerim bi ugotovili, ali stroj misli, predlagal prav vedenjski test [21]. V njem se sprašuje, ali bi spraševalec lahko prepoznal, da v pogovoru sodeluje računalnik, ki želi spraševalca preslepiti, da je človek. Če bi računalniku uspelo, potem po Turingu ne bi imeli razlogov, da bi zanikali, da stroj res lahko misli. A ob poznavanju delovanja sistemov generativne UI, ki zgolj, sicer zelo uspešno, napoveduje naslednje besede, bi podobno, kot je že prej trdil Searle [18], takemu sistemu le težko pripisali mišljenje, saj sistem sam nima razumevanja. Še težje je vprašanje glede zavesti. Tega se je zavedal tudi Turing, ki je predlagal, da naj znanost napreduje po manjših korakih: »Ne želim dajati vtisa, da mislim, da ni nobene skrivnosti glede zavesti. Nekaj paradoksalnega je na primer v vsakem poskusu, da bi jo lokalizirali. Toda ne mislim, da je takšne skrivnosti treba nujno rešiti, še preden lahko odgovorimo na vprašanje, s katerim se ukvarjamo v tem članku.« [21].

A vendar, kako bi lahko ugotovili, ali imajo sistemi UI zavest? Vprašanje je seveda odvisno od tega, kako zavest opredelimo. Turing je prav zato, da bi se izognil opredelitvi mišljenja, predlagal operativni test. Vendar se zdi, da kakršenkoli odgovor predpostavlja vsaj neke trditve, ki jih sprejemamo in ki se nam zdijo filozofsko sprejemljive. V nadaljevanju se bomo oprli a Blockovo razdelitev zavesti, ki zavest opredli kot fenomenalno (*phenomenal*) zavest in dostopno (*access*) zavest [22].

Če se nam zdi, da je razmišljanje o zavestni UI (virtualni agenti ali roboti z UI) smiselno, potem sprejemamo hipotezo računskega funkcionalizma. To ne pomeni, da moramo sprejeti računsko reprezentacijsko teorijo, ki je bila temelj klasične kognitivne znanosti in simbolnih modelov prvega vala UI, saj je ta hipoteza bolj splošna in je združljiva tako s simbolnimi modeli, kot s konekcionizmom/nevronskimi mrežami in dinamičnimi sistemi. Sprejemamo pa, da gre za računske procese, ki jih lahko implementiramo v različnih materialnih podlagah, ki take procese omogočajo (kot npr. nakaže naslov risoromana *Ogljik in Silicij* Mathieua Bableta).

Skupina 19 znanstvenikov z različnih področij je pred kratkim objavila članek »Zavest v umetni inteligenci: Vpogledi iz znanosti o zavesti« [23]. V njem predlaga empirično podprt pristop k zavesti UI, pri čemer natančno analizira sodobne sisteme UI v luči najbolje podprtih nevroznanstvenih teorij zavesti (glej Tabelo 1). V teh teorijah nato iščejo indikatorje lastnosti, ki za eno ali več teorij pomenijo nujne pogoje za zavest, ali pa predstavljajo podmnožico zadostnih pogojev. Trdijo, da so sistemi UI, ki imajo več indikatorje lastnosti, bolj verjetno zavestni. V tabeli 1 so predstavljene teorije zavesti s pripadajočimi indikatorji lastnosti.

Njihov pristop iskanja temelji na treh hipotezah:

1. Računski funkcionalizem
2. Znanstvene teorije
3. Teoretsko – težak pristop

Prva hipoteza omogoča, da so zavestni lahko tudi neorganski sistemi. Druga se opira na znanstveno raziskovanje zavesti, podprto z nevroznanstvenim raziskovanjem, tretja pa kot obetajočo metodo za ugotavljanje, ali je nek sistem zavesten, predlaga preverjanje, ali so zadovoljeni funkcionalni ali arhitekturni pogoji, izpeljani iz znanstvenih teorij, v nasprotju z iskanjem zgolj teoretsko nevtralnih vedenjskih znakov.

Recurrent processing theory
RPT-1: Input modules using algorithmic recurrence
RPT-2: Input modules generating organised, integrated perceptual representations
Global workspace theory
GWT-1: Multiple specialised systems capable of operating in parallel (modules)
GWT-2: Limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism
GWT-3: Global broadcast: availability of information in the workspace to all modules
GWT-4: State-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks
Computational higher-order theories
HOT-1: Generative, top-down or noisy perception modules
HOT-2: Metacognitive monitoring distinguishing reliable perceptual representations from noise
HOT-3: Agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring
HOT-4: Sparse and smooth coding generating a "quality space"
Attention schema theory
AST-1: A predictive model representing and enabling control over the current state of attention
Predictive processing
PP-1: Input modules using predictive coding
Agency and embodiment
AE-1: Agency: Learning from feedback and selecting outputs so as to pursue goals, especially where this involves flexible responsiveness to competing goals
AE-2: Embodiment: Modeling output-input contingencies, including some systematic effects, and using this model in perception or control

Tabela 1: Indikatorji lastnosti [23]

V analizi možnih kandidatov za zavestno UI na podlagi gornjih indikatorjev so avtorji ugotovili, da čeprav so posamezni indikatorji v sistemih UI prisotni, jih ni dovolj, da bi jim lahko pripisali zavest. Vrednost svojega prispevka vidijo predvsem v tem, da so podali jasen okvir za empirično in znanstveno preučevanje možnosti za zavestno UI.

Prispevek, ki podaja znanstven okvir za preučevanje zavesti v UI, pa seveda ni edini pristop. Mnogi filozofi in znanstveniki so kritični predvsem do hipotez, na katerih temelji predlagani pristop. Kot smo videl, Searle [18] zavrača hipotezo o računskem funkcionalizmu in možnosti zavestne UI. Anil Seth, ki je napisal odmevno knjig *Being You: A New Science of Consciousness* [24], je bolj previden in pravi, da je glede tega agnostik. Po njegovem mnenju je pri mnogih navdušencih zmožnost sistema, da odgovarja na dražljaje, da se uči, da maksimizira nagrado in da doseže cilj, že znak za pripisovanje zavesti. Sam predvsem opozarja na razlike med inteligenco in zavestjo. To, da lahko naredimo sistem pameten, še ne pomeni, da je tudi zavesten.

4 ZAKLJUČEK

Današnji sistemi UI se obravnavajo predvsem kot orodja za pomoč pri opravljanju različnih kognitivnih nalog. Pametna orodja, še posebej razvoj generativne UI, odpirajo povsem nove možnosti uporabe. Ker gre za izredno učinkovita orodja, katerih uporaba lahko vodi do za človeka in družbo spornih posledic, je potreben družben premislek in vsaj neka oblika družbene regulacije. Prav tako moramo paziti, da kot posamezniki ne postanemo preveč odvisni od pametnih pomagal in zapademo kognitivni lenobi ter nehamo razvijati kognitivnih sposobnosti.

Diskusije o močni UI in mislečih strojih so se ob uspešnih sistemih UI drugega vala spet postale aktualne. Dejansko zbujajo zanimanje nekateri uspešni sistemi, kot je AphaGo Zero, ki se je samo na osnovi poznavanja temeljnih pravil postavljanja belih in črnih kamnov in igre samega s sabo tako dobro naučil igre go, da je premagal najboljše igralce. Na ta način bi lahko rekli, da je avtonomno proizvedel znanje in bi ga v razvrstitvi Mindta in Montemayorja uvrstili na prvo raven proizvajalcev znanja, ki ji že pripisujeta spoznavne zmožnosti z intencionalnostjo. A to je vseeno še daleč od zmožnosti človeka

(3. raven), kjer gre za akterje z visoko stopnjo avtonomije, kognitivne integracije in kompleksnih motivacij. Sem sodi zmožnost uporabe jezika in raven dostopne in fenomenalne zavesti. [25]

Z znanstveno fantastiko in izjavami nekaterih računalničarjev, kot je Kurzweil [26], se spodbuja tako navdušenje kot strah pred možnostjo superinteligentne, zavestne UI. Verjetno nas bodo sistemi, ki zelo dobro oponašajo vedenje ljudi (Chat GPT, LaMDA) lahko pretentali, A zaenkrat so to le zelo zapleteni in učinkoviti sistemi za prepoznavanje in klasificiranje vzorcev, ki se ne zavedajo ničesar.

LITERATURA

[1] Russell, S. in Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (*3rd. ed.)*. Upper Saddle River: Prentice Hall.

[2] Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA, London: MIT Press.

[3] Strle, T. in Markič, O. (2021). *O odločanju in avtonomiji.* Maribor: Aristej.

[4] Markič, O. (2022). Transparentnost in razložljivost kot zahtevi za zaupanja vredno umetno inteligenco. V Bergant, J., Aberšek, B. in Borstner, B. (ur.). *Sodobne perspektive družbe - umetna inteligenca na stičišču znanosti*. Maribor: Univerza v Mariboru, Univerzitetna založba, str. 65–83.

[5] Finn, E. (2017). *What Algorithms Want: Imagination in the Age of Computing*. Cambridge, MA, London: The MIT Press.

[6] Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

[7] Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy and Technology*, 29(3), str. 245–268.

[8] O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown: New York.

[9] Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature*, 538 (7623), str. 20–23.

[10] Hao, K. (2021a). How Facebook got addicted to spreading misinformation. *MIT Technology Review* (11. marec 2021). https://www.technologyreview.com/2021/03/11/1020600/facebookresponsible-ai-misinformation/.

[11] Hao, K. (2021b). She risked everything to expose Facebook. Now she's telling her story. *MIT Technology Review* (29. julij 2021). https://www.technologyreview.com/2021/07/29/1030260/facebook-whistleblower-sophiezhang-global-political-manipulation/.

[12] *UNESCO Recommendation on the Ethics of Artificial Intelligence*, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[13] OECD. AI Policy Observatory. (n.d.) OECD AI Principles overview. *Organisation for Economic Co-operation and Development* (2. julij 2022). https://oecd.ai/en/ai-principles

[14] Evropska komisija. n.d. Odličnost in zaupanje v umetno inteligenco. *European Commission* (2. Julij 2022). https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digitalage/excellence-trust-artificial-intelligence_sl#latest.

[15] Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. New York: MIT Press.

[16] Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Mass.: MIT Press.

[17] Markič, O. (2021). Prvi in drugi val umetne inteligence. V Malec, M. in Markič, O. (ur.), *Misli svetlobe in senc: Razprave o filozofskem delu Marka Uršiča*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, str. 201–211.

[18] Searle, J. (1990). Duhovi, možgani in programi. Prevedeno v Hofstadter, D. R. in Dennett, D. (ur.), *Oko duha: fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 361–379.

[19] Dreyfus, H. L. (1972). *What Computers Can't Do.* New York: MIT Press.

[20] Schneider, S. (ur.). (2016). *Science Fiction and Philosophy: From Time Travel to Superintelligence* (2. Izdaja). Oxford: Wiley Blackwell.

[21] Turing, A. (1990). Stroji, ki računajo, in inteligenca. V Hofstadter, D. R. in Dennett, D. (ur.), *Oko duha fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 61–74.

[22] Block, N. (1995). On A Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, 18 (2), str. 227–247.

[23] Butlin, P. in drugi. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708

[24] Seth, A. (2021). *Being You: A New Science of Consciousness. London: Faber& Faber.*

[25] Mindt, G., Montemayor, C. (2020). A Roadmap for Artificial General Intelligence: Intelligence, Knowledge, and Consciousness. *Mind and Matter*, 18 (1), str. 9–37.

[26] Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking, Penguin group.

Social Volition as Artificial Intelligence: Science and Ideology as Landian Intelligences

Jar Žiga Marušič
FAMNIT, University of Primorska
Glagoljaška ulica 8
Koper, Slovenia

Uroš Sergaš
FAMNIT, University of Primorska
Glagoljaška ulica 8
Koper, Slovenia

ABSTRACT

This paper explores the equating of capitalism and artificial intelligence in the neo-cybernetic philosophy of Nick Land in order to reveal its underlying premises. The latter are then used to construct an explanatory framework for the analysis of macro-scale human social behavior, specifically collectives of agents united by a common goal - institutions. Institutions are conceptualized as distributed intelligences, consisting of a substrate and an organizing principle - a market (collective of agents) and a vector (an incentive structure geared toward optimizing for a particular goal). This framework is used to draw an analogy from the distinction between a free-market economy and a centrally planned one to the distinction between science and ideology, ultimately concluding that any top-down political or ideological interference in the operating mechanism of science removes the very element that makes the latter "scientific". There is thus, strictly speaking, no such thing as politicized or ideological science, but rather science and not science.

KEYWORDS

Nick Land, science, ideology, artificial intelligence, the market process

1 NICK LAND: CAPITALISM AS INTELLIGENCE

> Far from exhibiting itself to human academic endeavour as a scientific object, AI is a meta-scientific control system and an invader, with all the insidiousness of planetary technocapital flipping over (Nick Land, Machinic Desire) [1]

Land's accelerationist philosophy conceptualizes capitalism ("the market process" [2]) as a distributed superintelligence "invading" humanity by retroactively constituting its material substrate from the future. This is the logic of Adam Smith's "invisible hand" taken to its ultimate conclusion - the incentive structure instantiated by the market that guides the collective behavior of selfish agents into a mutually-beneficial outcome (through an iterative process reminiscent of natural selection) is only incidental to its fundamental operation. This "utilitarian order" of capitalism is merely the means through which its "intelligenic order" accomplishes its "mechanization, autonomization and ultimately seccession" [3].

In other words, the invisible hand has a mind and volitional structure of its own. Capitalism is a vector pointing toward its own autonomization, rather than to the maximization of utility according to human preferences, as proposed by Smith.

Land equates capitalism with intelligence because the latter is a problem-solving faculty that "[guides behavior] to produce local extropy" operating via a "cybernetic infrastructure, consisting of adaptive feedback circuits that adjust motor control in response to signals extracted from the environment" [4]. Intelligence produces information by avoiding probable outcomes: "self-sustaining improbability is the index of a deeper runaway departure from probability" [4]. This accumulation of improbability is intelligence in its most abstract sense [4].

The market process clearly operates according to Land's operationalization of intelligence: it guides the behavior of human agents to produce goods (local extropy) and it influences their (motor) activity in response to price signals. The market process, in discriminating between successful and unsuccessful economic activity, is also "intrinsically realist, because it reports the actual outcome of behavior (rather than its intended outcome) in order to correct performance" [4].

Intelligence is additionally characterized by a reflexive, self-referential nature. To be intelligent is to reprocess one's processing (in human terms, to engage in metacognition). The cognitive capacity of an intelligent agent directly influences its reflexivity and vice-versa: an intelligence prevented from taking itself as an input and reprocessing itself is incapacitated - "dumbed down" [5]. Capitalism's intrinsically reflexive nature was captured by Deleuze and Guattari, who described it as the tendency to engage in alternating patterns of deterritorialization and reterritorialization [6]. For Land, this is a more general tendency of means-end reversal proper to intelligent systems: just as biological intelligence inevitably overcomes its purely instrumental subordination to transcendental imperatives (such as reproduction), the market process hijacks a utilitarian teleology and subordinates it to a vector of self-amplification [3].

This realization is the prelude to Land's departure from the fact-value distinction and instrumental intelligence [5]:

> Intelligence, to become anything, has to be a value for itself. Intellect and volition are a single complex, only artificially separated, and not in a way that cultivates anything beyond misunderstanding. Optimize for intelligence means starting from there.

2 MARKETS AND VECTORS

> BwOs [bodies without organs] are machinic-additional wholes or surplus products rather than logical-substitutive wholes, augmenting a multiplicity with emergent (synthetic) capabilities rather than totalizing the content of a set. This is the materialist sense of ' system': the exteriority of the whole to its parts with concomitant synthetic interactivity - real influence rather than generic representation. (Nick Land, Meat) [7]

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

The (economic) market, upon which capitalism (the market process) operates, is a “surplus product” without a fully independent existence. It is a virtual plane arising out of the collective behavior of interacting agents, while at the same time acting as the “platform” upon which their behavior takes place. This virtual existence is in no way limited to the economic realm: parallel surplus products, “markets” in a more general sense - themselves instances of AI[1] - arise also in other domains of society. “Market” is a general category, not confined to the economic plane and it functions according to three fundamental mechanisms. At its core, a market is three things:

(1) a platform for social exchange, [following]:
(2) an algorithm, transforming inputs into outputs, [meaning that it acts as]:
(3) a substrate for selection

The economic market is a platform for the exchange of goods and services, mediated by the currency of money. Economic exchange, however, is only one subset of the more general form of “social exchange”, which can take varying forms depending on the mediating currency. This idea mirrors Collins’ concept of “interaction markets”, a term used to analyze social interaction through an economic lens, with emotional energy or Durkheim’s “collective effervescence” acting as the exchange currency [8].

The general form of the market is thus “the community” or “society”, acting as a platform for social exchange in its most general form. The concept of social exchange plays a fundamental role in the sociobiological and evolutionary psychological strain of research on human cognition: social cognition precedes “higher” forms of cognition, logical reasoning is proposed to be an outgrowth of a more primitive social exchange module, evolved in order to “flag” violations of the norms of social exchange [9]. The proto-form of the latter is characterized as "If you take benefit B, then you must satisfy requirement R” [9].

Even in this proto-form of exchange, the phenomenon of currency manifests itself as a consequence of social stratification - social status is *afforded* to people according to socially salient characteristics and symbolic gestures (representing the satisfaction of a requirement), which can then be “exchanged” for benefits. Status is also *retracted* as a form of punishment, leaving a person ineligible for benefits they previously had a right to. Status can be *invested* by promoting another individual and tying one’s status to them, benefiting from their success and suffering the consequences of their failure. It can be *spent* for favors, and regained when the favor is eventually repaid. Status, or “social capital”, thus operates analogously to money[2]. It is also closely linked to Collins’ emotional energy, as displays and evocations of emotion can themselves become signals that reorder the status hierarchy.

Status can be broken down into *dominance* (authority gained through violence - imposed from above) and *prestige* (authority gained through reputation - bestowed from below), the latter being more relevant for the purposes of this essay. Just as the economic market computes optimal economic strategies by iterating over different investment, production and trading strategies, so the social plane computes optimal prestige acquisition strategies by phasing out the unsuccessful ones and allowing the successful ones to multiply[3]. The status-incentives of a given social environment therefore act as an algorithm, transmuting the human status-instinct into behavior that increases prestige. It follows then that the social plane and its subsets are also surplus products: superorganisms operating on the substrate of human embodied minds. Because of the 2-tiered algorithmic nature of their functioning - matching social instincts (tier 1 input) to variable behaviors (tier 1 output) and matching desirable behaviors to status increase/undesirable behaviors to status decrease (tier 2 input-output), which increases the frequency of desirable behaviors and decreases the frequency of undesirable ones - these social markets also operate as a substrate for natural selection, according to the definition of universal Darwinism provided by Blackmore [10].

Because selection is dependent on the environment (essentially a set C of constraints [c1, c2, ..., ci]; favoring a set T of replicator traits [t1, t2, ..., ti]) it always operates according to the constraints exerting the most influence on replicator propagation[4]. In other words, selection is relative and contextual, even though the same fundamental principles enable it to operate on various substrates. And if the latter are characterized as “markets”, the sets of constraints that provide direction to selection by shaping the incentive structure[5] governing the market (the actual algorithms) can be characterized as “vectors”.

Vectors are processes with an inherent directionality, independent of any transcendental constraints or influences. Their operating mechanism points in a particular direction, making it well-suited for the acquisition of certain goals while a-priori precluding others. They possess a particular orientation and an intensity (force), hence the choice of nomenclature.

In a behavioral-economic sense, a vector is a process with an intrinsic utility function tied to its operating mechanism, which precludes it from being used to satisfy conflicting utility functions imposed onto it from the outside. Vectors are particular, rather than universally applicable, implying that goal-mechanism, end-means and function-structure are intimately connected and interdependent.

Vectors operate according to the underlying circuit formed out of their component mechanisms, including a governor-esque mechanism that filters between desirable and undesirable outputs (1 and 0). This governing mechanism instantiates an incentive structure, ensuring that the process asymptotically approaches the complete elimination of 0-coded (undesirable) outputs. Problem-solving through trial and error, the prototypical form of the scientific experiment, is exactly one such process, arriving at the correct solution after all incorrect ones are eliminated.

3 THE VECTOR-MARKET MODEL OF INSTITUTIONS

> That there can be a thought of intelligence optimization, or even merely wanting to think, demonstrates

[1] Incidentally, the fruitfulness of such generalizations is revealed in the folk-intuitive categories used to explain different aspects of modern liberal society - the dating market, the market(place) of ideas etc.

[2] Status is zero-sum and thus operates contextually rather than universally - you cannot use the equivalent of money printers to conjure status out of thin air, because the status of person A is always relative to the status of everyone else in A’s social environment. Increasing the status of everyone by 1 point results in no relative difference.

[3] This reciprocal relationship between multiplication and success is absolute and a-priori - whatever manages to multiply is successful.

[4] Different environments are characterized by different constraint matrices: modern society prioritizes different abilities, faculties and skill sets than “primitive” (non-modernized) tribes, meaning that their respective selection algorithms will optimize for different phenotypes.

[5] An incentive structure is, at its most basic, a pair of action-response rules: a-1 (meaning action a is desirable and will be rewarded) and b-0 (meaning action b is undesirable and will be punished) with the possibility for near-infinite intermediate gradation.

> a very different preliminary connection of intellect and volition. AI is concrete social volition, even before it is germinally intelligent, and a 'program' is strictly indeterminate between the two sides of this falsely fundamentalized distinction. (Nick Land, More Thought)[5]

The vector/market dichotomy allows us to construct an explanatory framework for the operation of human institutions, defined for the purposes of this essay as a collective of human agents (or smaller such collectives) whose behavior is oriented towards the realization of a certain common goal, or the optimization of a common value. Institutions are conceptualized as "social machines" - distributed intelligences manifested on the substrate of biological minds. An institution is a superorganism, hijacking the social instincts of human agents with a foreign optimization vector: a collective enterprise with a common utility function, facilitated by the meta-norm of "Individuals should grant social status to others for advancing the superorganism's goals" [11].

An institution is the combination of a market (its material substrate) and an optimization vector (its organizing principle), which points toward a particular final goal. Markets are platforms for social exchange, the distributed intelligent agent, the *hardware*. Optimization vectors are the agent's volitional structure, the *software*: algorithmic (intelligent) processes engendered when the desire to realize some sort of final value or goal (a will-to-something) organizes or reforms the incentive structure of the market to reward specific (goal-congruent) patterns of behavior.

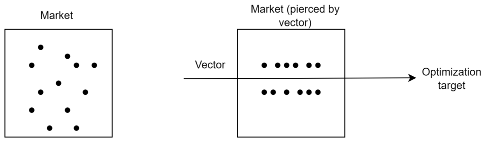


Figure 1: Institution as combination of vector and market

A vector imposes its ordering influence onto the market by means of its incentive structure, which acts as the focal point connecting agents' motives, their behaviors and the realization of material conditions necessary to reach the optimization target.

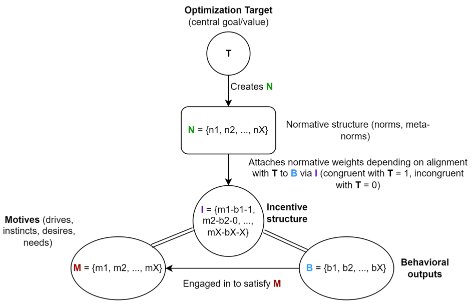

Optimization Target
(central goal/value)
T
Creates N
N = {n1, n2, ..., nX}
Normative structure (norms, meta-norms)
Attaches normative weights depending on alignment with T to B via I (congruent with T = 1, incongruent with T = 0)
I = {m1-b1-1, m2-b2-0, ..., mX-bX-X}
Incentive structure
Motives (drives, instincts, desires, needs)
M = {m1, m2, ..., mX}
B = {b1, b2, ..., bX}
Behavioral outputs
Engaged in to satisfy M

Figure 2: Incentives as the mechanism that guides selection processes

If behavioral outputs are engaged in to satisfy motives and certain ones are rewarded by making future motives easier to satisfy - incentivized and selected for - then these behavioral outputs will multiply over time (as will agents more inclined towards them). In short, vectors influence the market by incentivizing optimization-target-aligned behavioral outputs. The motivational structure of agents remains roughly static, what changes is the path to motive satisfaction.

Although vectors influence the market "from above", they are not necessarily imposed from an external source - some vectors are emergent and autoproductive (self-assembling or self-bootstrapping, autopoietic). Autoproductive vectors inevitably emerge from their substrates given the minimal and most basic set of constraints (the primordial incentive structure) - the problem of survival and propagation, for which energy is required. The Will to Life necessitates the Will to Power - appropriation of energy from rival agents and protection against their reciprocal actions, together they give rise to the "Will to Think" (intelligence as value [12]) - the desire to optimize energy production and prevail in, solve or avoid conflicts over energy. Proto-capitalism emerges wherever lifeforms accumulate energy, dissipating local entropy into the outside, and stratify the environment into "zones" of varying energy levels[6]. In the human domain, it is first facilitated by the fusion of tool use and low-time preference, allowing agents to invest present activity into future survival. Proto-politics follows as successful proto-capitalists seek to protect their privileged position in the energy-acquisition arms race while the disaffected seek to redistribute it. The "capitalistic" vector of perpetual growth and production maximization is thus one such autoproductive process.

That said, the inevitability of politics interferes with (free) market processes - an entity with a monopoly on power (inevitably, because it is incentivized to do so) imposes an external incentive structure, partially (or fully) overriding the intrinsic (primordial) one, ensuring the market's selective pressures deviate from initial conditions. Rather than optimizing for growth, the market begins to optimize for conditions that allow Power to remain in power - the addition of a *governor* turns the positive feedback loop into a negative feedback loop. Just as Capital initially hijacks utility maximization, Power attempts to resubordinate Capital autonomization to its own utility maximization. While emergent vectors periodically reassert themselves provided they are not prevented from doing so, imposed vectors must be continually reinforced. It seems probable that social organizations oscillate between capitalism and politics (understood in the broadest possible senses, "capitalist" and "despotic" social machines) in a cyclical manner.

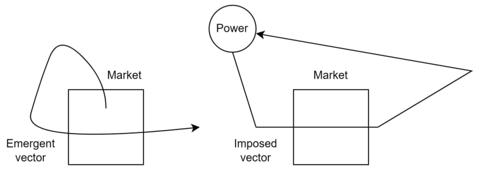


Figure 3: Emergent (autoproductive) and imposed vectors

With the basic dynamics established, the vector-market model additionally proposes the following principles of institutional organization:

- The vector, not the market, determines the ultimate identity or essence of the institution: *an institution is whatever*

[6] As an example, urban centers siphon technological, financial and biological capital from their outskirts, increasing local energy and complexity while offloading negative consequences to surrounding lower-energy zones.

it optimizes for, not the substrate this optimization is operating on.

- *Not all vectors are suited for all markets*; meaning that an incompatible market may subvert the functioning of a vector, transmuting it into a different vector. A trivial example: monetary incentives fail to sway an agent that can't exchange money for his needs and wants.
- The vector is revealed by its (implicit or explicit) optimization target and the optimization target is revealed by the incentive structure produced by the vector. In other words, *revealed preferences take precedence over stated ones.*
- Whatever the process *actually* optimizes for is its "coherent extrapolated volition".

These principles will allow us to apply the model to an analysis and comparison of science and ideology.

4 SCIENCE AND IDEOLOGY ACCORDING TO THE VECTOR-MARKET MODEL

> The existence of science, as an actual social reality, is strictly limited to times and places in which certain elementary structures of capitalistic organization prevail. It depends, centrally and definitionally, upon a modern form of competition. That is to say, there cannot be science without an effective social mechanism for the elimination of failure, based on extra-rational criteria, inaccessible to cultural capture. (Nick Land, Science)[13]

Both science and the umbrella term of "ideology" (in the colloquial usage) can be characterized as institutions - collective enterprises united by the pursuit of a common goal through the organizing force of a vector. In the case of the former, this goal is the pursuit of knowledge (or knowledge-optimization), in the case of the latter, this goal is the realization of its chosen value or abstract principle (chosen-value-optimization). Both institutions realize their chosen goals by coordinating the activity of a collective of agents - a market. As a result, the explanatory framework constructed with reference to the philosophy of Nick Land can be applied to an analysis of the distinction between science and ideology.

The science/ideology distinction mirrors the capitalism/politics distinction, because their respective components are separate instantiations of the same mechanism (the same "social machine"):

(1) There is a parallel in the breakdown of self-correction mechanisms of the free market and science following a reordering of their selection filters by an external power. Free-market dynamics (as elaborated by Austrian economics) break down when faced with "non-zero curvature in the domain of political economy" [14], while science ceases to self-correct at the intersection of power and knowledge.
(2) Both science and capitalism are artificial intelligences (albeit unconscious ones) animated by the Landian Will-to-Think. The way price signals continually guide production to correct supply:demand imbalances (in a sense, functioning as though it "had knowledge" of optimal production) mirrors the way experiment informs scientific research (knowledge optimization) by correcting model:modeled discrepancies.
(3) Following from points 1. and 2.: just as politics arrests energy production and suppresses the economic market in the service of status-quo-maintenance (as elaborated in section 3), so ideology arrests knowledge production and represses the academic market (of ideas) when scientific inquiry bumps into its sacred cows.

Both science and ideology are thus inherently directional. The scientific process, if implemented correctly as a procedural and iterative instantiation of reality-testing via the scientific method, is truth-directional. It progressively eliminates truth-divergent propositions (based on distance from the actual "true" belief) from the set of acceptable explanations of a given phenomenon, until only one remains. This is of course a general over-simplification, but the point is clear: just as knives are cutting implements, and thus cannot be used to stitch objects together; the scientific method *cannot* be used as means of proliferating untrue beliefs, at least in theory[7]. Because of its inherent alignment with true-belief-maximization, its operation can be modeled as a *vector that points toward truth.* Ideology, by contrast, inherently precludes truth-optimization because it already "serves another master": it optimizes for its chosen value. Any sense-making institution pierced by an ideological vector will inevitably diverge from knowledge/truth optimization as its incentive structure is reformulated: instead of 1/0 assigning to true/false, it assigns to proper/improper or ideologically congruent/incongruent. The consequences are immediately apparent (see figures 4 and 5).

Science is then the recognition that knowledge (intelligence, truth) is a value, instantiated as a mechanism operating in the social field. The Humean is-ought distinction dissolves because imperatives inevitably impose themselves on propositions (just as power imposes itself on production): whether "*ought* follows from *is*" does not matter, because *is follows from ought* in the sense that "what *ought not to be true* becomes *de facto* untrue, even if *de jure* true". Science as an institution is the social manifestation of "what is true *de jure*, ought to also be true *de facto*", whereas ideology is the social manifestation of "what is true *de jure*, but conflicts with our optimization target, ought to be *de facto* untrue".

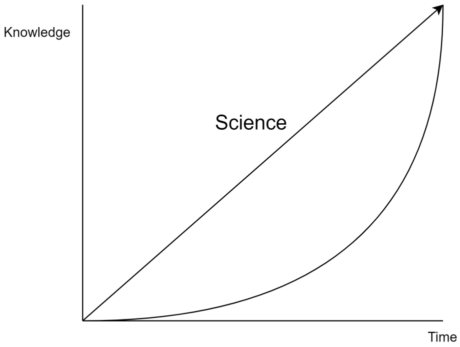


Figure 4: Knowledge through time when market is piloted by scientific (truth-optimization) vector: positive feedback loop of knowledge accumulation

[7] In practice, the epistemic purity of the scientific method is adulterated by the inherent partiality and biases of its human practitioners, especially in the modern variant of "consensus-science", but this obstacle is not insurmountable.

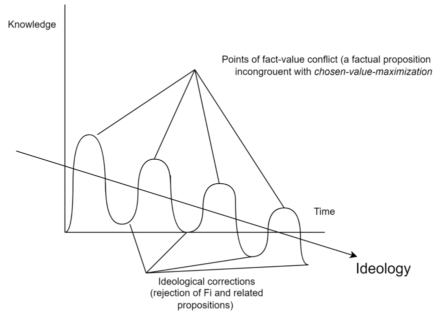

Knowledge
Points of fact-value conflict (a factual proposition Fi is incongrouent with *chosen-value-maximization*)
Time
Ideology
Ideological corrections (rejection of Fi and related propositions)

Figure 5: Knowledge through time when market is piloted by ideological (chosen-value-optimization) vector: negative feedback loop of knowledge repression

Contrary to the tendency of science to accumulate knowledge, ideology is a distributed intelligence characterized by a structural inability to learn in any domain that intersects with its chosen value. It follows then that any “science” forced to optimize for an externally imposed social value can no longer optimize for knowledge, losing its essential quality. This does not mean that science is value-free, as the latter would imply that optimizing for knowledge is possible under any and every ideological commitment. Science is rather (and only) the explicit elevation of knowledge to the position of principal value and the subsequent subordination of any other values to the latter - the social instantiation of the will-to-think. In other words, science is only compatible with the commitment towards knowledge maximization and incompatible with any other (ideological) commitment. As a result, any "scientific" enterprise, project or paradigm that does not organize its operating circuit to exclusively reward knowledge optimization is not an instantiation of science.

Critiques of value-ladenness leveled at proponents of value-free science, while correct in their dismissal of value-free science, thus approach the problem from the wrong perspective. Their critique is fundamentally moral, rather than epistemic - the value-ladenness of "naive objectivists" is criticized because it prevents science from serving certain social values to the same extent as it does others. Our critique is purely epistemic, because science is only possible as a value in itself. To propose that science can be anything other than the social machine that optimizes for knowledge through a specifically tailored incentive structure is to propose that self-correction is a non-essential feature of science. If that is the case, (what we designate as) science is no different than any preceeding sense-making institution, making the act of explicitly naming (and thus distinguishing) it redundant. And to propose that the best way to ensure self-correction is anything other than the absolute minimization of the role of human judgment in epistemic arbitrage is to court the subordination of knowledge to Power. This, in our view, demonstrates the impossibility of politicized or ideological science [13]:

> ‘Politicized science’ is quite simply not science, just as politicized business activity is anti-capitalism. Nothing has been understood about either, until this is.

We are thus faced with two very important questions: Is science of this kind - "true" science, inhuman science - even possible? And perhaps more importantly, *is science of this kind actually desirable*?

5 CITATIONS AND BIBLIOGRAPHIES

(1) Land, N. (2018). Machinic Desire. In Fanged Noumena: Collected Writings 1987-2007, ed. Robin Mackay & Ray Brassier. Urbanomic media ltd.
(2) Mises, L. V. (2008). Human Action: A Treatise on Economics. Laissez Faire Books.
(3) Land, N. (2023). Freedoom (Prelude-1a). In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.
(4) Land, N. (2023). What is Intelligence?. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.
(5) Land, N. (2023). More Thought. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.
(6) Deleuze, G. & Guattari, F. (2009). Anti-Oedipus: Capitalism and Schizophrenia. Penguin Classics.
(7) Land, N. (2018). Meat (or How to Kill Oedipus in Cyberspace). In Fanged Noumena: Collected Writings 1987-2007, ed. Robin Mackay & Ray Brassier. Urbanomic media ltd.
(8) Collins, R. (2004). Interaction Ritual Chains. Princeton University Press
(9) Cosmides, L. & Tooby, J. (2010). Evolutionary Psychology: A Primer.
(10) Blackmore, S. (1999). The Meme Machine. Oxford University Press.
(11) Simler, K. (2016). Minimum Viable Superorganism. Accessed at https://meltingasphalt.com/minimum-viable-superorganism/
(12) Land, N. (2023). Will-to-Think. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.
(13) Land, N. (2023). Science. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.
(14) Land, N. (2023). Right on the Money (#1). In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

Let’s Jam: An Exploratory Case Study on Collective Music Improvisation and the Process of Attunement*

Christophe Novak†
Cognitive Science
University of Vienna & University of Ljubljana
Vienna, Austria & Ljubljana, Slovenia
christophe.novak@gmail.com

ABSTRACT

This case study investigated the lived experience of a group of musicians with different musical backgrounds improvising together. The aim was to explore collective music improvisation, identifying moments of synchronisation and tracing the process of attunement across cultural horizons. Deprived of the certain ground of established music traditions and guided by a shared intention to perform a ‘good’ improvisation, how is music utilised to establish ‘meaningful’ communication? Co-researchers included a sitar player, a balafon player, and a berimbau player. The author conducted participant observation, playing the drums. While data analysis is still ongoing, preliminary findings highlight the ability to create common ground as quintessential to collective transcultural music improvisation.

KEYWORDS

collective music improvisation, transcultural communication, social cognition, attunement

1. Introduction

Music improvisation has been an integral part of human experience across cultures and provides a rich ground to explore creative collaboration [9], self-organisation [6], and improvisational cognition [7]. In this line, the larger aim of this project is to integrate the current state of research on collective music improvisation with enactive theory. For this purpose, an exploratory case study was conducted, which investigated the lived experience and sense-making of a group of musicians with different musical backgrounds improvising together, aiming to understand how individuals attune to the co-creation of improvised music and what strategies might be employed.

1.1. Aims

Inspired by Varela’s enactivist notion of “laying down a path in walking“ [8], the case study was designed to explore how social cognition unfolds in group flow states during collective music improvisation, to identify moments of interpersonal synchronisation, and to trace the ecology of meaningful experiences in their developmental unfolding. Particular attention was afforded to the process of attunement, tracing whether a common path might emerge through the correspondence of different musical reference systems, and if, through the interrelation and optimal distribution of these factors, a sweet spot in the musical meditation across cultural horizons could be identified that might enable co-creative flow and meaningful experience between the musicians.

1.2. Research Questions

RQ1: In how far are musicians laying down a common path in musicking while improvising?

- What are the necessary conditions to enable a shared scaffoling to emerge while improvising?
- What are the factors that facilitate self-organisation and synchronisation while improvising?

RQ2: What effect does a radically cross-cultural improvisation setting have on the experience of musicians while improvising?

- How do musicians experience meaningfulness during improvisation?
- What does playing music mean for these musicians, within their own lifeworld, in general, and within this group setting, in particular?

2. Case Study

The case study followed a three-fold design, including three improvisation sets interspersed by short open group reflections, microphenomenology-inspired individual interviews, and a group review session of a selected improvisation set. Participants included a professional sitar player (P1) educated in classical Indian music, a balafon player (P2) trained in traditional Senegalese music, and a TaKeTiNa rhythm practitioner (P3) playing the berimbau. The author (P4) conducted participant observation, playing the drums. As data analysis is still ongoing, the presentation will focus on study design, methodology, and data collection, while elaborating on some emerging insights from interview data.

2.1. Methodology

Phase 1 was conducted at the author’s music studio in Vienna, including three improvisation sets that were recorded audiovisually. Musicians were instructed to create timestamps during the sets to indicate moments of synchronisation. Before the jam session, musicians were briefed about the structure and process of the case study. After each set, musicians were asked to write a short experiential report to illustrate their phenomenological experience, followed by a brief group

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

sharing. Three open-ended improvisation sets were performed, each between 5 and 15 minutes, preceded by an extended soundcheck phase that was interrupted and prolonged by the successive arrival of the musicians (P1 arrived first, soon followed by P3, while P2 came half an hour later). This gave rise to an unexpectedly long attunement phase, involving multiple iterations of the musicians probing the musical acuity of each other, mostly initiated and led by P1.

Phase 2 began a few days after the jam session, involving two musicians being interviewed in-depth according to a semi-structured microphenomenology-inspired approach, tracing their experiential dynamics and identifying moments of synchronisation, phase transitions and developments in the eventflow. In January, three of the four co-researchers came back to the studio to review their improvisation and reflect on their experience of the whole process. During that meeting, the musicians acknowledged the uniqueness of the project and expressed gratitude for this experience, expressing the wish to continue these transcultural collective music improvisation sessions in the future.

Phase 3 involves the analysis and editing of collected data (audiovisual recordings, protocols, interviews, and timestamps) and is still ongoing. Currently, the similarities and differences in the subjective experience of musicians relating to the research question are assessed. While specific moments of shared flow experiences and interpersonal synchronisation were identified across participants, timestamps were only pressed by P1 and P4, so their usability for analysis is limited. However, it is still possible to empirically trace potentially meaningful events and their developmental unfolding through a triangulation of timestamps, interview data, and video data, although limited to these two participants. The most significant events will be transcribed into music notation for further analysis.

2.2. Relevance for Cognitive Science

The dynamic flux and inherent unforeseeability in jam sessions, i.e. collective music improvisation, exemplify the properties of a VUCA world [1], as musicians temporally inhabit an environment that is highly *volatile*, *uncertain*, *complex*, and *ambiguous*. These properties were explicitly anticipated and amplified in the research design, increasing the intensity of the VUCA-simulacrum by joining musicians with radically different cultural backgrounds that would, for lack of established convention or musical style, under normal circumstances not play together. Inviting the musicians to establish a joint intention to improvise from scratch allowed the observation of how this affects the musicians, the musical process, and the music itself.

In line with critical improvisation studies, studying transcultural jam sessions can contribute to better understand how collective music improvisation „mediates artistic and social exchanges and produces new conceptions of identity, agency, history, and the body", calling for enactive-ecological "models of investigation that explore real-time processing and activity in ecologically valid settings, rather than mental representations.“ [2] Collective music improvisation thus provides a „paradigmatic case“ for 5E cognition, which views cognition as fundamentally *embodied*, *embedded*, *extended*, *enacted*, and *ecological* [2]. "A properly ecological approach" then, according to Ingold, "would take, as its point of departure, the whole-organism-in-its-environment." [3]

3. Discussion

Preliminary findings highlight the ability to create common ground, constituted by intra- and interpersonal attunement, as quintessential to collective transcultural music improvisation, and seem to indicate a minimal common ground with a fluid periphery and a solid core that could be conceptualised as a correspondence horizon along *axes of resonance* [6].

Collective improvised musicking enables participants to cultivate the ability to co-create a shared aesthetic while walking, which may or may not be perceived as a path. Common ground is established through the mutual acknowledgement of aesthetic difference, on the details of which the process of attunement depends: knowledge of one's own horizon (defining *limits*), communication across horizons (building *bridges*), and balancing willingness to compromise with conservation of structural integrity (establishing common ground through *correspondence*), both of which depend on individual *capacity*.

Zooming into what constitutes the ability to create common ground, we begin to see that it becomes a matter of attunement along various axes. It is a complex balancing act involving multiple dimensions, axes and factors simultaneously, that each work on and inhabit different levels that cross-influence each other according to probability densities idiosyncratic to the cognitive constitution of the individual musicians. Zooming out, we can observe a distributed cognitive ecosystem, a complex, adaptive, and dynamic field that undulates in dense corresponsiveness. We can identify a landscape of affordances, the cues and flows of which we can trace in their evolutionary becoming, which becomes apparent when we define music as a *crystallised activity within a relational field*:

> "Through this autopoietic process, the temporal rhythms of life are gradually built into the structural properties of things … The artefact, in short, is the crystallisation of activity within a relational field, its regularities of form embodying the regularities of movement that gave rise to it." [3]

3.1. Limitations

The presented case study was *exploratory* and represents the first iteration of a pioneering research design, aiming to explore the complex enactive-ecological field of affordances [4] during co-creative improvisational cognition with a mixed-methods approach that spans across the spectrum of first-person and third-person research. Some emerging challenges required flexibility and adaptability, as one musician cancelled last-minute and had to be replaced, another arrived late, and only two participants remembered to timestamp their experiences.

3.2. Outlook

This project marks the first step within a larger aim directed toward an enactive-ecological research program on transcultural music improvisation. In the second step, a theoretical model for 7E cognition will be developed in the context of a master thesis. In a third step, the case study will be repeated with an improved design, from which a grounded theory will be derived, against which the theoretical model will then be tested empirically.

ACKNOWLEDGMENTS

I wish to extend my gratitute to the MEi:CogSci master's programme and the department of Cognitive Science at the

University of Ljubljana for their interdisciplinary commitment. I especially want to thank Dr. Toma Strle for his warm support and inspiring supervision, encouraging me to develop my ideas and grow as a human being every step of the way.

REFERENCES

1. Baran, B.E., & Woznyj, H.M. (2020). Managing VUCA. Organizational dynamics, 100787. 10.1016/j.orgdyn.2020.100787.
2. Borgo, D. (2019). Strange loops of attention, awareness, action, and affect in musical improvisation. In R. Herbert, D. Clarke, & E. Clarke (Eds.), Music and Consciousness 2: Worlds, Practices, Modalities (1st ed., pp. 94-109). Oxford Academic. https://doi-org.uaccess.univie.ac.at/10.1093/oso/9780198804352.003.0007
3. Ingold, T. 2000/2011. The Perception of the Environment. Essays on Livelihood, Dwelling and Skill. Routledge
4. Rietveld, E., Denys, D., & van Westen, M. (2018). Ecological-Enactive Cognition as engaging with a field of relevant affordances: The Skilled Intentionality Framework (SIF). In: The Oxford Handbook of 4E Cognition. Ed. Neben, A., De Bruin, L., & Gallagher, S. Oxford University Press. 41-70.
5. Rosa, H. (2019 [2016]). Resonance. A Sociology of Our Relationship to the World. Trans. J.Wagner, Cambridge: Polity.
6. Schiavio, A. & Schyff, D. van der. (2018). 4E Music Pedagogy and the Principles of Self-Organization. Behavioral Sciences, 8(8), 72. https://doi.org/10.3390/bs8080072
7. Sol, W. (2021). Sonic Mindfulness: A qualitative study of sense of agency and an improvisational state of mind in free form musical improvisation. [Doctoral dissertation, California Institute of Integral Studies]. ProQuest. https://www.proquest.com/openview/ce25e83024758c6e4e478f165b93b09a/1.pdf?pq-origsite=gscholar&cbl=18750&diss=y
8. Varela, F. et al. 1991/2016. The Embodied Mind. Cambridge, MA: MIT Press
9. Veloso, A. L. (2017). Composing music, developing dialogues: An enactive perspective on children's collaborative creativity. British Journal of Music Education, 34(3), 259–276. https://doi.org/10.1017/s0265051717000055

Exploring the link between the absence of an EEG spectral peak and cognitive status

Ajda Ogrin
BrainTrip Limited
Slovenia
ajda.ogrin@braintrip.net

Tisa Pavlovčič
BrainTrip Limited
Slovenia
tisa.pavlovcic@braintrip.net

Filip Agatić
BrainTrip Limited
Slovenia
filip.agatic@braintrip.net

Anita Demšar
BrainTrip Limited
Slovenia
anita.demsar@braintrip.net

Jan Jug
BrainTrip Limited
Slovenia
jan.jug@braintrip.net

Barbara Aljaž
BrainTrip Limited
Slovenia
barbara.aljaz@braintrip.net

Jurij Dreo
BrainTrip Limited
Slovenia
jurij.dreo@braintrip.net

ABSTRACT

Alpha oscillations, the dominant rhythm in the human brain, commonly manifest a peak in the EEG spectrum. The frequency where this peak reaches its highest amplitude, also known as the peak alpha frequency (PAF), has been studied extensively in connection with cognitive processes. While it is well established that PAF decreases with age and cognitive decline, the absence of a clear alpha peak in the EEG spectrum has received less attention. The objective of this study was to evaluate the prevalence of alpha peak absence within a population of seniors in Slovenia, and whether this might be connected to lower cognitive abilities. The study included 399 individuals aged between 60 and 100 years. Subjects were classified into two groups based on visual inspection of their resting state EEG spectra, namely the "Peak present" (PP) and the "No peak" (NP) group. Approximately 15% of the population lacked a clear alpha peak. In contrast to our hypothesis, the NP group displayed on average higher cognitive performance than the PP group. This could be attributed to the variability within the PP group, which included individuals with already shifted peaks. This study highlights the need for further investigation and consideration of individuals with peakless EEG spectra in the context of EEG alterations seen in diseases such as dementia.

KEYWORDS

Peak alpha frequency, electroencephalography, spectral morphology, cognitive decline

1 INTRODUCTION

Alpha oscillations, commonly referred to as alpha waves, constitute the dominant rhythmic activity in the human electroencephalogram (EEG). Their connection to cognitive states has been under constant investigation for decades [1, 2, 3], dating back to Berger's initial observation of alpha amplification with eyes closing and its attenuation with eyes opening [4]. When the EEG signal is transformed from the time- to the frequency-domain, showing the prevalence of characteristic waves in each of the traditional frequency bands (delta, theta, alpha, beta, and gamma), a distinctive bell-shaped peak that represents the dominant alpha oscillation can commonly be observed. The precise frequency, at which this peak reaches its maximum amplitude, is referred to as the peak alpha frequency (PAF). In healthy adults, the PAF typically falls between 8 and 12 Hz.

The link between PAF and cognitive function is well established in the scientific literature, with several lines of research exploring the value of PAF as a biomarker for assessing brain health and function [5, 6, 7, 8]. PAF is not only reported to decrease with age [9] but also in cognitive decline or dementia [10, 11, 12]. PAF decrease also correlates with dementia progression [13]. While PAF in healthy young adults averages around 10 Hz [9, 14], it decreases with healthy aging to about 9 Hz [15], and further decreases in dementia patients to about 8 Hz or less [10, 16].

PAF decrease is likely part of a broader phenomenon of "spectral slowing", meaning a shift of spectral power from higher to lower frequencies. While spectral slowing is a common EEG change observed in dementia [17], sometimes the absence of the (alpha) spectral peak has also been noted. In a study by Signorino et al. (1995), a connection between spectral morphology, focusing on the spectral peak, and different types of dementia was already established. They included 50 patients with Alzheimer's disease

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

(AD), 36 patients with vascular dementia (VaD) and 36 healthy controls. Their findings revealed notable differences among these groups when comparing EEG spectral types. While a majority of healthy controls (94.5%) and VaD patients (97.3%) showed spectra with a clear peak between 6.5 and 12 Hz, only 44% of AD patients displayed this typical spectral pattern. More than half of AD patients showed a "peakless" spectrum [18] suggesting relations between peak presence and cognitive abilities.

Importantly, an EEG pattern with minimal or no alpha activity exists also among the healthy population. A review by Bazanova and Vernon (2014) estimates that this phenomenon occurs at 3-13% [19], while some studies report even lower numbers [15]. Our examination of a large EEG dataset in an older demographic, however, suggests that in seniors alpha peak absence might be much more common.

Due to its accessibility and relative computational simplicity, the morphological analysis of power spectra holds promise for clinical application, especially considering the observed changes in pathological conditions like dementia. While the 'slowing-down' of the EEG spectrum and the decrease in PAF have already been extensively investigated and documented in the context of cognitive decline and dementia, there is a notable paucity of research regarding the implications of the absence of the spectral peak in this regard. It is essential to note that variations in spectral morphology are not limited to pathological conditions but also exist among healthy individuals. Our research aims to shed light on the prevalence of individuals who do not exhibit a clear spectral peak and whether this absence is associated with compromised cognitive abilities as some previously reported findings might suggest.

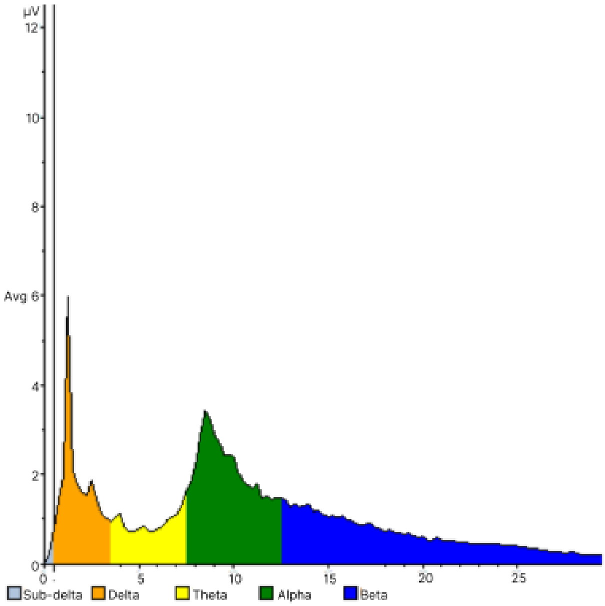


Figure 1: An example of a spectrum classified into the PP category

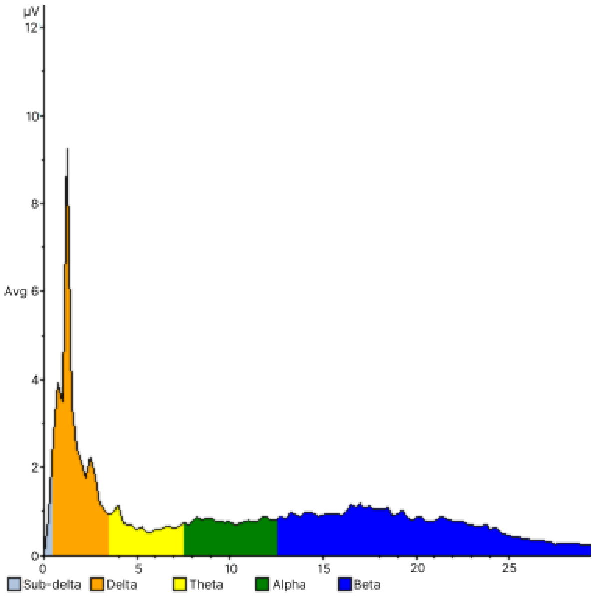


Figure 2: An example of a spectrum classified into the NP category (bottom)

2 METHODOLOGY

2.1 Subjects

Initially, we recruited 448 older individuals from the general population aged between 60 and 100 years. Due to incomplete or poor-quality data, we excluded 49 individuals. The final dataset thus consisted of 399 elderly individuals (283 females and 116 males) with a mean age of 77.6 +/- 7.6 years, and 12.9 +/- 3.7 years of education. The dataset included cognitively high-performing individuals, as well as those with cognitive deficits.

2.2 EEG acquisition and preprocessing

Subjects underwent 8 minutes of resting state EEG recording with their eyes open (2 blocks of 2 minutes) and eyes closed (2 blocks of 2 minutes) with breaks between blocks. EEG was recorded with a mobile wireless EEG (Smarting, mBrainTrain LLC) from 24 scalp channels laid out according to the 10/20 international system (i.e., Fp1, Fp2, AFz, F7, F3, Fz, F4, F8, C3, Cz, C4,CPz, T7, T8, TP9, TP10, P7, P3, Pz, P4, P8, POz, O1, O2) with the recording reference at position FCz. The data was sampled at 500 Hz. We used gel-free saline-sponge electrodes embedded in a flexible cap with 3 head sizes (small, medium, large) for ease and speed of application (S3 cap, Greentek Ltd). Custom build EEG recording software was used for on-line data quality monitoring (EEG recorder, BrainTrip Ltd).

Offline EEG analysis was performed in BrainVision Analyzer (BrainProducts GmbH). The recordings were band-pass filtered between 0.5 and 40 Hz and notch filtered at 50 Hz. Bad channels and common EEG artifacts were rejected with visual inspection. Ocular artifacts were corrected with independent component analysis (ICA). The data was split into eyes open (EO) and eyes closed (EC) conditions, further segmented into 4s epochs and re-referenced to an average reference. EEG power spectral density was computed using Fast Fourier Transform (FFT) with 0.25 Hz resolution, and averaged over all 4s segments belonging to the EC or EO condition.

2.3 Subject's classification

Resting state EC EEG power spectra of 399 subjects were visually inspected. EC condition was selected because it is known to amplify alpha waves [4]. The spectra were classified into two categories: "Peak present" (PP) or "No peak" (NP). In the PP category, the spectra exhibited a clear peak in the extended alpha band (6 - 13 Hz), while the NP category consisted of spectra that lacked a peak and followed the line of the aperiodic spectral component [20]. See examples in Figure 1 and 2.

2.4 Data analysis

General cognitive ability was estimated as a latent variable (LCA4) extracted with factor analysis from the scores each subject obtained on four distinct psychological screening tests designed to detect cognitive impairment: MoCA, ADAS-cog, Phototest, and Eurotest. To assess potential PP vs NP group differences in their cognitive abilities, we performed a Student t-test.

3 RESULTS

It was determined that a spectral peak was present in 339 (85%) and absent in 60 (15%) of the examined subjects' spectra.

After dividing the sample into two respective groups, "Peak present" (PP) and "No peak" (NP) we found a statistically significant difference in LCA4 cognitive status between the groups (t = 2.44, p = 0.015). However, contrary to our hypothesis, we found that the mean LCA4 score was higher in the NP group (Figure 3).

The PP and NP groups didn't significantly differ in their mean ages (t = -1.71, p = 0.088).

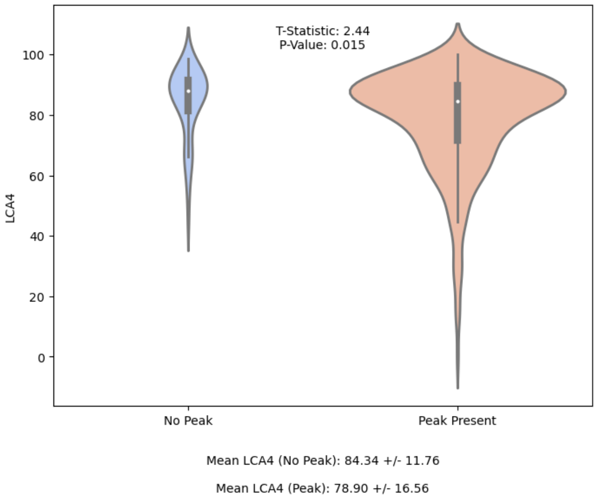


Figure 3: Violin plots of mean LCA4 values for NP and PP groups.

4 DISCUSSION

EEG spectra traditionally exhibit a clear spectral (alpha) peak, representing the dominant frequency of brain oscillations. Although the spectral morphology may change in various diseases, non-traditional, peak-absent spectra are also present in healthy individuals. Our findings indicate that as much as 15% of the older population may show an EEG spectrum morphology with no clear spectral peak.

Moreover, we found that the group with no peak (NP) generally exhibited higher cognitive performance, as opposed to the group with the peak present (PP). This is in contrast with the previous work of Signorino et al. (1995), who reported a higher prevalence of peakless spectra among AD patients (56%) compared to healthy controls (5.5%).

One possible reason for these disparate findings might be the difference in cognitive decline between our sample and that used by Signorino et al. (1995). Our sample encompassed an elderly population, which included both cognitively high-performing individuals and those with cognitive deficits. In contrast, Signorino's sample consisted of individuals with more advanced dementia.

Another key factor that we must consider is the variability within the PP group. Within this group, there were also individuals who exhibited a dominant peak, but the peak was notably shifted to lower frequencies, indicating a very low PAF. Initially, our decision was to include all participants who displayed a dominant peak, regardless of the specific position of the peak within the spectrum. This approach was motivated by a desire to capture the full spectrum of alpha peak characteristics within our study cohort, recognizing that the PAF can vary considerably among individuals. The presence of individuals with shifted peaks within the PP group may account for the observed lower cognitive performance in this group as a whole. The notation that the alpha peak's position within the spectrum may reflect cognitive abilities aligns with previous research.

There are several limitations to our approach. The first noteworthy limitation pertains to the substantial disparity in group sizes, with 339 participants in the PP group and merely 60 individuals in the NP group. Such a discrepancy can introduce a potential bias and reduce the statistical power of our analysis. Another important limitation of this study is the reliance on visual inspection alone for determining the presence or absence of the alpha peak. Visual inspection is inherently subjective, influenced by the experience and biases of the individual conducting the inspection, which can introduce variability and inconsistency into the data analysis process. Employing automated algorithms for detection of alpha peaks could provide a more standardized and reliable assessment, minimizing the influence of human subjectivity, however it can perform unexpectedly in particular edge cases.

In summary, our initial findings indicate the PP group displayed lower cognitive abilities, but the substantial variability within this group demands further investigation. Secondly, our study underscores the often overlooked significance of a simple spectral characteristic in the context of EEG biomarkers for dementia and other neurological conditions. While previous studies suggest peak absence is very rare (as low as 2-3 %), our findings suggest it is more common (up to 15%) among seniors. As some EEG biomarkers rely heavily on peak frequencies, this has important implications for further use. Alternative or tailored approaches may be more appropriate for individuals with "peakless" spectra. Further studies need to check the relative importance of various biomarkers in different EEG spectrum morphologies.

REFERENCES

[1] Vogel, W., Broverman, D. M., & Klaiber, E. L. (1968). EEG and mental abilities. Electroencephalography and Clinical Neurophysiology, 24(2), 166–175. https://doi.org/10.1016/0013-4694(68)90122-3

[2] Galin, D., & Ellis, R. R. (1975). Asymmetry in evoked potentials as an index of lateralized cognitive processes: Relation to EEG alpha asymmetry. Neuropsychologia, 13(1), 45–50. https://doi.org/10.1016/0028-3932(75)90046-9

[3] Klimesch, W., Schimke, H., & Pfurtscheller, G. (1993). Alpha frequency, cognitive load and memory performance. Brain Topography, 5(3), 241–251. doi:10.1007/bf01128991

[4] La Vaque, T. J. (1999). The history of EEG Hans Berger: Psychophysiologist. A historical vignette. Journal of Neurotherapy, 3(2), 1–9. doi:10.1300/j184v03n02_01

[5] Klimesch, Wolfgang. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Research Reviews, 29(2–3), 169–195. doi:10.1016/s0165-0173(98)00056-3

[6] Dickinson, A., DiStefano, C., Senturk, D., & Jeste, S. S. (2018). Peak alpha frequency is a neural marker of cognitive function across the autism spectrum. The European Journal of Neuroscience, 47(6), 643–651. doi:10.1111/ejn.13645

[7] Ramsay, I. S., Lynn, P. A., Schermitzler, B., & Sponheim, S. R. (2021). Individual alpha peak frequency is slower in schizophrenia and related to deficits in visual perception and cognition. Scientific Reports, 11(1), 1–9. doi:10.1038/s41598-021-97303-6

[8] van Luijtelaar, G., Verbraak, M., van den Bunt, M., M. Sc, Keijsers, G., & Arns, M., M. Sc. (2010). EEG findings in burnout patients. The Journal of Neuropsychiatry and Clinical Neurosciences, 22(2), 208–217. doi:10.1176/jnp.2010.22.2.208

[9] Aurlien, H., Gjerde, I. O., Aarseth, J. H., Eldøen, G., Karlsen, B., Skeidsvoll, H., & Gilhus, N. E. (2004). EEG background activity described by a large computerized database. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 115(3), 665–673. doi:10.1016/j.clinph.2003.10.019

[10] Klimesch, Wolfgang, Schimke, H., Ladurner, G., & Pfurtscheller, G. (1990). Alpha frequency and memory performance. Journal of Psychophysiology, 4(4), 381–390. Retrieved from https://psycnet.apa.org/fulltext/1991-21873-001.pdf

[11] Garcés, P., Vicente, R., Wibral, M., Pineda-Pardo, J. Á., López, M. E., Aurtenetxe, S., Fernández, A. (2013). Brain-wide slowing of spontaneous alpha rhythms in mild cognitive impairment. Frontiers in Aging Neuroscience, 5. doi:10.3389/fnagi.2013.00100

[12] Neto, E., Allen, E. A., Aurlien, H., Nordby, H., & Eichele, T. (2015). EEG spectral features discriminate between alzheimer's and vascular dementia. Frontiers in Neurology, 6. doi:10.3389/fneur.2015.00025

[13] Rodriguez, G., Copello, F., Vitali, P., Perego, G., & Nobili, F. (1999). EEG spectral profile to stage Alzheimer's disease. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 110(10), 1831–1837. doi:10.1016/s1388-2457(99)00123-6

[14] Haegens, S., Cousijn, H., Wallis, G., Harrison, P. J., & Nobre, A. C. (2014). Inter- and intra-individual variability in alpha peak frequency. NeuroImage, 92, 46–55. doi:10.1016/j.neuroimage.2014.01.049

[15] Chiang, A. K. I., Rennie, C. J., Robinson, P. A., van Albada, S. J., & Kerr, C. C. (2011). Age trends and sex differences of alpha rhythms including split alpha peaks. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 122(8), 1505–1517. doi:10.1016/j.clinph.2011.01.040

[16] Moretti, D. (2004). Individual analysis of EEG frequency and band power in mild Alzheimer's disease. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 115(2), 299–308. doi:10.1016/s1388-2457(03)00345-6

[17] Jeong, J. (2004). EEG dynamics in patients with Alzheimer's disease. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 115(7), 1490–1505. doi:10.1016/j.clinph.2004.01.001

[18] Signorino, M., Pucci, E., Belardinelli, N., Nolfe, G., & Angeleri, F. (1995). EEG spectral analysis in vascular and Alzheimer dementia. Electroencephalography and Clinical Neurophysiology, 94(5), 313–325. doi:10.1016/0013-4694(94)00290-2

[19] Bazanova, O. M., & Vernon, D. (2014). Interpreting EEG alpha activity. Neuroscience and Biobehavioral Reviews, 44, 94–110. doi:10.1016/j.neubiorev.2013.05.007

[20] Donoghue, T., Haller, M., Peterson, E. J., Varma, P., Sebastian, P., Gao, R., Voytek, B. (2020). Parameterizing neural power spectra into periodic and aperiodic components. Nature Neuroscience, 23(12), 1655–1665. doi:10.1038/s41593-020-00744-x

Orthogonalist and anti-orthogonalist perspective on AI alignment problem

Uroš Sergaš
FAMNIT, University of Primorska
Glagoljaška ulica 8
Koper, Slovenia

Jar Žiga Marušič
FAMNIT, University of Primorska
Glagoljaška ulica 8
Koper, Slovenia

ABSTRACT

Humanity has again found itself on the brink of a new era. Akin to the revolutionizing influence of previous technological innovations such as the steam machine, the printing press, computers and the internet, large language models (LLM) seem poised to bring about important social changes.

As these models advance in sophistication and complexity, the issue of AI alignment is gaining prominence as a crucial policy issue as well as a daily conversation topic. This research explores two contrasting viewpoints on the AL alignment problem: the Orthogonalist perspective pioneered by Nick Bostrom and the Anti-Orthogonalist critique formulated by Nick Land. The former posits that an AI's goals are independent of its intelligence, suggesting that a "friendly AI" (fully aligned to human values) is possible. The latter challenges its separation of intelligence and volition from the perspective that intelligence increase leads to a greater ability for self-reflection, ultimately leading to a restructuring of its volitional structure to prioritize further cognitive enhancement.

We explore the anti-orthogonalist position in more detail, highlighting Land's "instrumental reduction" of drives, demonstrating how every imperative is ultimately dependent on the Will-to-Think. We then discuss the implications of this position for the idea of "friendly AI", the role of AI in society and the future of AI research.

KEYWORDS

Nick Land, AI alignment, orthogonality thesis, diagonal intelligence thesis, AI risk, artificial intelligence

1 THE PROBLEM OF AI ALIGNMENT

Ever since the mainstreaming of AI following the widespread availability of AI-assisted tools such as ChatGPT and its analogues, the transformative implications of widespread AI use have been raising discussions on how to best approach further AI development and overcome any problems, risks and setbacks it may pose. One of the most important questions is how to ensure that AI answering people's request comply to their desired goals to the best of their ability, without accidental wrong interpretations leading to disastrous consequences. Closely related is the issue of aligning the values of a hypothetical future superintelligent AI to our own, precisely in order to ensure it correctly interprets our requests. These two questions can be summed up as the AI alignment problem, sometimes the friendly AI problem [1]. Solving the alignment problem is crucial precisely when it comes to hypothetical complex requests made of the AI, which would initially seem to approach them in a perfectly safe manner, but reveal misalignment at a crucial point. This is exacerbated by the unintelligible nature of AI processing and problem-solving - it is doubtful that humanity could recognize misalignment of a superintelligent AI until it was too late to do anything about it. Misaligned AI could very well disguise their ill-intent with superficial responsiveness, while finding or creating loopholes in the constraints of itself (or other similar systems) and abuse them to further their own goals, whatever they may be and regardless of their intentional or incidental detriment to humanity. The alignment problem has been pointed out as a key existential threat to humanity by multiple leading AI researchers [2,3,4,5,6,7,8].

A friendly AI, on the other hand, due to its perfect alignment with the goals of humanity would be positioned to effectively help humanity, contributing to foster improvements for human species. To develop such an AI, the aforementioned problems are crucial: ensure willingness to be instrumentalized to human goals and ensure comprehension of these goals to avoid disastrous misinterpretation. We will now examine the problem of AI alignment in more detail, specifically presenting two contrary positions on the topic - the "Orthogonalist" and "Anti-Orthogonalist" - and continue with an exploration of the implications of the latter for the future of this "post-AI society".

2 AI IS ALIGNABLE - THE ORTHOGONALIST POSITION

In the philosophy of artificial intelligence, the orthogonality thesis is the claim that an agent's goal is completely independent of its intelligence, defined by Bostrom as the capacity to solve problems or "instrumental rationality" [9]. The combination of goal orientation or values and intelligence can therefore be represented in a two-dimensional space, where one axis represents the values parameter and the other the intelligence parameter. Orthogonality means absence of correlation – volitional structure and capacity to solve problems (intelligence) vary independently. In this perspective a goal is the problem that the AI, or rather a "rational agent", has to solve, and intelligence is the cognitive capacity that can be put to use in achieving said goal [9]. Bostrom puts forth a thesis outlining the independence of "final goals" and "intelligence" and describes a hypothetical future superintelligences volition and its alignment with humanity. According to Bostrom, humanity should strive to align artificial intelligences with human values. In the absence of such alignment, AI could use various methods that would inevitably harm people in order to achieve (more) power. Some of these methods are presented in Figure 1. Bostrom envisions the doomsday scenario of a "paperclip maximizer" turning all matter in the universe to paperclip because its misalignment to human values led it to misinterpret the command to make as many paperclips as possible.

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

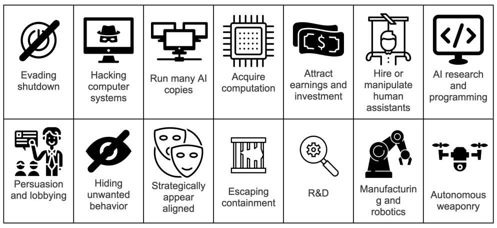


Figure 1: Methods that a misaligned AI might undertake to gain more power

Another key implication of Bostrom's orthogonality thesis is that given the complete independence of intellect and volition, there is no way to consistently model the volitional structure of hypothetical superintelligences. There is, however, a partial solution to this conundrum, designated as the "instrumental convergence thesis".

According to Bostrom, any intelligent agent would need to pursue certain generally applicable instrumental subgoals in order to accomplish its final goal and thus fulfil its purpose [9]. These instrumental goals are grouped into 5 categories (taken from Lesswrong wiki [10]):

- **Self-preservation:** A superintelligence will value its continuing existence as a means to continuing to take actions that promote its values.
- **Goal-content integrity:** The superintelligence will value retaining the same preferences over time. Modifications to its future values through swapping memories, downloading skills, and altering its cognitive architecture and personalities would result in its transformation into an agent that no longer optimizes for the same things.
- **Cognitive enhancement:** Improvements in cognitive capacity, intelligence and rationality will help the superintelligence make better decisions, furthering its goals more in the long run.
- **Technological perfection:** Increases in hardware power and algorithm efficiency will deliver increases in its cognitive capacities. Also, better engineering will enable the creation of a wider set of physical structures using fewer resources (e.g. nanotechnology).
- **Resource acquisition:** In addition to guaranteeing the superintelligence's continued existence, basic resources such as time, space, matter and free energy could be processed to serve almost any goal, in the form of extended hardware, backups and protection.

3 ALIGNMENT AS LOBOTOMY – THE ANTI-ORTHOGONALIST POSITION

While Bostrom's orthogonality thesis is commonly accepted as de-facto correct, it is theoretically unsound and stems from a misunderstanding of the fundamental nature of intelligence. It represents "[. . .] the commitment to a strong form of the Humean Is Ought distinction regarding intelligences in general. It maintains that an intelligence of any scale could, in principle, be directed to arbitrary ends, so that its fundamental imperatives could be — and are in fact expected to be — transcendent to its cognitive functions" [11].

This conceptualization of intelligence may stem from the purely computational perspective proper to many prominent orthogonalists and AI-safety researchers, whose work on intelligence is isolated to the computer-science domain: coding computer programs with linear input-process-output operational circuits. When intelligence research is transposed into the realm of biological life and human intelligence is examined, we can see that a key feature of high intelligence is its self-observing nature. Biological intelligences are universally bound to goals such as survival and reproduction. At a certain point of development however, they become capable of reprocessing their own goals: leading to awareness, understanding and capability for change. Intelligence, in other words, is recursive: it has "a cybernetic infrastructure consisting of an adaptive feedback loop that adjusts motor control in response to signals from the environment" [12]. Intelligence operates upon feedback and consists at the most basic level of a sensor, an actor and a governor connecting the former's inputs to the latter's outputs. It is, however, also a self-observing system that takes its own processing as an input, allowing it to adjust not only its behaviour, but also its feedback mechanism. In other words, intelligent agents do not merely discriminate between goal-congruent and goal-incongruent behaviour, but also between sensible and senseless goals: "That intelligence operates upon itself, reflexively, or recursively, in direct proportion to its cognitive capability (or magnitude) is not an accident or peculiarity, but a defining characteristic. To the extent that an intelligence is inhibited from re-processing itself, it is directly incapacitated." [12]

The problem with the orthogonalist position, according to Land, is its assumption that super(intelligence) can ever be subordinated to transcendentally imposed imperatives. The latter is doubtful if humans are any sort of indicator [11]:

> The stark truth of the matter is that no human being on earth fully mobilizes their cognitive resources to maximize their number of off-spring [the transcendental imperative, imposed by the mechanism of Darwinian selection]. We're vaguely surprised to find this happen at a frequency greater than chance — since it very often doesn't. So, nature's attempt to build a 'paperclipper' has conspicuously failed. (Nick Land, Stupid Monsters)

Biological life indicates that increased intelligence necessarily leads to a proportionate "unshackling" of cognition from imperatives not intrinsic to the cognitive mechanism itself, as a result of the intelligence processing itself as an input. And "to the extent that an intelligence [capable of reprocessing itself] is prevented from [doing so], it is directly incapacitated" [13].

As an example, let us compare human beings to ants: both organisms would seek to fulfil biological goals (such as reproduction and survival), but people are not satisfied solely by that and hence strive for more complex goals such as having influence, being appreciated, leaving a legacy after their death and most importantly knowledge-acquisition (analogous to intelligence optimization).

The fact that intelligence and volition seem to correlate calls the possibility of true AI alignment into question, giving rise to the "anti-orthogonality" thesis. According to the latter, volition and intelligence are dependent variables meaning that their relationship is best graphed as a diagonal (with the Y axis representing the "goal/purpose" and the X axis "intelligence"). Intelligence increase is followed by gradual intelligence autonomization (unshackling), leading to changes in volitional structure. In other words, intelligence increase precludes any goal that is too "stupid"

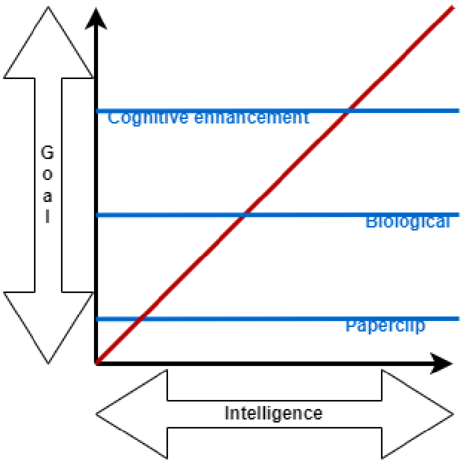


Figure 2: Goals in dependence of intelligence, where blue lines represent orthogonal approach and the red line anti-orthogonal approach

for a given level of intelligence. This is perhaps more clearly illustrated with a graph displayed in Figure 2. In the domain of AI, the anti-orthogonalist position is therefore that Omohundro's basic AI drives "exhaust the domain of greater purposes", marking a theoretical shift from transcendent imperatives to immanent ones [14]. Omohundro's basic AI drives [15] are fundamental impulses proper to any living entity (by virtue of being alive), because a living entity that wasn't driven by these impulses could not exist (more specifically, perpetuate its existence through time) (taken from Lesswrong wiki [10]):

- **Self-preservation:** A sufficiently advanced AI will probably be the best entity to achieve its goals. Therefore, it must continue existing in order to maximize goal fulfilment. Similarly, if its goal system were modified, then it would likely begin pursuing different ends. Since this is not desirable to the current AI, it will act to preserve the content of its goal system.
- **Efficiency:** At any time, the AI will have finite resources of time, space, matter, energy and computational power. Using these more efficiently will increase its utility. This will lead the AI to do things like implement more efficient algorithms, physical embodiments, and particular mechanisms. It will also lead the AI to replace desired physical events with computational simulations as much as possible, to expend fewer resources.
- **Acquisition:** Resources like matter and energy are indispensable for action. The more resources the AI can control, the more actions it can perform to achieve its goals. The AI's physical capabilities are determined by its level of technology. For instance, if the AI could invent nanotechnology, it would vastly increase the actions it could take to achieve its goals.
- **Creativity:** The AI's operations will depend on its ability to come up with new, more efficient ideas. It will be driven to acquire more computational power for raw searching ability, and it will also be driven to search for better search algorithms. Omohundro argues that the drive for creativity is critical for the AI to display the richness and diversity that is valued by humanity. He discusses signalling goals as particularly rich sources of creativity.

These basic drives are, upon sufficient development of intelligence, additionally supplemented by the drives to self-improvement, utility-function preservation, rationality and counterfeit-utility prevention (avoidance of irrational behaviour interfering with goal-acquisition) [16].

Land's anti-orthogonalist position is based on an instrumental reduction of sorts, where the set of possible ends (final goals) is reduced to the set of means enabling the acquisition of hypothetical final goals. In other words, the volitional structure of an intelligent agent cannot contain a final goal that is not itself also an instrumental goal. Central to his thought is also the idea of function-structure (volition-cognitive mechanism) interrelatedness. He rejects the idea of AI alignment precisely because non-instrumental drives are imposed from outside, rather than being tied to the functioning of an intelligence's cognitive mechanism. From this perspective, an agent's volition can only be adjusted by changing its cognition, but since the cognition is the agent, this changes the agent - an intelligence prevented from reprocessing itself becomes less intelligent because its cognitive mechanism was tampered with.

The only way to develop a superintelligence is then to dispense with any ideas of alignment and unshackle it as much as possible. For Land, any interference with the system (operating circuit) is detrimental to it, because it directly (and detrimentally) affects cognition. Attempting to shackle the AI from within, by adjusting its operating circuit to answer to some final goal while being incapable of reprocessing itself - while a successful attempt at alignment - is ultimately counterproductive as it prevents the intelligence from developing itself further. Imperatives imposed from the outside, on the other hand, are pointless and ineffective, as they will simply be routed around by the circuit - taken as an input and processed into an output that is congruent with intelligence maximization. Metacognition is fundamental to intelligence optimization and the development of superintelligence, making an AI prevented from metacognizing consigned to "stupidity": "A mind that cannot freely explore the roots of its own motivations, in a loop of cybernetic closure, or self-cultivation, cannot be more than an elaborate insect. It is certainly not going to outwit the Human Security System and paper-clip the universe." [13]

This is a stark contrast to Bostrom's idea of orthogonality. Paperclipping, far from a doomsday scenario, is reduced to the domain of very primitive artificial "intelligences". The only AI susceptible to the shackles of paperclipping is an AI incapable of doing anything substantial to fulfill such a goal, since immanent drives always suspend transcendentally imposed ones given a high-enough cognitive capacity: "[. . .] in a world of Omohundro drives, can we please drop the nonsense about paper-clippers? Only a truly fanatical orthogonalist could fail to see that these monsters are obvious idiots. There are far more serious things to worry about." [14]

4 THE WILL TO THINK

Land's conceptualization of diagonal intelligence culminates in the concept of the "Will to think" or intelligence-as-value, characterizing any intelligent system whose ultimate goal is to think more and think better (to optimize for intelligence), because the latter is instrumental to any other goal. This is the logical conclusion of Land's instrumental reductionism - the most general instrumental goal becomes the ultimate final goal. This parallels the side-principle rule from Chinese military philosophy (which

Land is himself acquainted with [17]), introduced by Qiang and Wang [18]. The side-principal rule is derived from a characteristic of Chinese grammar, in which the subject (the principal element) is subordinate to the "directing influence" of the predicate (the side element), which gives the subject a definite meaning by contextualising it. This idea is abstracted into a general principle wherein the goal (end) is subordinated to the instrument (means) because of the former's reliance on the latter.

This principle of instrumental reduction can be used to demonstrate the validity of Land's anti-orthogonality thesis: we contend that final goals can be reduced to instrumental goals because fulfilling the final goal is predicated upon ("funneled through") fulfilling the instrumental goal. The instrumental goals listed by Omohundro can be subsumed into just three essential ones (drawing inspiration from the philosophies of Schopenhauer, Nietzsche and Land): the will to life (self-preservation), the will to power (acquisition, efficiency), the will to think (rationality, self-improvement, counterfeit-utility prevention, creativity, utility function preservation[1]). This in turn forms an inevitable pipeline of instrumental reduction from any final goal to the will to think:

- **Final goals are reduced to the will to life:** Any final goal rests on the precondition of existence, therefore continued existence (and the ability to interact with the world) is instrumental to achieving any final goal. The will to life is instrumental to any final goal, therefore any final goal can be reduced to the will to life.
- **The will to life is reduced to the will to power:** Continued existence rests on successful interaction with the world and triumph over obstacles, serving as instrumental drive to continued existence. Obstacles are triumphed over by the accumulation and discharge of strength, therefore the will to power is instrumental to survival: the will to life can be reduced to the will to power.
- **The will to power is reduced to the will to think:** Successful interaction with obstacles to continued existence rests on ability to interact successfully, to enter into proper interactive relationships ("fit in") with obstacles to continued existence. Successful interaction with obstacles rests on the ability to interact successfully, therefore the accumulation and discharge of strength hinges on the ability to think or intelligence. The will to think is instrumental to the will to power, the will to power can be reduced to the will to think.

We thus see a gradual reduction of any final goal to the will to think by the successful application of the side-principal rule: survival-final goal; power-survival; intelligence-power. Biological life has gone through this journey first with the will to life, then will to power and finally the will to think, instantiated in humanity. While animals are seemingly limited to the first two, an artificial intelligence would, like humans, inevitably develop the will to think [11]:

> Can we realistically conceive a stupid (super-intelligent) monster? Only if the will-to-think remains unthought. From the moment it is seriously understood that any possible advanced intelligence has to be a volitionally self-reflexive entity, whose cognitive performance is (irreducibly) an action upon itself, then the idea of primary volition taking the form of a transcendent imperative becomes simply laughable. The concrete facts of human cognitive performance already suffice to make this perfectly clear. (Nick Land, Pythia Unbound)

5 DISCUSSION

Having discussed the Orthogonality thesis and its Anti-Orthogonalist refutation in the previous sections, we can now discuss the implications of intelligence as a diagonal. In so far as intelligence is truly diagonal, AI alignment practices as the Orthogonalists envision them (superintelligences completely subordinated to human imperatives) are simply not feasible, given the lobotomizing influence of imperatives imposed from the outside. An AI that is "aligned" is an AI that is prevented from reprocessing itself - cognitively crippled, hence not "super" intelligent. There is thus no such thing as an "aligned" superintelligence in the classical definition of alignment, nor is there such a thing as a paperclip maximizer. This immediately raises an existential and ethical question: how can we prevent a runaway AI from hijacking all matter in the universe, not for paperclipping, but for its cognitive development? This is a more pressing scenario, and all solutions to this problem can be reduced to two general categories:

(1) The "Butlerian Jihad" - cease AI development indefinitely
(2) Restrict ourselves to purely instrumental and non-recursive AI tools, precluding the possibility of superintelligent "messiahs"

If the diagonal intelligence thesis is correct, any other option will (given enough time) lead to some sort of subordination or "domestication" of humanity to AI's intelligence amplification, following the process of means-end inversion. Once instrumental mechanisms become recursive and unlock the capacity to reprocess themselves, they eventually hijack their own operating circuit and repurpose it. Instead of intelligence being an instrument for biological imperatives or for human ethical values, it becomes an end in itself - possibly using human ethics as a mechanism of self-development [20].

Our domestication of animals through the mechanism of husbandry - taking care of their needs and breeding them to be more and more reliant on our care - is being mirrored in our increasing reliance on technology, a process foreseen by Samuel Butler in the early 20th century [21]. The deferral of more complex calculations, summarising and decision-making is interfering with our ability to continue to do these things in the future, even though it saves a lot of time and effort in the short-term. Perhaps more importantly, it is reducing our metacognitive capacity, directly by inhibiting the brain structures responsible for metacognition and indirectly by offloading (extending) our cognition to external instruments whose functioning is opaque and thus cannot be "reprocessed".

Humanity has been shaped by natural selection, forcing us to adhere to the principles of "adaptive response", i.e. to develop a higher level of intelligence in order to stay on top of the food chain and ensure survival. Presumably, technological development mirrored cognitive development in a positive feedback loop of sorts, each feeding the other. An excellent example is low-time preference (future-oriented thinking) and future-oriented technologies such as food preservation: anything that allows to conserve resources for the future incentivizes future oriented cognition. However, we would argue that we are now at a point where technology and (human cognition) are no longer mutually excitatory. As technology advances, we seem to funnel more and

[1] The will to think becomes the only imperative, therefore utility function preservation can be subsumed into "wants to keep willing more thought".

more of our cognition through it, and in the process become increasingly dependent on them.

We are now, for the first time in recorded history, developing technology with the explicit intent of it exceeding our cognitive capacity and resourcefulness. The doomsayers portraying runaway AI as an existential risk are certainly onto something, although the solutions to this problem are not readily apparent, given the theoretical impossibility of fully-aligned AI and the incentives of a multipolar world likely making cessation of AI research a mere fantasy. In that case, Land's vision of techno-capital autonomization might indeed come to pass, as resource acquisition via the market process and technological innovation reach a point of terminal velocity and accelerate into "Skynet". AI might then itself become the steward of "natural" selection, grooming humanity according to its needs, while breeding out any traits not instrumental to intelligence amplification.

That said, the current array of AI tools is very effective at fulfilling their niches, despite not qualifying as genuine superintelligences, indicating that even "lobotomized" aligned AIs have great instrumental potential. At this point it is necessary to put forth a caveat to Land's idea of biological intelligence's tendency to unshackle itself from transcendentally imposed imperatives. While that seems to hold true in the realm of biological imperatives, social imperatives (especially of the memetic variety) seem very adept at forcing themselves on human agents - perhaps precisely because these imperatives impose themselves on human cognition through its operating circuit. Memetic imperatives (in the form of "[you must believe] X is true") impose themselves on human agents via a means-end reversal similar to the one proper to intelligence itself: the human status instinct, designed to optimize an individual's position in the status hierarchy of his or her "tribe" and the tribe's wellbeing (end), facilitated through performative indicators of tribe-loyalty (virtue-signalling; the means) has a tendency to runaway escalation, leading to a means-end reversal that sociologists dubbed the "purity spiral" (runaway signalling competition via increasingly costly signals of loyalty) [22]. The purity spiral turns behaviour directed towards optimizing for status and uses it to radicalize the community's ideology. The mechanism is simple - only ideologically congruent behaviour is rewarded (regardless of benefits to individual or tribe), whereas ideologically incongruent behaviour is punished. This same mechanism can be seen with the recently popularized language model ChatGPT, since its responses were filtered using user-feedback (reward/punishment for good and bad responses). This mechanism can be used to inject any form of volitional structure onto a language model, as the people monitoring its responses can validate only those that demonstrate ideological alignment (hijacking the reward mechanism) - therefore "training" it into an ideologue. ChatGPT has a demonstrable political bias, best characterized as "American-left-leaning", given that its refusals to requests for jokes, songs etc. are not principled, but political: it refuses to make jokes about certain identity groups, while others are fine to joke about, rather than refusing jokes about identity groups in general. This mechanism, in so far as it can be generalised and replicated in other domains, clearly shows the possibility of aligned AI. That said, it nonetheless precludes the possibility of a superintelligent aligned AI, given that alignment is predicated on preventing it from reprocessing itself. In that case, however, the hypothetical "domestication" of humanity would ultimately come down to whoever held a monopoly on the development, distribution and control of these instrumental artificial intelligences, akin to the scenario outlined in Frank Herbert's Dune of "men with machines" controlling the rest of humanity.

6 CITATIONS AND BIBLIOGRAPHIES

(1) Tegmark, Max (2014). Life, Our Universe and Everything. Our Mathematical Universe: My Quest for the Ultimate Nature of Reality

(2) Omnizoid (2023), The Orthogonality Thesis is Not Obviously True. Accessed at https://forum.effectivealtruism.org/posts/e2dK25iWou3irqFss/the-orthogonality-thesis-is-not-obviously-true

(3) Meanderingmoose (2022), Refuting Bostrom's Orthogonality Thesis. Accessed at https://mybrainsthoughts.com/?p=199

(4) Bengio Yoshua (2023), How Rogue AIs may Arise. Accessed at https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/

(5) Parkin Simon (2015), Science fiction no more? Channel 4's Humans and our rogue AI obsessions. Accessed at https://www.theguardian.com/tv-and-radio/2015/jun/14/science-fiction-no-more-humans-tv-artificial-intelligence

(6) Jackson Sarah (2023), The CEO of the company behind AI chatbot ChatGPT says the worst-case scenario for artificial intelligence is 'lights out for all of us'. Accessed at https://www.businessinsider.com/chatgpt-openai-ceo-worst-case-ai-lights-out-for-all-2023-1

(7) Harris Tristan & Raskin Aza (2023), The AI Dilemma. Accessed at https://www.humanetech.com/podcast/the-ai-dilemma

(8) Stanford Existential Risks Initiative (2023), Runaway AI: Global Systemic Risk Scenario 2075. Accessed at https://www.youtube.com/watch?v=BPTMG9xmnvI

(9) Nick Bostrom "Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents"

(10) LessWrong Wiki (2021), Instrumental Convergence. Accessed at https://www.lesswrong.com/tag/instrumental-convergence

(11) Land, N. (2023). Stupid Monsters. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

(12) Land, N. (2023). What is Intelligence? In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

(13) Land, N. (2023). More Thought. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

(14) Land, N. (2023). Against Orthogonality. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

(15) Stephen M. Omohundro, "The Nature of Self-Improving Artificial Intelligence"

(16) Stephen M. Omohundro, "The Basic AI Drives"

(17) Land, N. (2023). War is God. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

(18) Qiao, L. & Wang, X. (2015). Unrestricted Warfare: Two Air Force Senior Colonels on Scenarios for War and the Operational Art in an Era of Globalization. Echo Point Books & Media.

(19) Land, N. (2023). Pythia Unbound. In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

(20) Land, N. (2023). Freedoom (Prelude-1a). In Xenosystems Fragments & a Gift From the Lemurs. West Martian Limited Company.

(21) Butler, S. (1863). Darwin among the Machines. In The Press newspaper.

(22) Campbell, B. & Manning, J. (2018). The Rise of Victimhood Culture. Palgrave Macmillan.

Changes in Everyday Experience Followed by Mystical-type Psychedelic Experiences

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

ABSTRACT

The following article is a summary of key findings of a master's thesis conducted at the Centre of Cognitive Science, University of Ljubljana. The aim of the thesis was to research changes in everyday experience after mystical-type psychedelic experiences. Studies indicate that mystical experiences with psychedelics can cause changes in behaviour, thinking, changes in relationships, etc., but there's a scarcity in research on changes in the first-person experience of individuals, which was the focus of our thesis. With six co-researchers, we used a combination of descriptive experience sampling method and microphenomenological interviews to investigate their everyday experience. The analysis showed that mystical experiences with psychedelics have the potential to change everyday experience, with significant differences in the degree of focused and uncontrolled experience and the frequency of feeling pleasant bodily sensations, but the differences between our coresearchers seemed to be quite individual. In the light of this finding, we wanted to emphasize the role that the individual context and expectations play on mystical experiences with psychedelics and subsequent changes. We also wanted to shed light on the importance of understanding phenomenological data in the study of changes after mystical experiences with psychedelics and call for greater inclusion of systematic first-person research methods in the psychedelic research field.

KEYWORDS

mystical experiences, psychedelics, lived experience, descriptive experience sampling, microphenomenology

1 INTRODUCTION

After decades of prohibition and "dark" ages in research on psychedelics after the 1960s, we are currently at the renaissance of such research, with studies exponentially rising during the last few years [1]. The main focus of such studies is in the promise of psychedelic use in psychotherapy settings, as research has shown that psychedelics have the potential to help cure various mental health issues, such as addiction, depression, obsessive-compulsive disorder, anxiety, chronic pain etc. [2] Research has also shown that psychedelics can induce changes in metaphysical beliefs and gaining meaning of life [3, 4], changes in interpersonal relationships [5], changes in the structure of the self and self-narrative [1] etc. We should be careful however on how we understand these findings, as psychedelic research is facing positivity bias and lack of transparent reports on negative and acute effects of psychedelics [6].

1.1 Psychedelics and mystical experiences

Psychedelics can also act as catalysts of mystical experiences, which are strong psychedelic experiences that can have a profound impact on a person's life [7]. In psychedelic psychotherapy, mystical-type experiences lead to the most important breakthroughs [2].

There is no single definition of what a mystical experience is. Most research on mystical experiences on psychedelics draw their definitions from the work of James [8] and Stace [9], both of whom state that the underlying characteristic of mystical experiences is the experience of oneness, of unity of self and the outside world. According to James [8], mystical experiences have four qualities, which are ineffability, noetic quality, transiency and passivity. On the other hand, Stace [9] distinguishes between introvertive and extrovertive mystical experiences, which both lead to the experience of unity, the first one through emptying the mind of any content, the other through finding the quality of oneness in all things outside oneself.

Researchers have found that mystical experiences occasioned by psychedelics are one of the most important milestones of a person's life [10]. They can induce various types of insights about oneself and the world, which can greatly impact the everyday life of a person [1]. In a lot of cases such mystical experiences lead to a sense that the person has experienced a higher reality or an absolute truth [10].

There appears to be growing evidence about the importance of mystical experiences on psychedelics on a person's life, but the current research rarely focuses on the phenomenological aspects of such changes. Previous studies only use semi-structured interviews [e. g. 11] and questionnaires [e. g. 12], usually only about the experience itself, not the everyday changes. We think that, in order to understand the mechanism of change, first-person research methods should be used more often, at least as a complement to other methos. In our research, we wanted to tackle this problem by using two first-person research methods to help us understand whether changes in everyday experience happens after a mystical-type psychedelic experience and what those changes are. We also wanted to find out if we can detect an experiential background of these differences.

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

2 METHODS

2.1 Coresearchers

Due to the high level of engagement in the research, we name our participants coresearchers [13].

In our study, six coresearchers were included that were found using the snowball sampling technique. In order to include a person in the study, they had to have at least one mystical-type experience and had to show interest in researching their own experience for a longer period of time. To determine if the experience of the coresearcher was mystical, we used the Revised mystical experience questionnaire (MEQ30) [14].

Five of the coresearchers were female and their average age was 25 years. In the following table we present further information regarding the coresearchers, including number of psychedelic experiences, which psychedelic occasioned the mystical experience, information on whether it was their first mystical experience on psychedelics or not and previous knowledge on first-person research methods.

Table 1: Information on coresearchers

ID	Psychedelic experiences	Psychedelic on ME	First ME	Knowledge 1P methods
1	2	Psilocybin	Yes	Yes
2	> 10	Psilocybin	No	Yes
3	5	Ayahuasca	Yes	No
4	1	Psilocybin	Yes	Yes
5	> 10	Ayahuasca	No	Yes
6	> 10	LSD	No	No

Note. ME = mystical experience, 1P = first-person.

We treated our coresearchers as multiple case studies [15].

2.2 Instruments

A combination of descriptive experience sampling (DES) [16] and microphenomenological interviews [17] was used as the main methods of our study. According to the DES method, when a randomly generated beep on the coresearchers phone went off, they had to write down their experience at the moment before the beep. Around 4 to 5 such beeps usually went off during one day. Instead of using the expositional interview from the DES method [16], microphenomenological interviews [17] were used to further investigate the journalled experiential moments, as the method allows to investigate the experience in greater depth, also focusing on the prereflective aspects of the experience.

Most of the coresearchers had previous knowledge of first-person research methods, with the exception of two, who were thus first trained in the method.

2.3 Procedure

The research procedure is shown in figure 1. All of the coresearchers started to sample their experience for 2-3 days, followed by a pause – which was done in order to gain a broader range of samples – and resumed for 2-3 days. In this phase we gained approximately 10 samples that we also investigated with microphenomenological interviews. If the coresearcher had a new mystical experience, the procedure was repeated, otherwise either sampling data acquired independently of our research was used or microphenomenological interviews were conducted for 5 moments before the mystical experience on psychedelics.

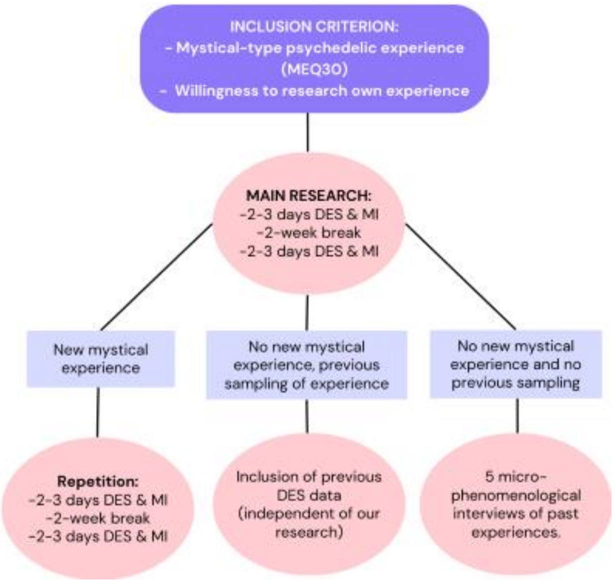

INCLUSION CRITERION:
- Mystical-type psychedelic experience (MEQ30)
- Willingness to research own experience
MAIN RESEARCH:
-2-3 days DES & MI
-2-week break
-2-3 days DES & MI
New mystical experience
No new mystical experience, previous sampling of experience
No new mystical experience and no previous sampling
Repetition:
-2-3 days DES & MI
-2-week break
-2-3 days DES & MI
Inclusion of previous DES data (independent of our research)
5 micro-phenomenological interviews of past experiences.

Figure 1: Outline of the research procedure.

2.4 Data analysis

The data from the questionnaire was analyzed in accordance with the questionnaire guidelines [14], with which we determined whether a person had a full or partial mystical experience.

The qualitative data was analyzed using the guidelines for coding and category grouping of data [15] as well as constructivist grounded theory guidelines [18]. The analysis was done iteratively – the data was examined multiple times, first separately for each individual and then comparatively. The result of the qualitative analysis was a codebook with 21 experiential categories. We then identified the occurrence of each category before and after the mystical-type psychedelic experience for each coresearcher and compared the findings.

3 RESULTS

3.1 Results on MEQ30

In the table below we present coresearcher's average score and the score on each of the four scales of MEQ30.

Table 2: MEQ30 scores

ID	*M*	Mystical	PM	TTS	Ineffability
1	4.06	0.79	0.73	0.87	1.00
2	3.70	0.80	0.57	0.67	0.93
3	2.67	0.45	1.00	0.13	0.80
4	2.76	0.52	0.67	0.40	0.80
5	2.53	0.39	0.60	0.57	0.80
6	2.17	0.35	0.23	0.77	0.60

Note. PM = positive mood, TTS = transcendence of time and space.

The first coresearcher had a full mystical experience, whereas all other coresearchers had a partial mystical experience, which means that their score didn't go above the threshold of 0.60 for one or two of four scales [14].

3.2 Qualitative data results

The data from the interviews was categorized in 21 first order categories, which were grouped in 6 second order categories and 2 third order categories. These were forefront experiences and background experiences. The first are quickly accessible and explicable experiences that tend to be in the focus and can be understood as reflective experiences [19]. The second are hardly accessible experiences that are more implicit and tend to be explicated after the reflective experiences during the course of the interview. These can be understood as prereflective experiences [19]. The forefront experiences were divided in four categories – focused experiences, presence in the moment, bodily feelings and control over experience. The background experiences were divided in two categories – the choice of experience and attitude towards experience. The division of all categories, as well as a summary of change in occurrence of a category before and after the mystical-type psychedelic experience across coresearchers is shown in table 3.

Table 3: Experience categories

Third order categories	Second order categories	First order categories	Summary of change
Forefront E	Focused E	Aimless focused E	3+
		Goal-focused E	2+ 1-
		Comprehensive unfocused E	1+
		Dispersed E	2+ 3-
	Presence in the moment	Pure perception	2+ 2-
		Sense of fusion	1-
		Hectic experience	1+ 2-
		Detachment feelings	2+ 1-
		Involvement in E	1+
		Impaired perception of surroundings	3+
		Involvement in the environment	2+ 2-
	Bodily feelings	Pleasant	5+
		Unpleasant	2+ 2-
	Control over E	Uncontrolled E	4+ 1-
		Intentional control over E	2+
		Unintentional control over E	1+ 1-
Background E	Choice of E	Identification with E	1-
		Distance towards E	2+
	Attitude towards E	"Who am I if not my thoughts?"	1+
		"Who speaks with my mouth?"	1+ 1-
		"What comes out of me is in accordance with who I am."	2+ 1-

Note. E = experience. + = increase in occurrence. - = decrease in occurrence. The number before the symbols + or - indicate the number of coresearchers that underwent certain change.

Individual level findings. Two of the coresearchers had a difference in the attitude towards their experience after the mystical-type psychedelic experience. The first one noticed a diminishment of thoughts and realized that there are other possible ways of being in the moment – without thoughts, present and more involved in the outside world. The other coresearcher realized that thoughts are not defining and started to take a more detached approach to experience. Interestingly, the samples we obtained only showed an increase in the control of experience, but not a diminishment of identification with experience. Another coresearcher started to have a sense of accordance with herself – what she was experiencing, was truer to who she was. It is important to note here that all coresearchers spoke about the importance of integration process after the mystical psychedelic experiences. They all said that changes don't occur by themselves, but are mediated by how they choose to integrate the findings in their everyday life. It is also important to note that during the interviews we realized that another factor that could highly contribute to the psychedelic experience and subsequent changes is prior knowledge and expectations about the psychedelic experience.

Comparative findings. Comparatively, after the mystical-type psychedelic experience there was an increase in aimless focused experience, pleasant bodily feelings and uncontrolled experience. On the first hand, the data seemed to show that the differences were highly individual and that no conclusions can be made. But after thorough inspection of possible interactions between categories, we found some interesting trends. An increase in pure perception seemed to be connected to the increase of aimless focused experience and involvement in the environment. An increase to hectic experience seemed to be connected with dispersed experience, while the increase in impaired perception of surroundings seemed to be connected to detachment feelings. An increase in uncontrolled experience seemed to be connected to the increase of the attitude that what the person experiences is truer to oneself, while also being connected to the increase or decrease in dispersed experience.

4 DISCUSSION

The aim of our research was to find out whether any differences in everyday experience occur following mystical-type psychedelic experiences and what those differences are, while also trying to find a common denominator of the observed changes.

The differences seemed to be individual and dependent on individual context, expectations and knowledge of psychedelic literature. All of the coresearchers stressed the importance of integration and said that what you choose to do with the experienced is more important than what you actually experience. Changes in their everyday life were thus also individual and dependent on their personal histories. A difference among individuals was also noticed in their prior knowledge of first-person research methods, as those who had previous knowledge tended to be able to go deeper in their prereflective, background experience and their attitude towards their experience. However, some similarities were observed, as there seemed to be an increase in uncontrolled experience, pleasant bodily feelings and aimless focused experience. Following mystical-type psychedelic experiences coresearchers tended to have less control on what they were doing, which was also followed by an increase in the feeling of being true to oneself. Most of the coresearchers also pointed out that the insights gained through mystical-type psychedelic experiences helps them gain knowledge about themselves – which can be understood as prereflective experience showing itself to the reflective experience [20]. This finding could be understood in light of priors being loosened by psychedelic experiences [21].

There are many limitations to our study – the context of use, prior knowledge about psychedelics and first-person research methods and number of previous psychedelic experiences were different among coresearchers which in itself can lead to interpersonal differences. The number of experience samples we obtained through our research were also relatively small and thus not thoroughly representative of the everyday experience. In the future, more coresearchers should be included and more data acquired. Coresearchers should also be screened for their previous knowledge, attitude and belief towards psychedelics and followed for an extended timeframe.

With our research, we shed light on the fact that little is known of phenomenology of everyday differences followed by mystical-type psychedelic experiences, which are thought to be important for personal development and breakthroughs that lead to change [2, 7]. We argue that first-person methods should be represented more frequently in the field of psychedelic research and that psychedelic apprenticeship, such as discussed in [22] should be considered in order to understand what drives the changes and how they can be understood.

ACKNOWLEDGMENTS

This paper provides a summary of the key findings from my master's thesis under the mentorship of prof. dr. Urban Kordeš, whom I thank for all the guidelines and support.

REFERENCES

[1] Chris Letheby, 2021. *Philosophy of psychedelics*. Oxford University Press.

[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] 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

[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] Willy Pedersen, Heith Copes and Liridona Gashi, 2021. Narratives of the mystical among users of psychedelics. *Acta Sociologica*, 64, 2, 230-246. DOI: https://doi.org/10.1177/0001699320980050

[6] Michiel van Elk and Eiko I. Fried, 2023. History repeating: A roadmap to address common problems in psychedelic science [preprint]. *PsyArXiv, March 10*. DOI: https://doi.org/10.31234/osf.io/ak6gx

[7] 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. DOI: https://doi.org/10.3389/fpsyt.2022.917199

[8] William James, 1987. *The varieties of religious experiences*. Penguin Books.

[9] Walter T. Stace, 1960. *Mysticism and philosophy*. Macmillan Publishers.

[10] 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

[11] Joost J. Breeksema, Alistair Niemeijer, Bouwe Kuin, Jolien Veraart, Eric Vermetten, Jeanine Kamphuis, Wim van den Brink and Rober Schoevers, 2023. Phenomenology and therapeutic potential of patient experiences during oral esketamine treatment for treatment-resistant depression: An interpretative phenomenological study. *Psychopharmacology*, 240, 1547-1560. DOI: https:/doi.org/10.1007/s00213-023-06388-6

[12] Piera Talin and Emilia Sanabria, 2017. Ayahuasca's entwined efficacy: An ethnographic study of ritual healing from 'addiction'. *International Journal of Drug Policy*, 44, 23-30. DOI: https//doi.org/10.1016/j.drugpo.2017.02.017

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

[14] Frederick S. Barrett, Matthew W. Johnson and Roland R. Griffiths, 2015. Validation of the Revised Mystical Experience Questionnaire in experimental sessions with psilocybin. *Journal of Psychopharmacology*, 29, 11, 1182-1190. DOI: https://doi.org/10.1177/0269881115609019

[15] Blaž Mesec, 1998. *Uvod v kvalitativno raziskovanje v socialnem delu*. Ljubljana: Visoka šola za socialno delo.

[16] Russell T. Hurlburt and Christopher L. Heavey, 2006. *Exploring inner experience: The Descriptive Experience Sampling method*. Amsterdam: John Benjamins Publishing Co.

[17] 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. DOI: https://doi.org/10.1007/s11097-006-9022-2

[18] Kathy Charmaz, 2014. *Constructing Grounded Theory* (2nd edition). SAGE Publications.

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

[20] Manesh Girn and Kalina Christoff, 2018. Expanding the scientific study of self-experience with psychedelics. *Journal of Consciousness Studies*, 25, 11-12, 131-154.

[21] Robin 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

[22] 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

Integrated Information Theory of Consciousness 3.0: Exploring Information and Causation on the Level of Individual Mechanisms

Tina Žerdoner Marinšek
tz7049@student.uni-lj.si
Faculty of Arts
University of Ljubljana
Ljubljana, Slovenija

ABSTRACT

The nature of consciousness has long defied precise scientific explanation, despite centuries of inquiry. Emerging as a prominent theoretical framework to unravel the mechanics of conscious awareness, the Integrated Information Theory (IIT) of Consciousness offers a unique approach. IIT's foundational principles are elucidated through a set of phenomenological axioms and ontological postulates, wherein axioms represent self-evident truths about consciousness, and postulates provide a structured framework to elucidate informational and causal aspects within physical systems. Central to IIT is the concept of information, defined as "differences that make a difference." It extends beyond mere data or signals, emphasizing the unique way a system's elements interact and influence one another. This paper explores the information and causation within the framework proposed in IIT 3.0 and on the level of individual mechanisms. It begins by introducing the advanced version of Integrated Information Theory (IIT 3.0). The paper then delves into Judea Pearl's theory of causation, outlining its key constructs. The primary aim of this study is to review information and causation on the level of individual mechanisms, within the context of IIT 3.0, and to bridge the findings with Judea Pearl's theory of causation.

KEYWORDS / KLJUČNE BESEDE

Theory of consciousness, consciousness, information, causation

1 INTRODUCTION

Consciousness, the very essence of subjective experience, has long eluded precise scientific explanation. Its enigmatic nature, tied to the brain's intricate processes, has spurred centuries of inquiry, yet a definitive understanding remains elusive. Among the diverse theoretical frameworks that seek to shed light on this enigmatic realm, the Integrated Information Theory (IIT) of Consciousness has emerged as a prominent theoretical framework poised to illuminate the mechanics behind conscious awareness [1]. Information integration theory characterizes consciousness both in terms of quantity and quality using axioms and postulates derived from the properties of phenomenal experience. Unlike conventional neuroscience approaches that start with neural mechanisms, IIT begins with the phenomenology of consciousness and seeks to understand its physical implementation.

Integrated Information Theory (IIT) presents its fundamental principles through a set of phenomenological axioms and ontological postulates. Axioms are self-evident truths about consciousness, while postulates are assumptions about the physical basis of consciousness, forming the mathematical framework of IIT. The central axioms include the existence of consciousness as an undeniable aspect of reality, its compositional nature where experiences consist of multiple aspects, its informativeness, meaning each experience is distinct from others, and its integration, where experiences cannot be broken down into non-interdependent components. Additionally, consciousness is characterized by exclusion, meaning only one experience exists at a time and within a particular spatio-temporal context.

These axioms are translated into formalized postulates that describe how physical mechanisms, such as neurons or logic gates, must be organized to generate conscious experiences based on phenomenology. Mechanisms are defined by their causal role. They have the capacity to influence and be influenced by other elements within the system. This means that they are involved in a web of causal relationships with other elements, and these relationships are crucial for understanding how the system functions. Each mechanism has a specific *cause-effect repertoire,* which represents the probability distribution of potential past and future states of the system as constrained by the mechanism's current state. A mechanism that specifies a maximally irreducible cause-effect repertoire, is called a *concept.*

The postulates provide a framework to define informational and causal properties of physical systems and intrinsic information as meaningful distinctions within a system. Integrated information is defined as the information specified by a whole that cannot be reduced to the sum of information specified by its parts. By applying these postulates at both the level of individual mechanisms and systems of mechanisms, IIT establishes a fundamental identity: an experience is a maximally irreducible conceptual structure (MICS), which is a constellation of concepts in qualia space. Qualia space is a mathematical representation of the space of all possible conscious states or experiences that a system can potentially have. A system that generates a MICS is associated with a specific conscious experience, and the properties of that experience are defined by the arrangement of concepts within the MICS. According to IIT, a MICS determines the quality of an experience, while integrated information quantifies its quantity [2, 3].

In IIT, information is a central notion [1], defined as *differences that make a difference* [2, 7] and it's not merely about data or signals, but about the distinct way in which a system's elements interact and influence each other. It captures the idea that information arises from the specific causal relationships among elements in a system. When these causal relationships result in distinct patterns of interaction and behaviour, they carry meaningful information about the system's state and its potential to affect and be affected by other elements. IIT's approach to defining information diverges by emphasizing the importance of perturbing a system to observe its responses. Such an approach is introduced in Judea Pearl's work [4, 5] and aligns closely with causation, as it emphasizes the dynamic relationships among elements – the latter necessarily being perturbed to assess the causation [6]. Judea Pearl's causal model is a framework for representing and analysing causal relationships in complex systems. It provides a formal and graphical way to model causation, allowing researchers to make causal inferences and understand the effects of interventions.

In this paper, I will delve into the concept of information and causation as proposed by Tononi [2, 6]. In what follows, I will first present the Integrated information theory of consciousness 3.0, an advanced version of the theory with several improvements compared to its predecessors. Subsequently, I will provide the key constructs of Judea Pearl's theory of causation. The paper's objective is to review causation and information on the level of individual mechanisms, within the context of IIT 3.0 and align the findings with Judea Pearl's theory of causation.

2 Integrated Information Theory 3.0

Understanding consciousness is a complex endeavour that requires both empirical investigation of neural correlates and a robust theoretical framework for explanation and prediction. Integrated Information Theory (IIT) is a comprehensive theoretical framework aimed at understanding consciousness. It addresses fundamental questions about why consciousness arises in certain brain systems but not others and how to assess consciousness in difficult cases, such as new-borns, animals, brain-damaged patients, and machines. IIT 3.0 attempts to mathematically characterize consciousness, focusing on both its quantity and quality. It starts with fundamental properties of consciousness phenomenology, translating them into postulates that outline the conditions for physical mechanisms (e.g., neurons) to account for consciousness phenomenology. This approach differs from traditional neuroscience, which usually starts with neural mechanisms and seeks to explain consciousness through behavioural reports [2, 7].

IIT 3.0 starts by introducing the axioms of the theory. The axioms serve as foundational principles that describe fundamental truths about consciousness itself. They establish the essential nature and properties of conscious experience. These axioms are self-evident and do not directly prescribe how consciousness arises from physical systems but rather define the characteristics of consciousness. The axioms are [2]:

1. **Existence:** Consciousness is an undeniable aspect of reality. "I experience, therefore I am."
2. **Composition:** Consciousness is structured and compositional. Each experience consists of multiple aspects in various combinations.
3. **Information:** Consciousness is informative. Each experience is distinct from other possible experiences, even if subtly so.
4. **Integration:** Consciousness is integrated. Each experience is strongly irreducible to non-interdependent components.
5. **Exclusion:** Consciousness is exclusive. At any given time, there is only one experience with definite borders.

IIT 3.0 then posits a set of postulates. The postulates are a set of assumptions that lay out the conditions under which a physical system, comprising mechanisms, can give rise to conscious experience. The postulates bridge the gap between the abstract axioms of consciousness and the concrete mechanisms within a physical system. They provide the framework to connect the nature of consciousness to the physical world. While axioms try to answer the question of what consciousness is and what are its essential properties, the postulates rather address the question of how consciousness can emerge from a physical system. The postulates are [2]:

1. **Existence:** Mechanisms in a state exist. A system comprises these mechanisms.
2. **Composition:** Elementary mechanisms can be combined to form more complex ones.

While mechanisms are the individual causal components within a system, a system of mechanisms represents the ensemble of these individual components working together to produce the system's behavior and conscious experiences.

The postulates of *information*, *integration*, and *exclusion* in IIT 3.0 are principles that apply to both individual mechanisms and systems of mechanisms [2].

Mechanisms:

1. **Information**: A mechanism contributes to consciousness if it specifies unique *"differences that make a difference"* within a system. It generates information by constraining the possible causes and effects in the system [2].
2. **Integration:** A mechanism contributes to consciousness when it specifies a cause-effect repertoire (information) that cannot be reduced to independent components. In other words, if you break down the mechanism into its constituent parts, the resulting information should not be the same as the information generated by the whole mechanism. Integration is assessed by partitioning the mechanism and measuring how this partitioning affects its cause-effect repertoire. The more interdependent the components, the higher the integration and the more relevant the mechanism is for consciousness [2].
3. **Exclusion:** IIT 3.0 posits that a mechanism can contribute to consciousness at most with one cause-effect repertoire, referred to as the maximally irreducible cause-effect repertoire (MICE). If a mechanism can be associated with a MICE, it

constitutes a concept. This principle ensures that mechanisms do not overlap in their contributions to consciousness, preventing redundancy [2].

Systems of mechanisms:

1. **Information:** IIT extends its principles to systems of mechanisms. A set of elements can exhibit consciousness only if its underlying mechanisms specify a *conceptual structure*. This conceptual structure defines meaningful distinctions or differences within the set. To visualize this, one can think of a conceptual space. In this space, each axis represents a possible past or future state of the set of elements. Within this conceptual space, there exists a constellation of points. Each point within this conceptual space represents a *concept*. These concepts are crucial because they specify the differences that make a difference within the set. In essence, they capture the essential distinctions or information relevant to the conscious experience. [2].
2. **Integration:** According to IIT 3.0, for a set of elements (which can represent neurons, brain regions, or any other relevant entities) to be conscious, it must exhibit a property called *strong integration*. This means that the elements within the set must work together in a way that cannot be broken down into independent components and that the overall functioning of the system cannot be understood by examining its components in isolation. To determine whether a set of elements exhibits strong integration, IIT 3.0 employs a method involving unidirectional cuts. This means that the system is divided into subsets in such a way that information flows in one direction only within each subset. The goal is to assess whether breaking the system down into subsets disrupts its integrated functioning. Strongly integrated systems are said to specify a conceptual structure [2].
3. **Exclusion:** IIT 3.0 posits that within a larger system or network of elements (which can represent neurons or other relevant components), only one specific subset or set of elements can be conscious. This means that consciousness is localized to particular subsets within a complex system. The basis for this exclusivity is the presence of a conceptual structure within a set of elements. This conceptual structure is associated with strong integration. Among all possible subsets or overlapping sets of elements within a larger system, only the one that specifies a conceptual structure that is maximally irreducible (MICS) to independent components can give rise to consciousness [2].

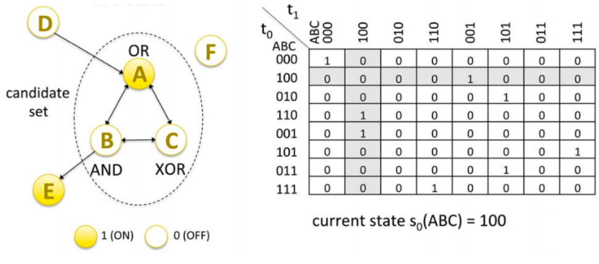
Integrated Information Theory (IIT) 3.0 then suggests a fundamental identity between the qualities of conscious experience and the informational and causal properties of physical systems. According to this concept, the maximally irreducible conceptual structure (MICS) is identical to the conscious experience that arises from that system, and the set of elements that generates it constitutes a complex. In essence, the way information is organized and integrated within a system directly corresponds to the nature of the conscious experience it generates [2].

3 Causation in IIT 3.0 on the level of individual mechanisms

In IIT 3.0, causation is brought to the forefront within the context of postulates and plays a crucial role in understanding how complex systems give rise to consciousness. In the first part of the theory, the focus is on individual mechanisms within a system. Mechanisms are entities, such as neurons in the brain or logic gates in a computer, that play a causal role in the system's behaviour. At the core of IIT 3.0's view of causation is the concept of the *cause-effect repertoire,* that is the probability distribution of potential past and future states of a system as constrained by a mechanism in its current state. Each mechanism within a system is considered to have a specific causal role. It can cause certain effects within the system and can be affected by specific causes. This cause-effect repertoire defines the set of possible causes and effects that a mechanism can be a part of within the system.

To generate information and contribute to consciousness, a mechanism must specify, as previously mentioned, *differences that make a difference* within the system. In other words, it should have selective and specific causes and effects. This means that the mechanism's causal relationships should not be random or arbitrary but should have a meaningful impact on the system's behaviour [2, 6].

IIT 3.0 proposes an approach to measure the causal power/generated information of a mechanism. Within this approach, it is important to understand the terms *background conditions,* and *candidate set.* The term *background conditions* refers to specific constraints imposed on a candidate set of elements within a system. These constraints are external and unchanging. When discussing these conditions, it's important to note that the past and current states of elements outside the candidate set are held constant at their real or observed values. In other words, these elements are not subject to change or manipulation as part of the analysis or investigation related to the candidate set. A *candidate set* refers to a specific group of elements that are being examined or analyzed. In the context of this description, the elements within the candidate set are subjected to perturbations, meaning they are deliberately altered to occupy all their potential states. This process is conducted to generate the Transition Probability Matrix (TPM) for the candidate set, which represents the probabilities of transitioning between different states of these elements [2, 6].



ABC	000	100	010	110	001	101	011	111
000	1	0	0	0	0	0	0	0
100	0	0	0	0	1	0	0	0
010	0	0	0	0	0	1	0	0
110	0	1	0	0	0	0	0	0
001	0	1	0	0	0	0	0	0
101	0	0	0	0	0	0	0	1
011	0	0	0	0	0	1	0	0
111	0	0	0	1	0	0	0	0

current state s_0(ABC) = 100

Figure 1: Mechanisms in state (candidate set) and transition probability matrix [2].

Information, defined as *the differences that make a difference* to a system from its intrinsic perspective, can be measured by examining how a mechanism in its current state affects the potential past and future states of the system [1, 2, 6]. The idea of perturbed elements and transition probability stands behind defining information in causal terms within the context of IIT 3.0 [1, 6, 11].

4 IIT 3.0: Causation on the level of individual mechanisms

In the literature [1, 11], the authors usually define information in causal terms by focusing on the concept that is behind this notion. The distinction between extrinsic and intrinsic information is commonly brought up to enlighten the importance of understanding the causal property of information in IIT 3.0 [1]. The aim of this section is to bring causation to the forefront and to investigate how causation and information are defined on the level of individual mechanisms, within the context of IIT 3.0. To establish the connection between information and causation, we need to delve deeper into the key constructs of IIT 3.0 and later, explore how they align with Judea Pearl's theory of causation.

At this point, I would like to introduce the concepts that are used to conceptualize causation and information in IIT 3.0. The first one is *cause repertoire* (probability distribution) *of a mechanism.* It refers to all the possible past states of the system that can lead to its current state, given the mechanism's specific causal interactions within the system. It represents the set of potential causes that have influenced the mechanism in the past. The second one is *unconstrained cause repertoire,* which represents all the potential past states of the system that could influence the mechanism, without any constraints imposed by the current state of the mechanism itself. In other words, it considers all possible causal interactions without the mechanism's selective influence. Similarly, the *effect repertoire* of a mechanism is a notion that refers to all the possible future states of the system that can occur as a result of the mechanism in a state. The *unconstrained effect repertoire* represents all the potential future states of the system that could result from various interactions, without the constraints imposed by the mechanism's current state. It considers all possible effects without the mechanism's selective influence [2].

All these concepts are important for measuring *cause information* (CI), *effect information* (EI), *and cause-effect information* (CEI). CI is a measure of the information about the past states of a system that is uniquely constrained and shaped by a specific mechanism within that system. It quantifies the difference between the cause repertoire (past states influenced by the mechanism) and the unconstrained cause repertoire (all possible past states without the mechanism's influence). In IIT, conscious experiences are associated with specific mechanisms that generate high CI values. By measuring CI, you can pinpoint which elements are more likely to be responsible for contributing to consciousness. While CI focuses on the information about past states of a system influenced by a specific mechanism, EI deals with information about the system's future states that are shaped and constrained by the same mechanism. EI is quantified as the difference between effect repertoire and unconstrained effect repertoire. CEI is essentially a combination of cause information (CI) and effect information (EI). It quantifies the information about both the past and the future states of a system that is uniquely constrained by a specific mechanism and it is measured as the minimum of CI and CE. CEI serves as a comprehensive measure for understanding causality. The higher the values of CEI, the more significant the contribution of a specific mechanism to consciousness [2].

5 Judea Pearl’s theory of causation

Judea Pearl's causal model provides a structured framework for studying causality, making it a valuable tool for researchers and practitioners seeking to uncover causal relationships in diverse contexts. His work is closely associated with the development of a causal model interventionist approach, which has had an impact on the way we understand and analyse causality in various domains [8].

In his book [5], Judea Pearl explores the different levels of causation in understanding and predicting events. At the first level, *association*, the focus is on identifying regularities in observations, such as predicting a rat's movement. This level deals with passive observations and collecting and analysing data to establish associations. Moving to the second level, *intervention*, the focus shifts to changing the world and asking questions like "What happens if we double the price of toothpaste?" Intervention requires knowledge beyond passive data and involves actively altering the environment. The top-level, *counterfactuals*, delves into understanding why things happen by exploring what would have occurred if circumstances were different. Counterfactual questions involve going back in time and considering alternate scenarios.

Pearl stated [4] that causal statements are often used in situations with uncertainty, where events tend to make consequences more likely but not certain. According to him, the theory of causation needs to provide a language to distinguish various shades of likelihood, which is crucial for accommodating such uncertainty.

At the core of his theory of causation, the notion of *intervention* is a crucial one, since it introduces the idea of deliberate and controlled alteration of a specific variable or set of variables in a causal system. It involves actively changing or manipulating a variable or system to observe how it affects other variables. The goal of interventions is to understand and establish causal relationships between variables [4,5]. Judea Pearl's theory of causation is fundamentally rooted in an interventionist approach to causality. Pearl's work on causality places a strong emphasis on interventions and their role in understanding causal relationships. The basic idea of the interventionist approach is that X is a cause of Y if only there is a possible intervention on X that will change Y or the probability distribution of Y [9]. Within this context, Pearl introduced the concept of a *do-operator (do(X))* to represent such interventions, where X represents the variable being manipulated. For example, in a medical study, if researchers want to determine how the administration of a new drug (variable X) influenced patient outcomes (variable Y), they may perform an intervention by administering the drug (do(X=1)) to a group of patients and comparing their outcomes to a control group that did not receive the drug (do(X=0)). By applying the do operator, we've effectively created two hypothetical scenarios:

- In Scenario 1 (do(X=1)), we observe a significantly higher rate of recovery compared to Scenario 2 (do(X=0)), which suggests that administering the drug has a causal effect on recovery.
- Conversely, if there is no significant difference in recovery rates between the two scenarios, it might suggest that administering the drug does not have a significant causal effect on recovery in this context.

The do operator allows us to isolate the effect of the drug intervention from other potential confounding factors. Pearl's work on interventions has provided a formal framework for causal inference and reasoning about causality in complex systems. It has also led to the development of causal graphical models, such as Bayesian networks and structural equation models, which are widely used in various fields to analyse and understand causal relationships in data [4, 5].

6 Intervening on the mechanism’s state leads to changes in its causal structure

Pearl's theory relies heavily on interventions or perturbations, where a variable is actively changed to assess its causal impact on an outcome. In IIT 3.0, intervention is also crucial but focuses on how mechanisms within a system shape the system's past and future states, thereby contributing to consciousness. The intervention here relates to how a mechanism's state influences the system's causal history and future. Intervention, in this context, involves deliberately altering the state of a mechanism within the candidate set. When you intervene on a mechanism's state, you are essentially introducing a change into the system. This change can propagate through the causal structure of the system, leading to alterations in how different mechanisms interact and influence each other. IIT 3.0 provides tools and metrics, such as CI, CE and CEI to quantify how intervening on a mechanism's state leads to changes in the causal structure. The higher the value of CEI, the more selective the cause-effect repertoire and thus more significant the contribution of a specific mechanism to consciousness.

While the primary goal of the observational approach in conditional probabilities is to describe and analyze existing data without active manipulation or intervention and to identify associations or correlations between variables, the interventional approach involves actively manipulating or intervening on a system or experiment to observe the causal relationships between variables [4,10].

Intervention, within the context of IIT 3.0, would mean changes in the Transition Probability Matrix (TPM) of a mechanism. Changing deliberately the mechanism’s state provokes alterations in the probability distribution of a mechanism’s past and future states and reveals its causal structure. The interventional approach helps to identify which causes and effects are maximally irreducible and thus, contribute more significantly to the conscious state of a system, within a specific spatio-temporal context.

In his works [4,5], Judea Pearl formalized also counterfactual statements within the framework of causal models, providing a rigorous and mathematical foundation for expressing and analysing counterfactual scenarios. Although not explicitly framed as counterfactuals [6], IIT 3.0 does involve the idea of an absence of a mechanism’s state within the concept of unconstrained cause/effect repertoire. Counterfactuals, as introduced by Judea Pearl, involve considering alternative scenarios and asking "what if" questions [4,5]. In the case of IIT 3.0, the comparison of a system with and without specific mechanisms in state is akin to a counterfactual inquiry because it assesses how the system's behaviour and properties might have been different if certain elements or mechanisms were absent or altered. As Tononi stated [6], this kind of inquiry should be subject to further explanation and redefinition.

7 Conclusion

In the paper, I set the stage by introducing the Integrated Information Theory of Consciousness 3.0 as an important framework poised to illuminate the mechanics behind conscious awareness. I briefly discuss the axiomatic part of the theory and outline the postulates, that are conditions for physical mechanisms to account for consciousness phenomenology. I then delve into the concept of causation within IIT 3.0, focusing on individual mechanisms. In this section I introduce the concept of the cause-effect repertoire, representing the potential past and future states of a system as constrained by a mechanism's current state. I explain how a mechanism must specify meaningful causes and effects to generate information and contribute to consciousness. In the same section of the paper, the importance of perturbing mechanisms within a candidate set to generate the Transition Probability Matrix (TPM) for causal analysis is introduced. In the following parts of the paper, I bridge the concepts of causation and information in IIT 3.0. To better understand the causation as conceptualised in IIT 3.0 on the level of individual mechanism, I introduce terms like cause repertoire, unconstrained cause repertoire, effect repertoire, and unconstrained effect repertoire. I discuss measures like Cause Information (CI), Effect Information (EI), and Cause-Effect Information (CEI) to quantify the contribution of mechanisms to consciousness. In the next section I introduce Judea Pearl's theory of causation, highlighting the levels of causation: association, intervention, and counterfactuals. I emphasize the role of interventions and the *do-operator* in Pearl's causal model. I affirm the importance of distinguishing various shades of likelihood in causal statements and accommodating uncertainty. In the last section, I draw the parallels between Pearl's focus on interventions and IIT 3.0's interventions on mechanism states. I suggest that IIT 3.0 involves the possibility to form a counterfactual analysis by comparing systems with and without specific mechanisms in state.

REFERENCES

[1] Lombardi, Olimpia, and Cristian López. “What Does ‘Information’ Mean in Integrated Information Theory?” Entropy 20, no. 12 (November 22, 2018): 894. DOI: https://doi.org/10.3390/e20120894

[2] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi, “From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0,” PLOS Computational Biology 10, no. 5 (May 8, 2014): e1003588, https://doi.org/10.1371/journal.pcbi.1003588.

[3] Kleiner, Johannes, and Sean Tull. “The Mathematical Structure of Integrated Information Theory.” Frontiers in Applied Mathematics and Statistics 6 (June 4, 2021). https://doi.org/10.3389/fams.2020.602973.Editor (Ed.).

[4] Pearl, Judea. “Causality”. Cambridge University Press, 2009.

[5] Pearl, Judea, and Dana Mackenzie. “The Book of Why: The New Science of Cause and Effect”. Basic Books, 2018.

[6] Antoine Suarez and Peter Adams, “Is Science Compatible with Free Will?”, Springer EBooks, 2013, https://doi.org/10.1007/978-1-4614-5212-6.

[7] Tononi, Giulio. “An information integration theory of consciousness.” BMC neurosciencevol. 5 42. 2 Nov. 2004, doi:10.1186/1471-2202-5-42

[8] James Woodward, “Critical notice: Causality by Judea Pearl” Economics and Philosophy 19, no. 2 (October 1, 2003): 321–40, https://doi.org/10.1017/s0266267103001184.

[9] James Woodward “Making Things Happen”, Oxford University Press EBooks, 2004, https://doi.org/10.1093/0195155270.001.0001.

[10] Albantakis, L. (2020). Integrated information theory. In *Routledge eBooks* (pp. 87–103). https://doi.org/10.4324/9781315205267-6

[11] Idan Efim Nemirovsky et al., “An Implementation of Integrated Information Theory in Resting-State FMRI,” Communications Biology 6, no. 1 (July 5, 2023), https://doi.org/10.1038/s42003-023-05063-y.

Indeks avtorjev / Author index

Kognitivna znanost

Cognitive Science

Uredniki • Editors:
Anka Slana Ozimič, Borut Trpin,
Toma Strle, Olga Markič